STA 302 / 1001

Note Title 9/22/2011

What to Read? Ch 1 – Motivation
Ch 2

Assignment 1 will be posted soon

Andriy will have CQUEST hours in RW 107 / 109
Tuesday Sept. 27 11:00 – 13:00
Tuesday Oct. 4 11:00 – 13:00
(so he won't be here on those 2 Tuesdays.)

Try SAS & the assignment before you ask him for help.

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + e$$

Y: r.v.

e: random var. $E(e) = 0$

X: sometimes random

β_0, β_1: parameters of model
Need to estimate from data

Estimators: $\hat{\beta}_0, \hat{\beta}_1$

Estimates: b_0, b_1

How to estimate β_0, β_1 from observed data:

Data: n pairs (x_i, y_i) $i = 1, \ldots, n$

Fitted value for each x_i:

$$\hat{y}_i = b_0 + b_1 x_i \quad \text{(estimate } \varepsilon \text{ by } 0\text{)}$$

"y_i hat"

Deviations from line: $y_i - \hat{y}_i = \hat{e}_i$ "residual"

Why interested in vertical errors?

- regression treats x and y differently
  - trying to predict y from x

Method of Least Squares

- makes no statistical assumptions

Least Squares: find b_0, b_1 that minimize sum of squares of residuals

Why Least Squares: - mean square error is most common way to measure error in statistics

- Gauss-Markov Theorem
  - least squares est. are "best" (min. variance)

Define RSS = "residual sum of squares"

$$\frac{\partial RSS}{\partial b_1} = -2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial RSS}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

Find b_0, b_1 that minimize this

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Set equal to 0:

b_0, b_1 must satisfy

"Normal equations"

$$\sum y_i = n b_0 + b_1 \sum x_i \quad (1)$$

$$\sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2 \quad (2)$$

Notation: $\bar{x} = \frac{1}{n} \sum x_i$, $\bar{y} = \frac{1}{n} \sum y_i$

(1) $\Rightarrow$ $b_0 = \bar{y} - b_1 \bar{x}$

(2) $\Rightarrow$ $\sum x_i y_i = n \bar{x} \bar{y} + b_1 \left(\sum x_i^2 - n \bar{x}^2 \right)$

$$b_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

also $$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

(EXERCISE: Show these 2 formulas for b_1 are equivalent)

$$b_1 = \frac{Sxy}{Sxx}$$, $Sxy = \sum (x_i - \bar{x})(y_i - \bar{y})$, $Sxx = \sum (x_i - \bar{x})^2$

For CFCs example:

Before MP, $b_0 = -19064$, $b_1 = 9.71152$

After MP, $b_0 = 3929.6775$, $b_1 = -1.83289$

Fitted model pre-MP:

$\hat{y}_i = -1906.4 + 9.7152\, x_i$

x_i: time in years

y_i: CFC conc in ppt

Fitted model post-MP:

$\hat{y}_i = 3929.68 + -1.834\, x_i$

- Estimate time of first manufacture of CFCs:
  - Pre-MP model, want x when $y=0$, gives ~1963
  - Actually first-produced in 1931
  - Why a discrepancy? Assumes same growth rate from beginning
    - we only know what happened 1977 →

- When will atmospheric concentrations of CFCs reach 0?
  Post-MP model want x when $y=0$
  Not a reasonable question
  (would have to assume linear decline continues at same rate)
- What does model say about 2011? A: nothing
- What is the practical interpretation of intercept?
  A: nothing, $x=0$ is not in range of data.

Our models ~~was~~ were based on data 1977-1990
1995-2004

We have no idea what happens outside that range

EXTRAPOLATION IS DANGEROUS

Properties of Fitted Regression Line (consequences of least squares (no statistical assumptions)

① Residuals $\hat{e}_i = y_i - \hat{y}_i$

$= y_i - (b_0 + b_1 x_i)$

$= y_i - (\bar{y} - b_1 \bar{x}) - b_1 x_i$

$= (y_i - \bar{y}) - b_1 (x_i - \bar{x})$

$\sum (x_i - \bar{x}) = 0$

$\sum \hat{e}_i = 0$

So $\bar{\hat{e}} = 0$ (average of residuals is 0)

② $\sum \hat{e}_i^2 \neq 0$ (unless data fit perfectly on line)

$= RSS$

③ $\sum \hat{e}_i x_i = 0$ } Exercise

④ $\sum \hat{e}_i \hat{y}_i = 0$ } Exercise

($\sum \hat{e}_i y_i \neq 0$ hard to show)

