The Life of a Longwood Statistic Student: Linear Regression of Sex and Height on Weight

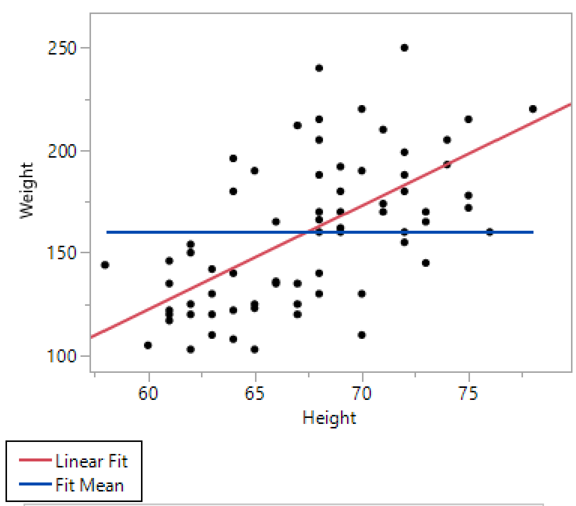
Authors: Natalie Wood & Kayla Cooksey

Introduction

We wanted to find the linear relationship given weight and height between males and females in Dr. Lunsford's fall semester statistics classes. We observed and analyzed a linear regression model to predict weight given height. The results observed showed that the linear relationship between males and females are positive, but not a good model given the low value of R2.

Data Collection and Description

The populations of interest included male and female statistics students in Dr. Lunsford's stats classes. The variables incorporated height as a quantitative type and explanatory role, weight as a quantitative type and response role, and sex as a categorical type and explanatory role. The Data were obtained from a representative sample of math 171 and 301 students from the Fall 2018 semester. The sample was large enough because n=71, which is greater than 30.



Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.611042	0.440541	0.738996	<.0001*
Covariance	106.9706			
Count	71			

Variable	Mean	Std Dev
Height	67.53521	4.597611
Weight	160.4225	38.07686

Figure 1 - Fit Y by X was run in JMP to display the linear model of Weight by Height of the data. The Summary Statistic was also displayed to show noteworthy data.

In Figure 1, the linear model showed a positive slope. This relationship was a positive direct relationship, meaning a change in height will produce a corresponding change in weight. The parameters of this model included H_0: $\beta_1=0$ and H_a: $\beta_1 \neq 0$. If β_1 is equal to 0, then there is no linear relationship between height and weight, and If β_1 is not equal to 0, then there is a linear relationship between height and weight, whether it is inverse or converse, depends on the direction of the slope.

According to Figure 1, the model is appropriate because the correlation coefficient is positive, which means the data points are closer to the line of best fit.

The simple linear regression statistical model was $Y=\beta_0+\beta_1X+\varepsilon$, $\varepsilon \sim N(0, \sigma)$ and the Predicted line: $\hat{y}=b_0+b_1x$. The parameters were β_0 and β_1, and the point estimates are b_0 (intercept of the regression line) and b_1 (slope for the regression line).

Analysis

The criteria for linear regression of the model included the confirmation of a random sample and linearity of the of the scatterplot. These can be confirmed because the sample was a random representative sample and that the scatter plot is linear because our R^2 was between 0 and 1. The criteria for linear regression of the error includes normality, zero mean, constant variance, and independence of the error (Lunsford, 2018). Normality constitutes that the random errors follow a normal distribution, the zero mean is where the error distribution is centered around 0, the constant variance is where the variance for Y is the same for all the X coordinates, and the independence of the error is that there is no relationship between the errors and the x-values. In Figure 2, approximately 24 out of 71 data points are within the mean confidence. This number reflects the R^2 coefficient of 34%.

