Assignment 2: Group 45

Daniel Engbert, Rik Timmer, Koen van der Pool

03 March 2023

Note: we made a function `checkNorm()` which prints a histogram, qqplot, and p-value from the shapiro-wilk normality test. And we made a function `printPval()` which simply prints a given p-value to 3 significant figures. We utilize both functions throughout this assignment.

Exercise 1: Trees

1 a)

```
trees = read.table("treeVolume.txt", header=T)
model = lm(volume~type, data=trees)
print("model coefficients:"); summary(model)$coefficients
```

```
## [1] "model coefficients:"
```

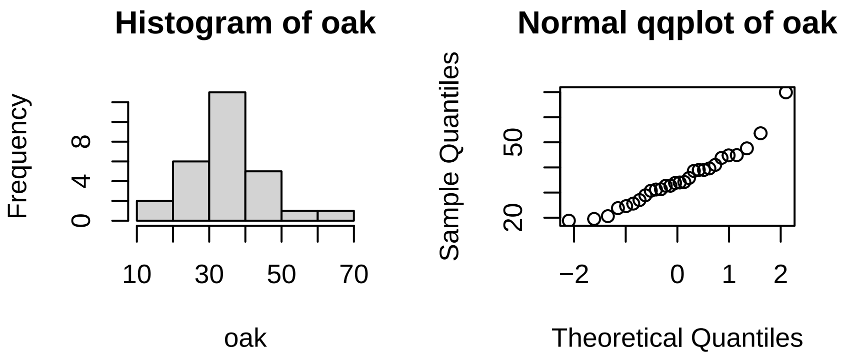
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.17       2.54   11.88 4.68e-17
## typeoak         5.08       3.69    1.38 1.74e-01
```

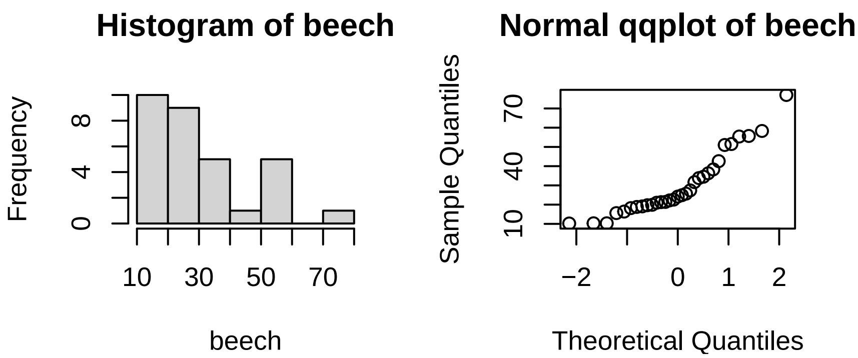
```
res = anova(model)
sprintf("ANOVA p-value for type = %.3f", res["type", "Pr(>F)"])
```

```
## [1] "ANOVA p-value for type = 0.174"
```

The p-value $0.174 > 0.05$ for the type in the ANOVA analysis of the linear model, suggests there's insufficient evidence to reject the H_0 (that tree type influences volume).

```
## [1] "Shapiro-Wilk normality p-value for oak: 0.082"
```






```
## [1] "Shapiro-Wilk normality p-value for beech: 0.004"
```

```
## [1] "oak mean volume = 35.250, beech mean volume = 30.171"
```

We can split the data into two samples of tree volume based on the tree types, and compare the means of the samples using a t-test to determine whether, based on this data, there is a significant difference in mean volume between the two tree types. As can be seen in the output of the t-test $0.166 > 0.05$, signifying once again that there is not enough evidence to reject the null hypothesis that the means of the samples are the same. This concurs with the results of the ANOVA.

```
new_oak = data.frame(type="oak"); new_beech = data.frame(type = "beech")
pred1 = predict(model, new_oak); pred2 = predict(model, new_beech)
sprintf("predicted volumes: oak = %.3f, beech = %.3f", pred1, pred2)
```

```
## [1] "predicted volumes: oak = 35.250, beech = 30.171"
```

1 b)

```
model = lm(volume~type*diameter + height, data=trees)
res = anova(model)
sprintf("ANOVA p-value for type:diameter = %.3f", res["type:diameter", "Pr(>F)"])
```

```
## [1] "ANOVA p-value for type:diameter = 0.474"
```

We built a linear model that added an interaction term between diameter and type, the p-value $0.474 > 0.05$ for this term suggests there's insufficient evidence to reject the H_0 (that the influence of diameter on volume is the same for both tree types).

