Overview of First Half of ECON 4050

Dr. Soliman

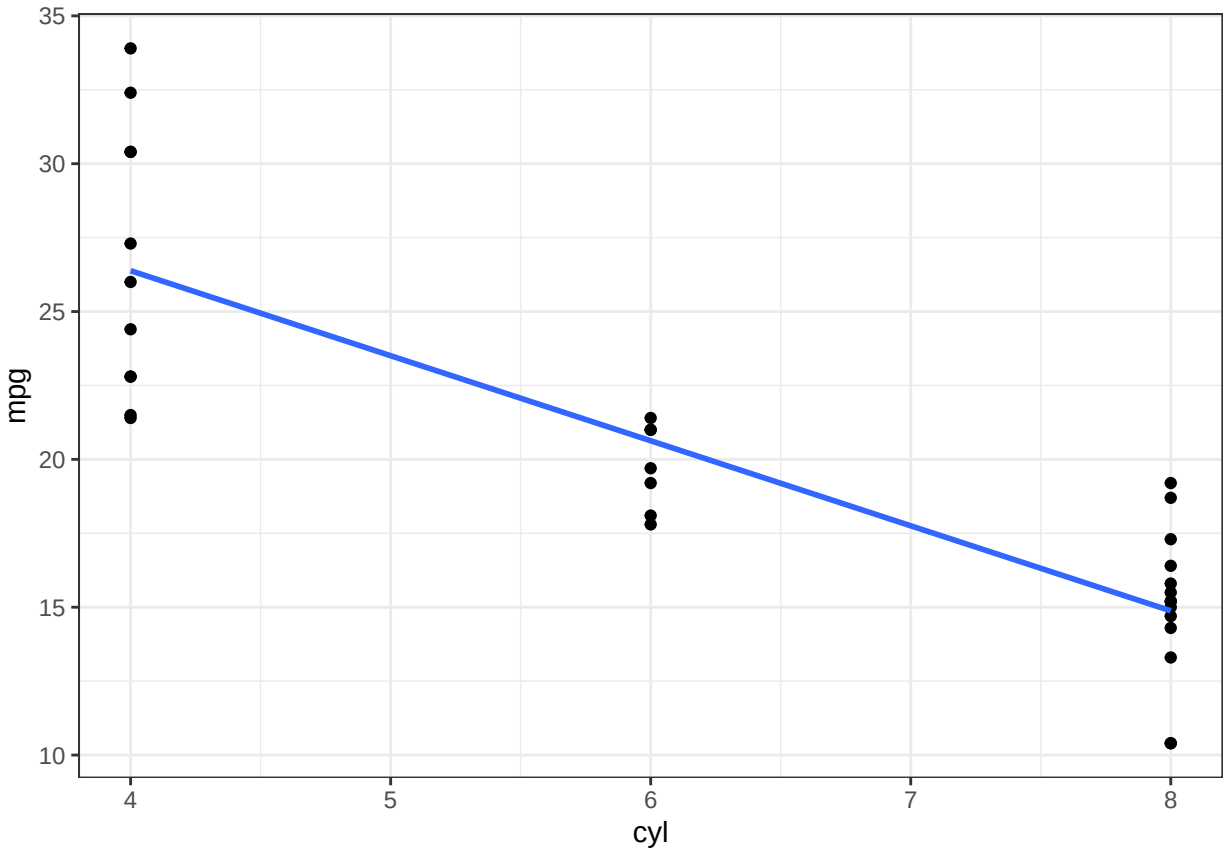
Always understand the data before using it by doing something like this:

```
summary(mtcars)
view(mtcars)
```

The Linear Regression

1. Create a scatter plot with mpg on the y-axis and cyl on the x-axis, which is the relationship between miles per gallon and cylinders for a given car.

```
## `geom_smooth()` using formula = 'y ~ x'
```



2. Regress mpg on cyl (dependent on independent) and interpret each coefficient.

```
lm(mpg ~ cyl, data = mtcars)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ cyl, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl
##      37.885       -2.876
```

Intercept: a car with no cylinders (what?!) has on average 38 miles per gallon.

cyl coefficient: on average, an increase of 1 cylinder is associated with a decrease of 2.9 miles per gallon.

