EEMB 146 Lab 7

Samantha Chen

5/14/2021

Question 1 Correlating Dandelions

Does the number of leaves in a dandelion rosette correlate with thediameter of the rosette?

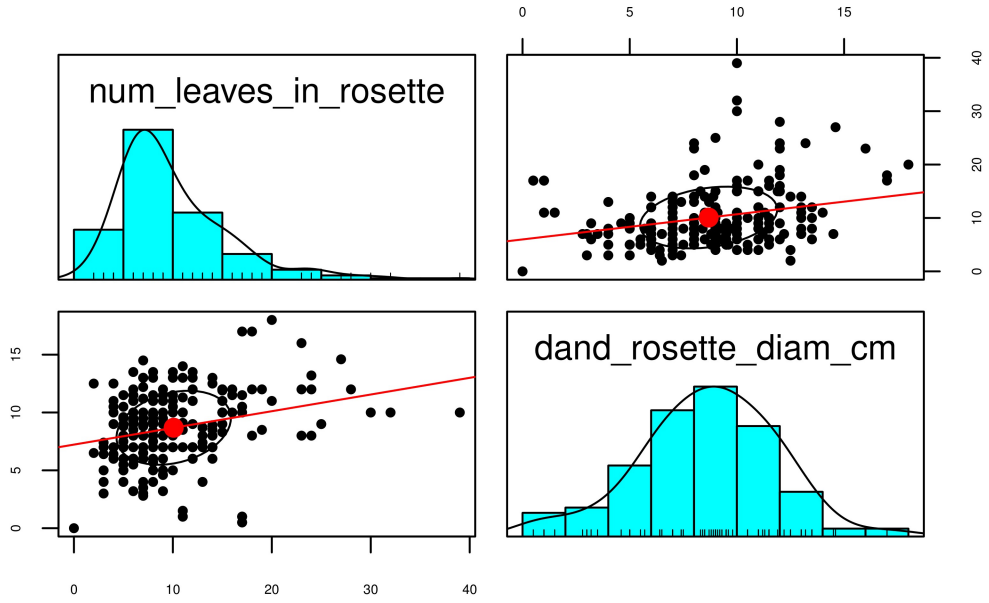
Clearly state your null and alternative hypotheses for the correlation test.

Null Hypothesis: The number of leaves in a dandelion rosette and diameter of the rosette are not related (H_0: rho = 0).

Alternative Hypothesis: The number of leaves in a dandelion rosette and diameter of the rosette are correlated (H_A: rho =/= 0).

Use a scatter plot matrix of num_leaves_in_rosette and dand_rosette_diam_cm to assess your assumptions for a parametric correlation test. Do you think your assumption of linearity and bivariate normality are met just based on the figure? *You don't need to run any Shapiro-Wilk tests to answer this.

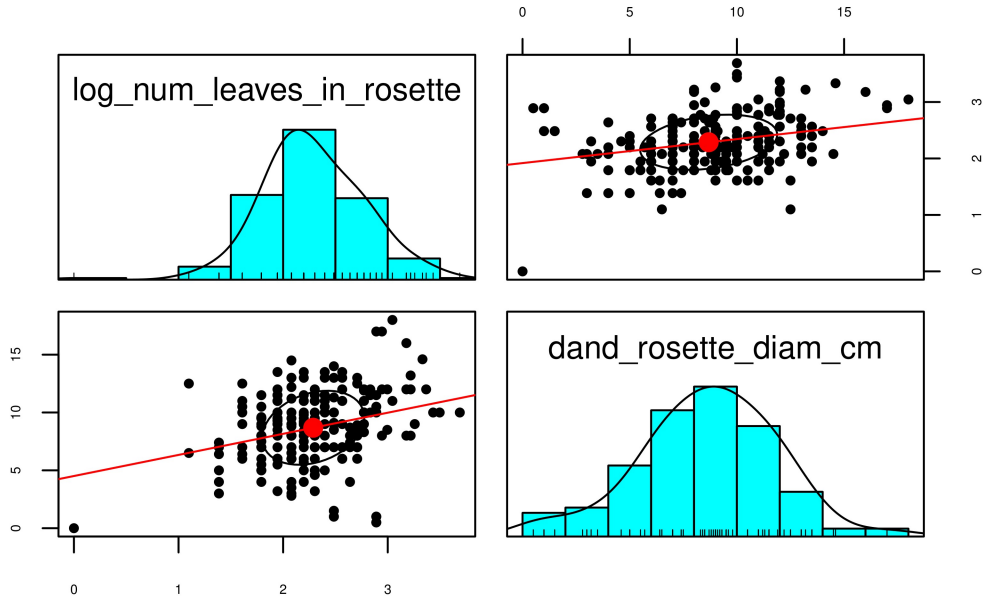
```
plant <- read.csv("~/EEMB 146 Lab Files/Lab 7 Data/plant_data.csv") #loading data
# str(plant) 259 rows, 10 variables
# subseting data
plant_plot <- subset(plant, select = c(num_leaves_in_rosette, dand_rosette_diam_cm))
# plotting scatterplot matrix
pairs.panels(plant_plot, density = TRUE, cor = FALSE,
             lm = TRUE, cex.axis = 0.5)
```



