Gov 51: Linear Regression with Multiple Predictors

Matthew Blackwell

Harvard University

Loading the midterms data

```
midterms <- read.csv("data/midterms.csv")
head(midterms)
```

```
##   year  president party approval seat.change
## 1 1946     Truman     D       33         -55
## 2 1950     Truman     D       39         -29
## 3 1954 Eisenhower     R       61          -4
## 4 1958 Eisenhower     R       57         -47
## 5 1962    Kennedy     D       61          -4
## 6 1966    Johnson     D       44         -47
##   rdi.change
## 1         NA
## 2        8.2
## 3        1.0
## 4        1.1
## 5        5.0
## 6        5.3
```

Fitting the approval model

```
fit.app <- lm(seat.change ~ approval, data = midterms)
fit.app
```

```
##
## Call:
## lm(formula = seat.change ~ approval, data = midterms)
##
## Coefficients:
## (Intercept)     approval
##      -96.84         1.42
```

Fitting the income model

```r
fit.rdi <- lm(seat.change ~ rdi.change, data = midterms)
fit.rdi
```

```
##
## Call:
## lm(formula = seat.change ~ rdi.change, data = midterms)
##
## Coefficients:
## (Intercept)   rdi.change
##       -27.4          1.0
```

Multiple predictors

- What if we want to predict Y as a function of many variables?

$$\texttt{seat.change}_i = \alpha + \beta_1 \texttt{approval}_i + \beta_2 \texttt{rdi.change}_i + \epsilon_i$$

- Better predictions (at least in-sample).
- Better interpretation as **ceteris paribus** relationships:
  - β_1 is the relationship between `approval` and `seat.change` holding `rdi.change` constant.

Multiple regression in R

```r
mult.fit <- lm(seat.change ~ approval + rdi.change,
               data = midterms)
mult.fit
```

```
##
## Call:
## lm(formula = seat.change ~ approval + rdi.change, data = midterms)
##
## Coefficients:
## (Intercept)     approval   rdi.change
##     -120.44         1.57         3.33
```

- $\widehat{\alpha} =$ -120.4: average seat change president has 0% approval and no change in income levels.
- $\widehat{\beta}_1 =$ 1.57: average increase in seat change for additional percentage point of approval, **holding RDI change fixed**
- $\widehat{\beta}_2 =$ 3.334: average increase in seat change for each additional percentage point increase of RDI, **holding approval fixed**

Least squares with multiple regression

- How do we estimate the coefficients?
- The same exact way as before: minimize prediction error!
- Residuals (aka prediction error) with multiple predictors:

$$\widehat{\epsilon}_i = \texttt{seat.change}_i - \widehat{\alpha} - \widehat{\beta}_1 \texttt{approval}_i - \widehat{\beta}_2 \texttt{rdi.change}_i$$

- Find the coefficients that minimizes the **sum of the squared residuals**:

$$\text{SSR} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = (Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2})^2$$

Model fit with multiple predictors

- R^2 mechanically increases when you add a variables to the regression.
    - But this could be overfitting!!
- Solution: penalize regression models with more variables.
    - Occam's razor: **simpler models are preferred**
- Adjusted R^2: lowers regular R^2 for each additional covariate.
    - If the added covariates doesn't help predict, adjusted R^2 goes down!

Comparing model fits

```r
summary(fit.app)$r.squared
```

```
## [1] 0.431
```

```r
summary(mult.fit)$r.squared
```

```
## [1] 0.445
```

```r
summary(mult.fit)$adj.r.squared
```

```
## [1] 0.366
```