Untitled

Marwin Carmo

2022-07-16

What is model selection

Much of the current research questions on behavioral and social sciences are investigated using statistical models [@flora2017statistical]. Models are simplified translations of a reality to mathematical expressions; its' aim is to express how data were generated. Regression-based models are vastly employed in empirical research and often used to estimate causal effects [@berk2010]. However, to perform causal modeling, researchers must specify a "correct" model (i.e., an accurate model of the data generating process) prior to data collection and use the obtained data only to estimate regression coefficients (see @berk2010 for further discussion). In practice, researchers usually only have a vague idea of the right model to answer their research questions, or even if such model can be estimated. Often, what is framed as statistical inference or causal modeling is in fact a descriptive analysis; what @berk2010 name as *Level I Regression Analysis*.

To determine which variables should be included in the model, a common solution is to resort to variable selection algorithms. The drawback is that whenever data-driven variable selection procedures are employed, classical inference guarantees are invalidated due to the model itself becoming stochastic [@berkValidPostselectionInference2013]. It means that if the model selection method evaluates the stochastic component of the data, the model is also considered stochastic.

Why is it a problem?

Variable selection procedures aren't in themselves problematic. Nevertheless, when the correct model is unknown prior to data analysis, and the *same* dataset is used for I) variable selection, II) parameter estimation, and III) statistical inferences, the estimated results can be highly biased. This is because we add a new source of uncertainty when performing model selection. These procedures discards parameter estimates from the model, and, as we shall see, the sampling distribution of the remaining regression parameters estimates gets distorted. In addition, the selected model isn't the same across samples, so there is another source of uncertainty to the estimates. [@berkStatisticalInferenceModel2010].

Consider a well defined population with its unknown regression parameter values. We draw a random sample and apply a model selection procedure. The "best" model found by the variable selection is sample specific and isn't guarantee to be the correct model (if we assume that such model in fact exists). Suppose we repeat the process of drawing a random sample and performing model selection 10,000 times. In this example there are six possible candidates, and only one correct model. We can simulate the expected frequency in which each of these models is chosen given a probability of 1/3 for the correct model and 1/9 otherwise. As shown in the following table, even if the correct model (in this case, $\hat{M}_2$) is three times more likely to be selected than the competing models, it is expected to be chosen more frequently but not at the majority of the time. Therefore, the expected selected model is an incorrect one.

Model	Frequency
M1	0.1226
M2	0.3773
M3	0.1271

Model	Frequency
M4	0.1262
M5	0.1270
M6	0.1198

To understand why the regression parameters estimates might be biased, recall that in a multiple regression we estimate *partial regression coefficients*: in a regression equation, the weight of independent variables are estimated **in relation to** the other independent variables in the model [@cohen1983]. For a dependent variable, Y, predicted by variables X_1 and X_2, $B_{Y1\cdot2}$ is the partial regression coefficient for Y on X_1 holding X_2 constant, and $B_{Y2\cdot1}$ is the partial regression coefficient for Y on X_2 holding X_1 constant. This regression equation is written as:

$$\hat{Y} = B_{Y0\cdot12} + B_{Y1\cdot2}X_1 + B_{Y2\cdot1}X_2 + \varepsilon(\#eq : multreg) \tag{1}$$

where $B_{Y0\cdot12}$ is the model intercept when X_1 and X_2 are held constant and ε is the error term.

The regression coefficient for X_i, for $i = \{1, 2\}$, is model dependent. To see why, let's take a look at the equations for the regression coefficients for X_1 ($B_{Y1\cdot2}$) and X_2 ($B_{Y2\cdot1}$)

$$B_{Y1\cdot2} = \frac{\rho_{Y1} - \rho_{Y2}\rho_{12}}{(1 - \rho_{12}^2)} \times \frac{\sigma_Y}{\sigma_1}(\#eq : beta12) \tag{2}$$

$$B_{Y2\cdot1} = \frac{\rho_{Y2} - \rho_{Y1}\rho_{21}}{(1 - \rho_{21}^2)} \times \frac{\sigma_Y}{\sigma_2}(\#eq : beta21) \tag{3}$$