Dand_rosette_diam_cm's histogram is symmetrical and looks normlly distributed but num_leaves_in_rosette's histogram looks asymmetrical and skewed to the right. These characteristics are also present in the scatterplot, where the bottom left plot has a clear pattern of clustering towards the left side. Therefore my data does not meet my assumption of linearity and bivariate normality and I will need to log transform the num_leaves_in_rosette.

If your assumptions of bivariate normality are not met (i.e. at least one of the variables is not normal), transform whatever variable is not normal so that the assumptions of linearity and bivariate normality are met. Assess these assumptions using a scatter plot matrix and briefly describe how the plot shows you that the assumptions are now met.

```
# log transformation
plant_plot$log_num_leaves_in_rosette <- log(plant_plot$num_leaves_in_rosette +1)
plant_plot2 <-  subset(plant_plot, select = c(log_num_leaves_in_rosette, dand_rosette_diam_cm))
# scatterplot with transformed data
pairs.panels(plant_plot2, density = TRUE, cor = FALSE,
             lm = TRUE, cex.axis = 0.5)
```



The num_leaves_in_rosette was log transformed and we can see a clear symmetrical bell-curve shape in the histogram. The scatterplot does not show any obvious pattern of a non-linear relationship so it's safe to say that the log transformed data meets the assumption of linearity and bivariate normality.

Run a Pearson's correlation test on the transformed data.

```
# running Pearson's test
cor.test(plant_plot2$log_num_leaves_in_rosette, plant_plot2$dand_rosette_diam_cm,
         method = "pearson", alternative = "two.sided")
```

```
##
##  Pearson's product-moment correlation
##
## data:  plant_plot2$log_num_leaves_in_rosette and plant_plot2$dand_rosette_diam_cm
## t = 4.2524, df = 215, p-value = 3.154e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1509495 0.3969917
## sample estimates:
##       cor
## 0.2785343
```

```
# t = 4.2524,
# p = 3.154e-05, r = 0.2785343
```

Pearson's correlation test on the transformed data produced a correlation value of r = 0.2785343 with t = 4.2524 and p-value of 3.154e-05. Because the p-value is much smaller than alpha = 0.05, I can reject

the null hypothesis and say that the correlation between the log transformed num_leaves_in_rosette and dand_rosette_diam_cm is significantly different from 0.

Run a Spearman's rank correlation test on the untransformed data.

```
# running a Spearman's rank
cor.test(plant_plot$num_leaves_in_rosette, plant_plot$dand_rosette_diam_cm,
         method = "spearman", alternative = "two.sided")
```

```
##
##  Spearman's rank correlation rho
##
## data:  plant_plot$num_leaves_in_rosette and plant_plot$dand_rosette_diam_cm
## S = 1254370, p-value = 8.568e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2634421
```

```
# S = 1254370,
# p-value = 8.568e-05, rho = 0.2634421
```

A Spearman's rank correlation test was performed on the untransformed data and produced a correlation coefficient of rho = 0.2634421 with a p-value of 8.568e-05. Because the p-value is much smaller than alpha = 0.05, I can reject the null hypothesis and say that there is a significant correlation between the ranks of num_leaves_in_rosette and ranks of dand_rosette_diam_cm.

