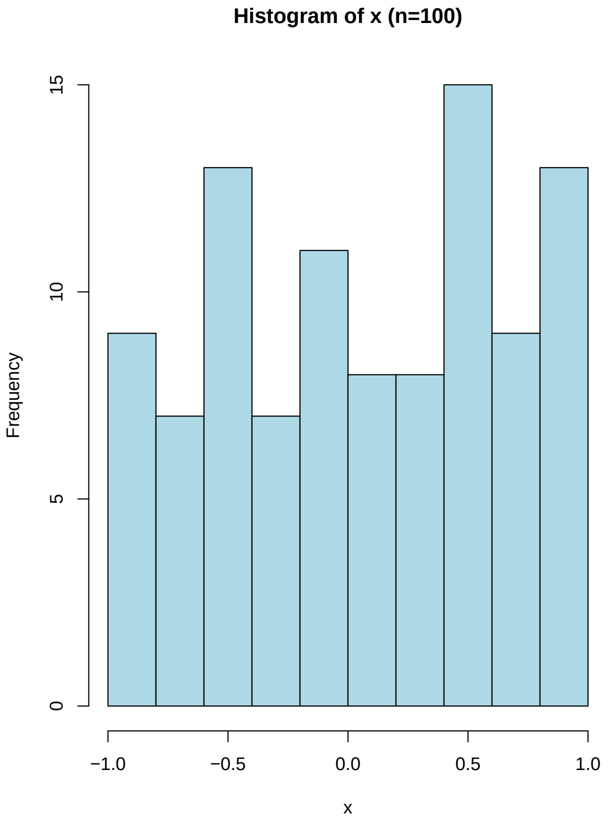
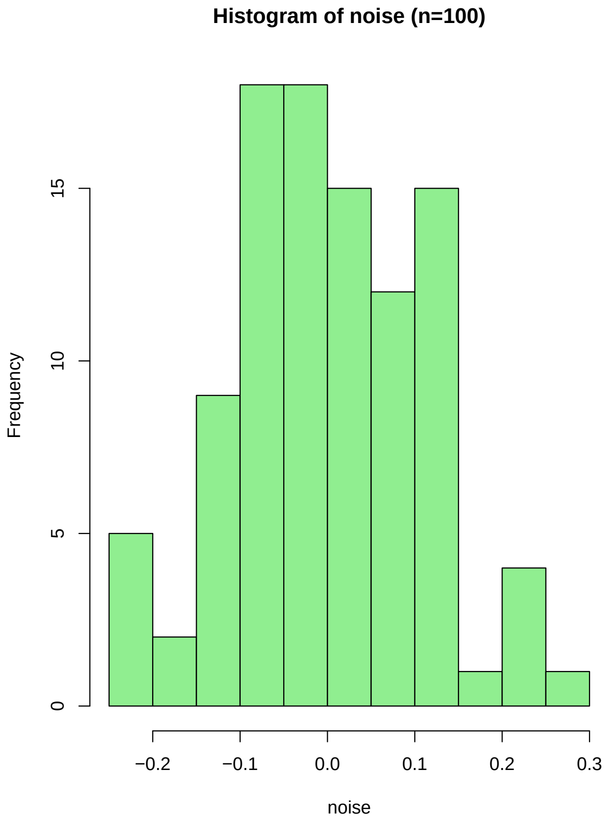
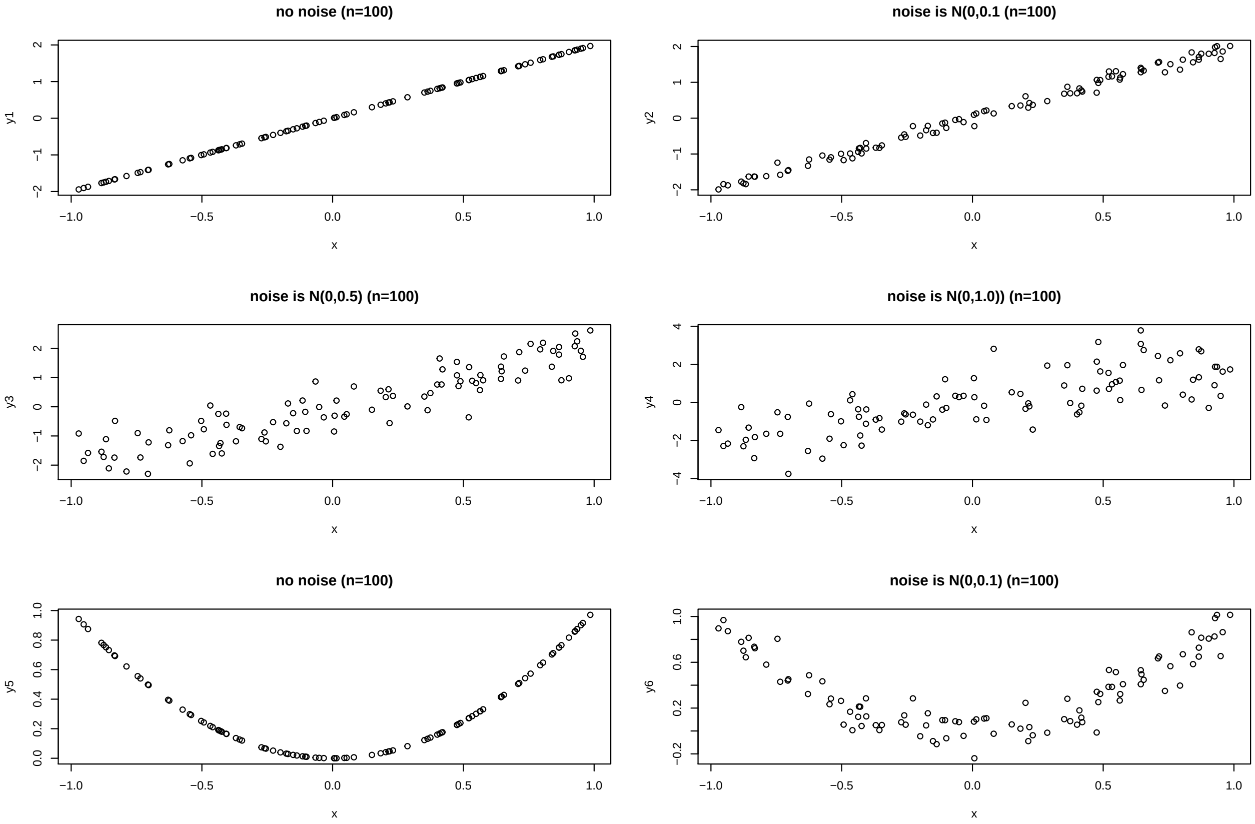
Question 1: We'll start with creating scatter plots of simulated data in R. (Section 3.1, Scatter plots)

1a) Insert the plots of the histograms of x and the noise.





1b) Insert the scatter plots. Describe the effect does the noise has on the scatter plots.



The noise scatters the data points up and down in a vertical motion. However, despite this, the trend is still clearly visible to the eye.

Question 2: We'll assess the correlation between the variables in the simulated data. (Section 3.2, Correlation)

2a) What do you expect cor(x, y1) to equal? What happens to corr(x, yi) as i increases to 4?

I expect cor(x, y1) to be very close, if not equal to 1. This is because the x and y1 is nearly completely modeled by a line (the linear regression model).

As I increases from 1 to 4, the correlation decreases. This is because with the increased variation, the linear regression does not fully account for the data, and there is some errors in its approximations. The data no longer follows a straight line as the error increases.

2b) What do you expect cor(x, y5) to equal? Explain.

I expect cor(x, y5) to be a very small number. This is because the data in the graph is a positive quadratic function, which is not well defined by using a linear regression. Therefore, the correlation coefficient should be very small.

2c) Report the five correlations, and explain their meaning, and determine whether they're appropriate.

Table 1: Table of the five correlation values

code	correlation
cor(x,y1)	1.0000000
cor(x,y2)	0.9956725
cor(x,y3)	0.9136715
cor(x,y4)	0.7691365
cor(x,y5)	0.1333678

Meaning: The meaning of these correlation coefficients, is that it represents the amount of variation that the model accounts for using a linear model. As the variation increases, the correlation decreases, as the linear model fits the data less and less.

Appropriateness:

For y1 through y4, these correlation coefficients are appropriate, because they are used on a linear model.

While it is technically appropriate to find the correlation of y5, because a linear model is used, it is not really that appropriate. This is because the data is clearly quadratic in nature, so there is no reason to use a linear model to model this data.

3a) Calculate the sample correlation coefficient for single data set consisting of the nine (x,y)

```r
# find the correlation between x and y
cor(xData, yData)
```

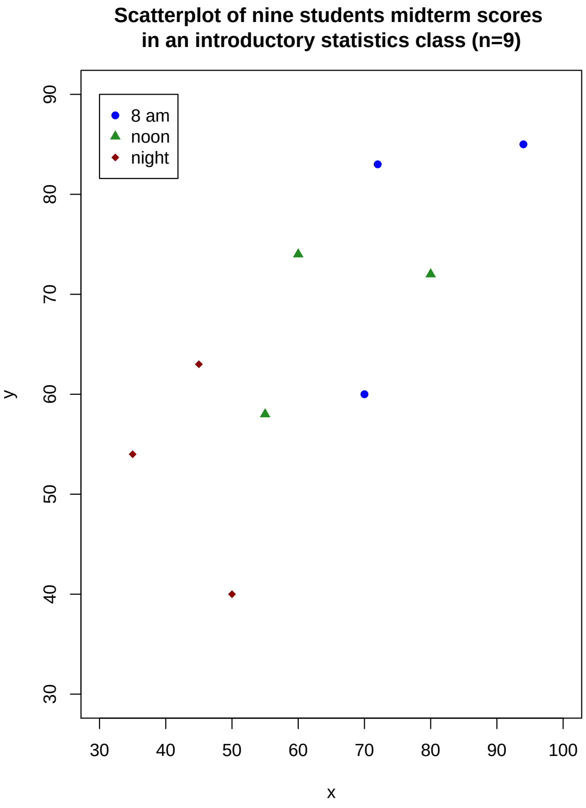
```
## [1] 0.7329053
```

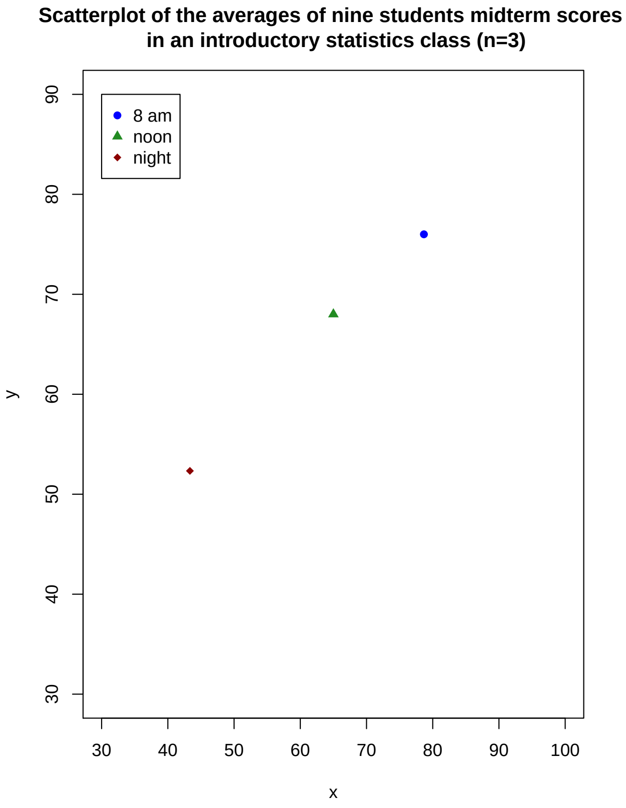
3b) Calculate r for the dataset of means.

```r
# calculate correlation of x_bar and y_bar
cor(x_bar,y_bar)
```

```
## [1] 0.9984929
```

3c) Insert the plots here. Use them to describe why r in part (a) is smaller than r in part (b). Does this suggest that a correlation coefficient based on averages (called an "ecological correlation" in statistics) might be misleading? Explain.





The r in part (a) contains 9 data points with a fairly large variation. Part (b) contains only 3 data points, with a smaller variation due to averaging. Because there are less points in part (b), it more easily fits a linear model, and thus its r value is a lot higher.

This does indeed suggest that a correlation coefficient base on averages might be misleading, as the original data itself might not have a high correlation, but the averages will. This can mislead.

Question 4: Now we'll make scatterplots and calculate correlation for a real data set. (Sections 3.1 and 3.2)

4a)

i. Find the sample size.

```r
# find number of rows in dataset
nrow(datSea)
```

```
## [1] 3650
```

ii. Find five-number summaries.

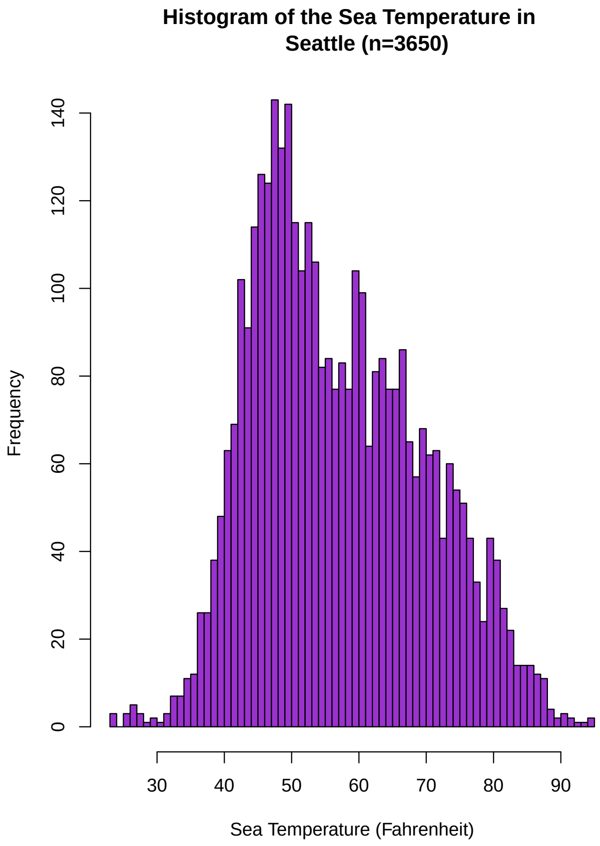
```
summary(datSea$V1)
```

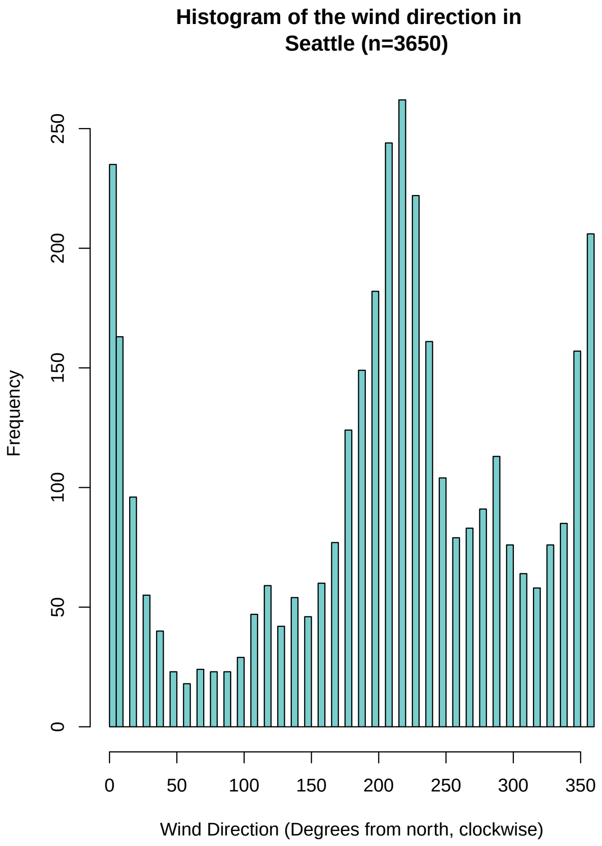
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   23.00   48.00   56.00   57.47   67.00   95.00 
```

```
summary(datSea$V2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   140.0   220.0   199.9   280.0   360.0 
```

iii. Create appropriately labeled histograms of both variables.





iv. Based on your summary, what are the units of the temperature?

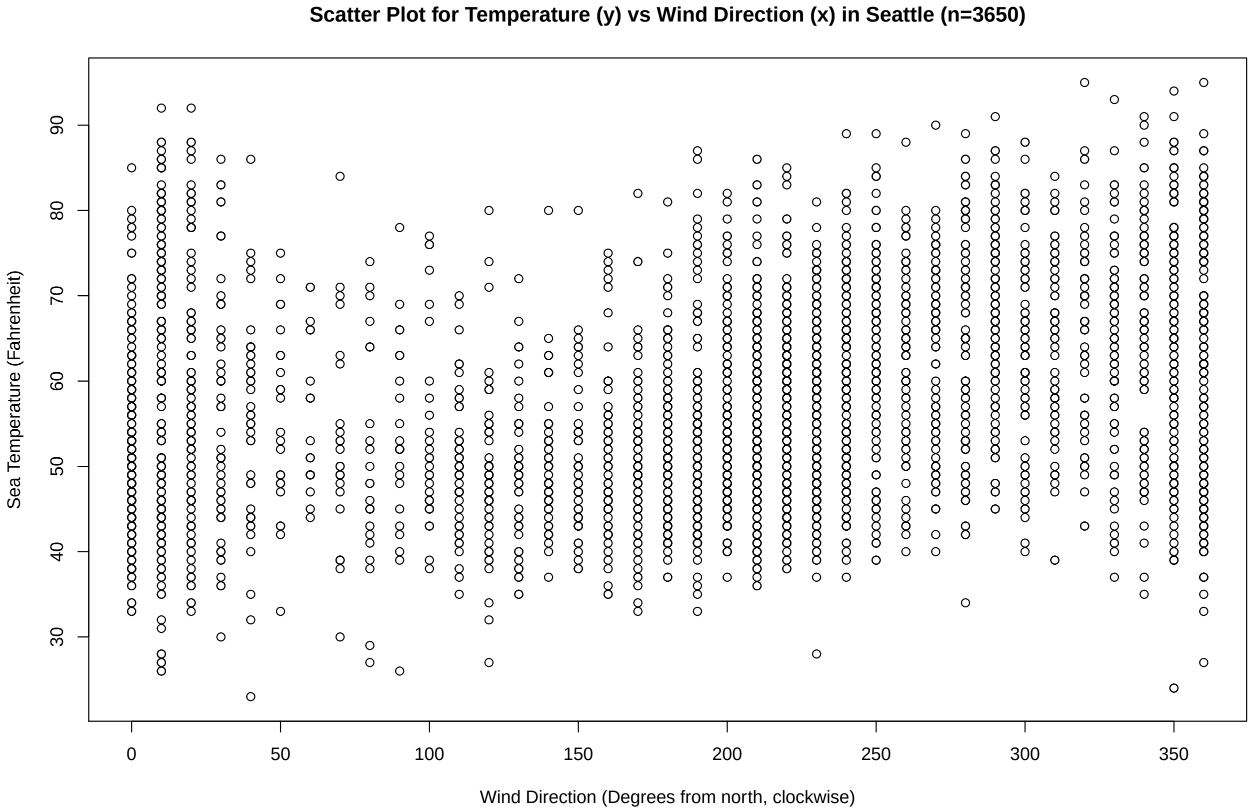
```r
summary(datSeaTemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   23.00   48.00   56.00   57.47   67.00   95.00
```

Because the minimum temperature is 23, and the max is 95, the temperature must be in Fahrenheit. If it was Celsius, this would mean that the water temperature in Seattle reaches a maximum of near boiling, which is not the case.

4b)

i. Make a labeled scatterplot of temperature (y) versus wind direction (x).



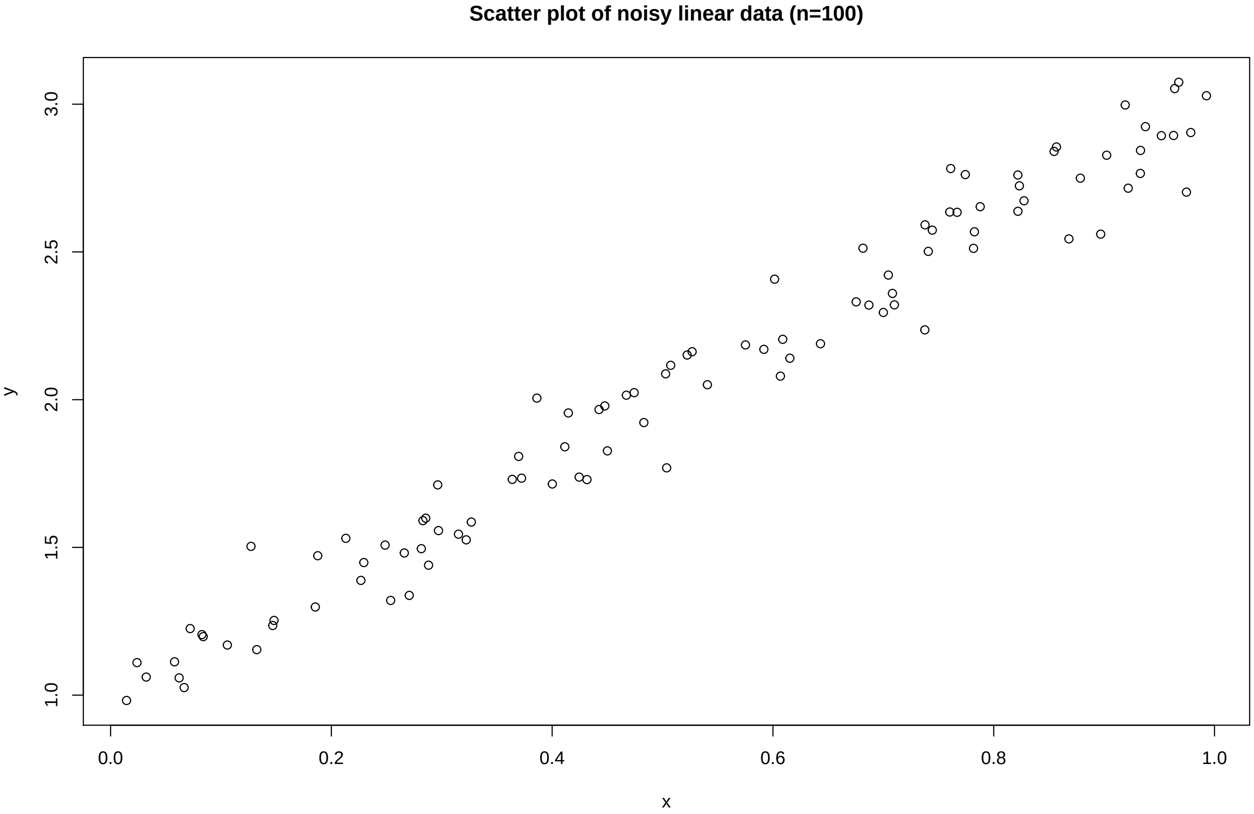
ii. Describe the general relationship between temperature and wind direction you observe.

There appears to be a vary slight positive association between wind direction and temperature in the center of the dataset's domain. However, it could be argued that there is not much of a relationship, and the plot looks very similar to an amorphous cloud.

iii. Does it make sense to calculate the correlation of temperature and wind direction? Explain.

No. Wind Direction is not a linear variable, so it does not make sense to find the correlation. We can see that 0 degrees and 360 degrees are identical, so we can not model this as a linear regression.

5. Next we'll investigate the defects of r. (Section 3.2, examples similar to the notes.)

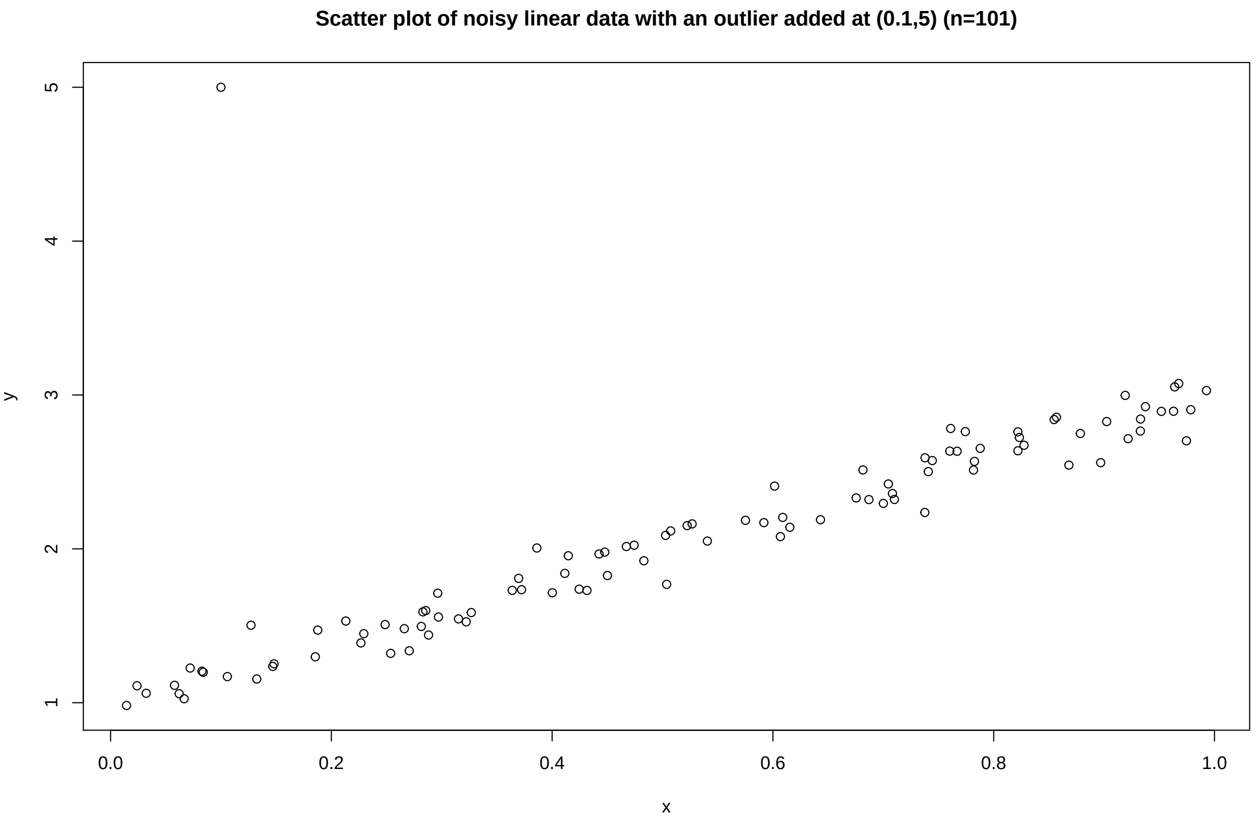


5a) Write down the correlation of x and y.

