Final

"I have not used any resources from outside the class or discussed the exam with anyone." - Martin Zanaj

1. For each of the following, answer the question and explain briefly how you find your answer.

a. What is the sample size n?

N is equal to **39**; this can be found by making use of df=n-p –> 35=n-4.

b. What is the residual sum of squares (RSS)?

The **RSS** is equal to **44382.52**. Now refering to Ch.2 of "Linear Models with R" on page 16, there is a clear formula for calculating tha variance of e. The estimate for such variance is represented by simga_hatˆ2 which is equal to RSS/df.residuals, where df.residuals is eqaul to (n-p) [total observations - (#predictors +1)]{39-4}.This formula can be rearranged, so as to solve for RSS by RSS= sigma_hatˆ2 * df.residuals. Both of these values can be found in the regression summary output where **sigma_hat = 35.61** and **df. residulas = 35**. Hence, our calculations can be carried as **35.61ˆ2 * 35 = 44382.52**.

c. What is AAA?

This is equal to **-5.077885**. In order to get the t-statistic, we simply divide the estimated B coefficient by the standard error, or simply -8.66277/1.70598.

d. What is BBB?

This is equal to **1.267768e-05**. In order to calculate the p-value (2-sided), we can call look up a t-table or make us of the r function with parameters 2*pt(-5.077885, df= 35).

2. For each of the following, answer the question and explain briefly how you find your answer.

a. What is CCC?

This is equal to **0.6905714**. Now refering to Ch.10 of "Linear Models with R" on page 155, there is a clear formula for calculating Adjusted Rˆ2 given by 1-((n-1)/(n-p) * (1-rˆ2)), which in the context of our problem is equal to 1-((39-1)/(39-4)*(1-0.715))= 0.6905714.

b. What is DDD?

This is equal to **3**. Since it is an F test, we can find its parameters according to F(p-1, n-p). In thisc case p-1 is the parameter in question. Hence, 4-1=3.

c. What is EEE?

This is equal to **35**. Since it is an F test, we can find its parameters according to F(p-1, n-p). In thisc case n-p is the parameter in question. Hence, 39-4=35.

d. What is FFF?

This is equal to **1.182043e-09**. The p-value can be found by running an F test with an F-statistic of 29.28 and df 3,35, or simply in r 1-pf(29.28, 3, 35).

3. If the p-value FFF is significant and we reject the corresponding null hypothesis, we could conclude which of the following? Choose only one of the answers below, and explain briefly.

a. At least one of the coefficients for wage, asset, or age is non-zero. It is useful to check our hypothesis H0: Bwage=Basset=Bage=0, HA: Bwage!= 0 or Basset!=0 or Bage!=0. By running an F-test,at an alpha level of 5%, we reject the null that all predicotrs are zero, in favor of the alternative at least one predictor is not zero.

4. What is the smallest possible value of ŷ given the information above? Explain briefly how you arrive at your answer.

The smallest possible value for yˆ could be **1808.004**. This takes in consideration the sign of each coefficient, and in order to find the smallest possible value for negative coefficients (wage/age) the maximum value is used for the x value, wherease for positive coefficients (asset) the minumum value is used. This configuration yields the smallest possible value. Of course, for this to be true an observation with this specfics must be present in the dataset. This configuration is equal to 2444.78795+(-47.61368* 3.636)+ (0.02641* 1370)+(-8.66277*57.70).

5.Suppose that the AGE variable, instead of being measured in years, is measured in months.

a. Which of the following is true? Explain briefly.

- None of the p-values will change. Age is a predictor, and changing the years to months will result in a change of scale. A change of scale in a predictor does not affect the p-values nor the test statistics for the T and F tests.

b. Which of the following is true? Explain briefly.

- None of the t statistics will change, and neither will the F statistic. To confirm the statement in the previous question, I ran a simple experiment on R with the pima dataset. As expected the p-values and t statistics for the predictors and the intercept did not change.

```r
library(faraway)
data(pima)

#Original Model
lm = lm(bmi~ triceps+glucose+pregnant, data=pima)

#Scaled model
scale=(pima$triceps)*12
lm2= lm(bmi~ scale+glucose+pregnant, data=pima)
#Output omitted for clarity
```

6. If you were to use the regsubset() function to find the best model with only two predictors,which two predictors would be included? Explain briefly how you know.

