AMS 380.01: Exam 1

Due on 03/25

Wei Zhu

Harris Temuri, 111354621

Problem 1

The built-in R data set iris was introduced by the British statistician and biologist Ronald A. Fisher in his 1936 paper entitled "The use of multiple measurements in taxonomic problems". It is also referred to as Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). This sums to 150 records under 5 attributes – Petal.Length, Petal.Width, Sepal.Length, Sepal.Width and Species.

Solution

(a) Please draw side-by-side box plots to visually compare the Sepal.Width of the 3 Species.

```
    # 1.1 Creating a box plot of the data
    ggplot(df, aes(x=Species, y=Sepal.Width)) + geom_boxplot()
```

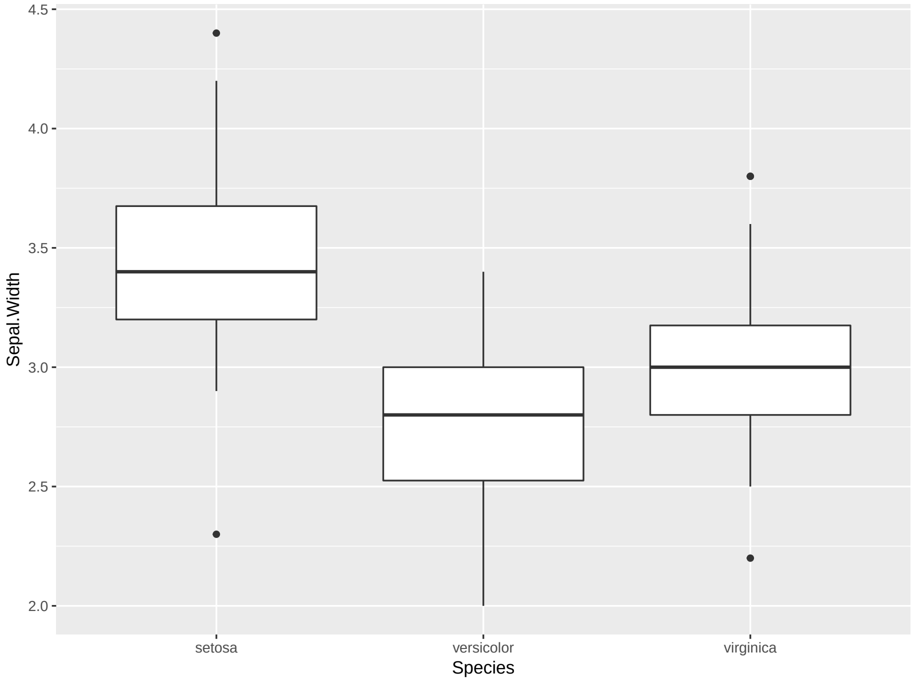

Figure 1: Box Plot comparing Sepal.Width to Species

Problem 1 continued on next page. . .

(b) Test at $\alpha = 0.05$ whether the mean Sepal.Width of the 3 Species are equal. What assumptions are necessary? Please test these assumptions.

```
# 1.2 Testing Normality and equal variance assumptions
# Normality Test
# for setosa
shapiro.test(df$Sepal.Width[df$Species=="setosa"])
# for versicolor
shapiro.test(df$Sepal.Width[df$Species=="versicolor"])
# for virginica
shapiro.test(df$Sepal.Width[df$Species=="virginica"])

# Equal variance assumptions test
bartlett.test(Sepal.Width ~ Species, data=df)

# 1.2 2 test for equal means
summary(aov(Sepal.Width ~ Species, data=df))
```

The p-values for the normality test are: 0.2715, 0.338, 0.1809 which are all above 0.05 so they are all normal. The p-value of the variance test is 0.3515 so do not reject the null hypothesis and the variances are equal. The p-value for the test for equal means is $<$ 2e-16 so reject the null hypothesis, the means are not the same.

(c) At the familywise error rate of $\alpha = 0.05$, please perform pairwise comparison of the Sepal.Width of the 3 Species using the Tukey HSD test.

```
# 1.3 Tukey Test
TukeyHSD(aov(Sepal.Width ~ Species, data = df))
```

Species versicolor and setosa: p-value is $<<$ 0.05 which means they are significantly different,
Species virginica and setosa: p-value is $<<$ 0.05 which means they are significantly different,
Species virginica and versicolor: p-value is $<<$ 0.05 which means they are significantly different.

(d) Please compare the Sepal.Width of the Species 'virginica' and 'versicolor' using the usual pooled-variance t-test at the significance level $\alpha = 0.05$. What assumptions are necessary? Please test these assumptions.

Already tested for normality. Testing for equal variances between the two and then t-test.

```
# 1.4 Comparing veriginica to versicolor
virgDf <- df$Sepal.Width[df$Species=="virginica"]
versDf <- df$Sepal.Width[df$Species=="versicolor"]
# Equal variance test
var.test(virgDf, versDf)
# T-test
```

```
    t.test(virgDf,versDf,var.equal = T)
```

The p-value for for the var test was above 0.05 so the variances are equal. The p-value for the t-test is 0.001819 which is less than 0.05 so the means are different.

Problem 2

What are the key factors behind a diamond's price? The data set diamond.csv contains the price (price in US dollars) of 1,000 diamonds along with 4 other variables: carat (weight of the diamond), color (diamond color), depth (total depth percentage) and table (width of top of diamond relative to widest point).

