PSC4375: Linear Regression Model Fit

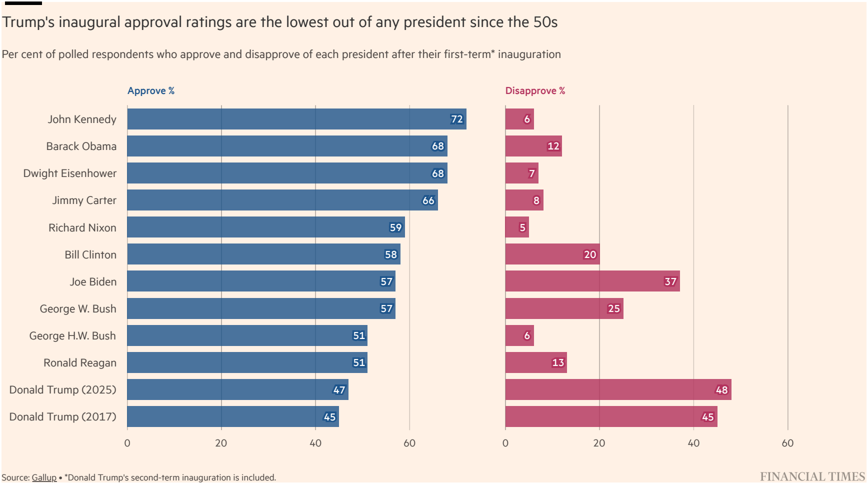
Week 6: Lectures 12 & 13

Prof. Weldzius

Villanova University

Slides Updated: 2025-02-26

Presidential Popularity and the Midterms



Presidential Popularity and the Midterms

- Does popularity of the president or recent changes in the economy better predict midterm election outcomes?

Name	Description
`year`	midterm election year
`president`	name of president
`party`	Democrat or Republican
`approval`	Gallup approval rating at midterms
`rdi.change`	change in real disposable income over the year before midterms
`seat.change`	change in the number of House seats for the president's party

Loading the data:

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

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

Fitting the Approval Model

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

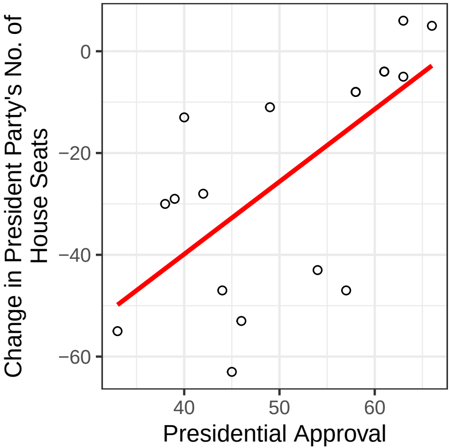
```
##
## Call:
## lm(formula = seat.change ~ approval, data = midterms)
##
## Coefficients:
## (Intercept)     approval
##     -96.845        1.424
```