Here ρ stands for the populational correlation coefficient and σ for the populational standard deviation. Unless we have uncorrelated predictors (i.e. $\rho_{12} = 0$), the value for any of the regression coefficients is determined by which other predictors are in the model. If either one is excluded from the model, a different regression coefficient will be estimated: excluding X_2, for example, would zero all the correlations involving this predictor, leaving,

$$B_{Y1\cdot2} = \frac{\rho_{Y1} - 0 \times 0}{(1 - 0^2)} \times \frac{\sigma_Y}{\sigma_1} = \rho_{Y1} \times \frac{\sigma_Y}{\sigma_1} = B_{Y1}(\#eq : beta12exc) \tag{4}$$

@berkStatisticalInferenceModel2010 warns that the sampling distribution of the estimated regression parameters is distorted because estimates made from incorrect models will also be included, resulting in a mixture of distributions. Therefore, the model selection process must be taken into account in the regression estimation whenever it is applied.

Illustration

We can illustrate the discussion above expanding an analytic example given by @berkStatisticalInferenceModel2010 with simulations. Consider a model for a response variable y with two potential regressors, x and z. Say we are interested in the relationship between y and x while holding z constant, that is, $\hat{\beta}_{yx\cdot z}$. Framing this as a linear regression model we have

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i(\#eq : m1) \tag{5}$$

Now, suppose that we're in a scenario where $\rho_{xz} = 0.5$, both β_1 and β_2 are set to 1 and $\varepsilon \sim N(0, 10)$. We'll use a sample size of 250 subjects, and 1,000 random samples will be drawn from this population. We'll calculate *coverage* and *bias* for each regressor. Coverage informs the frequency in which the true coefficient value is captured by the 95% confidence interval (CI) of the estimate. The bias of the estimations is calculated

as $\frac{1}{R}\sum(\hat{\theta}_r - \theta)$, where R is the number of repetitions, θ represents a population parameter, and $\hat{\theta}_r$ a sample estimate in each simulation.

Predictor	Coverage	Bias
x	0.945	0.0049418
z	0.949	0.0080319

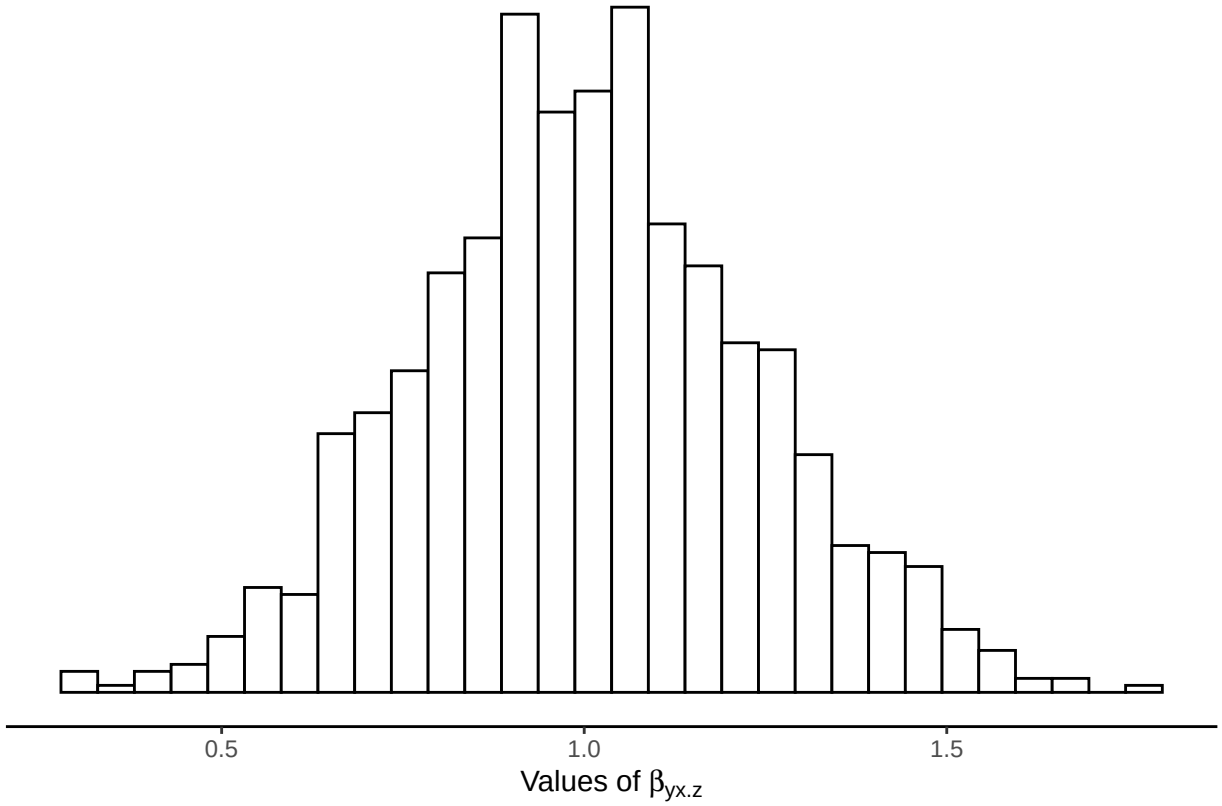

Figure 1: Sampling distributions of the regression coefficient for regressor x in the full model.

In this first scenario $\beta_{yx \cdot z}$ is estimated assuming x and z are always included in the model.

What would happen if by model selection we arrive at a model where z is excluded? As indicated in equation @ref(eq:beta12exc), if z is excluded, any correlation where z is involved is equivalent to zero, and we're left with,

$$\beta_{yx} = \rho_{xy}(\frac{\sigma_y}{\sigma_x})(\#eq : ex11) \tag{6}$$

Note that β_{yx} is not the same as $\beta_{yx \cdot z}$. If we do not have a model specified prior to data collection and analysis, it is not clear which definition of regression parameter for x we're trying to estimate, if $\beta_{yx \cdot z}$, as exemplified on equation @ref(eq:beta12), or if β_{yx} from @ref(eq:ex11). Therefore, the definition of $\hat{\beta}_1$ depends on the model in which it is placed.

Notice how far off the model estimates the coefficient for x when we fit our model as `y ~ x`. Under these conditions, we should expect a coverage of the true coefficient value of x of 0.317 and a bias of 0.29. Under

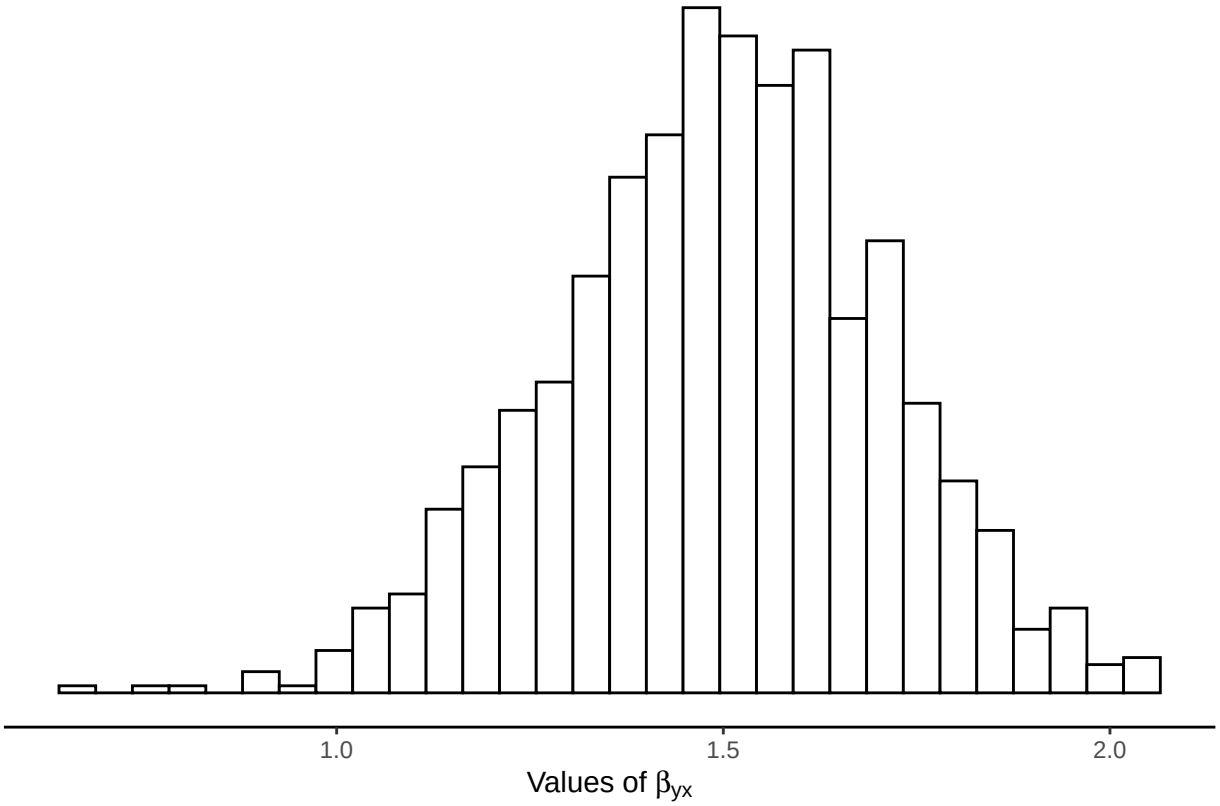

Figure 2: Sampling distributions of the regression coefficient for regressor X in a reduced model where Z is excluded.

the correct model, bias is negligible and coverage follows the Type I error rate of 5% that we've set for this exercise.

This simple example help us understand why estimating model parameters in the same single random sample used to find an appropriate model isn't a good idea (at least not without taking this process into account prior to making inferences). Discarding z from our model has distorted the sampling distribution of x. Thus, as @berkStatisticalInferenceModel2010 puts it: "when a single model is not specified before the analysis begins, it is not clear what population parameter is the subject of study. And without this clarity, the reasoning behind statistical inference becomes obscure".

Other issues

We already know that if two predictors are moderately correlated and one is dropped from the model, our estimation will be biased. But what else should we consider? If we are selecting the variable every time, then there is not an issue. In the following sections we'll consider what can make things go "bad". Before that, please consider the equation for the standard error of the regression coefficient, estimated in the example model selection,

$$SE(\beta_{yx\cdot z}) = \frac{\hat{\sigma_\varepsilon}}{s_x\sqrt{n-1}}\sqrt{\frac{1}{1-r_{xy}^2}}(\#eq:sebxy) \tag{7}$$

We have that $\hat{\sigma_\varepsilon}$ is an estimate of the residual standard deviation, s_x is the sample standard deviation of x, r_{xy}^2 is the square of the sample correlation between x and z, and n is the sample size. If the standard error for a regression coefficient is large, it means that its distribution will be more dispersed. So, from this equation, we identify as crucial parameters for a wider sampling distribution: larger residual variance, less variance in x, smaller sample size, and, stronger correlation between regressors and the response variable.

We'll use simulated data to aid our understanding. I have build a ShinyApp for that end. Its purpose is to illustrate the problems that arise when model selection, parameters estimation and statistical inferences are undertaken with the same data set. Although I will be working with code in this post, most of it can also be reproduced with this app.

For convenience purposes, I'll use the stepwise approach for model selection with AIC as model selection criterion. Some penalties are harsher than others, but the concerns raised here are irrespective of the chosen model selection method [@berkStatisticalInferenceModel2010]. In the app, I also provide the option to choose BIC or Mallow's Cp as selection method. In addition to bias and coverage, we'll also estimate the average Mean Squared Error (MSE), calculated as $\frac{1}{R}\sum[(\hat{\theta_r}-\theta)^2]$.

