

STAT 500: Fall 2018

MLR with categorical predictors

Motivating Example

- Students in STAT 101 at Iowa State University were asked in a recent semester to provide demographic data for use during the semester. A random sample of 75 students were selected and a few of the variables collected were:
 - Height (in inches)
 - Height of the student's ideal romantic partner (in inches)

SLR for Motivating Example

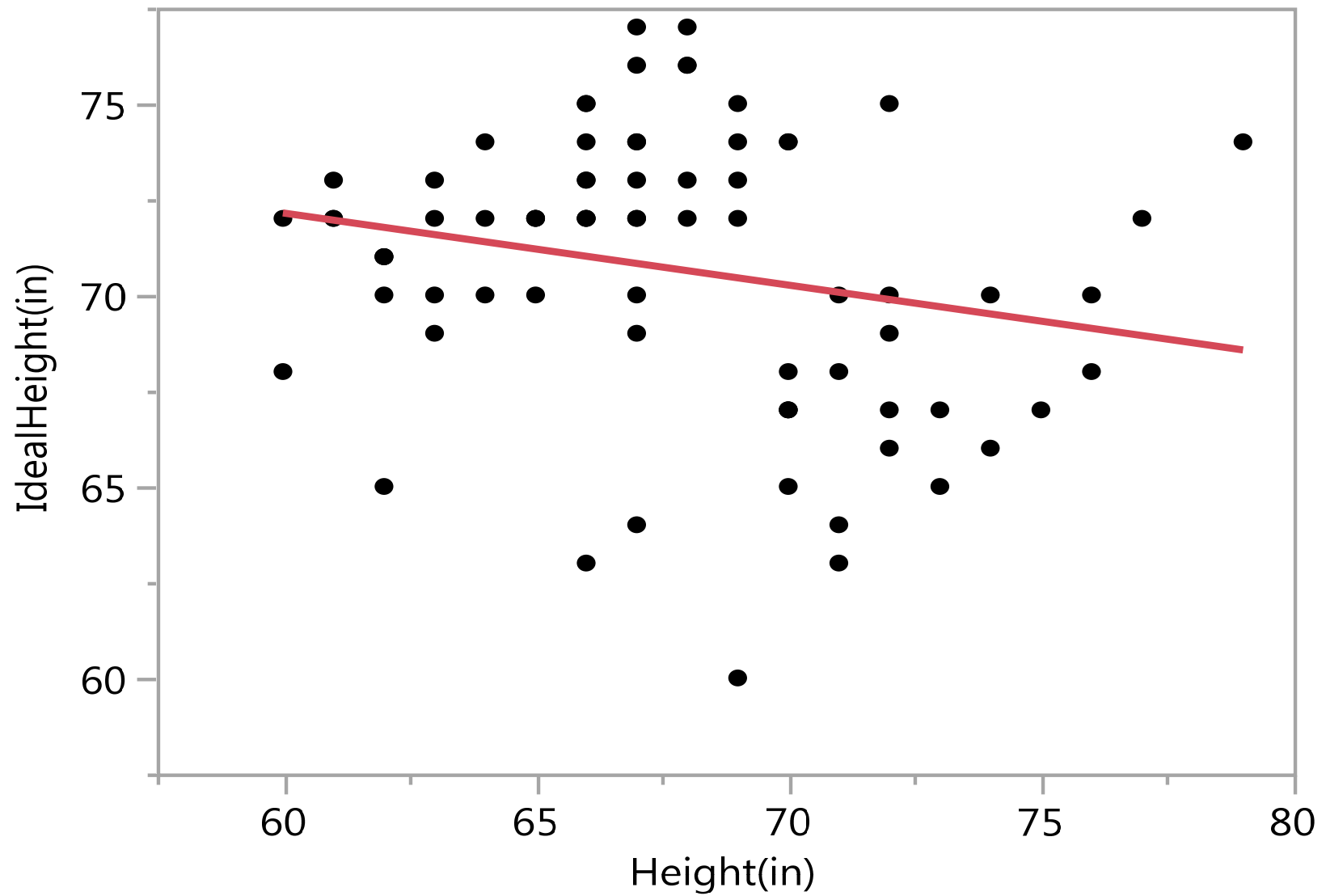
Model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

