Recitation 6: Introduction to Research Methods for Politics

Dept. of Politics, NYU

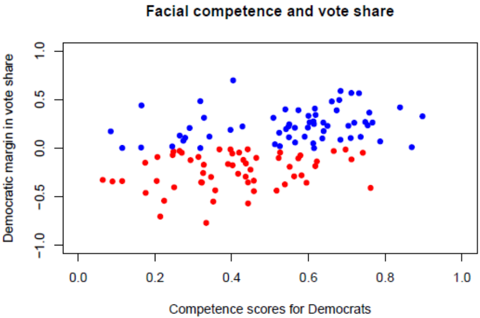
POL-850

Spring 2020

Linear Regression (QSS 4.2)

Modeling Relationships

We have seen two weeks ago how to depict a **bivariate relationship** using a scatter plot:



This gives us a sense of the **direction** of the relationship, but to summarize it into one measure we need a **statistical model**.

Using a line to predict[1]

- Prediction: for any value of X, what is the best guess about Y?
- Simplest way to relate two variables: a line
- Problem: for any line we draw, not all data is on the line
  - Some values will be above the line, some below
  - We need a way to account for **chance variation** away from the line

[1]from Matt Blackwell Gov 50: Lecture #11

The Linear Regression Model

The most intuitive model we can think of is a **linear** one:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where:

- Y_i is the outcome, or *dependent variable*.
- X_i is the predictor, or *independent variable*.
- α is an intercept, common to all units. (average value of Y when X is 0)
- β is the coefficient of our linear predictor, that tells **how X affects Y**.

Estimated coefficients[2]

- Parameters: α, β
  - Unknown features of the data-generating process
  - Chance error makes these impossible to observe directly
- Estimates: $\hat{\alpha}$, $\hat{\beta}$
  - An **estimate** is function of the data that is our best guess about some parameter
- **Regression line**: $\hat{Y} = \hat{\alpha} + \hat{\beta} * X$
  - Average value of Y when X is equal to x
  - Represents the best guess or **predicted value** of the outcome at x

[2] from Matt Blackwell Gov 50: Lecture #11

And what is ϵ_i?

It is our error term, i.e. the portion of the outcome that is left **unexplained** by the other components in the model. Think about this in terms of prediction:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

is the portion of Y_i that we manage to explain.

Then, on the other hand:

$$Y_i - \hat{Y}_i = \hat{\epsilon}_i$$

is the portion of Y_i that we do not manage to explain. $\hat{\epsilon}_i$ is the **residual**, the regression analogue of the error term ϵ_i,

Perceived Competence and Voting Behavior (1)

An experiment by Princeton researchers, asking people a within-a-second evaluation of unknown politicians' facial appearance:

Perceived Competence and Voting Behavior (2)

Name	Description
congress	session of congress
year	year of election
state	state of election
winner	name of winner
loser	name of runner-up
w.party	party of winner
l.party	party of loser
d.votes	number of votes for Democratic candidate
r.votes	number of votes for Republican candidate
d.comp	competence measure for Democratic candidate
r.comp	competence measure for Republican candidate

We are going to use linear regression to determine if and how much perceived competence (X) affects electoral performance (Y).

Ok, What Do We Do in Practice?

Note that Y_i and X_i are known, we have them recorded in our data. So our goal is to use what we have to get the remaining elements of the equation above: α, β, and ϵ_i. In *R*, we use `lm()`:

```
## lm(formula = diff.share ~ d.comp, data = face)
##
```

in our formula, we **only type variables**: the outcome to the left of the $\sim$ sign, and all the explanatory variables to the right.

Understanding Linear Model Results in *R*

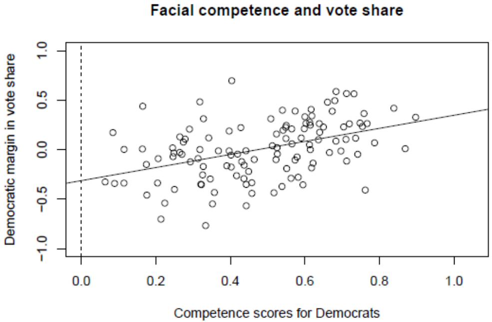
After running `lm()`, we get the following output:

```
##
## Coefficients:
## (Intercept)       d.comp
##     -0.3122       0.6604
```

which displays the two objects we were after: `Intercept` is $\hat{\alpha}$, our estimate of α, while `d.comp` is $\hat{\beta}$, the **coefficient** that measures the effect of the explanatory variable on the outcome.

Visualizing Regression (1)

Linear regression means fitting the **best possible straight line** based on our cloud of points:



Where "best" means it minimizes the cumulative distance between the points in the cloud and the line itself.

Behind Linear Regression: SSR

Therefore, our line is the one that minimizes the Sum of Squared Residuals (SSR):

$$SSR = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

Hence, every time we run `lm()` in R, we are simply asking it to find $\hat{\alpha}$ and $\hat{\beta}$ that makes SSR as small as possible.

Behind Linear Regression: RMSE

However, SSR is a bit hard to interpret. A nice alternative is to transform it to compute the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}SSR} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2}$$

Which represents the **average magnitude of the prediction error** for our regression model, that can be easily interpreted by referring to the outcome's scale.

RMSE in *R*

After running the regression via `lm()` we can easily compute the RMSE of our model with the following two steps:

```
epsilon.hat <- resid(fit)   # residuals
sqrt(mean(epsilon.hat^2))   # RMSE
```

```
## [1] 0.2642361
```

To understand whether this quantity is big (bad) or small (good), we need to know what is the **scale of our outcome variable**, Y. For instance, a RMSE of 0.264 could be pretty good if our outcome is vote share on a $1-100$ scale, but pretty bad if it is vote share on a $0-1$ scale!

Visualizing Regression (2)

We can finally trace back all the elements from the regression equation into our plot: