Exercise 9.24: Damages to grapes from predation is a serious problem for grape growers . . . Consider the following data on time spent on a single visit to the location,

1. Calculate the upper confidence bound for the true average time that blackbirds spend on a single visit at the experimental location.

   Solution:
   Finding the t-value for a 95% CI with 64 degrees of freedom we get that $t_{(.025,64)} = 1.671$. Calculating the upper bound of the confidence interval with the given $\bar{x} = 13.4$, $SE = 2.05$,

   $$\bar{x} + t_{(.025,64)}SE = 13.4 + 1.671(2.05) \approx 16.825.$$

2. Does it appear that the true average time spent by blackbirds at the experimental location exceeds the true average time birds of This type spend at the natural location? Carry out the test of hypothesis.

   Solution:
   First we must use the SE and the sample size to compute the sample standard deviation for each location,

   $$s_1 = 2.05(\sqrt{65}) \approx 16.527,$$
   $$s_2 = 1.76(\sqrt{50}) \approx 12.445.$$

   From the claim made in the problem statement we srt up the hypothesis test as follows,

   $$H_0 : \mu_1 = \mu_2,$$
   $$H_a : \mu_1 > \mu_2.$$

   Now we must solve for the test statistic t and it's degrees of freedom v,

   $$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{13.4 - 9.7}{\sqrt{\frac{16.527^2}{65} + \frac{12.445^2}{50}}} = 1.37.$$

   $$v = \frac{(SE_1^2 - SE_2^2)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}} = \frac{(2.05^2 - 1.76^2)^2}{\frac{2.05^4}{64} + \frac{1.76^4}{49}} \approx 112.$$

   Assuming a significance level of $\alpha = .05$. From our table A.5 we can see that our P is greater that our significance level and therefore we fail to reject the null hypothesis.

3. Estimate the difference between the true average time blackbirds spend at the natural location and the true average time that silvereyes spend at the natural location, and do so in a way that conveys information about reliability and precision.

Solution:
To proceed we will calculate a 95% CI for $\bar{x_1} - \bar{x_2}$, where the sample mean refers to the blackbirds and silvereyes respectively. Solving for the degrees of freedom of $\bar{x_1} - \bar{x_2}$,

$$v = \frac{(SE_1^2 - SE_2^2)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}} = \frac{(1.76^2 - 5.06^2)^2}{\frac{1.76^4}{49} + \frac{5.06^4}{45}} \approx 55.$$

Then using the SE and the sample size to compute the sample standard deviations,

$$s_1 = 1.76(\sqrt{50}) \approx 12.445,$$

$$s_2 = 5.06(\sqrt{46}) \approx 34.317.$$

Using our table we get that $t_{(.025,55)} = 2.006$. Finally we can calculate the 95% CI for the difference of means.

$$95\%CI = (9.7 - 38.4 \pm 2.009\sqrt{\frac{12.445^2}{50} + \frac{34.317^2}{46}}) \approx (-10.84, 10.68)$$

Exercise 9.28: As the population ages, there is increasing concern about accident-related injuries to the elderly... Does the data suggest that true average maximum lean angle for older females is more than 10 degrees smaller than it is for younger females? State and test the relative hypothesis at a significance level .10

Solution:
First we need to compute the sample mean and standard deviation for each set of data.
Console:

```
> YF =  c(29, 34, 33 ,27, 28, 32, 31, 34, 32, 27)
> OF =  c(18, 15,23, 13, 12)

> YF_mean = mean(YF)
    [1] 30.7

> OF_mean = mean(OF)
    [1] 16.2

> YF_sd = sd(YF)
```

```
[1] 2.750757

> OF_sd = sd(OF)
[1] 4.438468
```