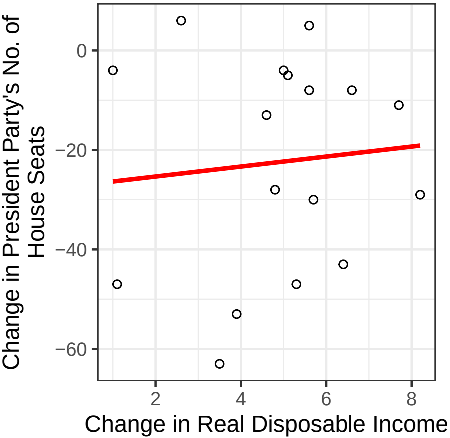
Fitting the Income Model

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

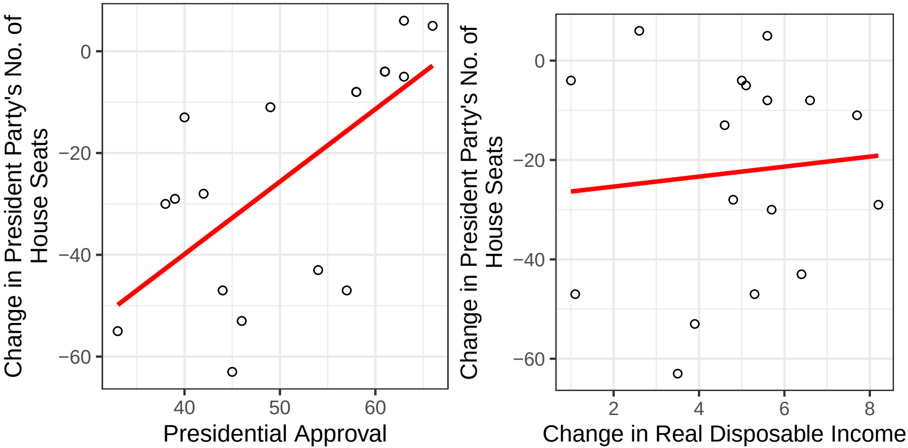
```
##
## Call:
## lm(formula = seat.change ~ rdi.change, data = midterms)
##
## Coefficients:
## (Intercept)   rdi.change
##     -27.354        1.004
```

Comparing Models



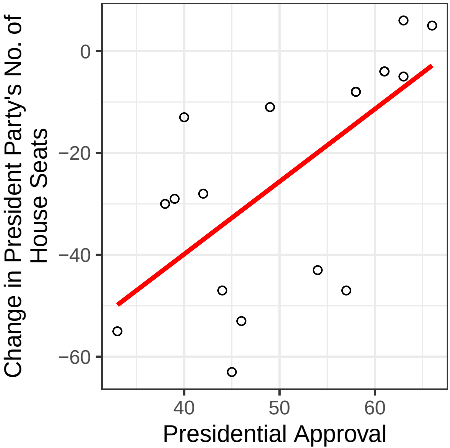


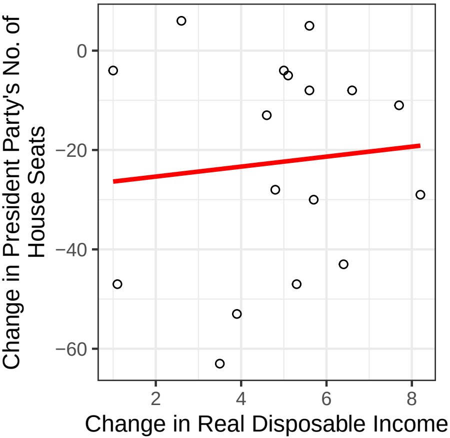
Comparing Models



- How well do the models "fit the data"?

Comparing Models





- How well do the models “fit the data”?
    - How well does the model predict the outcome variable in the data?

Model Fit

Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.

Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.
  - Measure of the **proportional reduction in error** by the model.

Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.
  - Measure of the **proportional reduction in error** by the model.
- Prediction error compared to what?

Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.
  - Measure of the **proportional reduction in error** by the model.
- Prediction error compared to what?
  - Baseline prediction error: **Total sum of squares**
    - $\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.
  - Measure of the **proportional reduction in error** by the model.
- Prediction error compared to what?
  - Baseline prediction error: **Total sum of squares**
    - $\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Model prediction error: **Sum of squared residuals**
    - $\text{SSR} = \sum_{i=1}^{n} \epsilon_i^2$

Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.
  - Measure of the **proportional reduction in error** by the model.
- Prediction error compared to what?
  - Baseline prediction error: **Total sum of squares**
    - $\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Model prediction error: **Sum of squared residuals**
    - $\text{SSR} = \sum_{i=1}^{n} \epsilon_i^2$
  - TSS - SSR: reduction in prediction error by the model.

Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.
  - Measure of the **proportional reduction in error** by the model.
- Prediction error compared to what?
  - Baseline prediction error: **Total sum of squares**
    - $\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Model prediction error: **Sum of squared residuals**
    - $\text{SSR} = \sum_{i=1}^{n} \epsilon_i^2$
  - TSS - SSR: reduction in prediction error by the model.
- R^2 is this reduction in error divided by the baseline error:

$$R^2 = \frac{\text{TSS} - \text{SSR}}{\text{TSS}}$$

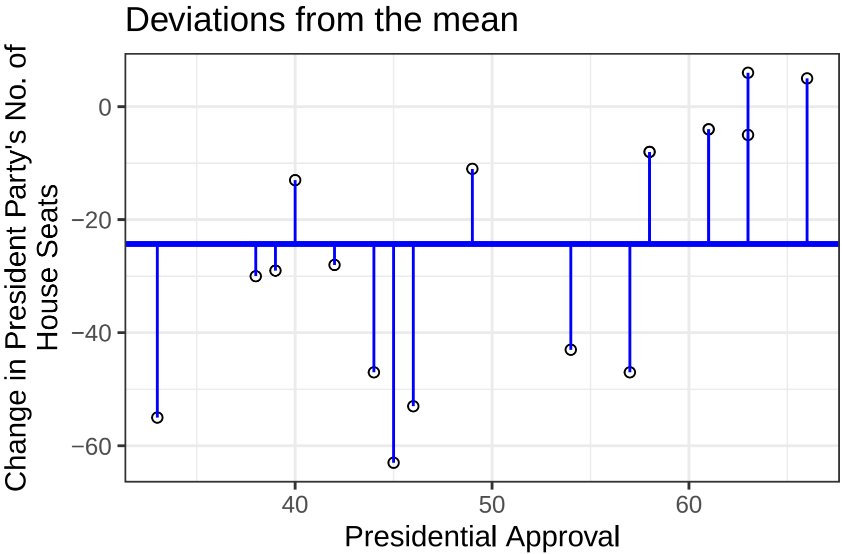
Model Fit

- One number summary of model fit: R^2 or **coefficient of determination**.
  - Measure of the **proportional reduction in error** by the model.
- Prediction error compared to what?
  - Baseline prediction error: **Total sum of squares**
    - $\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Model prediction error: **Sum of squared residuals**
    - $\text{SSR} = \sum_{i=1}^{n} \epsilon_i^2$
  - TSS - SSR: reduction in prediction error by the model.
- R^2 is this reduction in error divided by the baseline error:

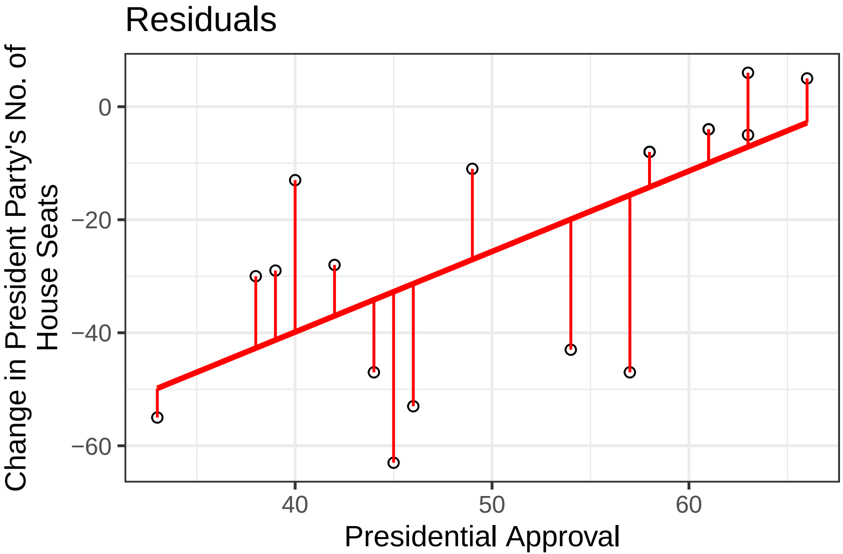
$$R^2 = \frac{\text{TSS} - \text{SSR}}{\text{TSS}}$$

