Linear Regression & ANOVA HW

Daphney

2024-05-21

Part A. Variables

In the field of psychology, much research is done using self-report surveys using Likert scales (look it up!).

A1

What type of variable is a Likert response? (1 pt) A Likert response is an ordinal variable because the responses represent categories with a specific order or ranking without equal intervals. It could also be a dependent variable, also known as the response variable. It measures the outcome that is influenced by the independent variable.

A2

What are some (at least 2) benefits of using Likert scales? (2 pts) Likert scales provide a straightforward way to quantify data for statistical analysis. They are also easy for participants to understand and complete it, leading to higher response rates and more accurate data collection.

A3

What are some drawbacks — and dangers — of using them? Make sure you mention at least one 'drawback' and one 'danger' (a 'drawback' is a shortcoming, while a 'danger' implies potential harm). (2 pts) Likert scales can overly simplify complex opinions. Due to the limited predefined options, It may not capture all of the participant's thoughts and feelings. If a participant has mixed feelings, they may be forced to choose a single option that does not fully represent their view. This can lead to a loss of valuable qualitative data. It also does not provide context for the reasons participants are choosing certain answers. This could make it difficult for us to understand the underlying reasoning.

Part B. Simple Linear Regression

Perform linear regressions on a dataset from a European Toyota car dealer on the sales records of used cars (Toyota Corolla). We would like to construct a reasonable linear regression model for the relationship between the sales prices of used cars and various explanatory variables (such as age, mileage, horsepower). We are interested to see what factors affect the sales price of a used car and by how much.

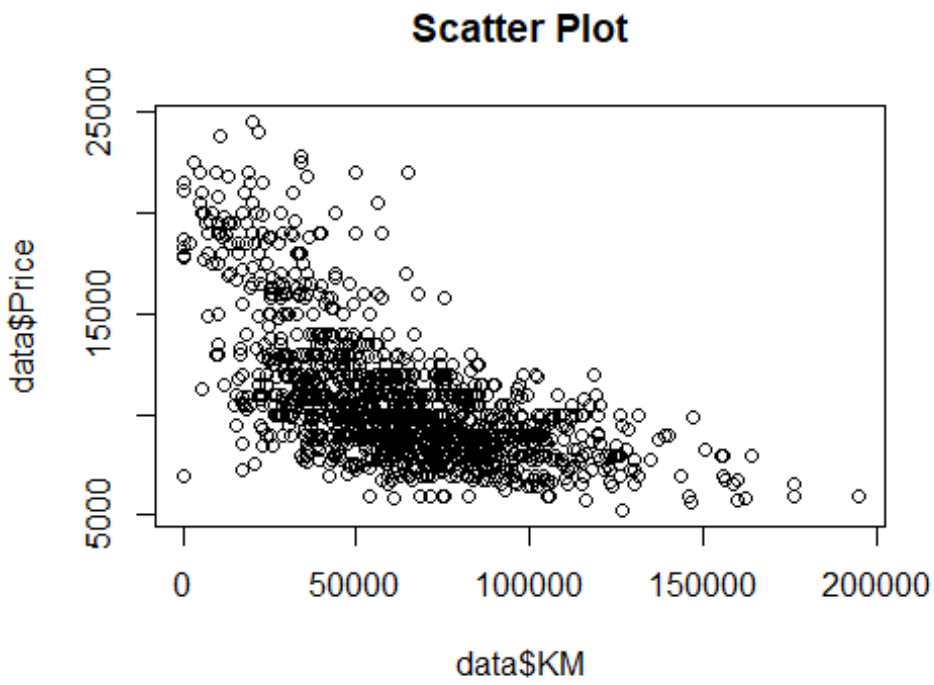
```r
# Read in the data
data = read.csv("UsedCars.csv",sep = ",",header = TRUE)
# The first three rows of data
head(data, 3)
```

```
##   Id                                                Model Price Age
KM  HP
## 1  1            TOYOTA Corolla 1800 T SPORT VVT I 2/3-Doors 21500  27
19700 192
## 2  2      TOYOTA Corolla 1.8 VVTL-i T-Sport 3-Drs 2/3-Doors 20950  25
31461 192
## 3  3 TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT BNS 2/3-Doors 19950  22
43610 192
##   Metallic Automatic   CC Doors Gears Weight
## 1        0         0 1800     3     5   1185
## 2        0         0 1800     3     6   1185
## 3        0         0 1800     3     6   1185
```

