1. No Perfect Collinearity Review (MLR3)[1]

i. Definition

Two (ore more) variables are said to be perfectly collinear if one variable is a linear combination of the other variable(s) (where a linear combination of x_1 would be any expression of the form $ax_1 + b$). When do we run up against multicollinearity?

- The simplest case of perfect collinearity is when one variable is a constant multiple of another. This might happen if a researcher inadvertently decides to include income measured in dollars as well as income measured in thousands of dollars.
- Perfect collinearity also arises when one variable can be expressed as an exact linear combination of two or more of the other independent variables. Wooldridge p.85 gives the following example: we want to estimate the effect of campaign spending on voting outcomes. Therefore, we run the following regression:

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u$$

  where $voteA$ is the percentage of vote for Candidate A, $expendA$ are campaign expenditures for Candidate A, $expendB$ are campaign expenditures for Candidate B, and $totexpend$ are total expenditures on both campaigns. Here $totexpend = expendA + expendB$, which means we run into perfect multicollinearity.

ii. Intuition

Why is multicollinearity a problem? So what if we have two variables that tell us the exact same information? Let's think about this in the "holding variables constant framework" that we introduced when interpreting parameters estimates in the multiple linear regression analysis (see Section 3 notes).

If I include two variables x_1 and x_2 that are perfectly collinear (which means they vary in the exact same way) and I hold one of these variables constant (i.e., I hold x_2 fixed and don't allow it to move) then I can't by definition allow my variable of interest x_1 to increase, which is what I want to do to determine it's effect on y! So in the previous example: β_1 is supposed to measure the effect of increasing expenditures on Campaign A by one dollar, holding total spending fixed (and Campaign B's expenditures fixed). But this doesn't make sense, because if $totalexpend$ are held fixed, then we can't increase $expendA$.

iii. Caveat

This assumption doesn't rule out correlation between variables, just perfect multicollinearity. Indeed, if we didn't allow for any correlation between variables, it wouldn't make much sense to perform multiple regression analysis.

[1] Many thanks to Erin Kelley for creating the original notes for this class.

iv. Exercise 3.12 Woolridge p.109

The following equation represents the effects of tax revenue mix on subsequent employment growth for the *population* of counties in the United States:

$$growth = \beta_0 + \beta_1 share_P + \beta_2 share_I + \beta_3 share_S + other factors$$

where *growth* is the percentage change in employment from 1980 to 1990, $share_P$ is the share of property taxes in total tax revenue, $share_I$ is the share of income tax revenues, and $share_S$ is the share of sales tax revenues. All of these variables are measured in 1980. The omitted share, $share_F$, includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

- Why must we omit one of the tax share variables from the equation? The shares add to one. If we do not omit one of the shares then we run into a problem of perfect multicollinearity
- Give a careful interpretation of β_1. Because each share is a proportion (and can be at most one, when all other shares are zero), it doesn't make sense to increase $share_P$ by one unit. If $share_P$ increases by .01 (= a one percentage point increase in the share of property taxes in total revenue) holding all other shares and all other factors constant, then growth increases by $\beta_1(0.01)$

v. Near Perfect Multicolinearity

Recall that (under assumption MLR5) the formula for the variance of the estimator $\hat{\beta}$ is given by

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

where $SST = \sum_j (x_{ij} - \bar{x})^2$ is the total sample variation in x_j, and R_j^2 is the R squared from regressing x_j on all other independent variables (and including an intercept).

If it is the case that x_j is a perfect linear combination of some of the other independent variables, then R_j^2=1: other variables explain all of the variation in x_j. You can see that if this is the case, the formula for $Var(\hat{\beta})$ has a zero in the denominator, and we can't actually calculate it. However, if we have highly correlated/collinear variables, this will still produce a high R_j^2 (e.g. 0.9) which will blow up the variance of our estimator. For example, taking an $R_j^2 = 0.9$ this implies that $1 - 0.9 = 0.1$, which in the denominator will increase the variance of $\hat{\beta}_j$ by 10 compared to an R_j^2 of zero.

2. Zero Conditional Mean and Omitted Variable Bias

Recall MLR4, which says that $\mathbb{E}[u|x_1 \cdots x_k] = 0$. This is the key assumption needed to get an unbiased estimate of β_1. If this assumption does not hold then we can't expect our estimate $\hat{\beta}_1$ to be close to the true value β_1. One way this assumption fails is by omitting important variables, leading to omitted variable bias (OVB). That is, failing to include a key variable in the model $\rightarrow \mathbb{E}[\hat{\beta}_1] \neq \beta_1$. A motivation of multiple regression is therefore to take this omitted variable out of the error term by including it in our estimation.

i. Example: Building Intuition

Suppose we want to understand how car theft rates (per capita) are affected by changes in financing for job training programs (per capita). Presumably, giving job training programs more resources will lower the rate of car thefts in a given area by increasing the economic return to legal employment. Let us imagine that the population model of car thefts looks like this, with an index of gang presence in a given district as an additional explanatory variable.

$$cartheft = \beta_0 + \beta_1 jobtrainingfinance + \beta_2 gangs + u$$

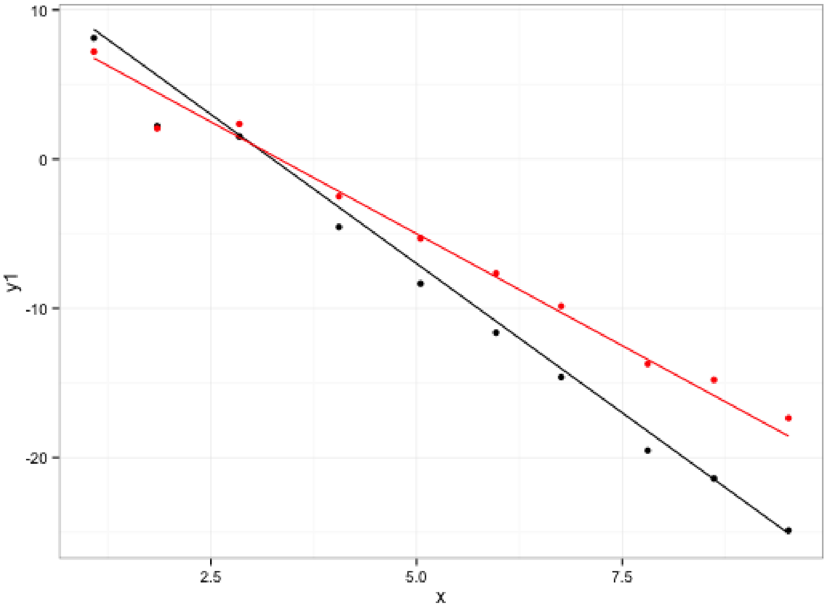
However, let's pretend that we didn't think to collect data on prevalence of gangs in each district, so that the model we have in mind is:

