Multivariate regression

Motivation

Previously, we assumed that the model for y is

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where

- β represents the causal effect of x on y, and
- ε is uncorrelated with x

We estimated this line as

$$\hat{y}_i = a + bx_i$$

when discussing the interpretation of our estimate b of β, we asked

> Can we always interpret b as an estimate of the causal effect of x on y?

The answer was

> **No.** b only represents a good estimate of the causal effect β if the assumption that x and ε are uncorrelated holds

The reason for this is that if x is correlated with ε, then when x changes, ε will change as well

In this case, b might tell us more about the effect of ε on y than the effect of x itself on y

Multivariate regression and omitted variable bias

Let's assume that the correct model actually involves *two* variables:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Note that if we *omit* x_2 from the model, the equation for y can also be written

$$y_i = \alpha + \beta_1 x_{1i} + \underbrace{u_i}_{=\beta_2 x_{2i} + \varepsilon_i}$$

Here, u is the error term in the model that omits x_2, and $\beta_2 x_2$ is part of u

Question: If the correct model is

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

but we *omit* x_2 from the model and estimate the line

$$\hat{y}_i = a + b x_{1i},$$

is b still a good estimate of β_1?

The answer comes from the **Omitted Variable Bias (OVB) formula**:

$$E(b) = \beta_1 + \beta_2 b_{x_2 x_1}$$

where $b_{x_2 x_1}$ is the slope coefficient from a regression of x_2 on x_1

Interpetation: If we omit x_2 from the model, the estimated coefficient on x_1 tells us

$$\text{Effect of } x_1 + \text{Effect of } x_2 \cdot \text{Relationship between } x_2 \text{ and } x_1$$

E.g.: Consider our wage regression, but now suppose that the *true* model is

$$wage_i = \alpha + \beta_1 school_i + \beta_2 ability_i + \varepsilon_i$$

If we omit ability from the regression, the coefficient on schooling will tell us

$$E(b) = \underbrace{\text{Effect of schooling}}_{\beta_1} + \underbrace{\text{Effect of ability}}_{\beta_2} * \underbrace{\text{Relationship between ability and school}}_{b_{ability,school}}$$

Intuitively, a regression of wages on schooling tells us how wages differ between people with different levels of schooling

But if schooling is correlated with ability, people with different amounts of schooling also have different levels of ability

So our regression is telling us a combination of how schooling itself affects wages and how ability affects wages

Let's quickly prove the omitted variable bias formula

$$
\begin{aligned}
E(b) &= E\left(\frac{\sum_i x_{1i} y_i}{\sum_i x_{1i}^2}\right) = \frac{1}{\sum_i x_{1i}^2} \sum_i x_i E(y_i) \\
&= \frac{1}{\sum_i x_{1i}^2} \sum_i x_{1i} E(\beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i) \\
&= \beta_1 \frac{\sum_i x_{1i}^2}{\sum_i x_{1i}^2} + \beta_2 \underbrace{\frac{\sum_i x_{1i} x_{2i}}{\sum_i x_{1i}^2}}_{=b_{x_2 x_1}} + \underbrace{\frac{\sum_i x_{1i} E(\varepsilon_i)}{\sum_i x_{1i}^2}}_{=0} \\
&= \beta_1 + \beta_2 b_{x_2 x_1}
\end{aligned}
$$

Multivariate regression

What can we do if we're worried that ε might be correlated with x_1 ?

The problem is that a regression of y on x_1 only tells us the *relationship* between x_1 and y: how y changes when x_1 changes

I.e., the regression compares y between people with different values of x_1

But when x_1 changes, x_2 might be changing as well

So our regression tells us a combination of how x_1 affects y and how x_2 affects y

What we need to do is look at how y changes when x_1 changes, *holding* x_2 constant (i.e, compare y between people with different values of x but the same values)

This is what a **multivariate** regression does[1]

In a multivariate regression,

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i}$$

We can actually include as many variables as we want:

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

1. Aka **multiple regression**

Estimation

How do we estimate a multivariate regression?

We still minimize the sum of squared residuals

That is, a and $b_1, b_2, \ldots, b_k$ are chosen to minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2$$

This is really the *only* difference

We still find the regression coefficients with calculus (now we take the partial derivatives with respect to a and $b_1, b_2, \ldots, b_k$ and set them equal to zero)

Although the algebra gets more complicated (and we won't go through it), computers have no problem handling the additional work

Interpreting multivariate regressions

Once we've estimated our multivariate regression

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

We can intepret the b_j as how we predict y will change if we increase x_k, *holding all of the other x's constant*:

$$b_j = \frac{\Delta \hat{y}}{\Delta x_j}\Big|_{\text{all other } x\text{'s constant}}$$

The fact that we are *controlling* for other x's by holding them constant is what solves the OVB problem

In our multivariate model,

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon$$

where ε is uncorrelated with $x_1, x_2, \ldots, x_k$, so that (conditonal on the x's) $E(\varepsilon) = 0$

The β's represent the **causal effects** of the x's on y

If x_j is uncorrelated with ε, b_j will be an unbiased and consistent estimate of β_j

The idea is that we want to take omitted variables *out of the error term* by controlling for them in the model

If we have done a good job of this, the part of the error term that remains will be uncorrelated with the x's, and so the b's will be good estimates of the β's (which represent the causal effects of the x's)

In terms of our previous example, we are now comparing wages for people with different amounts of education but the *same* amount of ability

Let’s do some simulations to illustrate OVB and multivariate regression

```r
b <- rep(0,1000)
b1 <- rep(0,1000)
for (i in 1:1000) {
  x1 <- rnorm(100)
  x2 <- 1 + 2*x1 + rnorm(100) #x1 is corr'd with x2
  y <- 2 + 2*x1 + 3*x2 + rnorm(100)
  b[i] <- lm(y~x1)$coef[2]
  b1[i] <- lm(y~x1 + x2)$coef[2]
}
mean(b) # By the OVB formula, E(b) = 2 + 3*2 = 8
```

```
[1] 7.992041
```

```r
mean(b1)
```

```
[1] 1.993346
```

