Lab 11

Adam Orr

November 17, 2017

Section: 91973 Friday 9am

Question 1

Script

```
# a. Make a scatter plot of the data. State the central assumptions of
# correlation analysis and examine the plot to determine whether these
# assumptions are met.
green <- read.csv('green.csv')
plot(green$attachment~green$birds,
     main = "Attachment to Green Spaces",
     xlab = "Number of bird species",
     ylab = "Attachment")

# b. Describe the pattern of the data in words. Is the relationship positive or
# negative? Is it linear? How strong is it?
#
# There seems to be a positive, linear relationship between the number of bird
# species and green space users' attachment. The relationship does not appear
# incredibly strong, but not weak either.
#
# c. Calculate an estimate of Pearson’s correlation coefficient. Do this
# calculation two ways: by hand, and with the R function cor.
#
# by hand
x <- green$birds
y <- green$attachment
xbar <- mean(x)
ybar <- mean(y)
xdiff<- x - xbar
ydiff<- y - ybar
(r <- sum(xdiff*ydiff)/sqrt(sum(xdiff^2)*sum(ydiff^2)))

# using cor
cor(x,y)

# d. Use the R function cor.test to perform a t-test of the null hypothesis that
# there is no correlation. Clearly state the conclusions of the test, including
# all key information (hypotheses, significance level, test statistic, degrees
# of freedom, and P-value). Also report the 95% confidence interval of rho.

#HO: The correlation between the number of bird species in a greenspace and
#greenspace user's attachment is 0.
```

```r
#HA: The correlation between the number of bird species in a greenspace and
#greenspace user's attachment is not 0.

cor.test(x,y)

# With a significance level of 0.05, we reject the null hypothesis that the
# correlation between the number of bird species in a green space and green
# space users' attachment is equal to 0 (t-test; t = 3.8595, df = 13, p =
# 0.002).

#The 95% confidence interval of rho is (0.349,0.904)
```

Output

```r
# a. Make a scatter plot of the data. State the central assumptions of
# correlation analysis and examine the plot to determine whether these
# assumptions are met.
```


```r
# b. Describe the pattern of the data in words. Is the relationship positive or
# negative? Is it linear? How strong is it?
#
# There seems to be a positive, linear relationship between the number of bird
# species and green space users' attachment. The relationship does not appear
# incredibly strong, but not weak either.
#
# c. Calculate an estimate of Pearson's correlation coefficient. Do this
# calculation two ways: by hand, and with the R function cor.
#
```

```r
# by hand
```

```
## [1] 0.7307406
```

```r
# using cor
```

```
## [1] 0.7307406
```

```r
# d. Use the R function cor.test to perform a t-test of the null hypothesis that
# there is no correlation. Clearly state the conclusions of the test, including
# all key information (hypotheses, significance level, test statistic, degrees
# of freedom, and P-value). Also report the 95% confidence interval of rho.

#H0: The correlation between the number of bird species in a greenspace and
#greenspace user's attachment is 0.

#HA: The correlation between the number of bird species in a greenspace and
#greenspace user's attachment is not 0.
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 3.8595, df = 13, p-value = 0.001972
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3491907 0.9044424
## sample estimates:
##       cor
## 0.7307406
```

```r
# With a significance level of 0.05, we reject the null hypothesis that the
# correlation between the number of bird species in a green space and green
# space users' attachment is equal to 0 (t-test; t = 3.8595, df = 13, p =
# 0.002).

#The 95% confidence interval of rho is (0.349,0.904)
```