- Roughly: proportion of the variation in Y_i "explained by" X_i

Total sum of squares vs. Sum of squared residuals



Total sum of squares vs. Sum of squared residuals



Model Fit in R

- To access R^2 from the `lm()` output, use the `summary()` function:

```
fit.app.sum <- summary(fit.app)
fit.app.sum$r.squared
```

```
## [1] 0.4307133
```

Model Fit in R

- To access R^2 from the `lm()` output, use the `summary()` function:

```
fit.app.sum <- summary(fit.app)
fit.app.sum$r.squared
```

```
## [1] 0.4307133
```

- Compare to fit using change in income:

```
fit.rdi.sum <- summary(fit.rdi)
fit.rdi.sum$r.squared
```

```
## [1] 0.008529029
```

Model Fit in R

- To access R^2 from the `lm()` output, use the `summary()` function:

```
fit.app.sum <- summary(fit.app)
fit.app.sum$r.squared
```

```
## [1] 0.4307133
```

- Compare to fit using change in income:

```
fit.rdi.sum <- summary(fit.rdi)
fit.rdi.sum$r.squared
```

```
## [1] 0.008529029
```

- Which does a better job predicting midterm election outcomes?

Fake data, better fit

Fake data, better fit

- Little hard to see what's happening in that example.

Fake data, better fit

- Little hard to see what's happening in that example.
- Let's look at fake variables x and y:

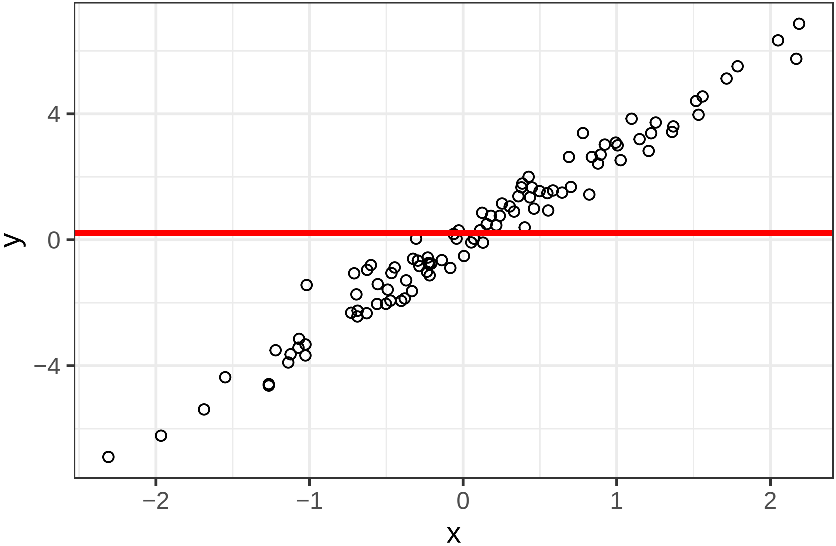
Fake data, better fit

- Little hard to see what's happening in that example.
- Let's look at fake variables x and y:

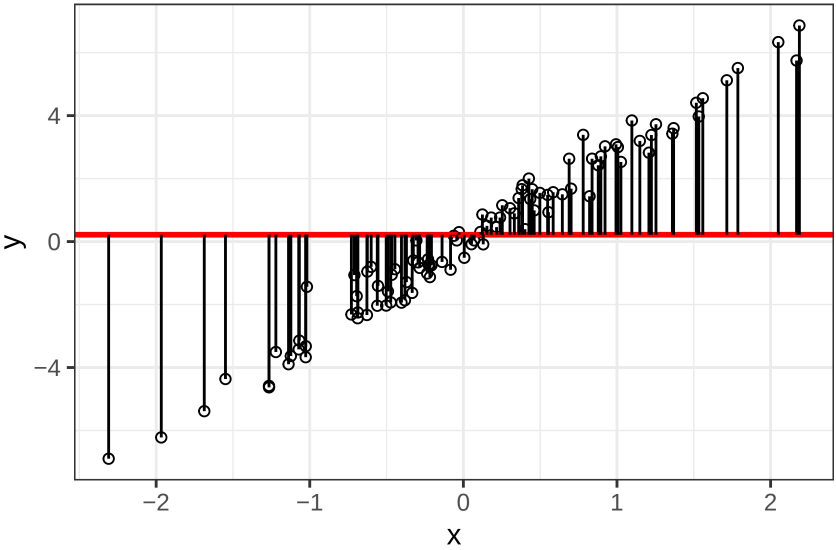
```
fit.x <- lm(y ~ x)
```

- Very good model fit: $R^2 \approx 0.95$

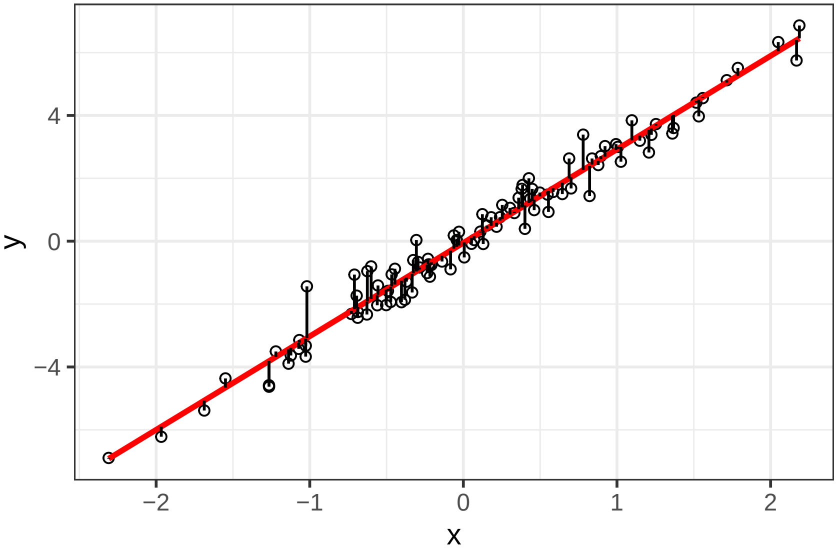
Fake data, better fit



Fake data, better fit

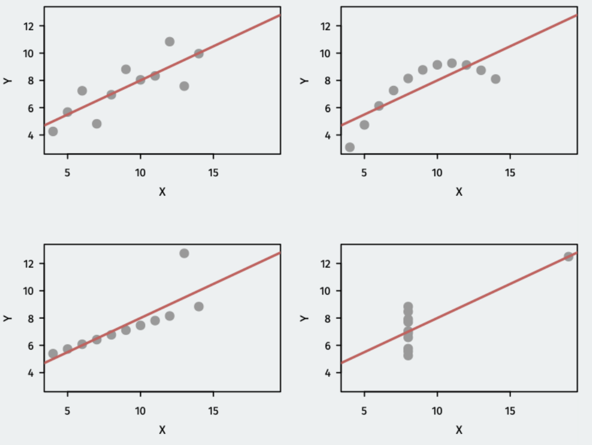


Fake data, better fit



Is R-squared useful?

- Can be very misleading. Each of these samples have the same R^2 even though they are vastly different:



Overfitting

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is a measure of in-sample fit.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
    - R^2 is a measure of in-sample fit.
- **Out-of-sample fit**: how well your model predicts new data.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is a measure of in-sample fit.
- **Out-of-sample fit**: how well your model predicts new data.
- **Overfitting**: OLS optimizes in-sample fit; may do poorly out of sample.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is a measure of in-sample fit.
- **Out-of-sample fit**: how well your model predicts new data.
- **Overfitting**: OLS optimizes in-sample fit; may do poorly out of sample.
  - Example: predicting winner of Democratic presidential primary with gender of the candidate.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is a measure of in-sample fit.
- **Out-of-sample fit**: how well your model predicts new data.
- **Overfitting**: OLS optimizes in-sample fit; may do poorly out of sample.
  - Example: predicting winner of Democratic presidential primary with gender of the candidate.
  - Until 2016, gender was a **perfect** predictor of who wins the primary.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is a measure of in-sample fit.
- **Out-of-sample fit**: how well your model predicts new data.
- **Overfitting**: OLS optimizes in-sample fit; may do poorly out of sample.
  - Example: predicting winner of Democratic presidential primary with gender of the candidate.
  - Until 2016, gender was a **perfect** predictor of who wins the primary.
  - Prediction for 2016 based on this: Bernie Sanders as Dem. nominee.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is a measure of in-sample fit.
- **Out-of-sample fit**: how well your model predicts new data.
- **Overfitting**: OLS optimizes in-sample fit; may do poorly out of sample.
  - Example: predicting winner of Democratic presidential primary with gender of the candidate.
  - Until 2016, gender was a **perfect** predictor of who wins the primary.
  - Prediction for 2016 based on this: Bernie Sanders as Dem. nominee.
  - Bad out-of-sample prediction due to overfitting!

Multiple Predictors (Multivariate Regression)

Multiple Predictors (Multivariate Regression)

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

Multiple Predictors (Multivariate Regression)

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

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

Multiple Predictors (Multivariate Regression)

- 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).

Multiple Predictors (Multivariate Regression)

- 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:

Multiple Predictors (Multivariate Regression)

- 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

Multiple regression in 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
##    -120.436        1.572        3.334
```

Multiple regression in 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
##    -120.436        1.572        3.334
```

- $\hat{\alpha} = -120.4$: average seat change president has 0% approval and no change in income levels.

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
##    -120.436        1.572        3.334
```

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

Multiple regression in 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
##    -120.436        1.572        3.334
```

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

Least squares with multiple regression

Least squares with multiple regression

- How do we estimate the coefficients?

Least squares with multiple regression

- How do we estimate the coefficients?
- The same exact way as before: minimize prediction error!

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:

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:

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

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:

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

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

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:

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

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

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

Model fit with multiple predictors

Model fit with multiple predictors

- R^2 mechanically increases when you add a variables to the regression.

Model fit with multiple predictors

- R^2 mechanically increases when you add a variables to the regression.
  - But this could be overfitting!!

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.

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**

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.

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 don't help predict, adjusted R^2 goes down!

Comparing Model Fits

Comparing Model Fits

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

```
## [1] 0.4307133
```

Comparing Model Fits

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

```
## [1] 0.4307133
```

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

```
## [1] 0.4448387
```

Comparing Model Fits

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

```
## [1] 0.4307133
```

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

```
## [1] 0.4448387
```

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

```
## [1] 0.3655299
```

Binary and Categorical Predictors

Binary and Categorical Predictors

- Political effects of government programs

Binary and Categorical Predictors

- Political effects of government programs
  - *Progresa*: Mexican conditional cash transfer program (CCT) from ~2000

Binary and Categorical Predictors

- Political effects of government programs
  - *Progresa*: Mexican conditional cash transfer program (CCT) from $\sim$2000
    - Welfare $ given if kids enrolled in schools, get regular check-ups, etc.

Binary and Categorical Predictors

- Political effects of government programs
  - *Progresa*: Mexican conditional cash transfer program (CCT) from $\sim$2000
    - Welfare $ given if kids enrolled in schools, get regular check-ups, etc.
  - Do these programs have political effects?

Binary and Categorical Predictors

- Political effects of government programs
  - *Progresa*: Mexican conditional cash transfer program (CCT) from ~2000
    - Welfare $ given if kids enrolled in schools, get regular check-ups, etc.
  - Do these programs have political effects?
    - Program had support from most parties.

