STAT 515 hw 11

Simple linear regression, contingency tables

1. Use the commands below to read into R a comma-separated-values data file:

```
data.url <- url("https://people.stat.sc.edu/gregorkb/data/ParticleBoard.csv")
data <- read.csv(data.url)
```

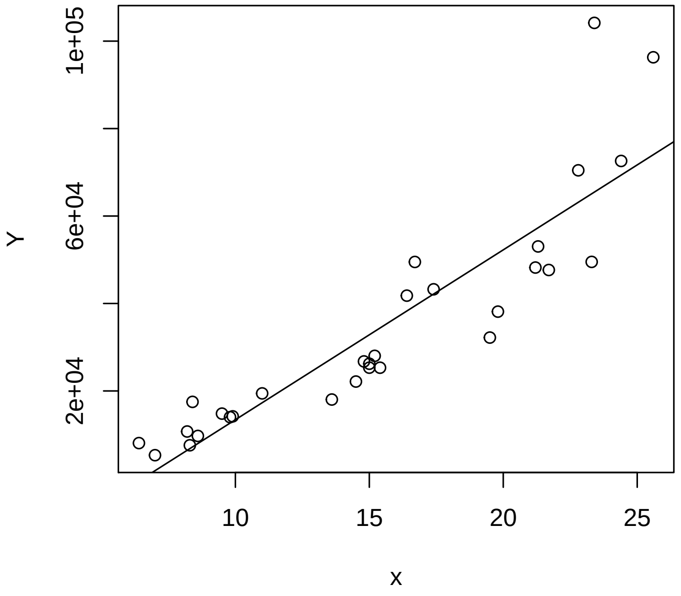
The data contain measurements of the stiffness (lbs per square inch) of particle boards produced at different densities (lbs per cubic foot) (Conners, 1979). Treat stiffness as Y and density as x.

(a) Make a scatterplot of the stiffness measurements versus the density measurements. Overlay the least-squares regression line.

The R commands

```
x <- data$Density
Y <- data$Stiffness
plot(Y~x)
abline(lm(Y~x))
```

produce the plot



(b) Write down the fitted least-squares regression model giving the estimated mean stiffness as a function of the density of the particle board.

We obtain from the R command

```
lm(Y~x)
```

the fitted model

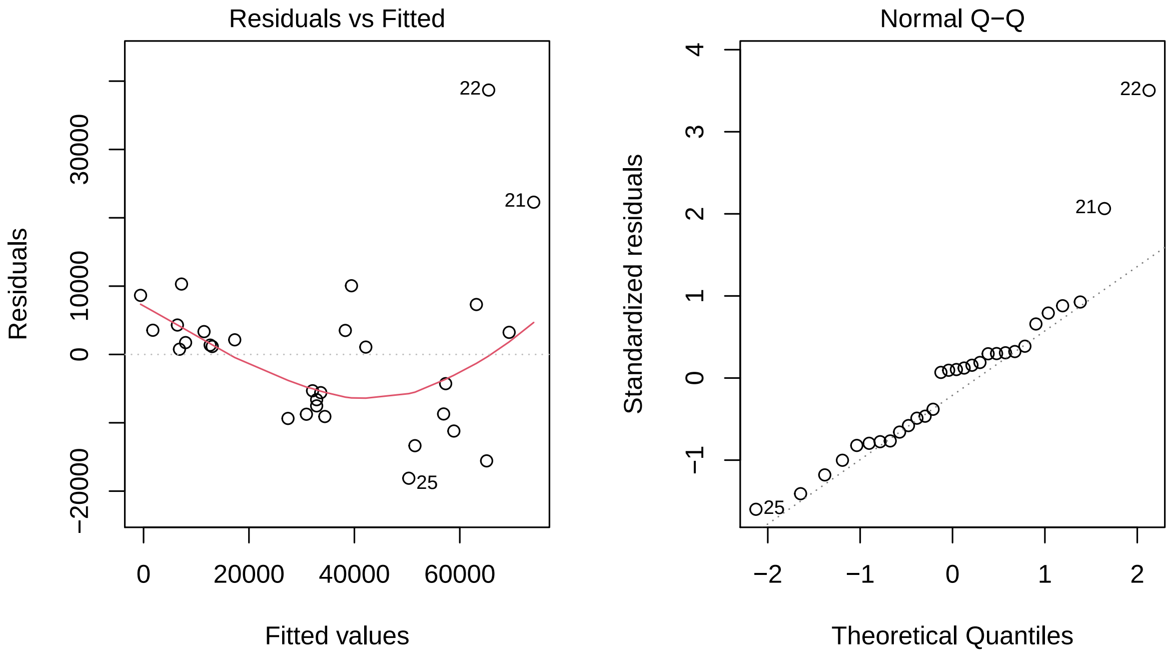
$$\texttt{Stiffness} = -25434 + 3885 \times \texttt{Density}$$

