

Final concepts of SLR

Author: Nicholas G Reich, Jeff Goldsmith

*This material is part of the **statsTeachR** project*

Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported License: <http://creativecommons.org/licenses/by-sa/3.0/deed.en-US>

Today's lecture

- Simple Linear Regression Continued
 - sums of squares, R^2
 - ANOVA
 - centering
- Multiple Regression Intro

Simple linear regression model

- Observe data (y_i, x_i) for subjects $1, \dots, I$. Want to estimate β_0, β_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 needed for inference]

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:* $RSS := \sum_{i=1}^n \hat{\epsilon}_i^2$
- *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

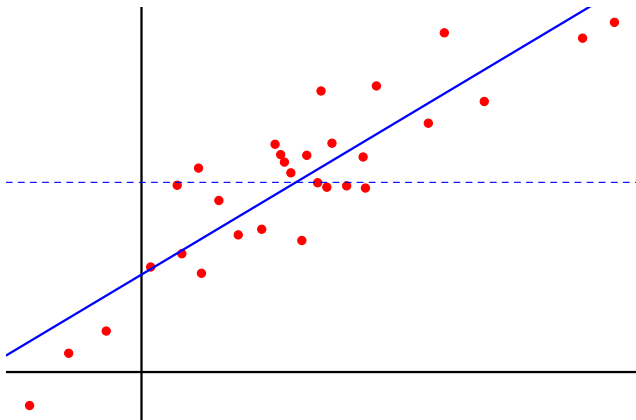
Analysis of variance (ANOVA) seeks to address goodness-of-fit by looking at these sample variances.

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 (more on this later)

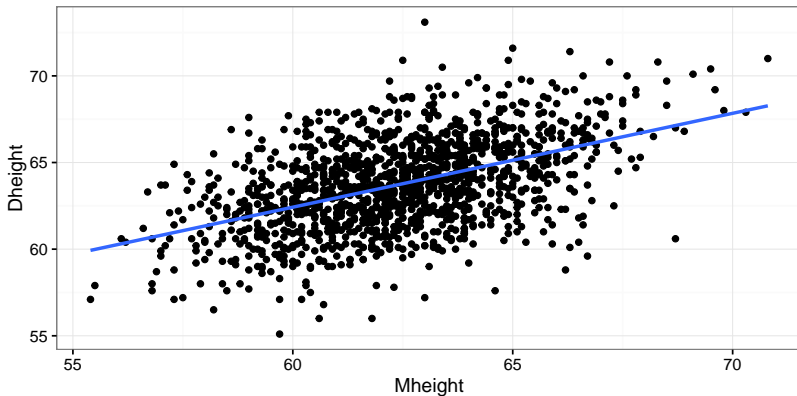
R example

```
library(alr3)
data(heights)
linmod <- lm(Dheight~Mheight, data=heights)
linmod

##
## Call:
## lm(formula = Dheight ~ Mheight, data = heights)
##
## Coefficients:
## (Intercept)      Mheight
##      29.9174      0.5417
```

R example

```
library(ggplot2)
theme_set(theme_bw())
ggplot(heights, aes(x=Mheight, y=Dheight)) + geom_point() +
  geom_smooth(method="lm", se=FALSE)
```



R example

```
summary(linmod)

##
## Call:
## lm(formula = Dheight ~ Mheight, data = heights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44  <2e-16 ***
## Mheight      0.54175    0.02596   20.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

R example

```
names(linmod)
```

```
## [1] "coefficients" "residuals"      "effects"        "rank"  
## [5] "fitted.values" "assign"         "qr"            "df.residual"  
## [9] "xlevels"      "call"          "terms"         "model"
```

R example

```
head(linmod$residuals)
```

```
##          1          2          3          4          5          6  
## -7.159733 -4.947113 -6.747306 -6.001480 -7.397402 -2.084396
```

```
head(resid(linmod))
```

```
##          1          2          3          4          5          6  
## -7.159733 -4.947113 -6.747306 -6.001480 -7.397402 -2.084396
```