Question B1: Exploratory Data Analysis

a. **3 pts** Use a scatter plot to describe the relationship between Price and the Accumulated kilometers on odometer. Describe the general trend (direction and form). Include plots and R-code used.

```r
library(ggplot2)
plot(data$KM, data$Price, main = "Scatter Plot")
```



It appears most of the data points are concentrated below 100,000 Kilometers. As the kilometers increase, the data points become more spread out. It appears there may be some outliers, particularly in the range above 150,000 kilometers where the data points are sparse.

b. **3 pts** What is the value of the correlation coefficient between *Price* and *KM*? Please interpret the strength of the correlation based on the correlation coefficient.

```
corr <- cor(data$KM, data$Price)
```

There is a moderately strong negative correlation in the data, meaning that as one variable increases, the other variable tends to decrease. The correlation is -0.61. As kilometers increase, price decreases.

c. **2 pts** Based on this exploratory analysis, would you recommend a simple linear regression model for the relationship?

Yes but I think the model would benefit from variable transformation. The model appears to violate the homoscedasticity assumption. Due to the bunching of data points, the variance does not appear to be constant. We can still run the model, but the coefficient may possibly be biased.

d. **1 pts** Based on the analysis above, would you pursue a transformation of the data? *Do not transform the data.*

I think the data can benefit from transformation. The spread of the data points increases as KM increases, suggesting heteroscedasticity. There also appears to be some outliers, which can skew the results of the regression. Transformation can remove or reduce outliers. We can "stretch" out the data more and improve the model.

Question B2: Fitting the Simple Linear Regression Model

Fit a linear regression model, named *model_1*, to evaluate the relationship between UsedCars Price and the accumulated KM. *Do not transform the data.* The function you should use in R is:

```
model_1 <- lm(Price ~ KM, data = data)
summary(model_1)
```

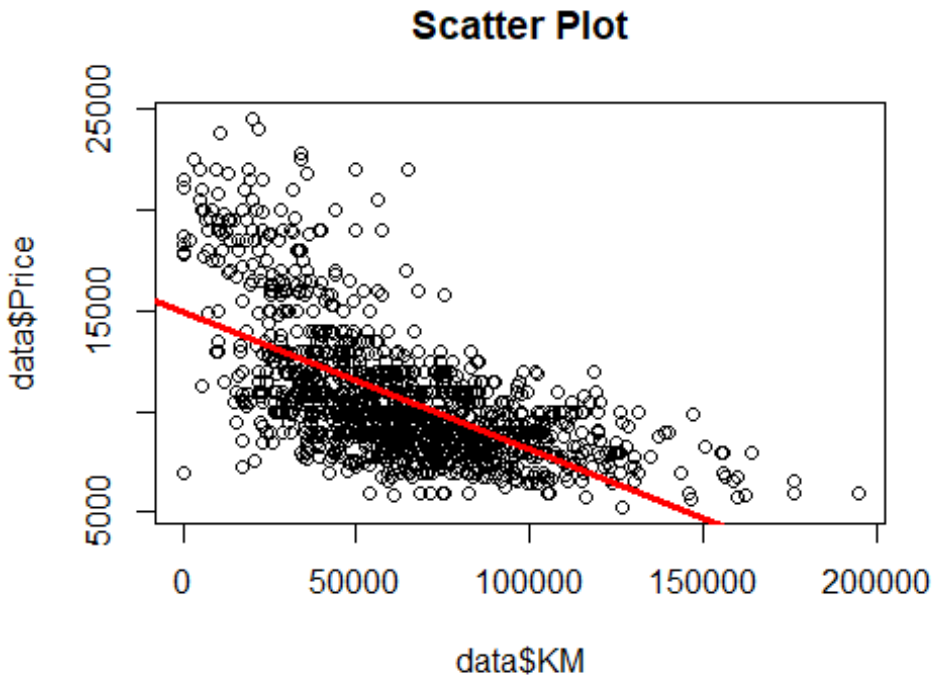
```
##
## Call:
## lm(formula = Price ~ KM, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7993  -1774   -457   1394  11437
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.494e+04  1.693e+02   88.24   <2e-16 ***
## KM          -6.817e-02  2.439e-03  -27.95   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2615 on 1262 degrees of freedom
## Multiple R-squared:  0.3824, Adjusted R-squared:  0.3819
## F-statistic: 781.4 on 1 and 1262 DF,  p-value: < 2.2e-16
```

```
model_1$coefficients
```

```
##   (Intercept)            KM
##  1.494316e+04 -6.817549e-02
```

```
plot(data$KM, data$Price, main = "Scatter Plot")
abline(model_1, col = "red", lwd =3)
```




```
cor.test(data$KM , data$Price, method ='pearson')
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$KM and data$Price
## t = -27.954, df = 1262, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6513175 -0.5831322
## sample estimates:
##       cor
## -0.6183873
```

a. **3 pts** What are the model parameters and what are their estimates?

