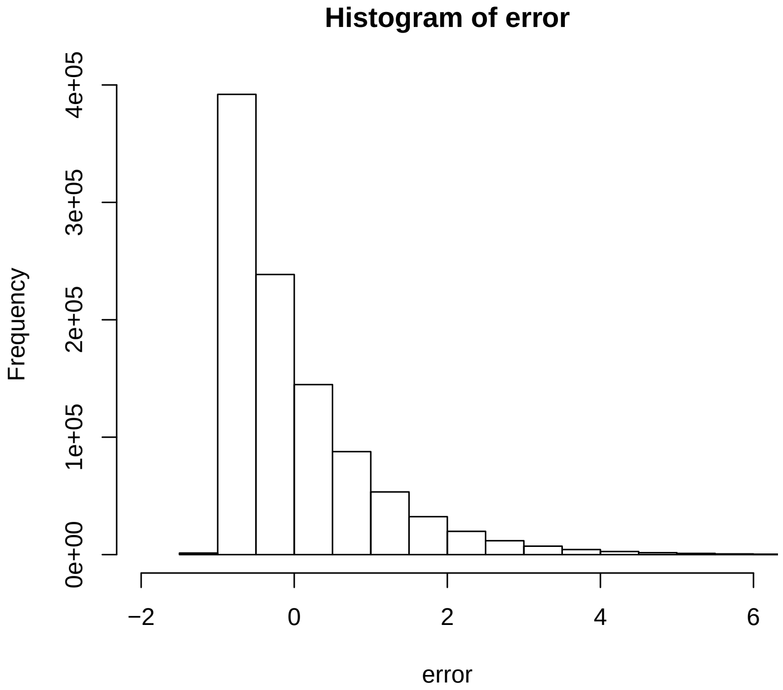
STA4241 Interactive Lab Week 7: Bootstrap and misspecification

Today we are going to learn more about the bootstrap. In particular, we will see how the bootstrap can affect inferences when modeling assumptions are incorrect.

(1) Let's investigate the bootstrap as a method for providing confidence intervals of linear regression parameters. We will explore three types of bootstraps here and discuss the various assumptions that are being made, and compare them with typical assumptions made in linear regression. We are going to explore the situation where we have one covariate, X, and the outcome is generated from

$$Y_i = 2X_i + X_i^3 + \epsilon_i$$

and our error distribution is a shifted and scaled version of a gamma distribution that is highly right skewed. The distribution of our errors looks like the following:



We are going to fit the traditional simple linear model

$$E(Y|X) = \beta_0 + \beta_1 X_i$$

We can see that two linear model assumptions are being broken here:

1. The linear relationship between X and Y does not hold
2. The error distribution is not normally distributed

We will focus on estimation of β_1. The error distribution should not affect our estimates of β_1, but it can affect our inferences about β_1 in small samples. The linear relationship not holding will affect our estimates. The linear model we're assuming will attempt to find the closest linear approximation to the nonlinear relationship. The limiting values (what we would get if we had an enormous sample) of β_0 and β_1 under model misspecification are the ones that minimize

$$E[(Y - \beta_0 - \beta_1 X)^2]$$

It turns out in this case that this value of β_1 is 5, though that won't be important for our discussion on bootstrap intervals. We will discuss three bootstrap approaches to estimating the standard error to see how they do in this case

1. Parametric bootstrap
2. Residual bootstrap
3. Nonparametric bootstrap

The steps of the parametric bootstrap are as follows:

1. Estimate $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2$ from the data
2. Generate new outcomes via $Y_i^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \epsilon_i^{(b)}$ where $\epsilon_i^{(b)} \sim N(0, \widehat{\sigma}^2)$
3. Estimate the linear model coefficients using the new outcome
4. Do this for bootstrap samples $b = 1, \dots, B$

The steps of the residual bootstrap are as follows:

1. Estimate $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ from the data
2. Calculate the residuals as $e_i = Y_i - \widehat{Y}_i$
3. Randomly sample n residuals with replacement. Call these $e_i^{(b)}$.
4. Create new outcomes via $Y_i^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i^{(b)}$
5. Estimate the linear model coefficients using the new outcome
6. Do this for bootstrap samples $b = 1, \dots, B$

And the nonparametric bootstrap is just as we discussed in class. These three bootstraps make varying degrees of assumptions about the model. The parametric bootstrap makes both key assumptions about the model: Both the linear model is correct, and that the distribution of the errors is normally distributed. The residual bootstrap also assumes that the linear model is correct, but does not assume normality of the errors. It uses a nonparametric estimate of the error distribution by simply sampling from replacement from the observed residuals. If the true errors are not normally distributed, then these residuals won't be normally distributed either, and the residual bootstrap should work better than the parametric bootstrap. The nonparametric bootstrap on the other hand makes no such assumptions about the linear model.

Let's look at simulation in the R code to see how these perform.

(Overview) Why do any of these approaches work at all? Our goal is to estimate the standard deviation of the sampling distribution of our test statistic, which in this case is the regression coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$. We assume that our data come from some unknown distribution F, and we need to approximate it in some way. Each bootstrap approximates it in different ways:

1. Nonparametric: $\widehat{F}_n \approx F$
2. Parametric: $\text{Normal}(\widehat{\beta}_0 + \widehat{\beta}_1 X, \widehat{\sigma}^2) \approx F$
3. Residual: $F(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{e}_i) \approx F$

Each making different assumptions when approximating F. There are other assumptions that the residual and parametric bootstrap make. For instance, they assume that the errors are not a function of X. It is possible that the residual variance is a function of X, in which case the nonparametric bootstrap would be best. Another subtle difference that we discussed in class is that the nonparametric bootstrap treats the X variables as random, while the other two bootstraps (and analytic standard errors) assume that X is fixed. Essentially, the nonparametric bootstrap states that

$$(X_i, Y_i) \sim \widehat{F}_n$$

and draws new values of both X and Y when re-sampling, which leads to additional uncertainty. The parametric bootstrap assumes that

$$Y_i | X_i \sim \text{Normal}(\widehat{\beta}_0 + \widehat{\beta}_1 X, \widehat{\sigma}^2)$$

and does not draw new values of X. As we have seen today, even if we want to treat the X variables as fixed and known quantities, it might be best to use the nonparametric bootstrap simply because it is robust to the strong assumptions that the other approaches to inference make. Additionally, we can see at the end of the code that when the linear model is correctly specified and all model assumptions are true, then the nonparametric bootstrap gives extremely similar results to the other approaches, even though it adds an additional source of uncertainty. This additional uncertainty is negligible in this case.