Binary and Categorical Predictors

- Political effects of government programs
  - *Progresa*: Mexican conditional cash transfer program (CCT) from $\sim$2000
    - Welfare $ given if kids enrolled in schools, get regular check-ups, etc.
  - Do these programs have political effects?
    - Program had support from most parties.
    - Was implemented in a nonpartisan fashion.

Binary and Categorical Predictors

- Political effects of government programs
  - *Progresa*: Mexican conditional cash transfer program (CCT) from ~2000
    - Welfare $ given if kids enrolled in schools, get regular check-ups, etc.
  - Do these programs have political effects?
    - Program had support from most parties.
    - Was implemented in a nonpartisan fashion.
    - Would the incumbent presidential party be rewarded?

The Data

- Randomized roll-out of the CCT program:
  - treatment: receive CCT 21 months before 2000 election
  - control: receive CCT 6 months before 2000 election
  - Does having CCT longer mobilize voters for incumbent PRI party?

Name	Description
`treatment`	early Progresa (1) or late Progresa (0)
`pri2000s`	PRI votes in the 2000 election as a share of adults in precinct
`t2000`	turnout in the 2000 election as share of adults in precinct

```r
cct <- read.csv("../data/progresa.csv")
```

Difference in Means Estimates

- Does CCT affect turnout?

Difference in Means Estimates

- Does CCT affect turnout?

```
cct.turn.ate <- cct %>% group_by(treatment) %>%
  summarize(t2000_mean = mean(t2000)) %>%
  pivot_wider(names_from = treatment, values_from = t2000_mean) %>%
  mutate(turnout_ate = `1` - `0`)
cct.turn.ate$turnout_ate
```

```
## [1] 4.269676
```

Difference in Means Estimates

- Does CCT affect turnout?

```
cct.turn.ate <- cct %>% group_by(treatment) %>%
  summarize(t2000_mean = mean(t2000)) %>%
  pivot_wider(names_from = treatment, values_from = t2000_mean) %>%
  mutate(turnout_ate = `1` - `0`)
cct.turn.ate$turnout_ate
```

```
## [1] 4.269676
```

- Does CCT affect PRI (incumbent) votes?

Difference in Means Estimates

- Does CCT affect turnout?

```
cct.turn.ate <- cct %>% group_by(treatment) %>%
  summarize(t2000_mean = mean(t2000)) %>%
  pivot_wider(names_from = treatment, values_from = t2000_mean) %>%
  mutate(turnout_ate = `1` - `0`)
cct.turn.ate$turnout_ate
```

```
## [1] 4.269676
```

- Does CCT affect PRI (incumbent) votes?

```
cct.pri.ate <- cct %>% group_by(treatment) %>%
  summarize(pri2000s_mean = mean(pri2000s)) %>%
  pivot_wider(names_from = treatment, values_from = pri2000s_mean) %>%
  mutate(pri_ate = `1` - `0`)
cct.pri.ate$pri_ate
```

```
## [1] 3.622496
```

Binary independent variables

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

Binary independent variables

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

- When independent variable X_i is binary:

Binary independent variables

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

- When independent variable X_i is binary:
- Intercept α is the average outcome in the $X = 0$ group.

Binary independent variables

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

- When independent variable X_i is binary:
- Intercept α is the average outcome in the $X = 0$ group.
- Slope β is the difference-in-means of Y between $X = 1$ group and $X = 0$ group.

Binary independent variables

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

- When independent variable X_i is binary:
- Intercept α is the average outcome in the $X = 0$ group.
- Slope β is the difference-in-means of Y between $X = 1$ group and $X = 0$ group.

$$\hat{\beta} = \bar{Y}_{treated} - \bar{Y}_{control}$$

Binary independent variables

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

- When independent variable X_i is binary:
- Intercept α is the average outcome in the $X = 0$ group.
- Slope β is the difference-in-means of Y between $X = 1$ group and $X = 0$ group.

$$\hat{\beta} = \bar{Y}_{treated} - \bar{Y}_{control}$$

- If there are other independent variables, this becomes the difference-in-means controlling for those covariates.

