GOVT 707 Lab 5, OLS Regression Part 2

Theodore Landsman

September 16, 2021

Problem Set 1 and 2 Review

- This class will feature a brief exercise outlining how to show and export regressions.
- But first, I wanted to use this as an opportunity to go through some of the problems from the problem sets that people had the most conceptual issues with.

Problem Set 1 Question 1: How to Define and Operationalize a Variable

1. Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

Figure 1: PS1

- Let's say we are interested in the *concept* of personal interactions between Professors and Undergraduates.
- We then need to think about what we can *measure.* One measure that could deal with this concept is *number of faculty*.
- But having more faculty is not going to increase faculty student interactions if the number of students also rises, so we can *operationalize* our variable as student-faculty ratio.

Operationalizing Variables Part 2

- Specification issues:
  1) Who counts as faculty, am I a faculty member? Are adjuncts faculty members? Maybe we specify the faculty side as tenure track faculty.
  2) Who counts as students, do you count as students? Probably we go with something like full time undergraduate students.
- However, particularly when specifying a dependent variable, we need to be very careful about understanding which pieces of our data are going into it.
- For example, we could include number of students as a independent variable if our dependent variable is number of professors, but not if our dependent variable is student faculty ratio.

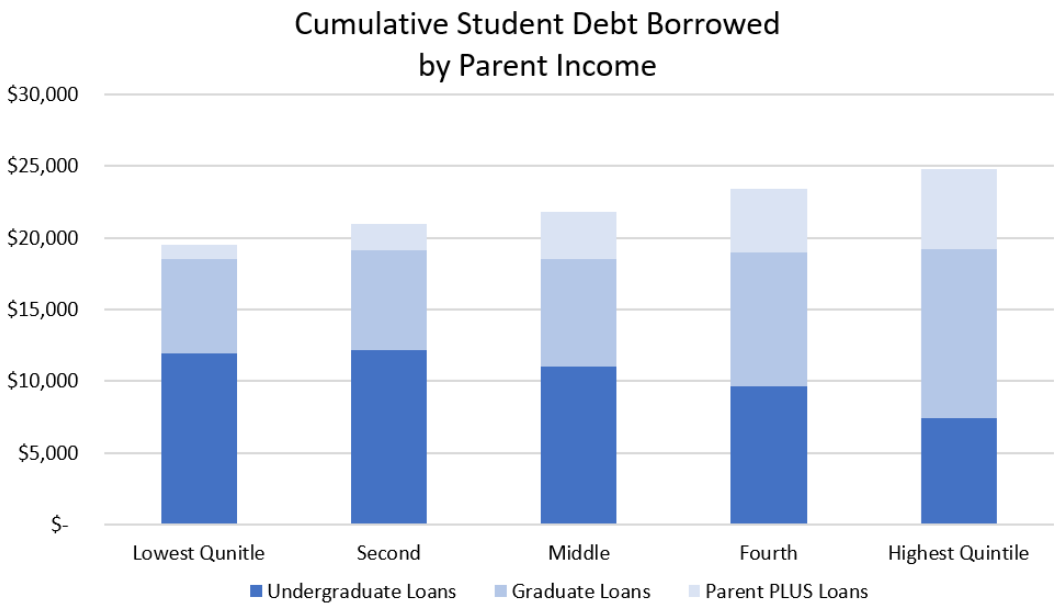
Operationalizing Variables Part 3

- There is not necessarily a right way to specify a variable once you get to this level, just tradeoffs.
- This is *very important* for those of you who are planning on using a dataset like VDem where a lot of potential DVs (like Democracy) are *very complicated*.
- Be careful that you are examining relationships that *might* be causal, not relationships that VDem *assumes* to be causal (like press freedom and polyarchy).

A Good Example: Problem Set 2 Question 2

2. How well does the income of a college student's parents predict how much the student will borrow for college? We have data on parents' income and college debt for a sample of 1200 recent college graduates. What are the explanatory and response variables? Are these variables categorical or quantitative? Do you expect a positive or negative association between these variables? Why?

- Most of you expected a negative relationship, a few of you expected a positive relationship. I expected a quadratic relationship with debt peaking in the middle, who is right?

