Political Science 209 - Fall 2018

Linear Regression

Florian Hollenbach

22nd October 2018

In-class Exercise Linear Regression

Please dowload *intrade08.csv* & *pres08.csv* from class website

- Read both data sets into *R*
- Create data summary for each data sets

Variables in the intrade data

- *day*: Date of the session
- *statename*: Full name of each state (including District of Columbia in 2008)
- *state*: Abbreviation of each state (including District of Columbia in 2008)
- *PriceD*: Closing price (predicted vote share) of Democratic Nominee's market
- *PriceR*: Closing price (predicted vote share) of Republican Nominee's market
- *VolumeD*: Total session trades of Democratic Party Nominee's market
- *VolumeR*: Total session trades of Republican Party Nominee's market

Variables in the pres08 data

- *state.name*: Full name of state (only in pres2008)
- *state*: Two letter state abbreviation
- *Obama*: Vote percentage for Obama
- *McCain*: Vote percentage for McCain
- *EV*: Number of electoral college votes for this state

Combining data sets

- First we have to combine the different data sets
- To do so, we need an identifier that tells *R* which observations to match to each other
- What could we use?

Combining data sets

- First we have to combine the different data sets
- To do so, we need an identifier that tells *R* which observations to match to each other
- What could we use?

state variable

Combining data sets

- Use *merge()* function

```
merge(x,y, by =)
intresults08 <- merge(intrade08, pres08, by = "state")
head(intresults08)
```

Question 1

Create a *DaysToElection* variable by subtracting the day of the election from each day in the dataset. Now create a *state margin of victory* variable to predict, and a *betting market margin* to predict it with.

election day in 2008: Nov, 4th

Solution 1

```
intresults08$DaysToElection
 <- as.Date("2008-11-04") - as.Date(intresults08$day)

intresults08$obama.intmarg <- intresults08$PriceD - intresults08$PriceR
intresults08$obama.actmarg <- intresults08$Obama - intresults08$McCain
```

Question 2

Considering only the trading one day from the election, predict the actual electoral margins from the trading margins using a linear model. Does it predict well? How would you visualize the predictions and the outcomes together? Hint: because we only have one predictor you can use *abline*.

Solution 2

```
latest08 <- intresults08[intresults08$DaysToElection == 1,]
int.fit08 <- lm(obama.actmarg ~ obama.intmarg, data = latest08)
coef(int.fit08)
summary(int.fit08)$r.squared
plot(latest08$obama.intmarg, latest08$obama.actmarg,
    xlab="Market's margin for Obama", ylab="Obama margin")
abline(int.fit08)
```

Question 3

What would be the prediction for the margin of victory if the InTrade margin was 25? Mark this point on the previous plot.

Solution 3

```
coef(int.fit08)[1] + coef(int.fit08)[2]*25

plot(latest08$obama.intmarg, latest08$obama.actmarg,
    xlab="Market's margin for Obama", ylab="Obama margin")
abline(int.fit08)
points(25,(coef(int.fit08)[1] + coef(int.fit08)[2]*25), col = "red")
```

Question 4

Even efficient markets aren't omniscient. Information comes in about the election every day and the market prices should reflect any change in information that seem to matter to the outcome.

We can examine how and about what the markets change their minds by looking at which states they are confident about, and which they update their 'opinions' (i.e. their prices) about. Over the period before the election, let's see how prices for each state are evolving. We can get a compact summary of price movement by fitting a linear model to Obama's margin for each state over the 20 days before the election.

We will summarise price movement by the direction (up or down) and rate of change (large or small) of price over time. This is basically also what people in finance do, but they get paid more...

Start by plotting Obama's margin in West Virginia against the number of days until the election and modeling the relationship with a linear model. Use the last 20 days. Show the model's predictions on each day and the data. What does this model's slope coefficient tells us about which direction the margin is changing and also how fast it is changing?

Solution 4

```
stnames <- unique(intresults08$state.name)
recent <- subset(intresults08, subset=(DaysToElection <= 20)
 & (state.name==stnames[1]))

recent.mod <- lm(obama.intmarg ~ DaysToElection, data=recent)
plot(recent$DaysToElection, recent$obama.intmarg,
xlab="Days to election", ylab="Market's Obama margin")
abline(recent.mod)
```

Question 5

Let's do the same thing for all states and collect the slope coefficients (β's). How can we modify the code from the answer to the previous question? Then plot the distribution of changes for all states.