```
head(linmod$fitted.values)
```

```
##          1          2          3          4          5          6  
## 62.25973 61.44711 62.74731 62.80148 63.39740 59.98440
```

```
head(fitted(linmod))
```

```
##          1          2          3          4          5          6  
## 62.25973 61.44711 62.74731 62.80148 63.39740 59.98440
```

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) 29.917437 1.62246940 18.43945 5.211879e-68  
## Mheight      0.541747 0.02596069 20.86797 3.216915e-84
```

```
summary(linmod)$r.squared
```

```
## [1] 0.2407957
```


R example

```
anova(linmod)

## Analysis of Variance Table
##
## Response: Dheight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Mheight      1 2236.7  2236.66  435.47 < 2.2e-16 ***
## Residuals 1373 7052.0     5.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R example

```
anova(linmod)

## Analysis of Variance Table
##
## Response: Dheight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Mheight      1 2236.7  2236.66  435.47 < 2.2e-16 ***
## Residuals 1373 7052.0     5.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(r2 <- 1-7052/(7052+2237))

## [1] 0.2408225
```

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:

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

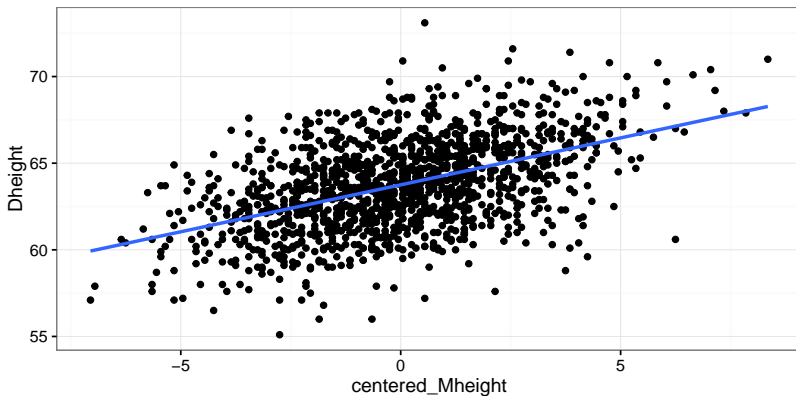
R example: centered xs

```
heights$centered_Mheight <- heights$Mheight - mean(heights$Mheight)
centered_linmod <- lm(Dheight ~ centered_Mheight, data=heights)
summary(centered_linmod)
```

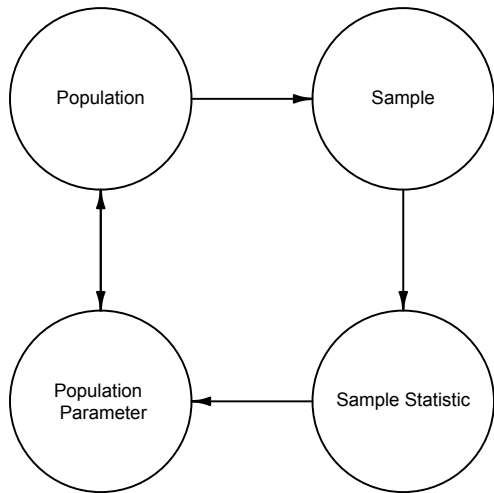
```
##
## Call:
## lm(formula = Dheight ~ centered_Mheight, data = heights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    63.75105    0.06112 1043.08  <2e-16 ***
## centered_Mheight  0.54175    0.02596   20.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

R example: centered xs

```
ggplot(heights, aes(x=centered_Mheight, y=Dheight)) + geom_point() +  
  geom_smooth(method="lm", se=FALSE)
```



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



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

Estimates are unbiased:

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

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

Variances of estimates

$$\text{Var}(\hat{\beta}_0) = \frac{\bar{x}\sigma^2}{\sum x^2}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}$$

where $SS_x = \sum (x - \bar{x})^2$

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 $\hat{\beta}_1$, increase variance of x

One slide on 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

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

- 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 $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$

Summary

Today's big ideas

- ▶ Simple linear regression definitions
- ▶ Properties of least squares estimates

Coming up soon

- ▶ More on MLR