Question 2

Script

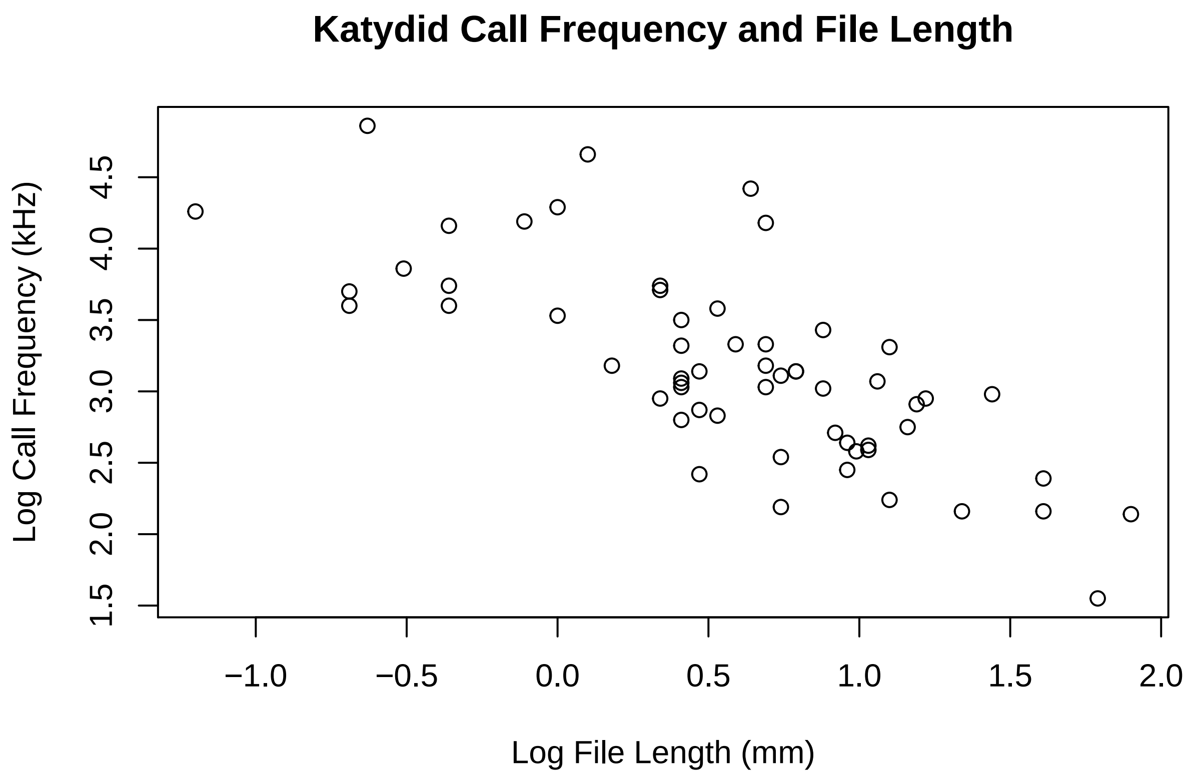
```r
# a. Make a scatter plot of log call frequency and log file length. Be sure that
# the independent variable is on the X-axis. Inspect the plot; does it look like
# there is a linear relationship?
katydid <- read.csv('katydid.csv')
x <- katydid$log.length
y <- katydid$log.freq
plot(x,y,
     main = "Katydid Call Frequency and File Length",
     xlab = "Log File Length (mm)",
     ylab = "Log Call Frequency (kHz)")

#There does appear to be a negative linear relationship between the log of file
```

```r
#length and the log of call frequency.
#
# b. Estimate the best-fit equation for the linear regression of log call
# frequency on log file length. Do this by hand.
xbar <- mean(x)
ybar <- mean(y)
xdiff <- x - xbar
ydiff <- y - ybar

b <- sum(xdiff*ydiff)/sum(xdiff^2)
a <- ybar - (b * xbar)

#The best fit linear equation is y = 3.627 - 0.795 * x

# c. Calculate the 95% confidence interval of the regression coefficient , by
# hand.
n <- length(x)
tcrit <- qt(c(.025,.975), n - 2)
yhat <- a + x * b
MSres <- sum((y-yhat)^2)/(n-2)
sb <- sqrt(MSres/sum(xdiff^2))
(CI <- b + tcrit * sb)