Intercept $\beta 0$ = 1.494316e+04 , Slope $\beta 1$= -6.817549e-02, Residual standard error σ^2= 2615

b. **2 pts** Write down the estimated simple linear regression equation.

Price = 14940 - 0.06817*KM

c. **2 pts** Interpret the estimated value of the β_1 parameter in the context of the problem.

For every kilometer driven, the expected decrease in the price is 0.06817 euros. The negative slope is consistent with the negative correlation observed between KM and price, indicating as mileage on a car increases, its price decreases.

d. **2 pts** Find a 95% confidence interval for the β_1 parameter. Is β_1 statistically significant at this level?

```
confint(model_1, level = 0.95)
```

```
##                    2.5 %        97.5 %
## (Intercept)  1.461094e+04  1.527539e+04
## KM          -7.296019e-02 -6.339078e-02
```

B1 is statistically significant. The p-value associated with KM, which is < 2e-16, indicating a very strong significance (less than the significance level of 0.05). This implies that the slope coefficient is significantly different from zero

e. **2 pts** Is β_1 statistically significantly negative at an α-level of 0.01? What is the approximate p-value of this test?

```
B1 = coef(model_1)["KM"]
```

B1 is indeed statistically significantly negative at an alpha level of 0.01. The coefficient sign is negative and the p-value associated with B1, in my output, is 2e-16 which is much less than the alpha level 0.01. We have evidence to reject the null hypothesis that suggests the coefficient is equal to zero.

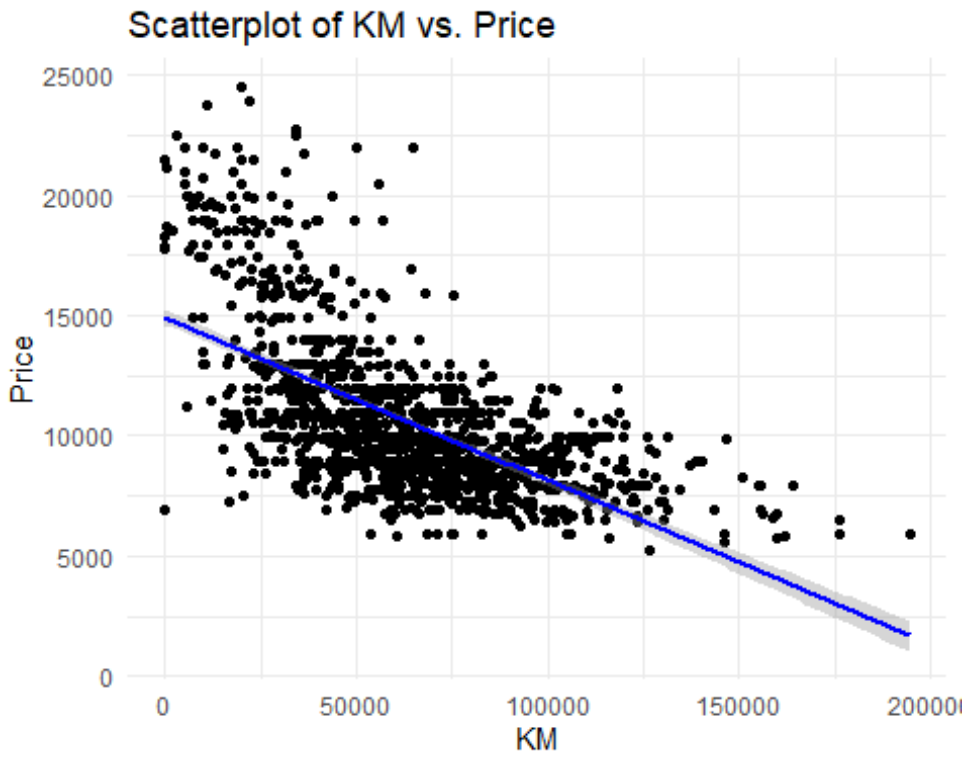
Question B3: Checking the Assumptions of the Model

