Methods-Related Excerpts from "Evalauting the ASVAB Armed Forces Qualfication Test for Differential Prediction" (Putka, Oppler, & Dahlke, 2022)
November 18, 2022

Chapter 3: Differential Prediction Analysis Methodology

Jeffrey A. Dahlke (HumRRO)

As we noted in Chapter 1, there are well-established practices for evaluating differential prediction; namely, using the Cleary framework for making contrasts among regression models and, more recently, computing effect sizes to characterize the magnitudes of differences in prediction. However, although these general analysis methodologies have been accepted within the I-O Psychology profession (SIOP, 2018), methodological challenges related to selection artifacts—commonly known as range restriction—remained largely unresolved. Selection artifacts occur when an observed sample is a non-random subset of a population, such that the systematic way in which the sample was selected limits the extent to which statistical results estimated from the sample accurately characterize relations among variables in the population.

In this research, we augmented the traditional Cleary framework's procedures to estimate models that account for selection artifacts and produce regression coefficients that more accurately characterize subgroup relations between AFQT scores and job-relevant criteria in the unrestricted applicant population. In this chapter, we describe our augmentations to the Cleary framework and summarize the methodology we used in our subsequent analyses.

Accounting for the Effects of Selection Artifacts on Differential Prediction Analyses

When a sample has been systematically selected in a way that makes it unrepresentative of its population, statistics computed using data from that sample will not generalize to the population of interest unless one takes steps to account for the ways in which the sample is unrepresentative. For example, when an organization conducts a study examining the criterion-related validity of scores from an assessment, analysts only have access to job performance data from incumbent employees; they have no performance data from applicants who were rejected during the selection process. If the organization fails to account for the ways in which their incumbent sample differs from the complete unrestricted applicant pool (namely, having less variance in predictor scores and having higher mean predictor scores than is typical of applicants), their incumbent data will give a misleading idea about the predictive value of their assessment. However, if they take advantage of tools and techniques to correct for this selection artifact (e.g., by applying range-restriction corrections or using modern missing data procedures, such as multiple imputation [MI] or full-information maximum likelihood [FIML] estimation), the organization can get a much better, more generalizable estimate of validity.

We considered a variety of ways to account for the known selection artifacts present in the criterion data we obtained from the Services. We explored options for using MI and FIML to account for missing criterion scores but determined that both methods posed limitations for the objectives of the present research. We ruled out FIML as an approach because, while FIML is highly effective at estimating models from data that exhibit missingness, it is not currently well-suited to evaluating differences among such models because the degrees of freedom for model contrasts are not readily available. We also ruled out MI because, although it would avoid the degrees-of-freedom problem we encountered with FIML, our experimentation with that method suggested it did not perform well in differential prediction analyses that exhibited the amounts of

missingness we anticipated in data from the Services. The challenges we faced when using MI were driven by the extremely high rates of missing data when we tried to use the applicant population to stabilize statistical estimates for individual occupations. Rather than stabilizing our estimates, MI had the opposite effect and made it likely we would erroneously detect differences between subgroups' regression lines.

Instead of using modern missing data methods such as FIML and MI, we relied on foundational principles of regression to account for selection artifacts. One of the interesting and useful characteristics of regression models fit using maximum-likelihood estimation is that a model's coefficients will not be biased by selection artifacts if all the variables involved in the selection process are included in the model. This property is known as invariance of coefficients under selection (Mulaik, 2009; pp. 408–414), and it is the basis for all range-restriction-correction formulas that researchers use to compute unrestricted estimates of correlations and Cohen's *d* values (e.g., Aitken, 1935; Lawley, 1944; Pearson, 1903; Thorndike, 1961; Wiernik & Dahlke, 2020).

Below, we offer a simple example illustrating the principles that underly our approach to controlling for selection artifacts in our differential prediction analyses.

Example Demonstrating the Principles Behind Our Analyses

As an example of our approach to controlling for selection artifacts, consider a case in which a researcher has a dataset containing a criterion variable called *Y* and two predictor variables called *X* and *Z*, and the researcher is primarily interested in the relation between *X* and *Y*. However, the researcher has incomplete observations for *Y* because data were only recorded for *Y* when cases had scores on *Z* that were above the mean, as might happen if *Y* were a job performance variable and if *Z* were used to make hiring decisions in a top-down selection system. Descriptive statistics for this example are presented in Table 3.1 for variables that have unrestricted means of 0, unrestricted standard deviations of 1, and unrestricted intercorrelations of .5. The restricted estimates in Table 3.1 represent descriptive statistics for the complete observed cases available for analysis (i.e., those for which *Y* was recorded) using listwise deletion. Table 3.1 also contains a modified version of *Z* labeled Z_{res} – this variable represents the unique variance of *Z* that is not shared with *X* after accounting for the relation between *X* and *Z* in the unrestricted data. The correlation between *X* and Z_{res} in the restricted data provides information about the severity with which selection artifacts have impacted the covariate, and this correlation is useful when using Z_{res} to control for selection artifacts in regression models.

Table 3.1. Descriptive Statistics for Range-Restriction Example

		Descriptive Statistics		Correlations		
Data Type	Variable	*M*	*SD*	*Y*	*X*	*Z*
Unrestricted	*Y*	0.00	1.00	---	---	---
	X	0.00	1.00	0.50	---	---
	Z	0.00	1.00	0.50	0.50	---
	Z_{res}	0.00	0.87	0.29	0.00	0.87
Restricted	*Y*	0.40	0.92	---	---	---
	X	0.40	0.92	0.41	---	---
	Z	0.80	0.60	0.33	0.33	---
	Z_{res}	0.60	0.63	0.02	-0.42	0.72

Estimating Regression Coefficients that are not Biased by Selection Artifacts

Table 3.2 shows regression coefficients for the example described above, with separate results reported for the unrestricted data that are unavailable to the researcher (these results represent the unobserved "truth" about how the variables relate) and the restricted data that *are* available to the researcher. The first row of Table 3.2 shows the unrestricted *X*-*Y* relationship the researcher is most interested in, while the second row shows the coefficients the researcher would get if they tested that relationship using their restricted data. Comparing these unrestricted and restricted coefficients makes it clear that the researcher will do a poor job of characterizing the true *X*-*Y* relationship unless they do something to account for the selection artifacts, as both the intercept and the slope are misestimated to a non-trivial degree.

By comparison, the researcher would have no problem estimating the *Z*-*Y* relationship (see rows 3-4 of Table 3.2). As *Z* was the sole selection variable that caused the missing values for *Y*, its inclusion in a model that predicts *Y* leads to unbiased estimation of the *Z*-*Y* relationship due to the invariance of coefficients under selection. By extension, the researcher would have no difficulty estimating coefficients that relate *X* and *Z* to *Y* in a multiple regression model (see rows 5-6 of Table 3.2). However, the coefficients from this type of multiple regression model are not appropriate for characterizing the direct bivariate *X*-*Y* relationship, as the model accounts for the shared variance between *X* and *Z* when estimating the coefficients and the *X* coefficient no longer describes the unique relationship between *X* and *Y*.

Ideally, the researcher would use a model that can account for *Z*'s influence in the selection process while attributing any variance shared between *X* and *Z* to *X* alone. Fortunately, this is possible to do, and it requires only a small amount of pre-processing to accomplish. To ensure that *Z* does not "steal" any variance from *X* in the regression model, the researcher can create a residualized version of *Z* that represents only the variance that is independent of *X*. This involves regressing *Z* on *X* using unrestricted predictor data, computing predicted/fitted Z estimates, and subtracting those predicted estimates from *Z* to obtain a vector of residual scores; these residuals are the Z_{res} variable we introduced earlier in Table 3.1. Including Z_{res} as a covariate in a model with *X* to predict *Y* will allow the researcher to control for the biasing effects of selection artifacts while obtaining coefficients that accurately reflect the direct relation between *X* and *Y* (see the "*X* & Z_{res}" results in rows 7-8 of Table 3.2 and compare them to row 1).

Table 3.2. Regression Coefficients for Range-Restriction Example

Predictor(s) in Model	Data Type	Regression Coefficients			
		Intercept	*X*	*Z*	Z_{res}
X	Unrestricted	0.00	0.50	---	---
	Restricted	0.24	0.41	---	---
Z	Unrestricted	0.00	---	0.50	---
	Restricted	0.00	---	0.50	---
X & *Z*	Unrestricted	0.00	0.33	0.33	---
	Restricted	0.00	0.33	0.33	---
X & Z_{res}	Unrestricted	0.00	0.50	---	0.33
	Restricted	0.00	0.50	---	0.33

The coefficients from the example in Table 3.2 gave a simple illustration of how one can include residualized covariates in a model to debias estimates of the coefficients one is most interested in. The principles from that example generalize to more complex regression models with multiple predictors of interest (such as those used in the Cleary framework), and they also generalize to logistic regression models.

Estimating the Unrestricted Variance of a Criterion Variable

After one has included covariates in a regression model to debias the estimates of coefficients, one can use those coefficients to estimate the unrestricted variance of a restricted criterion variable. In the context of our differential prediction analyses, this is valuable for estimating a more appropriate unrestricted scaling factor for the d_{Mod} effect sizes we discuss later. The process of obtaining this unrestricted variance estimate is based in the Pearson-Aitken-Lawley selection theorem (Aitken, 1935; Lawley, 1944; Pearson, 1903), which defines how to correct for (or induce) selection artifacts in a covariance matrix. To estimate the unrestricted variance of a continuous variable that is merely correlated with a selection variable, but was not involved in the selection process, one can use the following formula:

$$\widehat{\sigma}^2_{Y_{Unrestricted}} = \sigma^2_{Y_{Restricted}} - \sigma^2_{\hat{Y}_{Restricted}} + \sigma^2_{\hat{Y}_{Unrestricted}}$$

where $\widehat{\sigma}^2_{Y_{Unrestricted}}$ is the estimated unrestricted variance of *Y* (the variable that has been indirectly impacted by selection), $\sigma^2_{Y_{Restricted}}$ is the observed variance of *Y* in the restricted data, $\sigma^2_{\hat{Y}_{Restricted}}$ is the predicted variance of *Y* in the restricted data based on a regression model, and $\sigma^2_{\hat{Y}_{Unrestricted}}$ is the predicted variance of *Y* in the unrestricted data. This additive variance approach is only appropriate for continuous variables; we describe an approach for correcting the variance of a binary variable later in this chapter.

We demonstrate this in Table 3.3 using the same models as we presented earlier in Table 3.2. The first row of Table 3.3 shows the result our hypothetical researcher is most interested in: The relations between *X* and *Y* in the unrestricted population. However, the researcher only has restricted data for *Y*, so their analysis of *X* and *Y* would yield the estimates in the second row; these range-restricted results substantially underestimate the unrestricted variance of *Y*.

Once the researcher includes *Z* in their models (see Table 3.3; rows 3-8), they can estimate the unrestricted variance of *Y* without bias (i.e., they correctly recover the variance of 1.00). By including both *X* and Z_{res} as predictors in a model, the coefficient for *X* estimated from restricted data becomes an unbiased estimate of the unrestricted population parameter (see Table 3.2); this, combined with the fact that *X* and Z_{res} are uncorrelated in the unrestricted population, means that one can estimate the proportion of unique unrestricted criterion variance explained by *X*.

The methods we used in this example to residualize an auxiliary predictor and obtain estimates of regression coefficients and criterion variances that are unbiased by selection artifacts are central to the methods we used in our differential prediction analyses. We used these methods to debias the coefficients in our regression models whenever a variable (or set of variables) other than the predictor(s) of interest carried information about selection artifacts.

Table 3.3. Variance Estimates for Range-Restriction Example

		Variance							
		Analysis Data				Unrestricted Data*			
Predictors	Data Type	Y_{Obs}	$\hat{Y}$	$\hat{Y}_X$	$\hat{Y}_Z$	$\hat{Y}$	$\hat{Y}_X$	$\hat{Y}_Z$	Y_{Est}
X	Unrestricted	1.00	0.25	0.25	---	0.25	0.25	---	1.00
	Restricted	0.84	0.14	0.14	---	0.16	0.16	---	0.93
Z	Unrestricted	1.00	0.25	---	0.25	0.25	---	0.25	1.00
	Restricted	0.84	0.09	---	0.09	0.25	---	0.25	1.00
X & *Z*	Unrestricted	1.00	0.33	*0.11*	*0.11*	0.33	*0.11*	*0.11*	1.00
	Restricted	0.84	0.17	*0.09*	*0.04*	0.33	*0.11*	*0.11*	1.00
X & Z_{res}	Unrestricted	1.00	0.33	0.25	---	0.33	0.25	---	1.00
	Restricted	0.84	0.17	0.21	---	0.33	0.25	---	1.00

Note. *The unrestricted data is the same as the analysis data for the "Unrestricted" data type. Y_{Obs} = observed criterion scores available for use in an analysis. $\hat{Y}$ = predicted criterion scores using all predictors involved in a regression model. $\hat{Y}_X$ = predicted criterion scores using only predictor *X*. $\hat{Y}_Z$ = predicted criterion scores using only predictor *Z*. $\mathrm{Var}(Y_{Est}) = \mathrm{Var}(Y_{Obs}) - \mathrm{Var}(\hat{Y}_{Analysis\ Data}) + \mathrm{Var}(\hat{Y}_{Unrestricted\ Data})$. Values in italics would not be useful to a researcher because they represent partial estimates of variance that fail to account for the shared variance of predictors. Coefficients are not shown for Z_{res} because they are not of substantive interest when a residualized covariate is included in a model.

Procedures for Fitting Differential Prediction Regression Models

We used the principles described above to account for selection artifacts in our models and reduce selection's biasing effects on our regression coefficients. Below, we describe (a) the pre-processing steps we used to construct residualized covariates that can help account for selection artifacts, (b) the regression models we fit to our data, (c) additional steps we took to estimate versions of main-effect-only models that better reflect unrestricted relations, and (d) the model contrasts we used to detect intercept differences and slope differences.

Pre-Processing Procedure for Residualized Covariates

Many of the analyses we performed required multiple covariates to account for non-AFQT selection effects, and we combined these covariates into a composite before including them in our regression models.[1] We had initially planned to enter the individual covariates into our models, but we found this led to instability and overfitting. Some analyses required as many as four covariates, which would have greatly increased the complexity of our models by adding eight total predictors (four for main effects and four for covariate-subgroup interaction effects). Rather than limit our analyses to samples that had sufficient data to accommodate these high-complexity models, we developed a strategy for constructing composite covariates that would allow us to limit the complexity of our models while still accounting for the influence of variables

[1] The residualized covariate composites used in predictive bias analyses for each Service are documented in the substantive analysis chapters that follow. These composites reflected one or more ASVAB "line scores" (i.e., composites of ASVAB subtest score used by Services to inform occupation-specific assignment/qualification), or a composite of ASVAB subtests that do not contribute to the AFQT composite.

other than AFQT scores that contribute to selection artifacts via formal selection and classification processes in the Services.

Before computing a composite covariate, we had to put each of the covariates on a standardized scale so they could be combined in a meaningful way. We established this standardized scaling by computing the mean and standard deviation of each j^{th} covariate for the N applicants in the unrestricted population:

$$M_j = \frac{\sum_{i=1}^{N} Covariate_{ij}}{N}$$

$$SD_j = \sqrt{\frac{\sum_{i=1}^{N}\left(Covariate_{ij} - M_j\right)^2}{N}}$$

We then used these means and *SD*s to standardize the covariates in both the restricted and unrestricted data sets, and we averaged the resulting standardized scores across k covariates to arrive at a composite covariate we will call Z. For each i^{th} case in the restricted and unrestricted data sets, we computed Z as:

$$Z_i = \frac{\left(\sum_{j=1}^{k} \frac{Covariate_{ij} - M_j}{SD_j}\right)}{k}$$

After defining the composite covariate, we proceeded with the residualization process that would prepare the covariate and its interaction with the subgroup dummy variable for inclusion in our models. At this point in our process, we constructed product terms representing the predictor × subgroup interaction for scores on both the AFQT and Z:

$$Interaction_{AFQT} = AFQT \times Subgroup$$

$$Interaction_Z = Z \times Subgroup$$

Recall that the goal of this residualization procedure is to force covariates to have correlations of exactly zero with each of the primary predictors in the unrestricted population while preserving the covariates' unique variance that can help to account for selection artifacts. Residualizing the covariate and its interaction variable allowed us to account for both the main effect of the covariate and any subgroup slope differences that can be uniquely attributed to the covariate. We fit the following linear regression models predicting Z and its interaction variable from AFQT scores, subgroup membership, and AFQT scores' interaction variable in the unrestricted predictor data:

$$Z = b_{A0} + b_{A1} \times AFQT + b_{A2} \times Subgroup + b_{A3} \times Interaction_{AFQT} + e$$

$$Interaction_Z = b_{B0} + b_{B1} \times AFQT + b_{B2} \times Subgroup + b_{B3} \times Interaction_{AFQT} + e$$

where the capitalized letter subscripts on regression coefficients differentiate the models and indicate that they were only used in pre-processing; we use numeric subscripts to differentiate the models we tested for substantive analyses.

We then used these regression models to compute fitted/predicted values for *Z* and its interaction variable using restricted predictor data:

$$\hat{Z} = b_{A0} + b_{A1} \times AFQT + b_{A2} \times Subgroup + b_{A3} \times Interaction_{AFQT}$$

$$\widehat{Interaction}_Z = b_{B0} + b_{B1} \times AFQT + b_{B2} \times Subgroup + b_{B3} \times Interaction_{AFQT}$$

Finally, we computed residualized values for *Z* and its interaction variable using restricted predictor data:

$$Z_{Res} = Z - \hat{Z}$$

$$Interaction_{Z\_Res} = Interaction_Z - \widehat{Interaction}_Z$$

After constructing the Z_{Res} and $Interaction_{Z\_Res}$ variables, our data were ready to fit regression models and use the logic of the Cleary framework to conduct inferential tests of differential prediction.

