Linear Regression Models
P8111

Lecture 04

Jeff Goldsmith
January 28, 2016

THE DEPARTMENT OF BIOSTATISTICS
Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

Today's lecture

- Simple Linear Regression
- Least Squares Estimation

Regression modeling

$E(y|x) = ???$

- Want to use predictors to learn about the outcome distribution, particularly conditional expected value.
- Formulate the problem parametrically

$$E(y \mid x) = f(x; \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- (Note that other useful quantities, like covariance and correlation, tell you about the joint distribution of y and x)

Covariance and Correlation

$$\mathrm{cov}(x,y) = E\left(\underline{(x-\mu_x)}\,\underline{(y-\mu_y)}\right)$$

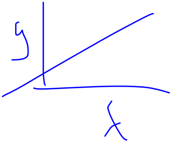
$$\widehat{\mathrm{cov}}(x,y) = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$\mathrm{cor}(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}$$

Simple linear regression

- Linear models are a special case of all regression models; simple linear regression is the simplest place to start
- Only one predictor:

$$E(y \mid x) = f(x; \beta) = \beta_0 + \beta_1 x_1$$



- Useful to note that $x_0 = 1$ (implicit definition)
- Somehow, estimate β_0, β_1 using observed data.

Coefficient interpretation

$$E(y|x) = \beta_0 + \beta_1 x$$

$$E(y|x=0) = \beta_0$$

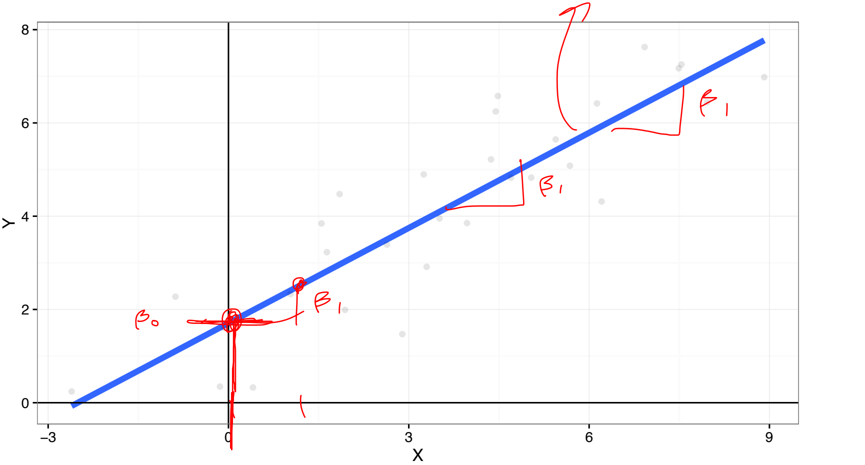
$$\beta_1 = (\beta_0 + \beta_1 17) - \beta_0 - \beta_1 16$$

$$= E(y|x=1) - E(y|x=0)$$

$$17 \qquad 16$$

Coefficient interpretation

$E(y|x) = \beta_0 + \beta_1 x$



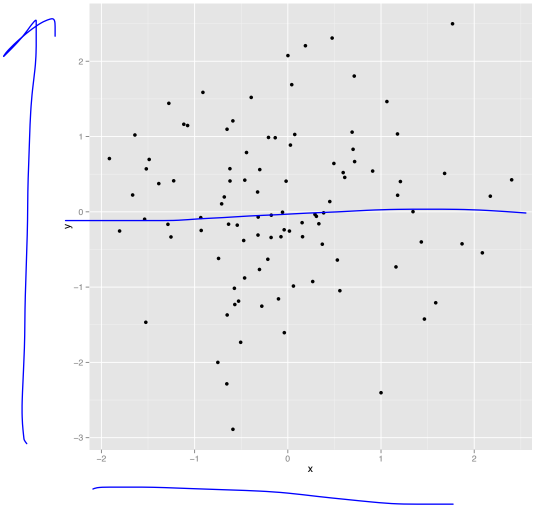
Look at the data

- Before Modeling

- Plot the data (using `ggplot` ...)
- Do the data look like the assumed model?
- Should you be concerned about outliers?
- Define what you expect to see before fitting any model.