Create and interpret the following graphs with respect to the assumptions of the linear regression model. In other words, comment on whether there are any apparent departures from the assumptions of the linear regression model. Make sure that you state the model assumptions and assess each one. Each graph may be used to assess one or more model assumptions.

a. **3 pts** Scatterplot of the data with *KM* on the x-axis and *Price* on the y-axis. Make sure you include a line showing the overall trend of the scatterplot

```
scatter_plot <- ggplot(data, aes(x = KM, y = Price)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Scatterplot of KM vs. Price",
```

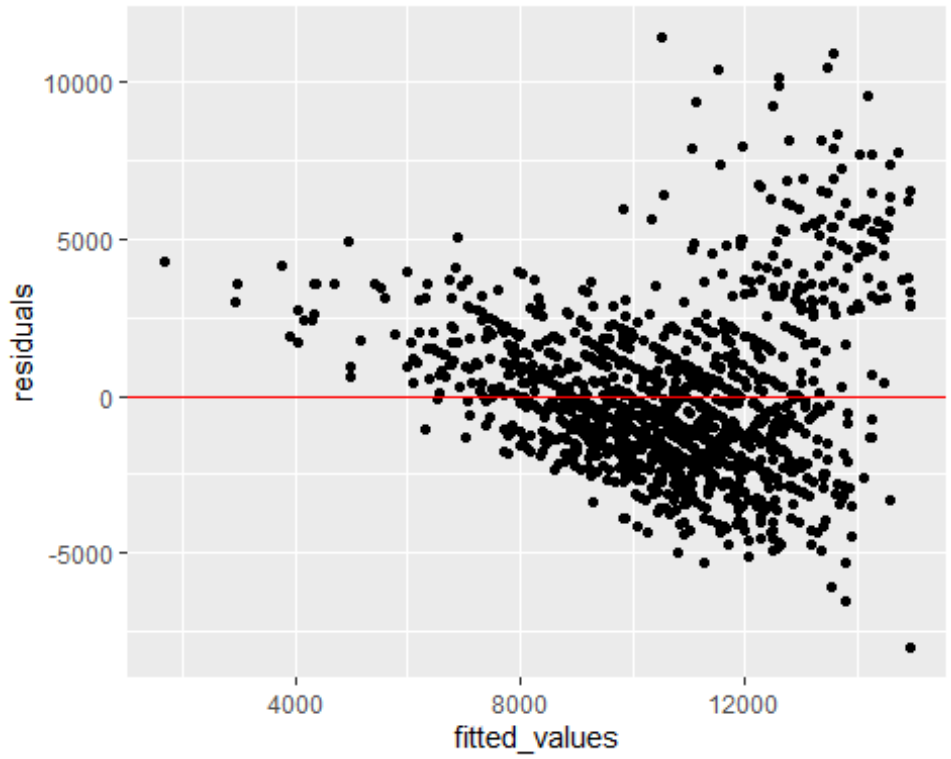
```
       x = "KM",
       y = "Price") +
  theme_minimal()
scatter_plot

## `geom_smooth()` using formula = 'y ~ x'
```



The scatterplot shows a negative linear trend between KM and Price, meaning as the kilometers driven increase, the price tends to decrease. This suggests that there is a linear relationship between the two variables and teh linearity assumption is met. However, the spread of the data points around the regression line shows variability, which will be further assessed in the residual plot and other diagnostics.

b. **4 pts** Residual plot - a plot of the residuals, $\hat{\epsilon}_i$, versus the fitted values, $\hat{y}_i$. Make sure you include a line showing the ideal baseline (hint: residual = 0) that serves as the comparison

```
residuals <- residuals(model_1)
fitted_values <- fitted(model_1)
residual_plot <- ggplot(data = data.frame(fitted_values, residuals), aes(x =
fitted_values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, col = "red")
residual_plot
```