Based on both tests, what do you conclude about the correlation (positive, negative, none) between number of leaves in a dandelion rosette and the diameter of a dandelion rosette?

Both my Pearson's correlation test and my Spearman's rank correlation test indicates that there is a positive, but relatively weak correlation between the number of leaves in a dandelion rosette and the diameter of a dandelion rosette (r = 4.2524; rho = 0.2634421). Because both test had a p-value of less than alpha = 0.05, I can reject the null hypothesis that says number of leaves in a dandelion rosette and diameter of the rosette are not related (H_0: rho = 0). There is a significant relationship between the number of leaves in a dandelion rosette and the diameter of the rosette.

Question 2 Social Spiders

Can you predict the number of spiders in a colony based on how high the web is off of the ground?

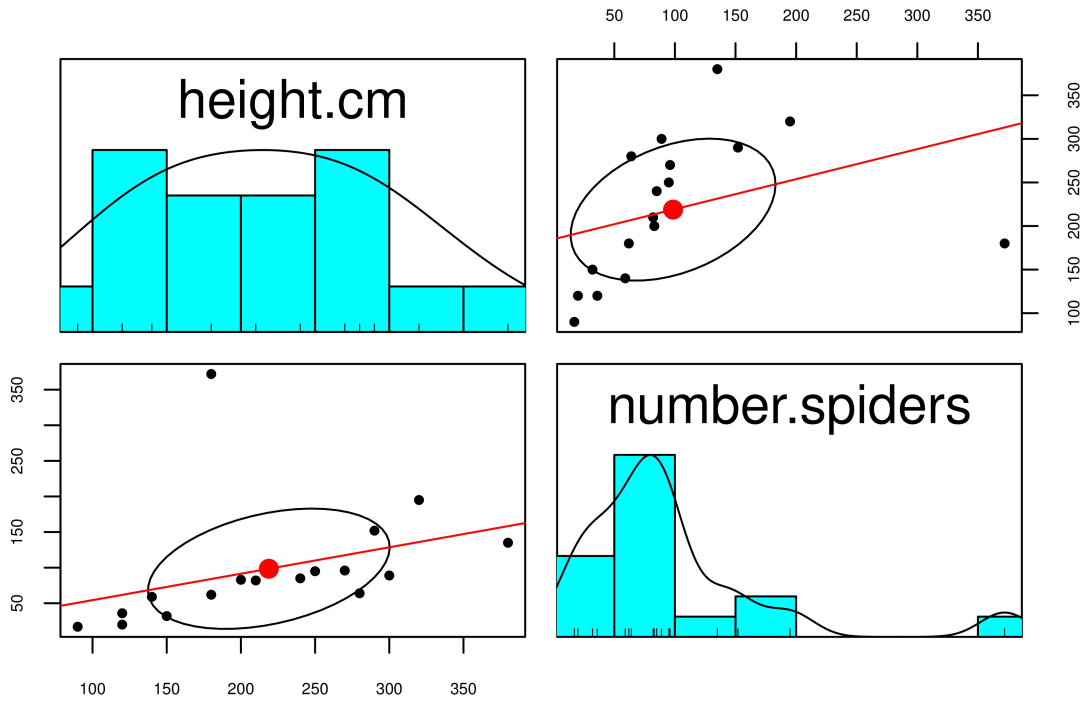
Clearly state your null and alternative hypotheses for a regression analysis.

Null Hypothesis: There is no effect of the spider web height on the number of spiders in a colony. (H_0: beta = 0)

Alternative Hypothesis: There is an effect of the spider web height on the number of spiders in a colony. (H_A: beta =/= 0)

Make a scatterplot of the data. What stands out to you about this scatterplot?

```
# loading data
spider <- read.csv("~/EEMB 146 Lab Files/Lab 7 Data/spiders.csv")
# dropping colony variable
spider_cleaned <- spider[c("height.cm", "number.spiders")]
# spider scatterplot
pairs.panels(spider_cleaned, density = TRUE, cor = FALSE, lm = TRUE, cex.axis = 0.6)
```