I expect the regsubset function would consider **Age** and **Asset** as the best 2model predictor this is due to their extremely significant p-values in comparision to the other predictor wage.

7.For (a) and (b) below, show how you find your answer, using only the information given above. Then answer (c). HINT: For (b), it will help to remember the equivalence of the t-test and F-test.

a.Calculate the AIC for the full model.

The AIC is **282.4445**. The formula for AIC is n* ln(RSS/n)+2(p+1), which is 39 * log(44382.52/39)+2* (3+1).

b.Calculate the AIC for the best model with only two predictors (the model from question 6).

What we can do is make use of the F statistic and determine whether a smaller model is better or not. We can set up an hypothesis H0: Bage!=0 Basset!=0 Bwage=0, Ha: Bage!=0 Basset!=0 Bwage!=0. As a result, if we fail to reject H0 than we can conclude that the model without Bwage is better, and consequently the AIC of this model ought to be bigger than the one without it. To get the F-statistics, we simply square the t-statisic 4.2849. To understand weather we should drop this predictor from the model we run an F(p-q,n-p) –> F(4-3, 39-4) –> F(1,35), and get its p-value, or simply in r 1-pf(4.2849, 1,35). This test provides a p-value of 0.045. As a result, we can reject the notion that Bwage=0, and determine that the model with all predictors will be better than the one with just two predictors. Hence, the AIC of the full model ought to be smaller (a.k.a better).

c.According to AIC, which model is better – the one in (a) or the one in (b)?

Model **b** due to the fact that it will have the smallest AIC.

Questions 8 and 9 below refer to the galapagos data. Before answering questions 8 or 9, load the data in R, and fit a poisson regression model, with Species as the response variable, and Area,Elevation, Nearest, Scruz, and Adjacent as the predictors. Display the summary of the model.

```
library(faraway)
data(gala)
model= glm(Species ~Area+Elevation+Nearest+Scruz+Adjacent, family = poisson, data=gala)
summary(model)
```

```
##
## Call:
## glm(formula = Species ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, family = poisson, data = gala)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.2752  -4.4966  -0.9443   1.9168  10.1849
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
## Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
## Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
## Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
## Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
## Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3510.73  on 29  degrees of freedom
## Residual deviance:  716.85  on 24  degrees of freedom
## AIC: 889.68
##
## Number of Fisher Scoring iterations: 5
```

8.For the island of Caldwell, area = 0.21, elevation = 114, Nearest = 2.8, Scruz = 58.7, and Adjacent = 0.78. What is the predicted value of the Poisson rate parameter for the island? (You may do this by hand, using the information in the model summary, or in R, but either way, show how you arrived at your answer.)

> The predicted value of Poisson rate parameter for the island is **3.247365**. Also, mu=exp(b0+XB) –> exp(3.1548078779-0.0005799429* 0.21+0.0035405940* 114+0.0088255719* 2.8-0.0057094223* 58.7-0.0006630311*0.78) –> 25.72248.

```
# make a dataframe with new data
newdata = data.frame(Area=0.21,Elevation=114, Nearest=2.8, Scruz=58.7, Adjacent=0.78)

# Prediction
predict(model, newdata = newdata,type='response')
```

```
##        1
## 25.72248
```

9.What is the estimated probability that you would observe 5 species on Caldwell island? Showany work.

> The estimated probability that you would observe 5 species on Caldwell island is **6.327787e-07**, or practically zero.

```
dpois(x=5, 25.72248)
```

```
## [1] 6.327787e-07
```

Questions 10 and 11 are general and do not refer to any specific dataset.

10.Deviance can be used... (choose one answer and explain briefly)

- All of the above. Deviance is used as a goodness of model fit. As in the case of the F test, we can compare multiple models with different parameters and determined which model is better according to the resulting p-vale. Perhaps, one can check a model with all parameters vs a model with all parameters-1, so as to understand the importance of one predictor. Altough, this is an indirect method, it is still possible to get an idea of the predictor.

11.In ridge regression, increasing shrinkage factor will typically (choose one answer and explain briefly)

- Decrease variance but increase bias. In ridge regression,we shrink coefficients, so as to decrease model complexity while keeping all variables in the model. When lamda equals zero, the coefficients are equal to the OLS estimates, but as we increase lambda (say infinity) our coefficients will go to zero. From this, we can infer an increase in bias (loss of true relationship) at the advantage of less variance.