Regression Models for Inferential Tests of Differential Prediction

We used a slightly modified version of the Cleary framework to test for differential prediction while accounting for the influence of *Z* in explaining selection artifacts. We fit the following three regression models:

Model 1: $Y = b_{10} + b_{11} \times Z_{Res} + b_{12} \times Interaction_{Z\_Res} + b_{13} \times AFQT + e$

Model 2: $Y = b_{20} + b_{11} \times Z_{Res} + b_{22} \times Interaction_{Z\_Res} + b_{23} \times AFQT + b_{24} \times Subgroup + e$

Model 3: $Y = b_{30} + b_{31} \times Z_{Res} + b_{32} \times Interaction_{Z\_Res} + b_{33} \times AFQT + b_{34} \times Subgroup + b_{35} \times Interaction_{AFQT} + e$

These models differ from a typical set of differential prediction models by the inclusion of Z_{Res} and $Interaction_{Z\_Res}$.

Although the models above are useful for evaluating model contrasts, the coefficients from Models 1 and 2 are not necessarily the best representation of unrestricted main effects when the AFQT interaction term is omitted. To offer a better characterization of main effects in the unrestricted data, we estimated alternate versions of Models 1 and 2 that explain the same amount of variance as Model 3 but partition the variance differently across predictors to produce debiased estimates of coefficients for main effects. We present the methods for estimating these alternate versions of Models 1 and 2 in the following subsection.

Procedures for Estimating Versions of Main-Effect Models with Coefficients that Generalize to Unrestricted Data

We used residualization procedures to debias the main effect estimates from Models 1 and 2 that were very similar to the procedures we used to prepare the Z_{Res} and $Interaction_{Z\_Res}$ variables. To debias the AFQT main effect coefficient from Model 1, we created versions of the $Subgroup$ and $Interaction_{AFQT}$ variables that were uncorrelated with AFQT scores in the unrestricted population. Likewise, to debias the AFQT and Subgroup main effects from Model 2, we created a version of the $Interaction_{AFQT}$ variable that was uncorrelated with AFQT scores and the subgroup-membership dummy variable.

The first step in preparing these residualized predictors was to fit the following three linear regression models using unrestricted applicant data:

$$Subgroup = b_{C0} + b_{C1} \times AFQT + e$$

$$Interaction_{AFQT} = b_{D0} + b_{D1} \times AFQT + e$$

$$Interaction_{AFQT'} = b_{E0} + b_{E1} \times AFQT + b_{E2} \times Subgroup + e$$

We then used those regression models to generate the following fitted/predicted values for the restricted predictor data:

$$\widehat{Subgroup} = b_{C0} + b_{C1} \times AFQT$$

$$\widehat{Interaction}_{AFQT} = b_{D0} + b_{D1} \times AFQT$$

$$\widehat{Interaction}_{AFQT'} = b_{E0} + b_{E1} \times AFQT + b_{E2} \times Subgroup$$

We used the fitted/predicted values to create residualized versions of the predictors for inclusion in Model 1′ and Model 2′ (Model 1 "prime" and Model 2 "prime"):

$$Subgroup_{Res} = Subgroup - \widehat{Subgroup}$$

$$Interaction_{AFQT\_Res} = Interaction_{AFQT} - \widehat{Interaction}_{AFQT}$$

$$Interaction_{AFQT\_Res'} = Interaction_{AFQT} - \widehat{Interaction}_{AFQT'}$$

Although the subgroup dummy variable is dichotomous, we used linear regression rather than logistic regression to predict it because our goal was simply to control for the correlation between variables. Using linear regression allowed us to "partial out" the variance shared with AFQT scores and produce residuals that could be used effectively within our modeling effort.

After preparing residualized versions of the $Subgroup$ and $Interaction_{AFQT}$ variables, we fit the following reformulations of Models 1 and 2 to obtain debiased estimates of main effects that account for AFQT-related selection artifacts:

Model 1′: $Y = b_{1'0} + b_{1'1} \times Z_{Res} + b_{1'2} \times Interaction_{Z\_Res} + b_{1'3} \times AFQT + b_{1'4} \times Subgroup_{Res} + b_{1'5} \times Interaction_{AFQT\_Res} + e$

Model 2′: $Y = b_{2'0} + b_{2'1} \times Z_{Res} + b_{2'2} \times Interaction_{Z\_Res} + b_{2'3} \times AFQT + b_{2'4} \times Subgroup + b_{2'5} \times Interaction_{AFQT\_Res'} + e$

We used Models 1′ and 2′ exclusively to obtain better estimates of coefficients from models that only involve main effects. These prime models have the same explanatory value and model fit statistics as Model 3, so they are not suitable for making model comparisons; we used the original versions of Model 1 and 2 in all model-comparison analyses.

Model Fit and Model Comparisons

We evaluated model fit and changes in model fit using conventional approaches for linear and logistic regression analyses. For linear models, we used *F* tests to evaluate the statistical significance of variance explained by the models and differences in variance explained between

models. For logistic regressions, we used deviance tests based on chi-squared distributions to evaluate the statistical significance of model fit and differences in model fit.

We used the regression models described above to evaluate model contrasts and determine whether and how subgroups' regression lines differed. We tested the following contrasts:

- **Omnibus Contrast:** Model 3 vs. Model 1
    - This model comparison establishes whether subgroups' regression lines differ in any way.
    - If it was not significant, we concluded that subgroups' regression lines do not differ.
    - If it was significant, we proceeded to test for slope differences.
- **Slope Contrast:** Model 3 vs. Model 2
    - This model comparison establishes whether subgroups' regression lines have different slopes.
    - If it was significant, we stopped our interpretation and concluded that there are slope differences.
    - If it was not significant, we proceeded to test for intercept differences.
- **Intercept Contrast:** Model 2 vs. Model 1
    - This model comparison establishes whether subgroups' regression lines have different intercepts.
    - If it was significant, we concluded that there are intercept differences.
    - If it was not significant, we concluded that subgroups' regression lines do not differ.

The outcome of this set of contrasts was a collection of model-difference tests (i.e., statistics for *F* tests or deviance tests) and a categorical decision about whether groups' lines were the same, had intercept differences, or had slope differences.

We also computed R^2 estimates to express model fit as an effect size, and we used these statistics to compute ΔR^2 estimates that represent the differences in model fit for each of the model contrasts. Linear regression analyses naturally lend themselves to use of R^2 statistics to characterize model fit in terms of an effect size, whereas logistic regression analyses do not; linear regression explicitly maximizes R^2 in a closed-form solution when fitting a model, while logistic regression uses iterative estimation to maximize a likelihood function.

Models that predict binary outcomes require additional steps to represent their fit in terms of a "pseudo" R^2 metric. There are many formulations of pseudo R^2 statistics available (e.g., Nagelkerke, McFadden, Efron, Cox and Snell), but we used a simple approach that we argue is the most conceptually similar to the R^2 estimates form linear models. Linear R^2 estimates are squared correlations between observed outcomes and model-predicted outcomes. Similarly, the pseudo R^2 estimates we used for the logistic models are squared point-biserial correlations between observed dichotomous outcomes and model-predicted probabilities of those outcomes. This approach to computing pseudo R^2 estimates is based on relevant model output, simple to

interpret, and consistent with our linear regression analyses. However, since R^2 is not a native concept for logistic regression analyses, it is possible for ΔR^2 estimates for model contrasts to be negative; this never happens with linear regression models because adding predictors can never worsen the maximum-likelihood R^2 (a larger model can be worse in terms of its shrunken R^2 estimate or a cross-validated R^2 estimate, but never in terms of its maximum-likelihood R^2).

Procedures for Computing d_{Mod} Effect-Size Estimates

As mentioned in Chapter 1, we supplemented our Cleary-based regression analyses with d_{Mod} effect-size estimates that characterize the standardized average magnitudes of differences between subgroup regression lines (Dahlke & Sackett, 2018; Nye & Sackett, 2016).

The d_{Mod} effect sizes are based on differences between predictions generated by the regression equations for the focal and referent subgroup, using the focal subgroup's predictor score distribution as the input to both regression equations. The mean differences between these distributions of predictions get standardized by dividing them by the standard deviation of observed criterion scores in the referent group. Using the focal predictor score distribution to generate predictions and using the referent group's criterion standard deviation for scale allows d_{Mod} to address two important questions:

1. How different are predictions for the focal group if we use a model that is specific to the focal group versus a model that characterizes predictor-criterion relations in the referent group?

2. How large are these differences relative to the amount of variability we observed in the referent group?

d_{Mod} effect sizes can be interpreted much like Cohen's *d*, Hedges' *g*, or Glass' Δ effect sizes. The difference between those metrics and d_{Mod} effect sizes is that *d*, *g*, and Δ are all computed from *observed* scores, whereas d_{Mod} is computed from *predicted* scores. As an example of interpretation, a d_{Mod} effect of .5 would indicate that, on average, the focal group's performance is overpredicted by half of a referent group standard deviation (traditionally interpreted as a medium effect), whereas a d_{Mod} effect of -.2 would indicate that, on average, the focal group's performance is underpredicted by one fifth of a referent group standard deviation (traditionally interpreted as a small effect).

The methods we used to estimate differences between groups' regression lines and the referent group's unrestricted criterion standard deviation are described in the following subsections.

Estimates of Differences in Prediction

The d_{Mod} effect sizes all involve dividing some difference in prediction by a scaling factor, and the numerator of that ratio is a function of subgroup regression equations. We used coefficients from Model 3 to construct the subgroup equations, and we used the focal group's applicant AFQT distribution to compute predictions as follows:

$$\hat{Y}_{Focal} = (b_{30} + b_{34}) + (b_{33} + b_{35}) \times AFQT_{Focal}$$

$$\hat{Y}^{*}_{Referent} = b_{30} + b_{33} \times AFQT_{Focal}$$

where $b_{30} + b_{34}$ gives the intercept for the focal group that accounts for the main effect of group membership, and $b_{33} + b_{35}$ gives the slope for the focal group that accounts for the interaction between AFQT scores and group membership. We represent $\hat{Y}^*_{Referent}$ with an asterisk as a reminder that, although we computed these predictions using coefficients from the referent group's regression line, the predictions are a function of the focal group's AFQT score distribution. It is important to emphasize that these predictions are based on the focal group's applicant distribution of AFQT scores; this, combined with the procedures we used to mitigate the impact of selection artifacts on our estimates of regression coefficients, means that $\hat{Y}_{Focal}$ and $\hat{Y}^*_{Referent}$ represent estimates of predictions from the focal group's unrestricted applicant population.

For logistic regression models, the initial $\hat{Y}$ values were in the logit metric. We used the following transformation to convert these estimates to probabilities:

$$\hat{Y}_{Probability} = \frac{e^{\hat{Y}_{Logit}}}{1 + e^{\hat{Y}_{Logit}}}$$

where $e^{\hat{Y}_{Logit}}$ is an intermediate conversion from the logit metric to the odds metric.

We used the unrestricted predictor distribution from the focal subgroup to compute conditional differences in prediction between the subgroups' regression lines:

$$\hat{Y}_{Difference} = \hat{Y}^*_{Referent} - \hat{Y}_{Focal}$$

The distribution of $\hat{Y}_{Difference}$ values provides the necessary data to compute estimates of average differences in prediction that can then be scaled into a standardized metric using an estimate of the referent group's unrestricted criterion standard deviation.

Estimates of Unrestricted Criterion Variances

After computing the distribution of differences in prediction we described in the previous subsection, the other input needed to compute d_{Mod} effect sizes is a scaling factor that reflects the standard deviation of criterion scores in the referent group. The most informative scaling factor is one that gives an *unrestricted* estimate of the referent group's criterion standard deviation so the resulting effect sizes properly quantify standardized differences in prediction in the applicant population.

Regardless of whether a criterion is continuous or dichotomous, it is possible to estimate the unrestricted variance of the criterion by capitalizing on principles of regression. We describe the procedures for doing so in the following subsections on linear and logistic regressions. Regardless of the type of regression model, the process for estimating the unrestricted variance of criterion scores in the referent group involves knowledge about the referent group's distribution of unrestricted predicted values. We used coefficients from Model 3 to compute predictions for the referent group as follows:

$$\hat{Y}_{Referent} = b_{30} + b_{31} \times Z_{Res\_Referent} + b_{33} \times AFQT_{Referent}$$

where $Z_{Res\_Referent}$ and $AFQT_{Referent}$ are unrestricted score distributions. For logistic models, we converted $\hat{Y}_{Referent}$ values from the logit metric to the probability metric using the transformation presented above.

Linear Regression Models

For linear regression models, we can estimate the unrestricted standard deviation of the criterion in the referent group using three variance terms that are simple to obtain: The variance of observed criterion scores in the restricted data ($SD^2_{Y_{Restricted\_Restricted}}$), the variance of predicted criterion scores in the restricted data computed using Model 3 ($SD^2_{\hat{Y}_{Restricted\_Restricted}}$), and the variance of predicted criterion scores in the unrestricted data computed using Model 3 ($SD^2_{\hat{Y}_{Restricted\_Unrestricted}}$). The terms can be combined as follows:

$$\widehat{SD}_{Y_{Referent\_Unrestricted}} = \sqrt{SD^2_{Y_{Referent\_Restricted}} - SD^2_{\hat{Y}Referent\_Restricted} + SD^2_{\hat{Y}_{Referent\_Unrestricted}}}$$

This formula is equivalent to the portion of the Pearson-Aitken-Lawley selection theorem that specifies how to estimate the unrestricted standard deviation of a variable for which unrestricted data are unavailable (Aitken, 1935; Lawley, 1944; Pearson, 1903). It relies on linear regression's assumption of homoscedasticity of residuals, which requires that the unexplained variance of criterion scores is consistent across the distribution of predictor scores. In other words, the overall variance of residual criterion scores is assumed to be equal to the *conditional* variance of residual criterion scores across the range of predictor scores. Additionally, this formula assumes that the variance of residual criterion scores is invariant to selection (i.e., it is the same when it is computed using restricted data or unrestricted data), which is true when the model includes all the variables that contributed to the selection process that produced the restricted sample. Thus, by subtracting the variance of restricted predicted values from the variance of restricted observed values, we get the residual variance of criterion scores (an invariant term). Then, by adding the variance of unrestricted predicted values, we get an estimate of the total variance of unrestricted criterion scores, which is the sum of explained and unexplained variance in unrestricted criterion data.

Logistic Regression Models

For logistic regression models, the process of estimating the unrestricted variance of a criterion is quite different. Unlike linear regression, logistic regression does not involve an assumption about the homoscedasticity of residuals, so the strategy of adding and subtracting variance terms does not generalize from continuous criteria to binary criteria. Instead, we can rely on important characteristics of (a) standard deviations of binary variables and (b) means of predicted criterion scores.

The mean of a 0/1 binary variable is the proportion of cases that have a score of 1 (e.g., the proportion of people who passed a class or the proportion of people who attritted), and this proportion (p) is the fundamental input for computing the standard deviation of the variable. The proportion of cases who have a score of 0 on the binary variable is $q = 1 - p$, and the product of $p \times q$ is the maximum-likelihood estimate of the binary variable's variance. This relationship between the mean and variance of a binary variable means that we only need to estimate the unrestricted mean of a binary criterion to arrive at an estimate of the unrestricted variance. Fortunately, this is simple to do: Because of the way we have formulated our regression models, we can use Model 3 to compute probability-metric predictions for the unrestricted data, and the mean of these probabilities should closely approximate the unrestricted mean of the criterion. With a reasonable estimate of the unrestricted mean in hand, it is straightforward to compute an estimate of the unrestricted standard deviation using the following equation:

$$\widehat{SD}_{Y_{Referent\_Unrestricted}} = \sqrt{\overline{\hat{Y}_{Referent\_Unrestricted}} \times \left(1 - \overline{\hat{Y}_{Referent\_Unrestricted}}\right)}$$

This variance-estimation approach for logistic regression models is very different from the approach for linear regression models, but is equally supported by the mechanics of the underlying modeling procedure.

Estimates of d_{Mod} Effects

We used the distribution of conditional differences in prediction represented by $\hat{Y}_{Difference}$ and the scaling factor represented by $\widehat{SD}_{Y_{Referent\_Unrestricted}}$ to compute d_{Mod} effect sizes for each subgroup contrast we evaluated. The d_{Mod} family of effect sizes consists of four main statistics, each of which offers a different but complementary characterization of differences in prediction. The four effect sizes can be considered two pairs of statistics: a pair of overall averages (d_{Mod_Signed} and $d_{Mod\_Unsigned}$) and a pair of directional averages (d_{Mod_Under} and d_{Mod_Over}).

Nye and Sackett (2016) introduced the overall average effects, and Dahlke and Sackett (2018) introduced the directional averages while sharing some general refinements to the original d_{Mod} effect sizes. These refinements included clarifications about the overall averages and new non-parametric algebraic formulas to supplement Nye and Sackett's parametric integration-based formulas, which assume a normal distribution for predictors. Our computational approach follows Dahlke and Sackett's non-parametric strategy, as we had large distributions of AFQT scores and wanted our d_{Mod} effect sizes to reflect any departures from normality present in the focal subgroups' applicant populations.

The two overall average effect sizes are known as d_{Mod_Signed} and $d_{Mod\_Unsigned}$ and they summarize difference in prediction across the entire range of predictor scores. The d_{Mod_Signed} effect size is the net average of all differences in prediction across all values of predictor scores and is computed as:

$$d_{Mod\_Signed} = \frac{\text{mean}\left(\hat{Y}_{Difference}\right)}{\widehat{SD}_{Y_{Referent\_Unrestricted}}}$$

As a net average, d_{Mod_Signed} can reveal the prevailing differential prediction effect within a sample. A positive d_{Mod_Signed} result indicates that, on average, overprediction is the most prevalent effect, while a negative d_{Mod_Signed} result indicates that, on average, underprediction is the most prevalent effect. However, because it gives a net average, any given d_{Mod_Signed} result could be obtained by an infinite number of possible regression-line configurations. For example, a d_{Mod_Signed} result of 0 could mean that the subgroup regression lines are exactly equal (same slopes and intercepts), but it could also mean that the lines cross and within the range of operational predictor scores and the overprediction effects simply cancel out the underprediction effects. It is important to resolve this ambiguity by reviewing the configuration of subgroups' regression lines and considering the pattern of results from other d_{Mod} statistics.