$$cartheft = \tilde{\beta}_0 + \tilde{\beta}_1 jobtrainingfinance + u$$

So what happens if we omit a variable like *gangs* above? Well we can sign the bias we expect by determining the signs of two correlations:

1. Think of how the omitted variable is correlated with our dependent variable y: here $Cov(cartheft, gangs)$ or $Cov(y, omitted\ variable) > 0$ because the presence of gangs is generally associated with higher car theft rates.
2. Think of how the ommitted variable is correlated with our independent variable x: here $Cov(jobtrainingfinance, gangs)$ or $Cov(x, omitted\ variable) > 0$ because we assume that areas with higher gang activity get more financing for job training programs.

When we forget to include gangs in the intial regression, we find that financing for job training programs reduces car thefts by -2 (red line in next graph).We might be tempted to say that a \$1000 increase in financing for job programs leads to a decrease in car thefts of 2. Is this correct?



Our intuition is No. So suppose we collect a little more data on gangs. We now run the regression and plot the relationship: we find it is actually -4: \$1000 increase in financing for job training programs leads to a decrease in car thefts of 4 (black line in graph). Why? What is the direction of the bias?

Well we knew that before we hadn't controlled for gangs. Moreover, there will be cities in our data where there are lots of gangs, which will also have high financing for job training programs and higher car thefts (because the correlations between the omitted variable and our independent/dependent variable are positive). So if we fail to include "gangs", it's going to look like more financing for job training programs in these cities is associated with higher rates of car thefts. Therefore we will have upward bias in our overall estimate of the effect of financing for job training programs on cities (if the biased estimate is -2 and the unbiased estimate is -4: $(-2-(-4)) = +2$ upward bias - you don't have to do this math just think about moving from your biased to your unbiased estimate).[2]

When you are doubting your intuition refer to the following table:

	$Cov(x, x_{ov}) > 0$	$Cov(x, x_{ov}) < 0$
$Cov(y, x_{ov}) > 0$	Upward bias	Downward bias
$Cov(y, x_{ov}) < 0$	Downward bias	Upward bias

Summary: same sign correlations → upward bias; different sign correlations → downward bias

ii. Exercise

Consider the following regressions:

$$\widehat{\ln(wage)} = 1.19 + 0.101educ + 0.011exp$$
$$\widehat{\ln(wage)} = 1.06 + 0.117educ + 0.011exp - 0.25female$$

- How does coefficient on education change when we add in a dummy variable for being female. Is the bias upward or downwar when we don't include $female$? The coefficient on education increased when we included female: we know the unbiased estimate is 0.117 while the bias estimate is 0.101: 0.101-0.117= -0.016, which is negative. Our estimate was downward biased.
- Is female positively or negatively correlated with education? The coefficient on female is negative: $cov(ln(wage), female) < 0$, and the bias on the estimated impact of education is negative (downward bias). Therefore the $cov(educ, female) > 0$. Intuition: women earn less in general but also tend to have higher education, so if we don't control for that we will see people with higher education getting lower wages, thereby downplaying the effect of education.

[2] If it helps, think of the case where \$1000 of financing for job training programs is associated with -4 car thefts (true effect) and 1 additional gang is associated with +2 car thefts. If I fail to include gangs, I confound the effects of gang and financing for job training programs, and I will find that the slope coefficient on financing for job training programs is -2(-4+2) rather than -4 (which is what I get if I hold the effect of gangs on car thefts constant)

iii. Caveat: Including Irrelevant Variables

What happens if we include a variable that doesn't have any effect on y (holding other variables constant)? In other words its population coefficient is zero, but we include it in our sample regression function anyways? For example, we estimate:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Even though the population regression function is $E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. There is no effect in terms of the unbiasedness of $\hat{\beta}_1$ and $\hat{\beta}_2$. It will though affect the variance of our estimator, making it less precise (see final section of notes).

iv. Some formal notation

Bias

We can view omission of a set of relevant variables as equivalent to imposing an incorrect restriction on the population model. Such an omission may be the consequence of an erroneous exclusion of a variable for which data are available or of exlusion of a variable that is not direclty observed.

The consequence: our estimator will be **biased**. In words the bias of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated.

Proving estimators are unbiased

An estimator is unbiased when:

$$E[\hat{\beta}_1] = \beta_1$$

So on average, your estimate of β_1 doesn't miss the mark: it's neither too high, nor too low.

Proof that $\hat{\beta}$ from our linear regression model is unbiased:

$$\begin{aligned}
E(\hat{\beta}_1) &= E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] \\
&= E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] \quad \text{See Appendix A Woolridge} \\
&= E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i+\mu_i)}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] \\
&= \beta_0 \cdot E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] + \beta_1 \cdot E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] + E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] \\
&= \beta_1 \cdot E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] + E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] \quad \text{Recalling that } \sum_{i=1}^{n}(x_i-\bar{x}) = 0 \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}\right] \quad \text{Canceling} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})x_i} E\left[u_i|x_1,\cdots x_n\right] \\
&= \beta_1 \qquad \text{SLR2 and SLR4}
\end{aligned}$$

Another common example of an unbiased estimator: sample mean $\bar{X} = \sum_{i=1}^{N} X_i$ is an unbiased estimator of the population mean μ of a random variable X. Saying that the sample mean is an unbiased estimate of the population mean simply means that there is no systematic distortion that will tend to make it either overestimate or underestimate the population parameter.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E(X_i) = \frac{1}{n}\sum_{i}^{n}\mu = \frac{1}{n}n\mu = \mu$$

Recall when we defined sample variance in Section 1, we divided by $n-1$. This was precisely to ensure that the estimator was unbiased. If we only divided by n the estimator was systematically undersestimating the population variance.