- Y = height of student's ideal romantic partner
- x = height of student

Scatterplot



SLR Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	47.08523	47.0852	3.6925	0.0586
Error	73	930.86144	12.7515		
C. Total	74	977.94667			

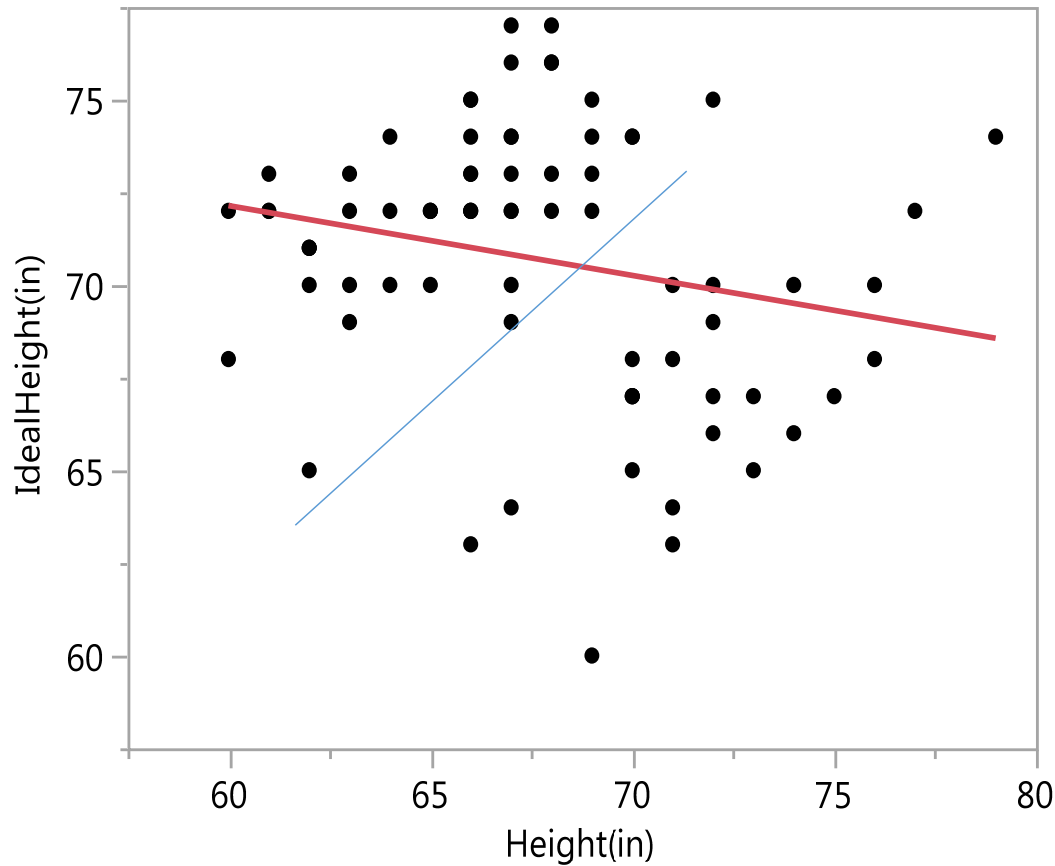
Root MSE	3.570928	R-Square	0.04815
Dependent Mean	70.69333	Adj R-Sq	0.03511

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	83.440038	6.646211	12.55	<0.0001
Height(in)	1	-0.188301	0.097992	-1.92	0.0586

Summary of SLR Model

- Linear relationship between student's height and the height of their ideal romantic partner is very weak and negative.
- $R^2 = 4.815\%$; Only 4.815% of the variation in the height of a student's ideal romantic partner can be explained by the simple linear regression with the student's height.
- Student's height is statistically significant at the 10% level, but not the 5% level.

What's going on?



- Two clusters of points in scatterplot.
 - To left of blue line
 - To right of blue line
- What are those clusters related to?

