Lab9

2024-11-11

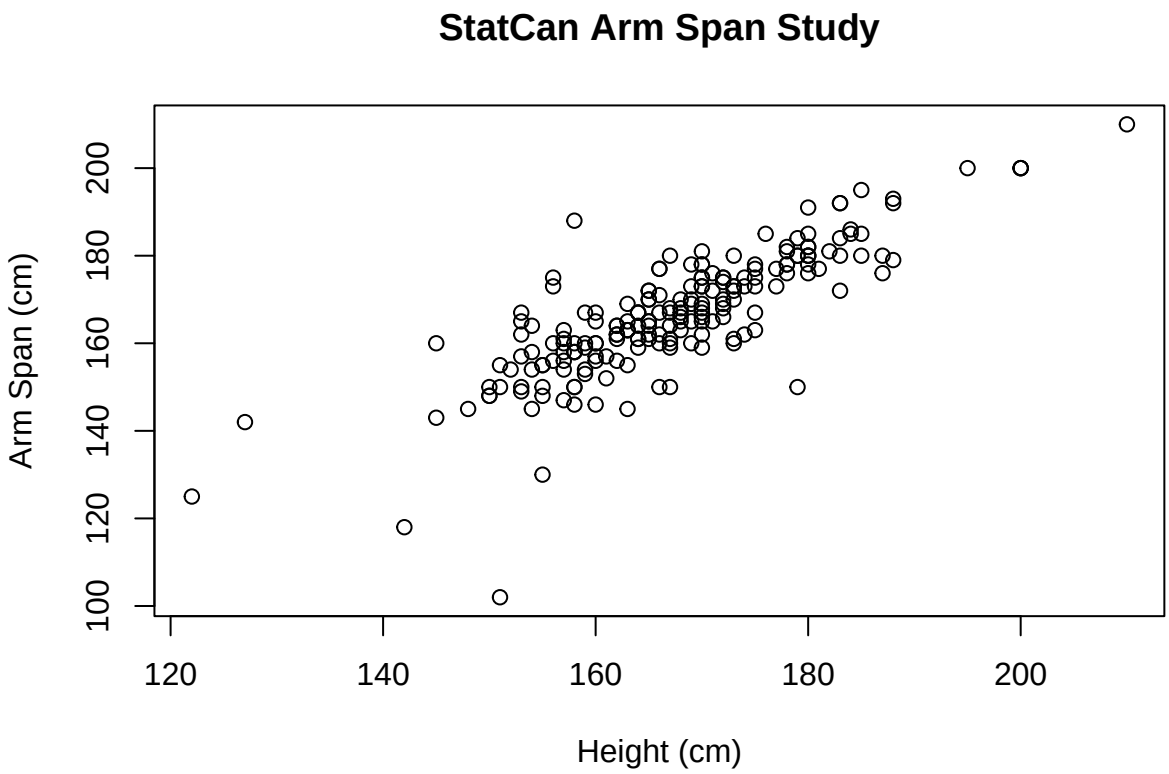
```
library(readr)
armspan <- read_csv("armspan.csv")
```

```
## Rows: 200 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Height, ArmSpan
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1.

Look at the scatterplot of the arm span and height values from the random sample of 200 students in Canada. What do you notice about the relationship between these two values?

```
plot(x = armspan$Height,
     y = armspan$ArmSpan,
     main="StatCan Arm Span Study",
     xlab="Height (cm)",
     ylab="Arm Span (cm)")
```




```
cor(armspan$Height, armspan$ArmSpan)
```

```
## [1] 0.8269918
```

The height and arm span appear to be positively and linearly related to one another.

2.

Write the simple linear regression (SLR) model for this problem. Give the definition of the parameter values $\beta_0, \beta_1,$ and σ^2 in the context of the response and explanatory variables.

$$Y_i = \beta_o + \beta_1 X_i + \epsilon_i$$

Where:

$$\epsilon \sim N(0, \sigma^2)$$

For Y (response) of arm span (cm) and X_i (explanatory) of height (cm).

β_0: mean arm span (cm) when height is 0 cm (the conditional mean of the response when the explanatory variable is 0). Yes, this is not especially realistic (having arm span at 0 height)

β_1: the increase (or change) in mean arm span (cm) (change in conditional mean of the response) when height (cm) increases by 1 cm

σ^2: the expected value of the mean-squared error of the residuals, or the variance of the residuals, where the residual is the vertical distance between observed value of arm span (cm) and the predicted value of arm span (cm)

3.

Give the equation of the least squares regression line to predict the value of a student's arm span from their height.