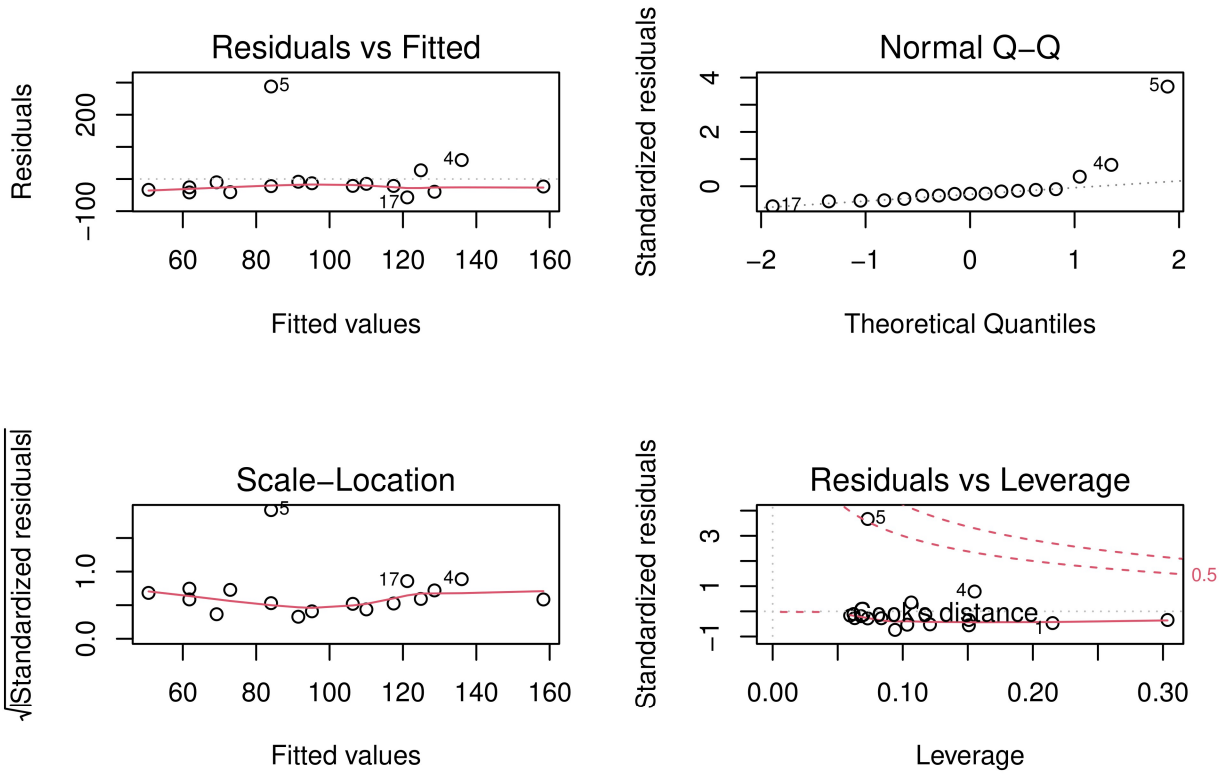
Looking at the histogram, height (cm) is symmetrical and normally distributed while number of spiders is asymmetrical and right skewed. The number of spiders also has an outlier present according to histogram. What stands out is that the scatterplots do indicate some linear relationship because the regression lines are at a slope and not completely a straight line. It also looks like the scatterplot on the right has a steeper slope.

Fit a linear regression to the data, and look at the diagnostic plots like we did earlier. Based on these plots, are the assumptions of normality and equal variance met? Briefly explain your answer.

```
# linear regression fit
spider_cleaned_lm <- lm(number.spiders~height.cm, data = spider_cleaned)
summary(spider_cleaned_lm) # F-statistic: 2.201 on 1 and 15 DF
```

```
##
## Call:
## lm(formula = number.spiders ~ height.cm, data = spider_cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57.18 -33.66 -21.47 -10.21 287.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.2501    58.2111   0.296    0.771
## height.cm     0.3712     0.2502   1.483    0.159
##
## Residual standard error: 81.53 on 15 degrees of freedom
## Multiple R-squared:  0.1279, Adjusted R-squared:  0.06981
## F-statistic: 2.201 on 1 and 15 DF,  p-value: 0.1587
```