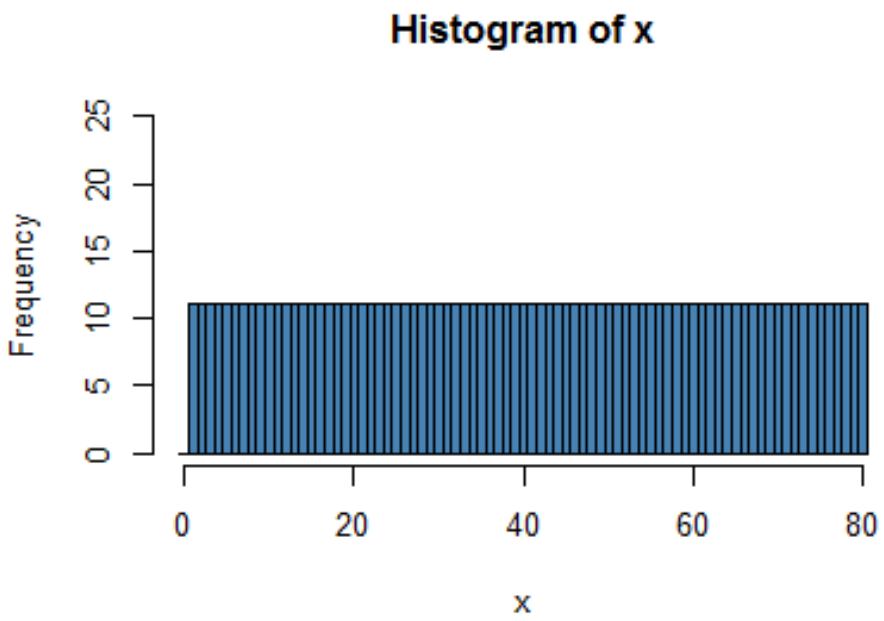


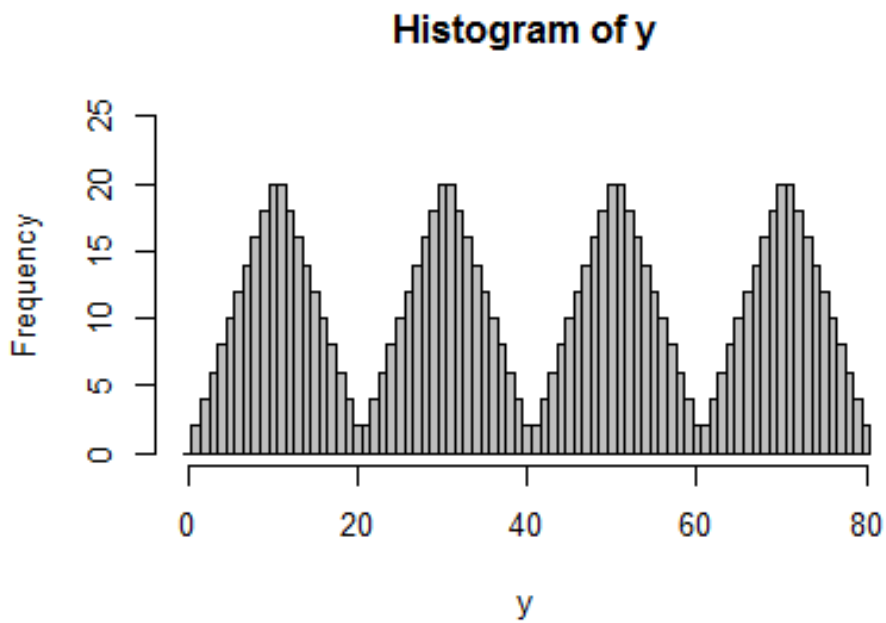
Review: Problem Set 2 Question 1

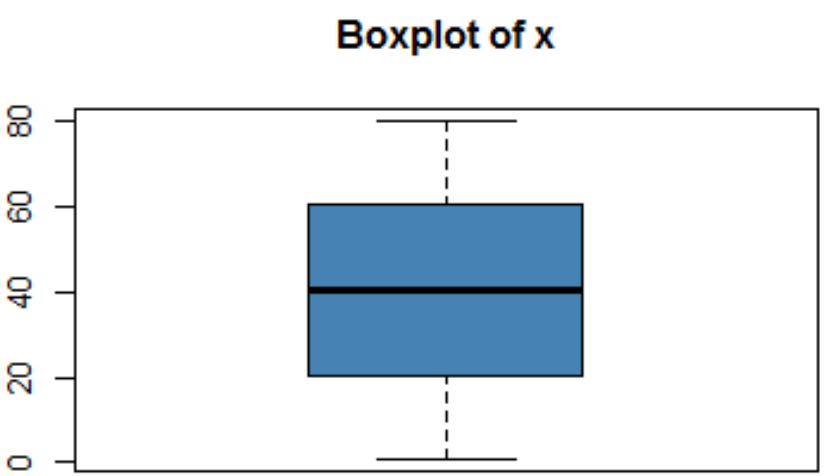
1. If two distributions have exactly the same mean and standard deviation, must their histograms have the same shape? If they have the same five-number summary, must their histograms have the same shape? Explain.