Look at the data

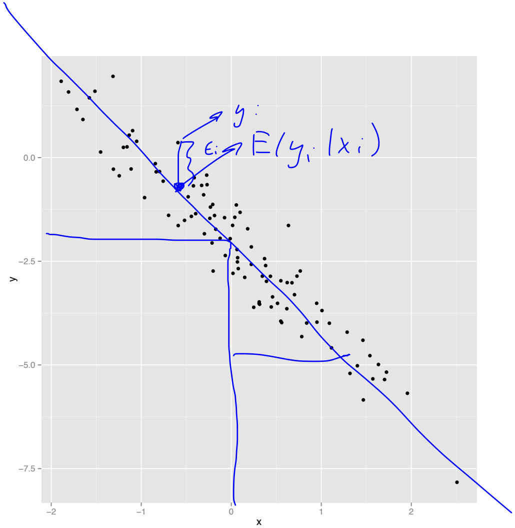
$E(y|x) = \beta_0 + \beta_1 x$



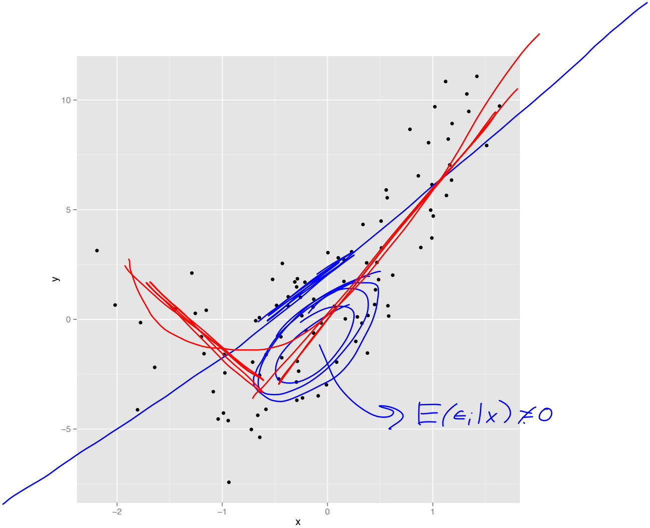
$\hat{\beta}_0 = 0$

$\hat{\beta}_1 \approx 0$

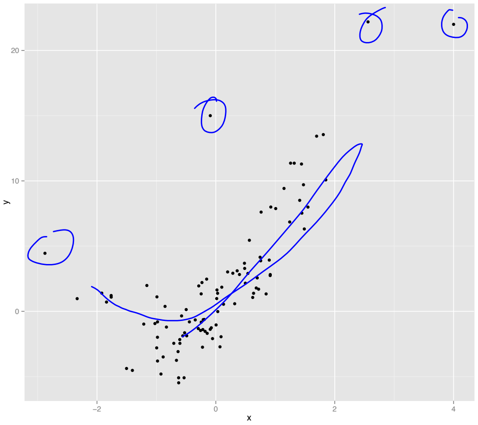
Look at the data



Look at the data



Look at the data



Least squares estimation

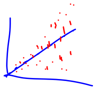
$E(y|x)$

- Observe data (y_i, x_i) for subjects $1, \ldots, n$. Want to estimate β_0, β_1 in the model

$y_i = E(y_i | x_i) + \epsilon_i$ N

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i;\ \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

- Note the assumptions on the variance:
  - $E(\epsilon \mid x) = E(\epsilon) = 0$
  - Constant variance