The $d_{Mod\_Unsigned}$ effect size represents the average absolute value of difference in prediction, meaning that it can only be zero or positive. It is computed as:

$$d_{Mod\_Unsigned} = \frac{\text{mean}\left(\left|\hat{Y}_{Difference}\right|\right)}{\widehat{SD}_{Y_{Referent\_Unrestricted}}}$$

This effect size can complement d_{Mod_Signed} by indicating the overall average in differences, regardless of their direction. If $d_{Mod\_Unsigned}$ equals the absolute value of d_{Mod_Signed}, it means there is a consistent direction of differences between subgroup regression lines (i.e., they do not cross within the operational range of predictor scores). However, if $d_{Mod\_Unsigned}$ is greater than the absolute value of d_{Mod_Signed}, it means subgroup regression lines cross within the operational range of predictor scores. This can be useful for generic MMR analyses, but it is admittedly not the most informative for differential prediction effects because, in this domain, the directions of the differences are of great importance. A more informative indicator of whether subgroup regression lines cross—and how much this impacts differences in prediction—is the pattern of directional d_{Mod} effect sizes.

Just as we can compute overall averages of differences in prediction, we can compute averages of differences for separate segments of the predictor distribution. The two segmented averages that have been formalized within the d_{Mod} framework represent separate effect sizes for underprediction and overprediction, called d_{Mod_Under} and d_{Mod_Over}, respectively. The d_{Mod_Under} effect size is computed as:

$$d_{Mod\_Under} = \mathrm{P}\left(\hat{Y}_{Difference} < 0\right) \times \frac{\mathrm{mean}\left(\hat{Y}_{Difference[<0]}\right)}{\widehat{SD}_{Y_{Referent\_Unrestricted}}}$$

where $\hat{Y}_{Difference[<0]}$ represents all negative differences in prediction observed within the sample and $\mathrm{P}\left(\hat{Y}_{Difference} < 0\right)$ represents the proportion of the sample associated with negative differences in prediction. We multiply the mean negative difference by the proportion of negative differences so that the effect size can properly characterize the magnitude and prevalence of underprediction effects. When there is no underprediction, d_{Mod_Under} is zero.

The overprediction counterpart to d_{Mod_Under} is d_{Mod_Over}, which is computed as:

$$d_{Mod\_Over} = \mathrm{P}\left(\hat{Y}_{Difference} > 0\right) \times \frac{\mathrm{mean}\left(\hat{Y}_{Difference[>0]}\right)}{\widehat{SD}_{Y_{Referent\_Unrestricted}}}$$

where $\hat{Y}_{Difference[>0]}$ represents all positive differences in prediction observed within the sample and $\mathrm{P}\left(\hat{Y}_{Difference} > 0\right)$ represents the proportion of the sample associated with positive differences in prediction. When there is no overprediction, d_{Mod_Over} is zero.

Since the d_{Mod_Under} and d_{Mod_Over} effect sizes are computed from non-overlapping segments of the predictor distribution and are weighted by the prevalence of their directional effects, they add up to d_{Mod_Signed} and their absolute values add up to $d_{Mod\_Unsigned}$. If only one of the directional effect sizes is non-zero, it means that subgroups' regression lines do not cross within the operational score range and the differences in prediction between subgroups are in a consistent direction. If d_{Mod_Under} and d_{Mod_Over} are both non-zero, we can infer that subgroups' regression lines cross. If they are both zero, we can infer that there are no differences in prediction across the operational range of predictor scores (this is rare, as the regression lines are likely to differ in some way, at least due to sampling error).

We supplemented our d_{Mod_Signed}, $d_{Mod\_Unsigned}$, d_{Mod_Under} and d_{Mod_Over} results with conditional d_{Mod} effect sizes that summarize the signed differences between subgroups' regression equations at key points in the AFQT distribution. Specifically, we computed conditional d_{Mod} effects for AFQT

scores of 10, 16, 21, 31, 50, 65, 93, which represent the lower bounds of the IVC, IVB, IVA, IIIB, IIIA, II, and I AFQT categories, respectively.

Proof-of-Methods Simulation

Given the changes we made to established procedures, we ran a targeted simulation to verify that these procedures functioned as anticipated. Due to the scope of the simulation, we present it in Appendix B rather that incorporate it directly into this chapter. As a high-level summary, the simulation supported the value of our residualized covariate approach for recovering estimates of unrestricted regression coefficients. Our results also showed that our methods helped in estimating d_{Mod} statistics, particularly d_{Mod_Signed}.

Procedures for Evaluating Post Hoc Power

Post hoc power (PHP) is a statistical concept that quantifies the probability of detecting an effect of a specific magnitude after an analysis has already been performed on a given data set. It is the post-analysis counterpart to *a priori* power analysis, in which one uses an anticipated effect size, a desired power level, and an alpha/significance level to establish a target sample size for a data collection effort. Although *a priori* power is arguably the more useful approach, as it informs the design of a study and ideally helps to collect enough data to stand a reasonable chance of detecting a hypothesized effect, PHP—if used properly—can provide helpful context when interpreting the results of an analysis. We examined PHP to provide context for the results of our differential prediction analyses, because these types of analyses tend to have difficulty achieving sufficient power to detect slope differences (Onwuegbuzie & Leech, 2004). We took measures to ensure our PHP analyses avoided the limitations that commonly undermine the value of PHP estimates.

When the observed effect size from an analysis is used to compute an estimate of PHP for that analysis, PHP provides no new information about the analysis (Hoenig & Heisey, 2001). This is because the power estimate is entirely determined if one knows the *p* value from the analysis that produced the effect size: An analysis that produced a small *p* value will have a higher level of estimated power, while a study that produced a large *p* value will have a lower level of estimated power. Although the relation between *p* values and post hoc power estimates is not linear, it is strictly monotonic, as such computing PHP for an analysis based on the results of that analysis does not increment one's understanding of the analysis.

For PHP to produce informative estimates, it is best if the effect size for the power calculations comes from an independent sample (e.g., a previous study on the same topic), an aggregate effect across multiple samples, or a determination of what constitutes a "meaningful" effect in one's research domain. For example, one could obtain an effect size from information presented in a published primary study, technical report, or meta-analysis; one could also set an effect size based on an average of effects observed in a set of samples from one's own study or based on an empirical definition of a meaningful effect (e.g., at least 1% increase in variance explained). In any case, the effect size used to evaluate power should be independent of (or at least not entirely determined by) the sample for which one wishes to evaluate power.

For our analyses, we evaluated PHP using effect sizes computed via two approaches: (1) aggregate effect sizes based on the weighted average magnitude of the top 10% of model differences (i.e., contrasts at the 90th percentile and above) observed in a given Service for a given criterion variable (e.g., Air Force occupations for which we compared the AFQT-awarding course grade relationship between White non-Hispanic Airmen and Black non-Hispanic Airmen)

and (2) fixed effect sizes that represent a magnitude of model difference that we judged could be meaningful in the context of differential prediction analyses.

Our linear regression analyses and logistic regression analyses were based on different probability distributions, so we present separate summaries of our methods for evaluating power for each type of regression analysis in the following subsections.

Post Hoc Power for Linear Regression Models

To evaluate PHP for each contrast between linear regression models, we used effect sizes based on (a) a normative sample-size weighted average of the top 10% of ΔR^2 values and (b) a fixed ΔR^2 value of .01. We sorted ΔR^2 values by descending magnitude and computed the average of the top 10% of values as:

$$\overline{\Delta R^2_{90\%ile}} = \frac{\sum_{i=90th\ \%ile}^{k}\left(R^2_{Full_i} - R^2_{Reduced_i}\right) \times n_i}{\sum_{i=90th\ \%ile}^{k} n_i}$$

The effect sizes needed to define the non-central *F* distribution are based on the ratio of ΔR^2 to the proportion of variance left unexplained by the larger (or "full") model. We computed the average R^2_{Full} value across all samples to use in the denominator of our effect size estimates, as using an average value in the denominator can provide a more stable power estimate than the sample-based R^2_{Full} value from any individual sample. We computed the average R^2_{Full} as:

$$\overline{R^2_{Full}} = \frac{\sum_{i=1}^{k} R^2_{Full_i} \times n_i}{\sum_{i}^{k} n_i}$$

We used the $\overline{R^2_{Full}}$ with the normative $\overline{\Delta R^2_{90\%ile}}$ value and the fixed .01 ΔR^2 value to compute the following effect size estimates:

$$ES_{90th\%ile} = \frac{\overline{\Delta R^2_{90\%ile}}}{1 - \overline{R^2_{Full}}}$$

$$ES_{Fixed} = \frac{.01}{1 - \overline{R^2_{Full}}}$$

We then estimated power to detect a normative $ES_{90th\%ile}$ effect as:

$$power_{90th\%ile_i} = 1 - P_F\left(F_{CV_i}; df_{1_i}; df_{2_i}; \lambda_i = ES_{90th\%ile} \times n_i\right)$$

where F_{CV_i} is the critical *F* value for an α value of .05 and degrees of freedom equal to df_{1_i} and df_{2_i}, and where λ_i is the non-centrality parameter of the *F* distribution. Likewise, we estimated power to detect a fixed ΔR^2 value of .01 as:

$$power_{Fixed_i} = 1 - P_F\left(F_{CV_i}; df_{1_i}; df_{2_i}; \lambda_i = ES_{Fixed} \times n_i\right)$$

These power estimates will provide helpful context for understanding the significance tests for our sets of nested linear models.

Post Hoc Power for Logistic Regression Models

Whereas the non-centrality parameters for linear regression PHP analyses are based on variance ratios that have conventional effect-size interpretations, the effect sizes that inform PHP analyses for logistic regressions are based on differences in deviance and are not as easy to interpret. The effect size that determines the non-central chi-square distribution for comparisons between logistic regression models is:

$$ES_i = \frac{Deviance_{Full_i} - Deviance_{Reduced_i} - df_{1_i}}{df_{2_i}}$$

where $Deviance_{Full_i}$ is the improvement in model deviance for the larger model over a null model, $Deviance_{Reduced_i}$ is the improvement in model deviance for the smaller model over a null model, df_{1_i} is the difference in the number of predictors between the larger model and smaller model, and df_2 is the residual degrees of freedom for the larger model. Conventionally, chi-squared-based models only have one value for degrees of freedom (what we call df_1; for a chi-squared distribution, this is also the expected value of the null distribution, which is why it is subtracted in the numerator), but the ANOVA-based residual degrees of freedom (df_2) are important for scaling the effect size according to a sample size that has been penalized for model complexity. This type of effect size is conceptually similar to a phi coefficient (also known as a Matthews correlation coefficient [MCC]), where one divides a chi-squared statistic by the corresponding sample size and takes the square root of the quotient. Since the effect size for PHP does not require taking a square root, it has qualities of an R^2-type statistic.

To evaluate PHP for each contrast between logistic regression models, we used effect sizes based on (a) a normative sample-size weighted average of the top 10% of sample-specific effects and (b) a fixed effect-size value (ES_{Fixed}) of .005. We determined this .005 effect size was roughly equivalent to a ΔR^2 value of .01 by plotting sample effect sizes against their corresponding pseudo ΔR^2 values across subgroup contrasts and model comparisons.

We sorted samples' effect-size values by descending magnitude and computed the average of the top 10% of values as:

$$\overline{ES_{90\%ile}} = \frac{\sum_{i=90th\ \%ile}^{k} ES_i \times n_i}{\sum_{i=90th\ \%ile}^{k} n_i}$$

We estimated power to detect a normative effect size of $\overline{ES_{90\%ile}}$ as:

$$power_{90th\%ile_i} = 1 - P_{\chi^2}\left(\chi^2_{CV_i}; df_{1_i}; \lambda_i = ES_{90th\%ile} \times n_i\right)$$

where $\chi^2_{CV_i}$ is the critical χ^2 value for an α value of .05 and degrees of freedom equal to df_{1_i}. Likewise, we estimated power to detect a fixed effect size of .005 as:

$$power_{Fixed_i} = 1 - P_{\chi^2}\left(\chi^2_{CV_i}; df_{1_i}; \lambda_i = ES_{Fixed} \times n_i\right)$$

These power estimates will provide helpful context for understanding the significance tests for our sets of nested logistic models.

Interpretation of Post Hoc Power

Given that our estimates of PHP were based on aggregated effects and rationally set benchmark values, we avoided the typical limitations of PHP that arise from estimating power based on an observed effect size. Our PHP analyses can provide helpful context for interpreting results from samples where we found non-significant differences between regression models. PHP can contribute to our understanding of results from MMR model contrasts, as these analyses have notoriously low power to detect interaction effects and PHP offers a way to rule out low power as an explanation for null findings.

If a sample has a high estimated level of power to detect our targeted effect sizes but has a non-significant observed result, it lends credence to the legitimacy of the non-significant finding because we can rule out low power as an explanation for the result. On the other hand, a non-significant finding from a sample that has low power to detect our targeted effect sizes inspires less confidence and stands a greater chance of being a Type II error (i.e., a false negative).

PHP is less relevant for samples that produced statistically significant findings, as these samples already produced the outcome that PHP is meant to evaluate. In other words, it is not terribly informative to evaluate the probability of obtaining a significant result in a sample that already produced such a result.

Aggregation of Results

Our analysis procedures generated a large volume of results across Services, as each occupation or course could potentially be examined for differential prediction in four subgroup contrasts for each criterion within its respective Service. To summarize results in a way that supports an understanding of overall trends, within each Service-specific chapter (i.e., Chapters 4 through 7), we aggregated the results across all occupations or courses for which analyses were conducted in that Service, shared a criterion, and were evaluated for the same subgroup contrast (e.g., male vs. female). Each of the Service-specific chapters in this report will present a collection of tables for regression analyses and d_{Mod} effect sizes summarizing statistical estimates and rates of significant results for each combination of subgroup contrast and criterion variable.

During a dry run of our analyses, we realized that not all occupations/courses that met the criteria for inclusion in our study (see Chapter 2) would be appropriate to include in aggregate summaries of results. This was only a concern for logistic regression models: Some occupation/courses had such extreme base rates on their binary criteria that, although they satisfied our requirement that the criterion have non-zero variance, the amount of variance was too small to fit a stable model. Furthermore, some occupations/courses produced regression models with coefficients that were implausibly large given the scaling of the predictors (e.g., regression coefficients with values with three digits or more). Such unstable models are misleading to include in aggregate summaries of results, and arguably cannot support valid insights into differential prediction for the samples they describe.

After running their analyses, our analysts reviewed their occupation/course-level results, flagged analyses with implausibly large regression coefficients, and explored their data for evidence of characteristics that could cause problems with model stability. The occupation/course-level analyses flagged during this process were retained for inclusion in our appendices of detailed results, but we excluded them from our tables of aggregate results. We also excluded these

problematic samples from the distributions of model statistics that contributed to the normative effect sizes used in our post hoc power analyses.

Summary

We augmented the traditional procedures for evaluating differential prediction with regression models and d_{Mod} statistics to mitigate the biasing impact of selection artifacts on our results. We developed a method for using residualized covariates to control for selection artifacts and estimate regression coefficients that do a better job of characterizing how AFQT scores, subgroup membership, and the AFQT-subgroup interaction relate to criterion scores in the unrestricted applicant population (see Appendix B for simulation evidence). We used the methods presented in this chapter in the analyses we summarize in Chapters 4 through 8.

Appendix B: Summary of Differential Prediction Analysis Methodology Simulation

Before we deployed our methodology for fitting and evaluating regression models introduced in Chapter 3, we ran a simulation to verify that the method had the intended effect of accurately recovering the unrestricted population parameters of regression coefficients. In this Appendix, we describe our simulation's methodology and findings.

Method

We used Monte Carlo methods to generate data with known patterns of differences in prediction, imposed selection artifacts on samples of incumbent data, and applied the approaches described in Chapter 3. We then compared the results from our modified Cleary analyses to the results from models we fit to unrestricted data without the use of covariates.

We generated Monte Carlo data in which 70% of applicants were from the higher-scoring referent group and 30% were from the lower-scoring focal group. We imposed standardized mean differences between these subgroups of .8 for the primary predictor *X* and the covariate *Z*. We made *X* and *Z* correlate .5 within each subgroup, and we also generated a continuous criterion variable *Y* that correlated .3 with *X* and *Z* in the referent group. We generated data in which subgroups had equal unrestricted standard deviations of 1.0. We introduced differential prediction by manipulating mean differences on *Y*, subgroup differences in the *X*-*Y* relationship, and subgroup differences in the Z-Y relationship using the conditions presented in Table B.1. We also fully crossed the subgroup difference conditions from Table B.1 with selection conditions in Table B.2, which define the selection ratios we applied to *X* and *Z*. We designed the selection conditions in Table B.2 so that each condition had an overall selection ratio of .50, regardless of which variables were used to make selection decisions.

The parameters for this simulation were designed for linear regressions, but we also simulated logistic regressions by dichotomizing the criterion variable (cases with z scores above zero were coded as 1 and cases with scores below zero were corded as zero). We ran all our analyses with and without selection artifacts, so we were able to evaluate the effectiveness of our estimation strategy against analyses that were not impacted by the challenges we were trying to overcome.

Table B.1. Differential Prediction Conditions for Simulation

Difference Condition	Differential Prediction Type	d_Y	r_{XY} Difference	r_{ZY} Difference
1	Equal Prediction	0.24	0.00	0.00
2	Intercept Difference	0.44	0.00	0.00
3	Slope Difference	0.24	0.15	0.00
4	Intercept and Slope Difference	0.44	0.15	0.00
5	Equal Prediction	0.24	0.00	0.15
6	Intercept Difference	0.44	0.00	0.15
7	Slope Difference	0.24	0.15	0.15
8	Intercept and Slope Difference	0.44	0.15	0.15

Table B.2. Selection Conditions for Simulation

Selection Condition	Selection Method	SR_X	SR_Z	$SR_{Overall}$
1	Selection on X Only	0.50	1.00	0.50
2	Selection on Z Only	1.00	0.50	0.50
3	Selection on X & Z (Equal)	0.64	0.64	0.50
4	Selection on X & Z (Mostly X)	0.52	0.85	0.50
5	Selection on X & Z (Mostly Z)	0.85	0.52	0.50

Note. SR_X and SR_Z represent selection ratios applied separately to *X* and *Z*, respectively. $SR_{Overall}$ represents the total selection ratio after accounting for the combined effect of selecting on *X* and *Z*.