Now let's look at a real world example. We'll start by regressing wages on schooling alone:

```r
library(tidyverse)
gril <- read_csv("griliches.csv")
gril$wage <- exp(gril$lw)
model1 <- lm(wage ~ s, data=gril)
summary(model1)
```

```
Call:
lm(formula = wage ~ s, data = gril)

Residuals:
    Min      1Q  Median      3Q     Max
-344.94  -84.91  -20.59   60.90  676.12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -125.295     28.458  -4.403 1.22e-05 ***
s             33.511      2.094  16.002  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 128.6 on 756 degrees of freedom
```

This model omits IQ from the equation. Next, let's investigate the relationship between IQ and schooling:

```
model2 <- lm(iq ~ s, data=gril)
summary(model2)
```

```
Call:
lm(formula = iq ~ s, data = gril)

Residuals:
    Min      1Q  Median      3Q     Max
-48.588  -7.326   0.346   7.543  34.543

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.8844     2.5886   23.91   <2e-16 ***
s             3.1311     0.1905   16.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.7 on 756 degrees of freedom
```

Finally, let's run a multivariate regression of wages on schooling *and* IQ:

```r
model3 <- lm(wage ~ s + iq, data=gril)
summary(model3)
```

```
Call:
lm(formula = wage ~ s + iq, data = gril)

Residuals:
    Min      1Q  Median      3Q     Max 
-357.58  -80.42  -23.46   58.60  691.68 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -198.0135    37.5193  -5.278 1.71e-07 ***
s             29.8315     2.4274  12.289  < 2e-16 ***
iq             1.1751     0.3978   2.954  0.00324 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In summary, a univariate regression of wages on schooling gives:

$$\widehat{wage}_i = -125.3 + 33.51 * s$$

But if IQ is related to both schooling and wages, this coefficient will only tell us a combination of the effect of schooling itself on wages and the effect of IQ on wages

If we run a multivariate regression that also controls for IQ, we get

$$\widehat{wage}_i = -198.01 + 29.83 * s + 1.18 * iq$$

The coefficient on schooling is now smaller, which suggests that the previous coefficient was biased up because we failed to control for IQ

If we believed that IQ was the only omitted variable, we could interpret this as the effect of schooling on wages

We also know that the relationship between IQ and schooling is given by

$$\widehat{iq}_i = 61.88 + 3.13 * s$$

The OVB formula tells us that the coefficient from a regression of wages on schooling alone should equal

$$E(b) = \beta_1 + \beta_2 * b_{iq,s} = 29.83 + 1.18 * 3.13 = 33.52,$$

which is very close to what we actually got (the difference is due to rounding error)

Correlation and causation revisited

Previously, we asked:

> Can we always interpret the regression coefficient as an estimate of the causal effect?

The answer was:

> *Not necessarily,* because there might be omitted variables

Now we know that if we're worried about OVB in a mode like

$$y_i = \alpha + \beta_1 x_{1i} + u_i$$

We can control for omitted variables (i.e., take them out of the error term and put them in the model) using a multivariate regression model:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

This is *very* helpful, but it does raise another question:

> How do we know if we have controlled for all of the relevant omitted variables?

Unfortunately, the answer is

> *We don't.* The best we can do is think about potential omitted variables and try to control for them

There is no way to guarantee that you are estimating a causal effect (the best you can do is try!)

In the wage example, when we omitted IQ, the coefficient on schooling was 33.51

When we controlled for IQ, the coefficient on schooling was 29.83

If we thought that IQ was the only omitted variable, this would tell us the causal effect

But we might be able to come up with other omitted variables that might be correlated with schooling

Regression anatomy

Recall that in the multivariate regression

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

we interpret β_1 as the effect of x_1 on y, *holding* x_2 *through* x_k *constant*

We also saw that we do this by minimizing the sum of squared residuals with respect to all of the coefficients (a and b_1 –b_k)

It turns out that there is another way of thinking about this, which also leads to another way of estimating multiple regression coefficients

According to the **regression anatomy formula**, we could also obtain the etiamte b_1 of β_1 using the following process:

1. Regress y on x_2 through x_k, and save the residuals $e_{y|x_2,\ldots,x_k}$
2. Regress x_1 on x_2 through x_k and save the residuals $e_{x_1|x_2,\ldots,x_k}$
3. Regress the "y" residuals ($e_{y|x_2,\ldots,x_k}$) on the "x" residuals ($e_{x_1|x_2,\ldots,x_k}$)

It turns out (although we won't derive it), that the coefficient on the x residuals from this regression is *exactly the same* as the coefficient b_1 from a multivariate regression of y on x_1 through x_k

Why is this useful? There are two reasons:

1. It reduces our multivariate regression (which is fairly complicated) to a univariate regression (which we understand pretty well)
2. It has an intuitive interpretation: In a multiple regression, β_1 is the effect of x_1 on y, *holding* $x_2 - x_k$ *constant.*

   When we regress the y residual on the x residual, we are looking at the relationship between the "parts" of these variables that are not explained by $x_2 - x_k$. The anatomy formula gives us another perspective on what it means to hold those variables constant

Two notes on this:

1. The same logic applies to estimating any of the β_j's, we only used β_1 as an example (e.g., to estimate β_j, you need the residuals from regressions of y and x_j on all of the x variables *except* x_j)
2. Although the coefficient estimate from the regression anatomy formula is the same as from a multivariate regression, the standard error from the regression anatomy version will not be right (this is because software like R doesn't know that the residuals are estimated quantities, rather than "known" data). Hence, you should always use multivariate regression to obtain regression estimates (the anatomy formula more useful as a *conceptual* tool)

Let's see the regression anatomy formula in action. Recall that, in the wage example, the multivariate regression coefficients are:

```r
model1 <- lm(wage ~ s + iq, data=gril)
summary(model1)
```

```
Call:
lm(formula = wage ~ s + iq, data = gril)

Residuals:
    Min      1Q  Median      3Q     Max 
-357.58  -80.42  -23.46   58.60  691.68 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -198.0135    37.5193  -5.278 1.71e-07 ***
s             29.8315     2.4274  12.289  < 2e-16 ***
iq             1.1751     0.3978   2.954  0.00324 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What about via regression anatomy?

```
yresid <- lm(wage ~ iq, data=gril)$resid
xresid <- lm(s ~ iq, data=gril)$resid
reg.anatomy <- lm(yresid ~ xresid)
summary(reg.anatomy)
```

```
Call:
lm(formula = yresid ~ xresid)

Residuals:
    Min      1Q  Median      3Q     Max
-357.58  -80.42  -23.46   58.60  691.68

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.625e-15  4.644e+00     0.0        1
xresid       2.983e+01  2.426e+00    12.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 127.9 on 756 degrees of freedom
```

The coefficient on `xresid` is exactly the same as before!