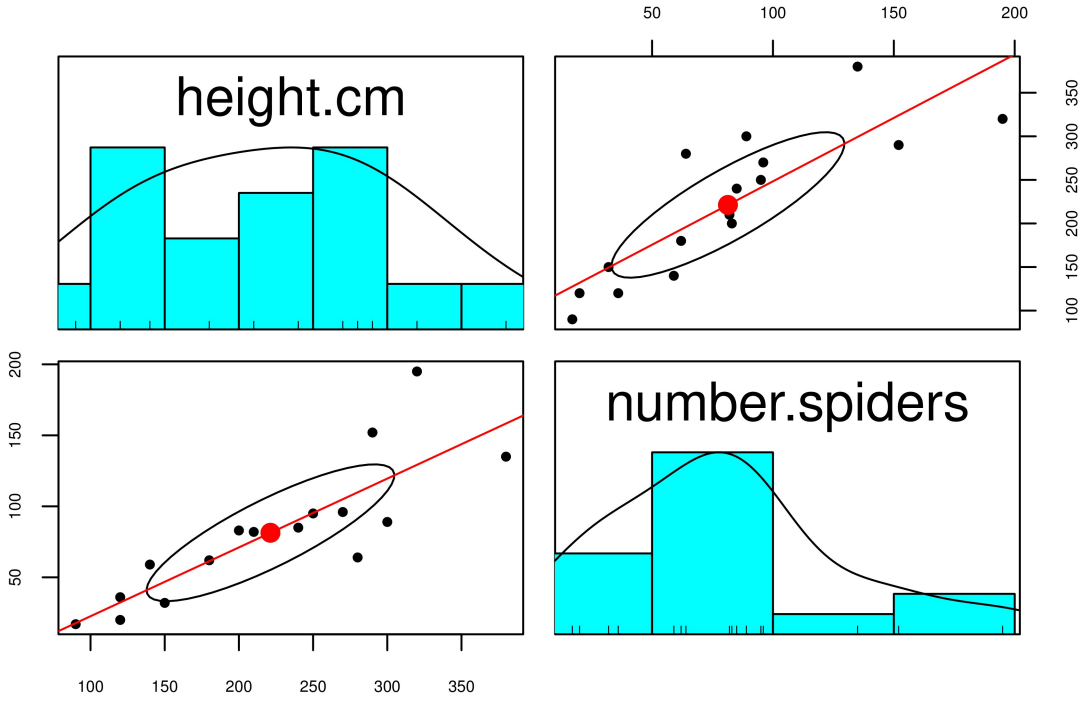
```
# p-value: 0.1587
par(mfrow = c(2,2))
plot(spider_cleaned_lm)
```



For the normal QQ plot most of the data points fall on the straight line with some deviating from the line and one outlier at 5. This means we can assume our residuals are normal. However, looking at the residual vs. fitted graph, there is a pattern in the residuals about the 0 line, all of them line up to the 0 line so this might mean our equal variance assumption is not met.

You learn that one of the research technicians miscounted observation 5 and you decide to drop it from your data and run a regression analysis. You can use the subset() function where colony!=5, and run your regression on the new subset. After you have fit the regression model, check your assumptions with diagnostic plots and report whether you think your assumptions for the linear regression are met.

```
# dropping colony 5 and making a scatterplot
spider_cleaned2 <- spider_cleaned[-c(5),] # drops row
pairs.panels(spider_cleaned2, density = TRUE, cor = FALSE, lm = TRUE, cex.axis = 0.6)
```