Results

Throughout our summary of results, we present figures that allow comparisons between analyses using (a) unrestricted data vs. restricted data and (b) a regular Cleary-based modeling strategy vs. our augmented Cleary approach that includes covariates computed from other selection variables (see Chapter 3 for a description of this approach). In each of these figures (see Figure B.1 for an example), we recommend making a set of four comparisons among "Analysis Types" for each cell of the plot grid. These comparisons are:

1. Comparing "Unrestricted (Covariate)" to "Unrestricted (Regular)" helps to confirm that the use of covariates does not affect coefficient estimates when there are no systematic selection effects. These conditions should produce the same results because the covariate is not necessary to properly estimate the coefficients in unrestricted data.
2. Comparing "Restricted (Regular)" to "Unrestricted (Regular)" reveals the estimation bias that can occur if selection variables are not included in regression models that are based on range-restricted data. If the "Restricted (Regular)" results differ from the "Unrestricted (Regular)" results, it indicates that using a traditional Cleary analysis without covariates can produce results that mischaracterize the configuration of subgroup regression lines in the applicant population.
3. Comparing "Restricted (Covariate)" to "Unrestricted (Regular)" helps to confirm that the methods described in Chapter 3 functioned properly. The "Restricted (Covariate)" results should closely approximate the "Unrestricted (Regular)" results if our method function as intended.
4. Comparing "Restricted (Covariate)" to "Restricted (Regular)" (and with "Unrestricted (Regular)" serving as an anchoring point) gives an idea of the improvement in estimation that can be achieved by including relevant covariates in models based on range-restricted data. The degree to which "Restricted (Covariate)" closes the distance between "Restricted (Regular)" and "Unrestricted (Regular)" represents the benefit of including informative covariates in one's regression analysis.

Collectively, these four comparisons are useful for establishing that the methods from Chapter 3 are effective at recovering estimates of unrestricted effects when there are selection artifacts that could bias those estimates.

Linear Regressions

The first set of results we examined were for linear regression models, and we grouped these results according to whether or not the covariate exhibited slope differences between subgroups in the unrestricted population.

Equal Subgroup Slopes for Covariate

The results of linear regression models featuring a covariate (*Z*) that has equal slopes between subgroups are summarized in Figures B.1–B.5. Figure B.1 shows results for conditions in which subgroups' unrestricted regression lines were the same for the primary predictor (*X*), Figure B.2 shows results for conditions in which subgroups' unrestricted regression lines had different intercepts for *X*, Figure B.3 shows results for conditions in which subgroups' unrestricted regression lines had different slopes for *X*, and Figure B.4 shows results for conditions in which subgroups' unrestricted regression lines had different intercept and different slopes for *X*.

The patterns of results are consistent across Figures B.1–B.5, such that all analysis types produce the same estimates when *X* is the only variable involved in selection, but using restricted data to run regular Cleary analyses can introduce biases into the estimates of intercepts, predictor main effects, group main effects, and predictor-group interactions when both *X* and *Z* are used to select applicants. In these analyses where the *Z* variable had the same slope in each subgroup, using Z as the sole selection variable did not noticeably bias the estimates of regression coefficients; however, this is not an observation from which one can safely generalize. Including covariates based on *Z* was effective at controlling for the biasing impacts of selection artifacts and helped to arrive at coefficient estimates that were better representations of the trends we observed in analysis of unrestricted data.

Figure B.5 shows the results for d_{Mod} effect sizes. Including covariates in our range-restricted models improved estimates of d_{Mod_Signed} in all conditions where selection artifacts had a biasing effect on estimates derived from regular Cleary analyses. Using a regular Cleary analysis tended to produce negatively biased estimates of d_{Mod_Signed}; this means that, when the d_{Mod_Signed} population parameter was zero, one would be at risk of erroneously detecting underprediction and, when that parameter was greater than zero, one would be at risk of underestimating the extent of overprediction.

Covariates also tended to aid in estimating d_{Mod_Under} and d_{Mod_Over} but, even with the covariates, these effects were more challenging to recover. The challenges associated with d_{Mod_Under} and d_{Mod_Over} also impacted $d_{Mod\_Unsigned}$, as that effect is the sum of absolute-value directional effects. At first, this appears to suggest a problem with our strategy for using covariates; however, as we describe below, these trends are more readily attributable to other sources.

Most of the difficulty associated with estimating d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ can be explained by the effect sizes themselves. Consider the results for the equal-prediction conditions: In these conditions, the directional d_{Mod} effects are zero in the population, but there is only one direction for the differences to err in samples drawn from that population. Since d_{Mod_Under} can only be zero or negative while d_{Mod_Over} can only be zero or positive, estimation errors for an equal-prediction scenario can only deviate from zero in one direction, which is reflected in the average results depicted in Figure B.5.

This type of single-direction opportunity for estimation errors is amplified in range-restricted samples. Even when we use covariates to control for selection artifacts, models fit using range-

restricted data will tend to have more sampling error, which means there is more room for departure from the average coefficients we depicted in Figures B.1–B.4 and, by extension, more opportunities for errors in estimates of directional effects. An estimate of d_{Mod_Signed} is the sum of d_{Mod_Under} and d_{Mod_Over}, which means that the directional errors associated with each of its component effect sizes can cancel each other out; this contributes to d_{Mod_Signed}'s stability and the relative ease of estimating the overall signed effect. Estimates of $d_{Mod\_Unsigned}$, however, do not provide a way for the directional errors to cancel out; instead, d_{Mod_Signed} inherits the directional errors from both d_{Mod_Under} and d_{Mod_Over} and can therefore be very difficult to interpret with any degree of confidence.

We find further evidence for our explanation of the d_{Mod} estimation challenges when we consider the difference between settings that should have d_{Mod_Signed} effects of zero (conditions with equal prediction conditions or slope differences only) and those that have should have non-zero effects (conditions with intercept differences, either alone or in combination with slope differences). We modeled our intercept differences as overprediction effects, meaning that the directional effects should be less volatile in intercept-difference conditions because there is a true non-zero d_{Mod_Over} effect and, because of this, the average observed d_{Mod_Under} effect should not deviate substantially from zero. Indeed, the amount of estimation bias attributable to range restriction is smaller in intercept-difference conditions than in the other conditions. In general, we find that the volatility of d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ estimates is inversely related to the absolute magnitude of the d_{Mod_Signed} effect.

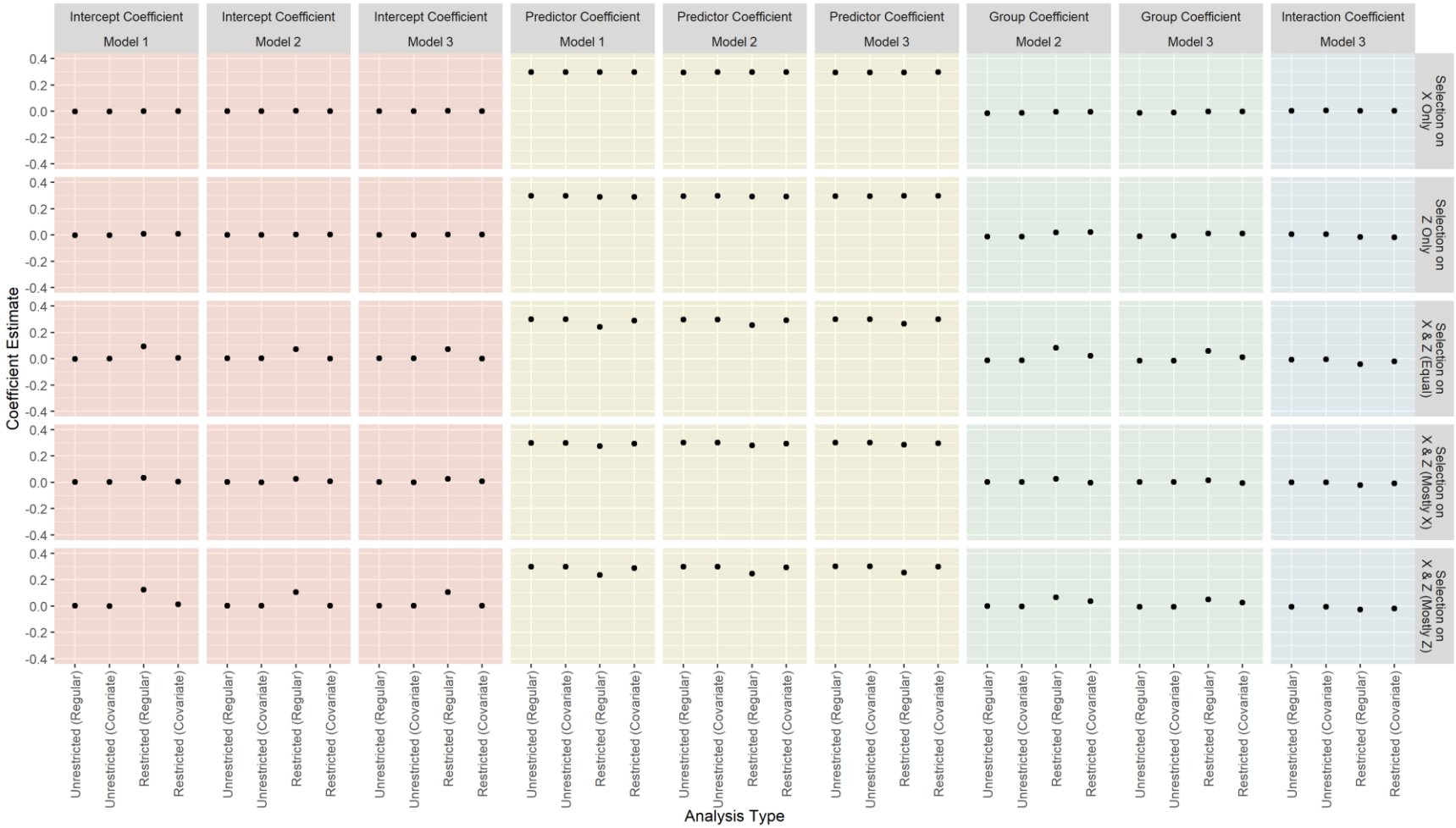

Figure B.1. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Equal Prediction Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

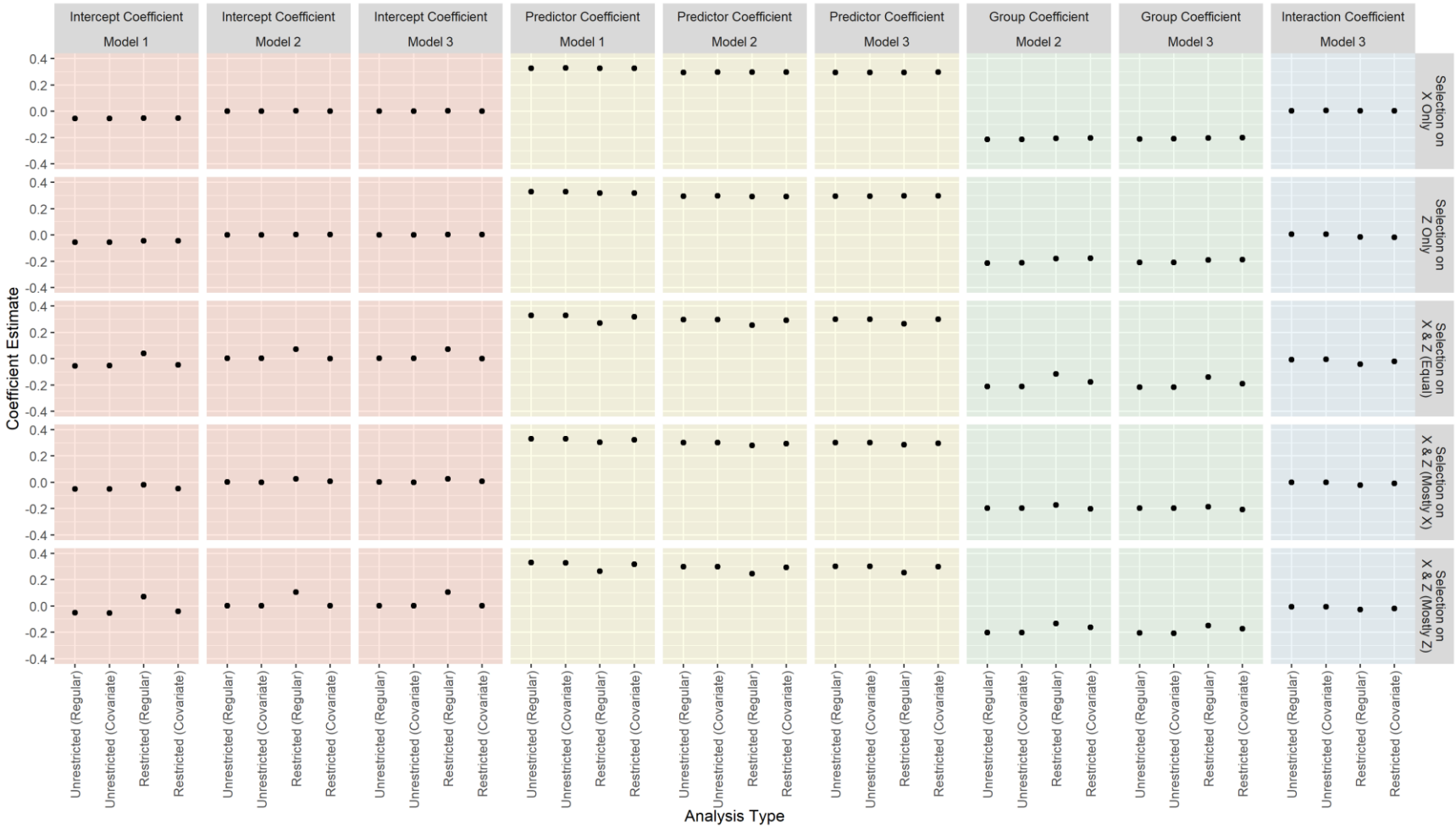

Figure B.2. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Intercept Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

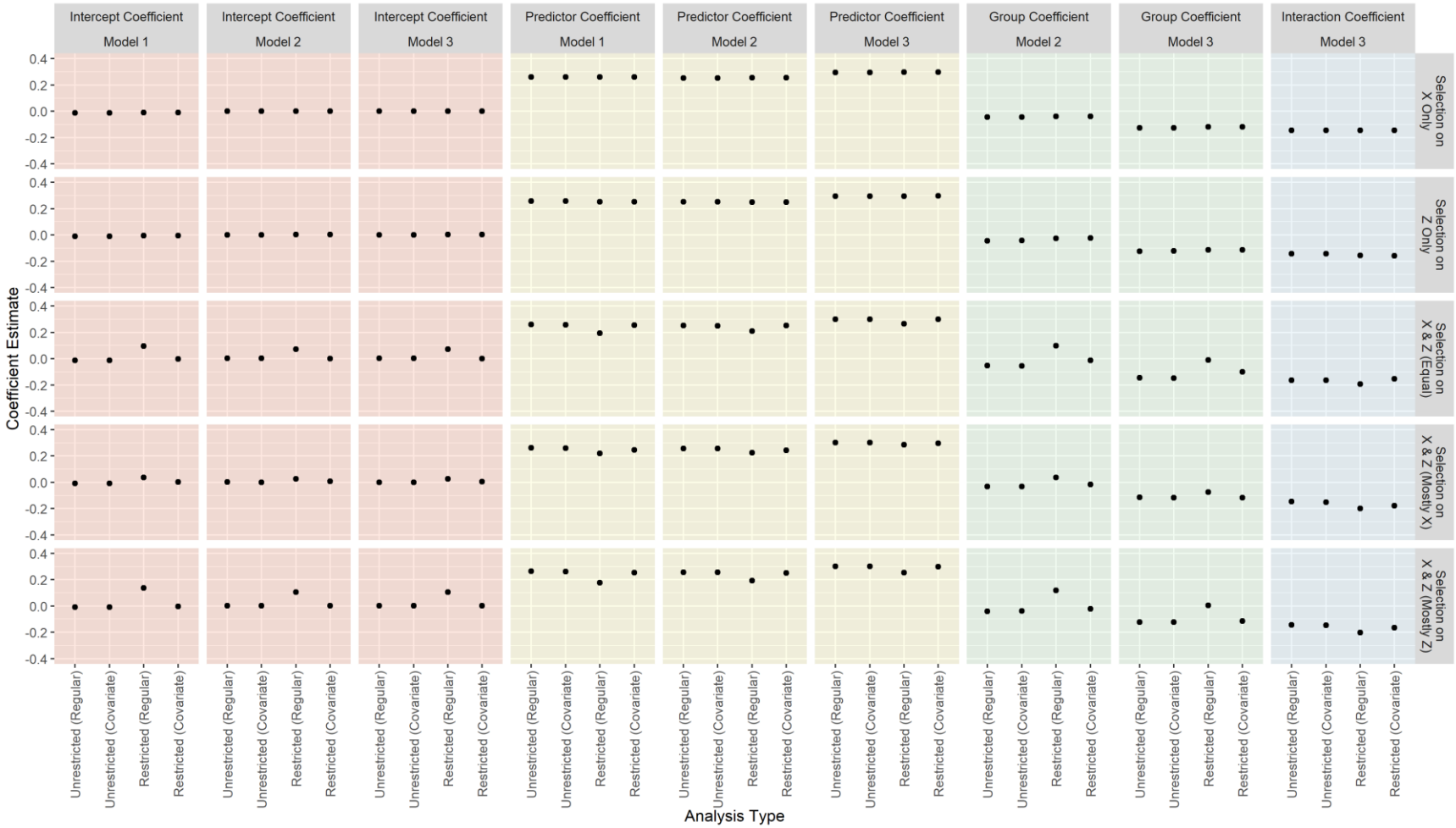

Figure B.3. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Slope Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