Properties of multiple regression

The regression anatomy formula shows that multivariate regression is really just univariate regression in disguise

Hence, it inherits all of the properties of univariate regression that we previously proved:

- It is unbiased: $E(b_j) = \beta_j$ (for $j = 1, \ldots, k$)
- It is consistent: $b_j \to \beta_j$

Similarly, the variance of a multivariate regression coefficient is

$$Var(b_1) = \frac{\sigma^2}{\sum_i e^2_{x_1|x_2,\ldots,x_k}}$$

where σ^2 is the variance of ε. Although we don't know this, we can estimate it using

$$s^2 = \frac{\sum_i e_i^2}{n-k-1}$$

Note that we divide by $n-k-1$ instead of $n-2$ because now we are using a multivariate regression

Multicollinearity

When there are only two x variables, there is another way of expressing the variance of the regression coefficients that is somewhat easier to interpret

It can be shown that

$$Var(b_1) = \frac{\sigma^2}{[1 - Corr(x_1, x_2)]^2 \sum_i (x_{1i} - \bar{x}_1)^2}$$

This allows us to answer the following question:

> How does the variance of b_1 change when we add a new variable to our regression?

If we start off with a regression of y on x_1, we know that

$$V(b_1) = \frac{\sigma^2}{\sum_i (x_{1i} - \bar{x}_1)^2}$$

If we add x_2 to the regression, we have

$$V(b_1) = \frac{\sigma^2}{[1 - Corr(x_1, x_2)]^2 \sum_i (x_{1i} - \bar{x}_1)^2}$$

When we add x_2 to the regression, two things change:

1. We are taking x_2 out of the error term and putting it into the model. Therefore, the σ^2 *decreases*, which tends to make $V(b_1)$ *smaller*
2. Now we have the term $[1 - Corr(x_1, x_2)]^2$ in the denominator. If x_1 and x_2 are highly corerlated, this term will be small, which tends to make $V(b_1)$ *larger*

Multicollinearity is when the second effect dominates, so that adding a new variable increases the variance of the estimated coefficient

Intuitively, a multivariate regression compares y for units with different values of x_1 but the same values of x_2. However, if x_1 and x_2 are highly correlated, there may not be many observations with different x_1 but the same x_2, so our effective sample size is small, leading to a large variance

As the intuition suggests, the solution to multicollinearity is to increase your sample size[1]

For this reason, multicollinearity has jokingly been called "micronumerosity," reflecting the fact that it isn't really a problem with your regression, just a challenge that can be overcome with a larger sample

1. Formally, because this will increase the term $\sum_i (x_{1i} - \bar{x}_1)^2$, offsetting the correlation term

There is a form of multicollinearity that is a little more problematic

Perfect multicolliearity occurs when two variables are *perfectly* correlated

Assuming a two-variable regression for simplicity,

$$V(b_1) = \frac{\sigma^2}{[1 - Corr(x_1, x_2)]^2 \sum_i (x_{1i} - \bar{x}_1)^2},$$

$V(b_1)$ is not even defined in this case

In fact, the problem is worse: we can't even estimate the regression

By regression anatomy, in the two-variable case:

$$b_1 = \frac{\sum_i e_{y|x_2} e_{x_1|x_2}}{\sum_i e^2_{x_1|x_2}}$$

But if x_1 and x_2 are perfectly correlated, x_2 perfectly predicts x_1, so $e_{x_1|x_2} = 0$ for all observations, and b_1 is not defined

Perfect multicollinearity can be a real problem, because in its presence, the regression coefficients can't be computed

However, in most cases, running into perfect multicollinearity means that we're doing something silly

For example, suppose we want to know the relationship between shoe size and height. We might estimate the equation:

$$\text{height} = \alpha + \beta_1 \text{Left shoe size} + \beta_2 \text{Right shoe size} + \varepsilon$$

In most datasets, left and right shoe size will be the same, so we'll have perfect multicollineariy

But the relationship between left shoe size and height is probably the same as the relationship between right shoe size and height, so we don't really need to know both (we can just drop one from the regression)

Goodness of fit with multivariate regression

Recall that

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

This formula still applies for a multivariate regression

Also recall, though, that since OLS minimizes the SSR, including a new variable can never *increase* the SSR (if adding a new variable were to increase the SSR, OLS could always set the coefficient on that variable to zero, so that the SSR wouldn't change)

But this means that adding a new variable can never *decrease* the R^2

This raises the concern that analysts might try to add a bunch of useless variables to their regression in order to get a large R^2

(Although there's no real point in dong this since no one really cares about the R^2)

To protect against this possibility, there is a variation on the R^2 that penalizes regressions for adding lots of variables

The **adjusted** R^2 is

$$\text{adj. } R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2} \left(\frac{n-1}{n-k-1} \right)$$

When k is large, $(n-1)/(n-k-1)$ will be large, so the adjusted R^2 will be smaller

Thus, the adjusted R^2 only increases if adding a new variable helps explain y "a lot"

Inference with multiple regression

Now that we know how to estimate and interpret multiple regressions, we need to learn how to do statistical inference using them

Fortunately, most of what we learned for the univariate case carries over, but there are a few new twists

Suppose that we want to test the null hypothesis that $\beta_j = 0$. The test statistic for this is still $t = b_j / se(b_j)$

However, in the univariate case, we said that this test statistic had a t distribution with $n - 2$ degrees of freedom

In the multivariate case, since we are estimating more parameters, this test statistic will have a t distribution with $n - k - 1$ degrees of freedom (note that if $k = 1$, this brings us back to the univariate case)

Essentially, all this means is that we have to look up a different row in the t table when finding the appropriate critical value

Another difference is that, when we are running a multivariate regression, we might want to test **joint** hypotheses about **combinations** of coefficients

For example, suppose we estimate the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Since this allows the coefficients on x_1 –x_3 to be any value, we call this the **unrestricted** model

Now suppose that we want to test the null hypothesis that $\beta_2 = 0$ and $\beta_3 = 0$

How can we do this?