```
cor(x,y)
```

```
## [1] 0.983245
```

5b) Write down the correlation of the updated x and y.


Scatter plot of noisy linear data with an outlier added at (0.1,5) (n=101)

```
cor(x,y)
```

```
## [1] 0.8089377
```

5c) Perform the IQR rule-of-thumb for detecting potential outliers updated x and y. Are any outliers flagged?

```
cat ("LL_Y is ", LL_Y)
```

```
## LL_Y is  -0.1244369
```

```
cat ("UL_Y is ", UL_Y)
```

```
## UL_Y is  4.2894
```

```
cat ("Y outliers:", y[y > UL_Y | y < LL_Y])
```

```
## Y outliers: 5
```

```
cat ("X outliers:", x[x > UL_X | x < LL_X])
```

```
## X outliers:
```

By looking at the plot, we can see that most of the data is within the upper and lower limits for X, so that is not an issue. However, we can see there is a point near 5 for the Y, which is well beyond the upper limit. Therefore, this point is likely an outlier. This point is (0.1, 5).

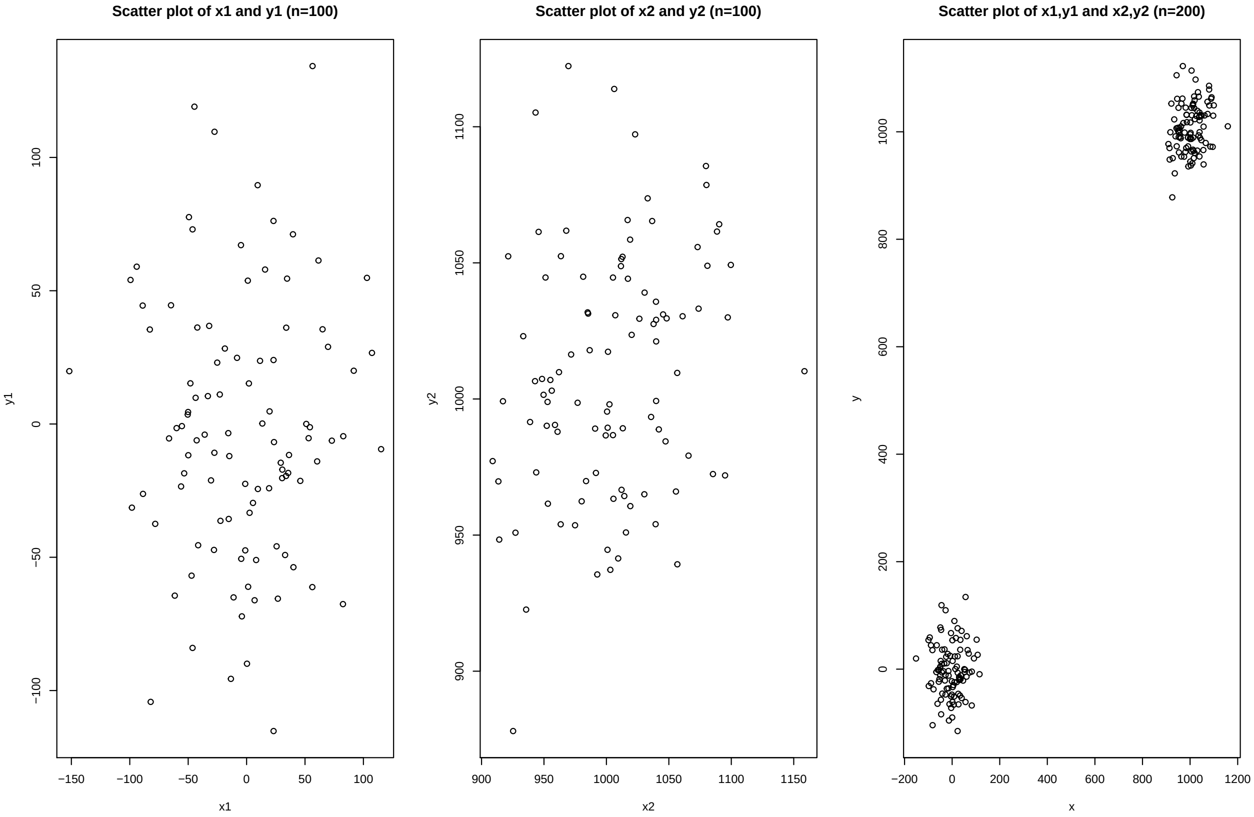
5d) Write down the correlation of the updated x and y. Would this outlier have been detected in a univariate EDA? Explain.

```
cor(x,y)
```

```
## [1] 0.8089377
```

Yes, we could have done the IQR Rule-of-Thumb on the univariate data, and found the outlier for the Y's, which is exactly what I did. This is because the IQR Rule-of-Thumb is part of univariate EDA analysis.

5e) Write down the correlation of the three pairs (the 1's, 2's and combined). How would you describe the distribution of x and y? Give an example from class where this kind of data was observed.



```
cor(x,y)
```

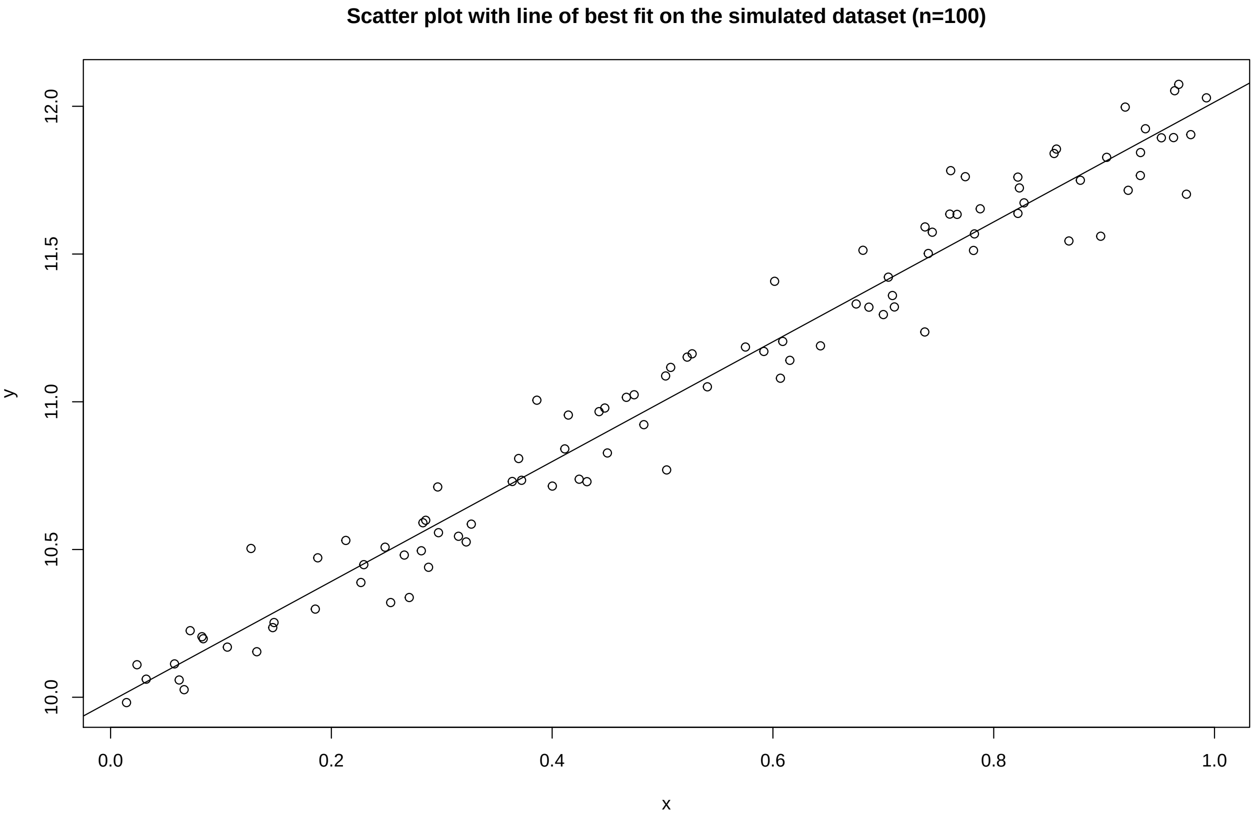
```
## [1] 0.991791
```

This distribution shown in the graph on the far right has two distinct clusters. This is often the case when the data is taken from two distinct and independent sample groups. An example of this that we discussed in class is the weight of pennies. The materials of the penny has changed over time, so taking samples at different years, results in clusters of masses due to the different weight of materials for the given time intervals.

6. Regression on the simulated dataset. (Section 3.3)

6a)

i. include the plot plus the line of best fit.



Scatter plot with line of best fit on the simulated dataset (n=100)

The line of best fit is: Y = 9.987 + 2.028x

ii. what is the class of the lm_1 object?

