Course Materials May Not Be Distributed or Posted Electronically

These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service.

© 2010–2017 Todd M. Gross

PSTAT 126
Regression Analysis

Dr. Todd Gross
Department of Statistics and Applied Probability
University of California, Santa Barbara

Lecture 4
Inference in Regression (con't)

Lecture Outline

- Review of Hypothesis Testing for β_1 (including R commands)

- Testing Regression Using Analysis of Variance

Example #1

- Research Question: Is GPA related to shoe size?
- Data:

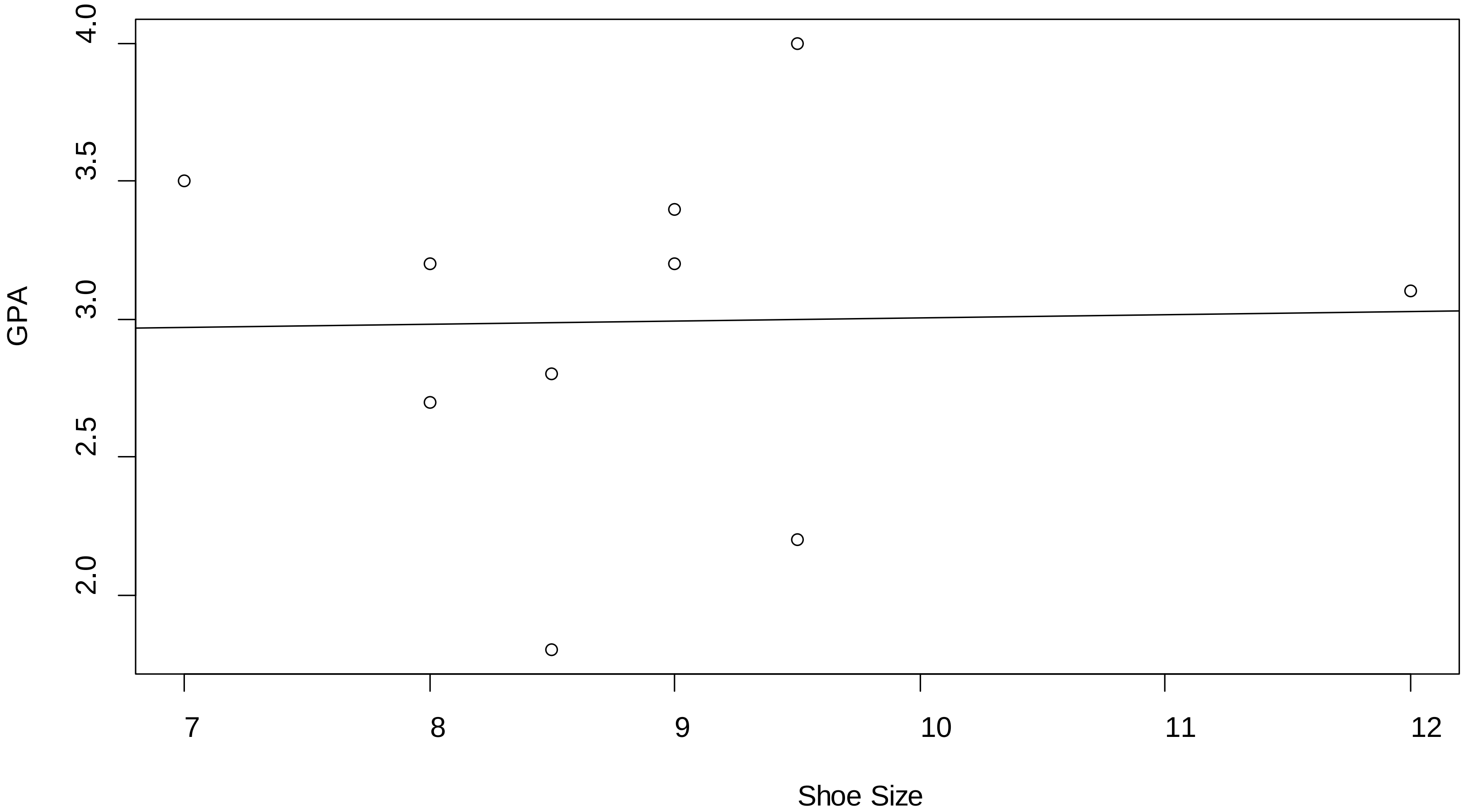
Shoe Size	7	8	8	8.5	8.5	9	9	9.5	9.5	12
GPA	3.5	2.7	3.2	1.8	2.8	3.2	3.4	4	2.2	3.1