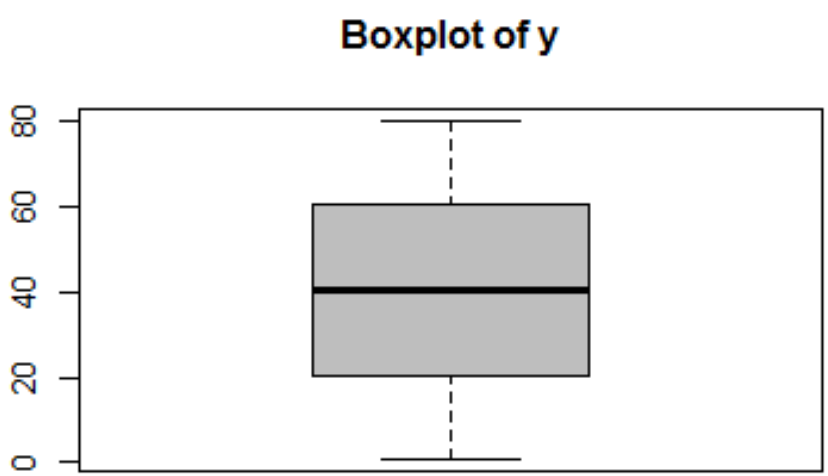
Figure 2: PS2

Review Problem Set 2 Question 1 Answer









Problem Set 2 Question 3

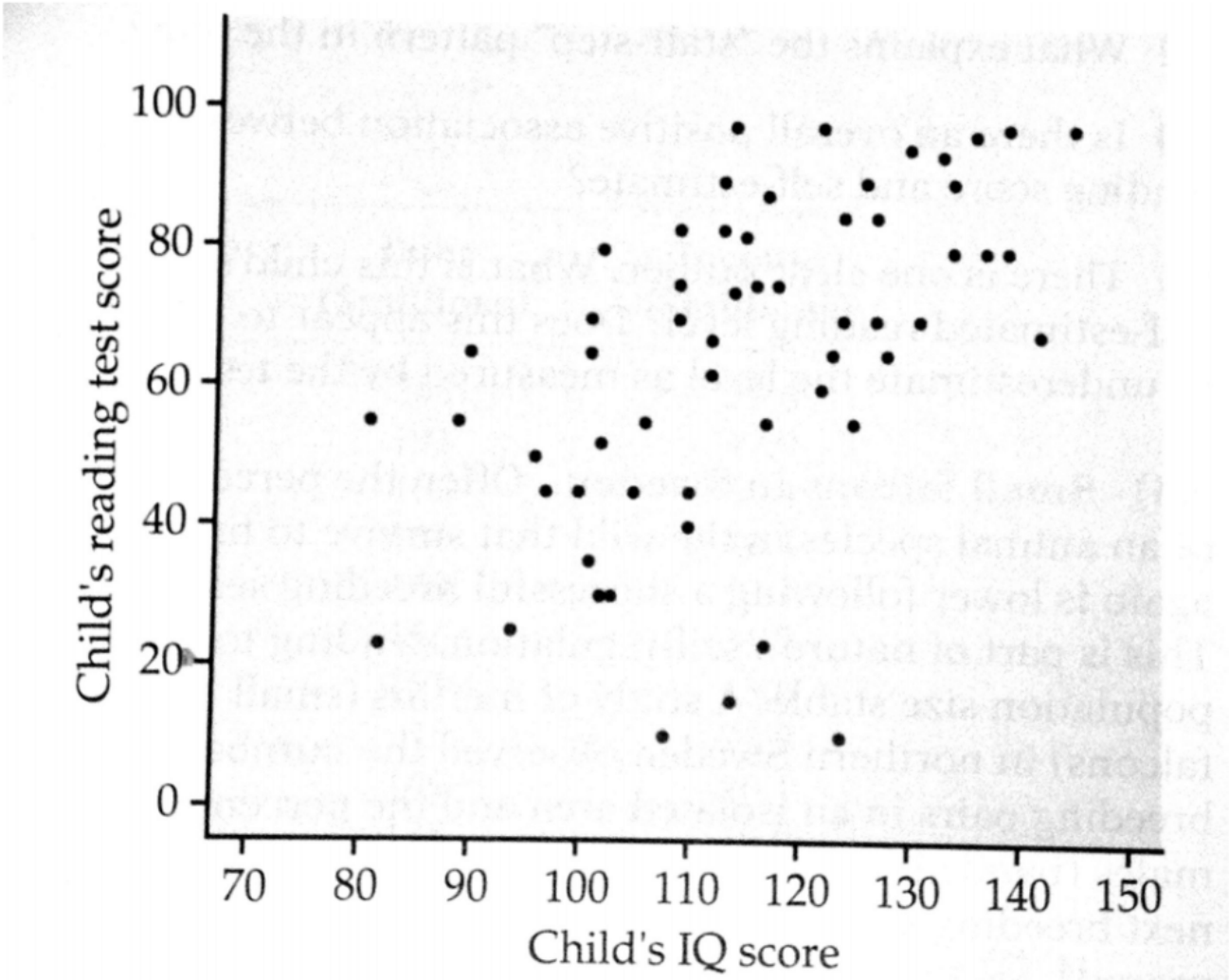


Figure 1: IQ and reading test scores for 60 fifth-grade children

Problem Set 2 Question 4

4. Investment reports often include correlations. Following a table of correlations among mutual funds, a report adds, "Two funds can have perfect correlation, yet different levels of risk. For example, Fund A and Fund B may be perfectly correlated, yet Fund A moves 20% whenever Fund B moves 10%." Write a brief explanation, for someone who knows no statistics, of how this can happen. Include a sketch to illustrate your explanation.