```
class(lm_1)
```

```
## [1] "lm"
```

The lm_1 object is of class “lm”.

iii. what information does the lm_1 object include?

The lm_1 includes the information of the slope and the intercept in the linear regression model for the data.

iv. what are the estimated slope and intercept of the line of best fit?

The estimated slope and intercept of the line of best fit are 2.028, and 9.987 respectively.

v. find and interpret the coefficient of determination.

The coefficient of determination is also known as the R^2 value, which is found by looking at the “Multiple R-squared” entry in the summary() function. This gives us the value of $\boxed{.9668}$.

6b)

i. What output to do you get when you compare the predicted values?

We get residual plots (scatter plots) that show the difference between the actual value and the predicted value. However, both methods produce identical outputs. They create residual plots that resemble amorphous clouds.

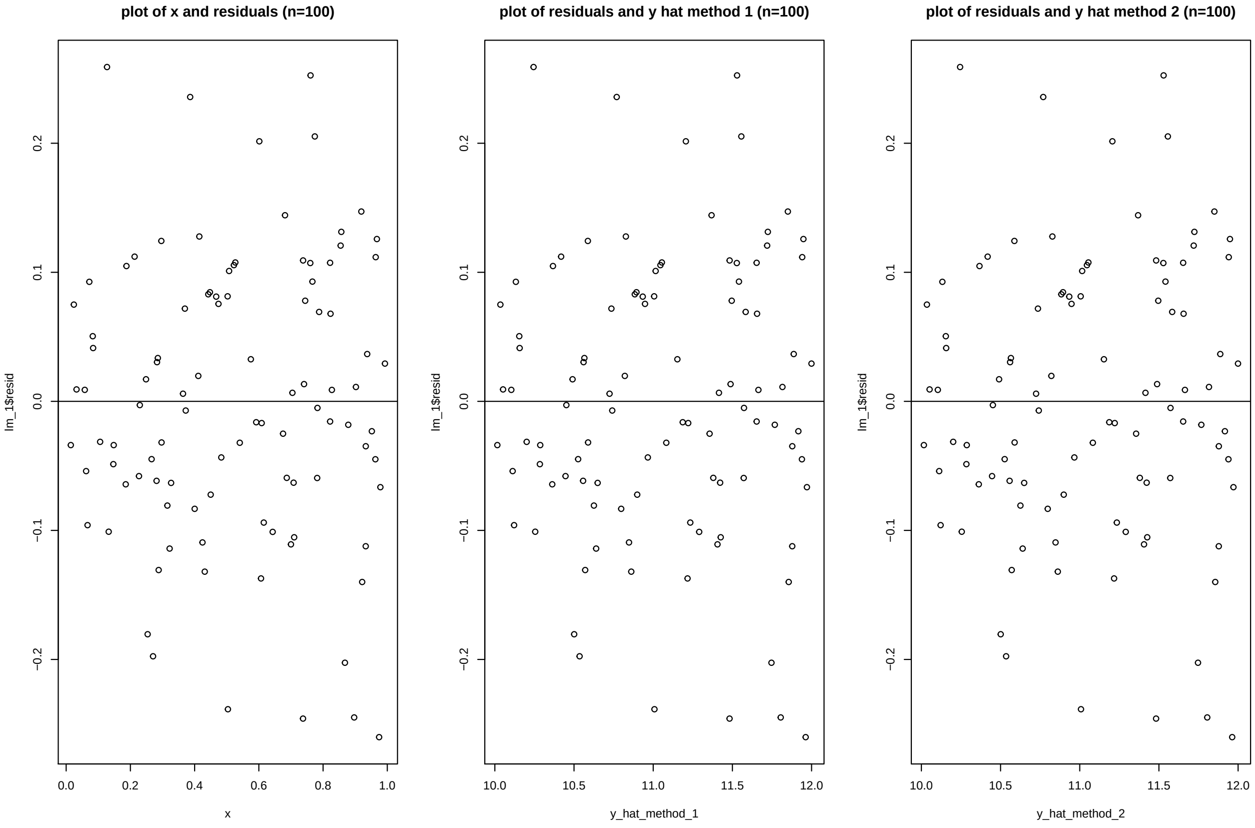
ii. Explain what is meant by 1e-13.

This is scientific notation for 1 x 10ˆ(-13).

iii.What do you conclude about the two methods used to calculate the predicted values?

Both methods produce identical outputs. They create residual plots that resemble amorphous clouds. Therefore, we can conclude that both methods give the same output, and either is correct to be used.

iv. Insert the plot here (there should be 3).



v. What do you observe about the 3 plots?

All have identical outputs. However, it should be noted that using one of the two y_hat_methods results in the domain changing. In spite of this, they create residual plots that resemble identical amorphous clouds. Therefore, we can conclude that all methods give the same output, and either method is correct to be used.

6c) In either case, what is the calculated correlation?

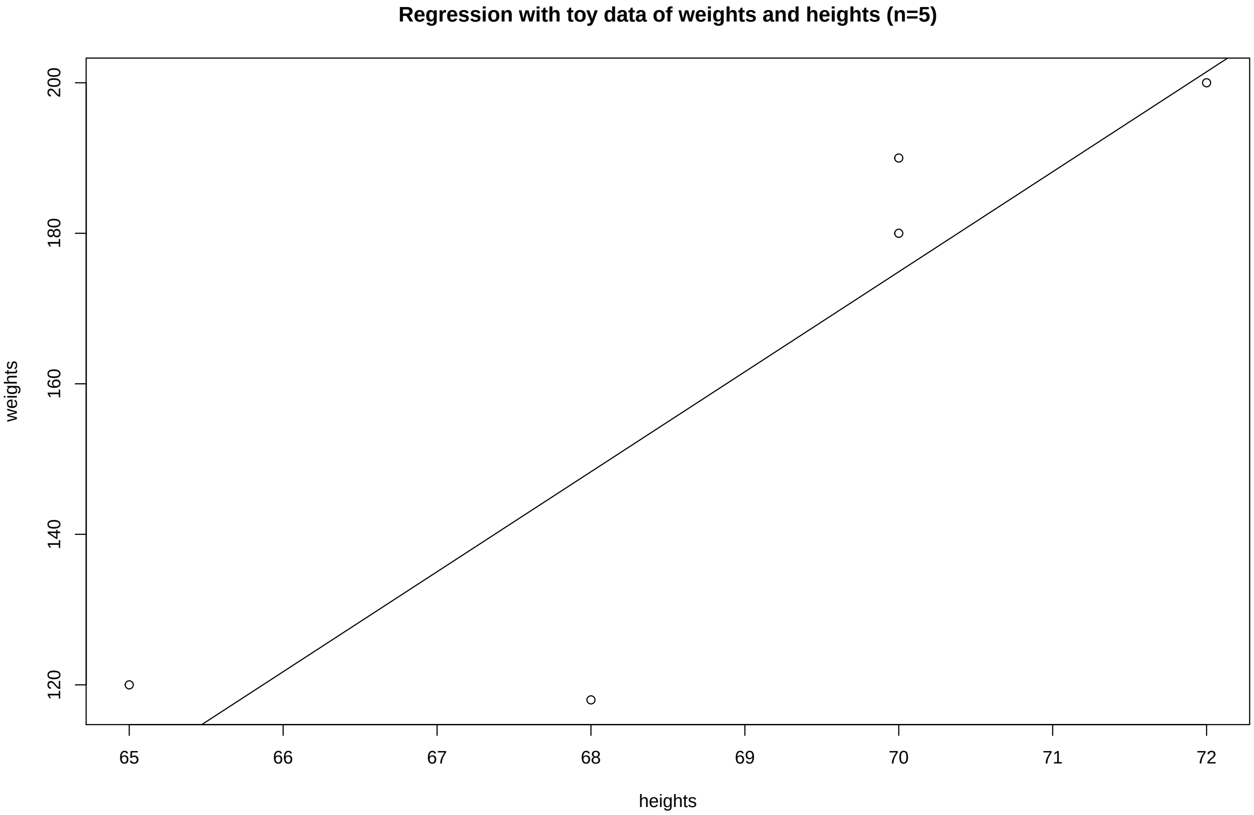
```
cor(x, lm_1$residuals)
```