Given the problem statement we get the following one sided hypothesis test,

$$H_0 : \mu_1 - \mu_2 = 10,$$

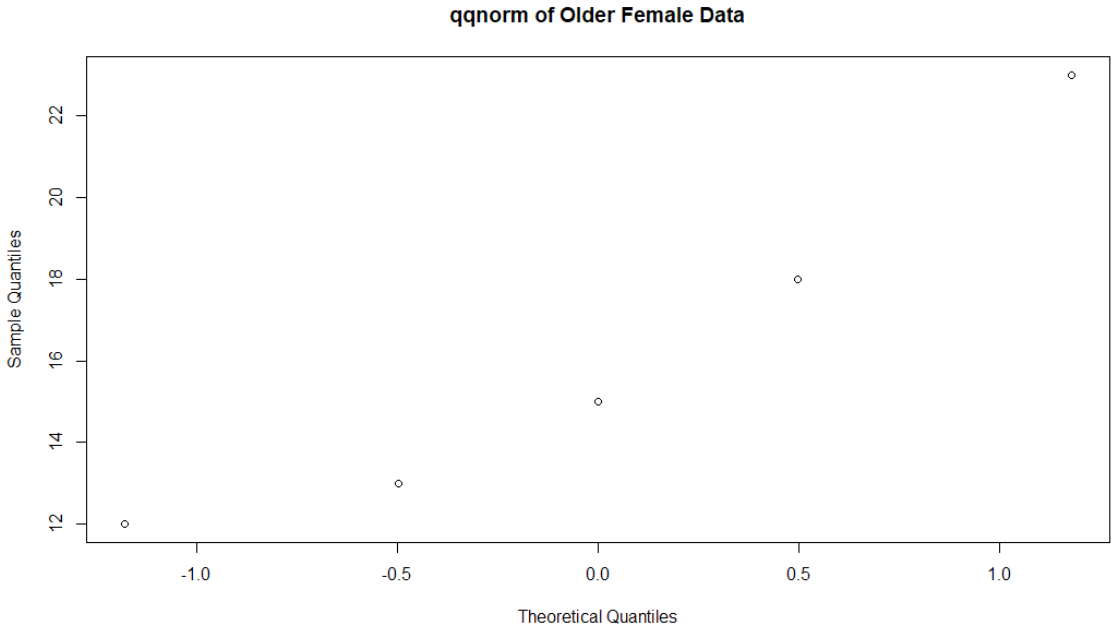
$$H_a : \mu_1 - \mu_2 > 10.$$

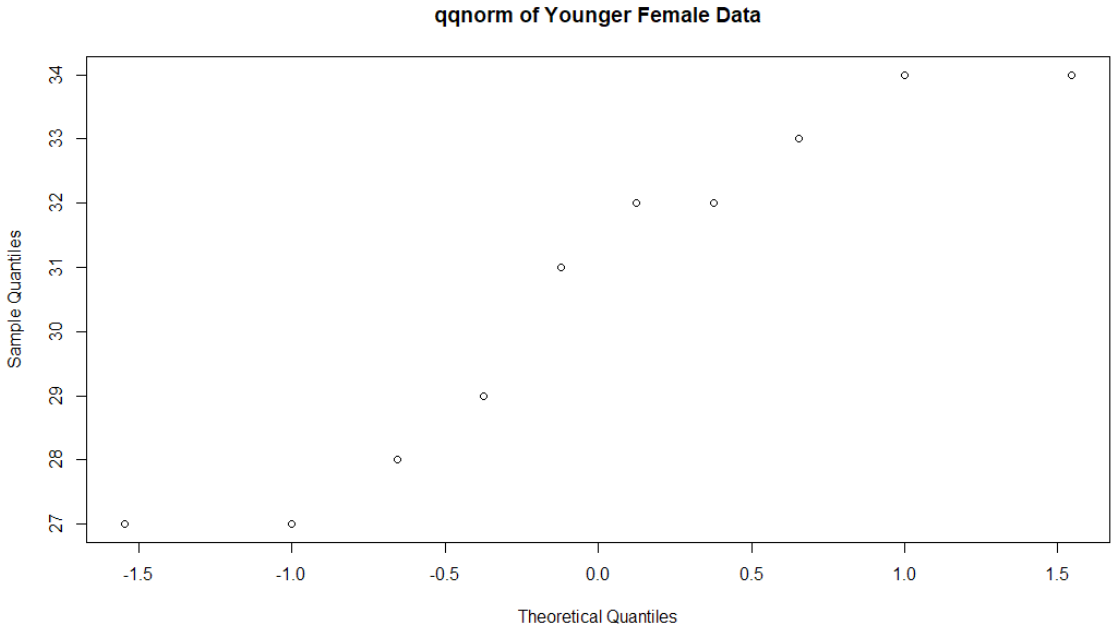
Finally determining the test statistic t and degrees of freedom df,
Console:

```
> t = (YF_mean - OF_mean)/sqrt((YF_sd^2/10) + (OF_sd^2/5))
[1] 6.690724

> df = (((YF_sd^2/10) + (OF_sd^2/5))^2)/
      (((YF_sd^2/10)^2/9) + (OF_sd^2/5)^2/4)
[1] 5.592239
```

Using r we get a $p - value = 0.0007221$ which is far below our significance level of .10 which means that we reject the Null hypothesis and claim that the true average maximum lean angle for older females is more than 10 degrees smaller than it is for younger females.