Multiple Linear Regression Model with Categorical Explanatory Variable

- Variable for student's Gender should be added to the model.
- x_{i1} = student's height
- $x_{i2} = \begin{cases} 1 & \text{if student is female} \\ 0 & \text{if student is male} \end{cases}$

Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

Two Models

- If student is female ($x_{i2} = 1$)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \varepsilon_i = (\beta_0 + \beta_2) + \beta_1 x_{i1} + \varepsilon_i$$

- If student is male ($x_{i2} = 0$)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

Comparison of Two Models

- Same Slope (β_1)
 - Assumes relationship between height of ideal romantic partner (response variable Y) and student's height (quantitative explanatory variable x_1) is the same for both groups in the categorical explanatory variable x_2 (Gender = female and Gender = male).
- Different Intercept
 - For females: $\beta_0 + \beta_2$
 - For males: β_0

MLR Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	675.04314	337.522	80.2287	<0.0001
Error	72	302.90353	4.207		
C. Total	74	977.94667			

Root MSE	2.051096	R-Square	0.69027
Dependent Mean	70.69333	Adj R-Sq	0.68166

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	27.664081	5.951059	4.65	<.0001*
Height(in)	1	0.5471394	0.082411	6.64	<.0001*
Gender(1/0)	1	8.9873486	0.735618	12.22	<.0001*

MLR Model

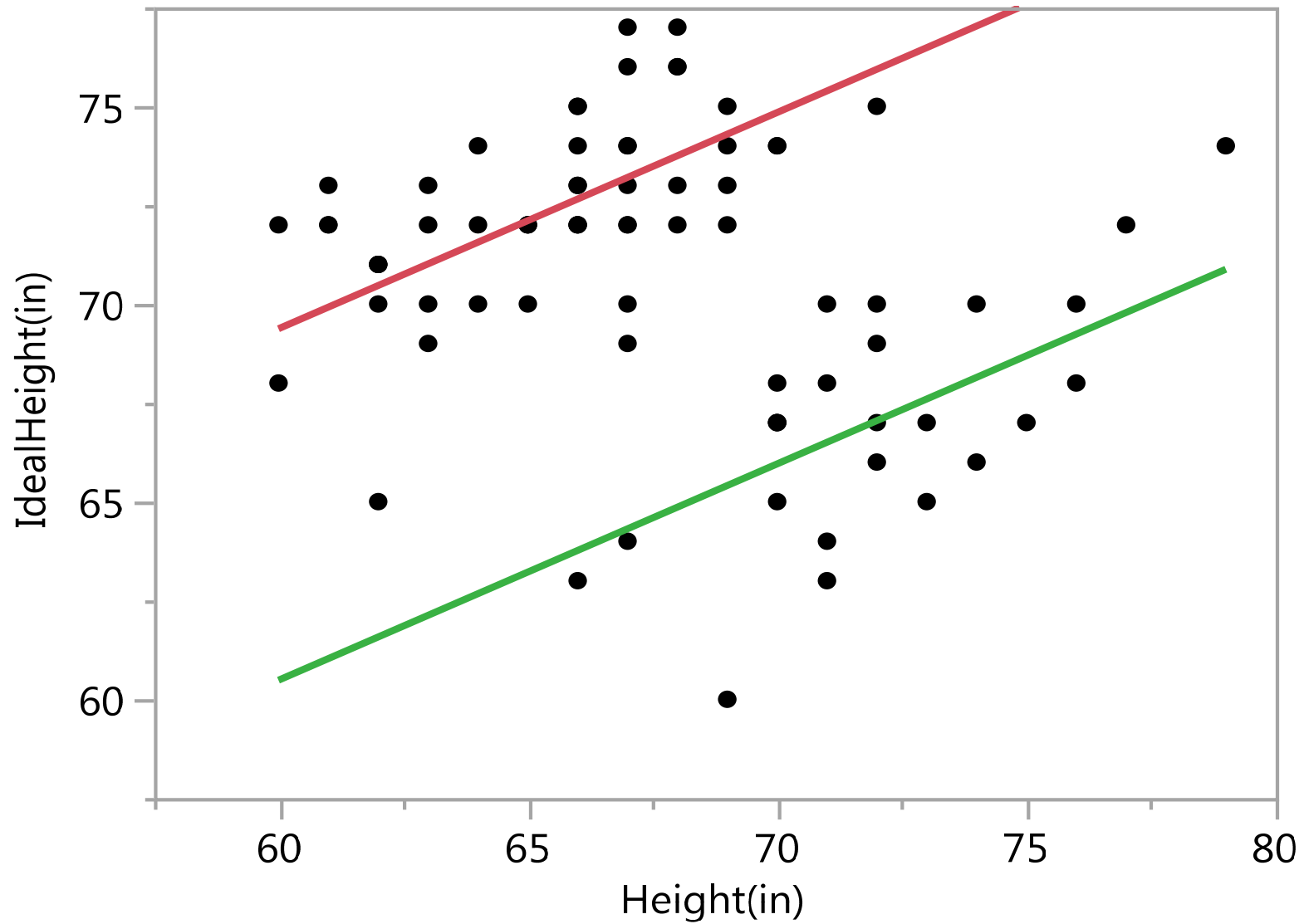
- Model is highly significant in explaining the height of the students' ideal romantic partner (p-value < 0.0001)
- $R^2 = 69.03\%$
 - 69.03% of the variation in the height of the students' ideal romantic partner can be explained by the multiple linear regression model with height and Gender of the student.
- Given height in the model, Gender is highly significant (p-value < 0.0001)
- Given Gender in the model, height is highly significant (p-value < 0.0001)

