Chapter 2

The Simple Regression Model

Jim Bang

Nonlinearities

- What do we mean by 'linear regression?'

a. that the population regression function is linear in the independent variable(s)
b. that the true relationship between the variables must be linear
c. that the population regression function is linear in the parameters
d. that the regression line minimizes the sum of squared residuals

Wages and Education

- Estimate a linear model for the log of wages on the level of education and call it lwage.lm1.
- Summarize the output

```
lwage.lm1 <- lm(log(wage) ~ educ, data = wage1)
summary(lwage.lm1)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21158 -0.36393 -0.07263  0.29712  1.52339
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
```

```
## educ        0.082744   0.007567  10.935  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4801 on 524 degrees of freedom
## Multiple R-squared:  0.1858, Adjusted R-squared:  0.1843
## F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16
```

Interpreting Regression Coefficients

- In the regression, $wage = \beta_0 + \beta_1 educ + u$, what is the economic interpretation of β_1?

a. that a one-year increase in education leads to a β_1 dollar increase in hourly wage on average
b. that a one-year increase in education leads to a β_1 percent increase in hourly wage on average
c. that a one percent increase in education leads to a β_1 percent increase in hourly wage on average
d. that a one-year increase in education leads to a β_1 dollar increase in hourly wage always

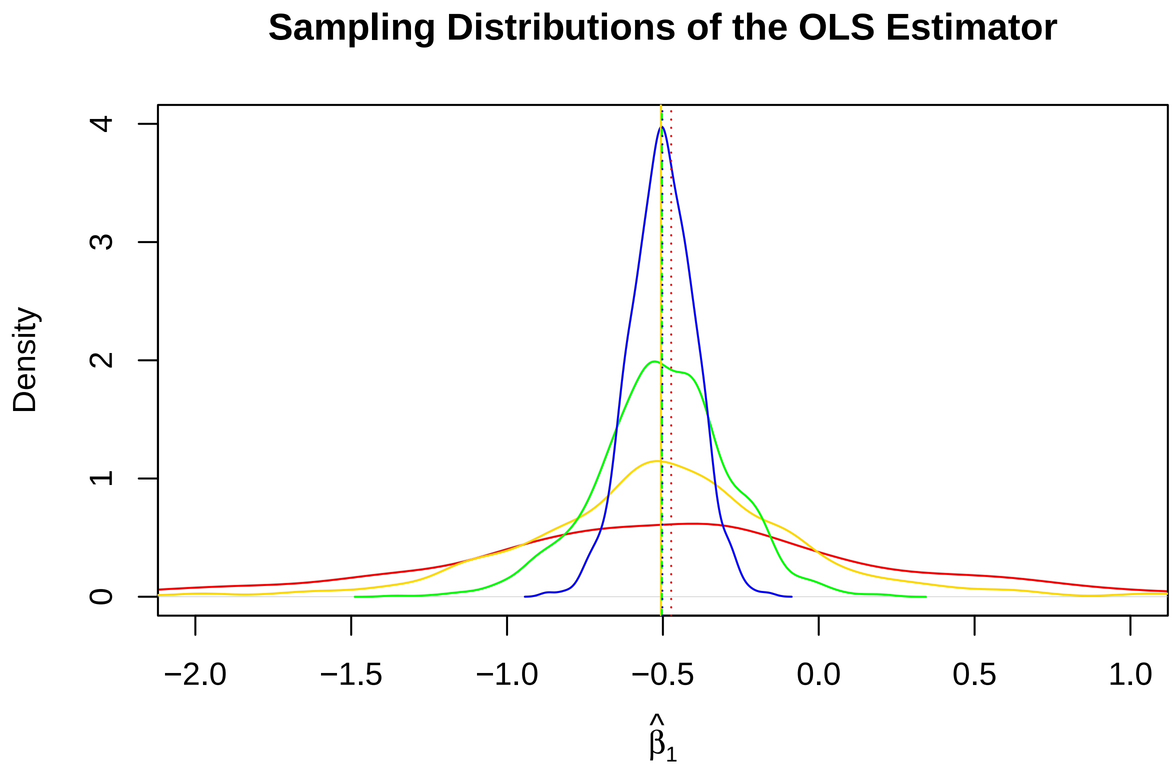
- In the regression, $\log(wage) = \beta_0 + \beta_1 educ + u$, what is the economic interpretation of β_1?