Consider the qqnorm plot for each data set.





Note that with the limited data points we can see the qqnorm plot for the older female data has a slight curve to it. This suggests that the data itself is skewed. However with so little data its difficult for me to asses the normality, visually.

Exercise 9.42: many freeways have service signs that give information on attractions, camping, lodging, food, and gas services prior to off-ramps. These signs typically do not provide information on distance. . . carry out a paired t test to determine whether there was any change in the mean number of crashes before and after the addition of distance information on the signs.

Solution:
From the problem statement we get the following hypothesis test,

$$H_0 : \mu_1 - \mu_2 = 0,$$

$$H_a : \mu_1 - \mu_2 \neq 0.$$

Solving for our test statistic t for a paired t test,

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} = .72562$$

Where D is the pairwise difference between datasets,
Console:

```
> Before =  c(15,26,66,115, 62, 64)
> After = c(16, 24, 42, 80, 78, 73 )
> D = Before - After

> D_mean = mean(D)
    [1] 5.833333

> D_sd = sd(D)
    [1] 19.69179

> t.test(D)

        One Sample t-test

data:  D
t = 0.72562, df = 5, p-value = 0.5006
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -14.83193  26.49860
sample estimates:
mean of x
 5.833333
```

Either using the table or r we get a $p - value \approx .5$. Compared to a significance level of $\alpha = .05$ we get that we fail to reject the Null hypothesis.
Calculating a 95% CI for $\overline{d}$ by using a 2- sided critical value of $z_{(.025,5)}$,

$$95\%CI = (5.83 \pm 2.571\frac{19.69}{\sqrt{6)}} = (-14.83, 26.49)$$

Exercise 9.50: Recent incidents of food contamination have caused great concern among consumers... 35 of 80 randomly selected Perdue brand broilers tested positively for either campylobacter or salmonella. Whereas 66 of 80 Tyson brans broilers tested positive.

1. Does it appear that the true proportion of non-contaminated Perdue broilers differs from that for the Tyson brand? Carry out the hypothesis test using a significance level of $\alpha = .01$

   Solution:
   From the problem statement we know the following sample proportions,

$$\hat{p}_1 = \frac{35}{80}$$

$$\hat{p}_2 = \frac{66}{80}$$

We also know that since they are asking if the true proportions differ, we set up the following Hypothesis test,

$$H_0 : p_1 - p_2 = 0,$$
$$H_a : p_1 - p_2 \neq 0.$$

Based on our assumption in the null hypothesis we get that the point estimator for the combined sample proportion is,

$$\hat{p} = \frac{35 + 66}{160} = \frac{101}{160}$$

Solving for our test statistic z,

$$z = \frac{\frac{35}{80} + \frac{66}{80}}{\sqrt{\frac{101}{160}\frac{59}{160}(\frac{1}{40})}} \approx -5.08$$

From our Table A.3 we can clearly see that a test statistic of $z = -5.08$ we get a $p - value$ that is guaranteed to be greater than α and thus there is sufficient evidence eot reject the null hypothesis and claim that the true proportion of non-contaminated Perdue broilers differs from that for the Tyson brand.

2. If the true proportions of non-contaminated chickens form the perdue and tyson brands are .50 and .25 respectively, how likely is it that the null hypothesis of equal proportions will be rejected when a .01 significance level is used and the sample sizes are both 80?

Solution:

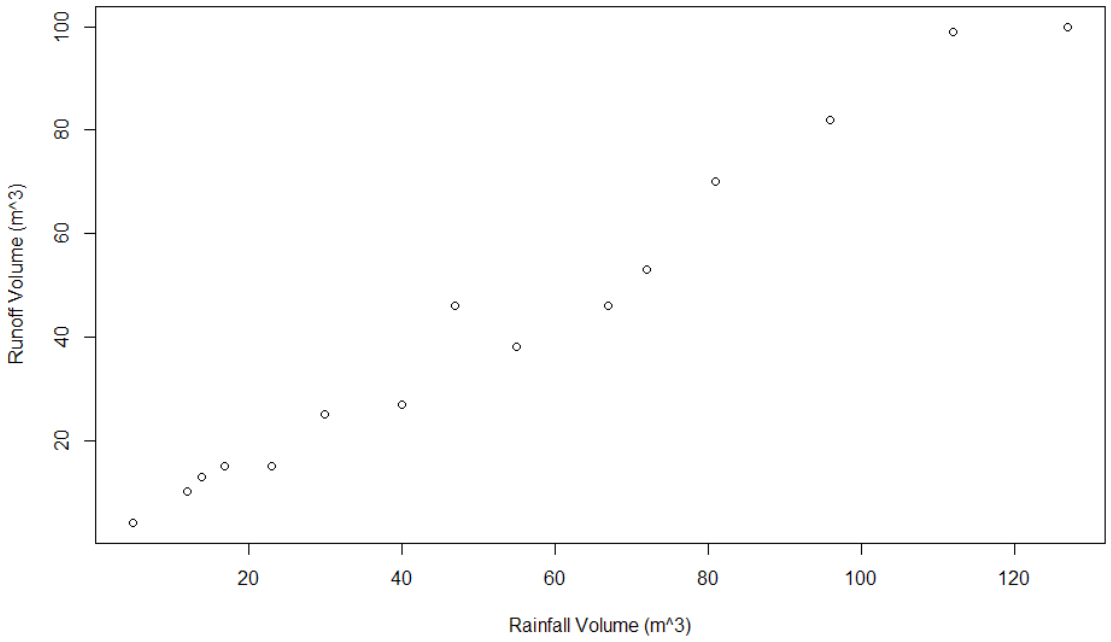
Exercise 12.6: One factor in the development of tennis elbow, a malady that strikes fear in the hearts of all serious tennis players, is the impact-induced vibration of the racket-and-arm system at ball contact…Consider the scatter plot, and discuss interesting features of the data and scatter plot.

Solution:
The data looks to have a lot of variation, with two data points in the lower right that would be considered outliers. Outside of those two points we can see that there is a negative linear relationship between the racket resonance frequency and sum of peal-to-peak acceleration.

Exercise 12.16: Given a scatter plot along with the least squares line of rainfall and runoff volumes for a particular location.

1. Does the scatter plot of the data support the use of the simple linear regression model.

   Solution:
   The scatter plot definitely supports the use of the simple linear regression model. There is no strong curvature, or even extreme outliers.



2. Calculate point estimate of the slope and intercept of the population regression line.

   Solution:
   Calculating the point estimate for the slope and intercept by using the following least squares solutions,

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$$

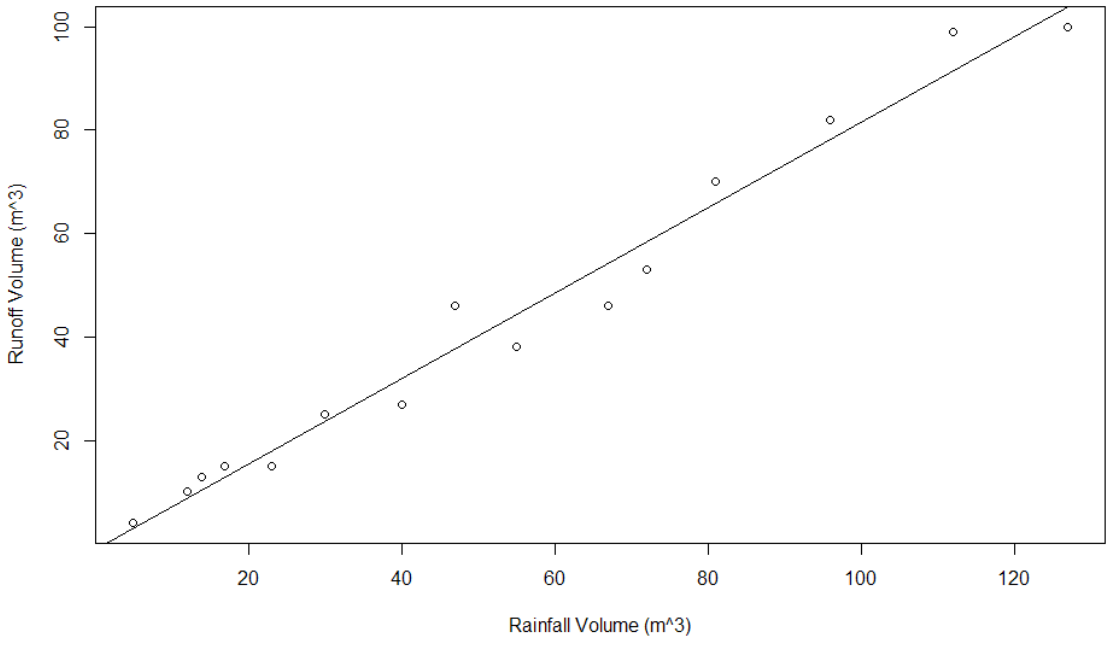
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