```
slr <- lm(ArmSpan ~ Height, data=armspan)
summary(slr)
```

```
##
## Call:
## lm(formula = ArmSpan ~ Height, data = armspan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.596  -3.329   0.539   3.630  30.480
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.22306    8.01667   0.153    0.879
## Height       0.98922    0.04779  20.698   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.046 on 198 degrees of freedom
## Multiple R-squared:  0.6839, Adjusted R-squared:  0.6823
## F-statistic: 428.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = b_0 + b_1 x = 1.22306 + 0.98922x$$

Where b_0, b_1 are the estimated values of the coefficients that minimize:

$$g(b_0, b_1) = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 x_i)]^2$$

4.

For the fitted SLR model, what is the interpretation of b1?

b1 is the model-estimated value of the parameter β_1, also know as the estimated increase (or change) in mean arm span (cm) when height (cm) increases by 1 cm (estimated change in conditional mean of the response when height increases by 1 cm)

We may also interpret b1 as the estimated value of β_1 that, along with b_0, minimizes the function of the data:

$$g(b_0, b_1) = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 x_i)]^2$$

5.

Give the ANOVA Table for the SLR model. Use the ANOVA Table to conduct a test of significance for the SLR model.

```
get.SS <- aov(ArmSpan ~ Height, data=armspan)
summary(get.SS)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Height        1  27738   27738   428.4 <2e-16 ***
## Residuals   198  12819      65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$

F = 428.4 with p-value <2e-16, provides overwhelming evidence (based on the thresholds we established previously for p-value interpretations) to reject the null hypothesis (can also say this meets the $\alpha = 0.05$ level significance), such that we have evidence that there is a significant linear relationship between arm span (in cm) and height (in cm).

6.

Give the value of R^2 for the SLR model. Show this value is equal to the ratio of the SS_{Model} to SS_{Total} using the ANOVA Table. Give an interpretation of this value.

```
summary(slr)
```

```
##
## Call:
## lm(formula = ArmSpan ~ Height, data = armspan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.596  -3.329   0.539   3.630  30.480
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.22306    8.01667   0.153    0.879
## Height       0.98922    0.04779  20.698   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.046 on 198 degrees of freedom
## Multiple R-squared:  0.6839, Adjusted R-squared:  0.6823
## F-statistic: 428.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
summary(get.SS)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Height        1  27738   27738   428.4 <2e-16 ***
## Residuals   198  12819      65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R^2: 0.6839

$\frac{SS_{Model}}{SS_{Total}} = \frac{SS_{Model}}{SS_{Model}+SS_{Residuals}} = 27738 \ / \ (27738 + 12819) = 0.6839263$

68.39% of the variation in arm span (cm) can be explained by the linear regression model with height (cm).

7.

Report the correlation coefficient between height and arm span. How does this value relate to the value of R^2 from the SLR model?

```
cor(armspan$Height, armspan$ArmSpan)
```

```
## [1] 0.8269918
```

Correlation Coefficient: 0.8269918

Correlation Coefficient^2 = R^2, as: 0.8269918^2 = 0.6839154

The values are related to one another, as noted above.

8.

Obtain the 95% confidence interval for the slope parameter in the SLR model. Give an interpretation of this interval.

```
confint(slr)
```

```
##                   2.5 %    97.5 %
## (Intercept) -14.5859436 17.032065
## Height        0.8949754  1.083472
```

We are 95% confident, that the true expected value of b1 is between 0.8949754 to 1.083472.

Also, we are 95% confident that the true expected increase (or change) in mean arm span (cm) when height (cm) increases by 1 cm (estimated change in conditional mean of the response when height increases by 1 cm) is between 0.8949754 cm to 1.083472 cm.