Solution 5

```
stnames <- unique(intresults08$state.name)
change <- rep(NA, length(unique(intresults08$state.name)))
names(change) <- unique(intresults08$state.name)

for(i in 1: length(unique(intresults08$state.name))){
recent <- subset(intresults08, subset=(DaysToElection <= 20)
& (state.name==stnames[i]))

recent.mod <- lm(obama.intmarg ~ DaysToElection, data=recent)
change[i] <- coef(recent.mod)[2]
}
hist(change)
```

Questin 5

Estimate a linear model using the intrade margin in the average intrade margin in the week before the election to predict vote margin in 2008. How well does the model predict?

Solution 5

```
latest08 <- intresults08[intresults08$DaysToElection <8,]
average.Intrade <- tapply(latest08$obama.intmarg, latest08$state, mean)
true.margin <- tapply(latest08$obama.actmarg, latest08$state, mean)

int.fit08 <- lm(true.margin ~ average.Intrade)
coef(int.fit08)
summary(int.fit08)$r.squared
```

Question 6

Next, we read in the same data for the 2012 election. Use the linear model created above to create predictions for the margin in 2012. Calculate and plot the prediction error.

Solution 6

```
data2012 <- read.csv("intresults12.csv")
data2012$DaysToElection <- as.Date("2008-11-06") - as.Date(data2012$day)
data2012$obama.intmarg <- data2012$PriceD - data2012$PriceR
data2012$obama.actmarg <- data2012$Obama - data2012$Romney
```

Solution 6

```
latest12
<- data2012[data2012$DaysToElection <8,]

average.Intrade12
<- tapply(latest12$obama.intmarg, latest12$state, mean, na.rm = T)

true.margin12
<- tapply(latest12$obama.actmarg, latest12$state, mean, na.rm = T)

prediction
<- coef(int.fit08)[1] + coef(int.fit08)[2]*average.Intrade12

error <- true.margin12 - prediction
hist(error)
```

Linear Regression and RCTs

Can we estimate regression models on data from experiments?

Linear Regression and RCTs

Can we estimate regression models on data from experiments?

Yes, treatment status as the independent variable (0 or 1)

Linear Regression and RCTs

- y $= \alpha + \beta$ * treatment $+ \epsilon$
- What is the interpretation of α here?

Linear Regression and RCTs

- y = α + β * treatment + ϵ
- What is the interpretation of α here?
- What is the interpretation of β?

Linear Regression and RCTs

- y = α + β * treatment + ϵ
- β = average treatment effect
- The two predicted values are the average outcome under each condition

Linear Regression and RCTs

- y $= \alpha + \beta$ * treatment $+ \epsilon$
- $\beta =$ average treatment effect
- The two predicted values are the average outcome under each condition

- β: Predicted change in Y *caused* by increase of T by 1

Linear Regression and RCTs

- y $= \alpha + \beta$ * treatment $+ \epsilon$
- $\beta =$ average treatment effect
- The two predicted values are the average outcome under each condition

- β: Predicted change in Y *caused* by increase of T by 1

Remember, generally regression coefficents are not to be interpreted as causal effects!

Race and Job Applications

```
resume <- read.csv("resume.csv")
head(resume)
```

```
  firstname    sex  race call
1   Allison female white    0
2   Kristen female white    0
3   Lakisha female black    0
4   Latonya female black    0
5    Carrie female white    0
6       Jay   male white    0
```

- Randomized "race" in job applications
- What is the effect of race on likelyhood of callback?

Marianne Bertrand and Sendhil Mullainathan (American Economic Review 2004)

Race and Job Applications

```
mean(resume$call[resume$race == "black"])
mean(resume$call[resume$race == "white"])
mean(resume$call[resume$race == "black"]) - mean(resume$call[resume$race == "white"])
```

```
[1] 0.06447639
```

```
[1] 0.09650924
```

```
[1] -0.03203285
```

Race and Job Applications

```
linear <- lm(call ~ race, data = resume)
coef(linear)
```

```
(Intercept)   racewhite
 0.06447639  0.03203285
```

R automatically turns the factor into a dummy (binary) variable

Race and Job Applications

```
linear <- lm(call ~ race, data = resume)
coef(linear)
```

```
(Intercept)   racewhite
 0.06447639  0.03203285
```

R automatically turns the factor into a dummy (binary) variable

- α is the intercept, when X = 0 (i.e. race is "black")
- β is change in when X is set to 1 (i.e. race is "white")

Linear Regression with multiple independent variables

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

- principle of regression model stays the same
- we attempt to draw the best fitting line through a cloud of points (now in multiple dimensions)

