Comparing Two Regression Lines

READING:	3.3
EXERCISES:	CH 3. 30,31,48
ASSIGNED:	HW 8
PRODUCER:	DR. MARIO

Example: Lego Sets

- Goal: *We want to use a linear regression model to predict the Amazon price of a Lego set based off the theme.*

```
library(mosaic)

lego = read.csv("lego.csv")
lego_2theme = subset(lego, Theme == "Star Wars"| Theme == "Friends")
lego_2theme = lego_2theme[,c("Theme","Pieces","Amazon_Price")]
str(lego_2theme)
```

```
## 'data.frame':    222 obs. of  3 variables:
##  $ Theme       : chr  "Friends" "Friends" "Friends" "Friends" ...
##  $ Pieces      : int  95 85 93 50 97 57 86 92 72 85 ...
##  $ Amazon_Price: num  7.71 7.99 7.99 8.99 8.99 ...
```

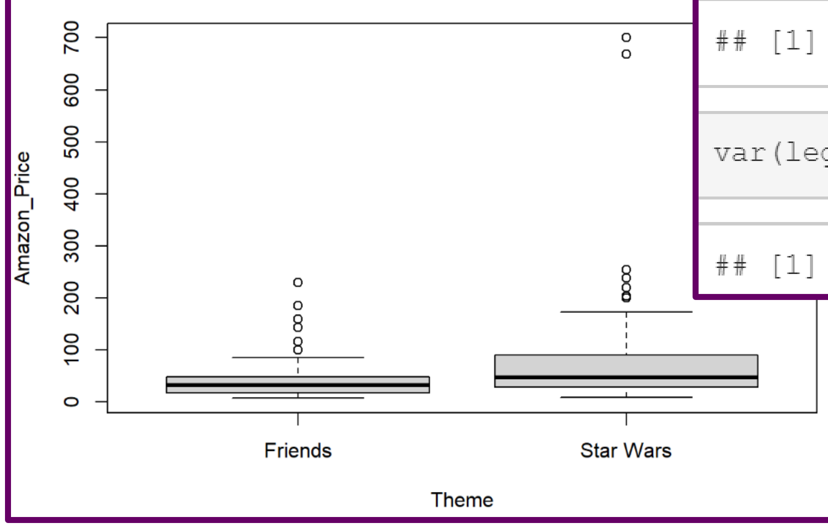
Example: Lego Sets

- Use Classic t-Test for Difference in Population Means

μ_1 = Average Price of Friends Theme $\quad$ μ_2 = Average Price of Star Wars Theme

- Assume Variance of Amazon Price of Both Groups are Equal

```
boxplot(Amazon_Price ~ Theme, data=lego_2theme)
```




```
var(lego_2theme$Amazon_Price[which(lego_2theme$Theme=="Friends")],na.rm=T)
## [1] 1442.47
var(lego_2theme$Amazon_Price[which(lego_2theme$Theme=="Star Wars")],na.rm=T)
## [1] 11011.63
```

Seems Like a Bad Assumption

Review of Pooled t-Test

- Use Classic t-Test for Difference in Population Means

$$\mu_1 = \text{Average Price of Friends Theme}$$
$$\mu_2 = \text{Average Price of Star Wars Theme}$$

- Pooled t-Test (Assuming Variances are Equal)

Hypotheses

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

Test Statistic

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{1/n_1 + 1/n_2}}$$

P-value

Use t-Distribution with $n_1 + n_2 - 2$ d.f.

