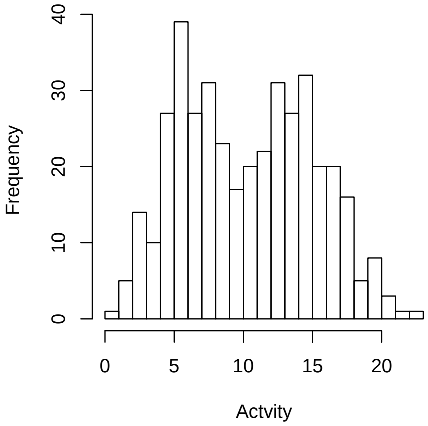
BIOS663 Homework 3 Solution

Dr. Cage wishes to study whether mice living in two types of cages have different levels of physical activity. One is a regular cage and the other is an enriched cage with more space and some toys. In addition, covariates including the age, sex, and strain of each mouse is available. The sample size is 400 and the following table shows the data of the first 6 samples. The covariate `sex`=0 or 1, for female or male mice, respectively. The covariate `strain`=0 or 1 for mouse strain B6 or Cast, respectively.

actv	sex	age	cage	strain
9.4	0	2.2	0	1
3.9	0	2.1	1	0
5.6	0	2.1	0	0
3.8	0	1.8	0	0
6.3	1	2	1	0
12.3	0	2.3	1	1

1. (6pts) Dr. Cage first plots the distribution of mouse physical activity, which is shown in the following figure. Does it look like that physical activity follows a normal distribution? If it does not follow normal distribution, what is the consequence for the following analysis?



No. The physical activity does not follow a normal distribution. This is not necessarily a problem since the normal distribution assumption is only applied to the residuals.

2. (6pts) Dr. Cage first did a t-test to compare physical activity between mice living in regular cages versus enriched cages. He obtains a p-value of 1.482e-05. Then he did an ANOVA analysis and obtains almost the same p-value.

```
            Df Sum Sq Mean Sq F value    Pr(>F)
cage         1  414.5  414.53  19.238 1.479e-05 ***
Residuals 398 8575.9   21.55
```

Explain under which assumption the t-test and ANOVA will give the same p-value. And based on this ANOVA table, what is the correlation between cage and mouse physical activity?

Homogeneity of variance assumption, i.e., the variance of physical activity are the same for mice living in regular cage or the enriched cage.

$R^2 = CSS/(CSS + SSE) = 414.5/(414.5 + 8575.9) = 0.046$, and thus correlation is $\sqrt{R^2} = 0.2147$.

3. (6pts) Next Dr. Cage did a regression analysis including all of the covariates. Here are the output of the Wald test p-values for each covariate and ANOVA table. .

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.7080     0.8668  -5.431 9.79e-08 ***
cage         -0.2774     0.2241  -1.238  0.21636
strain        8.0049     0.2183  36.670  < 2e-16 ***
sex           0.6849     0.2158   3.173  0.00162 **
age           5.4173     0.4215  12.852  < 2e-16 ***

Analysis of Variance Table

Response: actv
           Df Sum Sq Mean Sq   F value    Pr(>F)
cage        1  414.5   414.5   94.4     < 2.2e-16 ***
strain      1 6084.0  6084.0 1386.4372 < 2.2e-16 ***
sex         1   33.7    33.7    7.6816  0.005842 **
age         1  724.9   724.9  165.1841 < 2.2e-16 ***
Residuals 395 1733.3    4.39
```

Does this ANOVA table contain the type I SS or type III SS for the model? Fill in the blanks in the table. Why is the p-value for cage variable so different between the coefficient table and the ANOVA table? What is your conclusion now regarding whether cage has any effect on mouse physical activity?

This ANOVA table contains the type I SS. If it contains the type III SS, the p-values of F-test and Wald tests should be the same for all coefficients. Cage has strong correlation with physical activity, that is why it has a small p-value in the type I SS ANOVA table. However, after conditioning on other covariates, it is not significantly associated with physical activity.

4. (6pts) Next Dr. Cage did another regression analysis including all the covariates, and age^2. Here is the output of the Wald test p-value for each covariate and ANOVA table.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2271     4.3685   0.052  0.95857
cage         -0.2589     0.2245  -1.153  0.24965
strain        8.0183     0.2185  36.694  < 2e-16 ***
sex           0.6842     0.2157   3.171  0.00164 **
age           0.3999     4.3735   0.091  0.92720
age2          1.2516     1.0859   1.153  0.24977
```

```
Analysis of Variance Table

Response: actv
           Df Sum Sq Mean Sq   F value    Pr(>F)
cage        1  414.5   414.5   94.5434 < 2.2e-16 ***
strain      1 6084.0  6084.0 1387.5902 < 2.2e-16 ***
sex         1   33.7    33.7    7.6880  0.005823 **
age         1  724.9   724.9  165.3215 < 2.2e-16 ***
age2        1    5.8     5.8    1.3285  0.249772
Residuals 394 1727.5     4.4
```

Now both age and age^2 are insignificant in the coefficient table, should we drop both of them from the model? Why? Conduct an ANOVA test comparing this model with a model with variables intercept, cage, strain and sex. You already have all the numbers you need to calculate the test-statistic. You should provide H_0, the test statistic, the degrees of freedom but you do not need to calculate the p-value.

No. We should not drop both age and age^2. From the added in-order test, it is already obvious that age is significant given all the other variables except age^2.

$H_0 : \beta_{\text{age}} = \beta_{\text{age}^2} = 0.$
$F = \frac{[RSS(\text{smaller model}) - RSS(\text{larger model})]/2}{RSS(\text{larger model}/394)} = \frac{(724.9+5.8)/2}{1727.5/394} = 83.$ Under

H_0, F should follow a F-distribution with degrees of freedom 2 and 394.