Assignment 5

This assignment will give experience with power calculations and linear regression interpretation and calculations.

1. Ulcer recurrence (yes or no) was compared in two group of males randomized to conventional treatment (CT) versus conventional treatment plus antibiotics (CT+A) to eliminate the Heliobactor pylori organism. The results are below.

Treatment	n	num recur	percent recur	SE
CT	10	6	60%	15.5%
CT+A	10	4	40%	15.5%

1a. The difference in the two percentages is 20%. What is the standard error of this difference (SE_d)?

1b. Compute the Z statistic and corresponding p value for testing the null hypothesis that the true difference in the proportions is zero ($\pi_1 - \pi_2 = 0$ or equivalently $\pi_1 = \pi_2$). You may use the Gaussian table (or the EXCEL 'normsdist' function) to look up the p value.

1c. Is the difference statistically significant using the two sided $p < 0.05$ criterion?

1d. What is the power for detecting a 20% difference based on the above? Show your calculations or briefly explain why no calculations are needed. Are the p value results surprising in light of the power level?

1e. Using a table in Section V of the notes, determine the sample size needed for 80% power with a two sided alpha = 0.05. Is this consistent with the power calculation for the sample size of 10 per group?

2. Investigators are interested in the relation between height in cm (X), versus forced expiratory volume in one minute (FEV1, the Y variable) in liters in adult males and females age 25 or older. The mean FEV1 is 3.9 L with standard deviation for FEV1 (SD_y) of 0.5 liters. The mean height is 161 cm. They report the following equation obtained from a linear regression analysis. All terms in the equation are significant at $p < 0.05$. The variable "gender" is coded 0 for females and 1 for males.

Predicted FEV1 = $\hat{Y}$ = -9 + 0.08 height + 0.3 gender + 0.02 height x gender

They report the R square value as $R^2 = 0.60$.

2a. Assuming the equation is correct, is the relation between FEV1 and height parallel in males versus females? Within each gender, is the relation given by the equation between FEV1 and height linear?

2b. What proportion of the variation in FEV1 is accounted for by height and gender according to this equation?

2c. According to the equation, how much will FEV1 increase on average for a 10 cm increase in height?

2d. The intercept is -9.0 liters. Since FEV1 cannot be negative, is this a mistake or artifact? Briefly explain.

2e. To be clinically useful, the prediction ($\hat{Y}$) must be within 0.1 liter of the true value (Y) at least 95% of the time. Has this goal been met?

2f. If the investigators wish the prediction to be within 0.1 liter of the true value, circle all that are true

i. The R^2 must increase
ii The R^2 must decrease
iii The SD_e must increase
iv The SD_e must decrease
v. Cannot answer without more information

2g. If gender is removed from the equation and an equation relating FEV1 to height is computed using the same data, will the regression coefficient (b) for height (b=0.08) guaranteed to remain 0.08? Why or why not? Briefly explain.

3 Read the attached article on Testosterone vs Placebo treatment in older men (JAMA, 2 Jan 2008, Emmelot-Vonk et al.)

3a. What is the study design ?

3b. What test was used to generate the p values in Tables 2, 3 (excluding metabolic syndrome section) and 4

Is this the correct test? Briefly explain.

3c. A critic says that the results in Tables 2, 3 and 4 may be an artifact of multiple testing. Briefly comment as to whether this is a valid concern. Why or why not?

JMP assignment 5

A study was conducted on n=97 non smoking males to examine factors that may influence their semen volume (svol) in ml. The investigators were primarily interested in age (in years), but also considered alcohol use (1=yes or 0=no), whether the subject had abstained from sexual activity for 2 days or more (astn2d-1=yes or 0=no) and whether the subject had hypertension (htn-1=yes or 0=no).

Use the "svol.xls" dataset to determine how age, alcohol use, abstinence and hypertension are simultaneously related to semen volume. You need only consider interactions (if any) with age. Report on which of the variables are related to semen volume and the magnitude and direction of their effect. Be sure to report the final regression equation and its R square value and its interpretation.

Notes for JMP: If a 0, 1 coded variable is declared to be "nominal", it is internally recoded using effect coding, -1, 1. **The 0 is recoded to 1 and the 1 is recoded to -1!!** So, declare the variable to be "continuous" for the purpose of regression analysis even if it is not.

If the "cross" button in the "fit model" panel is used to make an interaction variable between two continuous variables, JMP will first subtract the mean from each variable. Thus for the interaction of X_1 with X_2, JMP creates $(X_1-M_1)(X_2-M_2)$ where M_1 is the mean for X_1 and M_2 is the mean for X_2.

Note that: $(X_1-M_1)(X_2-M_2) = X_1X_2 + X_1M_2 + X_2M_1 + M_1M_2$

To get around this, one can either **turn off the "center polynomial**" option in the fit model panel or "manually" make a new variable under the "cols" tab by multiplying X_1 times X_2. That is, in the formula panel make $X_3 = X_1 \times X_2$ and put X_3 in the model. X_3 corresponds to the first term only in the expression above. Using X_3 may make the interaction easier to interpret.