Recitation 7: Introduction to Research Methods for Politics

Dept. of Politics, NYU

POL-850

Spring 2020

Model Fit
(QSS 4.2.6)

Model Fit

- We now know how to estimate the linear regression model using `lm()` in R

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

- We can also check whether independent variable is a good predictor of dependent variable

R Squared

The proportion of the variance for a dependent variable that's explained by an independent variable in the linear regression model

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

- Sum of Squared Residuals (SSR)

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

- Total Sum of Squares (TSS)

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Model Fit

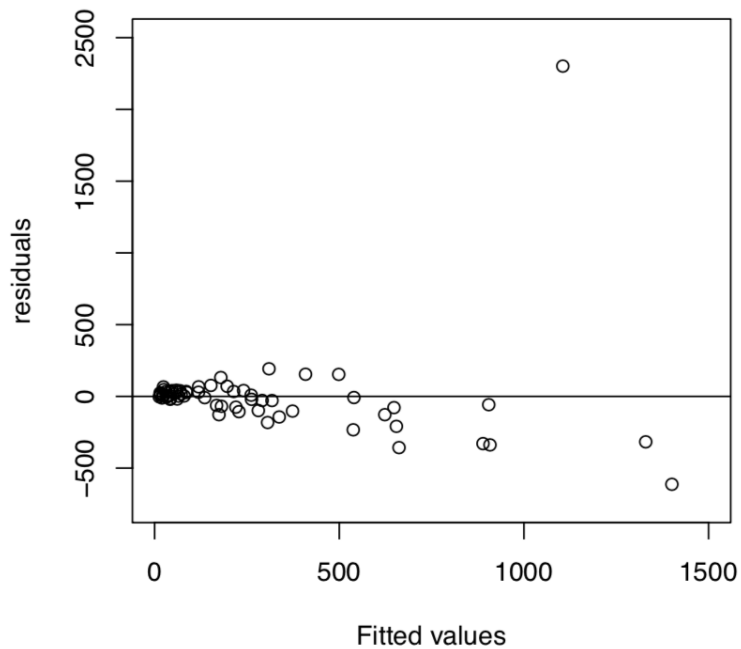
We examine the relationship between Perot's 1996 vote share and Buchanan's 2000 vote share

```
florida <- read.csv("data/florida.csv")
fit2 <- lm(Buchanan00 ~ Perot96, data = florida)
fit2

##
## Call:
## lm(formula = Buchanan00 ~ Perot96, data = florida)
##
## Coefficients:
## (Intercept)      Perot96
##      1.3458       0.0359

TSS2 <- sum((florida$Buchanan00 - mean(florida$Buchanan00))^2)
SSR2 <- sum(resid(fit2)^2)
## R^2 or coefficient of determination
(TSS2 - SSR2) / TSS2

## [1] 0.513
```

Model Fit

When plotting the model fit, we observe one outlier

```
plot(fitted(fit2), resid(fit2), xlim = c(0, 1500),
     ylim = c(-750, 2500), xlab = "Fitted values",
     ylab = "residuals")
abline(h = 0)
```



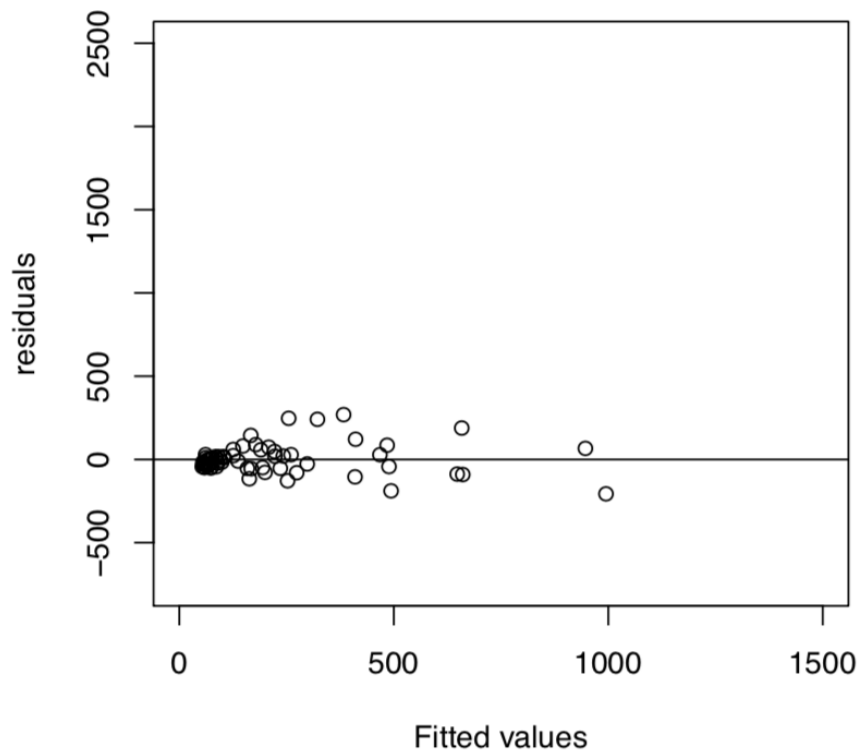
Model Fit

We run linear regression model without the outlier and see whether R^2 increases

```
florida.pb <- subset(florida, subset = (county != "PalmBeach"))
fit3 <- lm(Buchanan00 ~ Perot96, data = florida.pb)
fit3

##
## Call:
## lm(formula = Buchanan00 ~ Perot96, data = florida.pb)
##
## Coefficients:
## (Intercept)      Perot96
##     45.8419       0.0244

## built-in R function
summary(fit3)$r.squared

## [1] 0.851
```