Pooled Standard Deviation

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Example: Lego Sets

- Results from Pooled t-Test

```
t.test(Amazon_Price ~ Theme,var.equal=TRUE, data=lego_2theme)
```

```
##
##  Two Sample t-test
##
## data:  Amazon_Price by Theme
## t = -3.2299, df = 181, p-value = 0.001471
## alternative hypothesis: true difference in means between g
## 95 percent confidence interval:
##  -61.92623 -14.95793
## sample estimates:
##   mean in group Friends mean in group Star Wars
##                42.26448                 80.70656
```

Example: Lego Sets

- Results from Linear Regression

```
mod = lm(Amazon_Price ~ Theme, data=lego_2theme)
summary(mod)
```

```
##
## Call:
## lm(formula = Amazon_Price ~ Theme, data = lego_2theme)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -71.72 -35.91 -17.77   7.67 619.24
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        42.26       8.62   4.903 2.09e-06 ***
## ThemeStar Wars     38.44      11.90   3.230  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.4 on 181 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.0545, Adjusted R-squared:  0.04927
## F-statistic: 10.43 on 1 and 181 DF,  p-value: 0.001471
```

```
confint(mod)
```

```
##                    2.5 %   97.5 %
## (Intercept)     25.25524 59.27372
## ThemeStar Wars  14.95793 61.92623
```

Equivalent Results Because of the Homoscedastic Assumption in Linear Regression

Indicator Variables

- We Cannot Fit the Model Below if X is **Categorical**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- We Must Recode X to be **Numeric**

- Recoding X as an **Indicator Variable** (Dummy Variable)

$$X = \begin{cases} 0 & \text{if Theme} = \text{Friends} \\ 1 & \text{if Theme} = \text{Star Wars} \end{cases}$$

Indicator Variables

- Linear Regression Model is **Two Horizontal Lines** or **Two Dots**

$$\hat{\mu}_{Friends} = \widehat{\beta_0} = 42.26$$

$$\hat{\mu}_{Star\ Wars} = \widehat{\beta_0} + \widehat{\beta_1} = 42.26 + 38.44 = 80.7$$

- Slope is the Change in Predicted Y if We Switch from X = 0 to X=1

- t-Test: *If the indicator variable X is **not significant**, then we **don't** have evidence that there is a **difference** in the average amazon price between the two themes Friends and Star Wars.*

Indicator Variables

- The Two Models Have the Same Error Term

$$Y = \beta_0 + \beta_1(0) + \epsilon = \beta_0 + \epsilon$$

$$Y = \beta_0 + \beta_1(1) + \epsilon = (\beta_0 + \beta_1) + \epsilon$$

- Standard Error of the Regression is the Estimated Standard Deviation of that Error Term

- Homoscedasticity Assumption for Lego Model Led to $\hat{\sigma}_\epsilon = 80.4$

Indicator Variables

- Confidence Intervals
  - CI for Intercept Represents Where We Believe the Average of Y to be for the Group Recoded as 0
  - CI for Slope Represents What We Believe the Difference to be Between the Average Y for Group 1 and Average Y for Group 0

```
confint(mod)
```

```
##                     2.5 %   97.5 %
## (Intercept)      25.25524 59.27372
## ThemeStar Wars   14.95793 61.92623
```

Linear Regression with Indicator

- Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Variables Y and X_1 are Numeric (Continuous)
- Variable X_2 is Numeric (Binary) Based Off **Categorical** Variable

Linear Regression with Indicator

- Parallel Lines

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(0) + \epsilon = \beta_0 + \beta_1 X_1 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon$$

- Same Slopes But Different Y-Intercepts

Example: Lego Sets

```
mod2 = lm(Amazon_Price ~Pieces + Theme, data= lego_2theme)
summary(mod2)
```

```
##
## Call:
## lm(formula = Amazon_Price ~ Pieces + Theme, data = lego_2theme)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -173.463  -17.904   -2.446   11.628  255.582
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.082759   4.258263  -0.019   0.9845
## Pieces          0.143271   0.005468  26.203   <2e-16 ***
## ThemeStar Wars  9.803090   5.548089   1.767   0.0789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.75 on 180 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.8036, Adjusted R-squared:  0.8014
## F-statistic: 368.3 on 2 and 180 DF,  p-value: < 2.2e-16
```

```
confint(mod2)
```

```
##                     2.5 %     97.5 %
## (Intercept)    -8.4852953  8.3197764
## Pieces          0.1324816  0.1540601
## ThemeStar Wars -1.1445714 20.7507510
```

What Did We Learn About Theme?