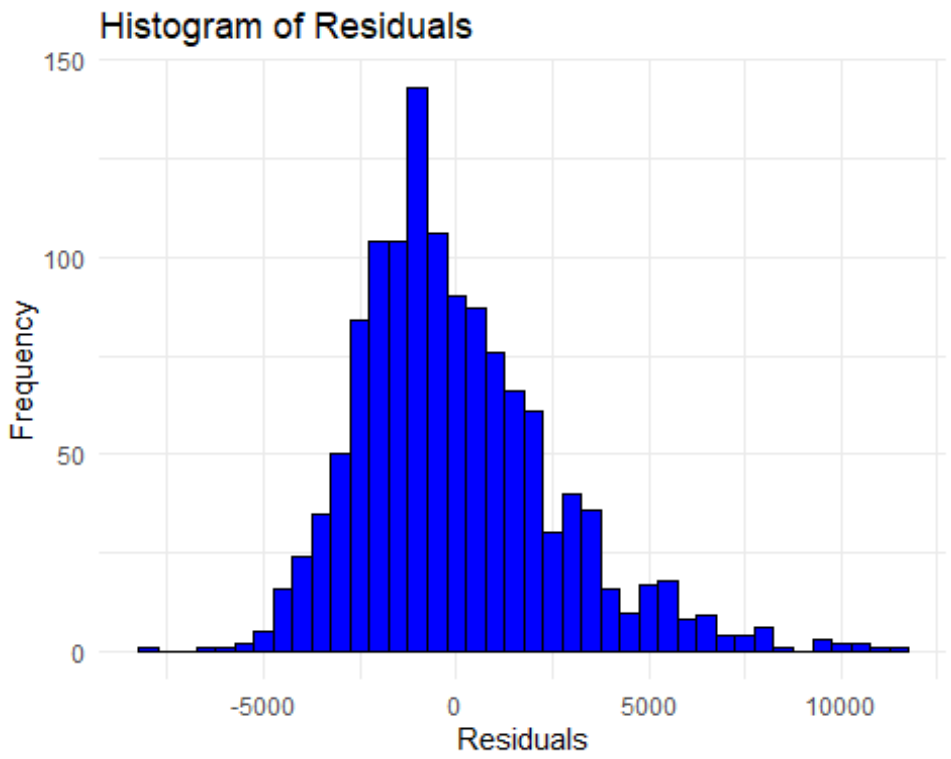


The residuals appear to be randomly scattered and do not show a clear pattern, indicating that there is no violation of the independence assumption.

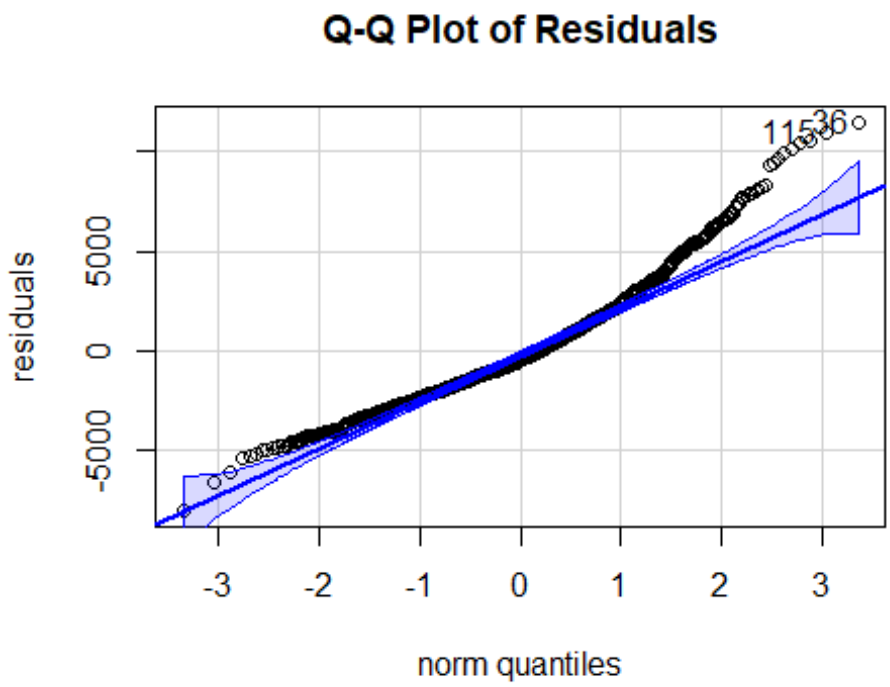
The homoscedasticity assumption requires that the residuals have constant variance across all levels of the independent variable. The residual plot should ideally show a "cloud" of points with no clear funnel shape. However, there appears to be some funneling which suggests a potential issue with heteroscedasticity.

