Final Exam Fall 2018

The following datasets and models are used to answer the questions below.

mydata contains the original variable values and the transformed values.

```
str(mydata)
```

```
'data.frame':  392 obs. of  6 variables:
 $ mpg             : num  18 15 18 16 17 15 14 14 14 15 ...
 $ horsepower      : num  130 165 150 150 140 198 220 215 225 190 ...
 $ weight          : num  3504 3693 3436 3433 3449 ...
 $ mpg_lp10k       : num  1.31 1.57 1.31 1.47 1.38 ...
 $ horsepower_100KW: num  0.969 1.23 1.119 1.119 1.044 ...
 $ weight_mtons    : num  1.59 1.68 1.56 1.56 1.56 ...
```

mydata_cent contains the values centered along the mean.

```
str(mydata_cent)
```

```
'data.frame':  392 obs. of  4 variables:
 $ mpg_lp10k_c       : num  0.182 0.443 0.182 0.345 0.259 ...
 $ horsepower_100KW_c: num  0.19 0.451 0.34 0.34 0.265 ...
 $ weight_mtons_c    : num  0.239 0.325 0.208 0.207 0.214 ...
 $ power2weight_c    : num  1.07 1.28 1.25 1.25 1.17 ...
```

mymodel1 has weight as explanatory variable and mileage as response variable.

```
mymodel1 = lm(mydata_cent$mpg_lp10k_c ~ mydata_cent$weight_mtons_c)
```

mymodel2 has weight and power in watts as explanatory variables and mileage as response variable.

```
mymodel2 = lm(mydata_cent$mpg_lp10k_c ~ mydata_cent$weight_mtons_c + mydata_c
ent$horsepower_100KW_c)
```

mymodel3 has weight and power2weight as explanatory variables and mileage as response variable.

```
mymodel3 = lm(mydata_cent$mpg_lp10k_c ~ mydata_cent$weight_mtons_c + mydata_c
ent$power2weight_c)
```

mymodel4 has Principal Components PC1 and PC2 as explanatory variables and mileage as response variable.

```
mymodel4 = lm(mydata_cent$mpg_lp10k_c ~ PCs1$scores[,1] + PCs1$scores[,2])
```

Model 1

Question 1: Why is the Intercept coefficient so close to zero?

Ans:

The intercept coefficient is close to zero because the dataset has been centered along the mean before conducting a linear regression on it. The first model uses mileage as the response variable and the y-intercept of the model shows the predicted response variable when the explanatory variable is zero. Here due to the centering of the data the mean of all the variables comes out to be zero which affects the value of the y-intercept.

Source Link: https://www.theanalysisfactor.com/center-on-the-mean/

Question 2: What proportion of fuel inefficiency variance is explained by weight variance?

Ans:

The correlation between the response variable (mileage) and the explanatory variable (weight) shows the proportion of fuel inefficiency variance explained by the model. From the output in the question set, it comes out to be 88.50%. The R output is shown below.

Output:

```
cor(mydata_cent$mpg_lp10k_c, mymodel1$fitted.values)
```

[1] 0.885056

Source Link: https://stats.stackexchange.com/questions/203540/explaining-the-variance-of-a-regression-model

Question 3: Vehicle weight in pounds has a variance of 721484.7, but converting to metric tons changes the variance to 0.148443. How does this transformation affect the proportion of variance explained by the model? Will that proportion be higher or lower than if the weight variable is expressed in pounds? Explain your reasoning using English sentences (instead of or in addition to calculations).

Ans:

There will be no change in the proportion of variance explained by the model due to this transformation. As the transformation is linear, only the variance of the variable changes keeping the amount of information contained in it intact. Such transformations can be said to be cosmetic as they help us see the data from a different angle or to standardize them before analysis.

Output:

```
cor(mydata_cent$mpg_lp10k_c, mymodel1$fitted.values)
```

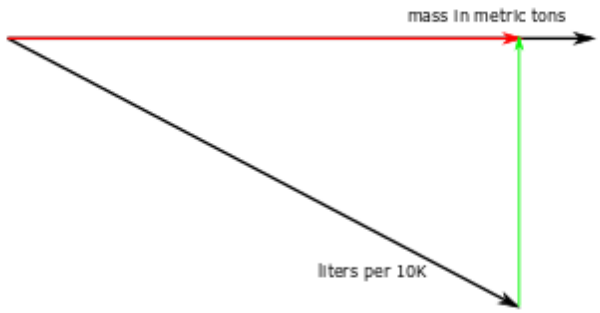
```
[1] 0.885056
```

```
cor(mydata_cent$mpg_lp10k_c, mydata$weight)
```

```
[1] 0.885056
```

As can be seen from the output above the amount of variance explained remains the same.

Source Link: https://blog.majestic.com/case-studies/correlation-data-transformations/



Question 4: Explain the relationship between the proportion of variance *not* explained by the model and the dimensions of the triangle in the diagram above.