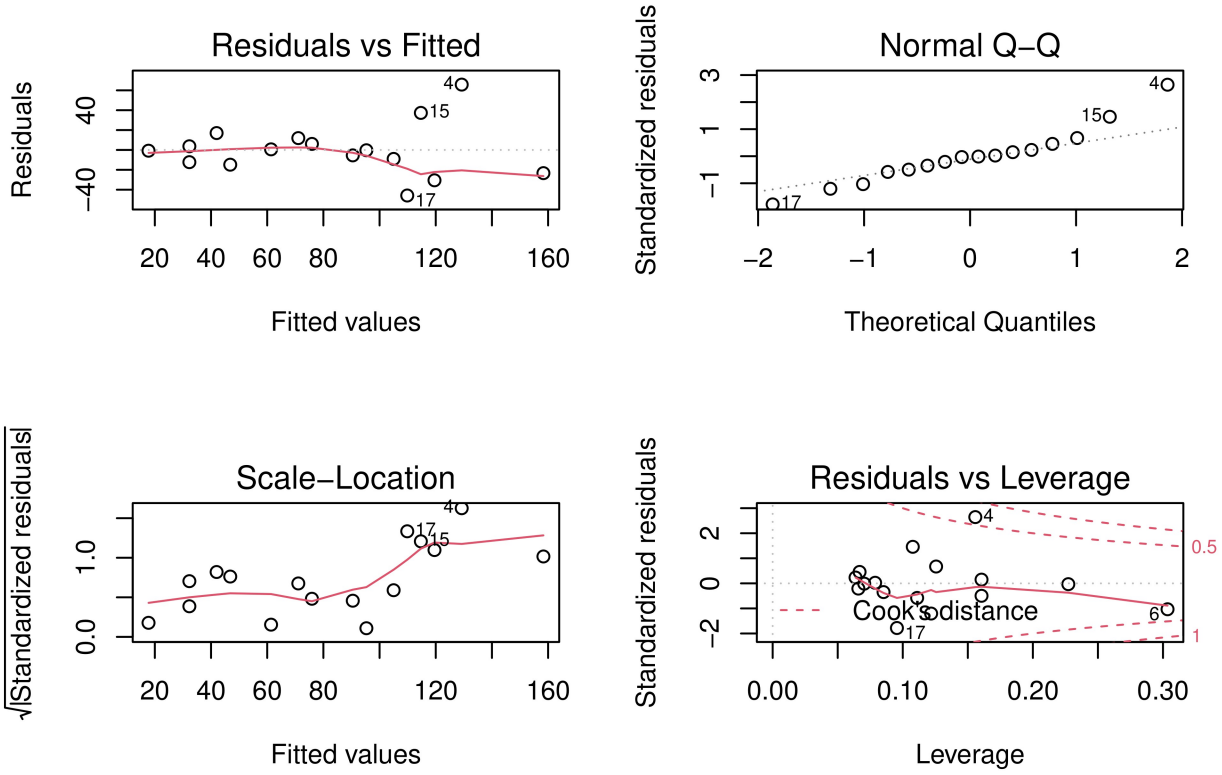



```
# linear regression fit
spider_cleaned2_lm <- lm(number.spiders~height.cm, data = spider_cleaned2)
summary(spider_cleaned2_lm)
```

```
##
## Call:
## lm(formula = number.spiders ~ height.cm, data = spider_cleaned2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.854 -12.930  -0.532   7.540  65.756
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.87545   19.72003  -1.312    0.211
## height.cm     0.48475    0.08372   5.790 4.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.07 on 14 degrees of freedom
## Multiple R-squared:  0.7054, Adjusted R-squared:  0.6844
## F-statistic: 33.53 on 1 and 14 DF,  p-value: 4.681e-05
```

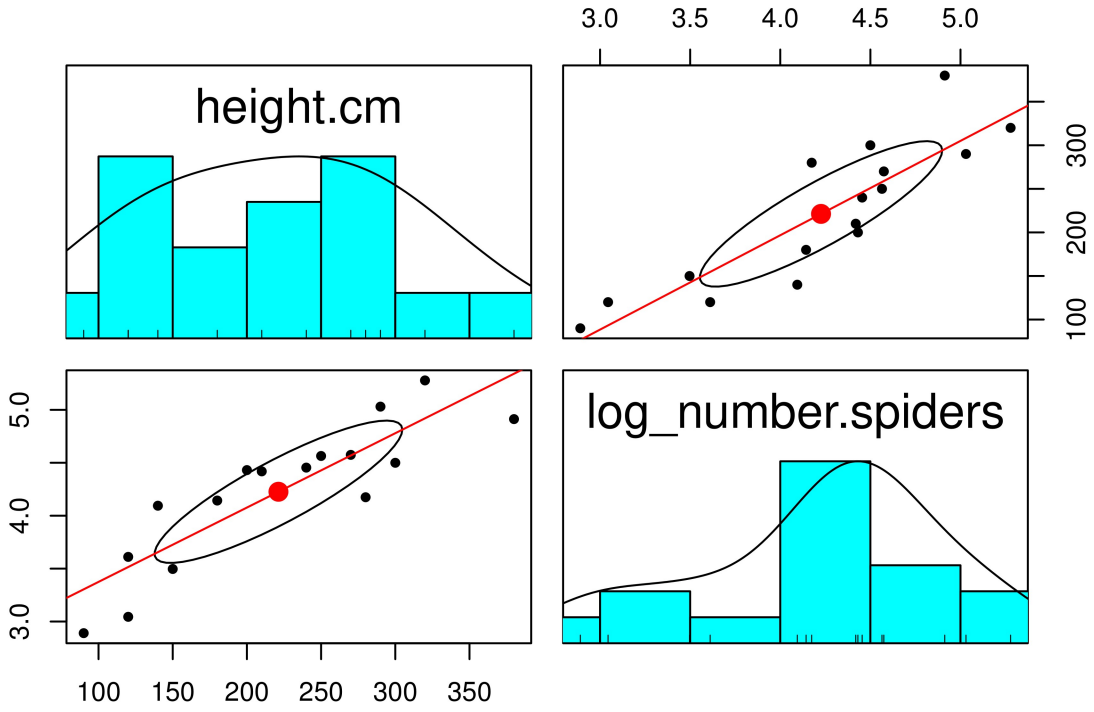
```
par(mfrow = c(2,2))
plot(spider_cleaned2_lm)
```



