STAT 3011 **Lab 10 Solution** (April 30–May 2, 2019) **Spring 2019**

Example 1: (CANCELLED)

Are the strengths of the Minnesota tornadoes dependent on the locations of occurrence? In this problem, we examine the following measurements for each of the 1,363 tornadoes that had made touchdowns in Minnesota between 1950 and 2006:

- `Strength`: The Fujita scale of the tornado (F0, F1, F2, and F3+.)
- `Location`: The region of Minnesota in which the touchdown occurred (NW = Northwest, N&NE = North Central and Northeast, W = West Central, C = Central, E = East Central, SW = Southwest, S = South Central, and SE = Southeast.)

The joint distribution of `Strength` and `Location` is summarized in the following contingency table:

		Strength				**Total**
		F0	**F1**	**F2**	**F3+**	
Location	**NW**	136	65	24	4	229
	N&NE	52	38	8	8	106
	W	86	64	26	7	183
	C	128	76	30	9	243
	E	54	42	17	14	127
	SW	62	59	20	9	150
	S	90	61	31	9	191
	SE	77	36	18	3	134
Total		685	441	174	63	1,363

(Source: NOAA/NWS Storm Prediction Center)
Use the following code to load the above contingency table into R:

```
tornado_table <- read.table("http://users.stat.umn.edu/~wuxxx725/tornadoes.txt",
                            header = T, check.names = F)
```

a. State the explanatory and response variables.

   Explanatory variable: Location of the tornado

   Response variable: Strength of the tornado

b. Use the following R codes to conduct the five-step hypothesis test for the association between the `Strength`s and `Location`s of Minnesota tornadoes at the significance level $\alpha = 0.05$.

```
> mytest <- chisq.test(tornado_table)
Warning message:
In chisq.test(tornado_table) : Chi-squared approximation may be incorrect
> mytest

        Pearson's Chi-squared test

data:  tornado_table
X-squared = 41.403, df = 21, p-value = 0.004998

> mytest$expected
            F0       F1       F2       F3+
NW   115.08804 74.09318 29.23404 10.584740
```

```
N&NE  53.27219 34.29640 13.53191  4.899486
W     91.96992 59.20983 23.36170  8.458547
C    122.12399 78.62289 31.02128 11.231842
E     63.82612 41.09098 16.21277  5.870139
SW    75.38518 48.53265 19.14894  6.933236
S     95.99046 61.79824 24.38298  8.828320
SE    67.34409 43.35583 17.10638  6.193690
```

i. Assumptions:

- Random sample
- All expected cell counts are at least 5. (Note: One of the cells has an expected cell count of 4.90. Since the sample size assumption is only slightly violated, we continue to proceed with the hypothesis test.)

ii. Hypotheses:

- $\mathrm{H_0}$: The locations and strengths of Minnesota tornadoes are independent.
- $\mathrm{H_a}$: The locations and strengths of Minnesota tornadoes are associated.

iii. Test statistic: $\chi^2 = 41.403$

iv. p-value: 0.004998

v. Conclusion and interpretation: Reject $\mathrm{H_0}$. There is a significant association between the locations and strengths of Minnesota tornadoes at the significance level $\alpha = 0.05$.

c. Use the `pchisq()` function to reproduce the p-value from the test statistic you obtained in part b).

```
> pchisq(41.403, df = 21, lower = F)
[1] 0.004997227
```

d. Find the estimated risks for a tornado to be F2 or higher in East Central Minnesota (E) and in North Central & Northeast Minnesota (N&NE), respectively.

East Central Minnesota (E):

$$\frac{17+14}{127} = \frac{31}{127} = 0.244.$$

The risk of a tornado to be F2 or higher in East Central Minnesota is 0.244.

North Central & Northeast Minnesota (N&NE):

$$\frac{8+8}{106} = \frac{16}{106} = 0.151.$$

The risk of a tornado to be F2 or higher in North Central & Northeast Minnesota is 0.151.

e. Calculate and interpret the relative risk for a tornado to be F2 or higher in East Central Minnesota (E) vs in North central & Northeast Minnesota (N&NE).

$$\frac{0.244}{0.151} = 1.62$$

Based on the sample, we estimate that a tornado in East Central Minnesota is 1.62 times as likely to have a strength of F2 or higher as a tornado in North Central & Northeast Minnesota does.

Example 2:

The R dataset `trees` in the `datasets` package contains the following measurements of 31 black cherry trees:

- `Girth`: Diameter of the tree at the height of 4 ft 6 in, measured in inches
- `Volume`: Lumber volume, measured in ft^3

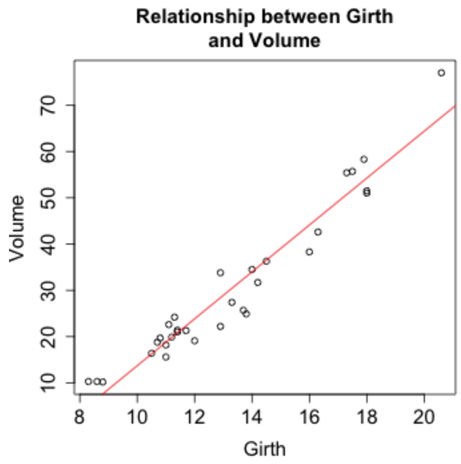
Please use the following command to load the dataset:

```
attach(trees)
```