- What question are we asking?
  - Is there a linear relationship between shoe size (x) and GPA (y)
  - $Y' = \beta_0 + \beta_1 X$
- What results do we need to answer research question?
  - Slope (b_1)

R Commands – Example #1

```
x<-c(7,8,8,8.5,8.5,9,9,9.5,9.5,12)
y<-c(3.5,2.7,3.2,1.8,2.8,3.2,3.4,4,2.2,3.1)
model1<-lm(y~x)
plot(x,y,xlab="Shoe Size",ylab="GPA")
abline(model1)
summary(model1)
```

R Plot – Example #1



R Output – Example #1

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1852 -0.2557  0.1409  0.3618  1.0028

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.88365    1.53447   1.879    0.097 .
x            0.01195    0.17071   0.070    0.946
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6807 on 8 degrees of freedom
Multiple R-squared:  0.0006121,      Adjusted R-squared:  -0.1243
F-statistic: 0.0049 on 1 and 8 DF,  p-value: 0.9459
```

t-test for the slope

Hypothesis Test – Example #1

- Is a linear relationship between Shoe Size (x) and GPA (Y)?
  - We want to know if $\beta_1 \neq 0$
  - Is b_1 (the sample result) big enough to conclude that $B_1 \neq 0$?
  - The null hypothesis is that there is NO linear relationship.

- H_O: $\beta_1 = 0$ vs. H_1:$\beta_1 \neq 0$
- $t^* = 0.07$, $p=0.946$
- Because $p > 0.05$, FAIL TO REJECT H_0
- “There is NOT sufficient evidence to conclude that there is a relationship between Shoe Size and GPA”

Example #2

- Research Question: Is Hours of Study per Week related to Units Taken?
- Data:

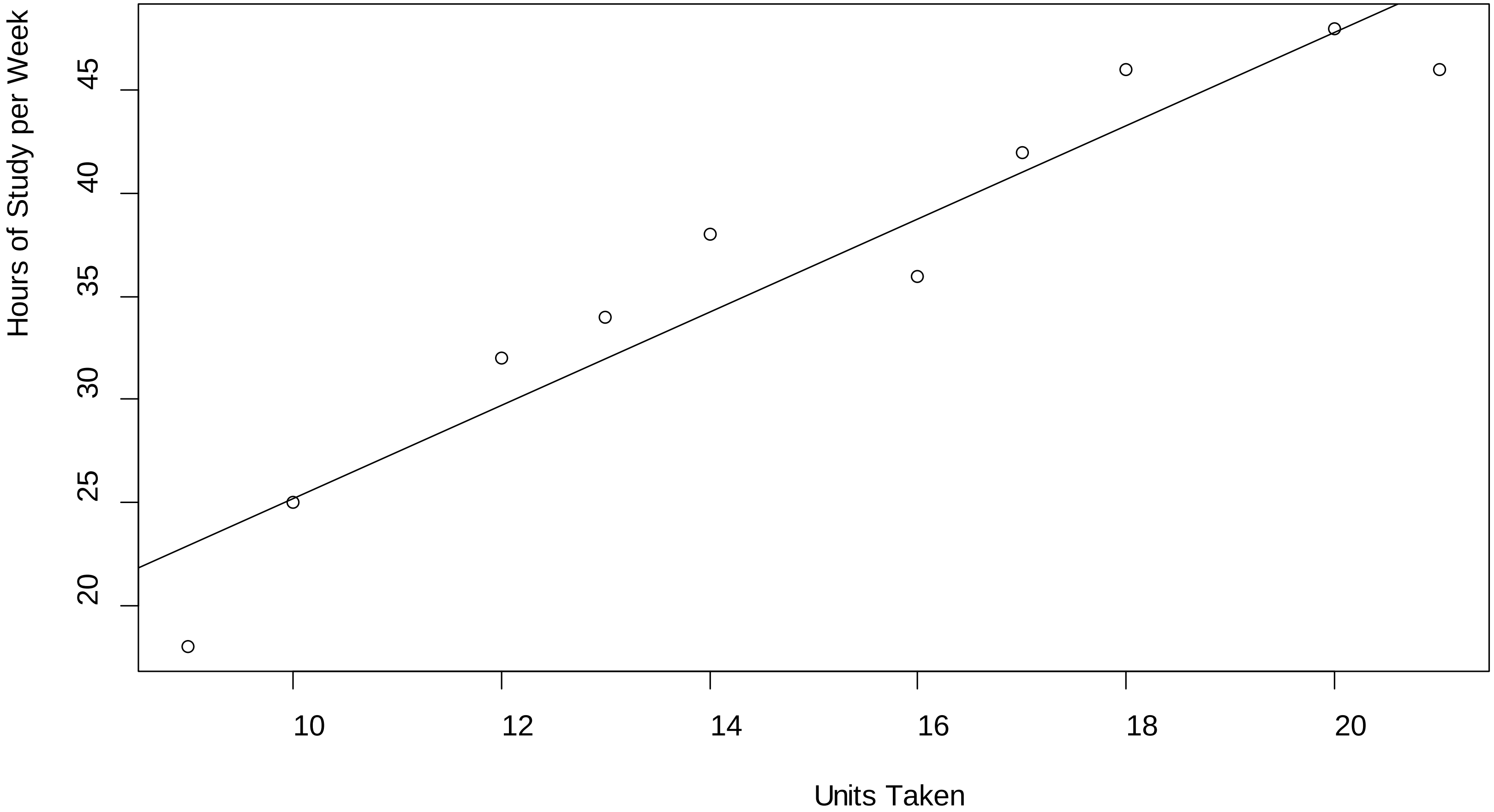
Units Taken	9	10	12	13	14	16	17	18	20	21
Hours of Study	18	25	32	34	38	36	42	46	48	46

- What question are we asking?
  - Is there a linear relationship between Units Taken (x) and Hours of Study per Week (y)
  - $Y' = \beta_0 + \beta_1 X$
- What results do we use to answer it?
  - Slope (b_1)

R Commands - Example #2

