Gov 51: Linear Regression Model Fit

Matthew Blackwell

Harvard University

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

```
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
```

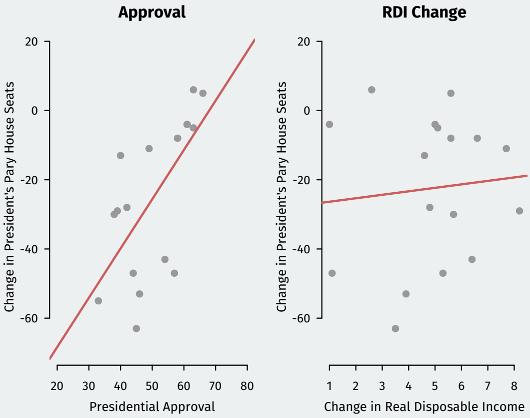
```
##
## Call:
## lm(formula = seat.change ~ approval, data = midterms)
##
## Coefficients:
## (Intercept)     approval
##      -96.84         1.42
```

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
```

Comparing models



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

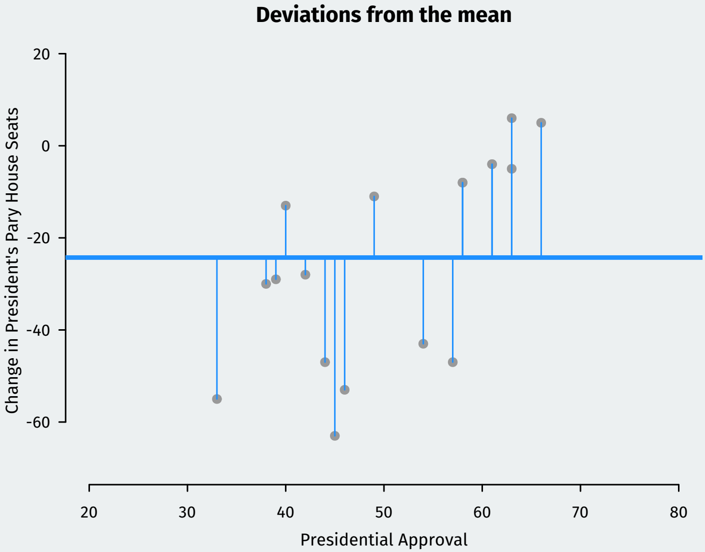
Model fit

- 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** TSS $= \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
    - Model prediction error: **Sum of squared residuals** SSR $= \sum_{i=1}^{n} \hat{\epsilon}_i^2$
    - TSS − SSR: reduction in prediction error by the model.
- R^2 is this reduction in error divided by the baseline error:

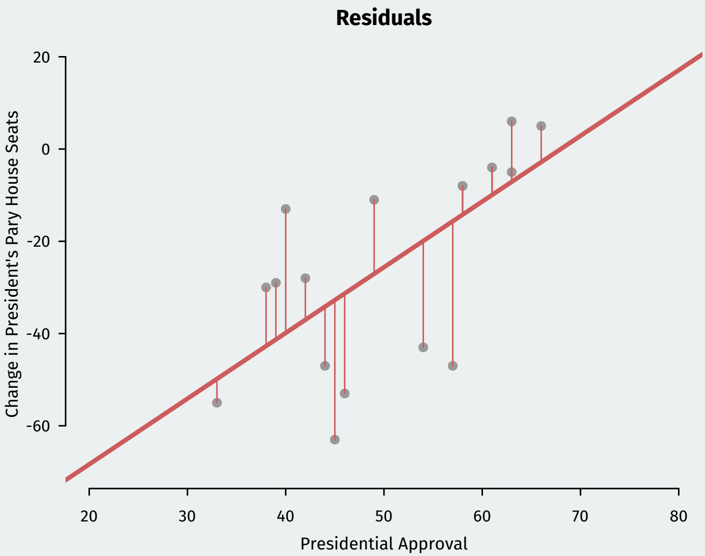
$$R^2 = \frac{TSS - SSR}{TSS}$$

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

Total SS vs SSR



Total SS vs SSR



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] 0.431
```

- Compare to the fit using change in income:

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

```
## [1] 0.00853
```

- Which does a better job predicting midterm election outcomes?