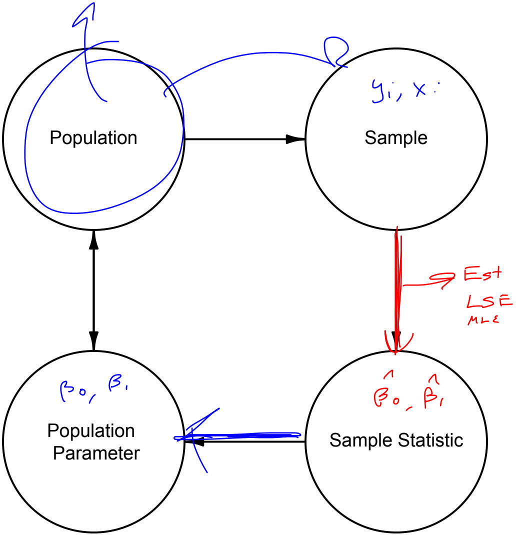


  - Independence
  - [Normally distributed is not needed for least squares, but is ~~needed~~ for inference]

nice/useful

need for MLE

Circle of Life



Least squares estimation

- Recall that for a single sample $y_i, i \in 1, \ldots, n$, the sample mean $\hat{\mu}_y$ minimizes the sum of squared deviations.

$$RSS(\mu_y) = \sum (y_i - \mu_y)^2 \longrightarrow E(y)$$

$$E(y|x) = \beta_0 = \mu_y$$

$$\frac{\partial RSS(\mu_y)}{\partial \mu_y} = -\cancel{2} \sum (y_i - \mu_y) = 0$$

$$\sum y_i - \sum \mu_y = 0$$

$$n\mu_y = \sum y_i$$

$$\hat{\mu}_y = \frac{\sum y_i}{n}$$

Least squares estimation

$$E(y|x) = \beta_0 + \beta_1 x$$

- Find $\hat{\beta}_0$.

$$RSS(\beta_0, \beta_1) = \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial RSS(\beta_0)}{\partial \beta_0} = -\cancel{2} \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum y_i - n\beta_0 - \beta_1 \sum x_i = 0$$

$$n\beta_0 = \sum y_i - \beta_1 \sum x_i$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

Least squares estimation

- Now find $\hat{\beta}_1$.

$$RSS(\beta_1) = \sum (y_i \overbrace{- \bar{y} + \beta_1 \bar{x}}^{\hat{\beta}_0} - \beta_1 x_i)^2$$

$$= \sum \left(\underline{(y_i - \bar{y})} - \beta_1 \underline{(x_i - \bar{x})} \right)^2$$

$$\frac{\partial RSS(\beta_1)}{\partial \beta_1} = -\cancel{2} \sum \left((y_i - \bar{y}) - \beta_1 (x_i - \bar{x}) \right)(x_i - \bar{x}) = 0$$

$$= \sum (y_i - \bar{y})(x_i - \bar{x}) - \beta_1 \sum (x_i - \bar{x})^2 = 0$$

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

Note about correlation

$$\rho = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}; \qquad \beta_1 = \frac{cov(x,y)}{var(x)}$$

R does exactly what we now expect

> data =

```
> linmod = lm(y~x, data = data)
> summary(linmod)

Call:
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5202 -0.5050 -0.2297  0.5753  1.8534

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.08743    0.22958   9.092 7.53e-10 ***
x            0.61396    0.05415  11.338 5.61e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8084 on 28 degrees of freedom
Multiple R-squared:  0.8211,Adjusted R-squared:  0.8148
F-statistic: 128.6 on 1 and 28 DF,  p-value: 5.612e-12
```

$\hat{\beta}_0$ $\hat{\beta}_1$

R does exactly what we now expect

library(broom)

```
> tidy(linmod)
         term  estimate  std.error statistic      p.value
1 (Intercept) 2.0874344 0.22958105  9.092364 7.529711e-10
2           x 0.6139621 0.05415004 11.338166 5.611585e-12
> glance(linmod)
  r.squared adj.r.squared     sigma statistic      p.value df   logLik ...
1  0.821148     0.8147604 0.8084399   128.554 5.611585e-12  2 -35.1538 ...
>
>
> beta1 = with(data, sum((x - mean(x))*(y - mean(y))) / sum((x - mean(x))^2))
> beta0 = with(data, mean(y) - beta1*mean(x))
> c(beta0, beta1)
[1] 2.0874344 0.6139621
```

data$x data$x data$y

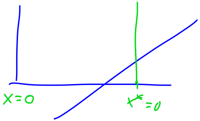
~~attach(data)~~

Note on interpretation of β_0

Recall $\beta_0 = E(y|x = 0)$

- This often makes no sense in context
- "Centering" x can be useful: $x^* = x - \bar{x}$
- Center by mean, median, minimum, etc
- Effect of centering on slope:

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

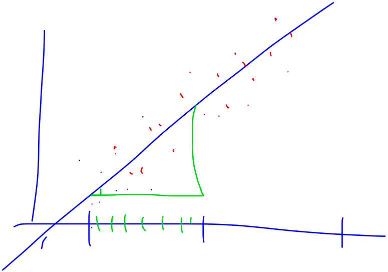


Note on interpretation of β_0, β_1

- The interpretations are sensitive to the scale of the outcome and predictors (in reasonable ways)
- You can't get a better model fit by rescaling variables

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad x^* = cx$$

$$\hat{\beta}_1^* = \frac{1}{c} \hat{\beta}_1$$



R example

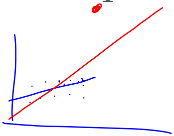
x.scaled.by.2

```
> data = mutate(data, x.cen = x - mean(x), x2 = x*2)
> linmod.cen = lm(y ~ x.cen, data = data)
> tidy(linmod.cen)
         term  estimate  std.error statistic      p.value
1 (Intercept) 4.0811993 0.14760027  27.65035 7.172437e-22
2       x.cen 0.6139621 0.05415004  11.33817 5.611585e-12
```

R example

```
> linmod.x2 = lm(y ~ x2, data = data)
> tidy(linmod.x2)
         term  estimate  std.error statistic      p.value
1 (Intercept) 2.0874344 0.22958105  9.092364 7.529711e-10
2          x2 0.3069811 0.02707502 11.338166 5.611585e-12
```

Least squares notes and foreshadowing



$$RSS(\beta_0,\beta_1) = \sum (y_i - E(y_i|x_i))^2 \nearrow \text{sample mean}$$

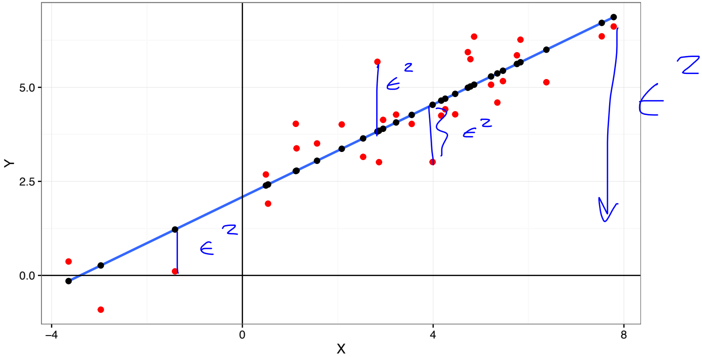
$$\sum |y_i - E(y_i|x_i)| \nearrow \text{sample median}$$

- Didn't have to choose to minimize squares – could minimize absolute value, for instance.
- Least squares estimates turn out to be a "good idea" – unbiased, BLUE.
- Later we'll see about maximum likelihood as well.

Geometric interpretation of least squares

Least squares minimizes the sum of squared vertical distances between observed and estimated y's:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{I} (y_i - (\beta_0 + \beta_1 x_i))^2$$



Least squares in regression generally

Broadly speaking, in regression we often are concerned with minimizing

$$E[\underline{f(x) + \epsilon} - \hat{f}(x)]^2$$

by choosing a "good" $\hat{f}$. For a given $\hat{f}$ this decomposes into

$$\underbrace{E[f(x) - \hat{f}(x)]^2}_{\text{Bias}^2} + \underbrace{\text{Var}(\epsilon)}_{+\ \text{Var}}$$

- Some variance isn't explainable (we just don't know how much)
- Focus on getting the left component right
- Minimizing squared error for *unseen* data is the real goal

Today's big ideas

- Simple linear regression – model and interpretation
- Least squares estimation

- Suggested reading: Faraway Ch 1, 2.1; ISLR 3.1