In the histogram, height looks symmetrical and normal but the number of spiders look right skewed. The scatterplots indicate a linear relationship between height of web and number of spiders because the line is sloped and not flat. For my QQ plot, Most of the data fall on line which means I can assume residuals are normal. In our residuals vs. fitted graph, there seems to be a similar pattern about the 0 line as the original data (with the outlier) so I cannot assume the equal variances. I will need to transform my data before performing a linear regression.

If necessary, try transforming your response and/or predictor variables. Report what transformations you tried and show the resulting diagnostic plots. Make sure you continue to exclude colony 5!

```
# log transforming spider
spider_cleaned2$log_number.spiders <- log(spider_cleaned2$number.spiders +1)
#subset
spider_cleaned3 <-spider_cleaned2[c("height.cm", "log_number.spiders")]
# scatterplot and linear regression check
pairs.panels(spider_cleaned3, density = TRUE, cor = FALSE, lm = TRUE)
```

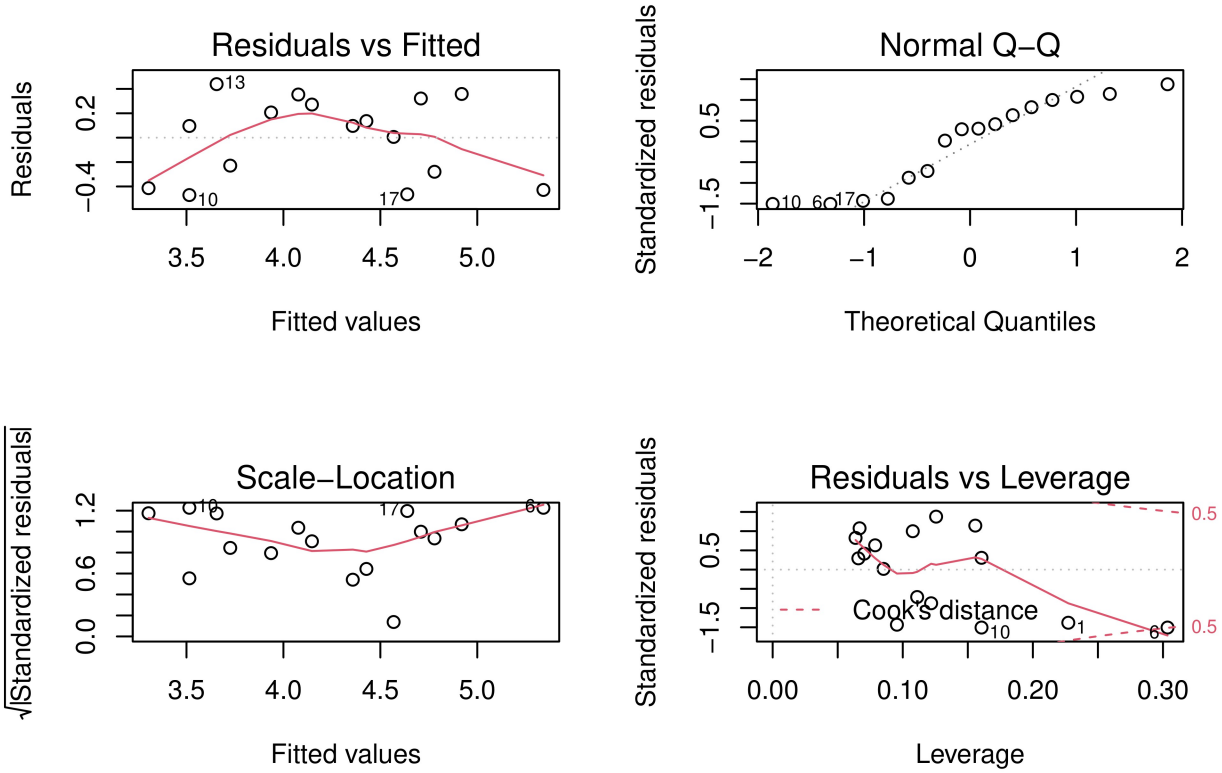



```
spider_cleaned3.lm <- lm(log_number.spiders~height.cm, data = spider_cleaned3)
summary(spider_cleaned3.lm) # F-statistic: 44.42 on 1 and 14 DF
```

```
##
## Call:
## lm(formula = log_number.spiders ~ height.cm, data = spider_cleaned3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47045 -0.31307  0.09622  0.28414  0.43889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 2.672114   0.248235  10.764 3.72e-08 ***
## height.cm   0.007024   0.001054   6.665 1.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3408 on 14 degrees of freedom
## Multiple R-squared:  0.7604, Adjusted R-squared:  0.7433
## F-statistic: 44.42 on 1 and 14 DF,  p-value: 1.069e-05
```

```
# p = 1.069e-05
par(mfrow = c(2,2))
plot(spider_cleaned3.lm)
```