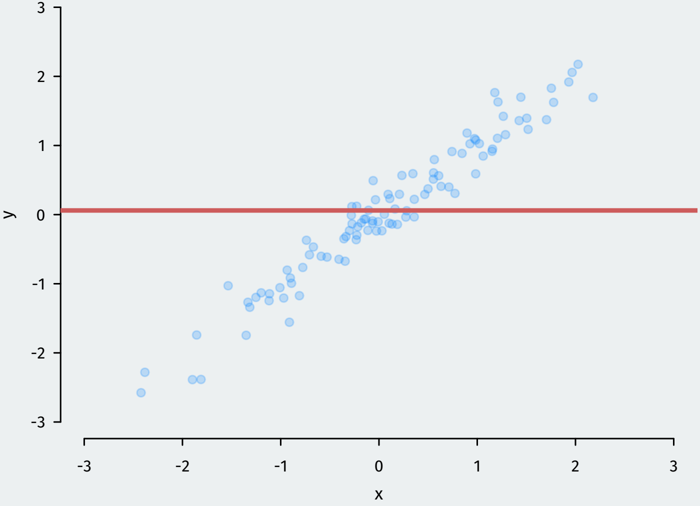
Fake data, better fit

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

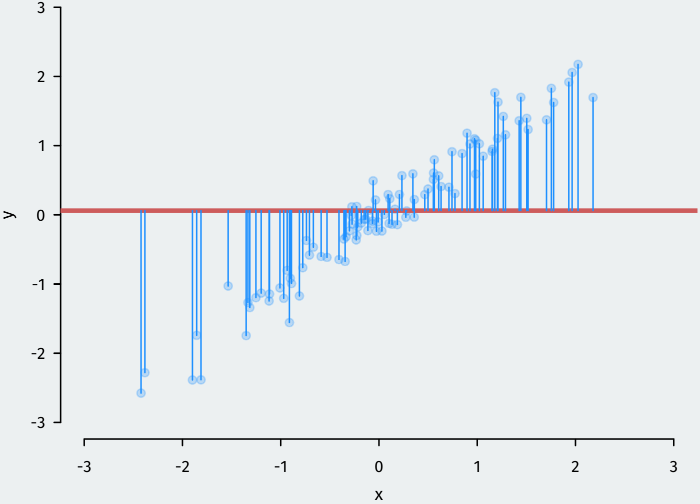
```
fit.x <- lm(y ~ x)
```

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

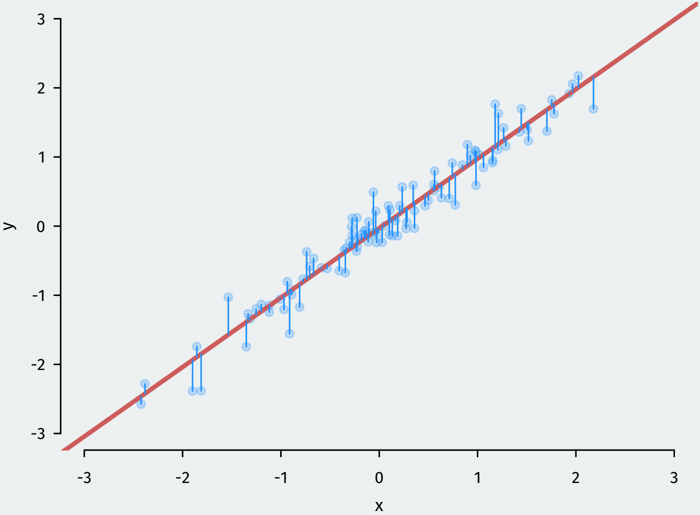
Fake data, better fit



Fake data, better fit

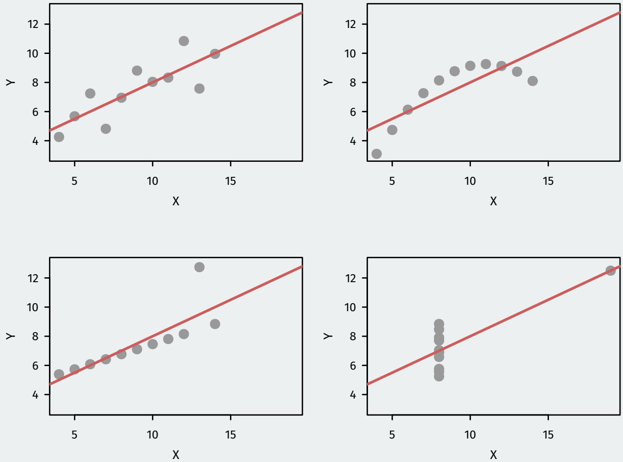


Fake data, better fit



Is R-squared useful?

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



Overfitting

- **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!