9.

Obtain a 95% confidence interval for the conditional mean arm span of all students in the population who are 170 cm tall. Give the interpretation of this interval.

```
predict.lm(slr, interval='confidence', newdata=data.frame(Height=170))
```

```
##        fit      lwr      upr
## 1 169.3911 168.2409 170.5413
```

We are 95% confident, that the true mean arm span for individuals who are 170 cm tall is between 168.2409 cm to 170.5413 cm.

10.

Obtain a 95% prediction interval for the predicted arm span of a student in the population who is 188 cm tall. Give the interpretation of this interval.

```
predict.lm(slr, interval='prediction', newdata=data.frame(Height=188))
```

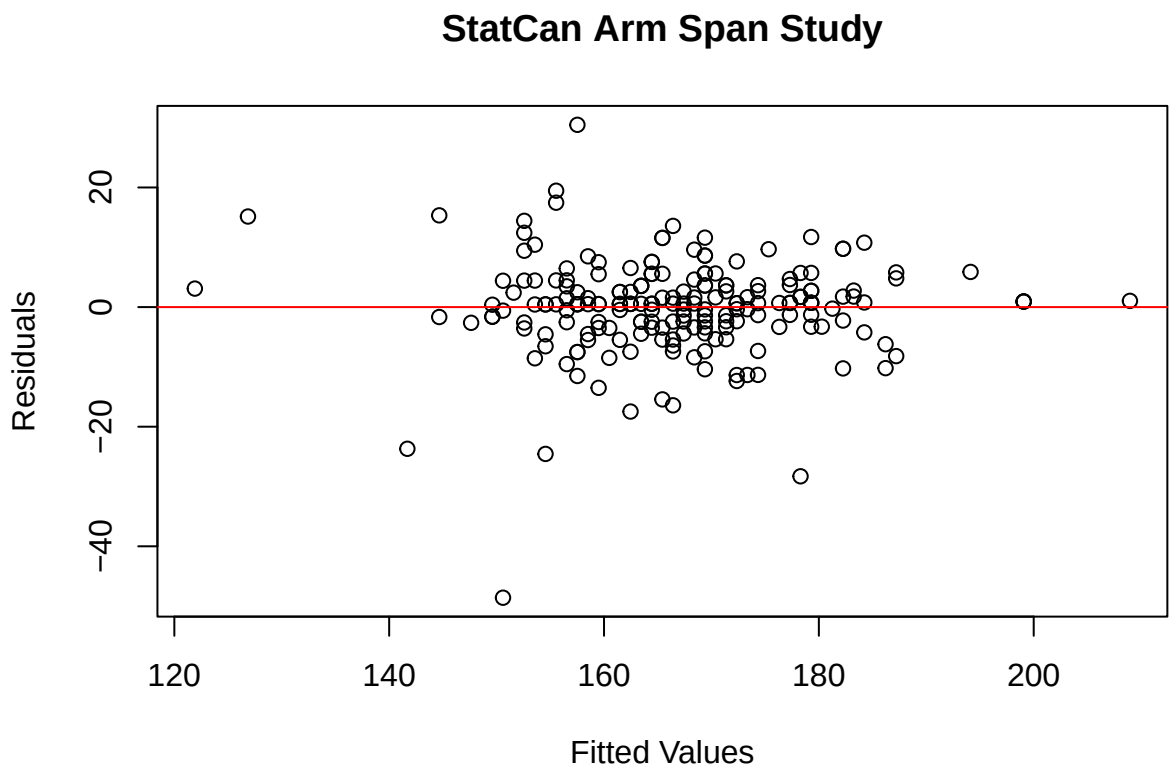
```
##        fit      lwr      upr
## 1 187.1971 171.1708 203.2234
```

We are 95% confident, that the true mean arm span for individuals who are 188 cm tall is between 171.1708 cm to 203.2234 cm.

