Assignment 4

Emil Hafeez (eh2928)

10/28/2020

Problem 1 (5p)

In the context of ANOVA models, prove the partitioning of the total variability (sum of squares).

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{\overline{y}}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_l\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_i-\overline{\overline{y}}\right)^2 \tag{1}$$

We assume that the samples are drawn independently from the underlying populations, that the distributions of the error terms are normal, and that the variances of the k populations are equal (and thus that the variance of the outcome does not depend on the sample).

Let n be the total number of observations

Now, we can write that the total variation is equal to $\sum_{i=1}^{n}(y_i-\overline{y})^2$, giving the sum of squares for these data where y_i is the ith data point, and where $\bar{y}$ is the estimate of the mean. Now, more specifically,

Let $\bar{y}$ be the mean from each of the the i^{th} group where i = 1,2,3,... k is the number of groups and let $\overline{\overline{y}}$ be the grand mean, such that $\overline{\overline{y}}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}}{n}$.

....Continue

Problem 2

Part a)

Generate descriptive statistics for each group and comment on the differences observed. (4p)

Descriptives: let's do mean for each group and sd, range, and median for each group, and comment.

Regarding participants' physical status before therapy, the "Above Average" group had a mean of 23.57 days (standard deviation = 4.20), and a median of 22 days. The "Average" group had a mean of 33 days to recover (standard deviation = 3.92 days), and a median of 32 days. The "Below Average" group had a mean of 38 days to recover (standard deviation = 5.478 days) and a median of 40 days to recover. Thus, it appears that worse prior physical status is associated with more days to recovery. Interestingly, it also appears that a worse prior physical status seems to have higher dispersion (as per the SD, as well). Finally, it appears visually that only the "Below" group had left skew, whereas the other two groups had right skew. This can be explored visually, as below.

Source	Sum of Square (SS)	Degrees of Freedom (df)	Mean Sum of Square	F-Statistic
Between	Between SS	k-1	Between SS / (k-1)	F = $\frac{(MeanBetweenSS/k-1)}{(WithinSS/(n-k)}$
Within	Within SS	n-k	Within SS / (n-k)	
Total	Between SS + Within SS	n-1		

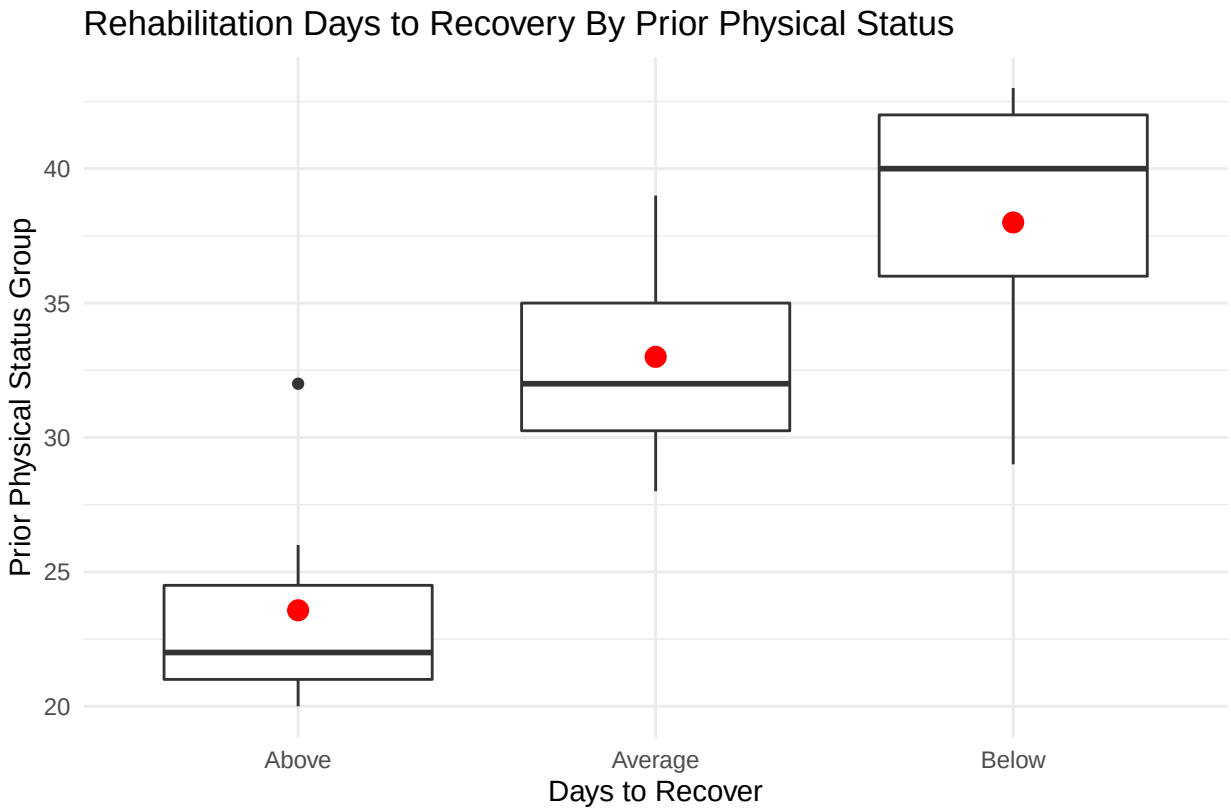

Figure 1, Biostatistics P8130 Assignment 4

Part b)

Let $\alpha = 0.01$.

$H_0 =$

$H_A =$

$F = \frac{(MeanBetweenSS/k-1)}{(WithinSS/(n-k)}$ $F_{critical} =$

Perform an ANOVA test: are the mean insulin levels significantly different? Need to mention the independent variable as a factor; o/w will be considered continuous Function lm() is broader, including linear regression models res<-lm(insulin~factor(ind), data=new_data)