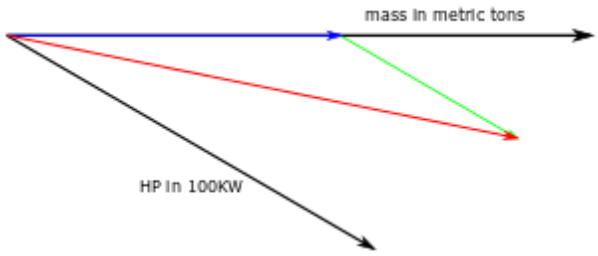
Ans:

The proportion of variance not explained by the model is 11.50% and it is called a residual. The relationship between them is that they form aright angled triangle. The actual values of the response variable is the hypotenuse, the fitted values and the residual the sides of the triangle. So, based on the Pythagorean theorem the relationship can be stated as:

$$\text{Model Values}^2 + \text{Residuals}^2 = \text{Actual Values}^2$$

Source Link: http://mosaic-web.org/go/StatisticalModeling/Chapters/Chapter-04.pdf

Model 2



Question 5: The fuel inefficiency response variable is not shown in the diagram above. Where is the vector for that fuel inefficiency variable in relation to the red vector of fitted values?

Ans:

The fuel inefficiency response variable is orthogonal to the red vector of the fitted values, so it is not shown in the diagram. The above diagram shows two explanatory variables and one response variable which requires a 3rd dimension is needed to show the response variable. So, the response variable is not shown in the above diagram.

Question 6: What do the blue and green vectors represent in the diagram above?

Ans:

The blue and the green vectors represent the proportion of weight and power variables in the fitted values. Since the model uses weight and power as explanatory variables so per vector addition they two will account for the proportion of the fitted values. As per the above diagram we can observe that they have different proportion of the fitted value.

Question 7: The second and third models each explain nearly the same amount of fuel inefficiency variance. Why is the lower correlation between the explanatory variables make the third model more attractive for understanding the relationship between engine power, vehicle weight, and fuel inefficiency?

Ans:

Having strong correlation between the explanatory variables is not desirable since a change in one variable can affect the other variable. Hence they can affect the regression coefficients of each of the explanatory variables. The third model is more attractive because it does not have high multicollinearity between the variables. They makes is more suitable to analyze the impact of the weight variable on the response variable. Having high multicollinearity in the model in not suitable to analyze the individual explanatory variables as the values of them become dependent on other explanatory variables in the model.

Source Link: https://onlinecourses.science.psu.edu/stat501/node/346/

```
> var(mpg$weight)
[1] 0.148443
> var(mpg$hp100KW)
[1] 0.08238541
```

Question 8: Are these two variances different enough to recommend using the correlation matrix instead of the covariance matrix? Justify your answer using evidence from the loadings.

Ans:

Correlation matrix is equal to the covariance matrix of the standardized dataset. This means that if the original dataset is standardized then we will get the same matrix as the result. Since our dataset in already centered along the mean but not standardized we will not get the same matrix. Using different matrix does not change the output of the analysis but the proportion of the variables in each component changes. One of the drawback of using a covariance matrix is that the variables having the maximum variance dominate the first principal component which limits our understanding of the relationships among the variables.

But as per the loadings given below both the components have proportional contribution on the components. Also, the difference of variance between them is not vast enough to use the correlation matrix. But I carried out a separate PCZ analysis of the original variables to find interesting results. Also, the variance difference between the original variables is very large.

```
> var(mydata$horsepower)

[1] 1481.569
> var(mydata$weight)

[1] 721484.7
```

PCs1 loadings:

```
Loadings:
                  Comp.1 Comp.2
horsepower_100KW_c  0.580  0.814
weight_mtons_c      0.814 -0.580

               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```

PCs2 loadings:

```
Loadings:
           Comp.1 Comp.2
horsepower         0.999
weight      0.999

               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```

```
summary(PCs2)

Importance of components:
                            Comp.1       Comp.2
Standard deviation     848.9695501 1.930491e+01
Proportion of Variance   0.9994832 5.168054e-04
Cumulative Proportion    0.9994832 1.000000e+00
```

As we can see from the above output each of the variables completely dominates the components but the second component account for very small proportion of the dataset which is not the case for our first PCA analysis. Hence I do not recommend using the correlation matrix for this dataset.

Question 9: The fourth model explains exactly the same proportion of variance as the second model. Why are those proportions exactly the same?

Ans:

To construct the forth model we used principal component analysis (PCA) to get our explanatory variables. The PCA was done by taking weights and power as the dataset. The PCA finds the direction of maximum variance among the dataset in the form of first principal component and then subsequent components find lesser variances direction orthogonal to each other. In our case the PC1 and PC2 components captures

93.82% and 6.18% of the variance of the dataset and together the capture 100% variance of the dataset. This means that together they capture all the information contained in both the variables (i.e. weight and power). As both these variables were used to construct the second model and both PC components were used to construct the forth model, they essentially use the same amount of information to model the response variable. So, the forth model and the second model explains exactly the same proportion of variance.

Source Link: https://onlinecourses.science.psu.edu/stat505/node/51/

Question 10: The Euclidean vector length of the second principal component is much shorter than the length of the first principal component. Under what circumstances would the second principal component have a longer vector than the first?

Ans:

The Euclidean vectors of the principal components signify the amount of variances captured by the principal components of the dataset. Since the first principal component always has a bigger eigenvalue than the rest of the components which signifies that it has captured the maximum amount of variance of the dataset so, the Euclidean vector length of second principal component can never be greater than the first principal component.

Question 11: In the fourth diagram the red vector of fitted values doesn't lie between the explanatory variables, as it does in the other diagrams. Why is that the case?

Ans:

The explanatory variables PC1 and PC2 here are orthogonal to each other due to the property of the principal component analysis. The vector of fitted values does not lie in between the explanatory variables because both the components point in the direction of the highest and second highest variance which may not lie on wither direction of the red vector. All the red vector of fitted values may not lie between the directions on maximum variance.