Noise

To express variability we can use a signal-to-noise ratio (SNR), defined here as:

$$\frac{S}{N} = \mathbf{b}\Sigma\mathbf{b}\sigma^{-2}(\#eq:snr) \tag{8}$$

where, $\mathbf{b}$ is a $(p+1)$ vector of regression coefficients, Σ is the covariance matrix of predictors and σ^2 is the error term variance. Note that "model selection bias also occurs when an explanatory variable has a weak relationship with the response variable. The relationship is real, but small. Therefore, it is rarely selected as significant" [@lukacsModelSelectionBias2009, p.118]. We can experiment with a range of values for the SNR to see how it affects the estimates. We'll use most values set in our first simulation exercise and SNR values ranging from 0.1 to 2. One thousand (1,000) simulations will be run for each of those values.

SNR	N	Predictor	Estimate	Coverage	Bias	MSE
0.01	250	x1	2.2595996	0.290	1.2595996	2.7185468
0.01	250	x2	2.3432900	0.312	1.3432900	2.7149654
0.10	250	x1	1.1592660	0.795	0.1592660	0.1518771
0.10	250	x2	1.1287940	0.802	0.1287940	0.1419542
0.50	250	x1	0.9987398	0.947	-0.0012602	0.0332223
0.50	250	x2	0.9965275	0.940	-0.0034725	0.0333397
1.00	250	x1	1.0042279	0.951	0.0042279	0.0173916
1.00	250	x2	0.9995938	0.955	-0.0004062	0.0160247
2.00	250	x1	1.0042212	0.943	0.0042212	0.0081511
2.00	250	x2	0.9957160	0.941	-0.0042840	0.0089674

As expected, with greater amount of noise in relation to signal, the selected model includes the true coefficient at smaller frequencies. Because model selection interferes with the true model composition, coefficient estimates deviate from its true value. In this particular setting, with $SNR \geq 0.5$, coverage frequency approximates 95% and the estimates get closer to their true value. Measures of bias and MSE are also useful to display how our uncertainty gets smaller with less variability in the data.

Sample size

Once we know @ref(eq:sebxy) it is not difficult to suppose that larger samples produce smaller standard errors for the regression coefficients. We can confirm that using our `sim_bias` function again, but this time varying sample sizes.

SNR	N	Predictor	Estimate	Coverage	Bias	MSE
0.5	10	x1	1.7607713	0.445	0.7607713	1.5161564
0.5	10	x2	1.7785770	0.414	0.7785770	1.6753835
0.5	50	x1	1.1401917	0.823	0.1401917	0.1452009
0.5	50	x2	1.1246416	0.780	0.1246416	0.1465185
0.5	100	x1	1.0129111	0.951	0.0129111	0.0706557
0.5	100	x2	1.0104064	0.939	0.0104064	0.0751206
0.5	200	x1	1.0071304	0.939	0.0071304	0.0445692
0.5	200	x2	0.9993969	0.957	-0.0006031	0.0398318
0.5	500	x1	1.0077521	0.965	0.0077521	0.0149319
0.5	500	x2	0.9943257	0.954	-0.0056743	0.0149595

Candidate predictors

OK, so far we've seen that with 2 predictors, each with true coefficient values of 1 and a correlation of 0.5, we get more precise estimates when $SNR \geq 0.5$ and $n \geq 100$. The number of candidate predictor variables and its covariance matrix are two important aspects not addressed yet. To demonstrate their influence we can run simulations with varying values for each, where each case will have a different combination of number of predictors and correlation between them.

For simplicity, all the off-diagonal elements the correlation matrix of predictors will be equal, meaning the that every predictor in the full model is correlated with the others by the same degree. We'll need to do a slight modification on our previous function so we can then apply a new function for each combination and build a dataframe summarizing our results.

To keep the computing load at reasonable levels, we'll limit the length of vectors of p and ρ to 10 values each. Our grid for ρ ranges from 0.1 to 0.9 and we'll increase p from 2 to 10. We'll set $SNR = 0.5$, $n = 100$ and set all true coefficient values to 1. Again, we're using AIC to select the best model, and performing 1,000 simulation replicates.