```
## [1] -1.188077e-16
```

The calculated correlation is very close to 0, which is correct as the data is an amorphous cloud.

7. Regression with the toy data of weights & heights from class notes. (Section 3.3)

7a) Include the plot and the model information in your lab write-up.

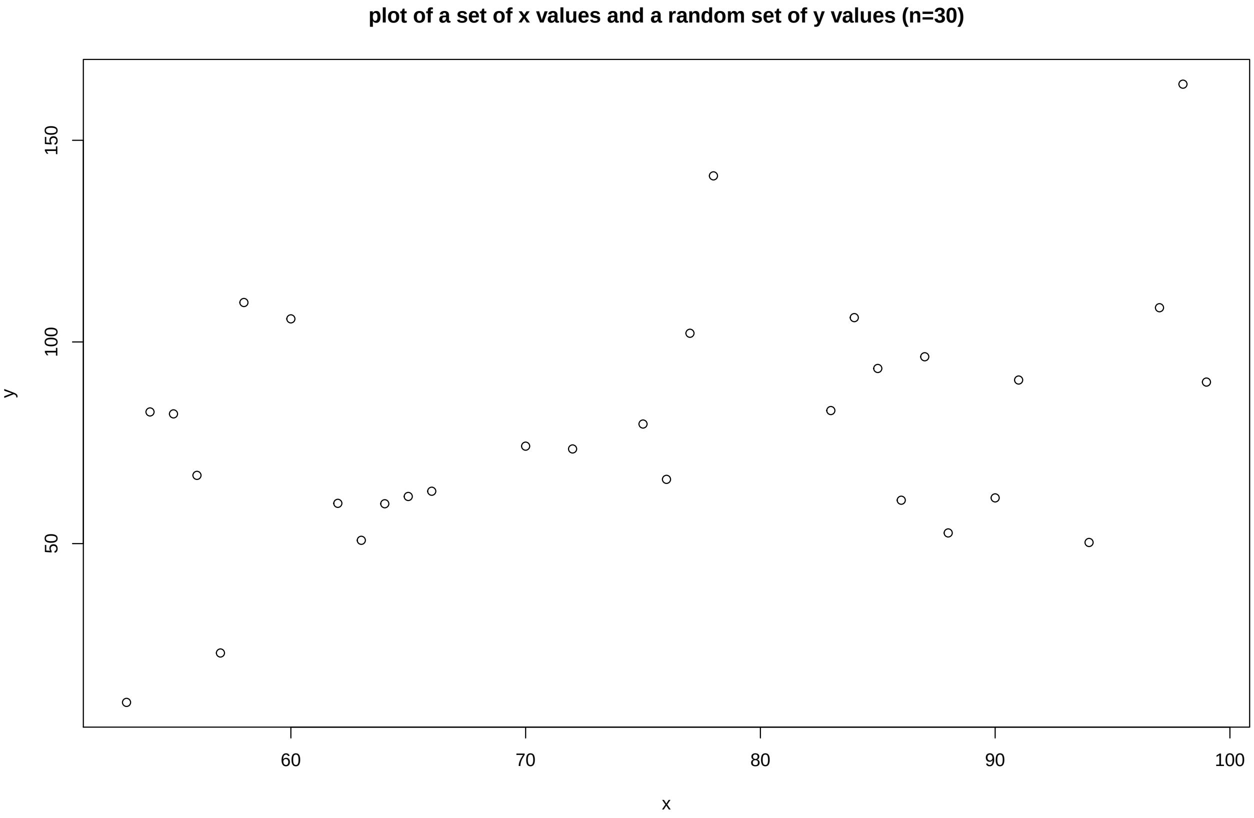


The model information is: Y = -755.11 + 13.29x

(Note: The units were not given)

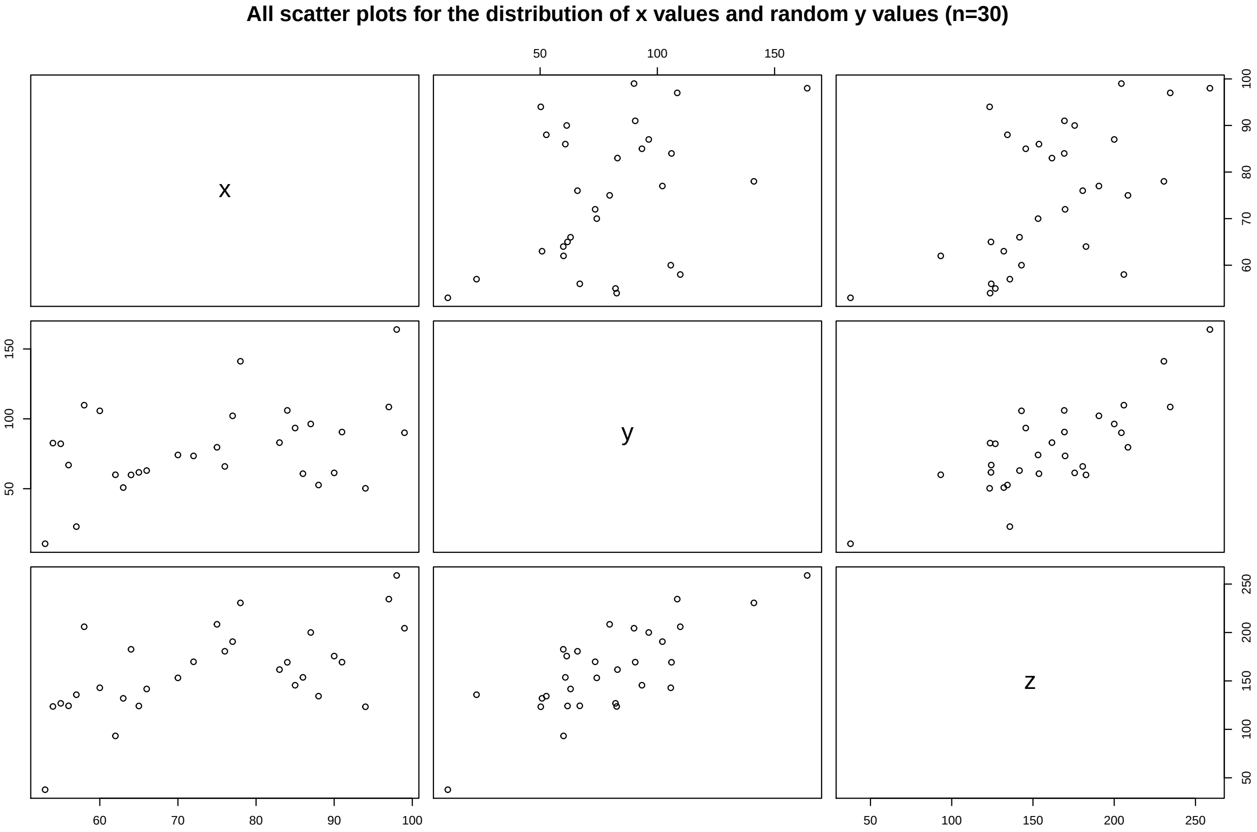
8. Example for later: Simultaneous scatter plots with 3 or more quantitative variables. (Section 3.5—multivariate regression)

8a) Describe what is plotted and include the plot.



Here we plot a range of x values from 50 to 100, with the y values being from this range of x values plus or minus a random value.

8b) Describe what is plotted and include the plot. This only works for data frames!



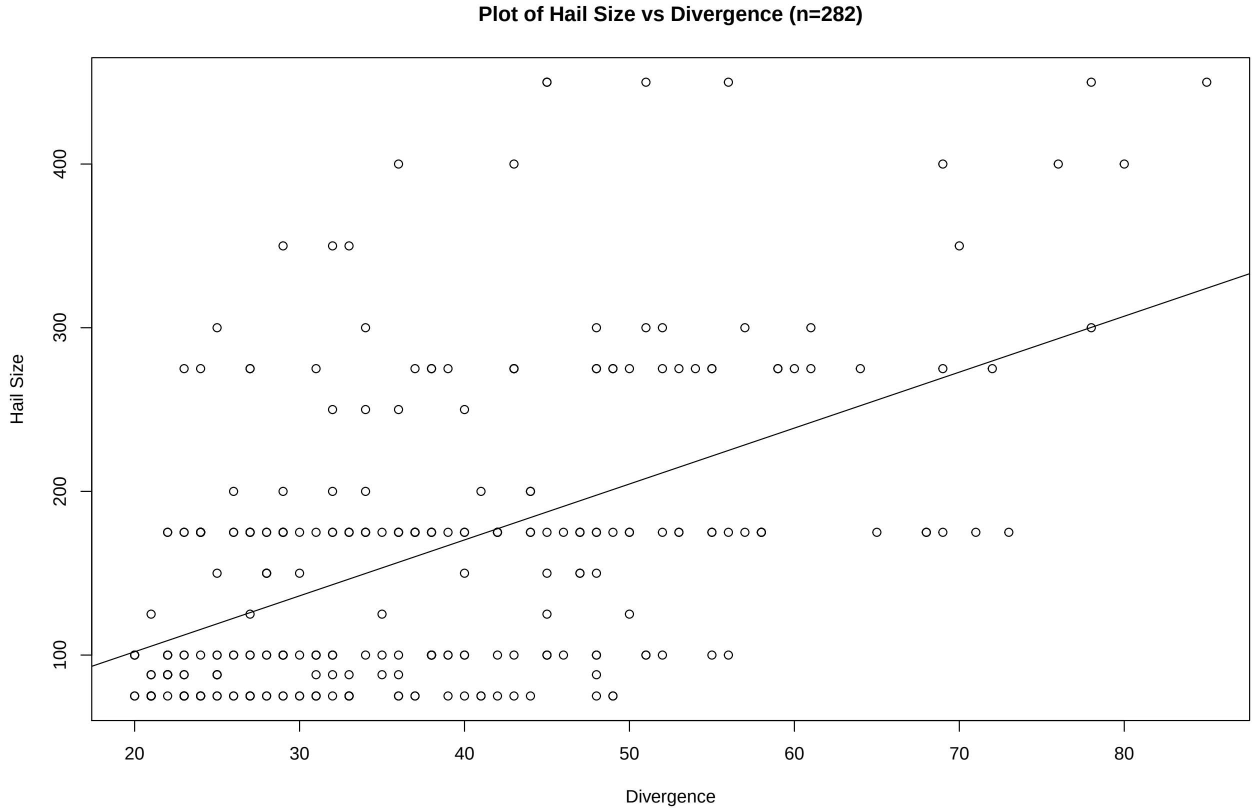
Here we make a scatter plot of every variable against every other variable in the model.

9a)

i. Do simple linear regression for predicting hail size from divergence.