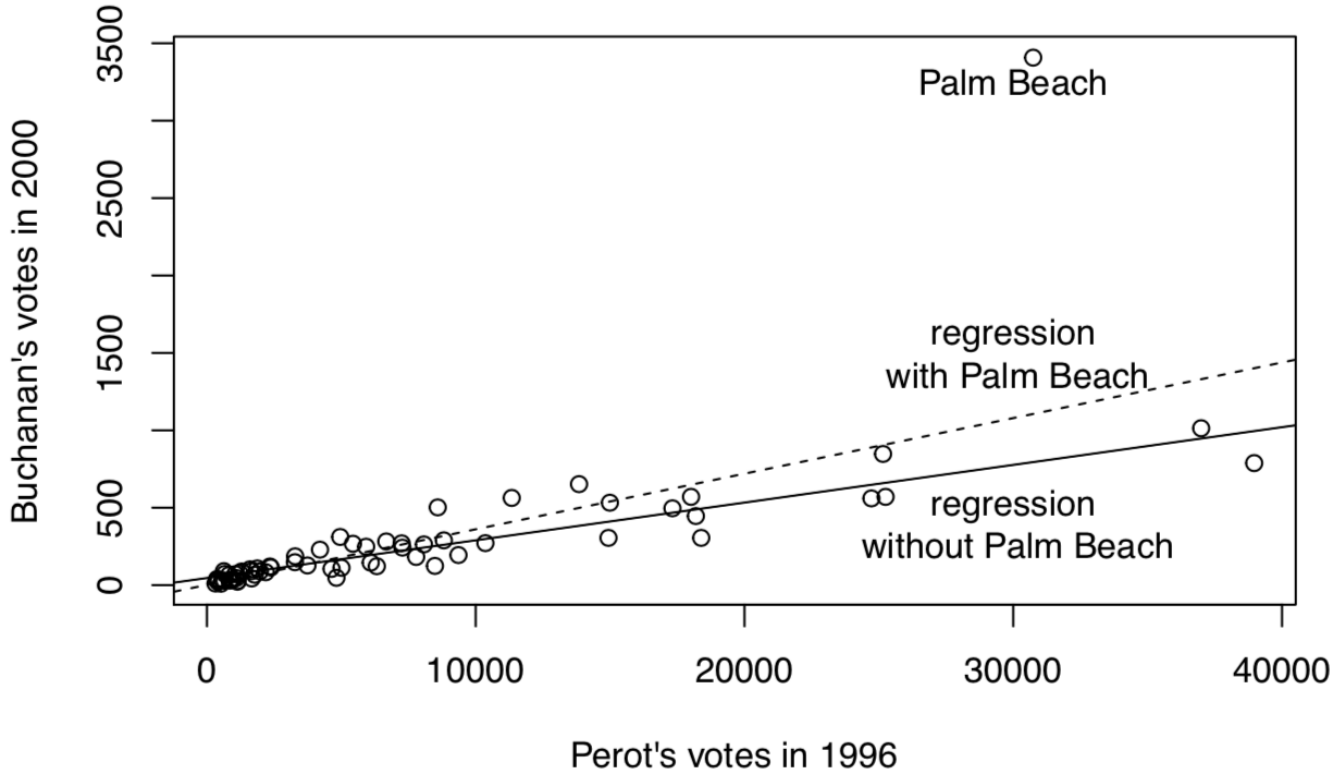
Model Fit

```
plot(fitted(fit3), resid(fit3), xlim = c(0, 1500),
     ylim = c(-750, 2500), xlab = "Fitted values",
     ylab = "residuals")
abline(h = 0)
```



Model Fit

Having outliers can change estimates of your regression model



How should researchers deal with outliers?

Regression with Multiple Predictors (QSS 4.3.2)

Model Fit

How does multiple linear regression differ from bivariate regression? Not much:

- The model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Sum of squared residuals (SSR):

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

Social Pressure Experiment

- August 2006 Primary Statewide Election in Michigan
- Send postcards with different (randomly assigned) messages
  1. no message (control group)
  2. civic duty message
  3. "you are being studied" message (Hawthorne effect)
  4. neighborhood social pressure message

Name	Description
`hhsize`	household size of voter
`messages`	GOTV messages voter received (`Civic`, `Control`, `Neighbors`, `Hawthorne`)
`sex`	sex of voter (`female` or `male`)
`yearofbirth`	year of birth of voter
`primary2004`	whether a voter turned out in the 2004 Primary election (`1`=voted, `0`=abstained)
`primary2006`	whether a voter turned out in the 2006 Primary election (`1`=voted, `0`=abstained)

Social Pressure Experiment

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____

Social Pressure Experiment

```
social <- read.csv("social.csv")
levels(social$messages) # base level is `Civic'
```

```
## [1] "Civic Duty" "Control"    "Hawthorne"  "Neighbors"
```

```
fit <- lm(primary2008 ~ messages, data = social)
fit
```

```
##
## Call:
## lm(formula = primary2008 ~ messages, data = social)
##
## Coefficients:
##       (Intercept)    messagesControl  messagesHawthorne
##          0.314538          -0.017899           0.007837
## messagesNeighbors
##          0.063411
```

Social Pressure Experiment

```
## ## create indicator variables
## social$Control <- ifelse(social$messages == "Control", 1, 0)
## social$Hawthorne <- ifelse(social$messages == "Hawthorne", 1, 0)
## social$Neighbors <- ifelse(social$messages == "Neighbors", 1, 0)
## ## fit the same regression as above by directly using indicator variables
## lm(primary2008 ~ Control + Hawthorne + Neighbors, data = social)

## create a data frame with unique values of `messages'
unique.messages <- data.frame(messages = unique(social$messages))
unique.messages
```

```
##     messages
## 1 Civic Duty
## 2  Hawthorne
## 3    Control
## 4  Neighbors
```