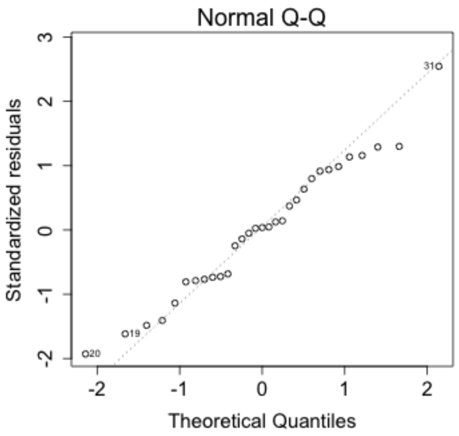
If the dataset fails to load, please copy and paste the following R codes to load the data manually:

```
Girth <- c( 8.3,  8.6,  8.8, 10.5, 10.7, 10.8, 11.0, 11.0, 11.1, 11.2,
           11.3, 11.4, 11.4, 11.7, 12.0, 12.9, 12.9, 13.3, 13.7, 13.8,
           14.0, 14.2, 14.5, 16.0, 16.3, 17.3, 17.5, 17.9, 18.0, 18.0, 20.6)
Volume <- c(10.3, 10.3, 10.2, 16.4, 18.8, 19.7, 15.6, 18.2, 22.6, 19.9,
            24.2, 21.0, 21.4, 21.3, 19.1, 22.2, 33.8, 27.4, 25.7, 24.9,
            34.5, 31.7, 36.3, 38.3, 42.6, 55.4, 55.7, 58.3, 51.5, 51.0, 77.0)
```

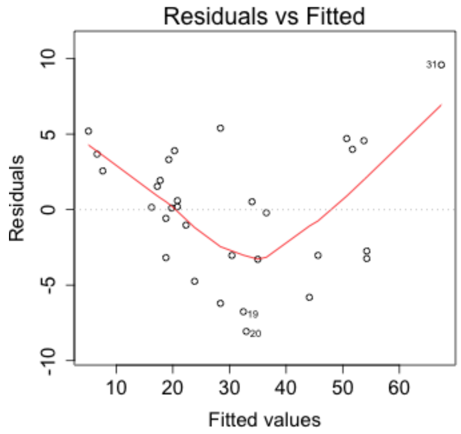
In this problem, we consider the linear regression model for `Volume` on `Girth`. The scatterplot for `Volume` vs `Girth`, the normal Q-Q plot for the errors, and the residual plot are given below.



Relationship between Girth and Volume



Normal Q-Q



Residuals vs Fitted

Please answer the following questions:

a. Fit a linear regression model for `Volume` on `Girth` and obtain its summary using the `lm()` and `summary()` functions.

```
> m1 <- lm(Volume ~ Girth, data = trees)
> summary(m1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max
-8.065 -3.107  0.152  3.495  9.587

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,       Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

b. State and interpret the value of r^2 from the model summary output in part a).

$r^2 = 0.9353$, which means that 93.53% of the variation in lumber volume is explained by its linear relationship with girth.

c. Calculate the correlation r between `Girth` and `Volume`, and state the strength and the direction of the correlation.

$r = \text{sign}(b)\sqrt{r^2} = 1\sqrt{0.9353} = 0.9671$. There is a strong, positive correlation between `Girth` and `Volume`.

d. State the estimated regression equation in the form $\hat{Volume} = a + b(Girth)$.

$$\hat{Volume} = -36.9435 + 5.0659(Girth).$$

e. Interpret the slope b.

For each inch of increase in a black cherry tree's girth, we expect its lumber volume to increase by an average of 5.0659 cubic feet.

f. Explain why it does not make sense to interpret the intercept a.

The intercept refers to the expected lumber volume of a black cherry tree with a girth of 0 inches. Since the girth of a black cherry tree cannot be 0 inches, it does not make sense to interpret the intercept.

g. Tree #15 has a `Girth` of 12 inches. Predict its lumber `Volume` using the estimated regression equation in part d).

$$\hat{Volume} = -36.9435 + 5.0659(Girth) = -36.9435 + (5.0659)(12) = 23.8473.$$

The predicted lumber volume for tree #15 is 23.8473 cubic feet.

h. The actual lumber volume of tree #15 is 19.1 cubic feet. Find the residual for tree #15.

i. Is it appropriate to use the estimated regression equation in part d) to predict the lumber volume of a black cherry tree with a girth of 25 inches? If so, give the estimated lumber volume. If not, please explain the reason.

No. A girth of 25 inches would be much greater than all the observed data. It is not appropriate to extrapolate far outside the observed range of data.

j. How would the correlation r change if `Girth` were given in centimeters and `Volume` were given in liters? Please explain. (1 in = 2.54 cm; 1 ft^3 = 28.3 L)

Changing the units of the variables does not affect the correlation r, thus r remains at 0.9671.

k. Conduct a five-step hypothesis test on whether the true population slope β is different from 0.

i. Assumptions:
   - Random sample
   - Linear trend between `Girth` and `Volume`
   - Normal conditional distribution for `Volume` at each value of `Girth`
   - Constant standard deviation for `Volume` at each value of `Girth`

ii. Hypotheses:
   - H_0: $\beta = 0$
   - H_a: $\beta \neq 0$

iii. Test statistic: $t = 20.48$

iv. p-value: $< 2 \times 10^{-16}$

v. Conclusion and interpretation: Reject H_0. There is strong evidence that an association exists between the girth and the lumber volume of a black cherry tree at the significance level $\alpha = 0.05$.

l. Check the linearity, normal error, and constant variance assumptions using the diagnostic plots.

- Linearity: From the scatterplot of `Volume` against `Girth`, a linear relationship is appropriate. (Note: It is arguable that a quadratic relationship might be more appropriate based on the curved pattern in the residual plot.)
- Normal error: From the normal Q-Q plot, the residuals are approximately normally distributed.
- Constant variance: From the residual plot, the variation of the residuals is approximately the same for all fitted values.