Linear Regression with multiple independent variables

We still minimize the sum of the squared residuals:

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Linear Regression with multiple independent variables

We still minimize the sum of the squared residuals:

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

Linear Regression with multiple independent variables

We still minimize the sum of the squared residuals:

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

And thus:

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

Linear Regression with multiple independent variables

Interpretation:

- α: Intercept or $\hat{y}$ when all $X_p = 0$

Linear Regression with multiple independent variables

Interpretation:

- α: Intercept or $\hat{y}$ when all $X_p = 0$
- β_p: Slope of predictor X_p

Linear Regression with multiple independent variables

Interpretation:

- α: Intercept or $\hat{y}$ when all $X_p = 0$
- β_p: Slope of predictor X_p
- β_p: Predicted change in $\hat{Y}$ when X_p increases by 1 and all other predictors are held constant!

Linear Regression with multiple independent variables

- β_p: Predicted change in $\hat{Y}$ when X_p increases by 1 and all other predictors are held constant!
- we can use the multiple regression to control for confounders

Linear Regression with multiple independent variables

- β_p: Predicted change in $\hat{Y}$ when X_p increases by 1 and all other predictors are held constant!
- we can use the multiple regression to control for confounders

- impact of each individual predictor when the other predictors do not change
- Example: Association between income and child mortality when regime type is not changing

Linear Regression with multiple independent variables in *R*

```
result <- lm(y ~ x1 + x2 + x3 + x4, data = data)
coef(result)
```

Linear Regression with multiple independent variables in *R*

```
data <- read.csv("bivariate_data.csv")
data2010 <- subset(data, Year == 2010)
bivar <- lm(Child.Mortality ~ log(GDP), data = data)
coef(bivar)
summary(bivar)$r.squared
```

Linear Regression with multiple independent variables in *R*

```
data <- read.csv("bivariate_data.csv")
data2010 <- subset(data, Year == 2010)
bivar <- lm(Child.Mortality ~ log(GDP), data = data2010)
coef(bivar)
summary(bivar)$r.squared
```

```
(Intercept)    log(GDP)
  276.58162   -26.12717

[1] 0.586953
```

Linear Regression with multiple independent variables in *R*

```
data <- read.csv("bivariate_data.csv")
data2010 <- subset(data, Year == 2010)
multiple <- lm(Child.Mortality ~ log(GDP) + PolityIV, data = data2010)
coef(multiple)
summary(multiple)$r.squared
```

Linear Regression with multiple independent variables in *R*

```
data <- read.csv("bivariate_data.csv")
data2010 <- subset(data, Year == 2010)
multiple <- lm(Child.Mortality ~ log(GDP) + PolityIV, data = data2010)
coef(multiple)
summary(multiple)$r.squared
```

```
(Intercept)    log(GDP)    PolityIV
 277.845620  -25.641789   -1.029062

[1] 0.6113747
```

Linear Regression with multiple independent variables in *R*

```
data <- read.csv("bivariate_data.csv")
data2010 <- subset(data, Year == 2010)
multiple <- lm(Child.Mortality ~ log(GDP) + PolityIV, data = data2010)
coef(multiple)
coef(bivar)
```

```
(Intercept)    log(GDP)    PolityIV
 277.845620  -25.641789   -1.029062

(Intercept)    log(GDP)
  276.58162   -26.12717
```

Linear Regression with multiple independent variables in R

- In multiple regression models we want to adjust the goodness of fit statistic by the number of variables included
- This is done via the degrees of freedom (DF) adjustment:

$$\text{adjusted}R^2 \quad = \quad 1 - \frac{SSR/(n-p-1)}{TSS/(n-1)}$$

Linear Regression with multiple independent variables in *R*

```
data <- read.csv("bivariate_data.csv")
data2010 <- subset(data, Year == 2010)
multiple <- lm(Child.Mortality ~ log(GDP) + PolityIV, data = data2010)
coef(multiple)
summary(multiple)$r.squared
summary(multiple)$adj.r.squared
```

Linear Regression with multiple independent variables in *R*

```
data <- read.csv("bivariate_data.csv")
data2010 <- subset(data, Year == 2010)
multiple <- lm(Child.Mortality ~ log(GDP) + PolityIV, data = data2010)
coef(multiple)
summary(multiple)$r.squared
summary(multiple)$adj.r.squared
```

```
(Intercept)    log(GDP)    PolityIV
 277.845620  -25.641789   -1.029062

[1] 0.6113747

[1] 0.6061582
```