Figure 5: PS_2_4_Review

Problem Set 2 Question 4 Answer

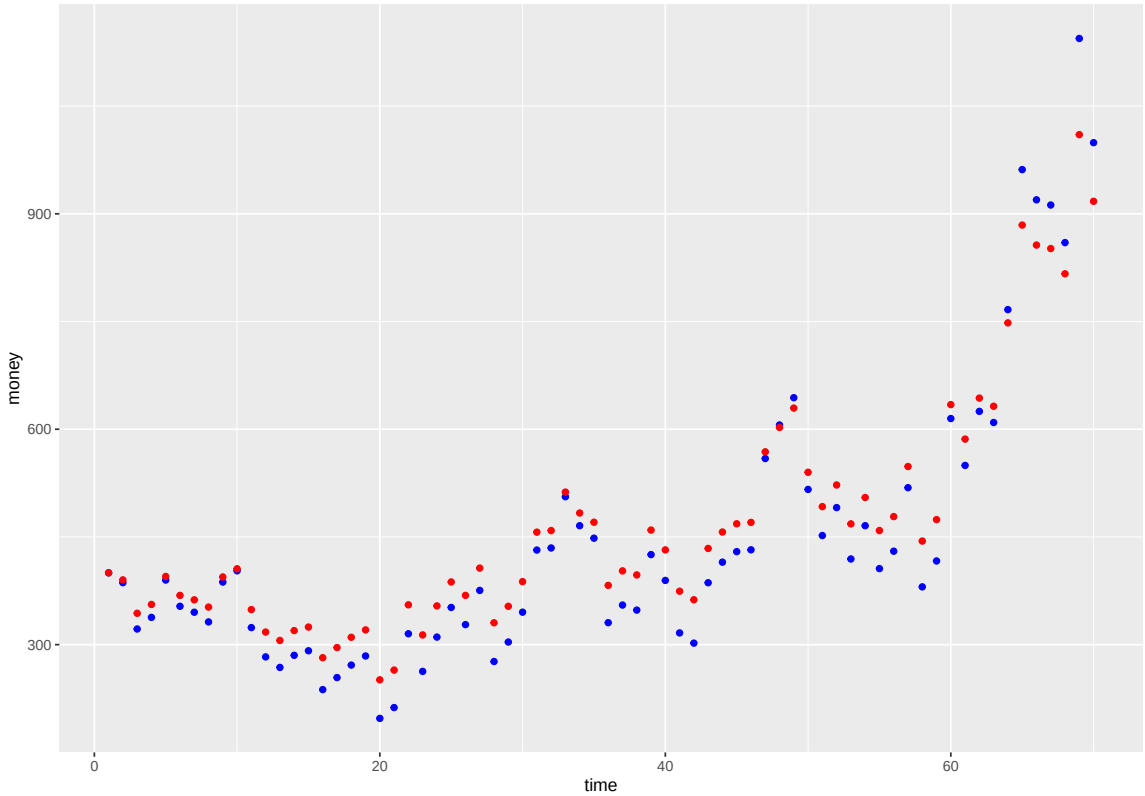
- Let's imagine that the stock market is a stochastic process with a mean of .05 and a SD of .1.
- In other words, stocks tend to grow by 5% a [unit of time] but can grow by up to 15% or decline by 5% in a typical [unit of time].
- Let's imagine that bonds are a safer asset class than stocks, meaning that they move by 1/2 of the amount that stocks move.

```
set.seed(1000)
market_growth <- rnorm(100, mean = .05, sd = .2)
stocks <- rep(100,100)
bonds <- rep(100,100)

for(i in 1:100) {
  stocks[i+1] <- stocks[i] + stocks[i]*market_growth[i]
  bonds[i+1] <- bonds[i] + bonds[i] * market_growth[i]/2
}
money <- (stocks*3) + bonds
moneyB <- stocks + (bonds*3)
time <- seq(1,101)
```

Problem Set 2 Question 4 Answer Graph

```
portfolios <- data.frame(time, money, moneyB)
ggplot(data = portfolios[1:70,],aes(x =time)) +
  geom_point(aes(y = money), color = "blue") +
  geom_point(aes(y =moneyB), color = "red")
```



Today's Exercise, Visualizing Regressions

- In general there are two values we care about when thinking about the strength of a regression relationship, the estimate (mean) and the standard error.
- Other things will show up on regression tables, such as stars, p-values and t-values, but these are all just products of the estimate, standard error, and number of observations.

```
Call:
lm(formula = numberofdeaths ~ cell_subscription, data = cellphones)

Residuals:
    Min      1Q  Median      3Q     Max
-844.04 -123.11  -56.48  151.64 1036.14

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      123.980493  54.644167   2.269   0.0278 *
cell_subscription   0.091145   0.006095  14.955 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 287.1 on 48 degrees of freedom
Multiple R-squared:  0.8233,    Adjusted R-squared:  0.8196
F-statistic: 223.6 on 1 and 48 DF,  p-value: < 0.00000000000000022
```

