Regression part I

Why regression?

The covariance and correlation tell us about the “relationship” between variables, but we want to know more:

- What is the **prediction** of y given a value of x?
- How does the prediction of y change when x changes?
- What is the **causal effect** of x on y

The last two aren’t always the same thing, we’ll have a lot to say about when a statistical relationship can also be interpreted as a causal effect

The population regression line

Let's assume that the equation for y is

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

In this equation,

- α is the **population intercept**. This is the value that y would take if x and ε were both zero
- β is the **population slope**. We interpret this as the **causal effect** of x on y
- ε is the **error term**. This reflects the fact that y is random, and affected by things other than x

α and β are the **population** intercept and slope

We can think of these as the intercept/slope in the true equation that affects the whole population

Since we don’t know the whole population, we want to come up with **estimate** of α and β based on our particular sample

Estimating the regression line

The population line is

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

and we want to use our sample to estimate α and β

Our estimated line will take the form

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

a will be our estimate of α and b will be our estimate of β

The “hat” indicates that our line will give us a **prediction** of y_i and not the actual value

If we have estimates a and b, the **residual** for our estimated line is the difference between the true value of y_i and our prediction $\widehat{y}_i$:

$$e_i = y - \widehat{y}_i$$

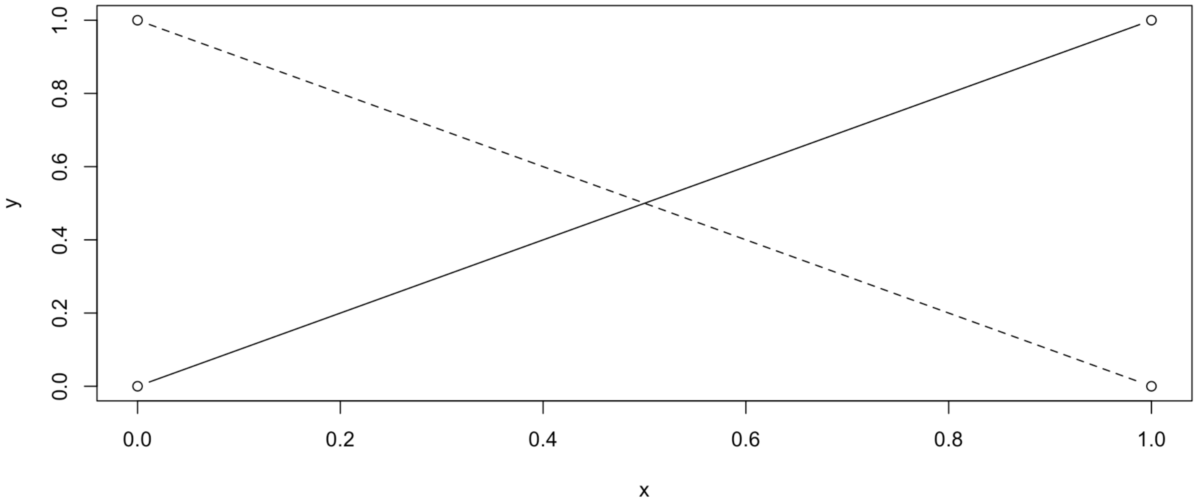
In general, we want our residuals to be as small as possible

However, we will have n residuals, one for each observation, so we need some way of making “all” of the residuals small

We could try to make the *sum* of the residuals as small as possible

However, it turns out that this doesn’t work:

```r
x <- c(0, 1)
y <- c(0, 1)
plot(x,y, type="b")
y1 <- c(1,0)
lines(x,y1, type="b", lty=2)
```



The problem is that large *negative* residuals cancel out with large *positive* residuals, so that even a bad line can have the residuals sum to zero

Instead, we will try to minimize the sum of *squared* residuals

We will choose our estimates a and b to be the values that minimize

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - a - bx_i)^2\end{aligned}$$

Finding the values of a and b that minimize the sum of squared residuals is a standard calculus problem

To solve it, we'll make a simplification: Assume that y and x are in *deviations-from-means* form (i.e., instead of y we have $y - \bar{y}$, and similarly for x)

Now, our model implies that the average of y is

$$\bar{y} = \alpha + \beta \bar{x}$$

However, if y is in deviation-from-mean form, the mean has to be zero (since the average of $y_i - \bar{y}$ is $\bar{y} - \bar{y} = 0$)

In this case, we can drop the constant a, so the model is just $y_i = \beta x_i$

Now we only need to find the value of b that minimizes

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - bx_i)^2$$

Recall from calculus that to minimize a function, we take the derivative with respect to the variable (which is b in this case), and set it equal to zero

Also recall that the derivative of a sum is the sum of the derivatives

The derivative of one term in the sum with is just

$$2(y_i - bx_i)(-x_i)$$

So the derivative of the SSR is

$$(-2)\sum_{i=1}^{n}(y_i - bx_i)x_i$$

At the minimum, this must equal zero:

$$(-2)\sum_{i=1}^{n}(y_i - bx_i)x_i = 0$$

We can divide both sides by -2 to get

$$\sum_{i=1}^{n}(y_i - bx_i)x_i = 0$$

We can rewrite this as

$$\sum_i y_i x_i + b \sum_i x_i^2 = 0$$

Solving for b:

$$b = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

What happens if y and x are not in deviations-from-means form?

