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 $1, \ldots, n$. Want to estimate β_0, β_1 in the model

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

- 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 nice for inference and needed for MLE]

Some definitions / SLR products

- *Fitted values*: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- *Residuals / estimated errors*: $\hat{\epsilon}_i = y_i - \hat{y}_i$
- *Residual sum of squares*: $\sum_{i=1}^{n} \hat{\epsilon}_i^2$
- *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

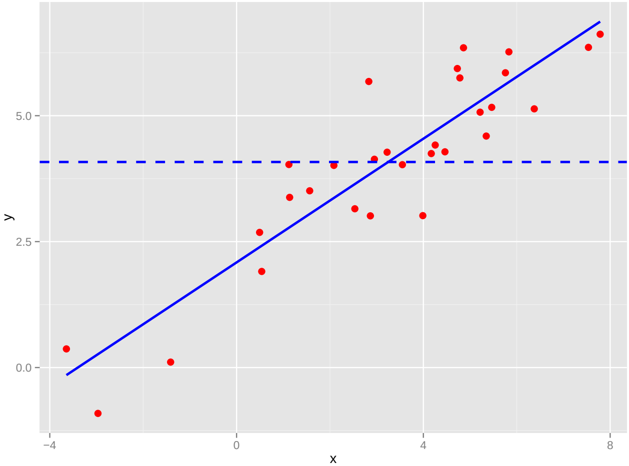
Analysis of variance (ANOVA) seeks to address goodness-of-fit by looking at these sample variances.

ANOVA

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

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

```
> linmod = lm(y ~ x, datat = 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
```

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

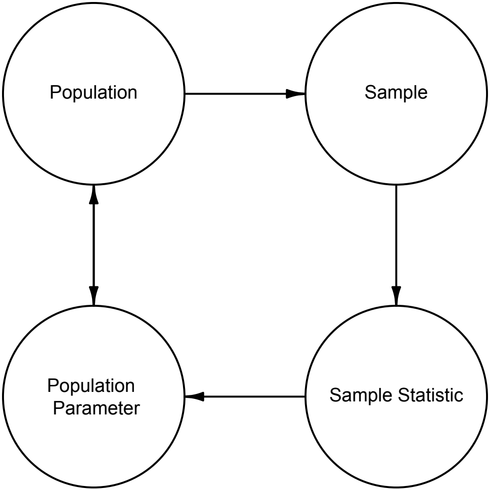
```
> names(summary(linmod))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
>
> summary(linmod)$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.0874344 0.22958105  9.092364 7.529711e-10
x           0.6139621 0.05415004 11.338166 5.611585e-12
>
> summary(linmod)$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 $\hat{\beta}_0, \hat{\beta}_1$

Estimates are unbiased:
$E(\hat{\beta}_0) =$

$E(\hat{\beta}_1) =$

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

Variances of estimates:

$Var(\hat{\beta}_0) =$

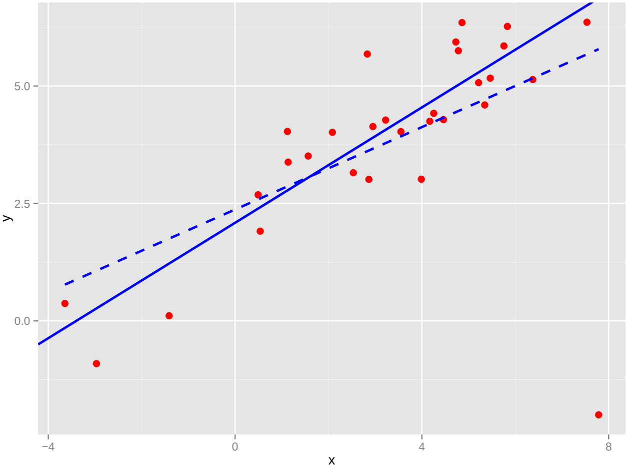
$Var(\hat{\beta}_1) =$

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

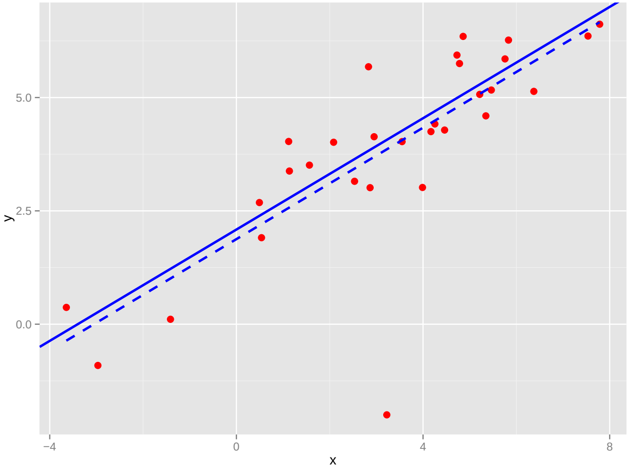
Note about the variance of β_1:

- Denominator contains $SS_x = \sum(x_i - \bar{x})^2$
- To decrease variance of β_1, increase variance of x

Effect of data on β_1



Effect of data on β_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

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

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

Matrix notation

- Let

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

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

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

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

Matrix notation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

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

Mean and Variance of a Random Vector

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

$$\begin{aligned} E(\boldsymbol{y})^T &= [E(y_1), \ldots, E(y_n)] \\ Var(\boldsymbol{y}) &= E\left[(\boldsymbol{y} - E\boldsymbol{y})(\boldsymbol{y} - E\boldsymbol{y})^T\right] = E(\boldsymbol{y}\boldsymbol{y}^T) - (E\boldsymbol{y})(E\boldsymbol{y})^T \end{aligned}$$

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

$$Cov(\boldsymbol{y}, \boldsymbol{z}) = E\left[(\boldsymbol{y} - E\boldsymbol{y})(\boldsymbol{z} - E\boldsymbol{z})^T\right]$$

Basics on Random Vectors

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

$$\begin{aligned} E(\boldsymbol{A}\boldsymbol{y}) &= \boldsymbol{A}E(\boldsymbol{y}) \\ Var(\boldsymbol{A}\boldsymbol{y}) &= \boldsymbol{A}Var(\boldsymbol{y})\boldsymbol{A}^T \\ Cov(\boldsymbol{A}\boldsymbol{y}, \boldsymbol{B}\boldsymbol{z}) &= \boldsymbol{A}Cov(\boldsymbol{y}, \boldsymbol{z})\boldsymbol{B}^T \end{aligned}$$

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