Regression part II

Introduction

Our population regression line is

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

As we have seen, we can use sample data to estimate this line as

$$\hat{y}_i = a + b x_i$$

We want to say more about how we can interpret our estimated regression line

Assumptions of the regression model

Before we discuss the interpretation of our estimated regression, we need to make a few more assumptions about the population regression line

Later, we will see how some of these assumptions can be relaxed

Assumption 1: Linearity. The line can be written

$$y_i = \alpha + \beta x_i + \varepsilon.$$

In other words, y is a linear function of x and not something funky like a polynomial (although later we will relax this)

Assumption 2: Mean-zero errors. *Conditional on all of the values of* x in our sample (i.e., $x_1, x_2, \ldots, x_n$), $E(\varepsilon) = 0$ (i.e., if we took the expected value of ε, holding the x's constant)[1]

Since the mean of ε is zero regardless of the values of the x's, this implies that $Cov(x, \varepsilon) = 0$

It also implies that the *unconditional* mean of ε is zero (i.e., if we took the expected value of ε without holding the x's constant)

As we'll see, this is they key assumption that allows us to interpret β as the **causal effect** of x on y

1. Technically, this is the conditional expectation $E(\varepsilon|x_1, x_2, \ldots, x_n)$

The real assumption here is that mean of ε doesn't depend on the x's, but as long as we have a constant α in our model, we can assume that the mean is zero

Suppose that the mean were μ instead. We could redefine the constant to be $\alpha + \mu$ and redefine the error term to be $\varepsilon - \mu$

Then the model would be

$$y_i = (\alpha + \mu) + \beta x_i + (\varepsilon - \mu)$$

and the mean of the error would be

$$E(\varepsilon - \mu) = \mu - \mu = 0$$

Assumption 3. Homoskedasticity. *Conditional on the x*'s in our sample, the variance of ε is constant:

$$Var(\varepsilon) = \sigma^2$$

The real assumption here is that the variance of ε does not change from observation to observation

Since $Var(\varepsilon) = E[\varepsilon - E(\varepsilon)]^2$ and $E(\varepsilon) = 0$, we can also write this as

$$E(\varepsilon^2) = \sigma^2$$

Is this a reasonable assumption?

Suppose that we are interested in the model

$$wage_i = \alpha + \beta_1 educ_i + \varepsilon_i$$

We might think that people with more education can work a greater variety of jobs, which would imply that the variance of their error term would be greater

Homoskedasticity would rule this out

We make this assumption mostly for simplicity, but later we'll see how to relax it

Assumption 4: Serially uncorrelated errors. *Conditional on the* x's in our sample, every observation's error is uncorrelated with with every other observation's error:

$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

for $i \neq j$

Is this a reasonable assumption?

In our wage regression, ε represents factors other than education that affect the wage.

We might expect people living in the same state to face similar local economic conditions, which might affect their wages similarly

Or we might have a model like

$$GDP_t = \alpha + \beta TaxRate_t + \varepsilon_t$$

where observations represent values for the US at different points in time

If the other determinants ε_t of GDP are correlated over time, this will be violated

Again, we make this assumption for simplicity, but we'll see how to relax it later

Interpreting the population regression line

Recall that the model is

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

Under our assumptions (and conditional on the x's),

$$E(y_i) = \alpha + \beta x_i + E(\varepsilon_i) = \alpha + \beta x_i$$

Since, *under our assumptions*,

$$E(y_i) = \alpha + \beta x_i$$

and

$$\hat{y}_i = a + b x_i$$

is our estimate of this line, we can also interpret $\hat{y}$ as an estimate of $E(y_i)$

Therefore, we can also interpret b as an estimate of $\Delta E(y_i)/\Delta x$

Properties of regression

We now discuss some formal properties of our regression estimates, *under the assumptions that we have made*

Property 1: Unbiasedness. *Conditional on the* x's, $E(b) = \beta$

Two ways to interpret this:

1. Since b is a function of the random variables $y_1, y_2, \ldots, y_n$, b is also a random variable that has a distribution. Unbiasedness means that the distribution of b is centered at β
2. If we had many different samples and estimated our regression in each of them, the average of the b's would be β

Let's *prove* that b is unbiased (using our deviations-from-means trick).