```
lm_hail = lm(hail$Hail_size ~ hail$Divergence)
```

ii. Draw the model on an appropriately labeled scatter plot.



iii. Write down the equation of the regression line.

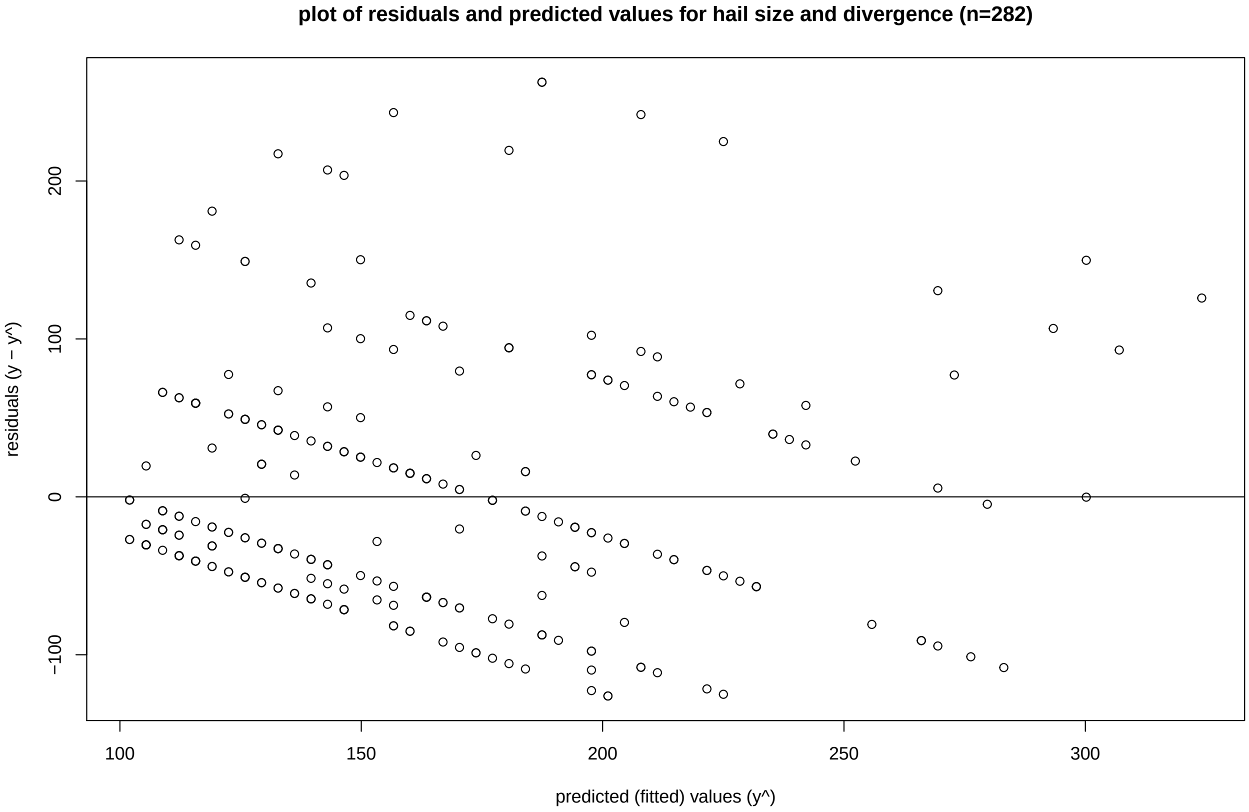
Y = 24.71176 + 0.07958x

(Note: Units not given)

iv. Find and interpret the coefficient of determination.

Rˆ2 is 0.2719. This means that 27.19% of the variability Hail Divergence is explained by a linear model of Hail Size. However, this does not make sense as divergence is not a linear variable. 360 is equivalent to 0, and it wraps around. Therefore, the coefficient of determination is irrelevant and is not useful.

v. Create a residuals plot and assess the model fit based on it.



We can see that the residual plot reasonably resembles an amorphous cloud, indicating that there is very little variation not accounted for by our linear regression model. I.E, our model fits our data very well. However, there does appear to be some sort of pattern, so depending on the analysis, it could be argued there is some variation not accounted for by our model.

vi. Find the correlation between the residuals and predicted values, and assess the model fit based on it.

