Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 7 (due: 11/11/16)

Please note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R.

1A. Using the summary function after fitting a multiple variable model, we reach the model $y = -165.5332 + 4.9605*\text{waist} + 2.4884*\text{height}$

(i) I am using outputs from R to calculate the following:
$\text{RSS} = \text{df} * (\text{RSE})^2 = 504 * (9.986)^2 = 50258.9$
$\text{SS}_{\text{reg}} = \text{F} * \frac{RSS}{n-2} = 1945 * \frac{50258.978784}{507-3} = 387917$
$\text{SYY} = \text{RSS} + \text{SS}_{\text{reg}} = 50258.9 + 387917 = 438176$

(ii) r^2 is 0.8853 and adjusted r^2 is 0.8848

(iii) Among people of the same waist size, people whose height is 1 inch more are, on average, 2.4884 pounds heavier.

(iv) No, it does not. People's height does not generally increase, so a one unit increase does not make sense. On the other hand, substituting height for waist size makes sense since, unfortunately, people's waist sizes can increase.

1B. This has been done in R.

(i) $\text{RSS} = \text{df} * (\text{RSE})^2 = 503 * (9.995)^2 = 50249.7$
$\text{SS}_{\text{reg}} = \text{F} * \frac{RSS}{n-2} = 1294 * \frac{50249.712575}{507-3} = 387929$
$\text{SYY} = \text{RSS} + \text{SS}_{\text{reg}} = 50249.7 + 387929 = 438176$

(ii) The RSS has slightly increased, while the SS_{reg} also decreased. The SYY is still the same. Moreover, if regression is useful, RSS will be small and SS_{reg} will be big, and thus, this model is slightly less valuable than the first one, meaning that the worthless data really has no significance in comparison to the height and waist data.

(iii) r^2 is 0.8853 and adjusted r^2 is 0.8846. The r^2 has remained the same while the adjusted r^2 has changed due to the addition of the worthless variable. Notice how the adjusted r^2 has actually decreased from before, indicative that the worthless variable does not belong in the model.

1C. This has been done in R

(i) $\text{RSS} = \text{df} * (\text{RSE})^2 = 503 * (9.995)^2 = 50249.712575$
$\text{SS}_{\text{reg}} = \text{F} * \frac{RSS}{n-2} = 1294 * \frac{50249.712575}{507-3} = 387929$
$\text{SYY} = \text{RSS} + \text{SS}_{\text{reg}} = 50249.712575 + 387929 = 438176$
Note that we get the exact same values as before, meaning that order does not matter in the summary function (this is not the case in sequential ANOVA!)

(ii) The RSS has slightly increased, while the SS_{reg} also decreased. The SYY is still the same. Moreover, if regression is useful, RSS will be small and SS_{reg} will be big, and thus, this model is slightly less valuable than the first one, meaning that the worthless data really has no significance in comparison to the height and waist data.

(iii) r^2 is 0.8853 and adjusted r^2 is 0.8846. The r^2 has remained the same while the adjusted r^2 has changed due to the addition of the worthless variable. Notice how the adjusted r^2 has actually decreased from before, indicative that the worthless variable does not belong in the model.

1D. Adjusted r^2 is definitely more valuable. When we added a worthless variable, the r^2 did not change. Whenever a new variable is added, either the r^2 will stay the same or increase; it's pretty optimistic, but in theory, this means that we can eventually get an r^2 of 1 just by adding random noise. The adjusted r^2, on the other hand, takes into account the value of each variable. It will not take into account random variables since it adjusts for the fact that some variables are useless. Every time a new variable is added, the adjusted r^2 will go down unless the value of the variable can overcome this penalty, since it divides by the degrees of freedom, which is correlated with the number of variables in the model. Thus, when comparing variables, the adjusted r^2 is more reliable than r^2.

1E. Looking at the SS_{reg} between two models is useless since you can't determine the significance of each variable, only its relative comparison. It won't let us know if the difference between two variables is actually significant. On the other hand, a partial test is useful because you can use the p-value to determine which variables are significant in a model at a 95% confidence level. In an F partial test, however, order does matter, so one must be careful of which variables are being inputted in which order in order to obtain the results one intended.

2A. I have fit the model in R. Using the summary function, the model is (bear with me):
y = 349.97628 + 1.05418*DealerCost – 32.24720*EngineSize + 228.32952*Cylinders + 2.36212*Horsepower – 16.74239*CityMPG + 46.75754*HighwayMPG + 0.69920*Weight + 27.05345*WheelBase – 7.32019*Length – 84.70850*Width

2B. Estimated slope for Cylinders variable: 228.32952
t-statistic for Cylinders variable: 3.171
p-value for Cylinders variable: 0.001730

From these values, we can interpret the slope to mean that for when all of the other variables held constant, then cars with 1 more cylinder, on average, have a suggested retail price of $228.33 more. Since the p-value is less than 0.05, at a significance level of 95% we conclude that this difference of $228.33 is significant, indicative of an actual association between cylinders and suggested retail price.

2C. In the ANOVA table, only the last variable corresponds to an F-statistics whose t-statistic will be corresponded. Remember that the F-statistics is the square of the t-statistics. Thus:

```
anova(lm(SuggestedRetailPrice~DealerCost+EngineSize+Horsepower+CityMPG+HighwayMPG+Weight+WheelBase+Length+Width+Cylinders, data = cars))
```

2D. From the above code, we get an F-statistic of 10.058, which is approximately the same as the (t-statistic)^2 = 3.171^2, and thus we can conclude that we did this correctly.

2E. In the full model, we obtain an adjusted r^2 value of 0.9989. In the model without the fuel consumption variables, we obtain an adjusted r^2 value of 0.9988. Although slightly better, the full model is actually a better representative of the data, since it explains 99.89% of the variability in the suggested retail price, which is quite amazing!