```r
model = lm(volume~type*height + diameter, data=trees)
res = anova(model)
sprintf("ANOVA p-value for type:diameter = %.3f", res["type:height", "Pr(>F)"])
```

```
## [1] "ANOVA p-value for type:diameter = 0.176"
```

Now running another linear model that includes an interaction term between height and type instead, the p-value $0.176 > 0.05$ for this term suggests there's insufficient evidence to reject the H_0 (that the influence of height on volume is the same for both tree types).

So based on the results from our two models above, there's insufficient evidence to suggest that the influences of diameter and height aren't similar for both tree types.

1 c)

We construct a linear model to investigate how diameter, height and type influence volume.

```r
model = lm(volume~type+height+diameter, data=trees)
print("model coefficients:"); summary(model)$coefficients
```

```
## [1] "model coefficients:"

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -63.781     5.5129  -11.57 2.33e-16
## typeoak       -1.305     0.8779   -1.49 1.43e-01
## height         0.417     0.0752    5.55 8.42e-07
## diameter       4.698     0.1645   28.56 1.14e-34
```

```r
print("anova:"); res = anova(model); res
```

```
## [1] "anova:"

## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value  Pr(>F)
## type       1    380     380    36.1 1.6e-07 ***
## height     1   2239    2239   212.9 < 2e-16 ***
## diameter   1   8577    8577   815.6 < 2e-16 ***
## Residuals 55    578      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA p-values, type is not a significant predictor for volume (p-value $0.143 > 0.05$), while height and diameter are significant (p-values less than 0.05). Diameter and height are both positively correlated with the volume, with diameter having the largest contribution (coefficient) of the two.

```r
# build better model where type isn't considered
modelC = lm(volume~height+diameter, data=trees)

avgTree = data.frame(height=mean(trees$height), diameter=mean(trees$diameter))
```

```r
pred = predict(modelC, avgTree)
sprintf("predicted volume of average tree = %.3f", pred)
```

```
## [1] "predicted volume of average tree = 32.581"
```

```r
# mean(trees$volume) # this also gives the same result as expected

r2 = summary(modelC)$r.squared; ar2 = summary(modelC)$adj.r.squared
sprintf("modelC: R^2 = %.3f, Adj. R^2 = %.3f", r2, ar2)
```

```
## [1] "modelC: R^2 = 0.949, Adj. R^2 = 0.947"
```

Using the resulting model, the volume of a tree with the average height and diameter is predicted to be 32.581 .

1 d)

We propose to transform the data to create a new column that contains the volume of a (theoretical) cylinder based on the tree's diameter and height. (Note we omit tree type from the model as we found it to not be a significant predictor above).

```r
# create predictor as cylinderical volume
trees$cylinder = trees$diameter * pi * trees$height

modelD = lm(volume~cylinder, data=trees)
print("model coefficients:"); summary(modelD)$coefficients
```

```
## [1] "model coefficients:"

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.8923   2.058137   -13.1 8.67e-19
## cylinder      0.0179   0.000603    29.7 2.47e-36
```

```r
r2 = summary(modelD)$r.squared; ar2 = summary(modelD)$adj.r.squared
sprintf("model: R^2 = %.3f, Adj. R^2 = %.3f", r2, ar2)
```

```
## [1] "model: R^2 = 0.939, Adj. R^2 = 0.938"
```

```r
print("ANOVA:"); anova(model)
```

```
## [1] "ANOVA:"

## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value  Pr(>F)
## type       1    380     380    36.1 1.6e-07 ***
## height     1   2239    2239   212.9 < 2e-16 ***
## diameter   1   8577    8577   815.6 < 2e-16 ***
## Residuals 55    578      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After constructing a linear model for predicting the actual tree volume from our proposed cylindrical estimator, we see that the cylinder variable is a significant predictor of volume ($p < 0.05$). However the adjusted R^2 values (and the regular R^2 values) for this model are less than that of the model in part c), so while cylinder is a useful predictor, it's still inferior to using just the provided height and diameter variables in the model.

Exercise 2: Expenditure on criminal activities

1 a)

```r
crimes = read.table("expensescrime.txt", header=T)
#crimes
```

Exercise 3: Titanic

1 a)

```r
titanic = read.table("titanic.txt", header=T)
#titanic
```

Exercise 4: Military Coups

1 a)

```r
coups = read.table("coups.txt", header=T)
#coups
```