3. Add the variable `am` to the previous model and rerun the regression. The `am` variable represents the type of transmission (0 = automatic, 1 = manual). What is the interpretation of each coefficient.

```
lm(mpg ~ cyl + am, data = mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + am, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl           am
##      34.522       -2.501        2.567
```

Intercept: a car with no cylinders (what?!) and automatic transmission has on average 35 miles per gallon.

cyl coefficient: holding transmission constant, on average, an increase of 1 cylinder is associated with a decrease of 2.5 miles per gallon.

am coefficient: holding the number of cylinders constant, manual cars have on average 2.6 more miles per gallon relative to automatic transmission cars.

4. Is this relationship causal? If not, what is a potential omitted variable that may be important to consider when examining the relationship between mpg and cyl?

No, it is not causal, just a correlation. The price of the car and thus the associated features could be associated with both miles per gallon and the cylinders in the car. However, there are many potential omitted variables, such as weight

5. Create a new object called `mpgextra` and add a variable that is the log of mpg. Why would we want to take the logaritm of a variable? Rerun the regression from question 2 but using the log of mpg as the outcome and interpret the coefficients.

```
mpgextra <- mtcars %>% mutate(logmpg = log(mpg))
lm(logmpg ~ cyl, data = mpgextra)
```

```
##
## Call:
## lm(formula = logmpg ~ cyl, data = mpgextra)
##
## Coefficients:
## (Intercept)          cyl
##      3.8393       -0.1425
```

Intercept: a car with no cylinders (cmon!) has a log miles per gallon of 3.8. More formally, it is exp(3.8393) = 46.5 miles per gallon.

cyl coefficient: on average, an increase of 1 cylinder is associated with a miles per gallon decrease of 14%. More formally, that is [exp(-0.1425) - 1]X100 = 13.3 percent decrease in miles per gallon.

Sampling, Confidence Intervals, and Hypothesis Testing

6. If you run the same regression on different samples, do you expect to get the exact same coefficient estimates or different ones? Why?

SAMPLING VARIATION! We would not expect to get the same estimates because we are dealing with samples, not the population. We observed in class that a random sample can produce a different estimate by chance, similar to the concept of flipping a coin two times: HH, TT, HT, TH are all possible.

7. When a sample size gets larger, what happens to the sampling variation? Regardless of how the underlying population distribution looks like, when sample means are based on larger and larger sample sizes, the sampling distribution of these sample means becomes both more and more like what type of distribution?

When a sample size gets larger, the sampling variation decreases, or put differently, our estimates of the true population parameter get more precise (smaller standard deviation). We know that due to the central limit theorem, it approaches a normal distribution.

8. How does the width of the confidence interval change as the confidence level increases? Why?

The greater the confidence level, the wider the confidence intervals, as a greater confidence level means the confidence interval needs to contain the true population parameter more often, and thus needs to be wider to ensure this.

9. How does the width of the confidence interval change as the sample size increases? Why?

The greater the sample size, the narrower the confidence intervals, as a larger sample size leads to less sampling variation and therefore a narrower boostrap distribution, which in turn leads to thinner confidence intervals.

10. Define the null and alternative hypothesis. What does it mean to have a very small p-value?

The null hypothesis (H0) is generally a hypothesis of no difference, while the alternative hypothesis (HA or H1) is the research hypothesis. A p-value is the probability of observing a test statistic just as or more extreme than the one we obtained, assuming the null hypothesis H0 is true. Therefore, the lower the p-value, the less consistent our null hypothesis is with the observed statistic. A small p-value means that it is unlikely our observed test statistic is due to chance/ random variation.

11. In economics, what action should you take with regards to the null and alternative hypothesis if you obtain a p-value of 0.007? How about if it was instead 0.7?

If we obtain a p-value of 0.007, as it is less than the alpha of 0.05, we would reject the null hypothesis. If it is instead 0.7, we would fail to reject the null hypothesis.

12. What elements do you need to construct a confidence interval? What us the formula for a 95% confidence interval and what does it mean?

We would need the sample estimate (b) and standard error (se(b)). The 95% CI is b+/-1.96Xse(b), or [b-1.96Xse(b), b+1.96Xse(b)].