First, let's use the fact that $y_i = bx_i + \varepsilon$ (plus a few sum properties) to rewrite b:

$$
\begin{aligned}
b &= \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\sum_i x_i(\beta x_i + \varepsilon_i)}{\sum_i x_i^2} \\
&= \frac{\sum_i (\beta x_i^2 + x_i \varepsilon_i)}{\sum_i x_i^2} = \frac{\sum_i \beta x_i^2}{\sum_i x_i^2} + \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2} \\
&= \beta + \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2}
\end{aligned}
$$

Now,

$$E(b) = E\left(\beta + \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2}\right) = \beta + E\left(\frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2}\right)$$

Conditional on the x's (treating them as constants), this becomes

$$E(b) = \beta + \frac{1}{\sum_i x_i^2} E\left(\sum_i x_i \varepsilon_i\right) = \beta + \frac{1}{\sum_i x_i^2} \sum_i E(x_i \varepsilon_i)$$

$$= \beta + \frac{1}{\sum_i x_i^2} \sum_i x_i \underbrace{E(\varepsilon_i)}_{=0} = \beta$$

Property 2. The variance of b is

$$Var(b) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}.$$

Although we know that b is unbiased, knowing its variance is helpful for telling us how far we can expect b to be from β simply because b is random

We can prove this, too. Conditional on the x's (and using properties of variances and our deviations-from-means trick)

$$
\begin{aligned}
Var(b) &= Var\left(\beta + \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2}\right) = Var\left(\frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2}\right) \\
&= \frac{1}{\left(\sum_i x_i^2\right)^2} Var\left(\sum_i x_i \varepsilon_i\right) = \frac{1}{\left(\sum_i x_i^2\right)^2} \sum_i Var(x_i \varepsilon \\
&= \frac{1}{\left(\sum_i x_i^2\right)^2} \sum_i x_i^2 Var(\varepsilon_i) = \frac{1}{\left(\sum_i x_i^2\right)^2} \sum_i x_i^2 \sigma^2 \\
&= \frac{\sigma^2}{\left(\sum_i x_i^2\right)^2} \sum_i x_i^2 = \frac{\sigma^2}{\sum_i x_i^2}
\end{aligned}
$$

Property 3: Consistency. b converges to β as the sample size grows[1]

Intuitively, this means that as our sample size gets larger, b gets closer and closer to β

Note the difference between unbiasedness and consistency. Unbiasedness says that if we had many samples, the average of the b's would be close to β. Consistency says that if we had *one* large sample, b would be close to β

1. Technically, converges *in probability*

We can prove this, too

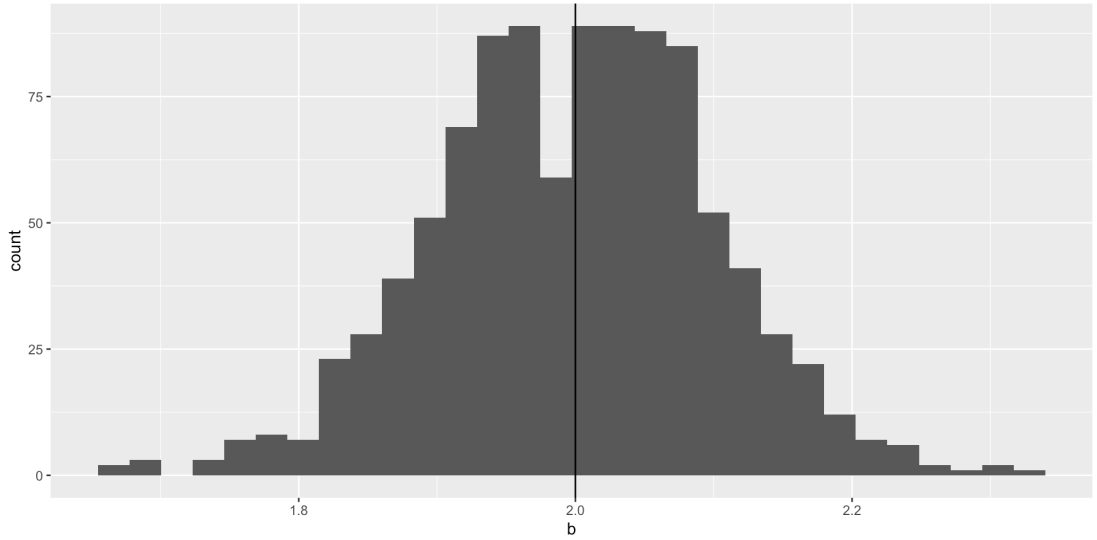
We know that $E(b) = \beta$ and $Var(b) = \sigma^2 / \sum_i (x_i - \bar{x})^2$