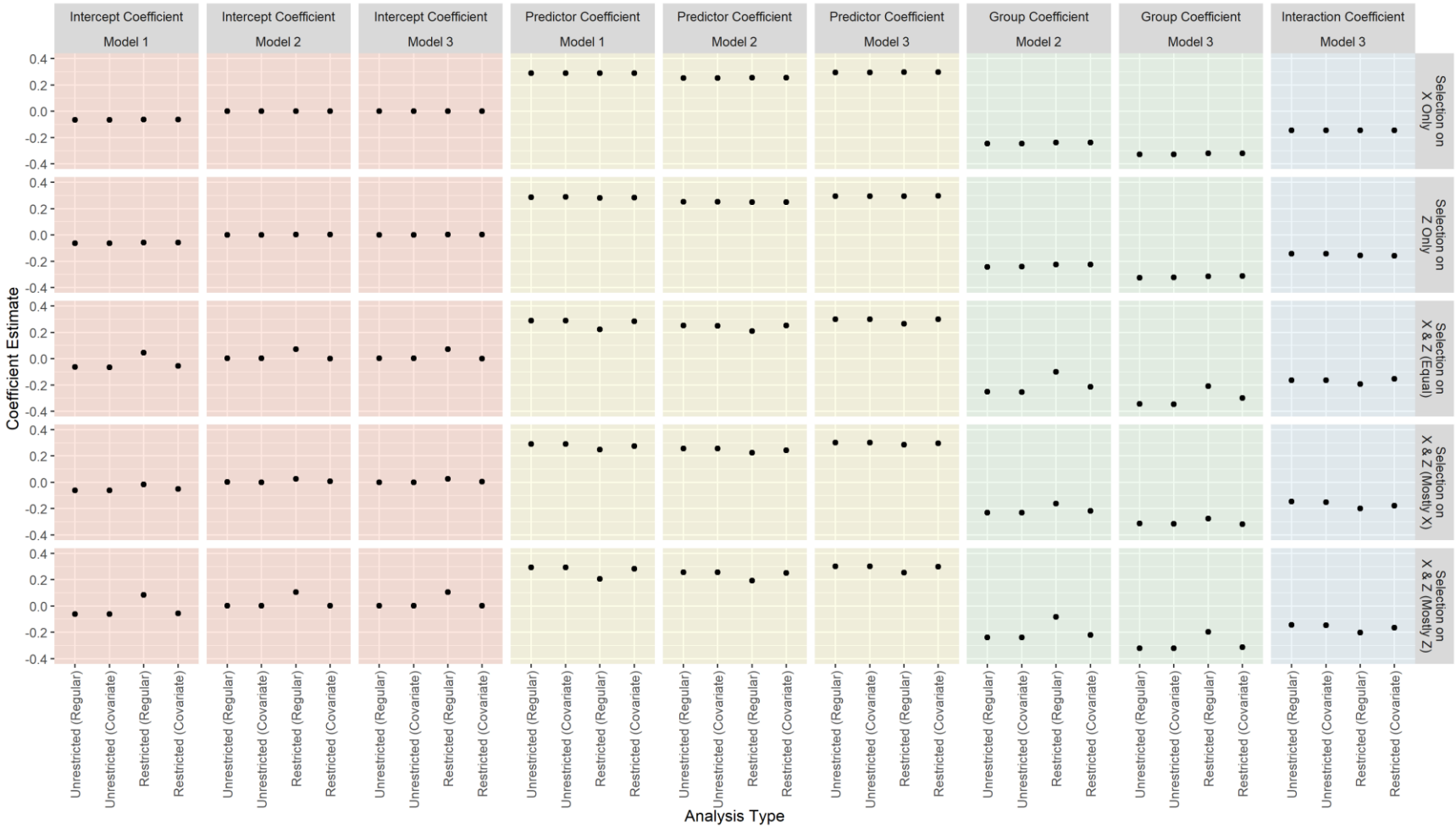

Figure B.4. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

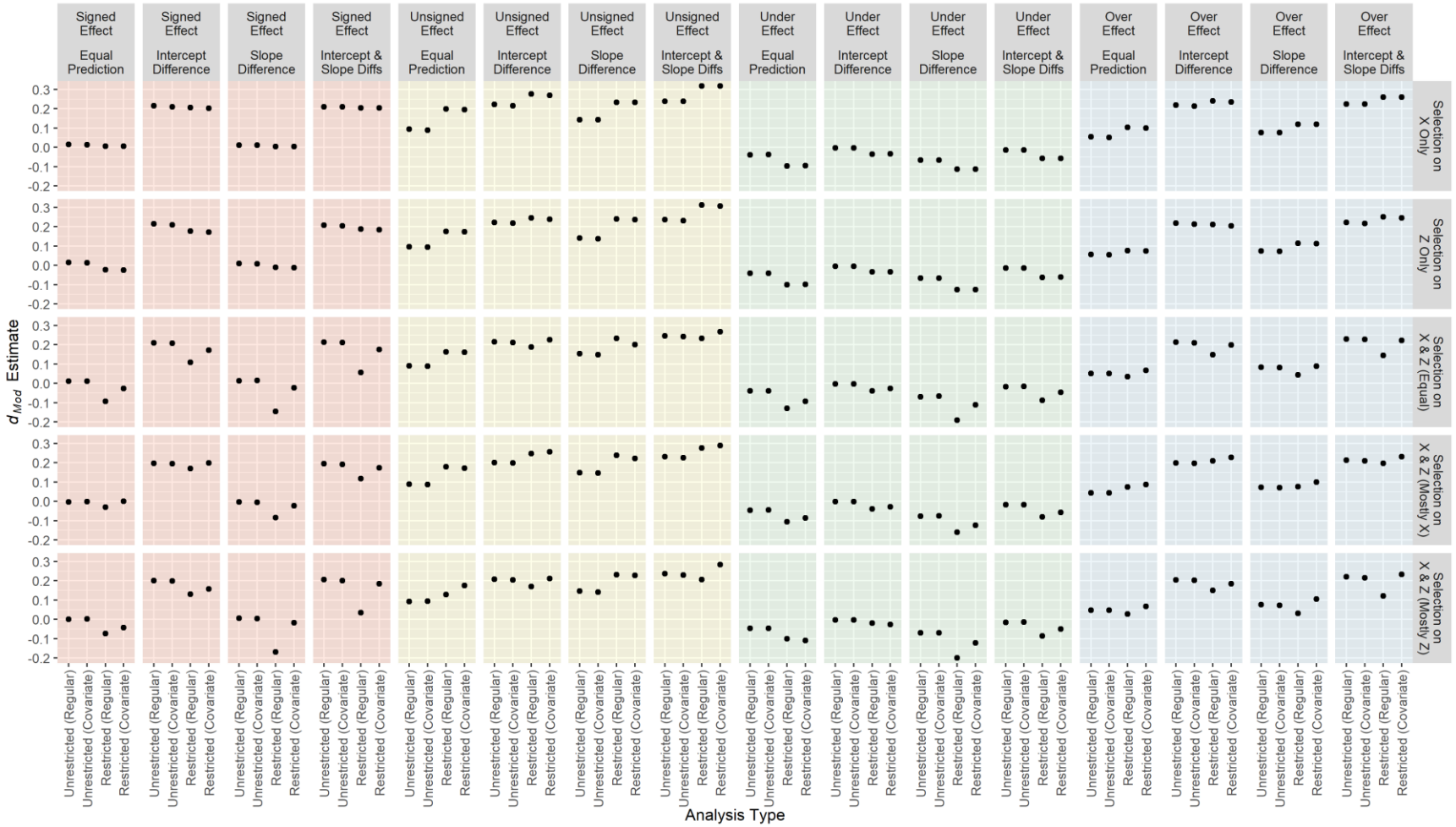

Figure B.5. Average Estimates of Linear Regression d_{Mod} Effect Sizes Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of d_{Mod} effect types and configurations of subgroup differences in the unrestricted population. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of d_{Mod} effect. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

Different Subgroup Slopes for Covariate

The results of linear regression models featuring a covariate that has different slopes between subgroups are summarized in Figures B.6–B.10. Figure B.6 shows results for conditions in which subgroups' unrestricted regression lines were the same for *X*, Figure B.7 shows results for conditions in which subgroups' unrestricted regression lines had different intercepts for *X*, Figure B.8 shows results for conditions in which subgroups' unrestricted regression lines had different slopes for *X*, and Figure B.9 shows results for conditions in which subgroups' unrestricted regression lines had different intercepts and slopes for *X*.

Consistent with our results for conditions in which subgroups had equal slopes for Z, controlling for selection artifacts was highly effective at debiasing estimates of linear regression coefficients when *Z* exhibited slope differences. When selection decisions involved *Z*, using a regular Cleary analysis without covariates could produce biased estimates of any regression coefficient; including covariates helped to avoid this.

Figure B.10 shows the results for d_{Mod} effect sizes. Whereas our results of conditions with equal Z slopes showed that range restriction had a negatively biasing effect on d_{Mod_Signed} effects, there was a positive bias when the referent group had a larger *Z* slope than the focal group. Using a regular Cleary analysis when an unmodeled selection variable exhibits slope differences puts one at risk of overestimating overprediction effects. Including covariates that capture selection artifacts helped to counteract this bias and bring the estimates of d_{Mod_Signed} into closer alignment with the unrestricted effect. Including covariates in regression models also tended to help when estimating d_{Mod_Over} and $d_{Mod\_Unsigned}$, but appears to have introduced some underprediction effects that were not present in the unrestricted data. We have previously described the challenges associated with estimating d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ effects in range-restricted settings, so we will not repeat those comments here.

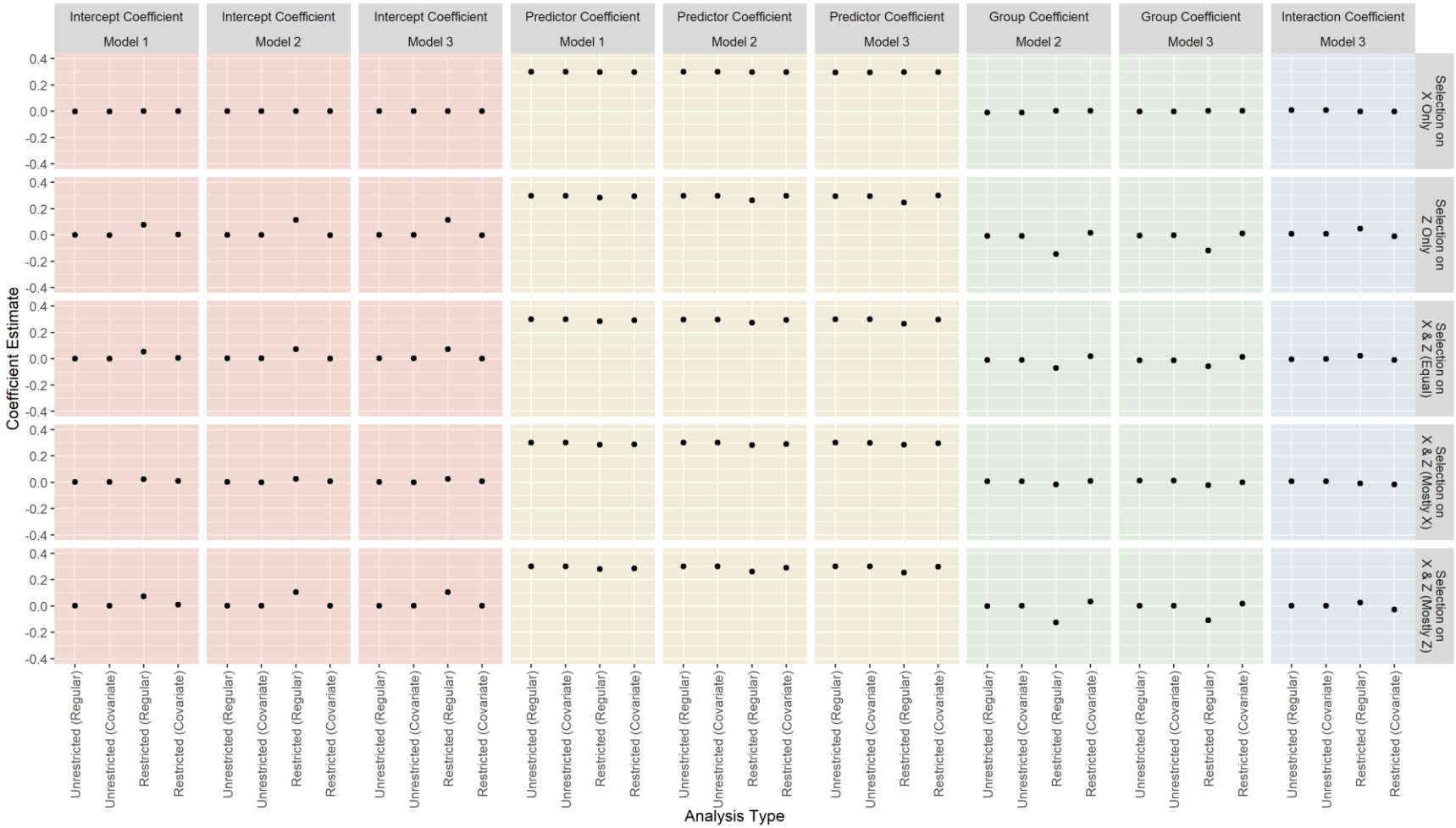

Figure B.6. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Equal Prediction Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

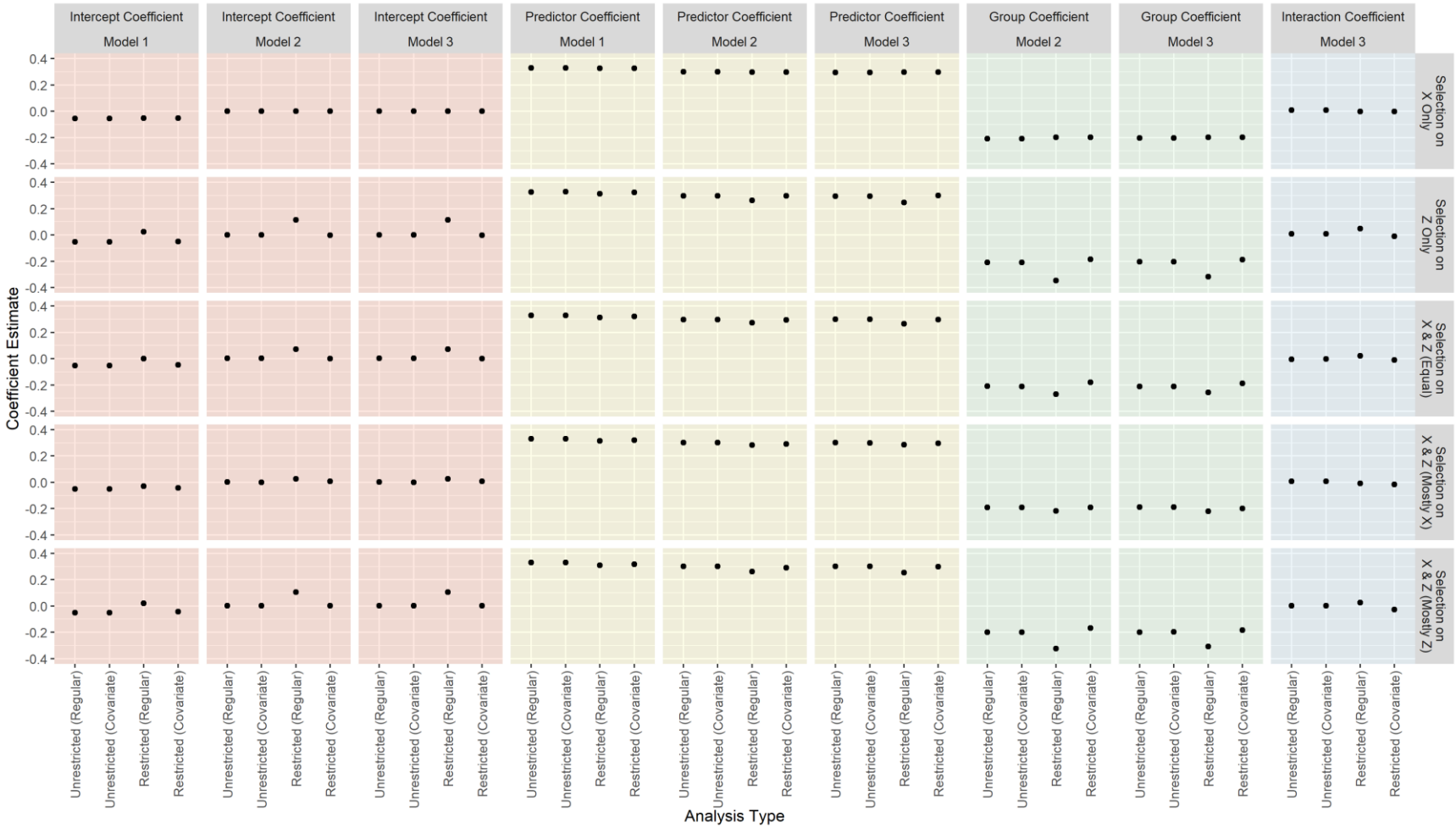

Figure B.7. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Intercept Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

