

# Linear Regression Models

## P8111

Lecture 05

Jeff Goldsmith  
February 4, 2016



THE DEPARTMENT OF  
**BIostatISTICS**



Columbia University  
MAILMAN SCHOOL  
OF PUBLIC HEALTH

# Today's lecture

- Simple Linear Regression Continued
- Multiple Regression Intro

# Simple linear regression model

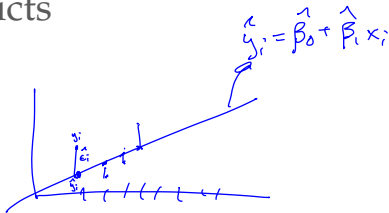
- Observe data  $(y_i, x_i)$  for subjects  $1, \dots, n$ . Want to estimate  $\beta_0, \beta_1$  in the model

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

- Note the assumptions on the variance:
  - $E(\epsilon | x) = E(\epsilon) = 0$  ✓
  - Constant variance ✓
  - Independence ✓
  - [Normally distributed is not needed for least squares, but is nice for inference and needed for MLE]

# Some definitions / SLR products

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



- Fitted values:  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residuals / estimated errors:  $\hat{e}_i = y_i - \hat{y}_i$
- Residual sum of squares:  $\sum_{i=1}^n \hat{e}_i^2$
- ✓ ■ Residual variance:  $\hat{\sigma}^2 = \frac{RSS}{n-2}$
- Degrees of freedom:  $n - 2$

Notes: residual sample mean is zero; residuals are uncorrelated with fitted values.

$R^2$ 

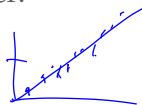
Looking for a measure of goodness of fit.

- RSS by itself doesn't work so well:

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

- Coefficient of determination ( $R^2$ ) works better:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$



# $R^2$

## Some notes about $R^2$

- Interpreted as proportion of outcome variance explained by the model.
- Alternative form

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- $R^2$  is bounded:  $0 \leq R^2 \leq 1$
- For simple linear regression only,  $R^2 = \rho^2$

# ANOVA

Lots of sums of squares around.

- Regression sum of squares  $SS_{reg} = \sum(\hat{y}_i - \bar{y})^2$
- Residual sum of squares  $SS_{res} = \sum(y_i - \hat{y}_i)^2$  ✓
- Total sum of squares  $SS_{tot} = \sum(y_i - \bar{y})^2$  ✓
- All are related to sample variances

Analysis of variance (ANOVA) seeks to address goodness-of-fit by looking at these sample variances.

# ANOVA

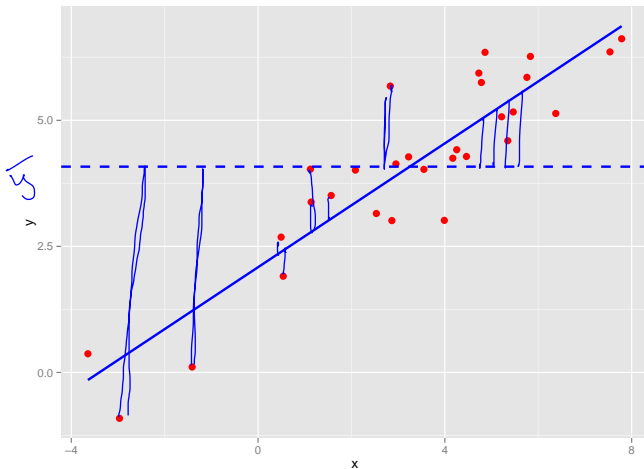
ANOVA is based on the fact that  $\underline{SS_{tot}} = \underline{SS_{reg}} + \underline{SS_{res}}$

HW 2



# ANOVA

ANOVA is based on the fact that  $SS_{tot} = SS_{reg} + SS_{res}$



# ANOVA and $R^2$

- Both take advantage of sums of squares
- Both are defined for more complex models
- ANOVA can be used to derive a “global hypothesis test” based on an F test

# R example

data =

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

Call:

```
lm(formula = y ~ x, data = data)
```

Coefficients:

```
(Intercept)      x  
    2.087         0.614
```

```
> 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

# R example

```
> 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
```

$n=30$

# R example

```
> names(linmod)
[1] "coefficients" "residuals" "effects" "rank"
[5] "fitted.values" "assign" "qr" "df.residual"
[9] "xlevels" "call" "terms" "model"
```

# R example

```
> linmod$residuals
      1          2          3          4          5          6
1.2555987 -0.2398006 0.2933523 -0.2499462 -1.5201821 -0.5099489
...
> linmod$fitted.values
      1          2          3          4          5          6
2.7754640 4.2675708 2.3901878 6.8676466 4.5362366 2.4181112
...
```

# R example