Omitted Variable Bias

Let's begin with a case where the true population model has two explanatory variables and an error term:

$$y = \beta_0 + \beta_1 x + \beta_2 z + u$$

We are interested in the effect of x on y. Think of y as hourly wage (or log of hourly wage), x = education, and z = a measure of innate ability. Further suppose that education and innate ability are correlated, with correlation coefficient ρ from a regression of z on x. To get an unbiased estimator of β_1 we need to regress y on x and z. However, we don't have information on ability (z), and so we estimate the model by *excluding* z. Instead we estimate

$$y = \beta_0 + \beta_1 x + v$$

where $v = \beta_2 z + u$. We perform the simple regression of y on x and obtain the equation (tilde to emphasize that $\tilde{\beta}_1$ comes from an underspecified model):

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$$

Now to see why this leads to bias, we need to plug in the true population model, $y = \beta_0 + \beta_1 x + \beta_2 z + u$, into our formula for $\tilde{\beta}_1$ and simplify:

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} \\
&= \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 z_i + u_i)}{\sum_i (x_i - \bar{x}) x_i} \\
&= \frac{\beta_0 \sum_i (x_i - \bar{x}) + \beta_1 \sum_i (x_i - \bar{x}) x_i + \beta_2 \sum_i (x_i - \bar{x}) z_i + \sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i} \\
&= \beta_1 + \beta_2 \frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i} + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i} \qquad \text{[Cancelling out and recalling that } \sum_{i=1}^n (x_i - \bar{x}) = 0]
\end{aligned}$$

There's an extra term! The second term $\beta_2 \frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i}$ is a result of our omission of the variable z that affects y. If we take the expectation of $\tilde{\beta}_1$:

$$\begin{aligned}
\mathbb{E}\left[\tilde{\beta}_1\right] &= \mathbb{E}\left[\beta_1 + \beta_2 \frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i} + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i}\right] \\
&= \beta_1 + \beta_2 \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i}\right] + \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i}\right] \\
&= \beta_1 + \beta_2 \rho
\end{aligned}$$

When $\mathbb{E}\left[\tilde{\beta}_1\right] \neq \beta_1$ then we say $\tilde{\beta}_1$ is biased. What this means is that on average, our regression estimate is going to miss the true population parameter by $\beta_2 \rho$. Note: ρ is the parameter estimate from a regression of z_i on x_i. You can always use this expression to sign the bias. Just recall that the sign of β_2 is obtained from thinking about how your dependent variable y is correlated with your omitted variable ($cov(y, x_{ov})$) and the sign of ρ is obtained from thinking about how your independent variable of interest x is correlated with your omitted variable ($cov(x, x_{ov})$).

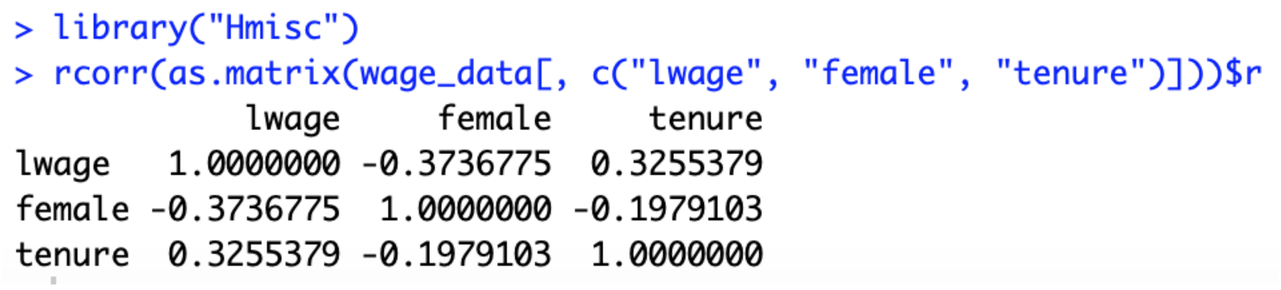
3. OVB Worked Examples

Example 1

In this section, we use the wage data (WAGE1.dta) from your textbook to demonstrate the consequences of omitted variable bias and prove to you that the OVB formula works. Let's pretend that this sample of 500 people is our whole population of interest, so that when we run our regressions, we are actually revealing the true parameters instead of just estimates. We're interested in the relationship between wages and gender, and our "omitted" variable will be tenure (how long the person has been at his/her job). Suppose our population model is:

$$\log(wage)_i = \beta_0 + \beta_1 female_i + \beta_2 tenure_i + u_i \qquad (1)$$

First let's look at the correlations between our variables and see if we can't predict how omitting tenure will bias $\hat{\beta}_1$:



```
> library("Hmisc")
> rcorr(as.matrix(wage_data[, c("lwage", "female", "tenure")]))$r
            lwage     female     tenure
lwage   1.0000000 -0.3736775  0.3255379
female -0.3736775  1.0000000 -0.1979103
tenure  0.3255379 -0.1979103  1.0000000
```

So we have

- $Cov(lwage, tenure)$ or $Cov(y, omitted\ variable) > 0$
- $Cov(female, tenure)$ or $Cov(x, omitted\ variable) < 0$

