ST 352

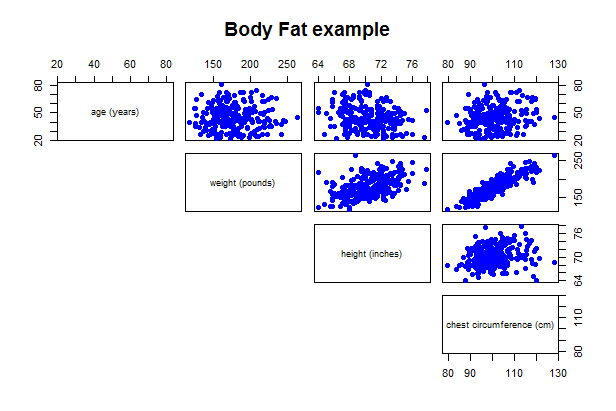
**Lab Assignment 4**

***33 points***

***Due 11:59 PM on Friday, November 1***

Use the graphs and output from the **bodyfat** data set that you obtained in the Lab 4 Notes to answer the following questions.

1. ***(2 points)*** Based on both the scatterplot matrix of the explanatory variables and the correlation matrix, which two explanatory variables are “highly correlated”? Explain. You must refer to the scatterplot matrix and the correlation matrix in your explanation of why these two variables are considered highly correlated. (Include the scatterplot matrix here. You do not have to include the correlation matrix.)



***Weight and chest circumference seem highly correlated from both the scatterplot matrix (tightly-packed points) and correlation coefficient (r = 0.89). None of the others appear to be highly correlated.***

***+1 for graph***

***+1 for support of their answer (i.e. “tightly-packed” points)***

2. ***(2 points)*** In the Lab 4 Notes, a strategy was shown to help with deciding which highly correlated explanatory variable to remove. From that strategy, which of the two highly correlated explanatory variables mentioned in the lab notes would be removed? Explain why that variable would be removed.

***I included the regression output below with just the two highly correlated explanatory variables (students should not include). Since “weight” has the higher p-value from the t-test, the strategy says to remove “weight”.***

***+0.5 for the correct variable to remove based on this strategy***

***+1.5 for correct support***

Coefficients:

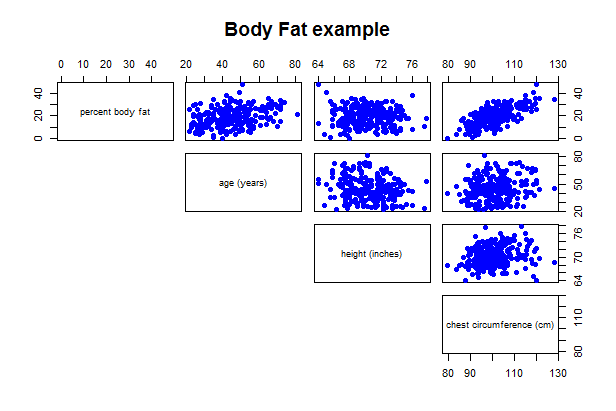
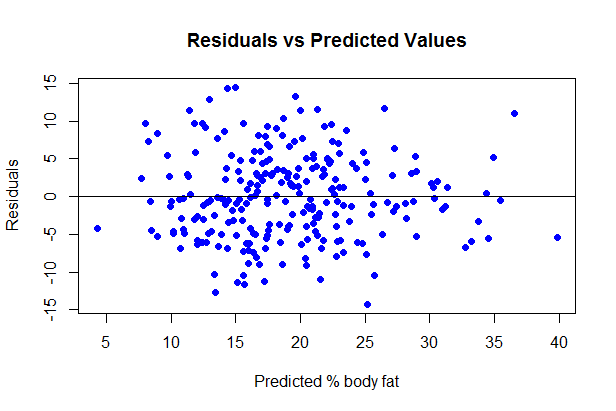
Estimate Std. Error t value Pr(>|t|)

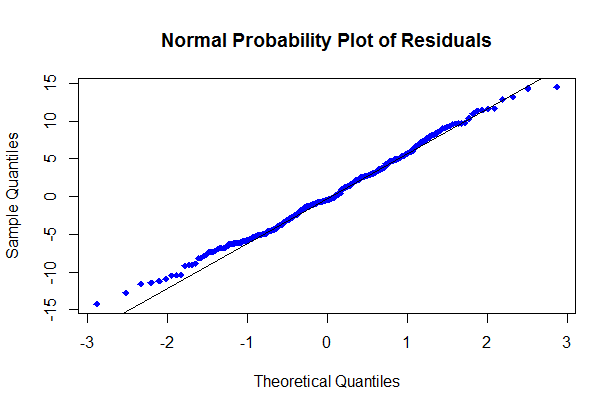
(Intercept) -53.99451 5.93133 -9.103 < 2e-16

weight -0.01074 0.03071 -0.350 0.727

chest 0.74446 0.10182 7.312 3.66e-12

3. ***(2 points)*** Include the graphical displays to assess the linearity, constant variation, and normality conditions (scatterplot matrix that includes the response variable, residual plot, and normal probability plot of the residuals).