```
> names(summary(linmod))
[1] "call"          "terms"          "residuals"      "coefficients"
[5] "aliased"       "sigma"          "df"             "r.squared"
[9] "adj.r.squared" "fstatistic"     "cov.unscaled"
>
> summary(linmod)$coef
      Estimate Std. Error  t value    Pr(>|t|)
(Intercept)  2.0874344  0.22958105  9.092364 7.529711e-10
x              0.6139621  0.05415004 11.338166 5.611585e-12
>
> summary(linmod)$r.squared
[1] 0.821148
```

# R example

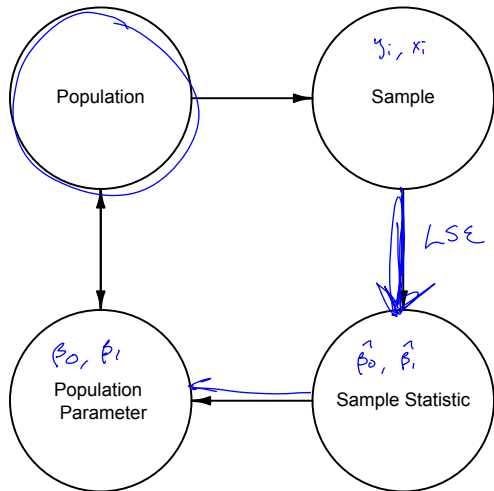
```
> anova(linmod)
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x         1  86.744   86.744  107.59 4.266e-11 ***
Residuals 28  22.575    0.806
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> 1 - 18.30 / (84.02 + 18.30)
[1] 0.8211493
```



# Properties of $\hat{\beta}_0, \hat{\beta}_1$



# Properties of $\hat{\beta}_0, \hat{\beta}_1$

$$y_i \sim (\beta_0 + \beta_1 x_i, \sigma^2) \sim$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Estimates are unbiased:

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

$$= E(\bar{y}) - E(\beta_1 \bar{x})$$

$$= E\left(\frac{\sum (\beta_0 + \beta_1 x_i + \epsilon_i)}{n}\right) - E\left(\beta_1 \frac{\sum x_i}{n}\right)$$

$$= E\left(\frac{\beta_0}{n}\right) + E\left(\frac{\beta_1 \sum x_i}{n}\right) + E\left(\frac{\sum \epsilon_i}{n}\right) - E\left(\beta_1 \frac{\sum x_i}{n}\right)$$

$$\underline{\epsilon_i \sim (0, \sigma^2)}$$

$$E(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} - E\left(\beta_1 \frac{\sum x_i}{n}\right)$$

$$\hat{\beta}_0 \sim (\beta_0, \text{---})$$

$$\hat{\beta}_1 \sim (\beta_1, \text{---})$$

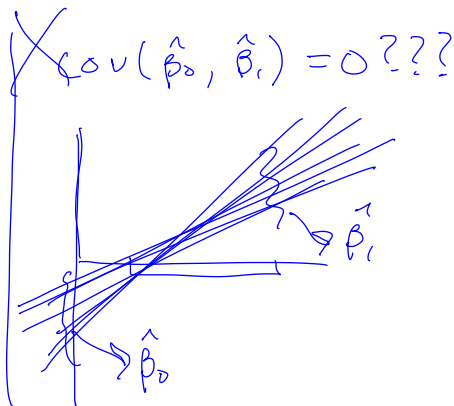
$$E\left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\right)$$

# Properties of $\hat{\beta}_0, \hat{\beta}_1$

Variances of estimates:

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$

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



$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

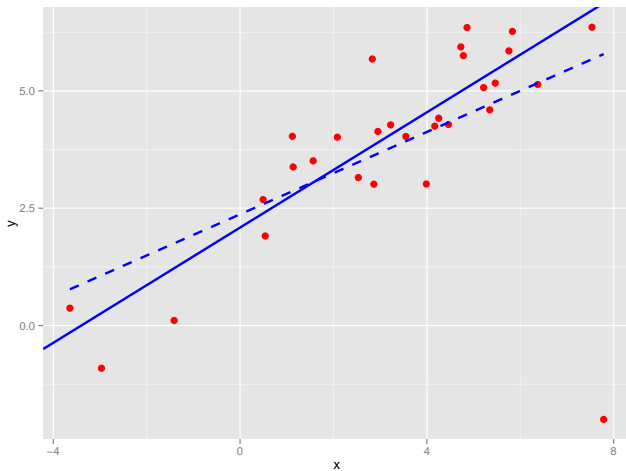
$$\text{Cov}(\hat{\beta}) = \begin{bmatrix} \text{var}(\hat{\beta}_0) & 0 \\ 0 & \text{var}(\hat{\beta}_1) \\ 0 & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ 0 & \text{cov}(\hat{\beta}_1, \hat{\beta}_0) \end{bmatrix}$$

# Properties of $\hat{\beta}_0, \hat{\beta}_1$

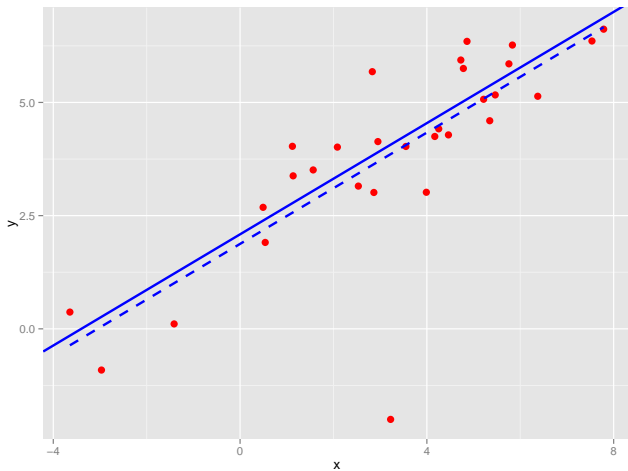
Note about the variance of  $\beta_1$ :

- Denominator contains  $SS_x = \sum(x_i - \bar{x})^2$