Linear regression for experiments

Linear regression for experiments

- Under **randomization**, we can estimate the ATE with regression:

Linear regression for experiments

- Under **randomization**, we can estimate the ATE with regression:

```
cct.pri.ate <- cct %>%
  group_by(treatment) %>%
  summarize(pri2000s_mean = mean(pri2000s)) %>%
  pivot_wider(names_from = treatment, values_from = pri2000s_mean) %>%
  mutate(pri_ate = `1` - `0`)
cct.pri.ate$pri_ate
```

```
## [1] 3.622496
```

Linear regression for experiments

- Under **randomization**, we can estimate the ATE with regression:

```
cct.pri.ate <- cct %>%
  group_by(treatment) %>%
  summarize(pri2000s_mean = mean(pri2000s)) %>%
  pivot_wider(names_from = treatment, values_from = pri2000s_mean) %>%
  mutate(pri_ate = `1` - `0`)
cct.pri.ate$pri_ate
```

```
## [1] 3.622496
```

```
lm(pri2000s ~ treatment, data = cct)
```

```
##
## Call:
## lm(formula = pri2000s ~ treatment, data = cct)
##
## Coefficients:
## (Intercept)    treatment
##      34.489        3.622
```

Categorical variables in regression

Categorical variables in regression

- We often have **categorical variables**:

Categorical variables in regression

- We often have **categorical variables**:
  - Race/ethnicity: white, Black, Latino, Asian.

Categorical variables in regression

- We often have **categorical variables**:
    - Race/ethnicity: white, Black, Latino, Asian.
    - Partisanship: Democrat, Republican, Independent

Categorical variables in regression

- We often have **categorical variables**:
  - Race/ethnicity: white, Black, Latino, Asian.
  - Partisanship: Democrat, Republican, Independent
  - Strategy for including in a regression: create a **series of binary variables**

Categorical variables in regression

- We often have **categorical variables**:
  - Race/ethnicity: white, Black, Latino, Asian.
  - Partisanship: Democrat, Republican, Independent
  - Strategy for including in a regression: create a **series of binary variables**

Unit	Party	Democrat	Republican	Independent
1	Democrat	1	0	0
2	Democrat	1	0	0
3	Independent	0	0	1
4	Republican	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Categorical variables in regression

- We often have **categorical variables**:
  - Race/ethnicity: white, Black, Latino, Asian.
  - Partisanship: Democrat, Republican, Independent
  - Strategy for including in a regression: create a **series of binary variables**

Unit	Party	Democrat	Republican	Independent
1	Democrat	1	0	0
2	Democrat	1	0	0
3	Independent	0	0	1
4	Republican	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Then include all but one of these binary variables:

$$turnout_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \epsilon_i$$

Interpreting categorical variables

Interpreting categorical variables

$$turnout_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \epsilon_i$$

Interpreting categorical variables

$$turnout_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \epsilon_i$$

- $\hat{\alpha}$: average outcome in the **omitted group/baseline** (Democrats).

Interpreting categorical variables

$$turnout_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \epsilon_i$$

- $\hat{\alpha}$: average outcome in the **omitted group/baseline** (Democrats).
- $\hat{\beta}$ coefficients: average difference between each group and the baseline.

Interpreting categorical variables

$$turnout_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \epsilon_i$$

- $\hat{\alpha}$: average outcome in the **omitted group/baseline** (Democrats).
- $\hat{\beta}$ coefficients: average difference between each group and the baseline.
  - $\hat{\beta}_1$: average difference in turnout between Republicans and Democrats

Interpreting categorical variables

$$turnout_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \epsilon_i$$

- $\hat{\alpha}$: average outcome in the **omitted group/baseline** (Democrats).
- $\hat{\beta}$ coefficients: average difference between each group and the baseline.
  - $\hat{\beta}_1$: average difference in turnout between Republicans and Democrats
  - $\hat{\beta}_2$: average difference in turnout between Independents and Democrats