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.

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

(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.

(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.

(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.

(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.

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

(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).

(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.

(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.

(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?

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.

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

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

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

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

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

   (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.

References

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