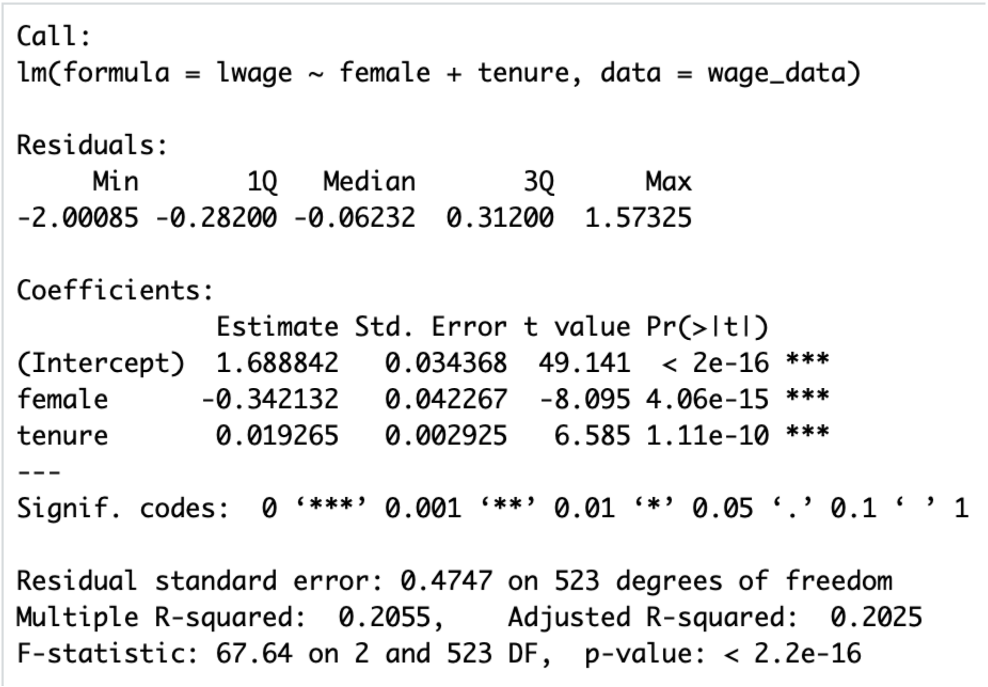
If we ran the regression instead:

$$\log(wage)_i = \tilde{\beta}_0 + \tilde{\beta}_1 female_i + v_i \qquad (2)$$

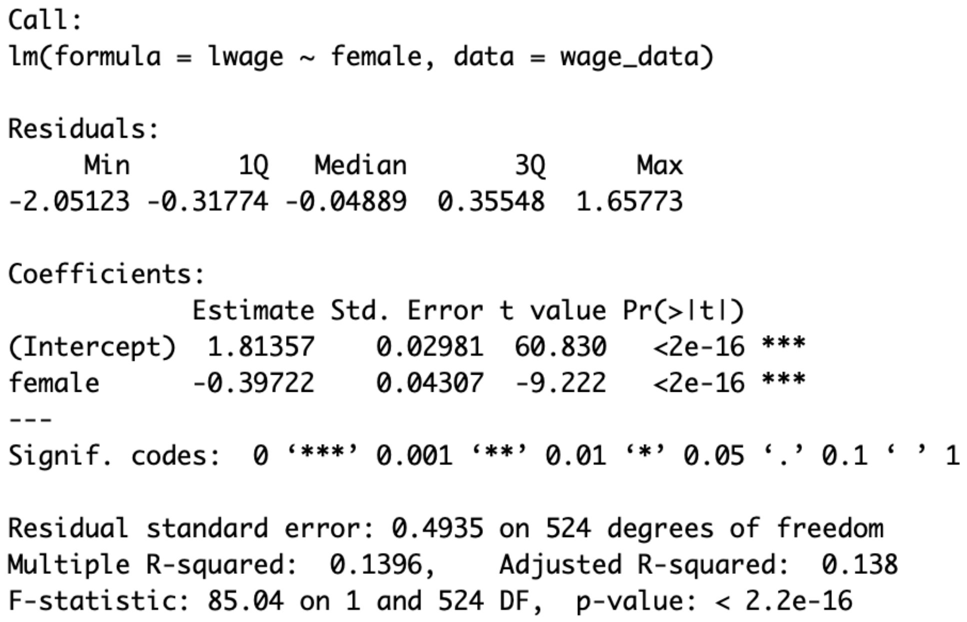
Using the table, we would expect: $E(\tilde{\beta}_1) < \beta_1$. Can we explain this in words? Well females have lower tenure, and lower tenure leads to lower wages - so if we fail to control for tenure then it will look like women have much lower wages than men because of their gender, when in reality women have lower tenure as well, and that contributes to lower wages - not just their gender.

Let's see if we were right. Below is the R output from running regressions (1) and (2).

STOP BEFORE PROCEEDING: can you interpret the coefficients in this regression?



```
Call:
lm(formula = lwage ~ female + tenure, data = wage_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.00085 -0.28200 -0.06232  0.31200  1.57325 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.688842   0.034368  49.141  < 2e-16 ***
female      -0.342132   0.042267  -8.095 4.06e-15 ***
tenure       0.019265   0.002925   6.585 1.11e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4747 on 523 degrees of freedom
Multiple R-squared:  0.2055,	Adjusted R-squared:  0.2025 
F-statistic: 67.64 on 2 and 523 DF,  p-value: < 2.2e-16
```