# d. Test the null hypothesis that  = 0, using a t-test. Be sure to clearly
# state your null and alternative hypotheses, as well as your conclusions.
t <- b/(sb)
#
#H0: The value of b is equal to 0.
#Ha: The value of b is not 0.
#
p <- 2*pt(t,n-2)

# With a significance of 0.05, I reject the null hypothesis that the value of b
# is equal to 0 (t-test, t = -8.39, df = 56, p = 1.8e-11).

# e. Calculate the coefficient of determination.
SSr <- sum((yhat-ybar)^2)
SSt <- sum((y-ybar)^2)
(r2 <- SSr/SSt)

# f. Re-do the regression, this time using the R functions lm and summary.
mod <- lm(y~x)
summary(mod)

# g. Overall, do the results support the use of file length to estimate call
# frequency? Explain why or why not.

#Overall, the results do support the use of file length to estimate call
#frequency, as the slope is significantly non-zero. Additionally, the
#coefficient of determination is somewhat large, meaning the regression explains
#a lot of the variance in frequency.

# h. Re-do the scatter plot from part a and add a regression line, using the
```

```
# command abline.
# plot(x,y,
#      main = "Katydid Call Frequency and File Length",
#      xlab = "Log File Length (mm)",
#      ylab = "Log Call Frequency (kHz)")
abline(mod)
```