As n grows, $\sum_i (x_i - \bar{x})^2$ will get larger and larger, so $Var(b)$ will decrease to 0

Since b is "centered" on β and the variance decreases, it must converge to β

Let’s illustrate these properties, starting with unbiasedness

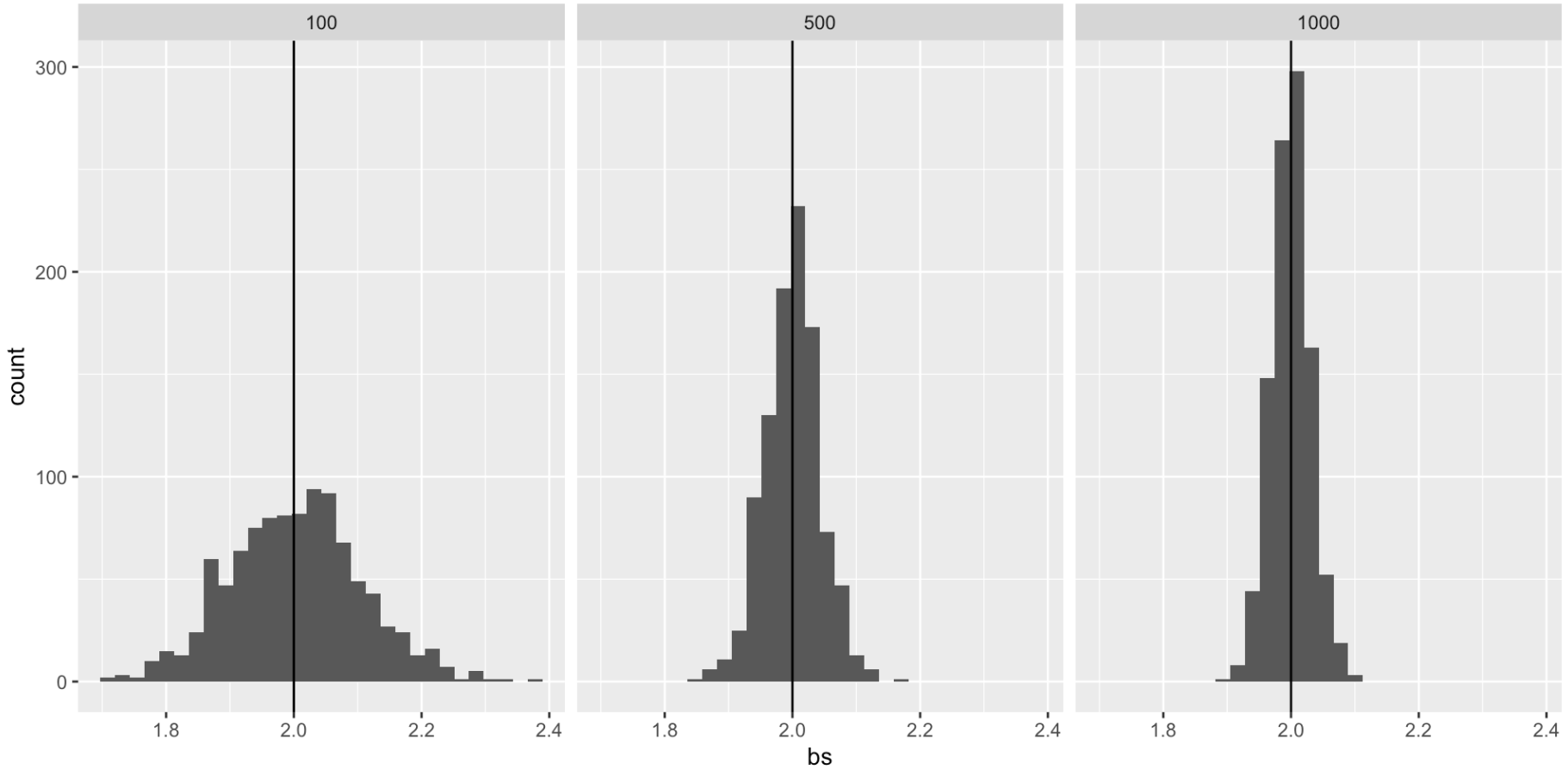
```
library(tidyverse)
b <- rep(0,100)
for (i in 1:1000) {
  x <- rnorm(100)
  e <- rnorm(100)
  y <- 1 + 2*x + e
  b[i] <- lm(y ~ x)$coef[2]
}
ggplot() + geom_histogram(aes(b)) + geom_vline(xintercept=2)
```