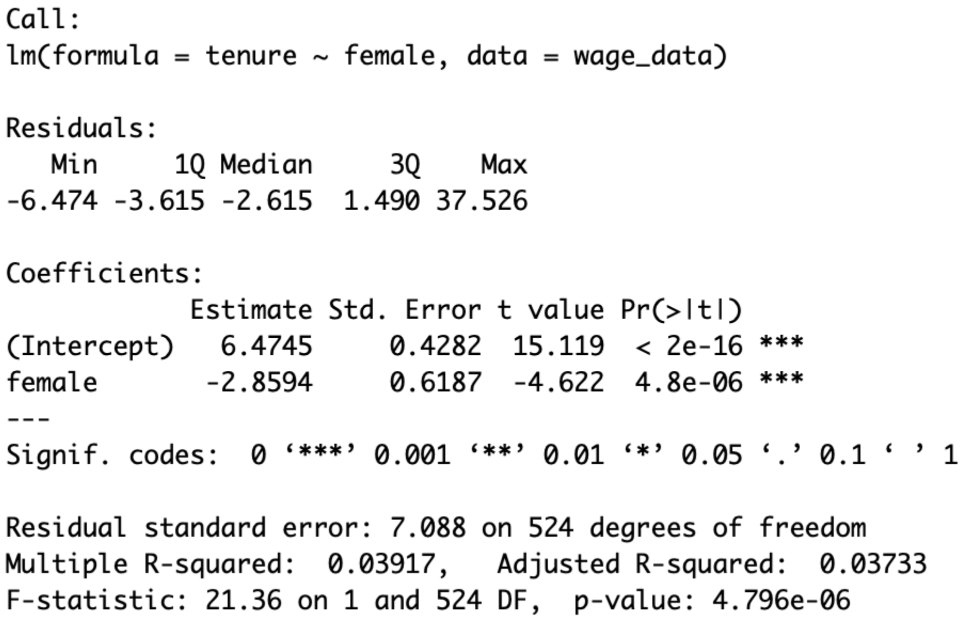
```
Call:
lm(formula = lwage ~ female, data = wage_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.05123 -0.31774 -0.04889  0.35548  1.65773 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.81357    0.02981  60.830   <2e-16 ***
female      -0.39722    0.04307  -9.222   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4935 on 524 degrees of freedom
Multiple R-squared:  0.1396,	Adjusted R-squared:  0.138 
F-statistic: 85.04 on 1 and 524 DF,  p-value: < 2.2e-16
```

We have confirmed our intuition - running the regression without tenure leads to a lower coefficient $-0.397 < -0.3421$ then when we run the true model. Our unbiased estimate is -0.3421 and our biased estimate is -0.397. As a result we have *downward* bias.

We can also use the OVB formula - by plugging in all required terms. We still need to run the following expression to calculate the OVB formula

$$tenure = \rho_0 + \rho_1 female + v$$

The R output from this regression is below:



```
Call:
lm(formula = tenure ~ female, data = wage_data)

Residuals:
   Min     1Q Median     3Q    Max 
-6.474 -3.615 -2.615  1.490 37.526 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.4745     0.4282  15.119  < 2e-16 ***
female       -2.8594     0.6187  -4.622  4.8e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.088 on 524 degrees of freedom
Multiple R-squared:  0.03917,	Adjusted R-squared:  0.03733 
F-statistic: 21.36 on 1 and 524 DF,  p-value: 4.796e-06
```

Just to clarify, our $\rho = -2.8594$

Now we can plug all of our parameters into the bias formula to check that it in fact gives us the bias from leaving out tenure from our wage regression:

$$\begin{aligned}\mathbb{E}[\tilde{\beta}_1] &= \beta_1 + \beta_2\rho_1 \\ &= -.342132 + (.019265)(-2.8593) \\ &= -0.3972164 \\ &< \beta_1\end{aligned}$$

And we have downward bias.

Example 2

Using data from Anderson (2008) for 49 US states, we can examine how primary seatbelt laws (an officer can pull you over just for not wearing your seatbelt) impact annual traffic fatalities. We have data on whether or not the state had a pimary seatbelt law in place, and the total population of the state. In 2000, just 35% of the 49 states had primary seatbelt laws (the rest had what's called a secondary seatbelt law). Suppose we run the following regression:

$$\widehat{fatalities} = \hat{\beta}_0 + \hat{\beta}_1 pop + \hat{\beta}_2 primary$$

We get

$$\widehat{fatalities} = 156.002 + 0.1232pop + 17.258primary$$

1. What do the results tells us about the relationship between seatbelt laws and fatalities? According to our estimates, predicted fatalities increase with the implementation of a primary seatbelt law. This is very surprising: omitted variables are the likely culprits here.
2. Think of another variable or factor that you think affects traffic fatalities: Two examples: alcohol laws and speeding laws might be different and affect fatalities. Let's stick with speeding laws for example.
3. Is this factor positively or negatively correlated with *fatalities*? Higher speed limits are associated with higher fatalities (+)
4. Is this factor positively or negatively correlated with *primary*? Higher speed limits are positively correlated with primary seatbelt laws (+)
5. Omitting this factor from our regression will bias $\hat{\beta}_1$: UPWARD OR DOWNWARD? States that have seatbelt laws will also have higher speed limits (the seatbelt laws may be a way to try and reduce fatalities driven by high speed limits); and higher speed limits are associated with more fatalities. So by failing to include speed limits we will misattribute the effect of higher speed limits to seatbelt laws. In other words some states with really high speeding limits will have high fatalities despite also having the seatbelt laws. In that case it will look like the seatbelt law is associated with high fatalities when in reality this is the effect of the high speeding laws. Our biased estimate is positive, while our unbiased estimate is almost surely negative. Therefore we have upward bias.

   When in doubt in terms of working through the intuition, refer to 1) the table, or 2) the formula. Taking each one in turn. Had we used the table, we would have looked at the entry that corresponds to $cov(y, x_{ov}) > 0$ and $cov(x, x_{ov}) > 0$, and we would have directly found Upward Bias. Similarly we could use the formula:

$$\mathbb{E}[\tilde{\beta}_1] = \underbrace{\beta_1}_{-} + \underbrace{\beta_2}_{+}\underbrace{\rho_1}_{+} > \beta_1$$

Which again tells us we have upward bias. (Note that with the above formula you could run the associated regressions to obtain the actual values of each term, like we did in the previous example, but for the purposes of signing the bias, knowing the signs of each term is sufficient.

Example 3

Take the following population regression equation of housing prices in a county on pollution levels in the county and a dummy for whether the county is rural or urban :

$$price = \beta_0 + \beta_1 pollution + \beta_2 rural + u$$

Imagine a scenario where your friend fails to include a dummy for rural, and estimates the model as:

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 pollution \qquad (1)$$

And gets $\hat{\beta}_1 = -2$.