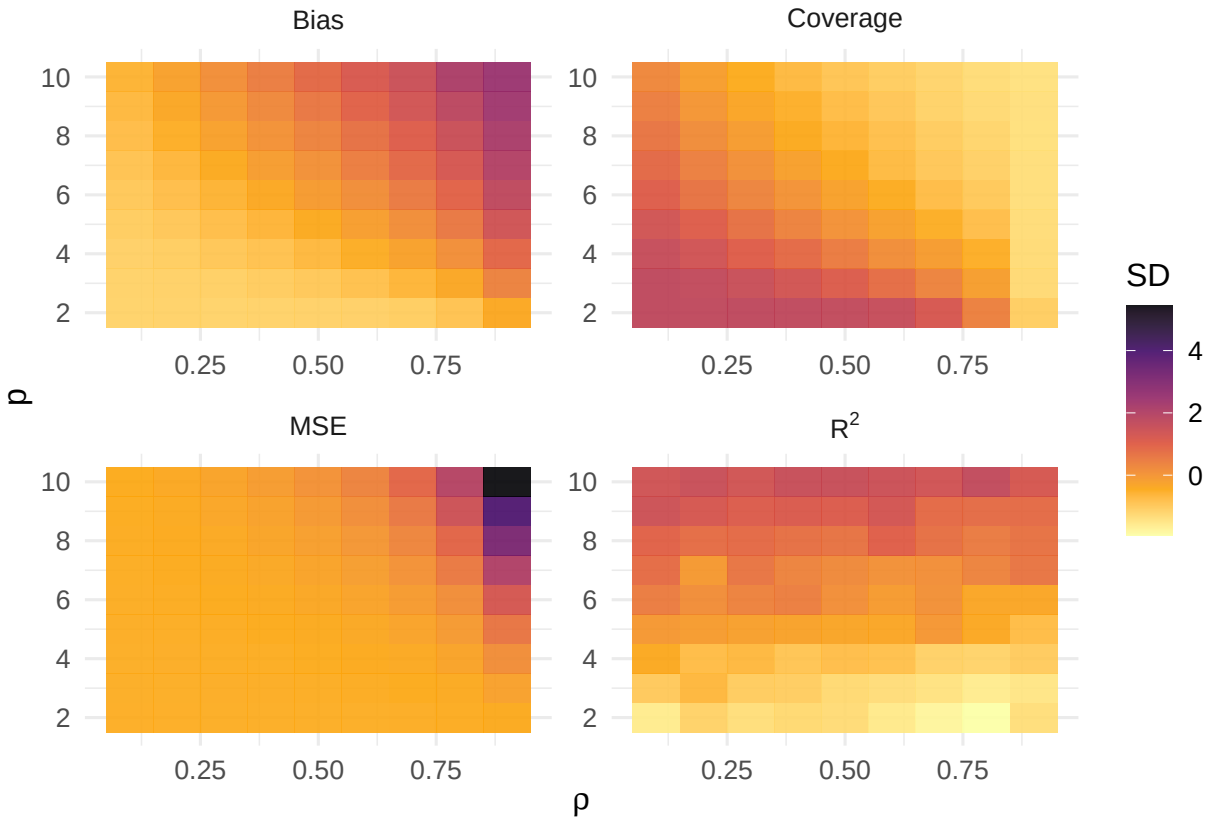

Figure 3: Standardized measures of Bias, Coverage, Mean Square Error (MSE) and R2, across each simulation of pairs of number of predictors and correlation values

Correlation	p	Predictor	Coverage	Estimate	Bias	
0.1	10	x1	0.565	1.48212455427486	0.482124554274857	0.4836 1
0.1	10	x2	0.534	1.50320660781006	0.503206607810065	0.4731
0.1	10	x3	0.575	1.527008856 0898	0.527008856089799	0.4998 4
0.1	10	x4	0.572	1.49374880165201	0.493748801652011	0.4638 0
0.1	10	x5	0.559	1.53159770593366	0.531597705933658	0.5426 4
0.1	10	x6	0.569	1.47138408431054	0.47138408431054	0.4501 7
0.1	10	x7	0.552	1.50399057598131	0.503990575981306	0.4731 8
0.1	10	x8	0.546	1.47555272871 69	0.475552728716897	0.5275 6
0.1	10	x9	0.591	1.52601723355963	0.526017233559626	0.5079
0.1	10	x10	0.553	1.49245028335111	0.49245028335111	0.4753 4

1–10 of 486 rows Previous **1** 2 3 4 5 ... 49 Next

To keep all four plots in the same scale, I have standardized the horizontal axis values. In this heat map, darker colors represent larger values (greater standard deviation). The pattern for Bias and MSE is similar: performing model selection with a great number of highly correlated predictors is likely to produced highly biased estimates. As expected, Coverage follows an inverted pattern from Bias and MSE: highly biased estimates fall out of the 95% CI coverage.