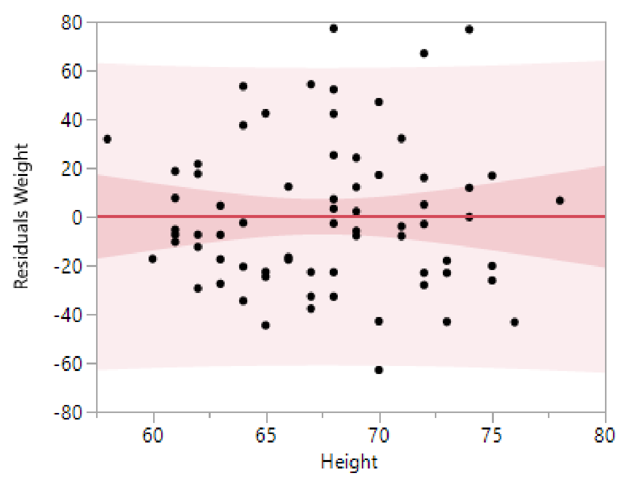


Figure 2 – Shows the bivariate fit of residuals for Weight by Height

When running the test for the hypotheses H_0: $\beta_1= 0$ and H_a: $\beta_1\neq0$ the t-test is $t(69)= 6.41$ with p-value $p=<.0001$ and f-test is $F(69)= 41.0881$ with p-value $p= <.0001$. $P<.05$, therefore, we rejected the null hypothesis and we're in favor of the alternative that males and females are not equal. when squaring t(69)= 6.41^2= 41.0881= F(69) (Figure 3)

Parameter Estimates						
Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-181.3446	53.42313	-3.39	0.0011*	-287.9208	-74.76839
Height	5.0605772	0.78924	6.41	<.0001*	3.4860864	6.6350681

Figure 3 – Parameter Estimates for Weight by Height

The p-value was significant because it was α<0.0001. Figure 3 showed the confidence interval, which relates to the outer ban in Figure 2. Weight increases with height because they were inversely proportional according to R^2.

The coefficient of determination is R^2 in this model was .393372 (Figure 3). This meant that the proportion of the variance between average heights on average weights was used to gauge whether the predicted number will be directly proportional to the prediction of the model via the amount of total variation.

Weight = -181.3446 + 5.0605772*Height

Summary of Fit	
RSquare	0.373372
RSquare Adj	0.364291
Root Mean Square Error	30.3592
Mean of Response	160.4225
Observations (or Sum Wgts)	71

Figure 4 – Summary for Fit and Predicted Model for the Linear Regression for Weight by Height

Prediction

The equation of the prediction line was: Weight = -181.3446 + 5.0605772*Height. The predicted weight was calculated by taking the equation of the predicted line and plugging in 68 inches in for height. The corresponding weight, according to that equation, of a Longwood Statistics student who was 68 inches tall was 162.775 lbs.

The residual equation is $Yi - \hat{y}i$ which was calculated by taking the observed data and subtracting it from the predicted calculation. The observed weight was 188 lbs and the predicted weight was 162.775 lbs, resulting in the residual being 25.225 lbs.

The predicted weight for a given height was calculated using the equation of our predicted line. We plugged 72.5 inches into the "Height" variable in the equation: Weight = -181.3446 + 5.0605772*(72.5 inches). This gave us the predicted weight of 185.5472 lbs. The predicted weight interval for this data point was (109.39348, 261.70092) because the standard deviation for the root mean square error was 38.07686, and for the data to be valid, it must be within 2σ of the mean. This was verified in Figure 5.

The average weight for students who are 72.5 inches tall is 188.6667 lbs. Weight = -181.3446 + 5.0605772*(72.5 inches) = 185.5472 lbs., this means that the average weight for the students that are 72.5 inches tall is more than the predicted weight.

In Figure 5 the confidence interval is the outer band (light red) and the mean confidence is the inner band (darker red).

Fit Mean	
Mean	160.4225
Std Dev [RMSE]	38.07686
Std Error	4.518892
SSE	101489.3

Figure 5 – Fit Mean of the Weight by Height data.

Discussion

Our model was not good for predicting weight in terms of height because our R^2 = .373372. R^2 was closer to 0, therefore was not a good model in terms of fit. If it were to be closer to 1, then it would be a better model and a perfect fit.