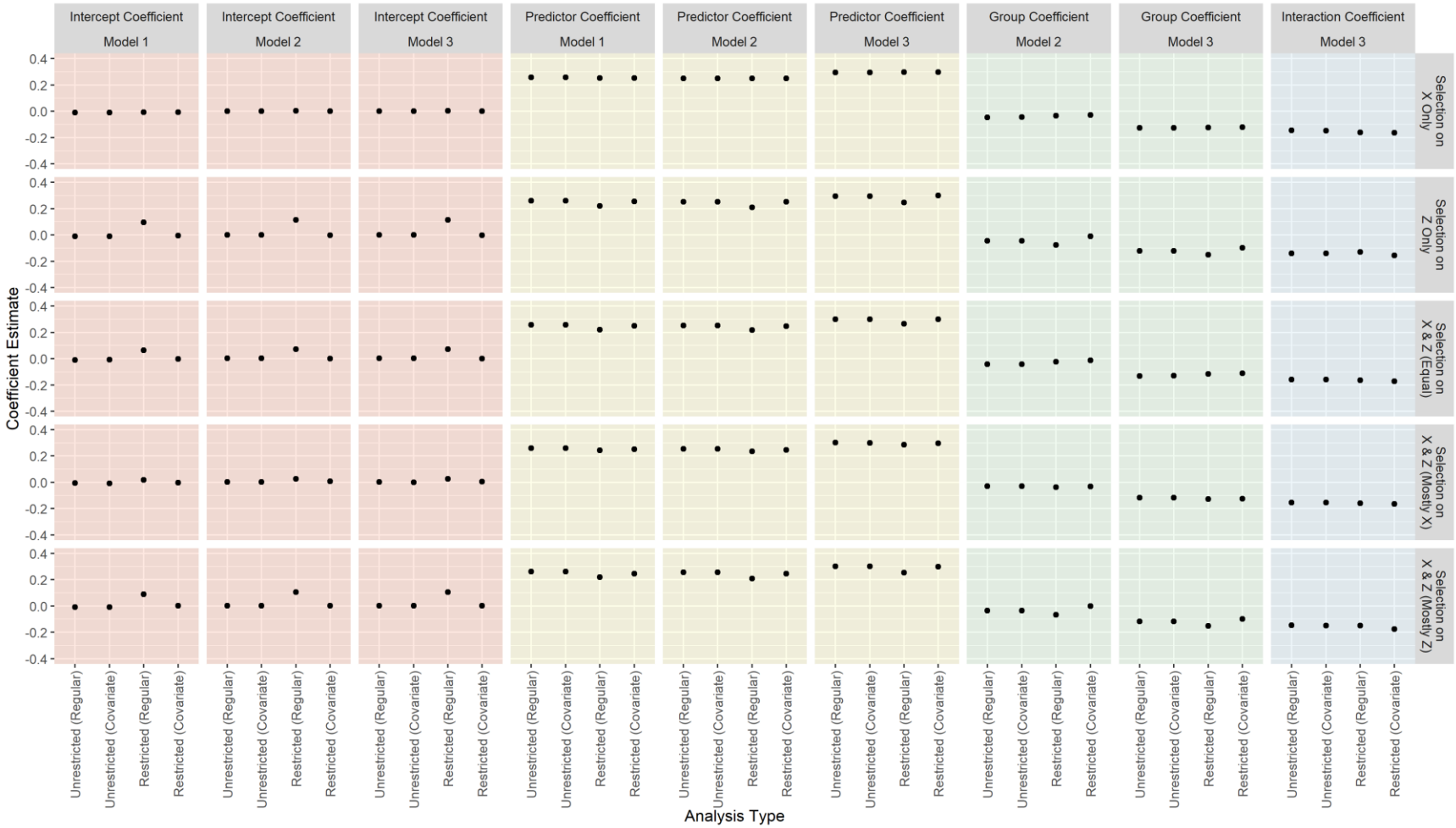

Figure B.8. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Slope Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

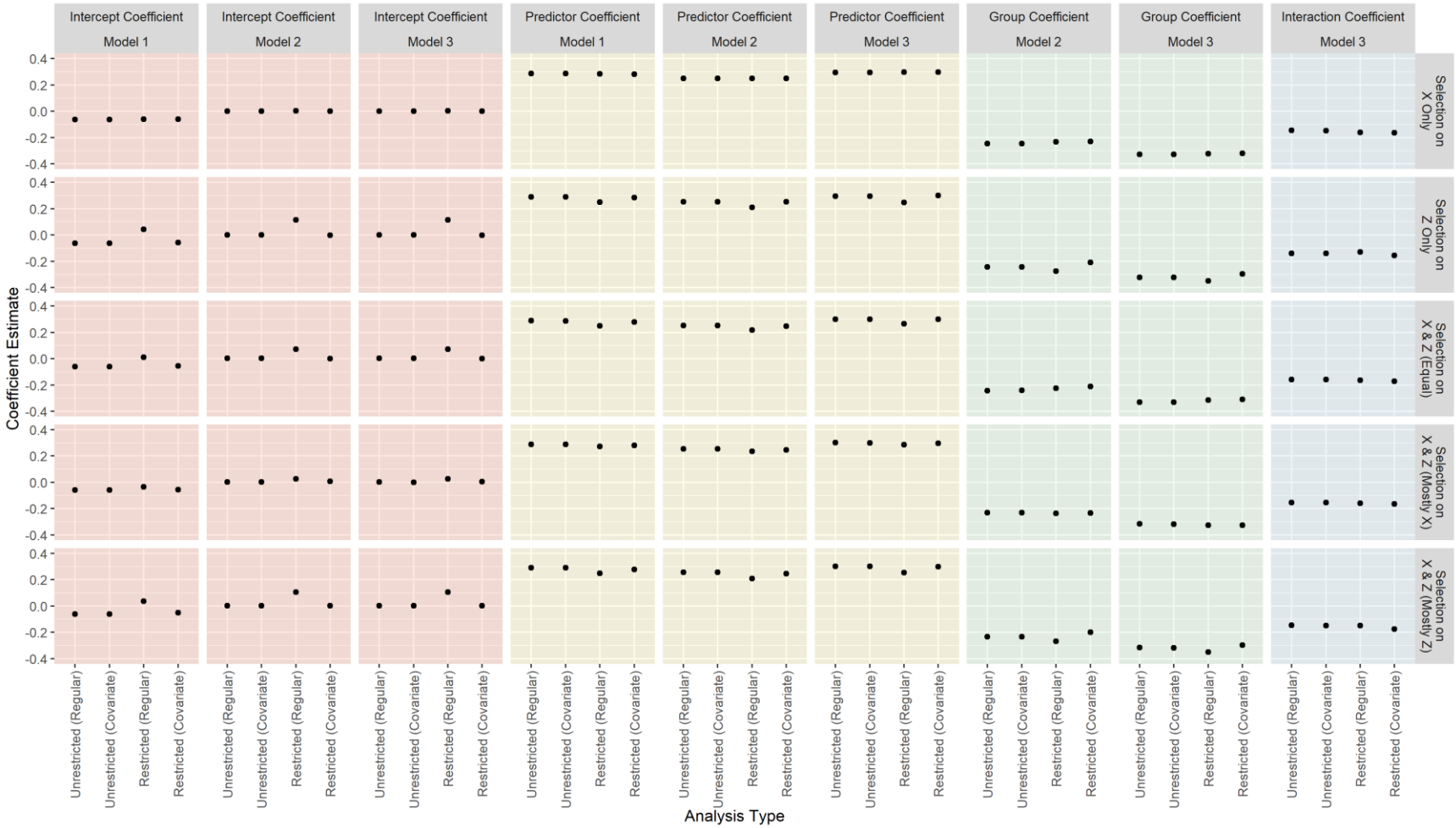

Figure B.9. Average Estimates of Linear Regression Coefficients Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

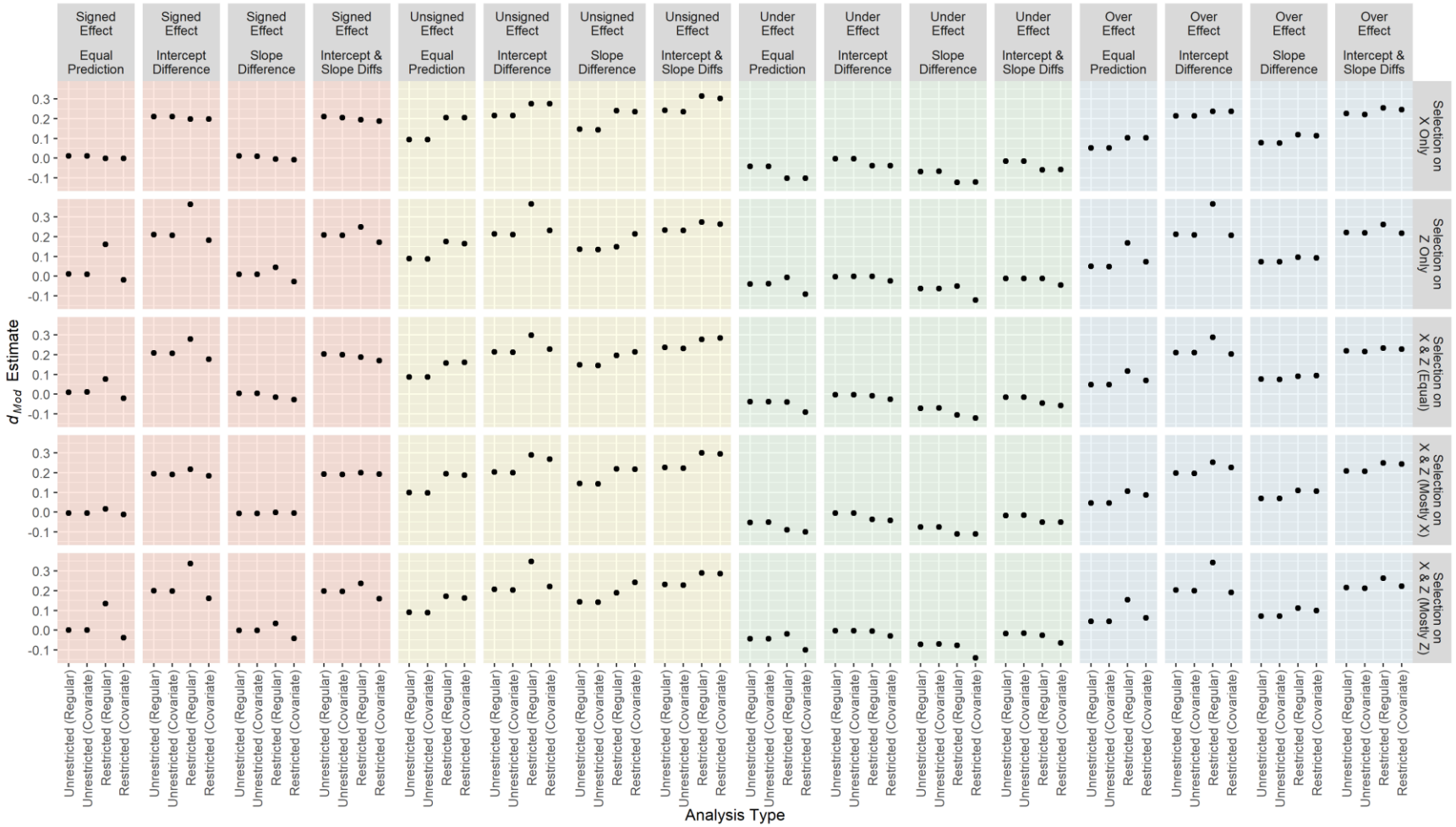

Figure B.10. Average Estimates of Linear Regression d_{Mod} Effect Sizes Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of d_{Mod} effect types and configurations of subgroup differences in the unrestricted population. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of d_{Mod} effect. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

Logistic Regressions

After finding that including appropriate covariates in linear regression models helped to recover estimates of unrestricted regression coefficients and d_{Mod} effects, we evaluated whether our approach generalized to logistic regression models. These results replicated those from our linear regression analyses, so we do not repeat our more detailed observations that we have already noted for the linear regressions.

Equal Subgroup Slopes for Covariate

The results of logistic regression models featuring a covariate that has equal slopes between subgroups are summarized in Figures B.11–B.15. Figure B.11 shows results for conditions in which subgroups' unrestricted regression lines were the same for *X*, Figure B.12 shows results for conditions in which subgroups' unrestricted regression lines had different intercepts for *X*, Figure B.13 shows results for conditions in which subgroups' unrestricted regression lines had different slopes for *X*, and Figure B.14 shows results for conditions in which subgroups' unrestricted regression lines had different intercept and different slopes for *X*. These results followed the same pattern as the corresponding results from our linear regression analyses. Whereas selection artifacts that involved a variable other than the predictor of interest caused misestimation of regression coefficients, adding covariates that captured those incidental selection effects brought the coefficient estimates into closer alignment with estimates computed from unrestricted data. Figure B.15 shows the results for d_{Mod} effect sizes, and the trends here replicated the corresponding linear regression d_{Mod} trends from Figure B.5.

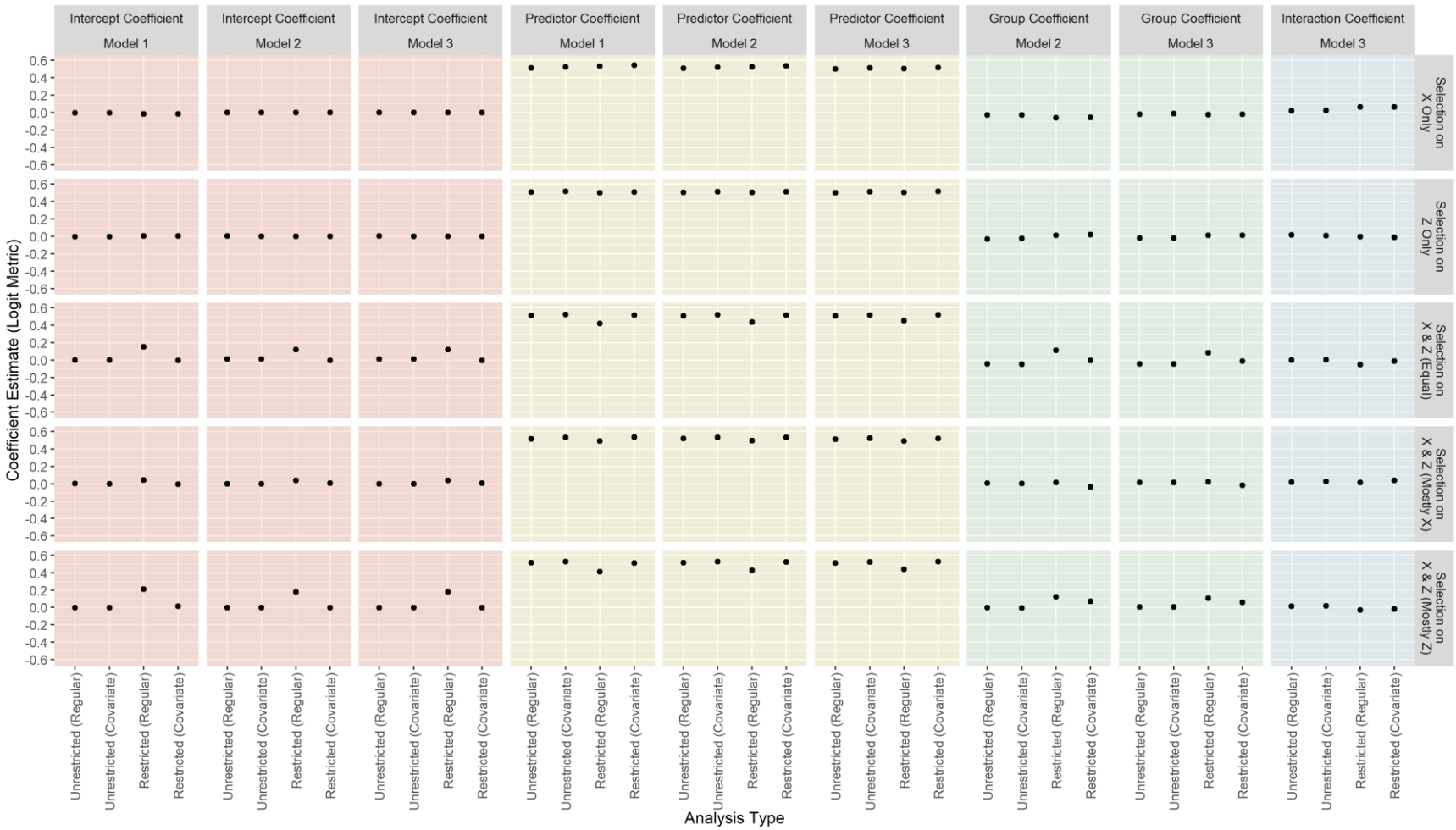

Figure B.11. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Equal Prediction Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

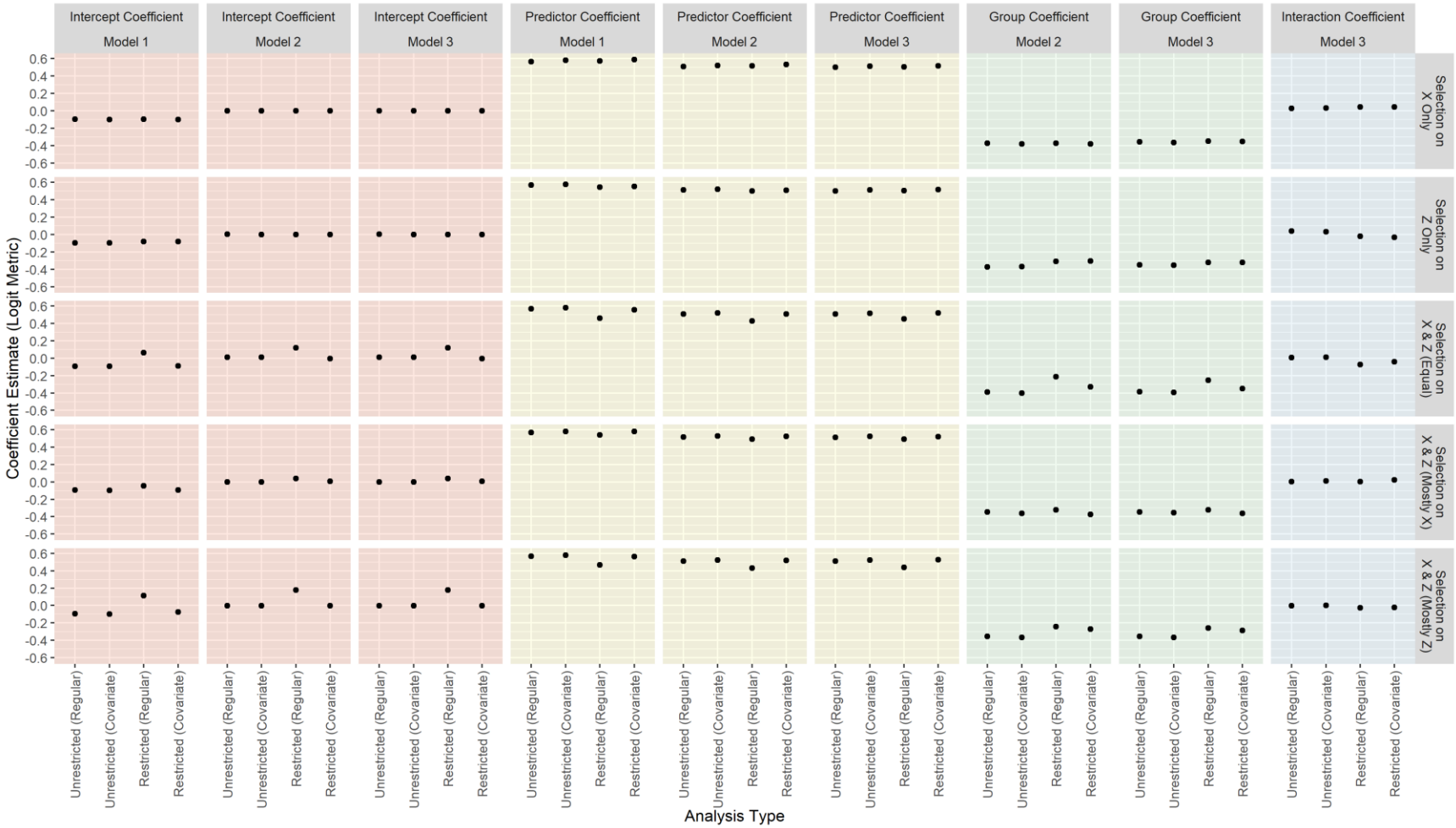

Figure B.12. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Intercept Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