```
cor(y_hat_method_1, lm_hail$resid)
```

```
## [1] -2.345215e-17
```

Therefore, we can see that there is nearly a coefficient of 0, meaning that linear regression model barely fits the data, and this suggests that the residual plot resembles an amorphous cloud. This is good. This means that there is very little variation not accounted for by our linear regression model. I.E, our model fits our data very well. However, I still have to look at the plot to determine this, as there could be variation in the residual plot, with the coefficient of determination still being close to 0. I have to look at the data to verify the meaning of this value.

vii. Compare your conclusions in v and vi.

Looking at vi, we would look at the correlation coefficient and see that our residual plot has a value close to 0, indicating that most of the variation was taken up by the model. However, looking at the residual plot, there appears to be some distinct pattern in the plot. This indicates that there might be some variation that was not accounted for by the model. So likely, both conclusions are not the same.

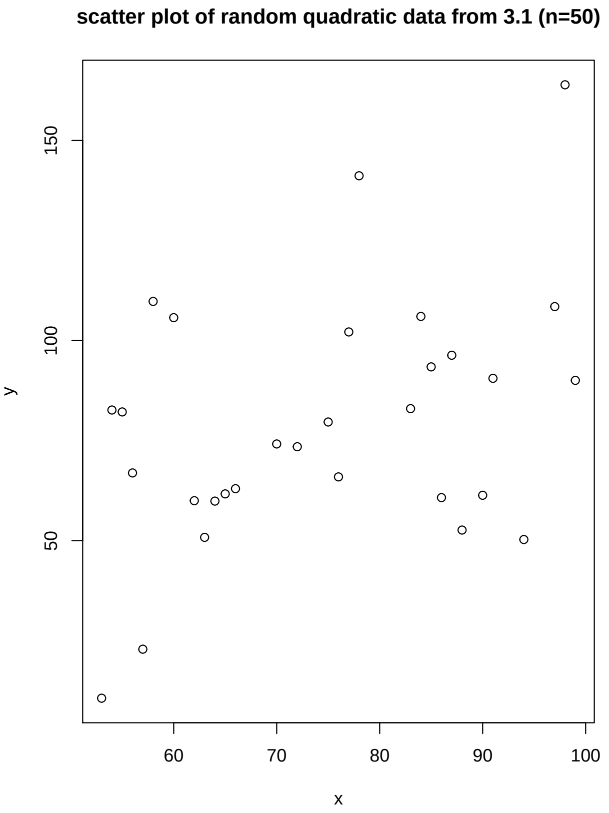
Take-home message: always look at the residuals!

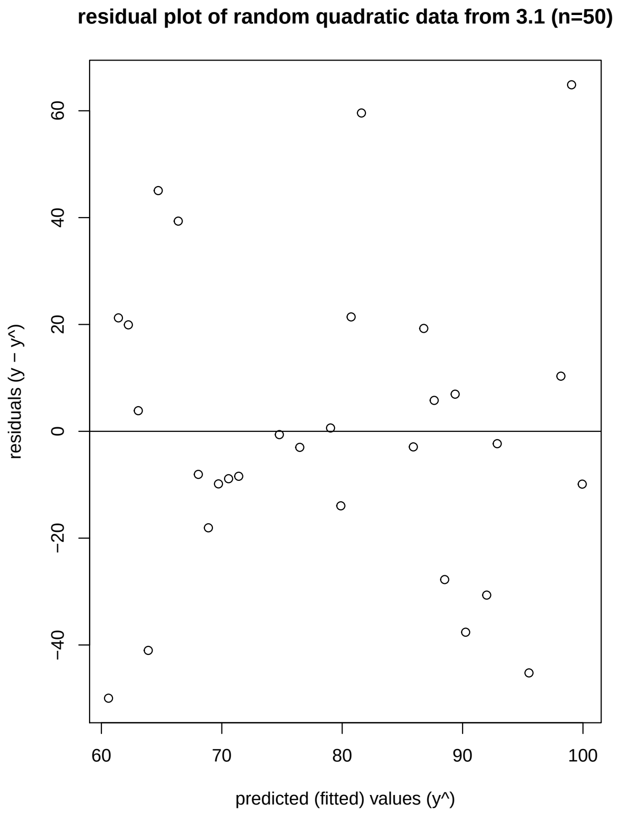
10. Nonlinearity example: quadratic model (Section 3.4)

10a) What do you note about the two models (lm_2a and lm_2b)

I notice that lm_2a and lm_2b are the exact the same model, despite one of them squaring the x value.

10b) Include your plot here, discuss the residual plot, and interpret the correlation found.


scatter plot of random quadratic data from 3.1 (n=50)


residual plot of random quadratic data from 3.1 (n=50)

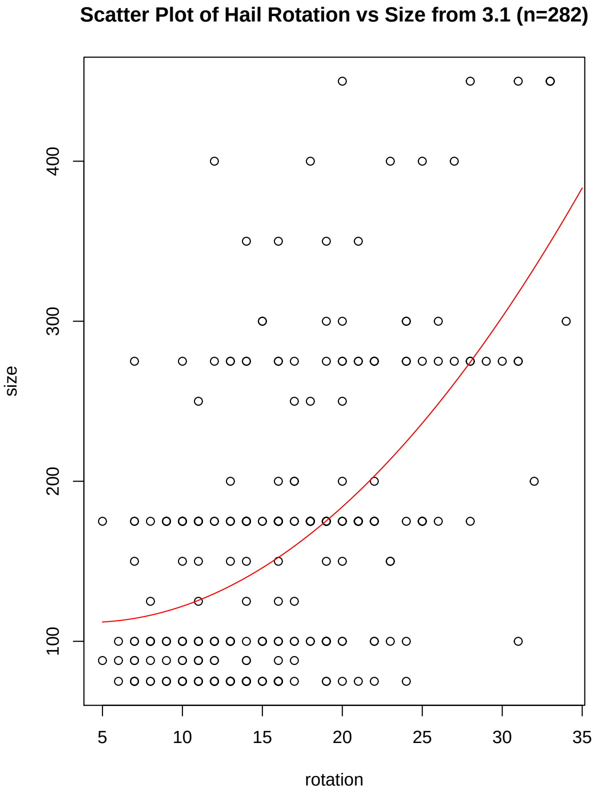
Here we can see that the residual plot appears to be an amorphous cloud, indicating that all variation was accounted for by the model. This is further verified by our correlation coefficient being close to 0 for our residual plot.

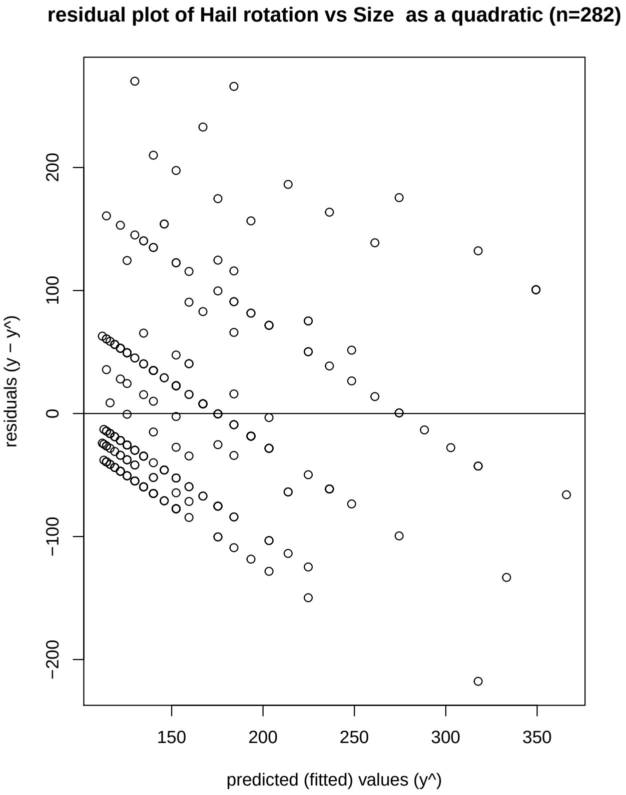
11. Go back to the hail data. (Sections 3.4 and 3.5)

The following code will get you started.

11a) Apply your knowledge to model hail size as a quadratic function of rotational velocity.

Include your code, the scatter plot with the plot of the model superimposed, and the residual plot. Note: if you use the code provided above, you'll need to choose an appropriate range of values for xlim.


Scatter Plot of Hail Rotation vs Size from 3.1 (n=282)


residual plot of Hail rotation vs Size as a quadratic (n=282)

```
summary (lm_hail_quad)
```

```
##
## Call:
## lm(formula = size ~ rotation + I(rotation^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217.72  -50.61  -18.39   40.47  270.16
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   116.35190   28.52568   4.079 5.91e-05 ***
## rotation       -2.26824    3.37281  -0.673  0.50182
## I(rotation^2)   0.28271    0.09197   3.074  0.00232 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.93 on 279 degrees of freedom
## Multiple R-squared:  0.3133, Adjusted R-squared:  0.3084
## F-statistic: 63.65 on 2 and 279 DF,  p-value: < 2.2e-16
```

12. Summarize the R functions you learned for performing the tasks listed above. Then decide whether anything is missing from the summary above, and write that here as well.

A Make sure you have two quantitative variables.

class(): If a given variable is quantitative, it will be classified as ‘numeric’ as if inputted into the class() function.

B Repeat the steps for univariate exploratory analysis (see the Summary in Lab 2).

summary(): Find the five number summary (Q_3, Q_2 (median), Q_1, min, max, IQR ($Q_3 - Q_1$))

Find IQR by doing Q3-Q1, and upper limit by Q1-1.5IQR and Q3+1.5IQR.

C Perform bivariate analyses of the variables.

Look at the scatter plot to see general distribution: plot(x,y)

Do a linear regression: lm(y~x)

Look at residual plot (want amorphous cloud): plot(lm$resid ~ lm$fitted.values)

i Determine which variable should be the response and which variable should be the predictor; explain your choice.

The cheaper variable should be the predictor (because it’s cheaper to gather data).

If the variable is independent, then it will be the predictor.

You have to read the question to find this out.

Furthermore, on a scatter plot, the predictor will be on the x-axis, while the response will be on the y-axis.

ii Look at a scatter plot of the data. Describe the pattern. Is it linear? Monotonic but nonlinear? Non-monotonic?

Look at scatter plot: plot(x,y)

Linear means that the line linear regressions fits the model, and that the r value is close to 1.

Monotonic but nonlinear means that the data is either always increasing or always decreasing, but is not at a linear rate (EX: xˆ2 quadratic).

Non-monotonic means that the data increases, and decreases.

iii If monotonic: if necessary, transform the predictor and/or response until the resulting scatter plot is roughly linear. Watch out for the domain of the data. Then perform a regression on the transformed data.

Transform: lm_hail_quad = lm(size ~ rotation + I(rotationˆ2))

Try different models (here we tried a quadratic)

Regression: lm(y~x)

iv If non-monotonic: decide on an appropriate degree polynomial to model the relationship. Perform a regression using the model you choose.

Count the number of turns you make, and add 1 to this:

For example, for a quadratic, you make 1 turn (one “U”), then add 1, and we get 2.

This is a second degree polynomial.

This same process works for polynomials of degree n.

Perform Regression: lm_hail_quad = lm(size ~ rotation + I(rotation^2))

Here we added the $I()^2$ term because we are doing a regression on a second degree polynomial.

v In either case, assess the model fit using the residual plot (residuals versus fitted values) and R 2 (and its adjusted version, if appropriate).

Look at the residual plot (want amorphous cloud): plot(lm$resid ~ lm$fitted.values)

If it does not resemble an amorphous cloud, some of the variation in the data was not fully accounted for by the model. I.E., you should try to find a better model.

R^2 should be close to 1 on the for the linear regression plot. You can find this by running summary(lm)