Regression Analysis

Te Rutherford

March 12, 2015

Agenda

- Last time

Agenda

- Last time
  - Method of Least Square

Agenda

- Last time
  - Method of Least Square
  - Testing whether β_1 significantly contributes to the prediction of y

Agenda

- Last time
  - Method of Least Square
  - Testing whether β_1 significantly contributes to the prediction of y
  - R-Squared

Agenda

- Last time
  - Method of Least Square
  - Testing whether β_1 significantly contributes to the prediction of y
  - R-Squared
- Today

Agenda

- Last time
  - Method of Least Square
  - Testing whether β_1 significantly contributes to the prediction of y
  - R-Squared
- Today
  - More on r-squared

Agenda

- Last time
  - Method of Least Square
  - Testing whether β_1 significantly contributes to the prediction of y
  - R-Squared
- Today
  - More on r-squared
  - Regression analysis steps

Agenda

- Last time
  - Method of Least Square
  - Testing whether β_1 significantly contributes to the prediction of y
  - R-Squared
- Today
  - More on r-squared
  - Regression analysis steps
  - Limitation of this approach

What is r-squared?

- Baseline model

$$y = \beta_0 + \epsilon$$

$$\hat{y} = \beta_0$$

What is r-squared?

- Baseline model

$$y = \beta_0 + \epsilon$$

$$\hat{y} = \beta_0$$

- Simple Linear Regression model

$$y = \beta_0 + \beta_1 * x + \epsilon$$

$$\hat{y} = \beta_0 + \beta_1 * x$$

What is r-squared?

- Baseline model

$$y = \beta_0 + \epsilon$$

$$\hat{y} = \beta_0$$

- Simple Linear Regression model

$$y = \beta_0 + \beta_1 * x + \epsilon$$

$$\hat{y} = \beta_0 + \beta_1 * x$$

- Definition of r^2

$$r^2 = \frac{SSE_{basemodel} - SSE_{model}}{SSE_{basemodel}}$$

What does it tell us? (1)

- Definition of r^2

$$r^2 = \frac{SSE_{basemodel} - SSE_{model}}{SSE_{basemodel}}$$

What does it tell us? (1)

- Definition of r^2

$$r^2 = \frac{SSE_{basemodel} - SSE_{model}}{SSE_{basemodel}}$$

- The percentage of SSE reduction from the base model if we use the regression model instead.

What does it tell us? (1)

- Definition of r^2

$$r^2 = \frac{SSE_{basemodel} - SSE_{model}}{SSE_{basemodel}}$$

- The percentage of SSE reduction from the base model if we use the regression model instead.
  - If $SSE_{basemodel} = 30$ and $SSE_{model} = 20$, then

$$r^2 = \frac{30 - 20}{30} = 1/3 \approx 0.33$$

What does it tell us? (1)

- Definition of r^2

$$r^2 = \frac{SSE_{basemodel} - SSE_{model}}{SSE_{basemodel}}$$

- The percentage of SSE reduction from the base model if we use the regression model instead.
  - If $SSE_{basemodel} = 30$ and $SSE_{model} = 20$, then

$$r^2 = \frac{30 - 20}{30} = 1/3 \approx 0.33$$

  - The percentage of SSE reduction $= r^2 * 100 = 33.33\%$

What does it tell us? (2)

- The percentage of SSE reduction from the base model if we use the regression model instead.

What does it tell us? (2)

- The percentage of SSE reduction from the base model if we use the regression model instead.
- r^2 is the percentage of the variation explained by the model

Example from 10.66

The strength of concrete pipes over time

```
library(gdata)
```

```
d = read.xls('http://www.typ-stats.com/xdatasets/PIPELOAD.X
head(d)
```

```
##    LOAD AGE
## 1 11450  20
## 2 10420  20
## 3 11142  20
## 4 10840  25
## 5 11170  25
## 6 10540  25
```

Example from 10.66

```r
mymodel = lm(LOAD~AGE, data=d)
summary(mymodel)
```

```
Residual standard error: 460.7 on 7 degrees of freedom
Multiple R-squared:  0.7311,  Adjusted R-squared:  0.6927
F-statistic: 19.03 on 1 and 7 DF,  p-value: 0.003305
```

What does it tell us? (2)

- 73.11 % of the SSE is reduced by using the linear model.

What does it tell us? (2)

- 73.11 % of the SSE is reduced by using the linear model.
- 73.11 % of the variation in the data is explained by the age.

What does it tell us? (2)

- 73.11 % of the SSE is reduced by using the linear model.
- 73.11 % of the variation in the data is explained by the age.
  - This is a pretty good model.

What does it tell us? (2)

- 73.11 % of the SSE is reduced by using the linear model.
- 73.11 % of the variation in the data is explained by the age.
  - This is a pretty good model.
- What if it's much lower? like 30% .

Steps in regression analysis

- Determine x and y

Steps in regression analysis

- Determine x and y
- Fit the linear regression model with `lm` command

Steps in regression analysis

- Determine x and y
- Fit the linear regression model with `lm` command
- Draw a scatterplot with linear trend line

Steps in regression analysis

- Determine x and y
- Fit the linear regression model with `lm` command
- Draw a scatterplot with linear trend line
- Interpret the regression coefficients

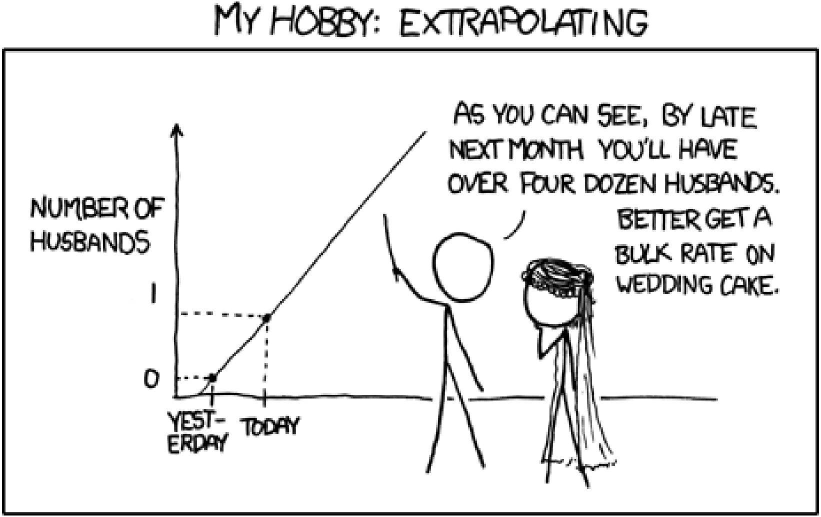
Steps in regression analysis

- Determine x and y
- Fit the linear regression model with `lm` command
- Draw a scatterplot with linear trend line
- Interpret the regression coefficients
- Interpret r^2

Steps in regression analysis

- Determine x and y
- Fit the linear regression model with `lm` command
- Draw a scatterplot with linear trend line
- Interpret the regression coefficients
- Interpret r^2
- Determine whether β_1 is significant. Interpret the significance.

Limitation of regression analysis



Limitation of regression analysis

- One outlier can really mess up your regression model. β_0, β_1, r^2 won't be accurate anymore.

```
d = read.xls('http://www.typ-stats.com/xdatasets/PIPELOAD.X
d[10,] = c(20000, 30) #<--- this will screw up the data
library(ggplot2)
ggplot(mapping=aes(x=AGE,y=LOAD),data=d) + geom_point() +
  geom_smooth(method='lm')
```