Solution

(a) Please fit a least squares regression between 'price' and 'carat', and write down the equation of the fitted line. What is the Pearson correlation between these two variables? What is the coefficient of determination for the given regression? Would you say there is a strong linear relationship between these two variables? What assumptions appear to be unattainable so that you cannot conduct the usual significance test of a linear relationship?

```
    # 2.1 Fit least squares regression
    fit <- lm(price ~ carat, data = diaDf)
    summary(fit)
    # Pearson Correlation Test
    cor.test(x=diaDf$carat,y=diaDf$price)
    # Normality Test
    shapiro.test(fit$residuals)
    # Homoscedasticity Test
    plot(fit)
    lmtest::bptest(fit)
```

$$Price = -2243.62 + 7644.24 * Carat$$

The Pearson correlation between the two variables is 0.9135685. The coefficient of determination is 0.8346 which is an okay linear relationship. The shapiro test and test for homoscedasticity however both had a p-value of less than 0.05 so it is not normal and not homoscedastic.

(b) Please fit the general linear model with the response variable being 'price' and a single predictor being 'color'. Please write down your regression model and point out which color group is the baseline group. Is color a significant predictor of price based on your analysis? Please report the p-value. What are the assumptions necessary for your test? Please test these assumptions.

```
# 2.2 Linear model fit price and color
fit2 <- lm(price ~ color, data = diaDf)
summary(fit2)
# Normality Test
shapiro.test(diaDf$price[diaDf$color=="D"])
shapiro.test(diaDf$price[diaDf$color=="E"])
shapiro.test(diaDf$price[diaDf$color=="F"])
shapiro.test(diaDf$price[diaDf$color=="G"])
shapiro.test(diaDf$price[diaDf$color=="H"])
shapiro.test(diaDf$price[diaDf$color=="I"])
shapiro.test(diaDf$price[diaDf$color=="J"])
# variance test
bartlett.test(price ~ color, data=diaDf)
```

$$Price = 3152.97 + 87.55I\textit{(colorE)} + 361.51I\textit{(colorF)} + 523.34I\textit{(colorG)}+$$

$$790.44I\textit{(colorH)} + 1707.34I\textit{(colorI)} + 1944.82I\textit{(colorJ)}$$

Color D is the baseline group. The colors I and J are significant predictors of the price. The p-value of the fit is 0.0005746. Tested for normality and equal variances. All of the colors had a p-value of less than 0.05 so they are not normal. The p-value of the bartlett test was also less than 0.05 so the variances were not equal.

(c) We shall use ‘price’ as the response variables, and there are a total of 4 regressors to choose from. Please note that we have one categorical variable ‘color’. Now based on the necessary dummy variables, how many regressors in total for us to choose from? Please use the R function regsubsets() [leaps package] for best-subset variable selection to identify different best models of different sizes ranging from 1 to the maximum number of regressors (in terms of the necessary dummy variables).

```
# 2.3 Test for best regressors
models <- regsubsets(price~., data = diaDf, nvmax = 9)
summary(models)
```

We have 9 regressors in total to choose from.

The best model with one variable is price carat

The best model with two variables is price carat + colorJ

The best model with three variables is price carat + colorI + colorJ

The best model with four variables is price carat + colorH + colorI + colorJ

The best model with five variables is price carat + colorH + colorI + colorJ + table

The best model with six variables is price carat + colorH + colorI + colorJ + depth + table

Problem 2 continued on next page. . .

The best model with seven variables is price carat + colorE + colorH + colorI + colorJ + depth + table

The best model with seven variables is price carat + colorE + colorG +colorH + colorI + colorJ + depth + table

The best model with seven variables is price carat + colorE + colorF + colorG +colorH + colorI + colorJ + depth + table

(d) Please use the 5-fold cross-validation to select the best overall model from all these best subset models identified in part (c) above. Please write down the equation of this best overall model.

```r
# 2.4 K-fold Cross validation
get_model_formula <- function(id, object, outcome){
  # get models data
  models <- summary(object$which[id,-1])
  # Get outcome variable
  form <- as.formula(object$call[[2]])
  outcome <- all.vars(form)[1]
  # Get model predictors
  predictors <- names(which(models == TRUE))
  predictors <- paste(predictors, collapse = "+")
  # Build model formula
  as.formula(paste(outcome, "~", predictors))
}

get_cv_error <- function(model.formula, data){
  set.seed(1)
  train.control <- trainControl(method = "cv", number = 5)
  cv <- train(model.formula, data = data, method = "lm",
              trControl = train.control)
  cv$results$RMSE
}

model.ids <- 1:9
cv.errors <- map(model.ids, get_model_formula, models, "price") %>%
  map(get_cv_error, data = diaDf) %>%
  unlist()
cv.errors
```

Keep getting error, " <text>:2:0: unexpected end of input1: price ".

(e) Please use the R function stepAIC() [MASS package] to identify the best model using the stepwise variable selection method. Please write down the equation of this best overall model.

```r
# 2.5 Step AIC
res.lm <- lm(price~., data = diaDf)
```

```
    step <- stepAIC(res.lm,direction = "both", trace = FALSE)
    step
```

The best function is:

$$Price = 11968.864 + 8062.951\,carat - 33.646\,I(colorE) - 3.684\,I(colorF) - 10.461\,I(colorG) -$$

$$544.680\,I(colorH) - 970.487\,I(colorI) - 1837.177\,I(colorJ) - 128.112 * depth - 110.764 * table$$