Two Estimated Models

- Using the parameter estimates from the MLR, we can determine the estimated intercept and slope of the two models – one for females and one for males. The two models are:

$$\hat{Y}_i = \begin{cases} 36.651 + 0.547x_i & \text{— for females} \\ 27.664 + 0.547x_i & \text{— for males} \end{cases}$$

Scatterplot



Interaction Model

- We can consider a model with an interaction term between x_1 and x_2 - height and Gender.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

- This model will allow for a different relationship between students' height and the height of their ideal romantic partner.

Two Models

- If student is female ($x_{i2} = 1$)

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1} + \varepsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} + \varepsilon_i \end{aligned}$$

- If student is male ($x_{i2} = 0$)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

Comparison of Two Models

- Different Slope
 - For females: $\beta_1 + \beta_3$
 - For males: β_1
- Different Intercept
 - For females: $\beta_0 + \beta_2$
 - For males: β_0

MLR Model with Interaction

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	684.77579	228.259	55.2796	<0.0001
Error	71	293.17088	4.129		
C. Total	74	977.94667			

Root MSE	2.032035	R-Square	0.70022
Dependent Mean	70.69333	Adj R-Sq	0.68755

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	15.451772	9.90122	1.56	0.1231
Height(in)	1	0.7166606	0.137325	5.22	<.0001*
Gender(1/0)	1	27.272359	11.93225	2.29	0.0253*
Interaction	1	-0.262206	0.170788	-1.54	0.1292