Important

Also Importan

Regression Tables

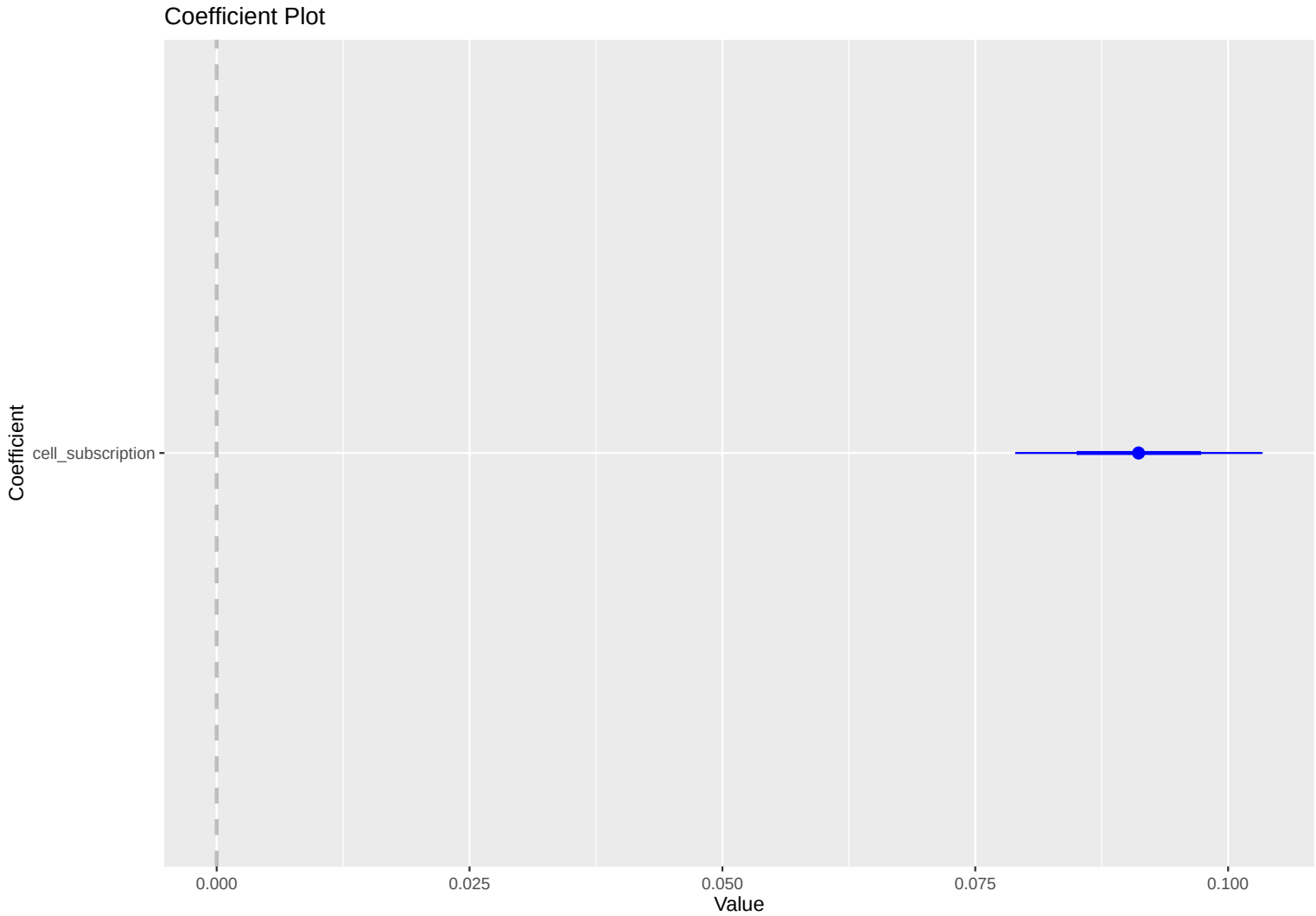
- There are several packages for visualizing regressions in R, one of the most prominent is `stargazer`

Table 1:

	Dependent variable:
	numberofdeaths
cell_subscription	0.091***
	(0.006)
Constant	123.980**
	(54.644)
Observations	50
R^2	0.823
Adjusted R^2	0.820
Residual Std. Error	287.091 (df = 48)
F Statistic	223.648*** (df = 1; 48)

Regression Graphs

- We can also display the estimate and standard error visually, what can we tell about the relationship between cellphone subscriptions and vehicle deaths from this graph?



Exercise

1) Using the cellphones dataset, articulate a relationship between the variables involving 3 or more IVs and the DV `number of deaths`.
2) Create a professional looking regression table and a coefficient plot. Write a very simple 'data essay' (1-2 paragraphs), articulating your hypothesis and model of the data and what the evidence shows.