c. **4 pts** Histogram and q-q plot of the residuals. Make sure you include a line in the q-q showing the ideal baseline that serves as the comparison in a q-q plot

```
histogram <- ggplot(data.frame(residuals), aes(x = residuals)) +
  geom_histogram(binwidth = 500, color = "black", fill = "blue") +
  labs(title = "Histogram of Residuals",
       x = "Residuals",
       y = "Frequency") +
  theme_minimal()
histogram
```




```
library(readr)
library(car)

## Loading required package: carData

qq_plot <- qqPlot(residuals, main = "Q-Q Plot of Residuals")
```




```
qq_plot
```

```
## [1]  36 115
```

Most of the points lie along the line, but there are deviations at the tails. This indicates that the residuals are approximately normally distrubted, but there are some deviations from normality, especially in the tails. Overall, the Q-Q plot suggests that the residuals are roughly normally distributed but with some outliers or deviations in the extremes.

Question B4: Prediction

Use the results from both model_1 to discuss the effects of KM on the dependent variable: Holding everything else equal, how much the sales price would decrease if a car accumulated 10,000 more kilometers? What observations can you make about the result in the context of the problem? (3 pts)

```
slope <- coefficients(model_1)["KM"]
decrease_per_10000_km <- slope * 10000
decrease_per_10000_km
```

```
##        KM
## -681.7549
```

The price would decrease by 681.7 euros. The negative correlation indicates a moderate inverse relationship between kilometers driven and the car's price. As the number of kilometers increases, the price tends to decrease.This information is useful for both sellers

and buyers in the used car market. Sellers can estimate the depreciation in value with increased usage, while buyers can anticipate how the price varies with the car's mileage.

Part C. Experiment!

You work for the National Park Service (NPS), and you absolutely love bears. Describe an imaginary (it can be realistic) scenario in which you get to run a one-way ANOVA on a few (3+) species of bears.

Part D1

What are you comparing (name the variable!)? What do you hope to learn from ANOVA? (2 pts)

I want to compare the behavior of three different bear species in various national parks. I will run a one way ANOVA to compare the average number of fish caught per day among the three species of bears. I am aiming to determine whether there are significant differences in the average number of fish caught per day among the three species of bear. This could help me to understand species-specific foraging behaviors and habitat requirements.

Part D2

Imagine that the results are "mixed", meaning you can draw some conclusions and not others. Describe your conclusions and make sure you detail, with reference to your ANOVA, why the results were "mixed." (3 pts)

I determined Grizzly bears and polar bears show a statistically significant difference in the average number of fish caught per day. However, black bears did not have any statistically significant differences with the other species groups. Based on what I found, I concluded that Grizzly bears and Polar bears show distinct foraging behaviors, possibly due to different habitats or physical adaptations. As for Black bears, it may have similar behaviors to both Grizzly and Polar bears, or the sample size for Black bears might not be sufficient to detect a significant difference.

Part D3

Now imagine that you have just been granted 3 months and $50,000 to continue this study (you're a great grant writer and a very likable member of the NPS!). Describe some next steps you would take to clarify, reinforce and/or further explore your nascent investigation. You MUST reference using a 'controlling' variable somehow in your response. (5 pts)

I would increase the sample size by collecting more data on the number of fish caught per day across a larger number of bears in each species. I would use several control variables in the experiment such as habitat type, fish availability, season, and the age of the bears. This will help to better understand the factors influencing hunting success.

Part D. Explain the meaning of a p-value!

Explain in detail what it means specifically — in a statistical sense — for any result to be "statistically significant" at a particular -level. In other words, explain the meaning and use of p-values. You should research this question, and you should expect your answer to be at least a paragraph long. (3 pts)

A p-value is a measure used in statistical hypothesis testing to determine the significance of the observed results. Specifically, it represents the probability of obtaining results at least as extreme as the ones observed, assuming that the null hypothesis is true. The null hypothesis represents a statement of no effect or no difference. A p-value of 0.03 means that there is a 3% probability that the observed results, or more extreme ones, could occur if the null hypothesis were true. When a result is "statistically significant" at a particular α-level, it means that the p-value is less than or equal to the predefined α-level. Statistical significance suggests that the observed effect is unlikely to be due to random chance. For example, with a p-value of 0.03 and an alpha level is 0.05, we could reject the null hypothesis and conclude that there are significant differences in the mean number of fish caught among the bear species since the p value is smaller than the alpha level. The lower the p-value, the stronger the evidence against the null hypothesis.