MLR Model with Interaction

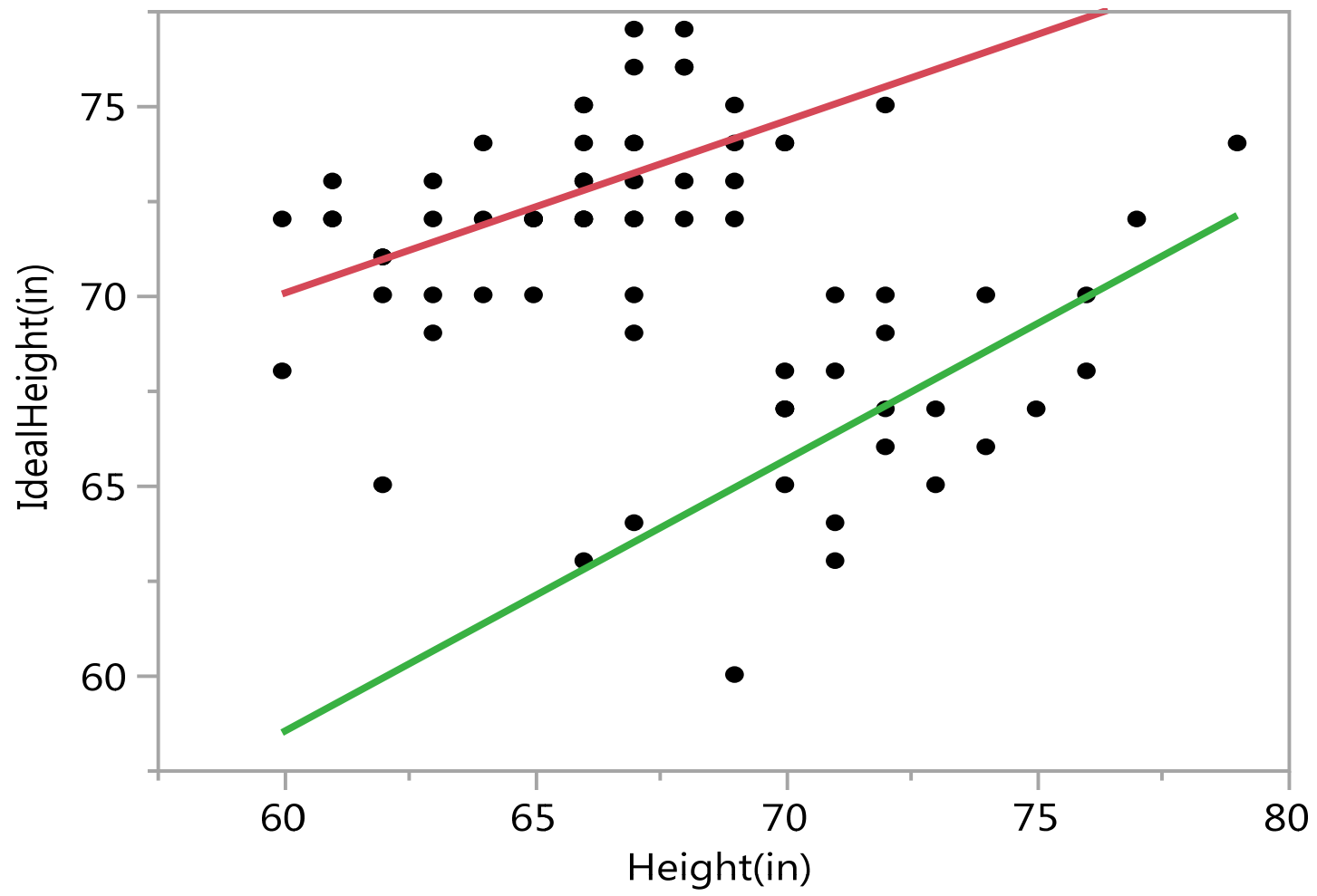
- Model is highly significant in explaining the height of the students' ideal romantic partner (p-value < 0.0001)
- $R^2 = 70.02\%$; 70.02% of the variation in the height of the students' ideal romantic partner can be explained by the multiple linear regression model with height and Gender of the student.
- The interaction term is not statistically significant (p-value = 0.1292).

Two Estimated Models

- Using the parameter estimates from the MLR model, we can determine the estimated intercept and slope of the models for females and for males. The two models are:

$$\hat{Y}_i = \begin{cases} 42.724 + 0.455x_i & \text{— for females} \\ 15.452 + 0.717x_i & \text{— for males} \end{cases}$$

Scatterplot



Model Comparison

- MLR Model
 - $SS_{\text{error}} = 302.90353$
 - $MS_{\text{error}} = 4.207$
 - $R^2 = 69.03\%$
 - $\text{adj } R^2 = 68.17\%$
- MLR Model with Interaction
 - $SS_{\text{error}} = 293.17088$
 - $MS_{\text{error}} = 4.129$
 - $R^2 = 70.02\%$
 - $\text{adj } R^2 = 68.76\%$

Which Model to Select?

- The two models are very similar.
- MLR model with interaction term has slightly better values for:
 - MS_{error}
 - $\text{adj } R^2$
- Interaction term is not statistically significant (p-value = 0.1292)
- Estimated MLR model with interaction term appears to fit the points in the scatterplot slightly better than the MLR model.