(c) Make a residuals-versus-fitted-values plot as well as a Normal quantile-quantile plot of the residuals. Comment on whether you think the assumptions of the simple linear regression model are satisfied for these data.

The R commands

```
par(mfrow = c(1,2))
plot(lm(Y~x), which = c(1,2))
```

produce the plots



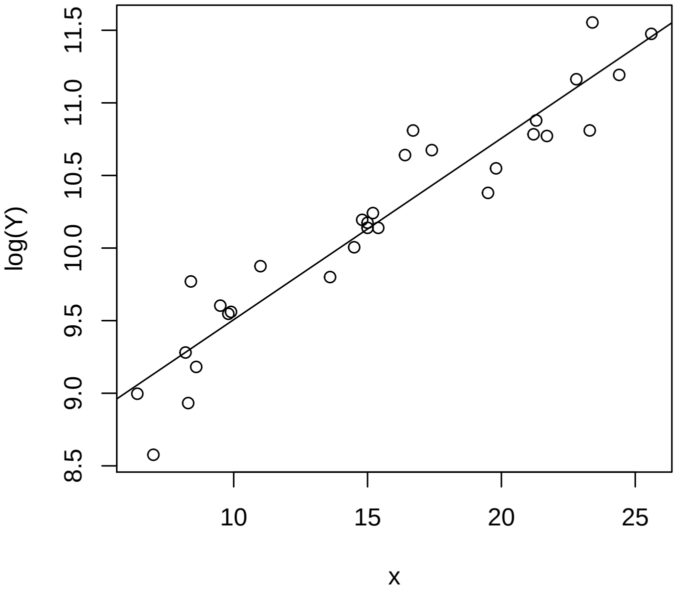
It appears that the relationship between the stiffness and the density of particle board may not be linear, as the points in the residuals-versus-fitted-values plot seems to follow a curve (which is indicated by the overlaid red line). We might suspect this from looking at the scatterplot of the stiffness versus the density values. The Normal quantile-quantile plot shows a couple of points quite distant from the straight line, suggesting that the residuals in this model may not follow a Normal distribution.

(d) If we discover that the relationship between our predictor and our response is nonlinear or if there is non-constant variance in the residuals, it is sometimes helpful to transform one or the other or both of the variables; we will try transforming the reponse variable. Make a scatterplot of the natural log of the stiffness measurements versus the density measurements. Overlay the least-squares regression line for the regression of the log-stiffness measurements on the density measurements.

The R commands

```
plot(log(Y)~x)
abline(lm(log(Y)~x))
```

produce the plot



(e) Write down the fitted least-squares regression model giving the estimated mean log-stiffness as a function of the density of the particle board.