```
x<-c(9,10,12,13,14,16,17,18,20,21)
y<-c(18,25,32,34,38,36,42,46,48,46)
model2<-lm(y~x)
plot(x,y,xlab="Units Taken",ylab="Hours of Study per Week")
abline(model2)
summary(model2)
```

R Plot - Example #2



R Output - Example #2

```
Call:
lm(formula = y ~ x)

Residuals:
   Min      1Q  Median      3Q     Max 
-4.940 -2.120   0.590   2.215   3.760 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.6000     4.0090   0.649    0.535    
x             2.2600     0.2588   8.733 2.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.169 on 8 degrees of freedom
Multiple R-squared:  0.9051,	Adjusted R-squared:  0.8932 
F-statistic: 76.27 on 1 and 8 DF,  p-value: 2.311e-05
```

t-test for the slope

Hypothesis Test – Example #2

- Is a linear relationship between Units Taken (x) and Hours of Study per Week (Y)?
  - We want to know if $\beta_1 \neq 0$
  - Is b_1 (the sample result) big enough to conclude that $\beta_1 \neq 0$?
  - The null hypothesis is that there is NO linear relationship.

- H_O: $\beta_1 = 0$ vs. H_1:$\beta_1 \neq 0$
- $t^* = 8.733$, $p=0.00002$
- Because $p < 0.05$, REJECT H_0
- "There IS sufficient evidence to conclude that there is a relationship between Units Taken and Weekly Study Hours"

Example #2 – Confidence Interval

- The University expects at least 3 hours of study per week for each unit taken.
- Do the data support the University's claim?
- Calculate a confidence interval for B_1