We can simply replace them with deviations from means:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Also, since we know that $\bar{y} = \alpha + \beta\bar{x}$, we can estimate α as

$$a = \bar{y} - b\bar{x}$$

These are the **Ordinary Least Squares (OLS)** estimates of α and β

Properties of the residuals

Let’s quickly discuss two properties of the OLS residuals

Property 1. $\bar{e} = 0$

Why? We know that

$$e_i = y_i - \widehat{y}_i = y_i - a - bx_i$$

which implies that

$$\bar{e} = \bar{y} - a - b\bar{x} = 0$$

since $\bar{y} = a + b\bar{x}$. Since $\bar{e} = (\sum_i e_i)/n$, this also implies that the residuals sum to zero

Property 2. $Cov(x, e) = 0$

Recall that, taking the derivative of the SSR, we found

$$\sum_i (y_i - bx_i)x_i = 0$$

But $y_i - bx_i = e_i$ so this implies that $\sum_i e_i x_i = 0$

But

$$Cov(x_i, e_i) = \frac{\sum_i (x_i - \bar{x})(e_i - \bar{e})}{n - 1} = \frac{\sum_i x_i e_i}{n - 1} = 0$$

where we have used the fact that x (which is in deviations from means form) and e have mean zero

Interpreting the regression line

Our estimated line is

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

We interpret the estimated slope coefficient b as

$$b = \frac{\Delta \widehat{y}}{\Delta x}$$

In words, b tells us how much we *predict* y will change when x increases by one unit

We also want to interpret b as our estimate of β, the *population* slope coefficient

However, in order to do that, certain assumptions must be satisfied

We will have more to say about this in the next section of the class

But we can always interpret b as the relationship between x and $\hat{y}$

Regression in action

E.g.:

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

```

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
```

Let’s run the regression “manually”:

```r
wbar <- mean(gril$wage)
wdev <- gril$wage - wbar
sbar <- mean(gril$s)
sdev <- gril$s - sbar
b <- sum(wdev*sdev)/sum(sdev^2)
b
```

```
[1] 33.51068
```

```r
a <- wbar - b*sbar
a
```

```
[1] -125.295
```

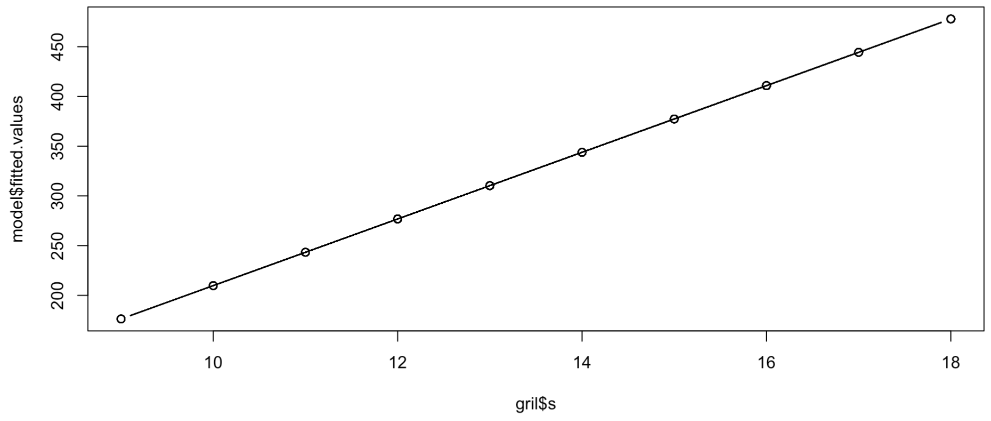
It works! In general, we still want to use the built-in function, because it does more for us and things get complicated with more variables

Our estimated regression line is

$$\widehat{wage}_i = -125.3 + 33.5 * school$$

Graphically:

```r
plot(gril$s, model$fitted.values, type="b")
```



The intercept of -125.3 means that for someone with no schooling, the wage sill be -125.3

This sounds ridiculous, since nobody can have a negative wage

However, nobody has zero schooling:

```r
summary(gril$s)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   12.00   12.00   13.41   16.00   18.00 
```

The smallest amount of schooling in the sample is 9 years

The predicted value for someone with 9 years of schooling is

$$\widehat{wage}_i\Big|_{s=9} = -125.3 + 33.5 * 9 = 176.2$$

Since

$$\widehat{wage}_i = -125.3 + 33.5 * school,$$

the effect of an increase in schooling on the predicted wage is

$$\frac{\Delta \widehat{wage}_i}{\Delta s} = 33.5$$

We interpret this to mean that *for every additional unit of schooling, the wage is predicted to increase by $33.50*

“Goodness of fit”

The sample variance of y is

$$Var(y) = \frac{\sum_i (y_i - \bar{y})}{n - 1}$$

We can think of the numerator of this as the “total variation” in y

When we run a regression, we might be interested in knowing how good a job the regression does of helping us explain this variation

It turns out (it's not hard to prove, but we'll skip it in the interest of time) that

$$\sum_i (y_i - \bar{y}) = b^2 \sum_i (x - \bar{x})^2 + \sum_i e_i^2$$

In other words, the total variation in y can be decomposed into the part that is explained by the regression and the unexplained part

The R^2 is the *fraction of the total variation in y* explained by the regression (or by x):

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

We usually use whichever form is easiest to work with

This gives us a sense of how good a job our regression does of “explaining” y

Don’t put too much emphasis on it. We’re usually interested in knowing whether x affects y, and this can be true even if x only explains a small fraction of the variation in y

E.g.:

```r
summary(model)
```

```

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
```

In the wages example, $R^2 \approx .25$, which means that education explains roughly 1/4 of the variation in y

Not bad, but while this is informative, we're more interested in understanding how education affects wages (more on this in the next section)