⑤ $\sum \hat{y}_i = \sum (b_0 + b_1 x_i)$

$= \sum (\bar{y} - b_1 \bar{x} + b_1 x_i)$

$= n\bar{y} - b_1 n\bar{x} + b_1 n\bar{x}$

$= n\bar{y} = \sum y_i$

So far: We assumed that linear model is appropriate

Now some statistical assumptions

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Gauss-Markov Conditions

1. $E(e_i) = 0$
2. $Var(e_i)$ constant (same for all observations)
3. e_i's uncorrelated

$$\Rightarrow E(Y_i \mid X_i = x_i) = \beta_0 + \beta_1 x_i$$

Gauss-Markov Theorem (without proof)

x_i's known, fixed

Linear model, with Gauss-Markov conditions

Can show: - least squares estimators are unbiased (will show)

- linear combinations of y_i's

(Exercise: show $b_1 = \sum d_i y_i$ where d_i function of x_i's)

Theorem says: Least squares estimators are BLUE

("Best linear Unbiased Estimators")

Best = minimum variance

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for x_i not random

estimate of $\beta_0, \beta_1 = b_0, b_1$ functions of observed data (numbers)

estimators of $\beta_0, \beta_1 = \hat{\beta}_0, \hat{\beta}_1$ functions of random variables, estimators of r.v.'s, can talk about their statistical properties

$b_0 = \bar{y} - b_1 \bar{x}$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum x_i Y_i - n \bar{x} \bar{Y}}{\sum x_i^2 - n \bar{x}^2}$$

Mean and Variance of $\hat{\beta}_1$

Model: $Y_i = \beta_0 + \beta_1 x_i + e_i$

Going to treat x's as not random

If x's r.v.'s, do everything the same, but condition on $X_1 = x_1, \ldots, X_n = x_n$

$$E(\hat{\beta}_1) = E\left(\frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum x_i^2 - n\bar{x}^2} \right), \quad SXX = \sum x_i^2 - n\bar{x}^2 = \sum (x_i - \bar{x})^2$$

$$= \frac{1}{SXX} E\left(\sum x_i Y_i - n\bar{x}\bar{Y} \right)$$

$$\left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right]$$

$$= \left[\sum_{i=1}^{n} (y_i - (a + b x_i))^2 \right]$$

$$= \frac{1}{SxS} \left[\sum_{i=1}^{n} (y_i - (a + b x_i))^2 \right] SxS$$

$$= \frac{1}{SxS} \left[\sum_{i=1}^{n} (y_i - (a + b x_i))^2 \right] SxS$$

$$= \frac{\left[\sum_{i=1}^{n} (y_i - (a + b x_i))^2 \right]}{SxS}$$

$$= \frac{\text{SSE}}{SxS}$$

$$= \frac{\text{SSE}}{n-2}$$

$$= \text{MSE}$$

$$= \text{Estimate of } \sigma^2$$

$$\text{So } E(\hat{\beta}_1) = \frac{1}{S_{xx}}\left[\beta_0 n\bar{x} + \beta_1 \sum x_i^2 - n\bar{x}\beta_0 - \beta_1 n\bar{x}^2\right]$$

$$= \frac{1}{S_{xx}}\beta_1\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$= \beta_1$$

So $\hat{\beta}_1$ is an UNBIASED estimator of β_1

Exercise: Show $\hat{\beta}_0$ is an unbiased estimator of β_0

Want: $\mathrm{Var}(\hat{\beta}_1)$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x}) Y_i - \sum (x_i - \bar{x}) \bar{Y}}{S_{xx}}$$

$$= \frac{\sum (x_i - \bar{x}) Y_i}{S_{xx}}$$

Recall U, V r.v.'s, a, b constants

$$\mathrm{Var}(aU + bV)$$

$$= a^2 \mathrm{Var}\, U + b^2 \mathrm{Var}\, V + 2ab\, \mathrm{Cov}(U, V)$$

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{\sum (x_i - \bar{x}) Y_i}{S_{xx}} \right)$$

(treating x's as const; otherwise condition on $X_1 = x_1, \ldots, X_n = x_n$)

$$= \frac{1}{S_{xx}^2} \sum \left[(x_i - \bar{x})^2 \, \mathrm{Var}(Y_i) \right]$$

$\mathrm{Var}(\sum Y_i)$

$= \sum \mathrm{Var}(Y_i)$?

Yes, because assumed e_i's uncorr.

$Y_i = \beta_0 + \beta_1 x_i + e_i$

$\mathrm{Var}(Y_i) = \mathrm{Var}(e_i)$

Const

$= \sigma^2$

(σ^2 is the var of the error terms)

$$= \frac{1}{SXX^2} \sum (x_i - \bar{x})^2 \sigma^2$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SXX} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

So the more spread out the x_i's, the smaller the variance of the estimator of the slope

Exercise: Show $\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right]$

Show $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$