These plots make clear that our best case scenario is to have few, weakly correlated predictors. Of course, this is a setting rarely seen in empirical research – whats the point of going through model selection if you have only a handful of variables? What should really get our attention here is that bias in estimates is likely to get bigger if model selection is naively performed with little prior information about the data generating system to filter out candidate models. We have seen that sample size, SNR, number of predictors, and correlation between parameters with themselves and with the response variable, are characteristics of the model that each in their own way contributes to producing biased estimates. I encourage you to use this companion shinyapp to play with different scenarios to practice what we've been discussing so far.

Visualizing distributions

We've discussed earlier how dropping a predictor from the model can distort the coefficient sampling distribution of the remaining ones when they're not orthogonal. However, even when the preferred (or "correct") model is selected, there is no guarantee about obtaining sound regression coefficient estimates. In this final example, we'll show that if by chance the "correct" model is achieved by model selection, the sampling distributions of the resulting regression coefficients might be different whether we condition on arriving at the correct model or on the correct model being known in advance.

Consider this example from @berkStatisticalInferenceModel2010. As with the other examples, we'll implement forward stepwise regression using the AIC as a fit criterion. The full regression model takes the form of

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \beta_3 z_i + \varepsilon_i (\#eq : b10) \tag{9}$$

where $\beta_0 = 3.0$, $\beta_1 = 0.0$, $\beta_2 = 1.0$, and $\beta_3 = 2.0$. The variances and covariance are set as: $\sigma_\varepsilon^2 = 10.0$, $\sigma_w^2 = 5.0$, $\sigma_x^2 = 6.0$, $\sigma_z^2 = 7.0$, $\sigma_{w,x} = 4.0$, $\sigma_{w,z} = 5.0$, and $\sigma_{x,z} = 5.0$. The sample size is 200.

Note that @berkStatisticalInferenceModel2010 uses the term *preferred* model instead of *correct* model. They do so because a model that excludes W can also be called correct the same as one like @ref(eq:b10), as long as $\beta_1 = 0$ is allowed. To be consistent with the original text, we'll use *preferred* to refer to the model with W excluded. This model is preferred because it generates the same conditional expectations for the response using up one less degree of freedom. The plots show the regression estimates t-values. "A distribution of t-values is more informative than a distribution of regression coefficients because it takes the regression coefficients and their standard errors into account" [@berkStatisticalInferenceModel2010, p. 266].

In both figures @ref(fig:predinc) and @ref(fig:prefcond) the red density plot represents the regression estimates t-values distribution when no model selection is performed. In @ref(fig:predinc) the blue density plot show the distributions when either X or Z are included in the model. For @ref(fig:prefcond) the blue distribution refers to distributions of either X or Z t-values when the preferred model is selected.

The contrast between red and blue curves are apparent. The difference that is the most striking are in t-values distributions post-model-selection for regressor Z when conditioned on the regressor being included in a model. This curve displays a bimodal distribution and highly biased mean and standard deviaton, as summarized below.

Model	M_x	σ_x	M_z	σ_z
Full model	2.166	1.020	4.784	1.039
Preferred model	2.550	0.776	4.487	0.934
X included	2.550	0.776	4.487	0.934

Model	M_x	σ_x	M_z	σ_z
Z included	2.550	0.776	5.757	2.593

It is especially telling observing those plots that the assumed underlying distribution can be very different from what is obtained. Statistical inference performed in such scenarios would be misleading. Figure @ref(fig:prefcond) confirms that conditioning on arriving at the preferred model does not guarantee trustable estimates.