Example: Lego Sets

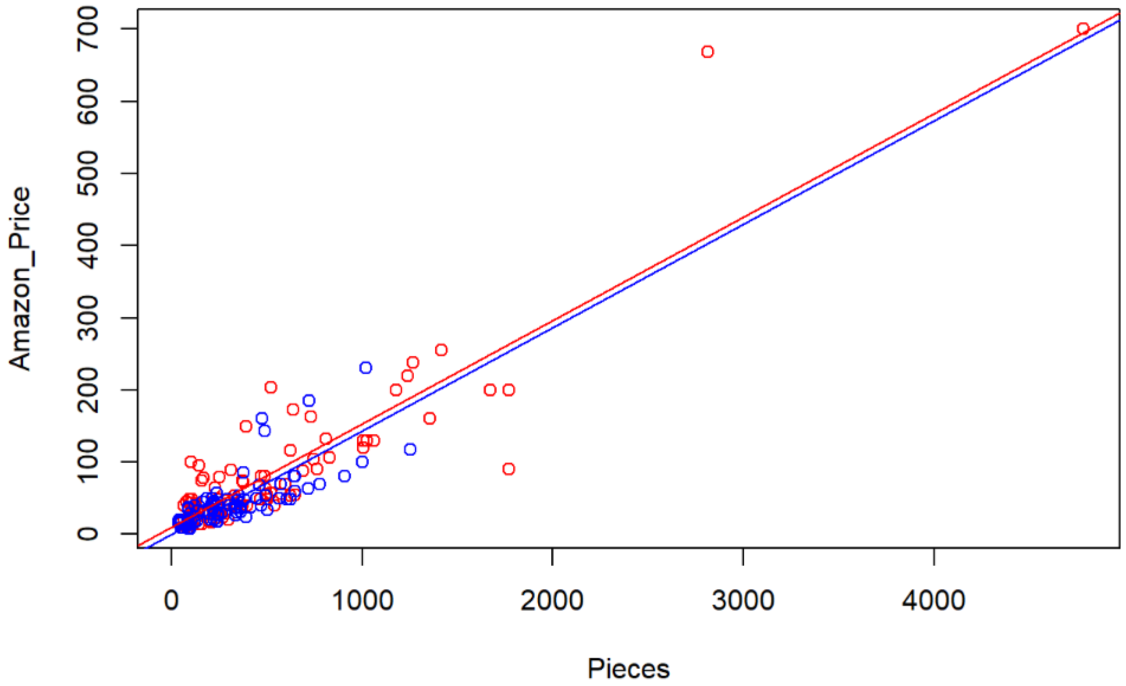
```
plot(Amazon_Price~Pieces, col="red",
     data=subset(lego_2theme,Theme=='Star Wars'))

points(Amazon_Price~Pieces, col="blue",
       data=subset(lego_2theme,Theme=='Friends'))

B_Int = summary(mod2)$coef[1,1]
B_Pieces = summary(mod2)$coef[2,1]
B_Theme = summary(mod2)$coef[3,1]

abline(
  B_Int,
  B_Pieces,
  col = "blue",
  )

abline(
  B_Int + B_Theme,
  B_Pieces,
  col = "red",
  )
```



Are You Surprised?

Linear Regression with Indicator

- Model with **Interaction Variable**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \epsilon$$

- Not Parallel Lines and Still Same Error Term (Homoscedasticity)
  - Model for Group 0

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

  - Model for Group 1

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon$$

Example: Lego Sets

```
mod3 = lm(Amazon_Price ~Pieces + Theme + Pieces*Theme, data= lego_2theme)
summary(mod3)
```

```
##
## Call:
## lm(formula = Amazon_Price ~ Pieces + Theme + Pieces * Theme,
##     data = lego_2theme)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -179.043  -17.467   -3.589    8.259  245.448
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.06395    5.99060   1.513    0.132
## Pieces                  0.11233    0.01538   7.302 8.96e-12 ***
## ThemeStar Wars         -1.51114    7.60839  -0.199    0.843
## Pieces:ThemeStar Wars   0.03532    0.01643   2.149    0.033 *
```