   Console:

```
> x = c(5, 12, 14, 17, 23, 30, 40, 47, 55, 67, 72, 81, 96, 112, 127)
> y = c(4, 10, 13, 15, 15, 25, 27, 46, 38, 46, 53, 70, 82, 99, 100)


> x_mean = mean(x)
> y_mean = mean(y)

> Diff_x = x - x_mean
> Diff_y = y - y_mean

> Numerator = Diff_x*Diff_y

> Denominator = Diff_x*Diff_x

> slope = sum(Numerator)/sum(Denominator)
    [1] 0.8269731

> intercept = mean(y) - slope*mean(x)
    [1] -1.128305
```

Finally we get that our simple linear regression is,

$$f(x) = 0.8269x - 1.1283.$$



3. Calculate the point estimate of the true average runoff volume when rainfall volume is at 50.

Solution:
Simply plugging $x = 50$ into our linear regression will yield the desired runoff point estimate.

$$\hat{y} = 0.8269(50) - 1.1283 = 40.2204.$$

4. Calculate the point estimate for the standard deviation.

Solution:
Calculating the point estimate for the standard deviation using our linear regression,

$$\sigma = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}.$$

Console:

```
> Point_Estimates = slope*x + intercept

> SSE = sum((y - Point_Estimates)^2)

> sigma = sqrt(SSE/(length(y) - 1))
    [1] 5.049835
```

5. What proportion of the observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall?

Solution:
The coefficient of determination is calculated by,

$$r^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2}.$$

Console:

```
> SSE = sum((y - Point_Estimates)^2)

> SST = sum((y - y_mean)^2)

> r2 = 1 - (SSE/SST)
    [1] 0.9752689
```

With a $r^2 = 0.9752$ it means that approximately 97.5% of observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall.

Exercise 12.24: the invasive diatom species...Consider the relation between y = colony density, and x =rock surface area.

1. Fit the simple linear regression model to this data, predict colony density when surface area = 70 and when surface area = 71. Calculate the corresponding residuals, how do they compare?

   Solution:
   Plugging the data into r and running a similar instruction set to the previous problem we get a linear regression of

   $$f(x) = 9.963x - 305.881$$

   Using our regression to get a point estimate for the density of the colony when $x = 70, 71$,

   $$f(70) = 9.963(70) - 305.881 = 391.55,$$

   $$f(x) = 9.963(71) - 305.881 = 401.51.$$

   Calculating the corresponding residuals,

   $$\epsilon_7 0 = 13 - f(70) = -378.55,$$

   $$\epsilon_7 1 = 1929 - f(71) = 1527.49.$$

   Clearly we can see that $\epsilon_7 1$ is much larger than $\epsilon_7 0$. We could see that in general when a linear regression has such large residuals, the coefficient of determination is likely to be very small, making this regression unsuccfull at modeling the relationship between surface area and colony density.

2. Calculate and interpret the coefficient of determination.

   Solution:
   Again calculating the coefficient of determination in r with the same formula described in the previous problem.

   $$r^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} \approx 0.12371.$$

With a $r^2 = 0.12371$ it means that our linear regression can explain approximately 12.4% of the variation between colony density and rock surface area.

3. Remove the outlier and recalculate the regression. Is it substantially different? What is the impact on r^2 and s?

Solution:
After removing the outlying data point and recalculating the linear regression in r, we get the following function,

$$f(x) = 0.779x + 34.373.$$

Note that this function is considerably different from the previous one that we computed. Both the slope and intercept are on the order of 10 times smaller than before. Calculating the coefficient of determination r^2 and σ for both data sets, we get that, $r^2 = 0.0238$ and $\sigma = 87.222$ for the new data set, and $\sigma = 5.049835$ for the old data set. We can observe that our new model is even worse than before after removing an extreme outlier.

Exercise 12.34: Electro magnetic technologies offer effective nondestructive sensing techniques for determining characteristics of pavement ... Consider that y = dielectric constant and x = air void (%).

1. Obtain the equation of the least squares and interpret its slope.

Solution:
From the given r output we can see that the least squares equation is,

$$f(x) = -.074676x + 4.858691$$

Since the slope is negative we know that the lower the air void (%) the greater the dielectric constant, and on average for an increase of 1% in air void we can expect the dielectric constant to decrease by .0746.