Aside: What is the difference between testing the joint null that $\beta_2 = 0$ and $\beta_3 = 0$ and doing individual tests of the nulls that $\beta_2 = 0$ and $\beta_3 = 0$?

If x_2 and x_3 are highly correlated, the SEs for b_2 and b_3 might be very large because of multicollinearity. But then the t-stats $b_2/se(b_2)$ and $b_3/se(b_3)$ would be small, and we might not be able to reject either null hypothesis, even if both x_1 and x_2 affect y

We can circumvent this problem by testing the joint null hypothesis that "$\beta_2 = 0$ and $\beta_3 = 0$"

Under the null hypothesis that $\beta_2 = \beta_3 = 0$, the model becomes

$$y = \alpha + \beta_1 x_1 + \varepsilon$$

Since this restricts the coefficients on x_2 and x_3 (to be zero in this case), we call it the **restricted** model

Here is the intuition behind our joint null hypothesis test: If the restriction that $\beta_2 = \beta_3 = 0$ is true, then when we drop x_2 and x_3 from the model, it shouldn't really matter, so the SSR shouldn't change. So if the SSR changes a lot when we drop them, we will reject the null

Formally, let SSR_R be the SSR from the restricted model and SSR_U be the SSR from the unrestricted model

The **F-statistic** for our test is

$$F = \frac{(SSR_R - SSR_U)/j}{SSR_U/(n-k-1)},$$

where j is the *number of restrictions* (two in our example)

Under the null hypothesis, this statistic has an $F(j, n-k-1)$ distribution[1]

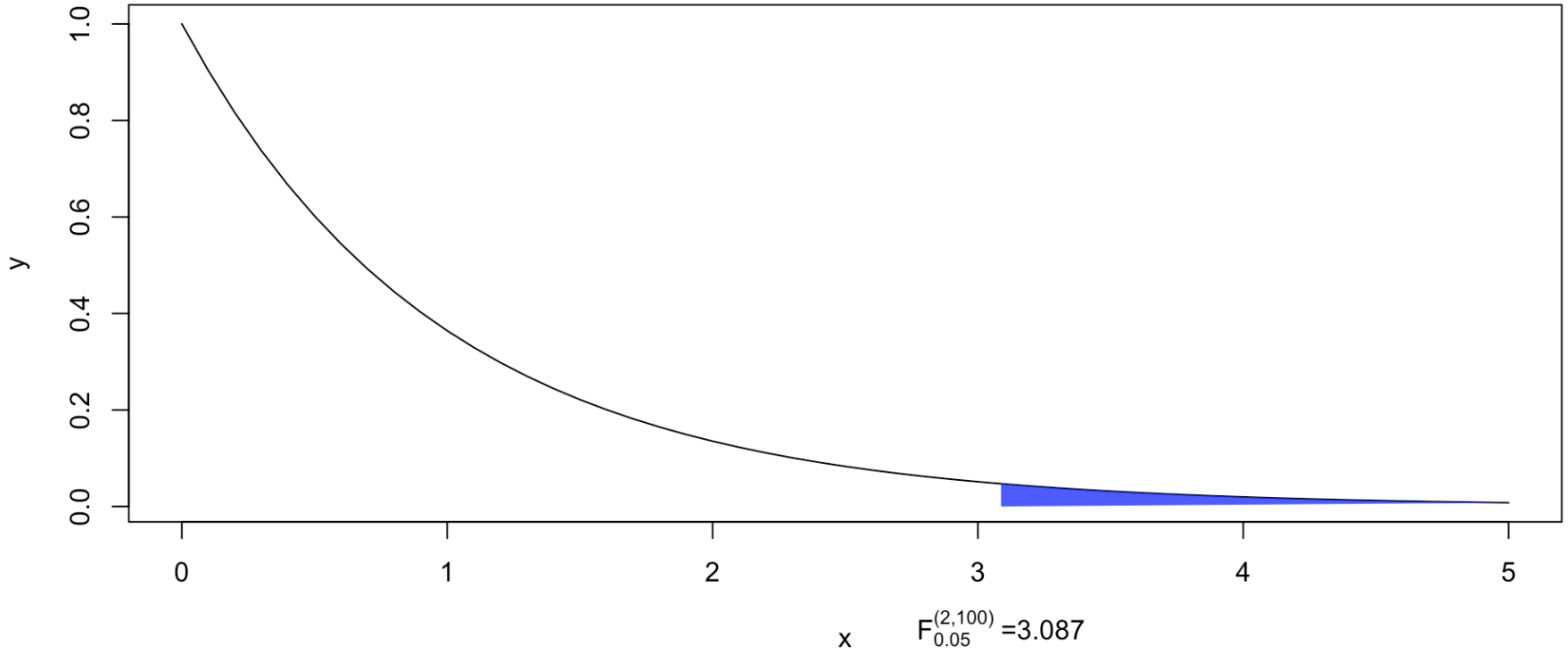
1. An F-distribution with j *numerator* degrees of freedom and $n-k-1$ *denominator* degrees of freedom. We divide by SSR_U to ensure that the F-stat has a known distribution.

To implement the test, we

1. Run the restricted and unrestricted models, and save the SSR from each
2. Compute the F-statistic
3. We reject the null if the F-statistic exceeds the appropriate critical value $F_{\alpha}^{(j,n-k-1)}$ for our α level and degrees of freedom (we just look this up in an F-table, like this one)[1]

1. Recall that adding variables never increases the SSR, so $SSR_U < SSR_R$, so $F > 0$

Here's what this looks like for $\alpha = .05$ and an $F(2, 100)$ distribution:



Let’s use our wage example to test the null that neither experience nor tenure affect wages:

```
model.u <- lm(wage ~ s + expr + tenure, data=gril)
model.r <- lm(wage ~ s, data=gril)
ssr.u <- sum(model.u$residuals^2)
ssr.r <- sum(model.r$residuals^2)
f.stat <- ((ssr.r - ssr.u)/2)/(ssr.u/754)
f.stat
```

```
[1] 34.15699
```

From this F table, the critical value for a test with α-level .05, 2 numerator degrees of freedom, and 754 denominator degrees of freedom[1] is about 3.087