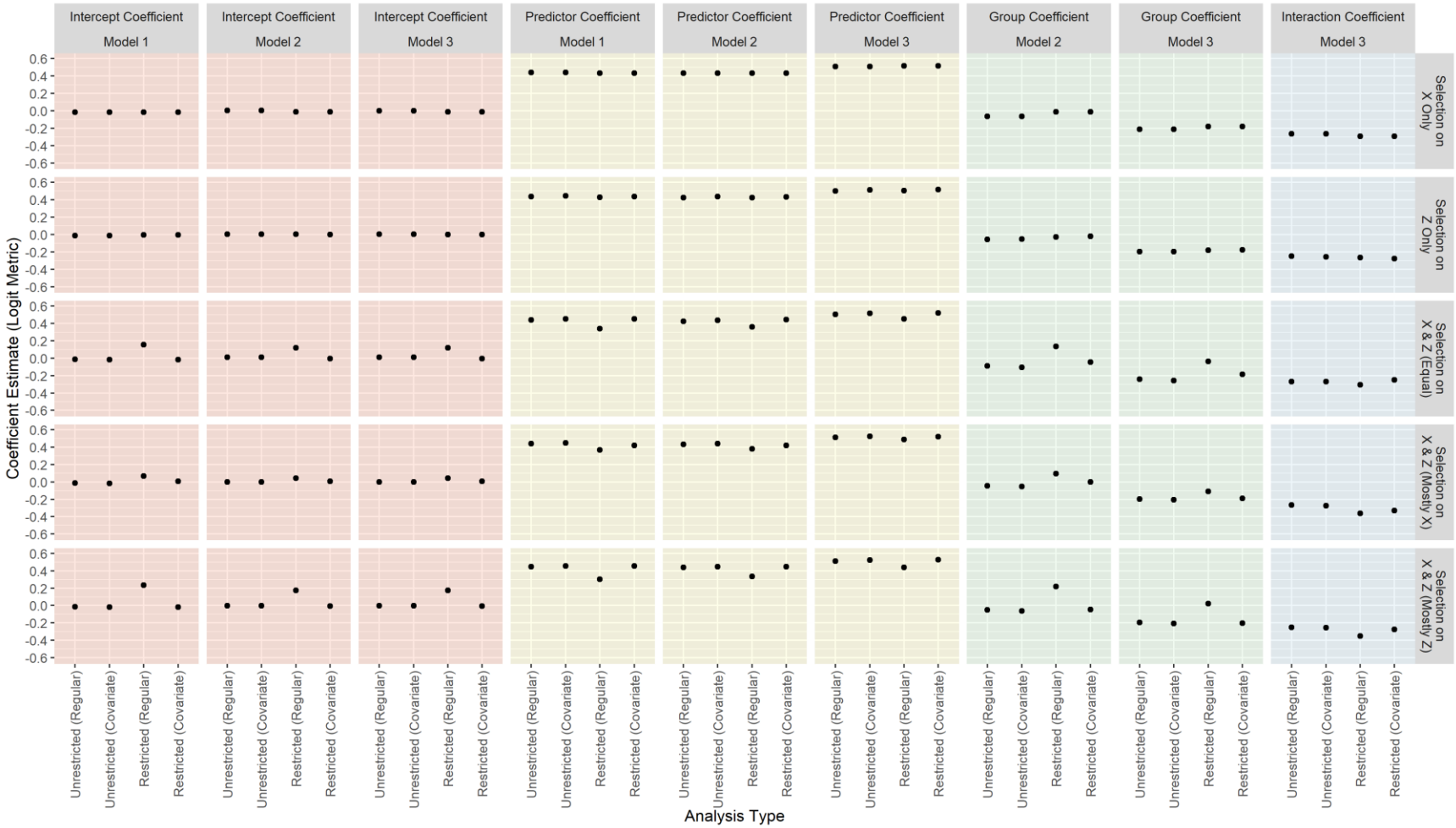

Figure B.13. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Slope Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

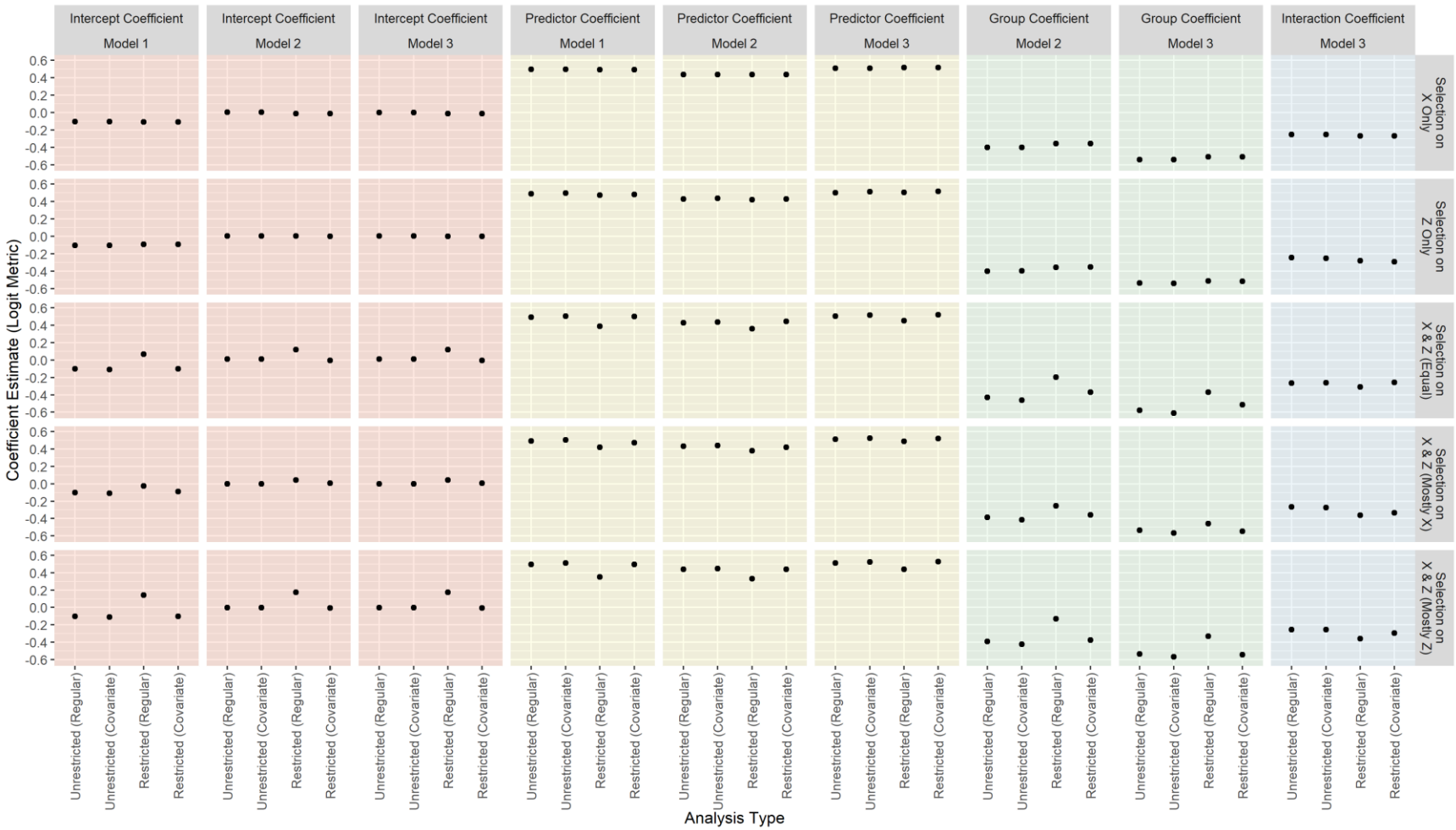

Figure B.14. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

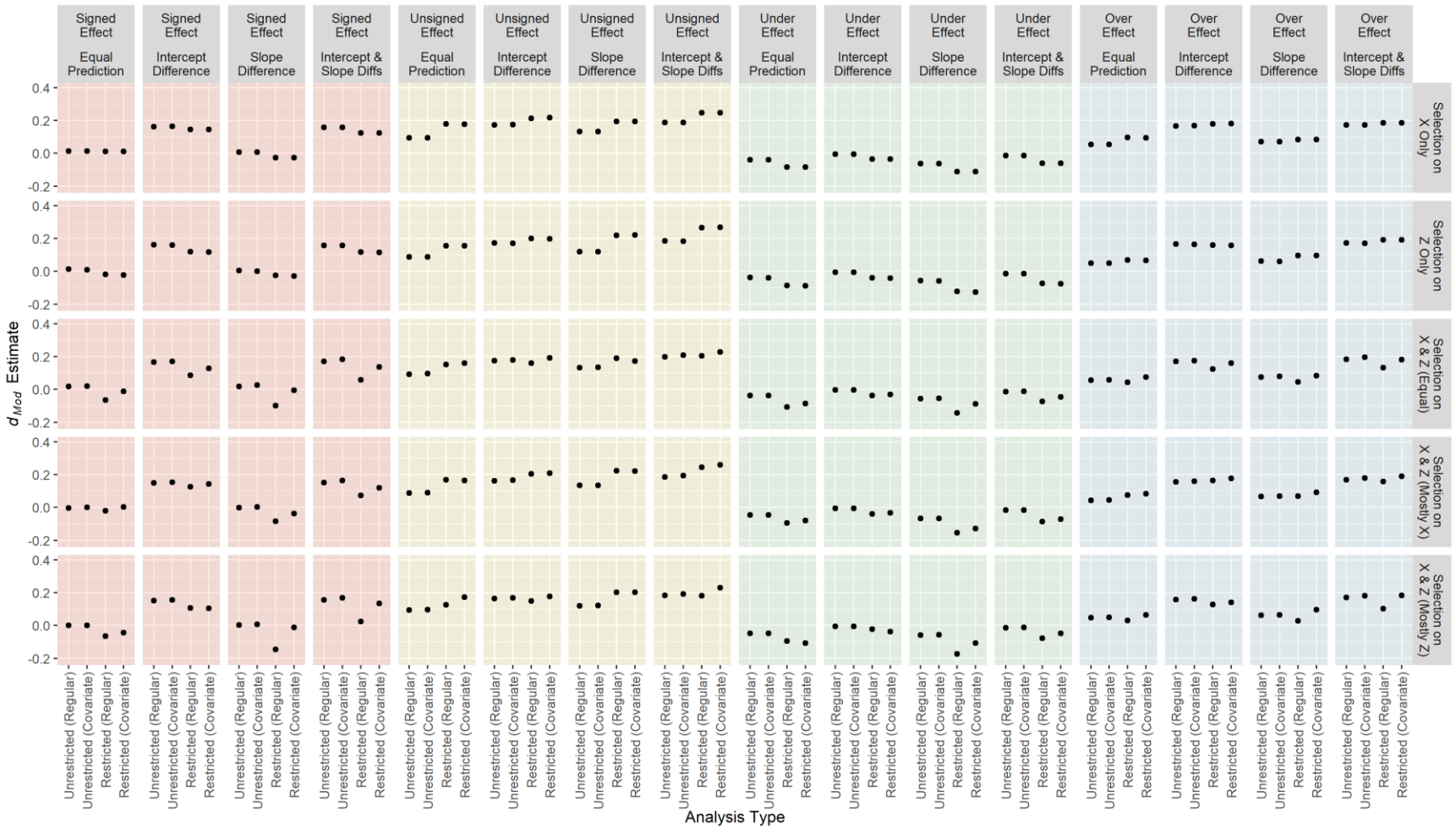

Figure B.15. Average Estimates of Logistic Regression d_{Mod} Effect Sizes Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Equal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of d_{Mod} effect types and configurations of subgroup differences in the unrestricted population. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of d_{Mod} effect. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

Different Subgroup Slopes for Covariate

The results of logistic regression models featuring a covariate that has different slopes between subgroups are summarized in Figures B.16–B.20. Figure B.16 shows results for conditions in which subgroups' unrestricted regression lines were the same for *X*, Figure B.17 shows results for conditions in which subgroups' unrestricted regression lines had different intercepts for *X*, Figure B.18 shows results for conditions in which subgroups' unrestricted regression lines had different slopes for *X*, and Figure B.19 shows results for conditions in which subgroups' unrestricted regression lines had different intercepts and slopes for *X*. These results followed the same pattern as the corresponding results from our linear regression analyses. Whereas selection artifacts that involved a variable other than the predictor of interest caused misestimation of regression coefficients, adding covariates that captured those incidental selection effects brought the coefficient estimates into closer alignment with estimates computed from unrestricted data. Figure B.20 shows the results for d_{Mod} effect sizes, and the trends here replicated the corresponding linear regression d_{Mod} trends from Figure B.10.

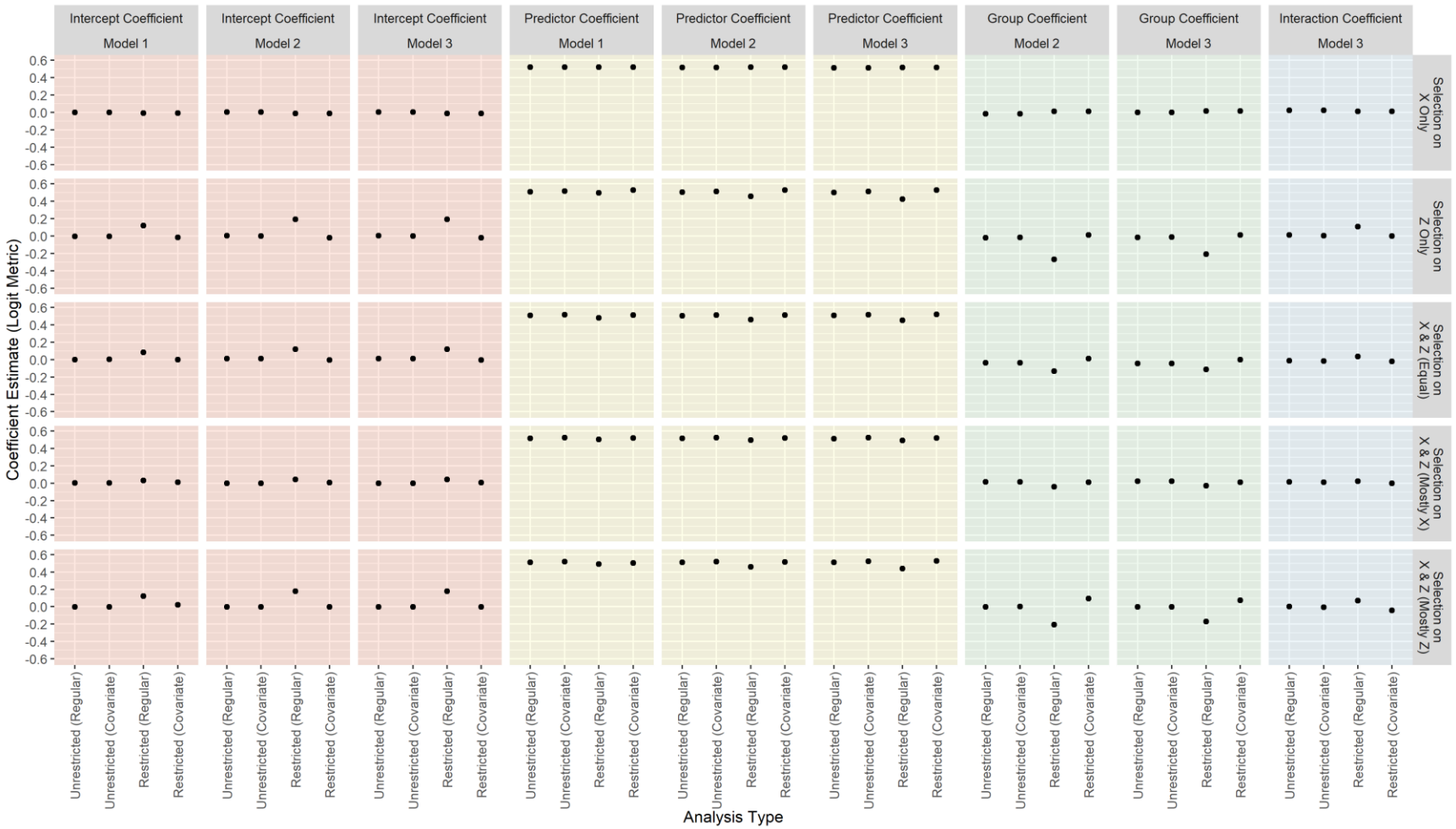

Figure B.16. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Equal Prediction Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

