GOVT 707 Lab 4, OLS Regression Part 1

Theodore Landsman

September 16, 2021

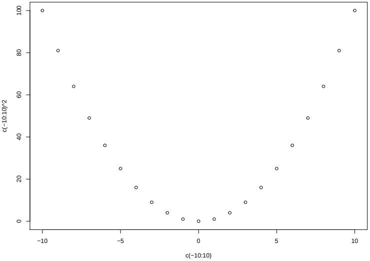
What is OLS Regression

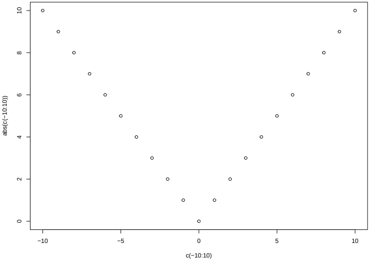
- OLS stands for **O**rdinary **L**east **S**quares
- **Ordinary**: We are not doing any fancy manipulations.
- **Least**: We are minimizing something.
- **Squares**: The thing we are minimizing is a squared term.
- Why is it helpful to square things before taking the sum of them? What else could we do?

Squares

- Defined at 0.
- No discontinuity.
- Plays well with other mathematical operations.

```r
par(mar = c(4, 4, .1, .1))
plot(c(-10:10),c(-10:10)^2)
plot(c(-10:10), abs(c(-10:10)))
```





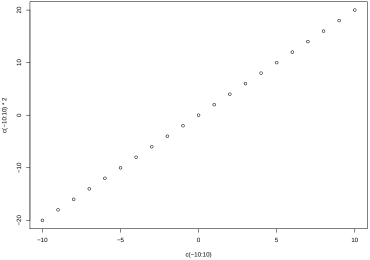
Why to use OLS

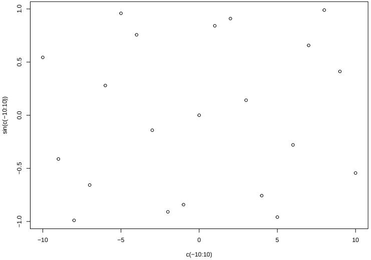
- OLS is used by quatitative methods practitioners for two distinct purposes:
- Social scientists primarily use OLS to assess the cause of something the happened in the past.
- Data scientists primarily use OLS to predict what will happen in the future.
- These uses lead to different practices around data use and model specification, but all within the OLS umbrella.

Other terms for OLS

- Among data scientists, OLS is sometimes treated as part of the broader family of ‘machine learning’ algorithms.
- OLS is also referred to as Linear Regression, because the processes we describe with it are linear (ie additive) and therefore the types of relationships described are linear.

```r
par(mar = c(4, 4, .1, .1))
# Linear relationship
plot(c(-10:10),c(-10:10)*2)
# Non-linear relationship
plot(c(-10:10), sin(c(-10:10)))
```





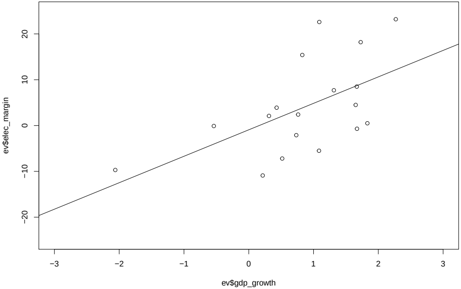
What are we Minimizing?

- We can think of OLS as generating fitted values of our Y variable that are perfect linear manipulations of our X variables.
- When we subtract the value of the fitted term from the value of the actual term, we get the residual, ie the difference between the best value of Y that OLS can predict using X and the actual value of Y for a given X.
- OLS mimizes this residual.

Mimizing the Residual, Example

- Remember when I said that social scientists look at causes and Data Scientists try to predict, sometimes Social Scientists try to predict too!
- However, when Social Scientists do predictions, they are usually trying to show that something is easy to predict because one thing is so causally significant on it.

```r
# Example, economic voting, the effect of gdp_growth on the incumbe
ev_fit <- lm(formula = elec_margin ~ gdp_growth, data = ev)
plot(ev$gdp_growth, ev$elec_margin, xlim = c(-3, 3), ylim = c(-25, 2
abline(ev_fit)
```



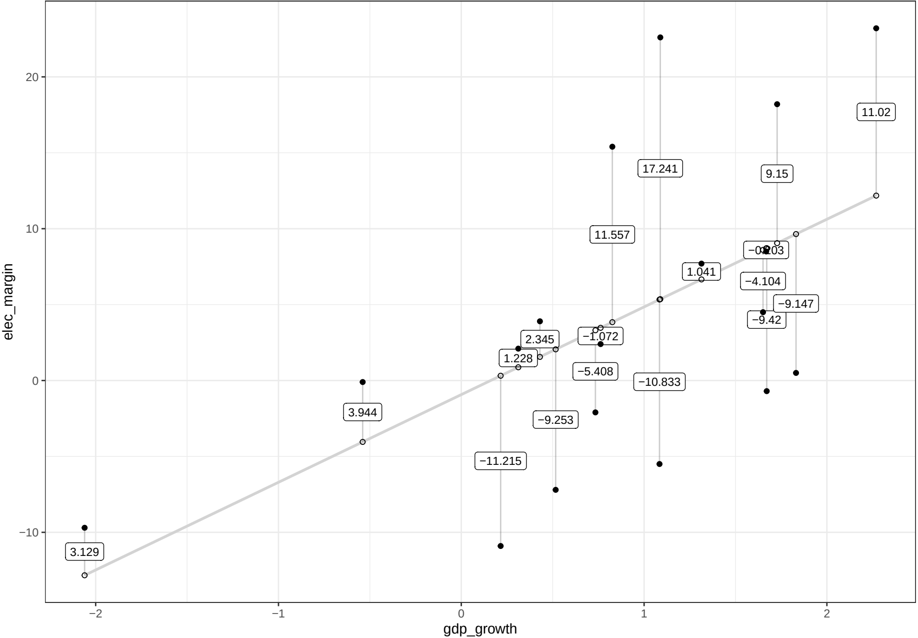
Minimizing the residual, example part 2

```
# 2020 prediction (gdp_growth = -8.9861168)

# alpha_hat + beta_hat * gdp_growth_2020

-0.9292 + 5.7754 * -8.9861168
```

```
## [1] -52.82762
```

```
## The actual value of elec_margin was around -1,
## that would be a HUGE residual for our model.
```

Minimizing the residual, example part 3



Exercises:

1. Load the dataset `cellphones` from Canvas, this is a dataset vehicle fatalities and cellphone use from 2012, when there was a lot of concern about texting and driving.
2. Articulate a hypothesis about number of vehicle deaths and cellphone adoption. Express that hypothesis in a dataset assignment like:

```r
cellphones$predicted_deaths <- cellphones$cell_subscription/10000
```

3. Examine how well your hypothesis predicts the actual values of Y by taking the sum of the square of the actual values - your predicted values .
4. Now run OLS on the same relationship, do the same thing with the residuals. Based on these values and the summary of the regression, how close were you?