If an Interaction Variable is Significant, Don't Remove the Individual Variables No Matter if They are Significant

Example: Lego Sets

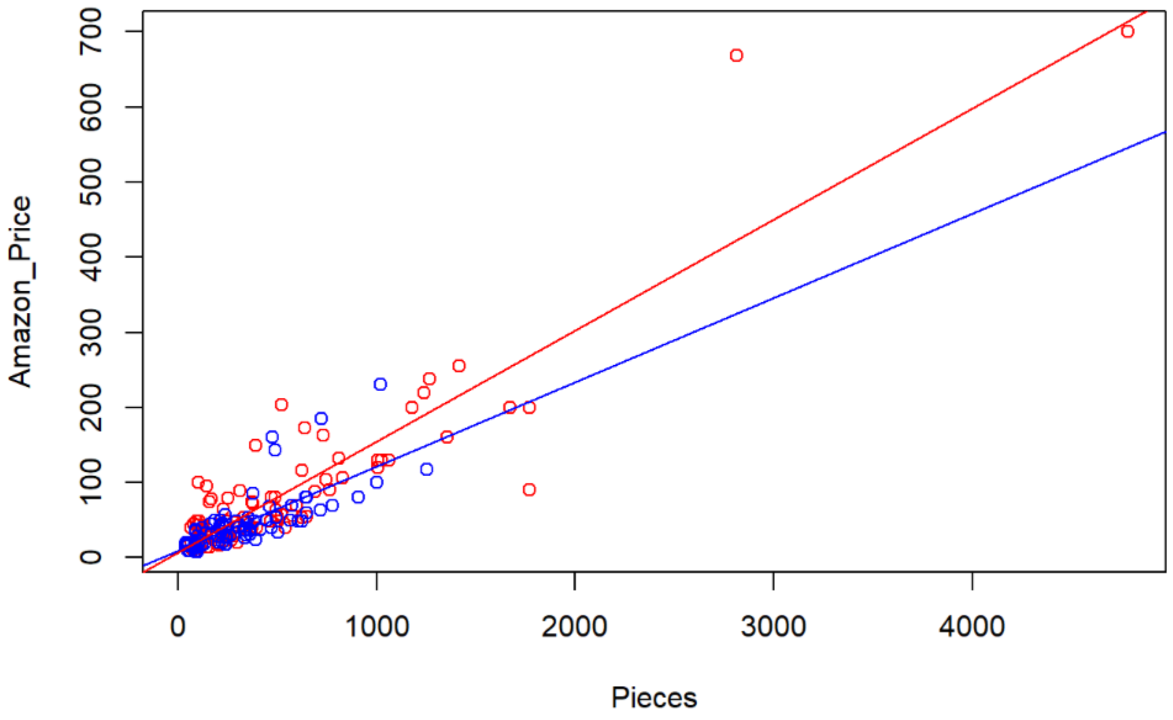
```
plot(Amazon_Price~Pieces, col="red",
     data=subset(lego_2theme,Theme=='Star Wars'))

points(Amazon_Price~Pieces, col="blue",
       data=subset(lego_2theme,Theme=='Friends'))

B_Int = summary(mod3)$coef[1,1]
B_Pieces = summary(mod3)$coef[2,1]
B_Theme = summary(mod3)$coef[3,1]
B_PiecesXTheme = summary(mod3)$coef[4,1]

abline(
  B_Int,
  B_Pieces,
  col = "blue",
  )

abline(
  B_Int + B_Theme,
  B_Pieces + B_PiecesXTheme,
  col = "red",
  )
```



Lines Have Different Slopes and Intercepts

Conclusion

- Textbook Splits Dataset Into Two Groups and Fits Separate Regression Lines to Each of the Groups
- What is Better?
  - Two Separate Models Fitted to Two Separate Datasets
  - One Model with Interaction Fitted to Full Dataset
- Read Textbook to See Checking of Assumptions From Residuals

Thank You

Make Reasonable Decisions