Conclusion

Model selection methods are routine in research on the social and behavioral sciences, and commonly taught in applied statistics courses and textbooks. However, little is mentioned about the biased estimates obtained when such procedures are carried. With simulated data we have identified specific characteristics of the data generating model that can potentially increase the bias in estimates obtained through variable selection. Thus, we can conclude that post-model-selection sampling distribution can deviate greatly from the assumed underlying distribution, even when the best representative model of the data generation process has been selected.

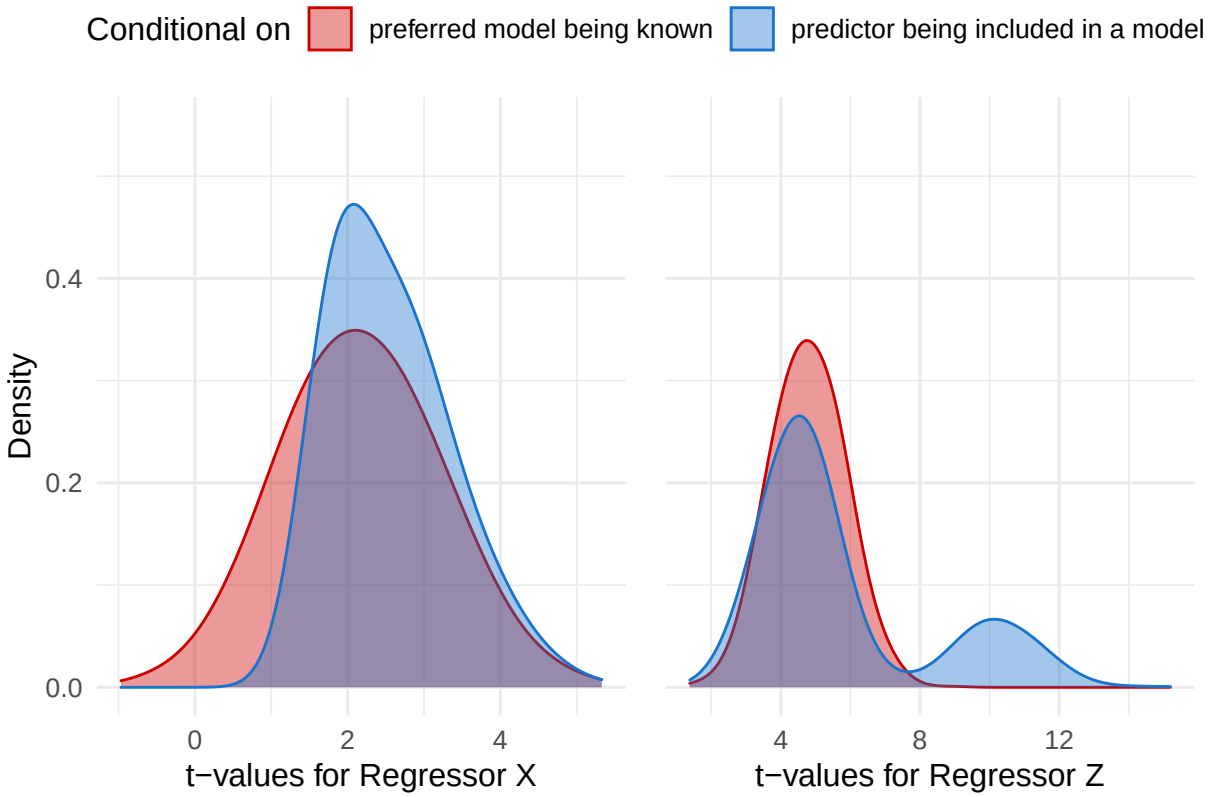

Figure 4: Stepwise regression sampling distributions of the regression coefficient t-values for regressors X and Z. Red density plot is is conditional on the preferred model being known. The blue density plot is conditional on the regressor being included in a model

Figure 5: Stepwise regression sampling distributions of the regression coefficient t-values for regressors X and Z. Red density plot is is conditional on the preferred model being known. The blue density plot is conditional on the preferred model being selected