***+ 2 for all three correct graphs***

***+ 1.5 for two correct graphs***

***+ 1 for one correct graph***

4. ***(2 points)*** Using the appropriate graph, discuss whether or not the linearity condition is satisfied. Make sure to reference which plot you are using in assessing this condition.

***Linearity condition seems satisfied as there is no curvature in the residual plot or in scatterplots of body fat versus each explanatory variable.***

***+ 0.5 for reference to the correct plot or plots. (We’ll allow reference to just one plot here.)***

***+ 1 for correctly using the plot(s) to support their answer***

***+ 0.5 for an answer of “yes” or “no” that is based on their support***

5. ***(2 points)*** Using the appropriate graph, discuss whether or not the constant variation condition is satisfied. Make sure to reference which plot you are using in assessing this condition.

***From residual plot above, the constant variation condition is pretty well satisfied. (Perhaps an argument could be that there is slightly less spread for higher predicted values, but for the most part, this condition is satisfied. Look for support when grading.)***

***+ 0.5 for reference to the correct plot.***

***+ 1 for correctly using the plot to support their answer***

***+ 0.5 for an answer of “yes” or “no” that is based on their support***

6. ***(2 points)*** Using the appropriate graph, discuss whether or not the normality condition is satisfied. Make sure to reference which plot you are using in assessing this condition.

***An argument could be made either way. Most points fall on or near the reference line, supporting that the normality condition is met. There are a number of points off the line at either end, which could support that the normality condition is not perfectly met. Both arguments are acceptable, but this is an example where we should feel comfortable that the condition is satisfied.***

***+ 0.5 for reference to the correct plot.***

***+ 1 for correctly using the plot to support their answer***

***+ 0.5 for an answer of “yes” or “no” that is based on their support***

Answer the remaining questions using a model with no transformation regardless of whether all of the conditions are satisfied or not.

***Here is the output (students should not include here!)***

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -22.56283 10.43100 -2.163 0.031501

age 0.08304 0.03016 2.753 0.006338

height -0.49966 0.14713 -3.396 0.000797

chest 0.72514 0.04654 15.581 < 2e-16

Residual standard error: 5.65 on 246 degrees of freedom

Multiple R-squared: 0.5415, Adjusted R-squared: 0.5359

F-statistic: 96.85 on 3 and 246 DF, p-value: < 2.2e-16

7. Perform an F-test to determine if at least one explanatory variable is helpful in predicting the response.

a. ***(2 points)*** State the null and alternative hypotheses **in words**.

**H0: *None of the explanatory variables help to predict the percent body fat.***

**HA: *At least one of the explanatory variables helps to predict percent body fat.***

***+ 1 for each hypothesis***

* *Deduct ½ point if do not say “at least one” in the alternative hypothesis*
* *Deduct 1 point if have null and alternative hypotheses switched*
* *Deduct ½ to 1 point for other errors*
* *Note: do not deduct if student tries to write hypotheses in notation and doesn’t write the notation correctly (since they are not asked to write the hypotheses in notation)*
* *It’s okay to just say “explanatory variables” as I did above. Students could list the explanatory variables as well.*

b. ***(1 point)*** Using the regression output, report the F-statistic with degrees of freedom and the p-value. (No calculations are necessary!)

***F-stat(3,246) = 96.85 p-value < 0.0001***

***+ 1/4 each for 1) correct F-statistic, 2) correct numerator df, 3) correct denominator df, and 4) correct p-value.***

* ***If student just says 246 degrees of freedom, count this as the correct denominator df and take off ¼ point for not having the numerator df***

c. ***(3 points)*** State a conclusion from the F-test in the context of the problem. **(1 point)**

***There is strong evidence to indicate that at least one of age, height, and/or chest circumference help to predict a person’s percent body fat (p-value < 0.0001).***

8. For ***each*** explanatory variable,

a. ***(2 points)*** report the t-statistic with degrees of freedom and the p-value for the t-test (testing if that explanatory variable explains the response variable after accounting for the effects of the other variables)

***age: t-stat246 =2.753, p-value = 0.0063***

***height: t-stat246 = -3.396, p-value = 0.0008***

***chest: t-stat246 = 15.581, p-value < 0.0001***

***+ 1/2 for correct degrees of freedom***

***+ 1 for correct t-statistics***

***+ 1/2 for correct p-values***

* ***If student made an error in recording just one of the t-statistics or p-values, take 1/4 point off.***

b. ***(3 points)*** state a conclusion in the context of the problem based on the p-value from the t-test. (You should have three separate sentences, one for the conclusion for each explanatory variable. Much of each sentence may contain the same wording, but you still need to write three separate sentences with three separate conclusions.)