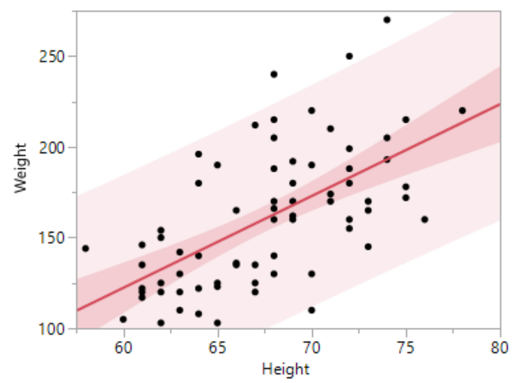
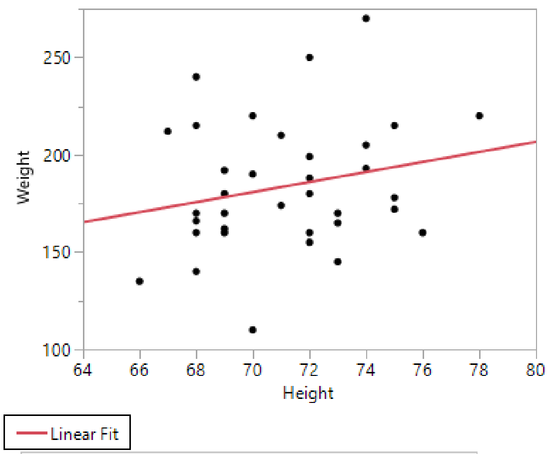


Figure 6 - The Bivariate Fit of Weight by Height. This shows the Confidence Shaded Fit and Individuals.

When we ran separate tests for the sexes, the P-values were significantly different. As shown in Figure 7, the P-value for males was significantly higher than .05, therefore we would fail to reject the null hypothesis. Found in Figure 8, the females P-value was significantly less than the males P-value and because it was less than .05 we reject the null hypothesis. You can also see that the graphs are not linear which shows no correlation between weight and height.

R^2 for males was .04923 (Figure 7) and .117663 for females (Figure 8). Both values were very small, which meant that both models for the individual sexes were very uncoordinated, therefore weak. Performing linear regression on the sexes under one model showed that the R^2 = .373372 (Figure 4). This value was closer to 1 than to the separated R^2's between the two sexes, however, was still not a significant correlation overall.



Linear Fit

Weight = 0.2371859 + 2.579397*Height

Summary of Fit

RSquare	0.049423
RSquare Adj	0.020617
Root Mean Square Error	33.20243
Mean of Response	183.7429
Observations (or Sum Wgts)	35

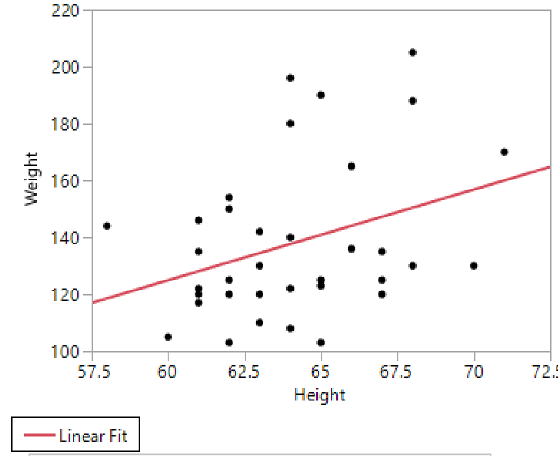
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1891.435	1891.43	1.7157
Error	33	36379.251	1102.40	Prob > F
C. Total	34	38270.686		0.1993

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.2371859	140.2077	0.00	0.9987
Height	2.579397	1.969212	1.31	0.1993

Figure 7 – Bivariate Fit of Weight by Height of Males



Linear Fit

Weight = -66.47891 + 3.1896923*Height

Summary of Fit

RSquare	0.117663
RSquare Adj	0.091712
Root Mean Square Error	25.81456
Mean of Response	137.75
Observations (or Sum Wgts)	36

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	3021.436	3021.44	4.5340
Error	34	22657.314	666.39	Prob > F
C. Total	35	25678.750		0.0406*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-66.47891	96.00896	-0.69	0.4934
Height	3.1896923	1.497983	2.13	0.0406*

Figure 8 – Bivariate Fit of Weight by Height of Females

References

Lunsford, M Leigh. "Inference for Linear Regression." *Math 301: Applied Statistics*. Longwood University, Virginia. 30 Oct. 2018.