2. What proportion of observed variation in dielectric constant can be attributed to the approximate linear relationship between dielectric constant and air void.

Solution:
From the given r output we can see the coefficient of determination is $r^2 = 0.7797$.

Therefore approximately 77.9% of observed variation in dielectric constant can be attributed to the simple linear regression relationship between dielectric constant and air void.

3. Does there appear to be a useful linear relationship between dielectric constant and air void? State and test the appropriate hypothesis?

Solution:
For this problem we want to use the Model Utility Test, recall the following,

$$H_0 = \beta_1 = 0.$$

$$H_0 = \beta_1 \neq 0.$$

Note that when $\beta_1 = 0$ there is no useful linear relationship between dielectric constant and air void. We can see that our test statistic t and $p - value$ are calculated in the given r output,

$$\frac{\hat{\beta_1} - 0}{s_{\hat{\beta_1}}} = \frac{-0.074676}{0.009923} = -7.526,$$

$$p = 1.21e - 06.$$

With such an incredibly small $p - value$ we reject the null hypothesis and therefore it must be the case that there is a significant linear relationship between dielectric constant and air void.

4. Suppose it had previously been believed that when air void increased by 1 percent the associated true average change in dielectric constant would be at least -.05. Does the sample data contradict this belief? Carry out a test of hypothesis with a significance level of .01

Solution:
From the problem statement we set, $\beta_{10} = -.05$ and consider the following hypothesis test,

$$H_0 = \beta_1 = \beta_{10},$$

$$H_0 = \beta_1 \neq \beta_{10}.$$

Calculating our t statistic using $\hat{\beta_1}$ and $s_{\hat{\beta_1}}$ from the given r output.

$$\frac{\hat{\beta_1} - \beta_{10}}{s_{\hat{\beta_1}}} = \frac{-0.074676 - (-.05)}{0.009923} = -.249,$$

Using our r calculate a $p - value = 0.8065$ on 16 degrees of freedom. Since our $p - value$ is considerably larger that our significance level we cannot reject our null and therefore the sample data does not contradict the belief.

Exercise 12.46: Astringency is the quality in a wine that makes the wine drinker's mouth feel slightly rough, dry, and puckery…

1. Fit the simple linear regression model to this data, Then determine the proportion of observed variation in astringency that can be attributed to the model relationship between astringency and tannin concentration.

   Solution:
   From the given summary quantities amd the fact that $\overline{x} = \frac{19.404}{32}$ and $\overline{y} = \frac{-.542}{32}$ and we can quickly calculate the slope and intercept of linear regression,

   $$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}} = 2.585,$$

   $$\hat{\beta_0} = \overline{y} - 2.585\overline{x} = -1.585.$$

   Thus the function for our linear regression is,

   $$f(x) = 2.585x - 1.585.$$

   To solve for the coefficient of determination we must first solve for, Error sum of squares and total sum of squares. By the definition and the linearity of sums we reduce to the following formulas,

   $$SSE = S_{yy} - \hat{\beta_1} S_{x,y} = 1.9243.$$

   $$SST = S_{yy} = 11.8263.$$

   Finally computing r^2,

   $$r^2 = 1 - \frac{1.9243}{11.8263} \approx .84.$$

   Thus 84% of the proportion of observed variation in astringency that can be attributed to the model relationship between astringency and tannin concentration.

2. Calculate and interpret a confidence interval for the slope of the true regression line.

   Solution:
   In order to calculate a 95% CI we must first calculate the point estimate $s_{\hat{\beta_1}}$. Recall that,

   $$s_{\hat{\beta_1}} = \frac{s}{\sqrt{S_{xx}}}.$$

   Also note that the estimate s of standard deviation for σ,

   $$s = \sqrt{\frac{SSE}{n-1}}.$$

Using our given data we can calculate $s_{\hat{\beta}_1}$,

$$s_{\hat{\beta}_1} = \frac{1}{\sqrt{S_{xx}}}\sqrt{\frac{SSE}{n-1}} = 0.204664.$$

Note that the critical value $t_{.025,30} = 2.042$. Finally calculating our $95\%CI$,

$$95\%CI = (2.585 \pm 2.042(0.204664)) \approx (2.16, 3.01).$$

Interestingly even at 95% we are getting a relatively large interval, likely because our sample standard deviation is still relatively large.