- To decrease variance of  $\beta_1$ , increase variance of  $x$

# Effect of data on $\beta_1$



# Effect of data on $\beta_1$



# Switching to multiple linear regression

- Observe data  $(y_i, x_{i1}, \dots, x_{ip})$  for subjects  $1, \dots, n$ . Want to estimate  $\beta_0, \beta_1, \dots, \beta_p$  in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i; \epsilon_i \stackrel{iid}{\sim} (0, \sigma^2)$$

- Assumptions (residuals have mean zero, constant variance, are independent) are as in SLR
- Notation is cumbersome. To fix this, let
  - $\mathbf{x}_i = [1, x_{i1}, \dots, x_{ip}]$
  - $\boldsymbol{\beta}^T = [\beta_0, \beta_1, \dots, \beta_p]$
  - Then  $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$

# Matrix notation

- Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & x_{ij} & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Then we can write the model in a more compact form:

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\epsilon}_{n \times 1}$$

- $\mathbf{X}$  is called the *design matrix*



# Matrix notation

$$y = X\beta + \epsilon$$

- $\epsilon$  is a random vector rather than a random variable
- $E(\epsilon) = 0$  and  $Var(\epsilon) = \sigma^2 I$
- Note that *Var* is potentially confusing; in the present context it means the “variance-covariance matrix”

# Mean and Variance of a Random Vector

- Let  $\mathbf{y}^T = [y_1, \dots, y_n]$  be an  $n$ -component random vector. Then its mean and variance are defined as

$$E(\mathbf{y})^T = [E(y_1), \dots, E(y_n)]$$

$$\text{Var}(\mathbf{y}) = E[(\mathbf{y} - E\mathbf{y})(\mathbf{y} - E\mathbf{y})^T] = E(\mathbf{y}\mathbf{y}^T) - (E\mathbf{y})(E\mathbf{y})^T$$

- Let  $\mathbf{y}$  and  $\mathbf{z}$  be an  $n$ -component and an  $m$ -component random vector respectively. Then their covariance is an  $n \times m$  matrix defined by

$$\text{Cov}(\mathbf{y}, \mathbf{z}) = E[(\mathbf{y} - E\mathbf{y})(\mathbf{z} - E\mathbf{z})^T]$$

# Basics on Random Vectors

Let  $A$  be a  $t \times n$  non-random matrix and  $B$  be a  $p \times m$  non-random matrix. Then

$$E(A\mathbf{y}) = AE(\mathbf{y})$$

$$\text{Var}(A\mathbf{y}) = A\text{Var}(\mathbf{y})A^T$$

$$\text{Cov}(A\mathbf{y}, B\mathbf{z}) = ACov(\mathbf{y}, \mathbf{z})B^T$$

# Today's big ideas

- Simple linear regression definitions
- Properties of SLR least squares estimates
- Matrix notation for MLR

- 
- Suggested reading: Faraway Ch 2.2 - 2.3; ISLR 3.1