I log transformed the number of species because the untransformed data is right skewed. The data looks more symmetrical and normal compared to that of the untransformed data. In addition, the Residual vs. Fitted plot looks like there is no distinct pattern about the 0 line so I can assume the variances are equal. Some of the ppoints on straight line of the normal QQplot deviates from the straight line but it is still good enough to assume residual data is normal. I have met all assumptions and can carry forward with my linear regression model.

Report the resulting linear regression model in the form: response variable = b0 + b1 * explanatory variable. Interpret b1 in a sentence.

My response variable is number of spiders, b_0 is the intercept which is 2.672114 and b_1 is the slope of height.cm (explanatory variable) which is 0.007024. Therefore the resulting linear regression model is

log_number.spiders = 2.672114 + 0.007024*height.cm. The value b_1 tells us that for every one cm increase in spiderweb height, there is a corresponding 0.007024 increase in the number of spiders.

Use the model you wrote down in the question above to predict the expected number of spiders in a colony 230cm off of the ground. Hint: if you log transformed your data, remember that log in R is by default the natural log.

```
# using model to calculate for a colony 230 cm off ground
log_number.spiders = 2.672114 + 0.007024*230
log_number.spiders
```

```
## [1] 4.287634
```

```
exp(log_number.spiders)
```

```
## [1] 72.79403
```

The model predicts that a spider colony 230 cm off the ground has approximately 73 spiders (y = 72.79403 spiders).

Is web height a significant predictor of the number of spiders in a colony? What is the value of the p-value that is allowing you to make that conclusion?

Hypotheses for Transformed Data: Null Hypothesis: There is no effect of the spider web height on the log transformed number of spiders in a colony. (H_0: beta = 0)

Alternative Hypothesis: There is an effect of the spider web height on the log transformed number of spiders in a colony. (H_A: beta =/= 0)

The linear regression model indicates that the p-value for the slope b_1 p = 1.07e-05 which is less than alpha = 0.05. Because it is smaller than alpha = 0.05, I can reject the null hypothesis that the slope is equal to 0 and can confidently say that the height of spider web (cm) is a good predictor of the number of spiders in a colony.

Finally, report the R2 value of the model and interpret it in a sentence.

The linear regression model produced a multiple R-squared value of R^2 = 0.7604. This means that 76.04% of the variation in the log transformed number of spiders (Y) can be explained by the spider web height of of the ground (X).