We obtain from the R command

```
lm(log(Y)~x)
```

the fitted model

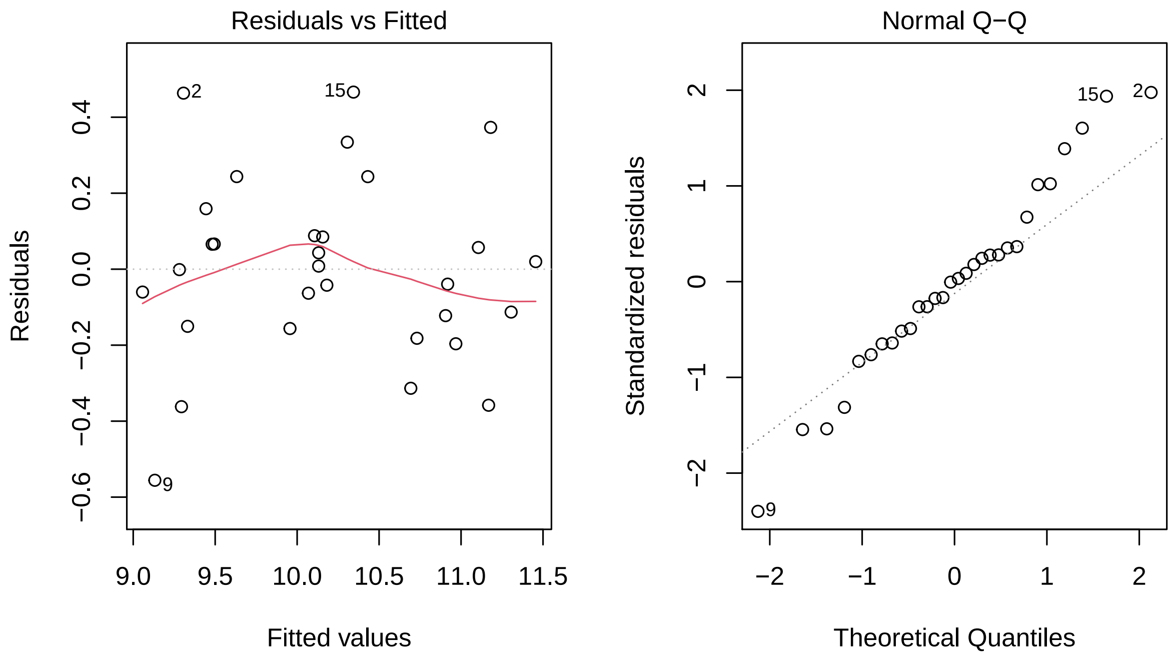
$$\texttt{log(Stiffness)} = 8.2574 + 0.1249 \times \texttt{Density}$$

(f) Make a residuals-versus-fitted-values plot as well as a Normal quantile-quantile plot of the residuals. Comment on whether you think the assumptions of the simple linear regression model are satisfied for these data—when the natural log of the stiffness measurements is used as the response.

The R commands

```
par(mfrow = c(1,2))
plot(lm(log(Y)~x), which = c(1,2))
```

produce the plots



The residuals-versus-fitted-values plot show a much more random scattering of the points, and the points Normal quantile-quantile plot of the residuals fall closer to the straight line. It seems that the assumptions of the simple linear regression model of the natural log of the stiffness measures versus the density measurements are satisfied.

(g) Give a 95% confidence interval for the value of the slope parameter in the log-stiffness versus density regression model.

The R command

```
confint(lm(log(Y)~x))
```

returns the 95% confidence interval $(0.109, 0.141)$ for the slope parameter.

(h) Give a 99% confidence interval for the mean log-stiffness of a pieces of particle board produced at a density of 10.5 pounds per cubic foot. Then exponentiate the lower and upper bounds to give the interval on the scale of the original stiffness measurements. *That is, having obtained the log-scale interval, say* (a, b)*, give the back-transformed interval* (e^a, e^b)*.*

The R command

```
predict(lm(log(Y)~x),
        newdata = list(x = 10.5),
        interval = "confidence",
        level = 0.99)
```

returns the 95% confidence interval $(9.406, 9.733)$ for the mean log-stiffness of particle board with density equal to 10.5. Back-transforming this interval from the log-scale back to the original scale by exponentiating both bounds gives the interval $(12156, 16859)$.

(i) Give a 99% prediction interval for the log-stiffness of a single piece of particle board produced at a density of 10.5 pounds per cubic foot. Then exponentiate the lower and upper bounds to give the interval on the scale of the original stiffness measurements.

The R command

```
predict(lm(log(Y)~x),
        newdata = list(x = 10.5),
        interval = "prediction",
        level = 0.99)
```

returns the 95% prediction interval $(8.873, 10.265)$ for the mean log-stiffness of particle board with density equal to 10.5. Back-transforming this interval from the log-scale back to the original scale by exponentiating both bounds gives the interval $(7135, 28721)$.