Now let's illustrate consistency

```
runreg <- function(n) {
  x <- rnorm(n)
  y <- 1 + 2*x + rnorm(n)
  b <- lm(y ~ x)$coef[2]
}
b1 <- rep(0,1000)
b2 <- b1
b3 <- b1
for (i in 1:1000) {
  b1[i] <- runreg(100)
  b2[i] <- runreg(500)
  b3[i] <- runreg(1000)
}
bs <- c(b1, b2, b3)
n <- c(rep(100, 1000), rep(500, 1000), rep(1000, 1000))
df <- data.frame(bs, n)
```

```r
ggplot(df, aes(bs)) +
  geom_histogram() + geom_vline(xintercept=2) + facet_wrap(n ~ .)
```



Property 4. The Gauss-Markov Theorem

Recall that

$$b = \frac{\sum x_i y_i}{\sum_i x_i^2} = \frac{x_1}{\sum_i x_i^2} y_1 + \frac{x_2}{\sum_i x_i^2} y_2 + \cdots + \frac{x_n}{\sum_i x_i^2} y_n$$

The point is that b is a *linear function* of the y, which means that it is a linear estimator

The **Gauss-Markov Theorem** states that among all *unbiased* and *linear* estimators of β, b is the one that has the smallest variance

This is often remembered by saying that OLS is *BLUE*: the Best Linear Unbiased Estimator (where “best”=“smallest variance”)

We care about this because the variance of our estimator tells us how far away from β it will tend to be just because of randomness. Having the best unbiased estimator means that our estimator will tend to be as close to β as possible

The proof of this is not more difficult than the other proofs that we’ve seen, but since it would take a lot of time, we will skip it

Correlation, causation and all that: A preview

In our wage example, we estimated that

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

We interpreted this as a *relationship* between x and $\widehat{y}$: for every additional year of schooling, we predict the wage will increase by $33.5

Question: Since b is our estimate of β, which represents the **causal effect** of education on earnaings, can we also interpret b as an estimate of the causal effect?

Answer: *Not necessarily.* In our model, β represents the causal effect of x

Our proof that b is and unbiased and consistent estimate of β assumed that $E(\varepsilon_i) = 0$ (conditional on the x's)

Recall that this is the same as assuming that ε and x are uncorrelated

We can only interpret b as an estimate of the causal effect if this assumption holds

Why might this assumption fail?

Recall that our model is

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

where ε_i represents all of the *other factors besides* x that affect y

It's possible that these factors actually *are* correlated with x

Intuitively, our regression is telling us how y changes when x changes

I.e., the regression is comparing y for units with different values of x

But if x is correlated with ε, which represents all of the other things that affect y, those changes might not be *because* of x

In other words, b might tell us the *relationship* between x and y, but not the *effect* of x on y

In our wage example, ε might include things like ambition, which probably affects wages, but may also be correlated with education

If this is true, our regression may be telling us something about the effect of ambition (which is related to education) on wages, rather than the effect of ambition itself

We will have much more to say about this later

For now, the message is: We can only interpret regression coefficients as causal effects if we believe that ε is not correlated with x

Let’s conclude with one more example (one of the granddaddy’s of regression examples)

```
galt <- read_csv("galton.csv")
summary(galt)
```

```
     parent          child      
 Min.   :64.00   Min.   :61.70  
 1st Qu.:67.50   1st Qu.:66.20  
 Median :68.50   Median :68.20  
 Mean   :68.31   Mean   :68.09  
 3rd Qu.:69.50   3rd Qu.:70.20  
 Max.   :73.00   Max.   :73.70  
```

```r
galton_model <- lm(child ~ parent, data=galt)
summary(galton_model)
```

```
Call:
lm(formula = child ~ parent, data = galt)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.8050 -1.3661  0.0487  1.6339  5.9264 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
```

The estimated relationship between child and parental height is

$$\widehat{child}_i \approx 24 + .65parent$$

Thus, for every additional inch of parental height, we expect the child to be .65 inches taller

This is where regression gets its name – the relationship between parental and child height is not 1:1

A tall parent will still have a tall child, but whey won't be quite as tall as their parent – their is "regression to the mean"