```
> confint(model2,level=.95)
                2.5 %    97.5 %
(Intercept) -6.644747 11.844747
x            1.663254  2.856746
```

- "We are 95% confident that students study LESS than 3 hours per week for each unit taken."

Testing Regression Using Analysis of Variance

ANOVA Hypothesis Test

- We can test the same hypothesis for slope using Analysis of Variance (ANOVA).
  - The ANOVA will yield the same conclusion as the t-test
- However, ANOVA will be much more useful when we move to multiple regression in the 2^{nd} half of the course

Consider the hypothesis

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

Analysis of Variance

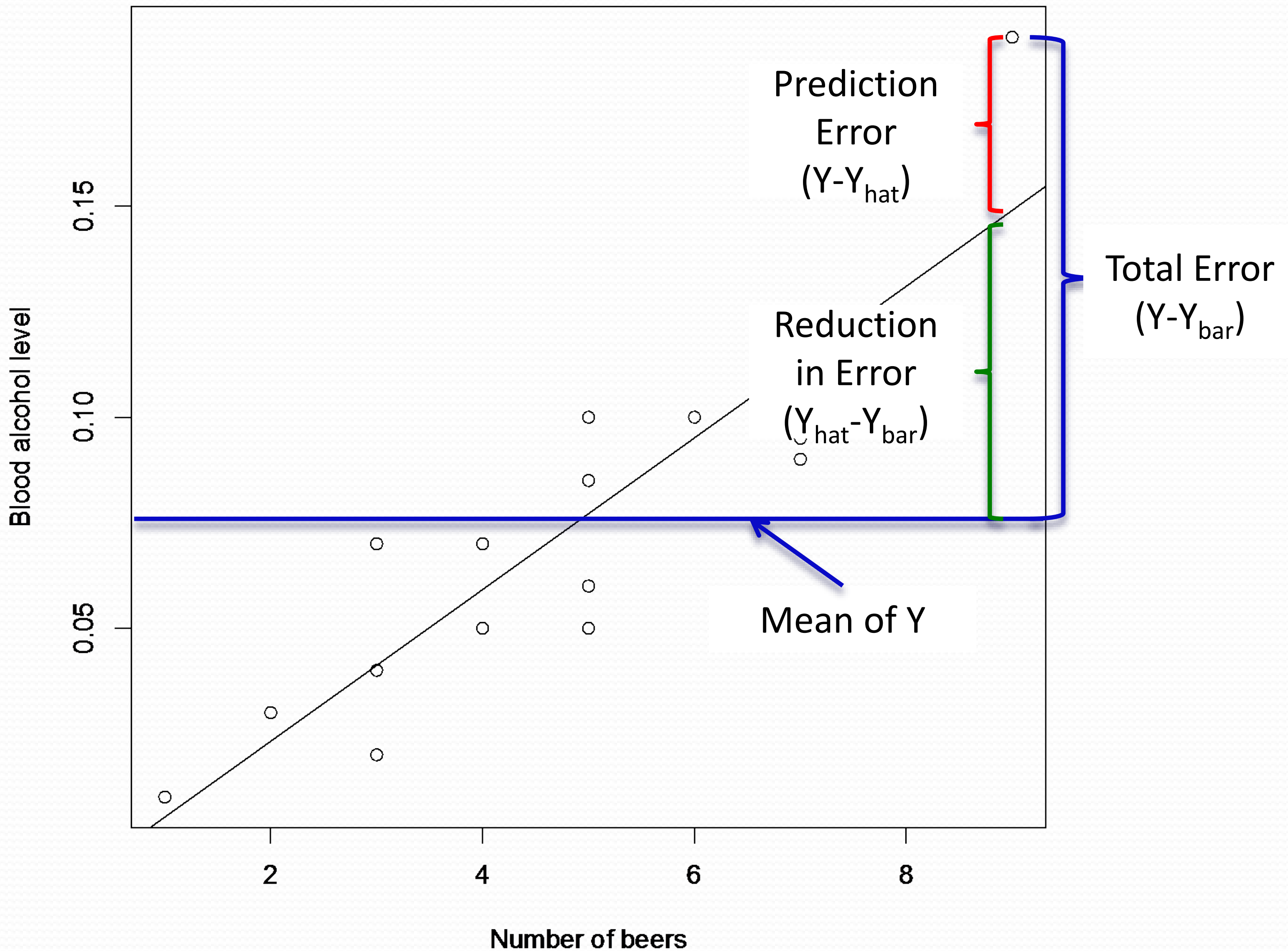
- Analysis of Variance (ANOVA) is an alternative method of testing hypotheses

- ANOVA is performed by conducting an F test (similar to a t-test)

$$F = \frac{Variance(Effect)}{Variance(Error)} = \frac{MSR}{MSE}$$

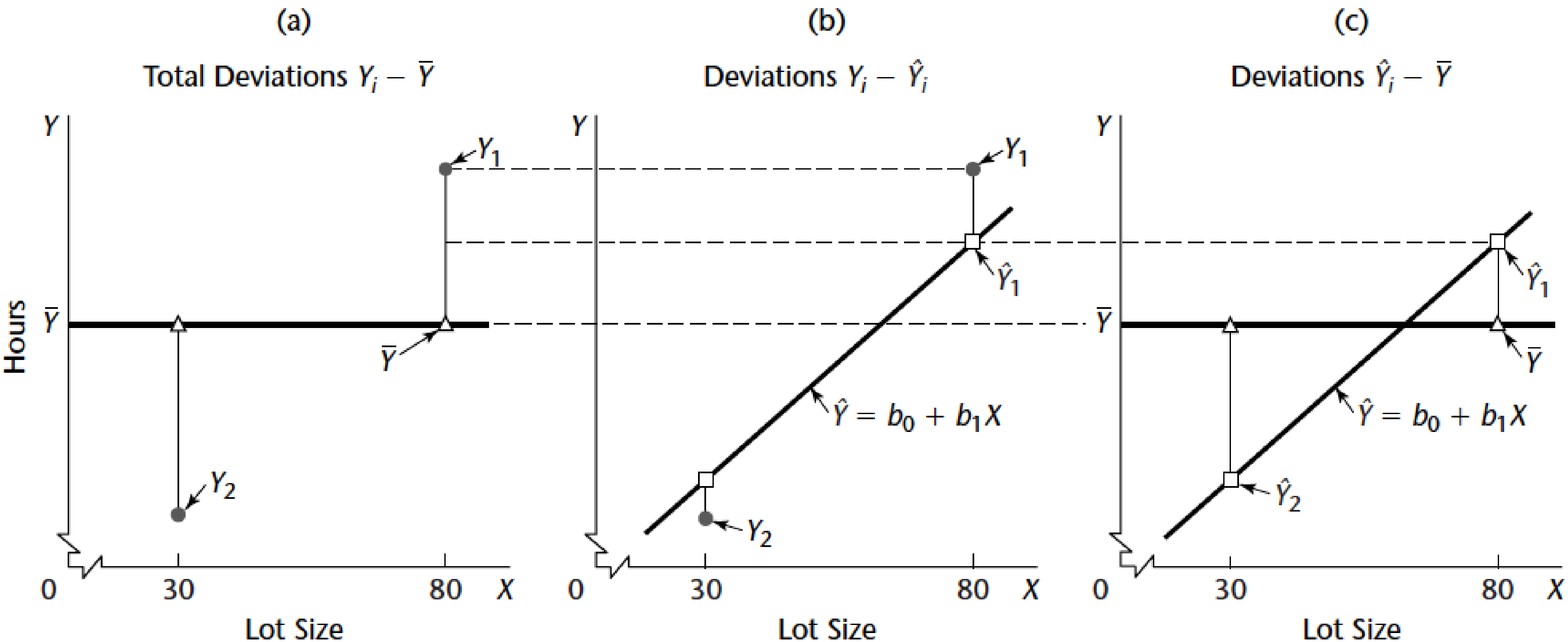
- A stronger relationship between X and Y tends to increase the numerator, resulting in a larger value of F

Partitioning Deviations



Partitioning Deviations (Y – Ybar)

FIGURE 2.7 Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$—Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).



Partitioning Sum of Squared Deviations

- Partitioning the Total Deviation

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

- Partitioning the Sum of Squared Deviations

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

- Sum of Squares (SS) Notation

  SSTotal = SSRegression + SSError

- Partitioning Degrees of Freedom

  $df_T = df_R + df_E$