(j) Report the value of the coefficient of determination R^2 from the regression of the log-stiffness measurements on the density measurements. Give an interpretation of the value of R^2.

The out from the R command

```
summary(lm(log(Y)~x))
```

gives $R^2 = 0.9016$. This means that about 90% of the variability in the log-stiffness of the particle board is explained by the density of the particle board.

(k) Suppose you were asked to predict the stiffness of particle boards produced at a density of 30 lbs per cubic foot based on these data. How would you respond?

The particle boards in the study had densities between 6.4 and 25.6 lbs per cubic foot; although the model for the log-stiffness versus the density of particle boards appears to fit the data in this range, we cannot be certain that the model holds for densities outside of this range. Predicting the stiffness of particle boards produced at a density of 30 lbs per cubic foot based on these data would be extrapolation, which we should be very cautious about.

2. Rick and Jane Wilson of San Antonio, TX played many games of Qwirkle, recording each time which player got the first “qwirkle” and which player won or whether it was a “tie”. They sent the data to their son-in-law (your humble instructor) for analysis, with the question in mind: Does the person who gets the first qwirkle tend to win the game? Read the data, which has the “ties” removed, into R with the commands

```
data.url <- url("https://people.stat.sc.edu/gregorkb/data/RJQwirkle.csv")
data <- read.csv(data.url)
```

(a) Formulate the null and alternate hypotheses of interest to Rick and Jane Wilson. Ignore the one-sided nature of their research question and consider whether the player who gets the first qwirkle could be more or possibly less likely to win the game.

The null and alternative hypotheses of interest are

H_0: There is no association between getting the first qwirkle and winning.

H_0: There is an association between getting the first qwirkle and winning.

(b) Using the data, construct the relevant contingency table.

From the data we obtain the 2×2 table of counts

	Jane won	Rick won	
Jane 1st Qwirkle	23	18	41
Rick 1st Qwirkle	11	16	27
	34	34	68

(c) Give the table containing the counts expected under the null hypothesis.

The 2×2 table of the counts we would expect if there were no association between getting the first qwirkle and winning is

	Jane won	Rick won	
Jane 1st Qwirkle	20.5	20.5	41
Rick 1st Qwirkle	13.5	13.5	27
	34	34	68

(d) Give the value of the test statistic for Pearson’s chi-squared test of association.

We have

$$W_{\text{test}} = \frac{(23-20.5)^2}{20.5} + \frac{(18-20.5)^2}{20.5} + \frac{(11-13.5)^2}{13.5} + \frac{(16-13.5)^2}{13.5} = 1.535682.$$

(e) Give the p-value for testing the null hypothesis based on these data.

The p-value is given by $P(W > W_{\text{test}})$, where $W \sim \chi_1^2$, where the degrees of freedom of the relevant chi-squared distribution is 1, since the table has dimension 2×2. We have

$$P(W > W_{\text{test}}) = \texttt{1 - pchisq(1.535682)} = 0.215.$$

(f) Summarize your results with respect to the research question of your humble instructor's in-laws.

Dear Rick and Jane Wilson,

According to my analysis of the Qwirkle data you so painstakingly collected, I am unable to find a significant association between getting the first "qwirkle" and winning the game. In spite of the disappointment you may feel at this result, I wish you many more fun games of Qwirkle in the future. Who knows—if you keep on collecting data, the sample size might one day be large enough for an association to become detectable!

Kind regards and happy Qwirkling.

(g) Suppose Rick and Jane Wilson go on collecting data until they have played three times as many games. Moreover, suppose they go on winning and/or getting the first qwirkle the same proportion of times as they have so far. Give the p-value they would obtain based on their three-times-larger data set. *Hint: Just multiply all your counts by three and repeat the analysis.* Explain why multiplying the counts by three would change the p-value.

With the R commands

```
table(data)
chisq.test(3*table(data),correct = FALSE)
```

we obtain the p-value 0.03184, so that we could reject the null hypothesis at any significance level α greater than 0.03184.

References

Conners, T. E. (1979). *Investigation of Certain Mechanical Properties of a Wood-foam Composite*. PhD thesis, University of Massachusetts.