Output

```
# a. Make a scatter plot of log call frequency and log file length. Be sure that
# the independent variable is on the X-axis. Inspect the plot; does it look like
# there is a linear relationship?
```




```
#There does appear to be a negative linear relationship between the log of file
#length and the log of call frequency.
#
# b. Estimate the best-fit equation for the linear regression of log call
# frequency on log file length. Do this by hand.
#The best fit linear equation is y = 3.627 - 0.795 * x

# c. Calculate the 95% confidence interval of the regression coefficient , by
# hand.
```

```
## [1] -0.9845774 -0.6050498
```

```
# d. Test the null hypothesis that  = 0, using a t-test. Be sure to clearly
# state your null and alternative hypotheses, as well as your conclusions.
#
#H0: The value of b is equal to 0.
```

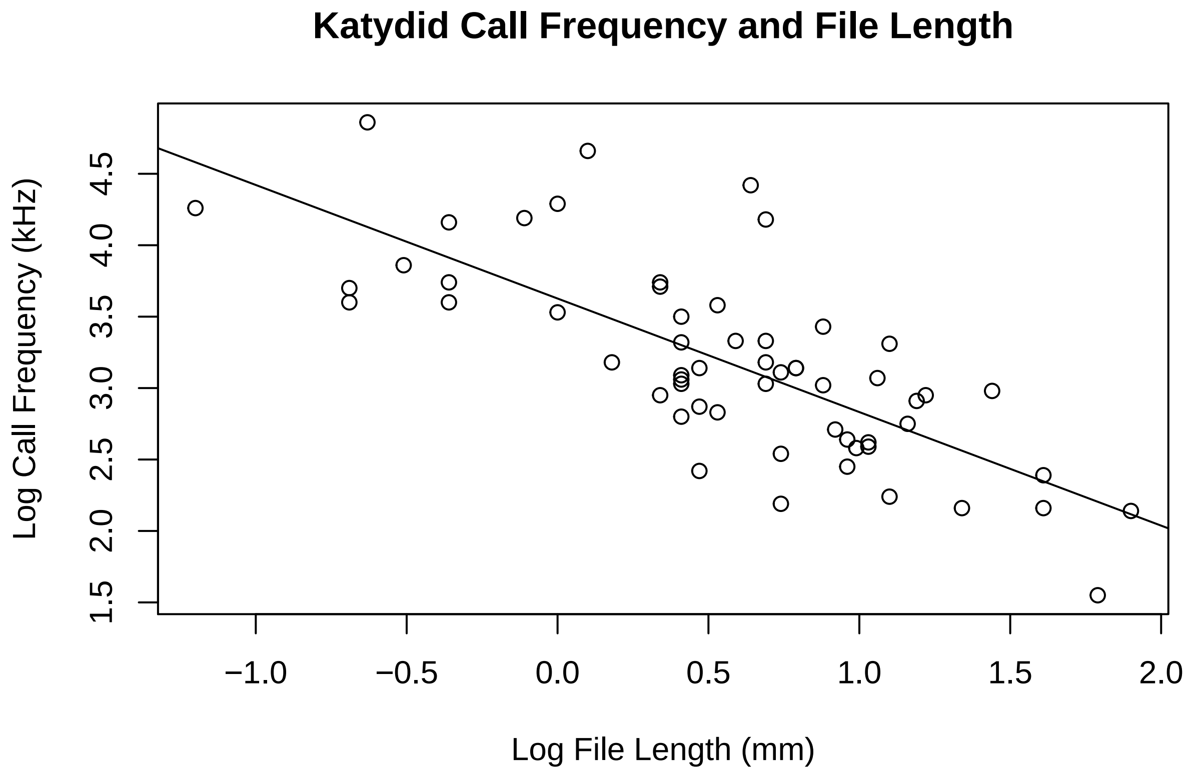
```
#Ha: The value of b is not 0.
#
# With a significance of 0.05, I reject the null hypothesis that the value of b
# is equal to 0 (t-test, t = -8.39, df = 56, p = 1.8e-11).

# e. Calculate the coefficient of determination.
```

```
## [1] 0.5569606
```

```
# f. Re-do the regression, this time using the R functions lm and summary.
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84876 -0.31075 -0.07271  0.25036  1.30176
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.62692    0.08121   44.66  < 2e-16 ***
## x           -0.79481    0.09473   -8.39 1.77e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4595 on 56 degrees of freedom
## Multiple R-squared:  0.557,  Adjusted R-squared:  0.549
## F-statistic:  70.4 on 1 and 56 DF,  p-value: 1.769e-11
```

```
# g. Overall, do the results support the use of file length to estimate call
# frequency? Explain why or why not.

#Overall, the results do support the use of file length to estimate call
#frequency, as the slope is significantly non-zero. Additionally, the
#coefficient of determination is somewhat large, meaning the regression explains
#a lot of the variance in frequency.

# h. Re-do the scatter plot from part a and add a regression line, using the
# command abline.
```