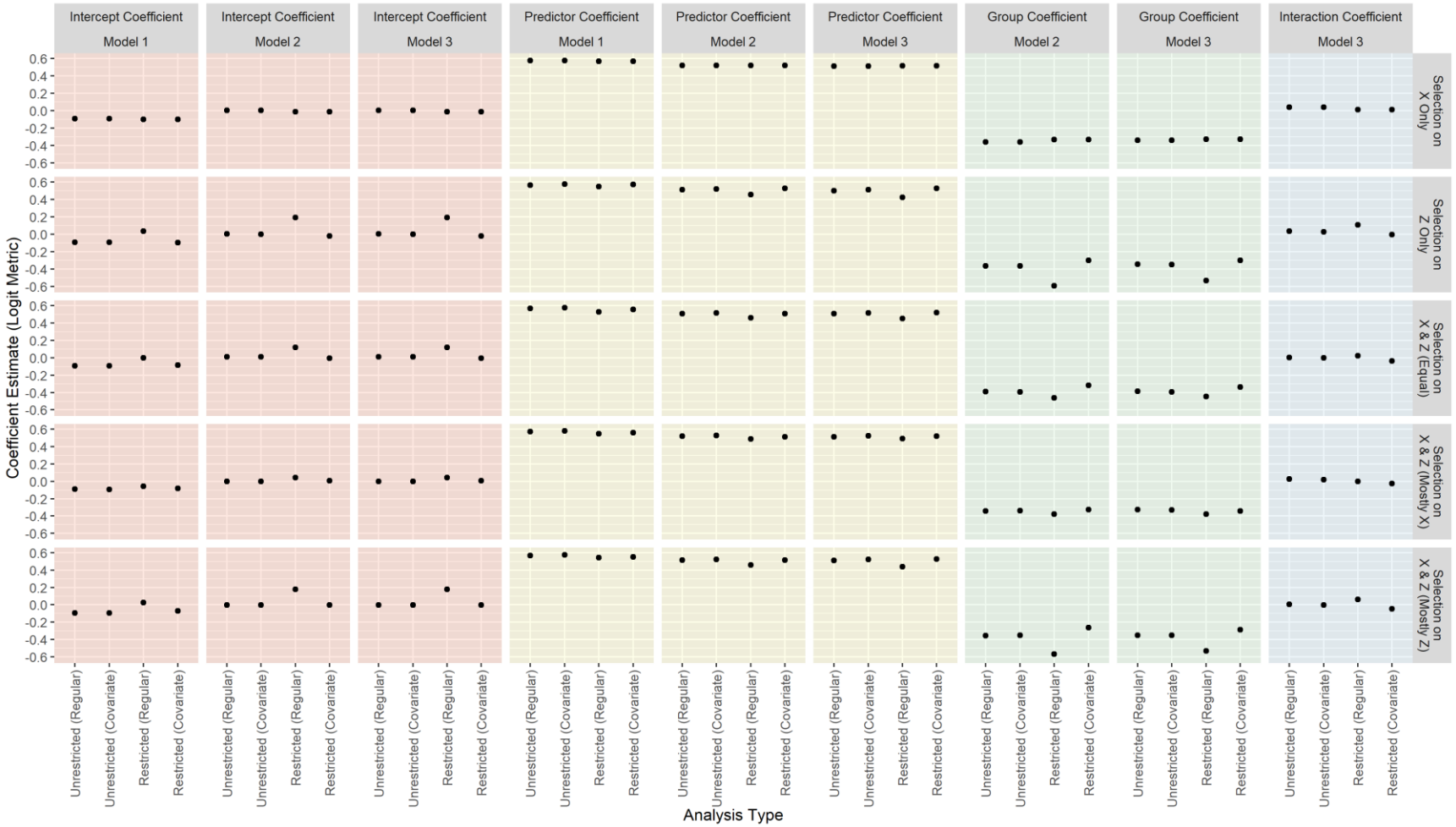

Figure B.17. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Intercept Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

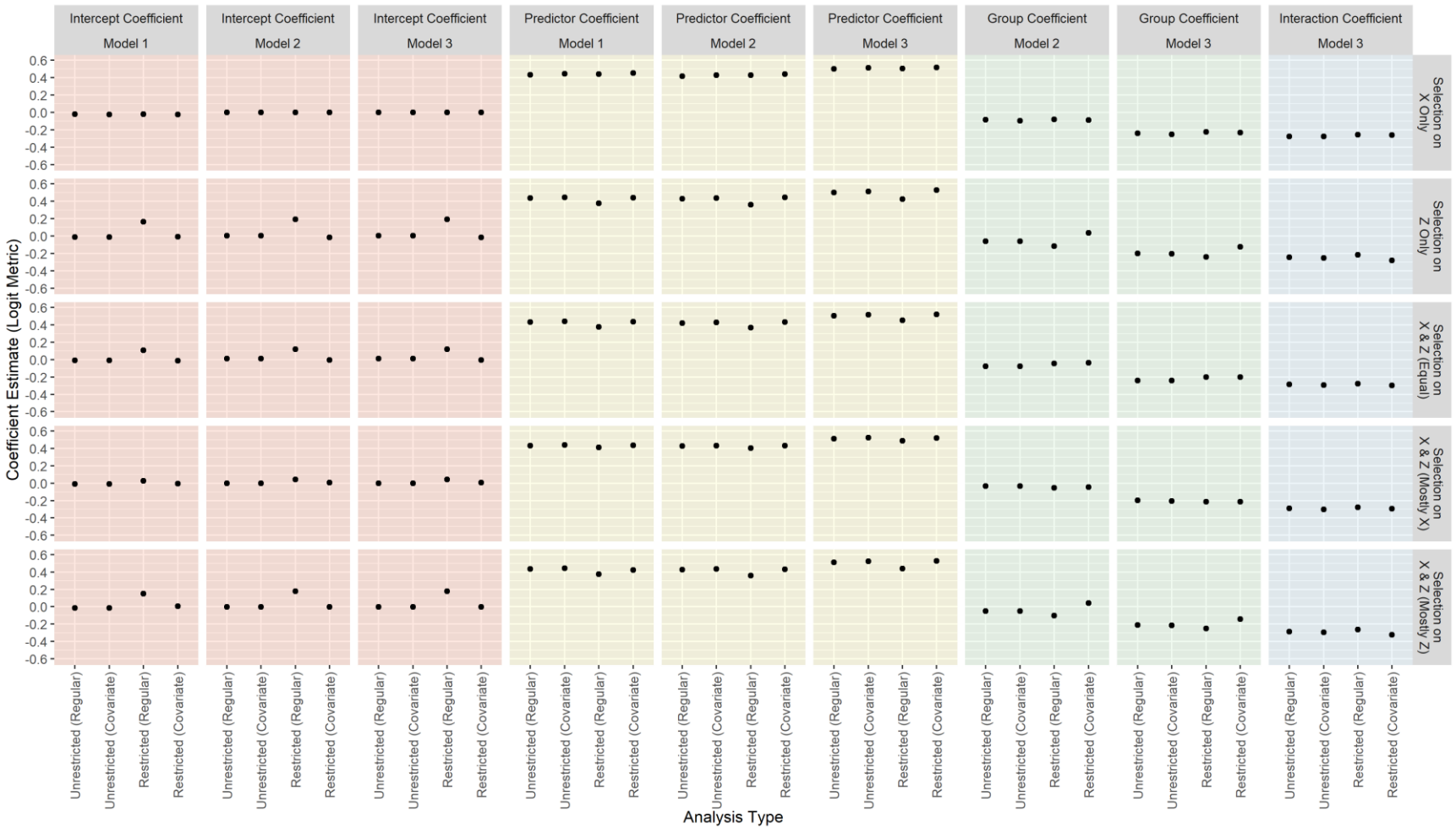

Figure B.18. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Slope Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

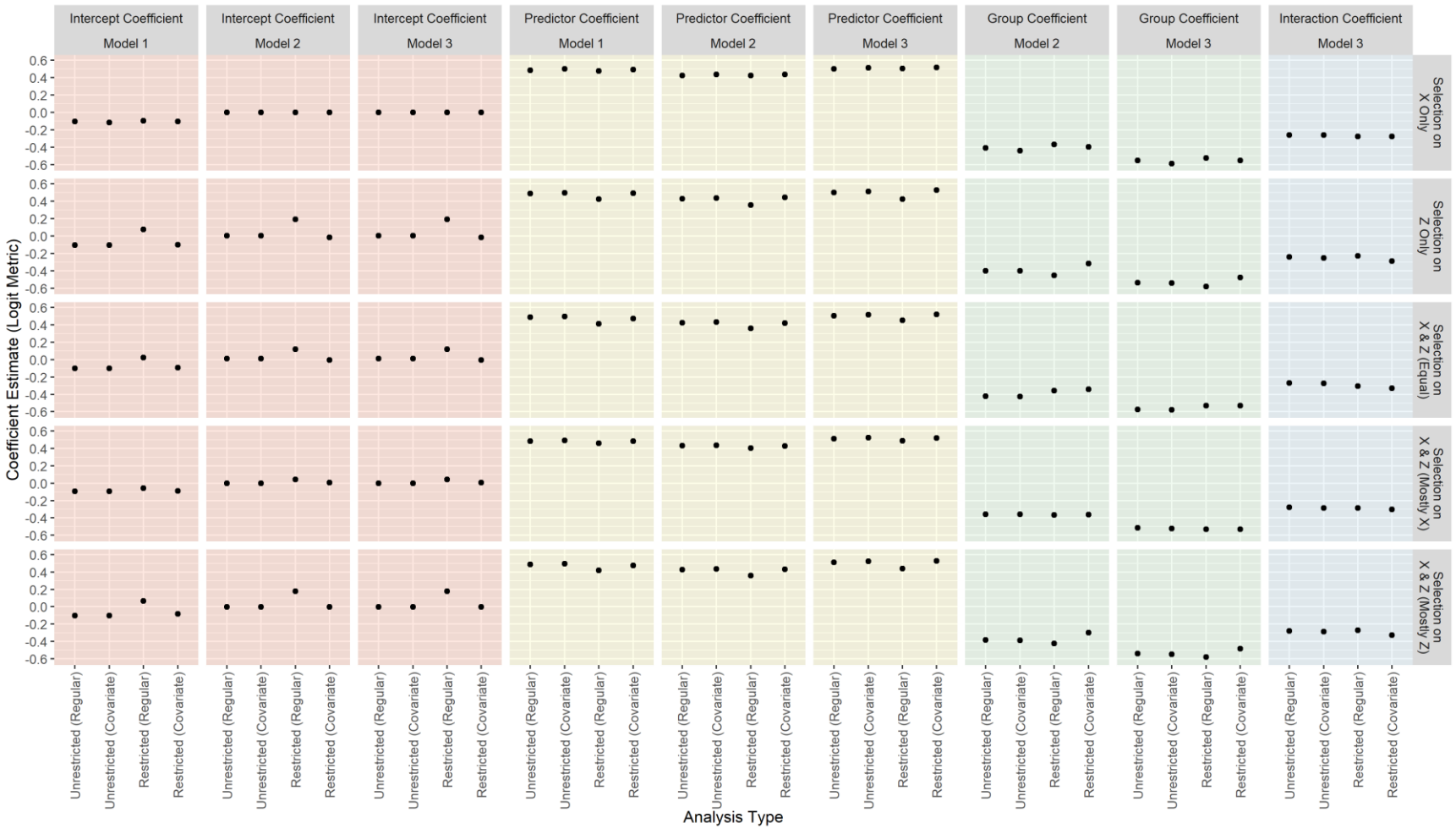

Figure B.19. Average Estimates of Logistic Regression Coefficients Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of regression coefficients and Models 1, 2, and 3 from our Cleary-based analyses. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of regression coefficient. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

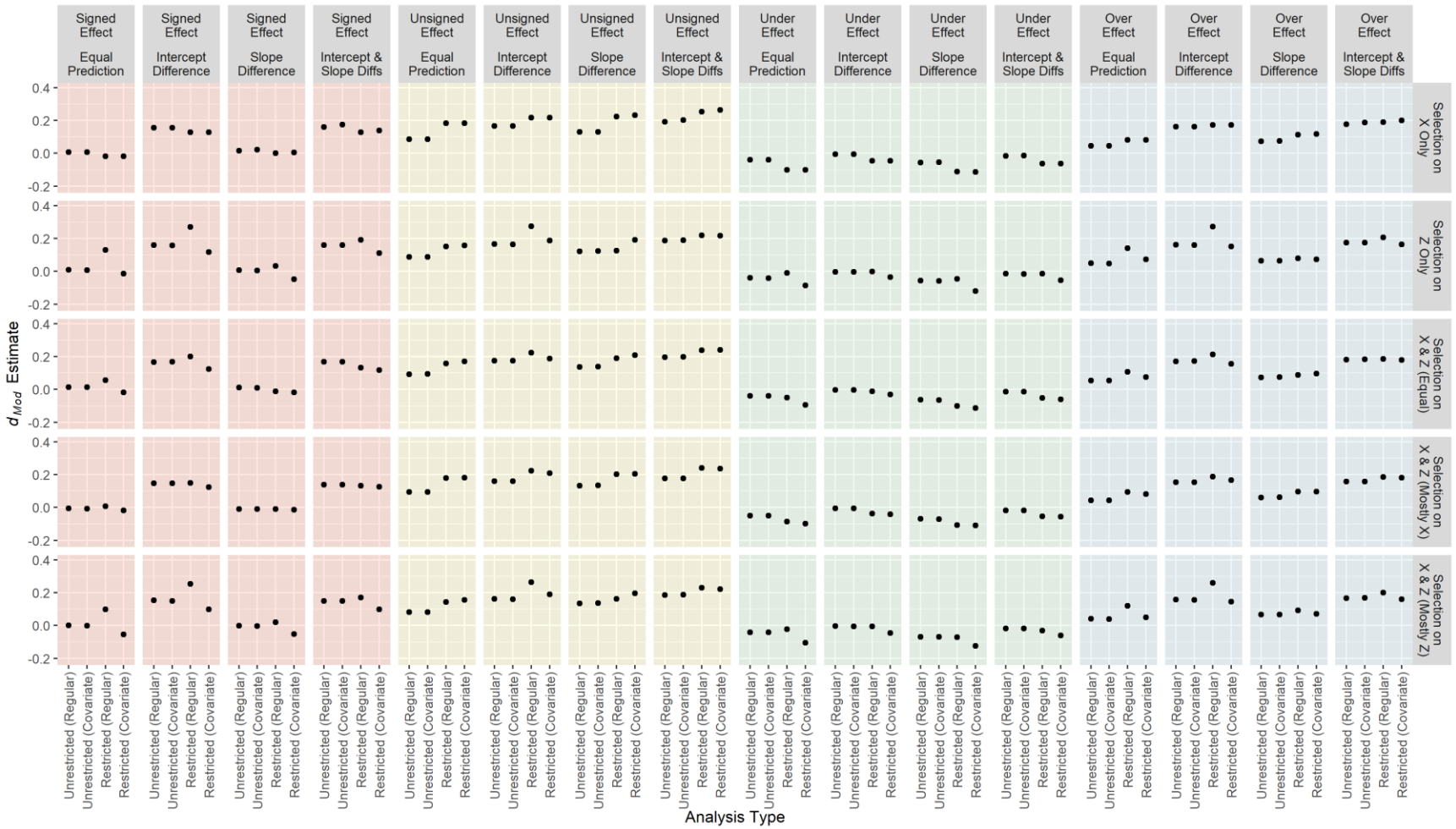

Figure B.20. Average Estimates of Logistic Regression d_{Mod} Effect Sizes Across 100 Simulated Samples from a Population with Intercept and Slope Differences Between Subgroups for the Primary Predictor and Unequal Slopes for the Covariate

Rows of the plot grid represent different methods of introducing selection artifacts, and the columns of the plot grid represent different combinations of d_{Mod} effect types and configurations of subgroup differences in the unrestricted population. Columns of the grid are color-coded to help distinguish among results that correspond to the same type of d_{Mod} effect. Analysis Types arrayed on the *X* axis are either "unrestricted" (i.e., there are no selection artifacts because cases were selected at random) or "restricted" (i.e., the data are impacted by selection artifacts because cases were selected systematically, and, parenthetically, they are also designated as either "regular" (i.e., a traditional Cleary analysis was run without including a covariate) or involving a "Covariate" (i.e., the residuals of the *Z* variable and its interaction with group membership were included as predictors to aid in estimating the coefficients of interest, as described in Chapter 3).

Discussion

We conducted a targeted simulation to evaluate the efficacy of our analysis method for estimating regression coefficients and d_{Mod} effect sizes using data impacted by selection artifacts. We simulated examples of equal subgroup prediction, intercept differences, slope differences, and the co-occurrence of slope and intercept differences for both linear and logistic regression models using four different configurations of selection artifacts. By including covariates that capture information about selection effects that occurred independently of the predictor of interest, we were able to improve our recovery of parameters from the unrestricted applicant population. Our method performed similarly well in both linear and logistic analyses.

During this simulation, we identified characteristics of the d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ effect sizes that make them challenging to estimate and interpret when one's data have been impacted by range restriction. Whereas d_{Mod_Signed} is an efficient estimator that performed quite well with our covariate-based Cleary analyses, the other d_{Mod} effects are more difficult to estimate because of their focus on directional effects. Directional and overall unsigned effects are more volatile to estimate than overall signed effects: Whereas overall signed effects are freely estimated averages that allow positive and negative errors to cancel each other out, directional effects capture these errors independently and overall unsigned effects inherit errors from both directions.

Although we believe d_{Mod_Under} and d_{Mod_Over} have value for understanding differential prediction effects, our simulation results suggest that these effects should be interpreted with caution, especially when d_{Mod_Signed} is small in magnitude. The $d_{Mod\_Unsigned}$ effect size, however, is of limited value in understanding differential prediction because evaluating differential prediction presupposes an interest in knowing the direction of subgroup differences; this, combined with the challenges in estimating $d_{Mod\_Unsigned}$, substantially limits its usefulness in the context of differential prediction research.

This simulation demonstrated that our analysis approach from Chapter 3 functioned as intended. Our covariate-based approach produced coefficient estimates that were much better representations of unrestricted data than what we might achieve by using a traditional formulation of the Cleary method without covariates. Based on the results of this simulation, we proceeded with our analysis approach and applied our methodology to data from the U.S. Armed Services.