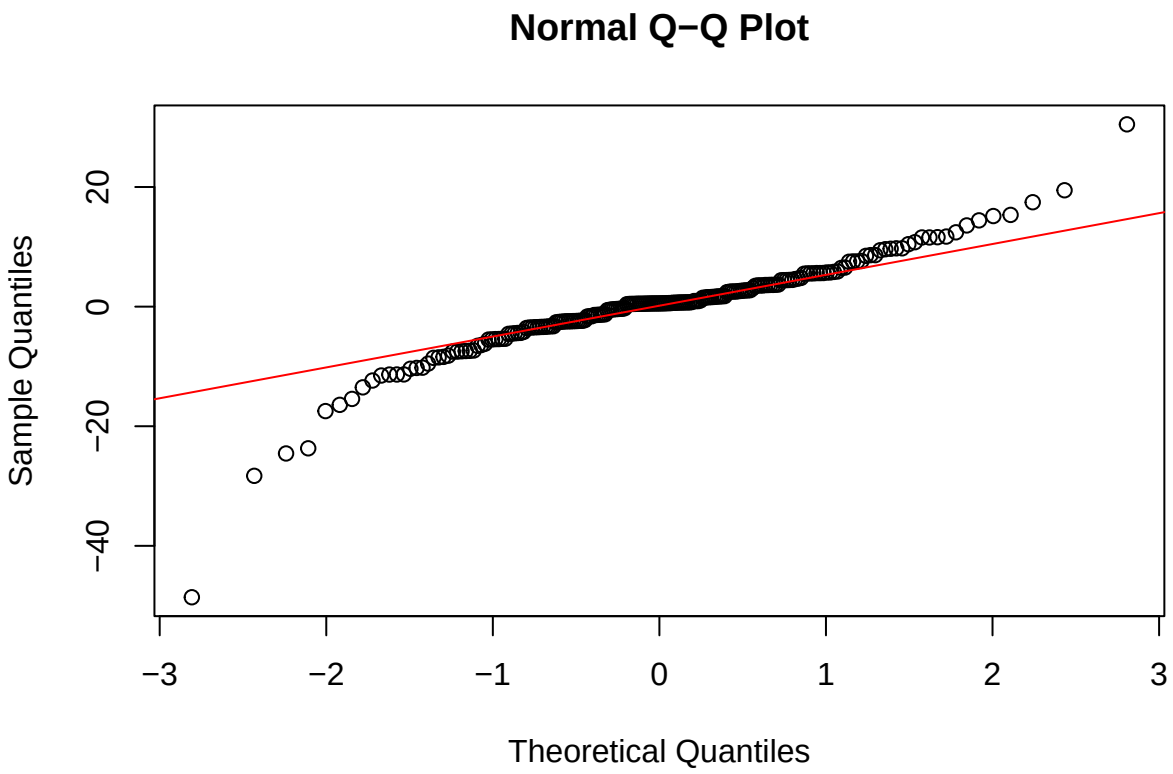
11.

Assuming that the independence and fixed-values-for-x assumptions are met, check the assumptions of linearity, constant variance, and normality. Summarize your findings.

```
plot(slr$fitted.values, slr$residuals, main="StatCan Arm Span Study",
xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
```