Question 3

Script

```
# a. Use the R functions residuals and fitted to calculate residuals and fitted
# values from the linear model you made in question 2.

#residuals
(res <- residuals(mod))

#fitted values
(fit <- fitted(mod))

# b. Plot the residuals vs. the fitted values.
plot(res~fit,
     xlab = "Fitted values",
     ylab = "Residuals",
     main = "Residuals vs. Fitted Values")


# c. Make a normal probability plot of the residuals.
qqnorm(res)

# d. Evaluate the plots you have made for adherence to normality, linearity, and
# homogeneity of variances.

#The residuals vs. fitted values plot shows that the residuals are somewhat
```

```
#evenly dispersed around 0, indicating that the regression meets the assumption
#of linearity and homogeneity. The normal probability plot shows a straight
#line, indicating adherence to normality. Overall, there is not much evidence
#for departure from normality, linearity, or homogeneity of variances.
```

Output

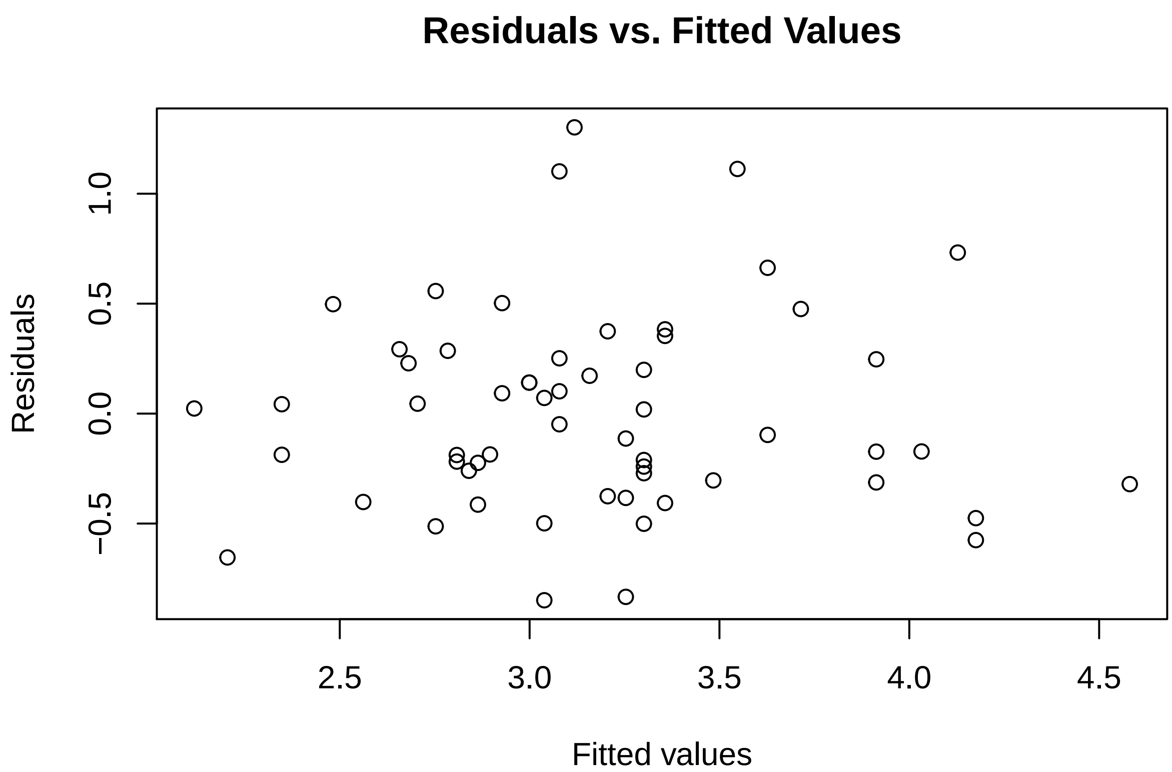
```
# a. Use the R functions residuals and fitted to calculate residuals and fitted
# values from the linear model you made in question 2.

#residuals
```

```
##           1           2           3           4           5           6
## -0.65420432  0.02322517 -0.40187043 -0.18727077 -0.84875857 -0.51262569
##           7           8           9          10          11          12
##  0.04272923 -0.83335824 -0.41389959 -0.49875857 -0.26005518 -0.21826264
##          13          14          15          16          17          18
## -0.18826264 -0.22389959 -0.18569213  0.04506313 -0.50104705 -0.37566942
##          19          20          21          22          23          24
## -0.38335824  0.22890753 -0.40668400  0.29275194  0.49761093  0.09251533
##          25          26          27          28          29          30
## -0.04849925 -0.27104705 -0.24104705  0.28558177 -0.21104705  0.07124143
##          31          32          33          34          35          36
##  0.14098211 -0.11335824  0.14098211  0.10150075 -0.30385417  0.55737431
##          37          38          39          40          41          42
##  0.01895295  0.17201939  0.25150075  0.50251533  0.19895295 -0.09692061
##          43          44          45          46          47          48
##  0.37433058 -0.31305350 -0.57534198 -0.47534198  0.35331600  0.38331600
##          49          50          51          52          53          54
## -0.17305350 -0.17227553  0.24694650  1.10150075  0.47564989 -0.32069690
##          55          56          57          58
##  0.66307939  1.30176007  1.11256074  0.73234684
```