***There is strong evidence to indicate age helps to predict a person’s percent body fat after accounting for the effects of height and chest circumference (p-value = 0.0063).***

***There is strong evidence to indicate height helps to predict a person’s percent body fat after accounting for the effects of age and chest circumference (p-value = 0.0008).***

***There is strong evidence to indicate chest circumference helps to predict a person’s percent body fat after accounting for the effects of height and age (p-value < 0.0001).***

***Things for which to deduct points:***

* ***-1 if forget “after accounting for the effects of the other explanatory variables”***
* ***-1 if always used wrong adjective (such as “not enough” evidence instead of “strong” evidence. Note: we’ll accept just “evidence” or “some evidence” in place of “strong evidence”***
* ***-1 if do not have correct phrasing of “helps to predict a person’s percent body fat”***
* ***Note: if a student made some sort of other minor error, take off ½ point. The above deductions do total three points, but it’s unlikely that they would make all of the above errors in addition to some other minor error.***
* ***Note 2: again, the above deductions do total three points. If you feel a student should receive at least some credit even though they made all of these errors, you can give up to ½ point. But, if they made all three of these errors, it’s unlikely they really had anything correct in their conclusion.***
* ***Note 3: a statement of just “strong evidence to reject the null hypothesis” without writing anything else should receive a 2.5 point deduction!!!***
* ***Note 4: a statement of “strong evidence to reject the null hypothesis” followed by the null hypothesis should receive a 1.5 point deduction! Such a statement tells me that they have decided to reject the null hypothesis but doesn’t tell me what they conclude (i.e. what the alternative hypothesis is).***

9. ***(1 point)*** Suppose a backwards selection process was performed. Would any of the explanatory variables drop out? Why or why not? If so, which one would drop out first? Why?

***No. Since all p-values from the t-test are less than 0.05, all three explanatory variables remain in the model.***

***+ 1/2 for correct answer and + 1/2 for correct reasoning***

10. If needed, perform a backwards selection process to obtain a model with only significant explanatory variables. Use the model with only significant explanatory variables to answer these questions:

a. ***(1 point)*** Include the regression output of your final model.

**It is here students should include the output I had above**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -22.56283 10.43100 -2.163 0.031501

age 0.08304 0.03016 2.753 0.006338

height -0.49966 0.14713 -3.396 0.000797

chest 0.72514 0.04654 15.581 < 2e-16

Residual standard error: 5.65 on 246 degrees of freedom

Multiple R-squared: 0.5415, Adjusted R-squared: 0.5359

F-statistic: 96.85 on 3 and 246 DF, p-value: < 2.2e-16

b. ***(2 points)*** Write the least-squares regression equation. Define the terms in the equation (i.e. what the x’s and represent in the context of the problem).

**= -22.56283 + 0.08304x1 – 0.49966x2 + 0.72514x3**

**Where = predicted percent body fat**

**x1 = age (in years)**

**x2 = height (in inches)**

**x3 = chest circumference (in cm)**

**+ 1 for correct equation**

**+ 1 for defining the terms in the context of the problem**

* **Note: if student put “age” in place of x1 in the equation (for example), they should still define “age”.**

c. ***(2 points)*** Interpret the coefficient of *age* in the context of the problem.

***For a particular person, body fat is predicted to increase by 0.08304% as the person ages by 1 year.***

***+ 0.5 points for “keeping the other variables the same”***

***+ 0.5 points for increasing age by 1 year***

***+ 0.5 points for correctly stating the response variable is “predicted” (i.e. they must include “predicted”, “expected”, or “on average”)***

***+ 0.5 points for correctly stating “increase” by the value of the coefficient***

* ***Note: “For a particular person” is an efficient way of keeping the other variables the same. Students could (and will probably) write it differently, such as “keeping height and chest circumference the same”. It doesn’t matter how a student writes it as long as they are indicating that we’re not changing the values of the other variables.***

d. ***(1 point)*** Predict percent body fat for a 23 year old who is 73 inches tall, and has a chest circumference of 120 cm. Use R to find this predicted value.

***about 29.89%***

e. ***(3 points)*** Report and interpret a 95% prediction interval for the person in question 10c.

***(18.51%, 41.27%)***

***We’re 95% confident that a person 23 years old who is 73 inches tall with a chest circumference of 120 cm will have body fat between 18.51% and 41.27%.***

***+ 1 for correct prediction interval (including writing it in proper syntax)***

***+ 2 for correct interpretation***