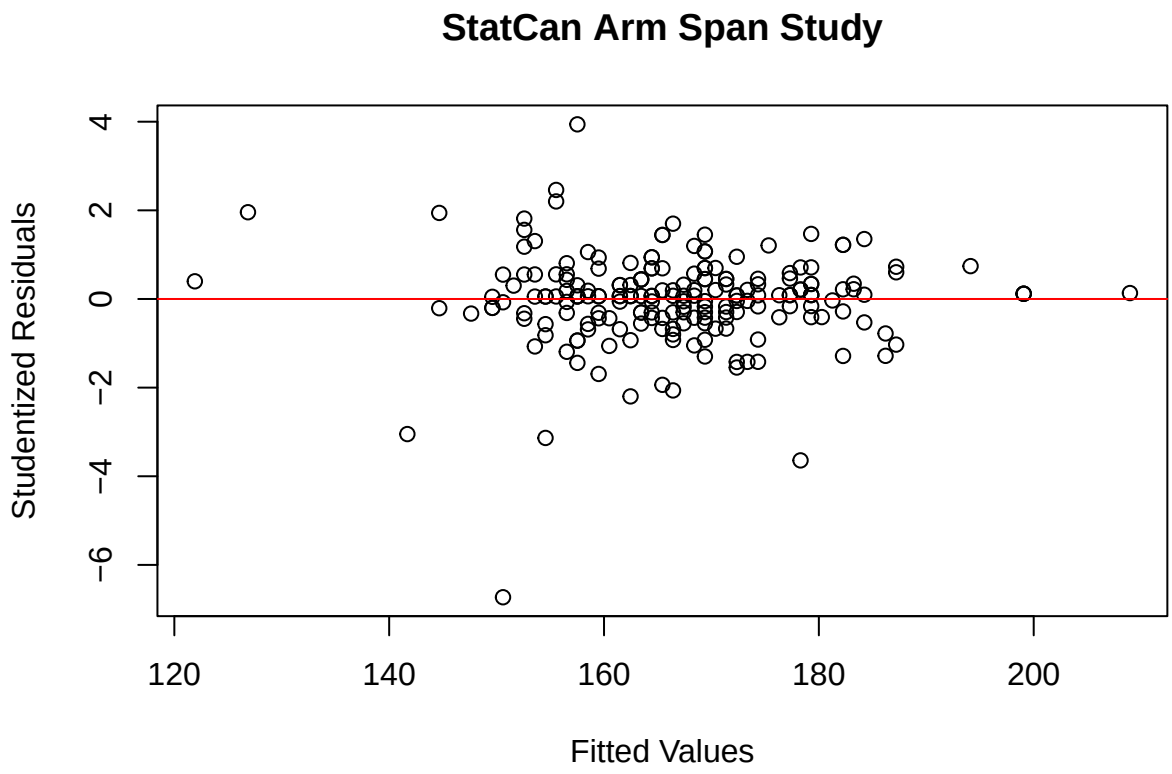



```
qqnorm(slr$residuals)
qqline(slr$residuals, col="red")
```

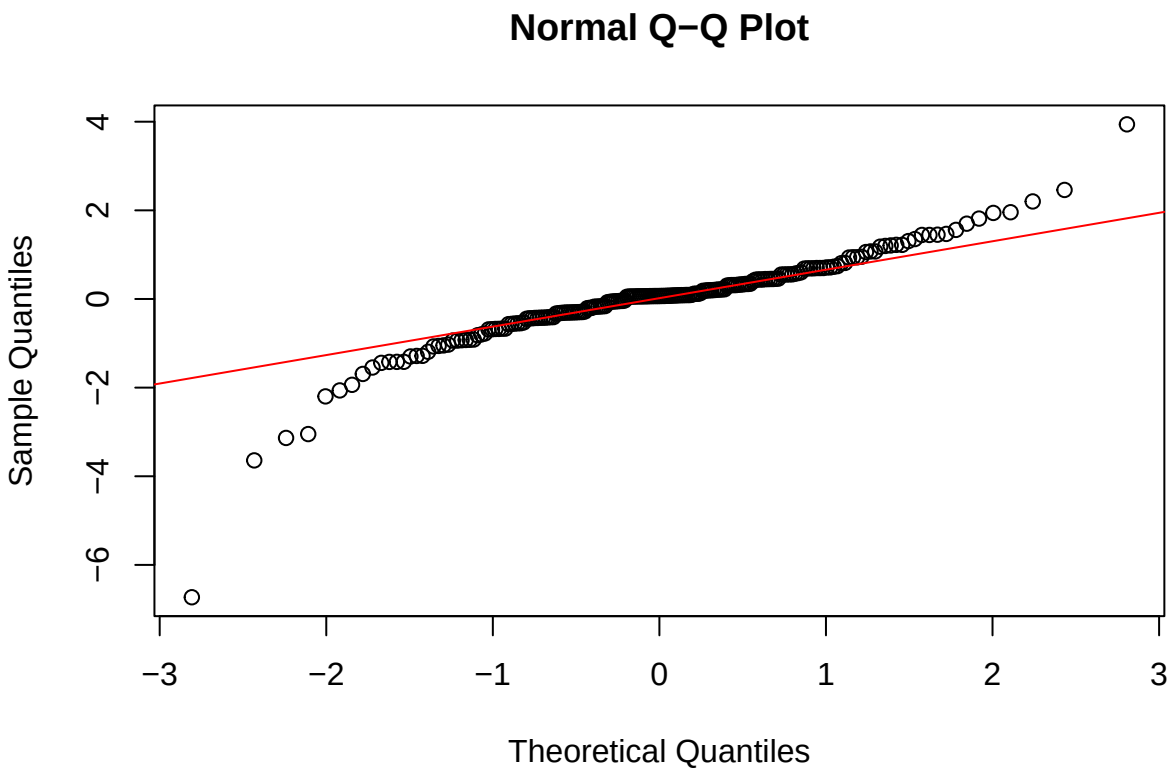



```
library(MASS)
stdresids <- studres(slr)
```

```
plot(slr$fitted.values, stdresids, main="StatCan Arm Span Study",
xlab="Fitted Values", ylab="Studentized Residuals")
abline(h=0, col="red")
```




```r
qqnorm(stdresids)
qqline(stdresids, col="red")
```