Since 34.16 > 3.087, we reject the null that $\beta_2 = 0$ and $\beta_3 = 0$

1. Actually, the highest it goes is 100, which is what I'm using

It’s nice to know how to do this “by hand,” but we usually let software help us implement the test. In this case, we can use

```r
library(car)
linearHypothesis(model.u, c("expr = 0", "tenure = 0"))
```

```
Linear hypothesis test:
expr = 0
tenure = 0

Model 1: restricted model
Model 2: wage ~ s + expr + tenure

  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    756 12501021
2    754 11462495  2   1038526 34.157 6.307e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we don’t really need to estimate the restricted model (R will handle that automagically)

You might have noticed that R automatically reports something called “F-statistic” when you run a regression. This is the F-statistic for the joint null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (i.e., all of the variables are jointly statistically insignificant)

```
1 summary(model.u)
```

```
Call:
lm(formula = wage ~ s + expr + tenure, data = gril)

Residuals:
    Min      1Q  Median      3Q     Max 
-315.21  -76.17  -18.11   56.94  639.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -214.793     29.510  -7.279 8.49e-13 ***
s             36.779      2.069  17.773  < 2e-16 ***
expr          12.135      2.252   5.390 9.44e-08 ***
tenure        13.451      2.752   4.888 1.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 0.1 ' ' 1
```

We can also use this to test other kinds of joint hypotheses

Suppose that our null is that the coefficients on experience and tenure are the same. Now the restricted model is

$$y = \alpha + \beta_1 educ + \beta_2 expr + \beta_2 tenure + \varepsilon$$

We could do this manually by defining a new variable $z = expr + tenure$, then regressing y on $educ$ and z to get SSR_R

```r
gril$z <- gril$expr + gril$tenure
model.r2 <- lm(wage ~ s + z, data=gril)
ssr.r2 <- sum(model.r2$resid^2)
F.2 <-((ssr.r2 - ssr.u)/1) / (ssr.u /754)
F.2
```

```
[1] 0.1121167
```

```r
qf(.05, 1, 754, lower.tail=FALSE)
```

```
[1] 3.853821
```

Now there is only one restriction, so from this table the critical value is 3.936. Since .11<3.936, we fail to reject

Actually, the last line shows that we can also get R to tell us the exact critical value (3.85 in this case). We still fail to reject

In practice, though, we can just use the canned function:

```r
linearHypothesis(model.u, c("expr = tenure"))
```

```
Linear hypothesis test:
expr - tenure = 0

Model 1: restricted model
Model 2: wage ~ s + expr + tenure

  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    755 11464199
2    754 11462495  1    1704.4 0.1121 0.7378
```

Here, we can see from the p-value that we fail to reject the null

MOAR examples!

Now that we understand multivariate regression, let's look at some more empirical examples

First, we'll examine the effect of capital punishment (i.e., the death penalty) on murder rates

Our file is in Stata format, so we'll use the haven package to import it into R:

```
library(haven)
mur <- read_dta("MURDER.dta")
```

We can use the `summary` command to get some descriptive statistics for the data in the file (it's always a good idea to get to know your data before you dive into any analysis)

The key variables are the murder rate per 100K (`mrdrte`) and the number of executions in the past three years (`exec`)

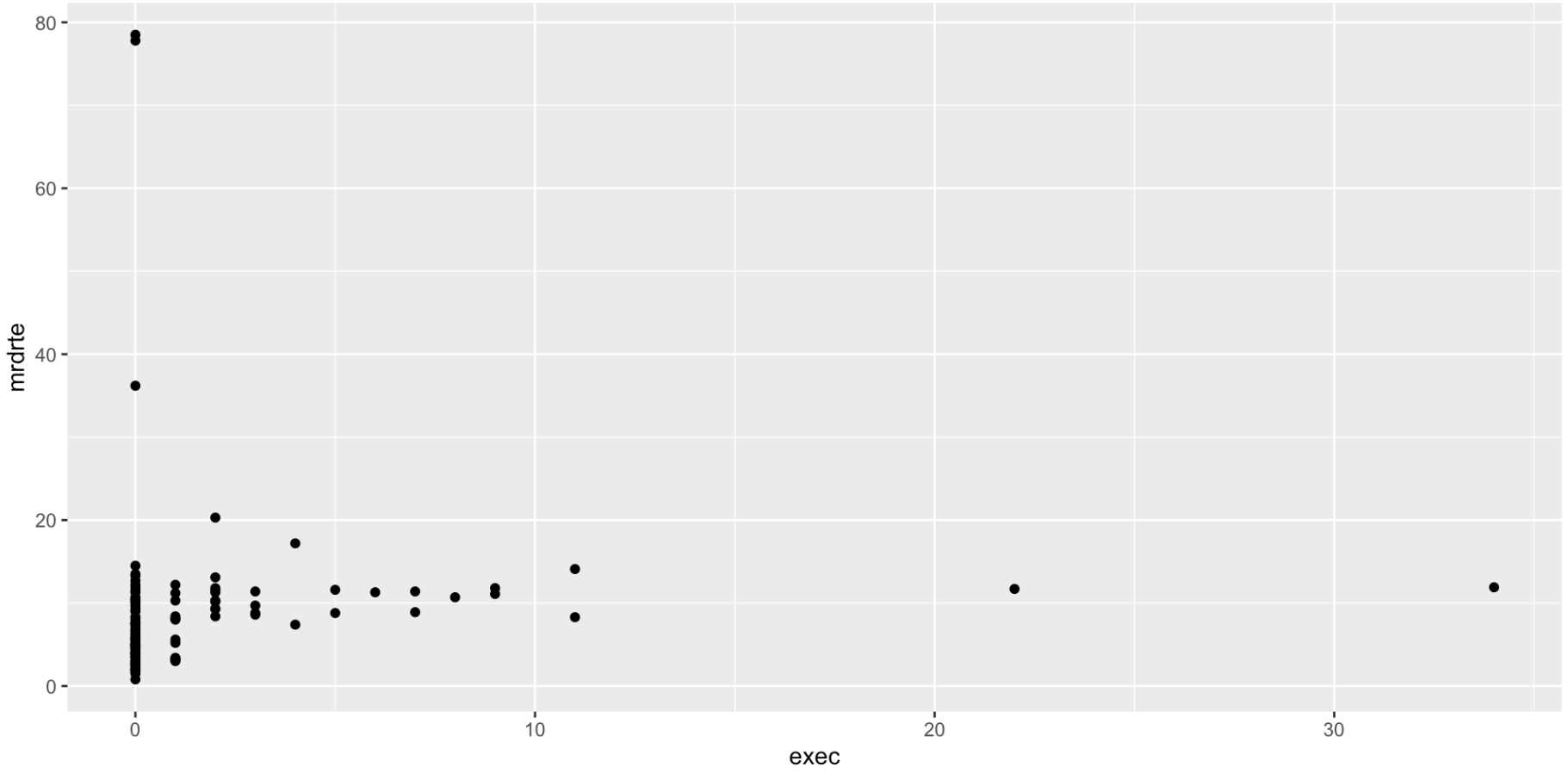
```
1 summary(mur)
```