  $(n – 1) = 1 + (n - 2)$

Partitioning SS (proof)

The following equality always holds:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

So we have

$$\begin{aligned}
& \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \\
= \;& \sum_{i=1}^{n} \{(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\}^2 \\
= \;& \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})\}^2 + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
= \;& \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2
\end{aligned}$$

Proof (con't)

We used the fact that

$$
\begin{aligned}
& \sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
= \;& \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\bar{Y} \\
= \;& \sum_{i=1}^{n} e_i \hat{Y}_i - \bar{Y}\sum_{i=1}^{n} e_i \\
= \;& 0
\end{aligned}
$$

Partitioning Sum of Squared Deviations (SS)

- The term $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is called the `total sum of squares` (SSTO)
- The term $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ is the `regression sum of squares` (also called `explained sum of squares`) (SSR)
- The term $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is the `error sum of squares` (also called `residual sum of squares`) (SSE)
- The above equation decomposes SSTO into two parts: explained by the linear regression model and unexplained:

$$\text{SSTO} = \text{SSR} + \text{SSE}$$

The Logic of ANOVA for Regression

- SSR is the sum of squares due to regression. So large SSR provide evidence against H_0
- How large is large? Magnitude of SSR is not enough because it depends on scale. We want to use a relative quantity
- The F statistic

$$F^* = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}}$$

- From the ANOVA table, F^* is close to 1 under H_0. Thus a value much larger than 1 provide evidence against H_0
- Under H_0, $F^* \sim F(1, n-2)$
- Reject H_0 if $F^* > F(1-\alpha; 1, n-2)$
- $F^* = (t^*)^2$, thus F test and t test are equivalent for the simple linear regression

ANOVA Source Table

Source	SS	df	MS	F*	P-value
Model	SSR	1	MSR = SSR/1	MSR/MSE	From Statistical Table
Error	SSE	n-2	MSE = SSE/(n-2)	---	
Total	SSTO	n-1			

ANOVA (F statistic)

- Mean Squared Deviation (aka Variance) is the Sum of Squared Deviations divided by degrees of freedom

$$MSR = \frac{SSR}{1} = SSR \qquad MSE = \frac{SSE}{n-2}$$

- The ratio of two variances is distributed as F

$$F^* = \frac{MSR}{MSE}$$

- F* is compared to F(0.05,dfR/dfE)

ANOVA - R

```
> anova(model2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x          1 766.14  766.14  76.271 2.311e-05 ***
Residuals  8  80.36   10.04                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $F(1,8) = 76.271$, $p < 0.05$, therefore REJECT H_O
- "There is sufficient evidence to conclude that there is a positive relationship between number of units taken and hours of study per week."

ANOVA Source Table Based on R Output

Source	SS	df	MS	F*	p-value
Model	766.14	1	766.14	76.271	p<0.0001
Error	80.36	8	10.04	---	
Total	846.5	9	---		

- $F(1,8) = 76.271$, $p < 0.05$, therefore REJECT H_O
- "There is sufficient evidence to conclude that there is a positive relationship between number of units taken and hours of study per week."