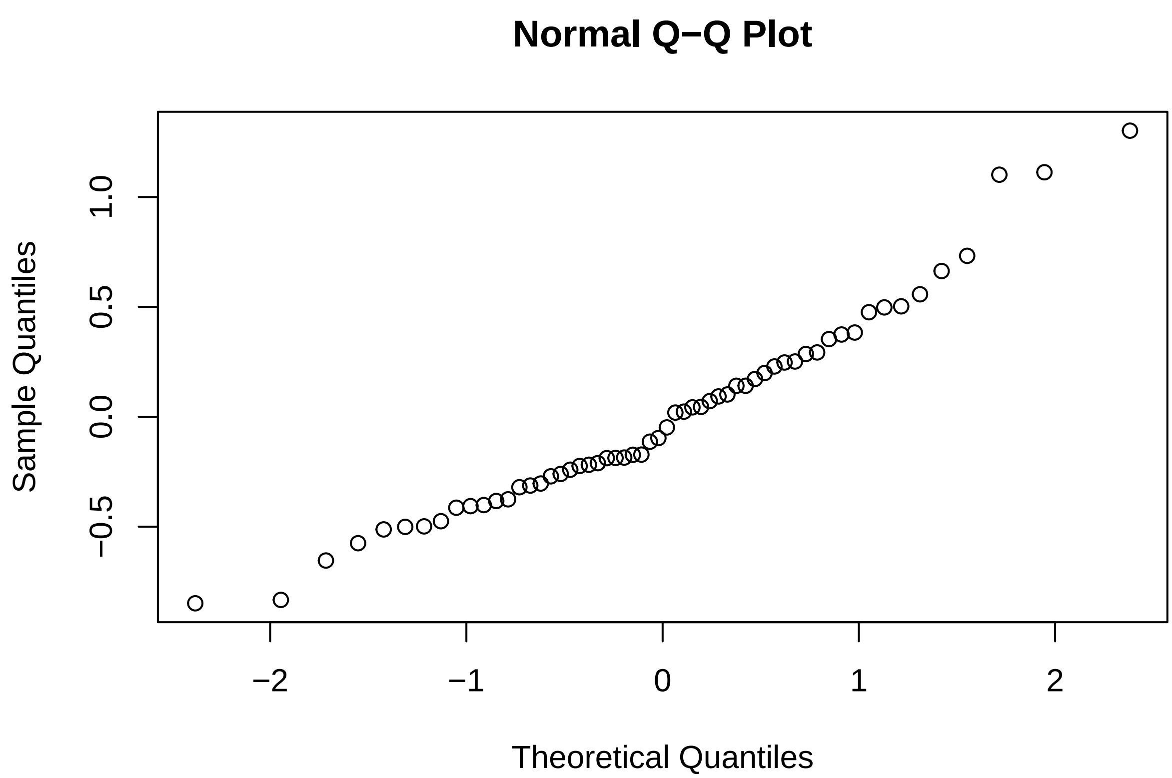
```
#fitted values
```

```
##        1        2        3        4        5        6        7        8
## 2.204204 2.116775 2.561870 2.347271 3.038759 2.752626 2.347271 3.253358
##        9       10       11       12       13       14       15       16
## 2.863900 3.038759 2.840055 2.808263 2.808263 2.863900 2.895692 2.704937
##       17       18       19       20       21       22       23       24
## 3.301047 3.205669 3.253358 2.681092 3.356684 2.657248 2.482389 2.927485
##       25       26       27       28       29       30       31       32
## 3.078499 3.301047 3.301047 2.784418 3.301047 3.038759 2.999018 3.253358
##       33       34       35       36       37       38       39       40
## 2.999018 3.078499 3.483854 2.752626 3.301047 3.157981 3.078499 2.927485
##       41       42       43       44       45       46       47       48
## 3.301047 3.626921 3.205669 3.913053 4.175342 4.175342 3.356684 3.356684
##       49       50       51       52       53       54       55       56
## 3.913053 4.032276 3.913053 3.078499 3.714350 4.580697 3.626921 3.118240
##       57       58
## 3.547439 4.127653
```

```
# b. Plot the residuals vs. the fitted values.
```




```
# c. Make a normal probability plot of the residuals.
```




```
# d. Evaluate the plots you have made for adherence to normality, linearity, and
# homogeneity of variances.
```

```
#The residuals vs. fitted values plot shows that the residuals are somewhat
#evenly dispersed around 0, indicating that the regression meets the assumption
#of linearity and homogeneity. The normal probability plot shows a straight
#line, indicating adherence to normality. Overall, there is not much evidence
#for departure from normality, linearity, or homogeneity of variances.
```