a. that a one-year increase in education leads to a β_1 dollar increase in hourly wage on average
b. that a one-year increase in education leads to a β_1 percent increase in hourly wage on average
c. that a one percent increase in education leads to a β_1 percent increase in hourly wage on average
d. the log-linear model is better than the linear model

Expected Values and Variances of OLS Estimators

Recall in the practice for Appendix C we simulated 1000 resamples of a simple OLS regression for sample sizes from 1 to 100. Run the following code to view the density plots for this simulation with vertical lines colored corresponding to the density they describe and textured differently for visibility.

```
set.seed(8675309)
X <- NULL
u <- NULL
Y <- NULL
b1 <- matrix(NA, nrow = 1000, ncol = 100)
for(i in 1:100) {
  for(j in 1:1000) {
    X <- rexp(n = i, rate = 1)
    u <- rnorm(n = i, mean = 0, sd = 1)
    Y = 2 - 0.5*X + u
    b1[j, i] = lm(Y ~ X)$coefficients[2]
  }
}
plot(density(b1[,5]), xlim = c(-2, 1), ylim = c(0, 4), col = 'red',
  main = "Sampling Distributions of the OLS Estimator", xlab = expression(hat(beta)[1]))
lines(density(b1[,10]), col = 'gold')
lines(density(b1[,30]), col = 'green')
lines(density(b1[,100]), col = 'blue')
abline(v = mean(b1[,5]), col = 'red', lty = 'dotted')
abline(v = mean(b1[,10]), col = 'gold', lty = 'solid')
abline(v = mean(b1[,30]), col = 'green', lty = 'dashed')
abline(v = mean(b1[,100]), col = 'blue', lty = 'dotted')
```



Unbiasedness of the OLS Estimator

Assumptions:

1. Linearity in Parameters
2. Random Sampling: (X_i, Y_i) are independently and identically distributed
3. Sample Variation: x_is vary - rules out perfect collinearity with constant
4. Zero Conditional Mean of u. $E(u|x) = 0$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$E[\hat{\beta}_1] = E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i+u_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

$$= E\left[\frac{\beta_0\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\beta_1\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2})\right]$$

The following properties give us the result: 1. $\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n} x_i - n\sum_{i=1}^{n}\frac{x_i}{n}) = 0 \Rightarrow E\left[\frac{\beta_0\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = 0$ 2. $\sum_{i=1}^{n}(x_i-\bar{x})x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 \Rightarrow \frac{\beta_1\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1$ 3. $E\left[\sum_{i=1}^{n}(x_i-\bar{x})u_i\right] = 0$ by the assumption that the expectation of the errors conditional on x equals zero.

Variance of the OLS Estimator

Additional Assumption:

- Homoskedasticity: $u_i{}'s$ have constant variance regardless of the value of x.

$$Var(u|x) = 0$$

Standard Error of the Regression: $\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} u_i^2}$ - the standard deviation of the residuals

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\hat{\sigma}^2}{(n-1)\hat{\sigma}_x^2}$$

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \frac{\hat{\sigma}^2\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\hat{\sigma}^2}{(n-1)\hat{\sigma_x}^2}\cdot\frac{\sum_{i=1}^{n} x_i^2}{n}$$

Sampling Distribution of $\hat{\beta}_1$ & $\hat{\beta}_0$

By the Central Limit Theorem, for sufficiently large n,

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_{\hat{\beta}_1}^2}{n})$$

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_{\hat{\beta}_0}^2}{n})$$