```
       id         state                year         mrdrte            exec       
 Min.   : 1   Length:153         Min.   :87   Min.   : 0.800   Min.   : 0.000  
 1st Qu.:13   Class :character   1st Qu.:87   1st Qu.: 3.900   1st Qu.: 0.000  
 Median :26   Mode  :character   Median :90   Median : 6.400   Median : 0.000  
 Mean   :26                      Mean   :90   Mean   : 8.071   Mean   : 1.229  
 3rd Qu.:39                      3rd Qu.:93   3rd Qu.:10.200   3rd Qu.: 1.000  
 Max.   :51                      Max.   :93   Max.   :78.500   Max.   :34.000  

      unem             d90              d93             cmrdrte        
 Min.   : 2.200   Min.   :0.0000   Min.   :0.0000   Min.   :-2.6000  
 1st Qu.: 4.900   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:-0.4000  
 Median : 5.800   Median :0.0000   Median :0.0000   Median : 0.3000  
 Mean   : 5.973   Mean   :0.3333   Mean   :0.3333   Mean   : 0.8422  
 3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 1.3000  
 Max.   :12.000   Max.   :1.0000   Max.   :1.0000   Max.   :41.6000  
```

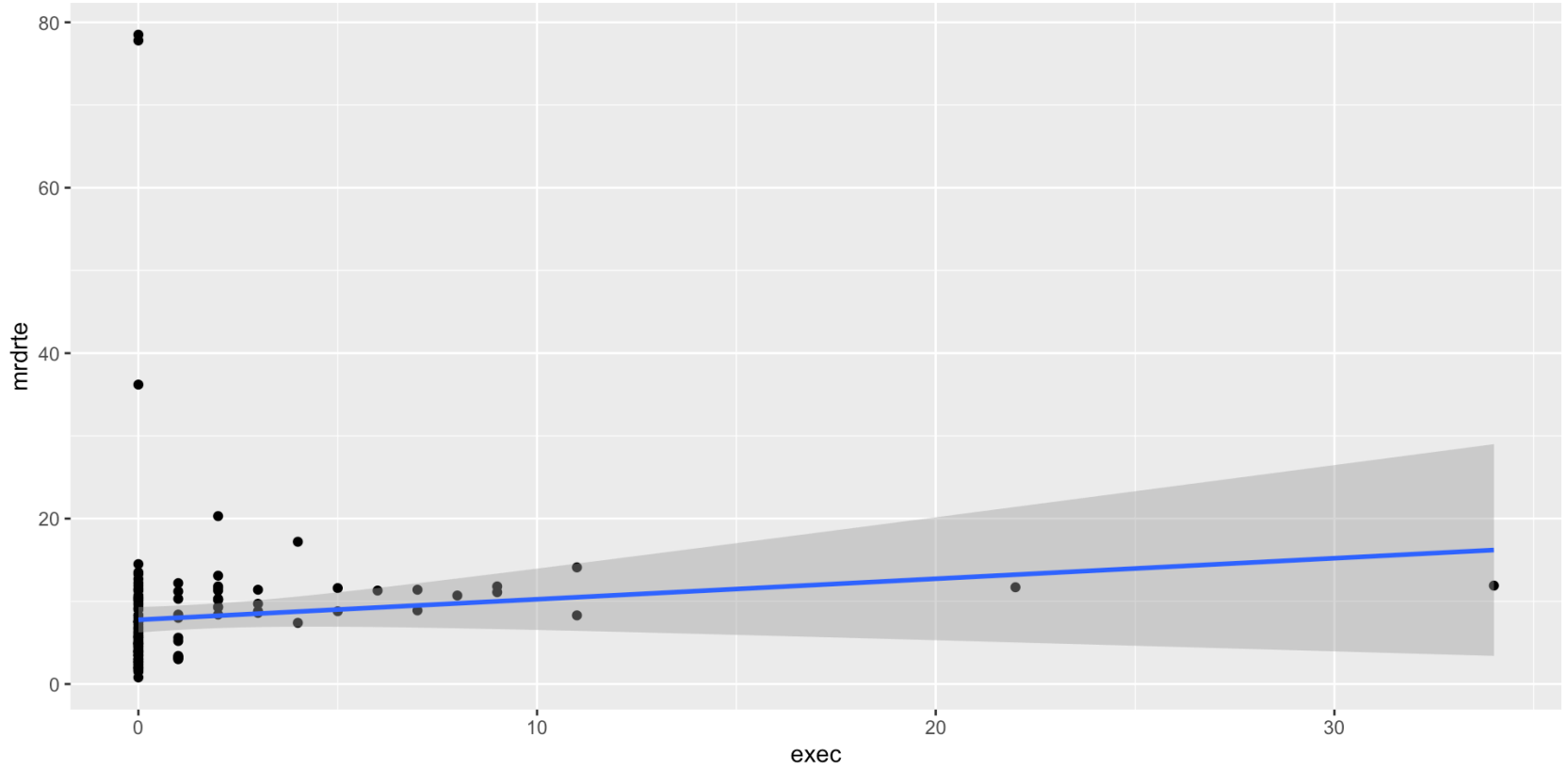
Let's look at a plot of murders against executions:

```
1 ggplot(data=mur, aes(x=exec, y=mrdrte)) + geom_point()
```



The relationship appears positive, but we can clarify by adding a “line of best fit” (which is really just a regression line)

```
1 ggplot(data=mur, aes(x=exec, y=mrdrte)) + geom_point() + geom_smooth(method
```



Now let's examine the relationship using a regression:

```r
mur.mod1 <- lm(mrdrte ~ exec, data=mur)
summary(mur.mod1)
```

```
Call:
lm(formula = mrdrte ~ exec, data = mur)