You tell him his mistake and he goes back to collect data on rural/urban status. He then estimates the model as:

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 pollution + \hat{\beta}_2 rural \qquad (2)$$

And gets $\hat{\beta}_1 = -5$, $\hat{\beta}_2 = -3$

1. What is the direction of the omitted variable bias? You can compare the two results and see that $\hat{\beta}_1$ from the biased model (1) is "not as low as it should be" compared to the unbiased model (2). So you would infer there is upward bias.

2. What is the sign of $cov(rural, pollution)$? you can do this in the following way:

   - You know you have upward bias by comparing the estimates.
   - You can argue that $cov(y, x_{om}) = cov(prices, rural) < 0$ since price of houses are lower in rural areas.
   - Therefore you infer $cov(x, x_{om}) = cov(pollution, rural) < 0$
   - Conclude: when we include a variable "rural" our estimate becomes more negative. Given that rural areas have lower housing prices, in order to get a more negative coefficient when we control for rural, it must be the case that rural areas also have lower pollution levels, so that when I omit rural, I dampen the effect of pollution on prices.

3. What if I hadn't given you the values of the estimates and I had asked you to infer the likely direction of bias? You need two arguments:

   - You can argue that $cov(y, x_{om}) = cov(prices, rural) < 0$ since price of houses are lower in rural areas.

- You can argue that $cov(x, x_{om}) = cov(pollution, rural) < 0$ since pollution is lower in rural areas.
- Therefore you infer upward bias
- You could also have applied the formula $E[\tilde{\beta}_1] = \underbrace{\beta_1}_{-} + \underbrace{\beta_2}_{-} \cdot \underbrace{\rho}_{-} > \beta_1$
- Conclude: By failing to include rural, you may see low pollution counties in your data, and you won't see a strong effect on prices but that's because these counties are also rural, and rural areas have lower housing prices.

4. Confidence Intervals

We now move from talking about assumptions for unbiased estimates in regression results and revert back to basic statistics and hypothesis testing. Why? The reason is because we haven't really needed to use statistics to calculate a fitted regression line. Now, we need to use statistics to tell us how much confidence we have that our results aren't just random noise. A point estimate from a particular sample does not, by itself, provide enough information for testing economic theories or for informing policy discussions. A point estimate may be the researcher's best guess at the population value, but it provides no information about how close the estimate is "likely" to be to the population parameter. So OLS gives us the best possible fit of our model to our sample data, but it may not be capturing the true relationship.

For example, suppose you receive some data on a random sample of workers from New York. You find that workers who receive on-the-job training have higher hourly wages (5%). Can you say anything about whether or not this is close to the effect we would detect in the population of workers who could have been trained? Not as is, no. However, we can make statements involving probabilities: enter confidence interval estimation.

i. Central Limit Theorem

The central limit theorem (CLT) states that the average from a random sample for any population (with finite variance), when standardized, has an asymptotic standard normal **distribution**. Consider a random sample $X_1, \cdots, X_n$ from a population with mean μ and variance σ^2, then

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma_X / \sqrt{n}} \xrightarrow{d} N(0,1)$$

has an asymptotic standard normal distribution, regardless of the population distribution of X. The variable Z_n is the standardized version of $\bar{X}_n$: we have subtracted off the mean $E(\bar{X}_n) = \mu$, and divided by the standard deviation $Sd(\bar{X}_n) = \sigma_X / \sqrt{n}$. Thus, regardless of the population distribution of X, Z_n has mean zero and variance one, which coincides with the mean and variance of the standard normal distribution. The entire distribution of Z_n gets arbitrarily close to the standard normal distribution as n gets large.

This holds more generally for parameters other than the sample mean. For a given sample estimator $\hat{\gamma}$ with true population value γ and sample variance σ^2, the CLT tells us that $\hat{\gamma} \sim N(\mu, \sigma^2)$, or after standardizing, $(\hat{\gamma} - \mu)/\sigma \sim N(0,1)$. It is important to note that σ^2 here refers to the variance of the estimator.

Intuition: if we take the average of enough values that are sampled randomly from the same distribution with a finite mean and variance, then this aggregate starts to behave as if it is normally distributed. If we then demean and divide by the standard deviation, then this aggregate starts to behave as if standard normally distributed.

Wooldridge Appendix C.5: "We already know one way of assessing the uncertainty in an estimator: find its sampling standard deviation. Reporting the standard deviation of the estimator, along with the point estimate, provides some information on the accuracy of our estimate. However, even if the problem of the standard deviation's dependence on unknown population parameters is ignored, reporting the standard deviation along with the point estimate makes no direct statement about where the population value is likely to lie in relation to the estimate." This is why we turn to confidence intervals.

ii. Confidence Intervals: Defined

Now we can use what we know about the distribution of standard normal variables to help us say something meaningful about what the true population mean, μ, might be:

- We know that for any standard normal variable v, $Pr(-1.96 < v < 1.96) = 95\%$
- We know from the CLT $\frac{\bar{X}-\mu}{\sigma_X/\sqrt{n}}$ is distributed standard normal

But we're not *really* interested in the variable $\frac{\bar{X}-\mu}{\sigma_X/\sqrt{n}}$, rather, we'd like to learn more about μ. Let's assume for now that $(\sigma_X/\sqrt{n})$ is known;

$$Pr(-1.96 < \frac{\bar{X}-\mu}{\sigma_X/\sqrt{n}} < 1.96) = 0.95$$
$$Pr(\bar{X} - 1.96\frac{\sigma_X}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma_X}{\sqrt{n}}) = 0.95$$

This equation tells us that the probability that the random interval $\left[\bar{X} - 1.96\frac{\sigma_X}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma_X}{\sqrt{n}}\right]$ contains the population mean μ is 0.95 or 95%.

When we say that $\left[\bar{X} - 1.96\frac{\sigma_X}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma_X}{\sqrt{n}}\right]$ is a **confidence interval**, we mean that the **random interval** $\left[\bar{X} - 1.96\frac{\sigma_X}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma_X}{\sqrt{n}}\right]$ contains μ with probability 0.95. In other words, *before* the random sample is drawn, there is a 95% chance that the interval contains μ. This is a random interval, since the endpoints change with different samples (the confidence limits contain random variables $(\bar{X}, \frac{\sigma_X}{\sqrt{n}})$ that vary in value from one sample to another). BUT for any one sample of data of size n, and corresponding estimates of $(\bar{X}, \frac{\sigma_X}{\sqrt{n}})$, the confidence limits are **fixed**.