The following assumptions were primarily evaluated using the studentized residuals, though the non-studentized residual diagnostics are also included above and are consistent with the overall observations noted below.

Taking as a given that the independence and fixed-values-for-x assumptions are met, we evaluate the following assumptions of our linear model.

Linearity: Upon reviewing the residual plot (by fitted values) we generally observe a random spread of residuals across the range of fitted values (there is not a clear trend present in the residual plot above). We do not observe any trends or noticeable patterns in the above plots, such that we have reason to believe our linearity assumption is not being violated. Furthermore, when again reviewing the scatterplot from Q1, we see a general linear trend between our observed responses (arm span) and our one explanatory variable (height), further suggesting evidence that linearity is not being violated.

Constant Variance: We observe in the residual plot (by fitted values) that the spread of residuals tends to decrease somewhat following predicted (fitted values) greater than 160, though the spread may be thought of as relatively consistent after removing outliers from consideration. Generally, it appears that we do have constant variance, though with the possibility of some outliers in our data such that I have vague anxiety that this assumption is possibly being violated.

Normality: Based on the above QQ (Quantile) plot of the residuals, we tend to observe our residuals plot fairly closely with the reference line, at least within the range of +1 to -1 theoretical quantiles. However, outside of these quantiles we observe deviations from the reference line, especially when looking at the tails of the distribution. This generally is evidence to suppose that normality is possibly being violated.

12.

Conduct the F-test for lack-of-fit and report the results.

```
cell.means <- aov(ArmSpan ~ as.factor(Height), data=armspan)
summary(cell.means)
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Height)  45  31054   690.1   11.18 <2e-16 ***
## Residuals         154   9503    61.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(slr, cell.means)
```

```
## Analysis of Variance Table
##
## Model 1: ArmSpan ~ Height
## Model 2: ArmSpan ~ as.factor(Height)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    198 12819.4
## 2    154  9503.4 44    3316.1 1.2213 0.1883
```

```
qf(df1 = 44, df2 =  154, p = .95)
```

```
## [1] 1.455475
```

$H_0 : E(Y_{ij}|X_i) = \beta_0 + \beta_1 X_i$ $H_A : E(Y_{ij}|X_I) = \mu_i = \beta_0 + \beta_1 X_i + g(X_i)$

$F = \frac{MS_{\text{Lack of Fit}}}{MS_{\text{Pure Error}}} = \frac{(\frac{3316.1}{44})}{61.7} = 1.2213$ with p-value 0.1883.

Where $F_{(df_{LoF}, dfPE),0.95} = 1.455475$, we have $1.2213 < 1.455475$ ($F < F_{(df_{LoF}, dfPE),0.95} = F_{(44,154),),0.95}$) such that we do not reject the null hypothesis stated above at the $\alpha = 0.05$ level.

Our conclusion is that using Y = Arm Span (cm) as the response, the data are consistent with the straight line model using height (cm) as our explanatory variable, specifically the model as specified by:

$$Y_{ij} = \beta_0 + \beta_1 X_i + \epsilon_i$$