Residuals:
   Min     1Q Median     3Q    Max
-6.966 -3.866 -1.566  1.898 70.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7658     0.7800   9.957   <2e-16 ***
exec          0.2481     0.1963   1.264    0.208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.175 on 151 degrees of freedom
```

This suggests that the murder rate is positively (and statistically significantly) associated with the number of executions

Taken literally, the coefficient of .248 would mean that for every additional execution, there are an additional .248 murders per 100K people

To put this into context, in 2023, the population of Memphis was 618,639, or about 6.19 thousand. Thus, each additional execution would *increase* the number of murders by $6.19 * .248 \approx 1.5$

The R^2 of .01 means that executions only explain about 1% of the variation in the murder rate

Does it make sense that executions increase murders? Some people think capital punishment deters crime, but worst-case scenario, it probably has no effect

What's probably happening is that areas that already have high crime also tend to execute more prisoners

I.e., if the model is

$$mrdrte = \alpha + \beta exec + \varepsilon,$$

the other factors ε that determine the murder rate are *correlated* with the number of executions

Let's see what happens if we control for some of these factors. We'll start with the unemployment rate

```
1 mur.mod2 <- lm(mrdrte ~ exec + unem, data=mur)
2 summary(mur.mod2)
```

```
Call:
lm(formula = mrdrte ~ exec + unem, data = mur)

Residuals:
   Min     1Q Median     3Q    Max
-9.175 -3.472 -1.416  1.114 69.143

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3481     2.6872   0.130  0.89710
exec          0.1650     0.1939   0.851  0.39601
unem          1.2589     0.4374   2.878  0.00458 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now the murder rate is positive, but statistically insignificant. This suggests that areas with worse economic conditions have more murders but also tend to execute more people

It's also possible that both the murder rate and the number of executions just happen to change over time. We can control for this by including variables indicating whether an observation corresponds to a particular year

```
1 mur.mod3 <- lm(mrdrte ~ exec + unem + as.factor(year), data=mur)
2 summary(mur.mod3)
```

```
Call:
lm(formula = mrdrte ~ exec + unem + as.factor(year), data = mur)

Residuals:
   Min     1Q Median     3Q    Max
-9.130 -3.119 -1.211  1.379 67.810

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.8644     3.0695  -0.607  0.54452
exec                0.1628     0.1939   0.839  0.40268
unem                1.3908     0.4509   3.085  0.00243 **
as.factor(year)90   2.6753     1.8169   1.472  0.14302
as.factor(year)93   1.6073     1.7748   0.906  0.36659
---
```

The coefficient on `exec` is still statistically insignificant

Finally, we might think that certain states just always tend to have more murders and execute more prisoners. We can control for this by including variables that indicate whether a particular observation corresponds to a particular state

```
mur.mod4 <- lm(mrdrte ~ exec + unem + as.factor(year) + as.factor(state), d
summary(mur.mod4)
```

```
Call:
lm(formula = mrdrte ~ exec + unem + as.factor(year) + as.factor(state),
    data = mur)

Residuals:
     Min       1Q   Median       3Q      Max
-26.6858  -0.6584  -0.0657   0.6747  13.3941

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.9038     3.3143   1.781  0.07796 .
exec                -0.1383     0.1770  -0.781  0.43642
unem                 0.2213     0.2964   0.747  0.45701
as.factor(year)90    1.5562     0.7453   2.088  0.03939 *
as.factor(year)93    1.7332     0.7004   2.475  0.01506 *
```

Ok, that’s a lot of variables, but after including all of these control variables, the coefficient on executions is no longer positive

This makes more sense, since it no longer suggests that executions increase murders

Unfortunately for proponents of capital punishment, the coefficient is still statistically insignificant, so there is no evidence of a deterrent effect either