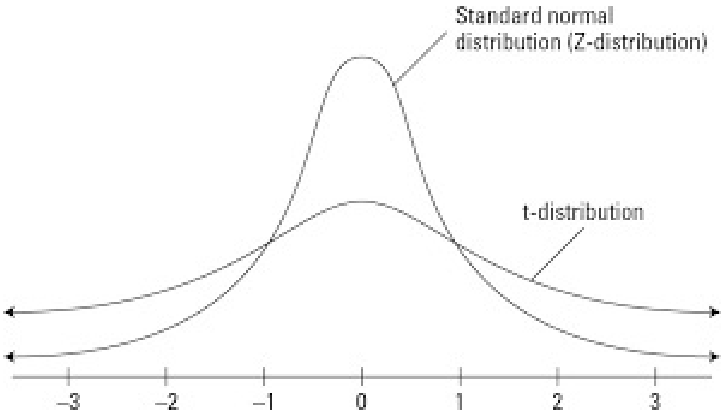
Therefore any one confidence interval calculated for a particular sample of data is a fixed, meaning **nonrandom interval**. It either does or doesn't contain the true mean.

Wooldridge explains this clearly (p.763): "A confidence interval is often interpreted as follows: "The probability that μ is in the interval is 0.95". This is incorrect. Once the sample has been observed and $\bar{X}$ has been computed, the limits of the confidence interval are simply numbers. The population parameter, μ, though unknown is also just some number. Therefore, μ either is or is not in the interval (and we will never know with certainty which is the case). Probability plays no role once the confidence interval is computed for the particular data at hand. The probabilistic interpretation comes from the fact that for 95% of all random samples, the constructed confidence interval will contain μ".

The issue is that we don't observe $\frac{\sigma_X}{\sqrt{n}}$, which means that we can't calculate the confidence interval as is. We need to replace σ_X with the sample standard deviation s_X. Then we will get the expression $\frac{s_X}{\sqrt{n}}$, which is referred to as the **standard error** of $\bar{X}$, $se(\bar{X})$. Therefore, rather than using the standard normal distribution, we rely on the **t-distribution**. The t-distribution arises from the fact that:

$$\frac{\bar{X}-\mu}{s_X/\sqrt{n}} \sim t_{n-1} \qquad (1)$$

We don't prove (1) - but check out Larsen and Marx (1986, Chapter 7) if you want to see the derivation. The pdf of the t distribution has a shape similar to that of the standard normal distribution, except it is more spread out and therefore has more area in the tails. As n grows, the t distribution approaches the standard normal. In the picture below we depict a t-distribution overlaying a normal distribution



Summary 1: when you replace the standard deviation of x (σ_X) with our estimate for the standard deviation (s_X), the 95% confidence interval you get with a t-distribution will generally be wider than for the normal (because of these fatter tails). So let's use an example: let's assume we have 30 observations, and we want a 95% confidence interval. Under a normal distribution, we know: $Pr(-1.96 < v < 1.96) = 95\%$ Under a t-distribution, we find that $Pr(-2.042 < v < 2.042) = 95\%$. When we proceed to calculate the confidence interval, it should become apparent that the **t-distribution will give you a wider confidence interval**.

Summary 2: the reason we want to use the t distribution rather than the normal is that we now have to estimate σ^2 with s^2 (so we have gone from a variable in the population to something that depends on the sample we drew). This estimate introduces some uncertainty and we can no longer preserve the 95% level of confidence because s^2 depends on a particular sample. Thus, in order to have an accurate confidence interval for our estimator we need to use a distribution that has a wider distribution (i.e fatter tails i.e larger confidence interval).

iii. Confidence Intervals: Construction

Suppose I took a random sample of 121 UCB students' heights in inches, and found that $\bar{x} = 65$ and $s_x^2 = 4$. Now, I'd like to construct a 95% confidence interval for the average height of UCB students.

Step 1: Determine the confidence level.

We want to be 95% confident that our interval covers the true population parameter, so our confidence level is 0.95. We commonly also use confidence levels of 0.90 and 0.99.

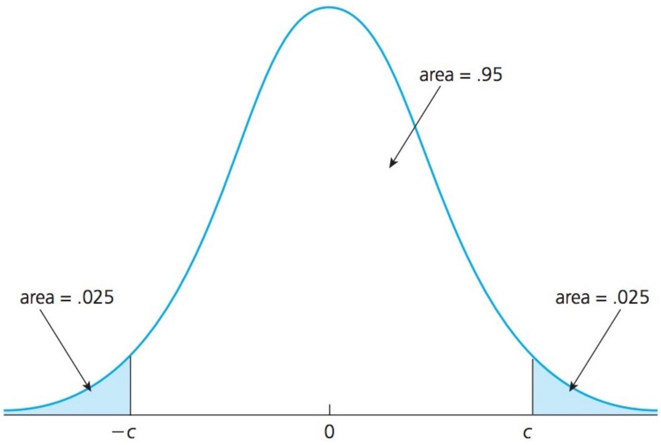
Step 2: Compute your estimates of $\bar{x}$ and s_x.

We know that from the sample we drew that $\bar{x} = 65$ and $s_x^2 = 4$ (no additional calculations are needed here - though you could imagine having to apply the formulas reviewed in previous sections).

Step 3: Find critical value, c, from the t-table.

The value of c will depend on both the sample size (n) and the confidence level (always use 2-tailed for confidence intervals). You can find these in the tables in the back of the Woolridge book or on the Internet.