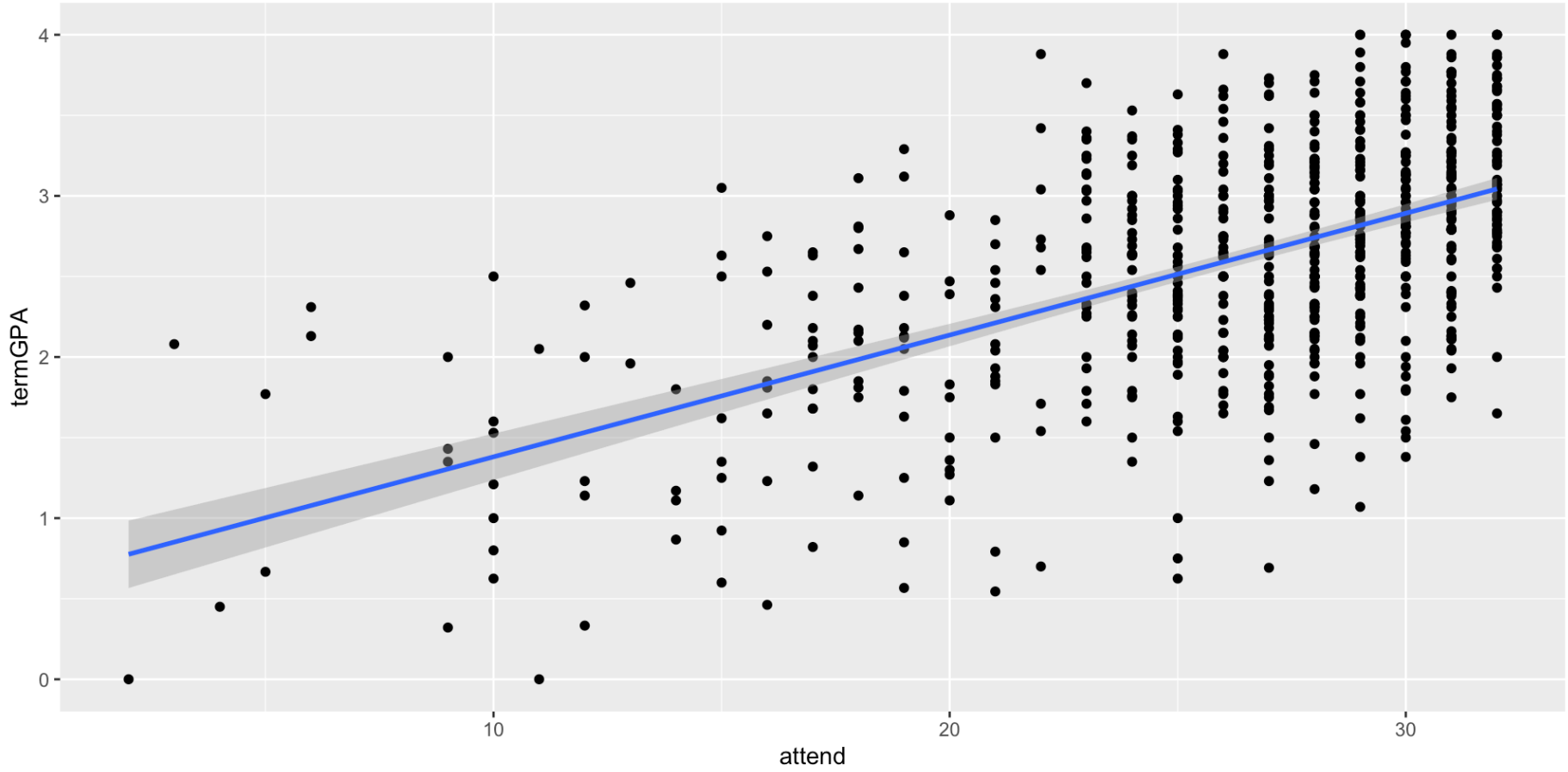
Let's do one more example, this time focusing on the relationshp between attendance and academic performance. Here the key variables are the number of classes attended out of 32 (`attend`) a student attended and their GPA for the term (`termGPA`)

```
1 att <- read_dta("ATTEND.dta")
2 summary(att)
```

```
     attend          termGPA          priGPA           ACT       
 Min.   : 2.00   Min.   :0.000   Min.   :0.857   Min.   :13.00  
 1st Qu.:24.00   1st Qu.:2.138   1st Qu.:2.190   1st Qu.:20.00  
 Median :28.00   Median :2.670   Median :2.560   Median :22.00  
 Mean   :26.15   Mean   :2.601   Mean   :2.587   Mean   :22.51  
 3rd Qu.:30.00   3rd Qu.:3.120   3rd Qu.:2.942   3rd Qu.:25.00  
 Max.   :32.00   Max.   :4.000   Max.   :3.930   Max.   :32.00  

     final           atndrte           hwrte            frosh       
 Min.   :10.00   Min.   :  6.25   Min.   : 12.50   Min.   :0.0000  
 1st Qu.:22.00   1st Qu.: 75.00   1st Qu.: 87.50   1st Qu.:0.0000  
 Median :26.00   Median : 87.50   Median :100.00   Median :0.0000  
 Mean   :25.89   Mean   : 81.71   Mean   : 87.91   Mean   :0.2324  
 3rd Qu.:29.00   3rd Qu.: 93.75   3rd Qu.:100.00   3rd Qu.:0.0000  
```

```
1 ggplot(data=att, aes(x=attend, y=termGPA)) + geom_point() + geom_smooth(met
```



The graph suggests it's a good thing you're here. Let's take a look at the regression:

```r
att.mod1 <- lm(termGPA ~ attend, data=att)
summary(att.mod1)
```

```
Call:
lm(formula = termGPA ~ attend, data = att)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97347 -0.39223  0.00394  0.44460  1.59246

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.624657   0.114773   5.443 7.35e-08 ***
attend      0.075586   0.004297  17.590  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6108 on 678 degrees of freedom
```

Attendance is positive and highly statistically significant

The coefficient of .08 means that for every additional class attended, GPA increases by .08. Perfect attendance would increase your GPA by 2.56 points – not bad

The R^2 of .31 suggests that attendance alone explains 31% of the variation in GPA

But what if there is omitted variable bias? Maybe really good students just happen to attend class more, but they would have done well in their classes even if they didn't attend them. Let's try to control for this by including ACT score and prior GPA as regressors:

```
1 att.mod2 <- lm(termGPA ~ attend + ACT + priGPA, data=att)
2 summary(att.mod2)
```

```
Call:
lm(formula = termGPA ~ attend + ACT + priGPA, data = att)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.71915 -0.27337  0.01315  0.29626  1.55887 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.060329   0.168597  -6.289 5.73e-10 ***
attend       0.054412   0.004173  13.040  < 2e-16 ***
ACT          0.033403   0.006303   5.299 1.57e-07 ***
priGPA       0.574736   0.044125  13.025  < 2e-16 ***
```

The regression suggests that there's a *bit* of that going on, because now the coefficient on `attend` shrinks to .05. But it's still positive and statistically significant

If we believed that `ACT` and `priGPA` were the only omitted variables, we could conclude that each additional class attended increases one's GPA by about .05 points (or an extra 1.5 GPA points for perfect attendance)

I guess I'll see you next time

Let's test the hypothesis that the effect of attendance is the same as the effect of one's ACT score (i.e., $\beta_{\text{attend}} = \beta_{\text{ACT}}$):

```
1 library(car)
2 linearHypothesis(att.mod2, c("attend = ACT"))
```

```
Linear hypothesis test:
attend - ACT = 0

Model 1: restricted model
Model 2: termGPA ~ attend + ACT + priGPA

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    677 171.6
2    676 168.7  1    2.8977 11.611 0.0006944 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null is easily rejected

Presenting regression results

This is a good place to discuss how to communicate the results from your regressions

Results from simple regressions are often presented as equations, with standard errors in parentheses, like this:

$$\widehat{gpa}_i = \underset{(0.114)}{.625} + \underset{(0.004)}{.076} \cdot attend_i$$

Of course, if you use this, you might want to also note the sample size, R^2, and any other important stats (like F tests, etc.)

When you have more than one regression, it's helpful to present the results as a table that compares different models

There are a few different R packages that allow you to do this

```r
library(texreg)
screenreg(list(att.mod1, att.mod2))
```

```
===================================
              Model 1     Model 2
-----------------------------------
(Intercept)    0.62 ***   -1.06 ***
              (0.11)      (0.17)
attend         0.08 ***    0.05 ***
              (0.00)      (0.00)
ACT                        0.03 ***
                           (0.01)
priGPA                     0.57 ***
                           (0.04)
-----------------------------------
R^2            0.31        0.54
Adj. R^2       0.31        0.54
Num. obs.    680         680
```

This also gives you extra info like sample size, R^2, etc. You can also use these packages to export to formats that can be pasted into Word processors