- We look for the value with $n - 1$ degrees of freedom[3] . Our confidence level is 95%, with a sample size of 121: $c = 1.98$



[3] Degrees of Freedom: A common way to think of degrees of freedom is as the number of independent pieces of information available to estimate another piece of information. The denominator of the random variable $(\bar{X} - \mu)/s_X$, contains the expression $\sum_{i=1}^{n}(X_i - \bar{X})$, and we have a constraint on this expression $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. The first $n - 1$ components of this vector can be anything. However, once you know the first $n - 1$ components, the constraint tells you the value of the nth component. Therefore, this vector has $n - 1$ degrees of freedom.

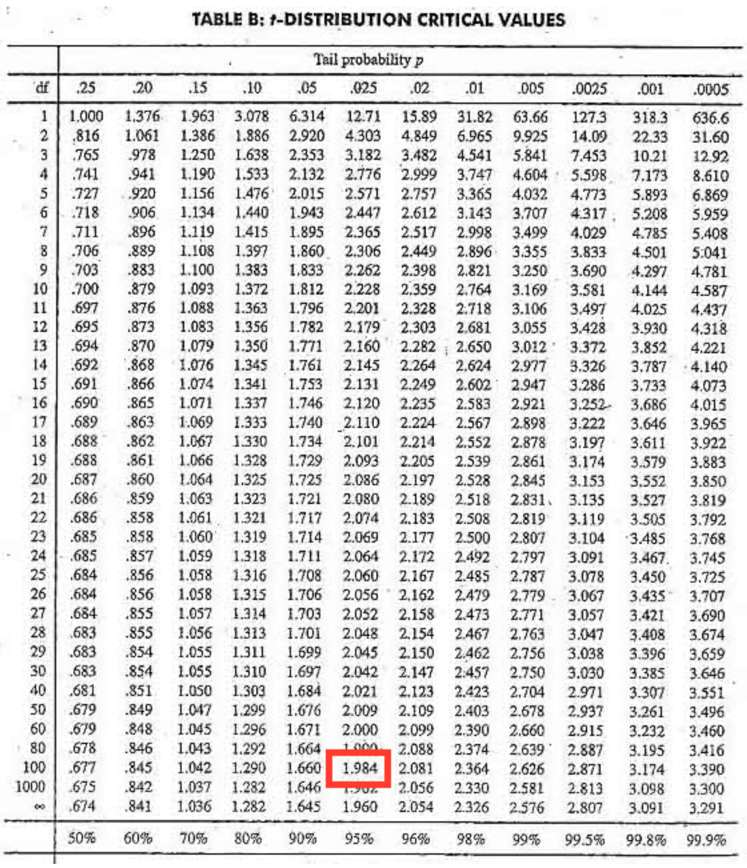


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Step 4: Plug everything into the formula

The formula for a confidence interval based on a t-distribution is:

$$CI = \left[\bar{x} - c \cdot \underbrace{\left(\frac{s_x}{\sqrt{n}} \right)}_{se(\bar{X})}, \bar{x} + c \left(\frac{s_x}{\sqrt{n}} \right) \right]$$

In this case, plugging everything in yields:

$$CI = \left[65 - 1.98 \left(\frac{2}{\sqrt{121}} \right), 65 + 1.98 \left(\frac{2}{\sqrt{121}} \right) \right]$$

the 95% confidence interval is $[64.64, 65.36]$.

Step 5: Interpret

This interval has a 95% chance of covering the true average height of the UCB student population.

Practice: Work through the following example

Use the Stata output below to construct a 90% confidence interval for Michigan State University average undergraduate GPA from a random sample of the MSU student body:

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      colGPA |        101       2.984    .3723103        2.2          4
```

1. Confidence level: 90%
2. $\bar{x}, s_x$: $\bar{x} = 2.984$ and $s_x = 0.3723$
3. Find the critical value for $n = 101$ and confidence level of 90%: $c = 1.662$
4. Compute :

$$\begin{aligned}&\big[2.984 - 1.662\frac{0.3723}{\sqrt{101}}, \ \ 2.984 + 1.662\frac{0.3723}{\sqrt{101}}\big]\\ &\big[2.922, \ \ 3.046\big]\end{aligned}$$

5. Interpret: We are 90% confident that this interval covers the true MSU average GPA.

iv. Confidence intervals and Binary Variables

These are constructed in much the same way, though we want to be careful about the formulas we apply. The setup is as follows: we have a binary variable X which takes on the value 1 with success probability p, and value 0 with failure probability (1-p). With binary data, the sample mean is the observed $\hat{p}$ share of successes, and the sample variance is given by $(\hat{p}(1-\hat{p}))/n$.

Take the following example: we are interested in knowing what proportion (p) of voters in the US that would approve the construction of the Keystone pipeline. Let's imagine that we randomly draw a sample (n=444) from the population of voters. We find that the proportion who would vote yes in our sample is 76%.

What is the 95% confidence interval for p ? We know that $\hat{p} = 76\%$ and $n = 444$. Then, we run through the same steps:

1. Confidence level: 95%
2. $\bar{x}, s_x$:

   We have $\bar{x} = \hat{p} = 0.76$. Then, recall our formulas:

$$\begin{aligned}Var(\bar{X}) &= \frac{\sigma_X^2}{n}\\ Sd(\bar{X}) &= \frac{\sigma_X}{\sqrt{n}}\end{aligned}$$

   We don't observe σ_X, so we replace with s_X, and we calculate $se(\bar{X}) = \frac{\hat{\sigma_X}}{\sqrt{n}}$. Doing all this with a binary variable yields:

$$\begin{aligned}se(\hat{p}) &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\\ &= \sqrt{\frac{0.24 \times 0.76}{444}}\\ &= \sqrt{0.00041}\\ &= 0.02\end{aligned}$$

3. Find the critical value. $c = 1.96$

4. Compute:

$$[\bar{x} - 1.96se(\hat{p}), \bar{x} + 1.96se(\hat{p})]$$
$$[0.76 - 1.96 \times 0.02, 0.76 + 1.96 \times 0.02]$$
$$[0.7208, 0.7992]$$

5. Interpretation:

   We are 95% confident that this interval covers the true proportion of voters that would approve construction of the Keystone pipeline.