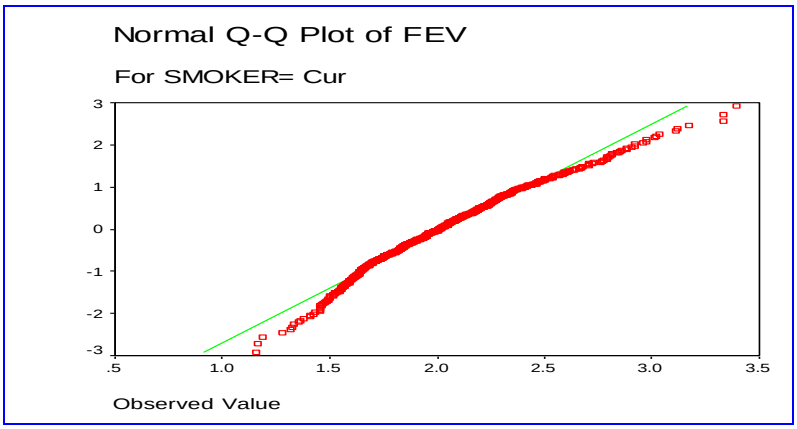
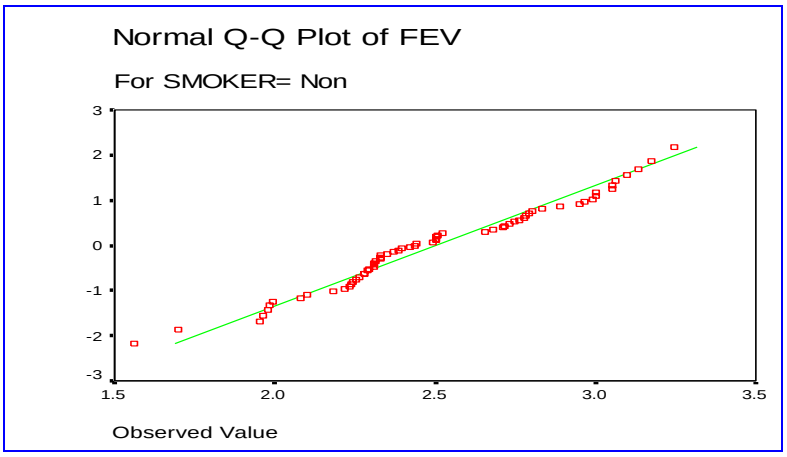
# SOLUTION TO THE LAB ASSIGNMENT 3

1)

(a) In the study a sample of 654 children was obtained and age, height, FEV, gender and smoking status of their parents was recorded for each subject. No random mechanism was used to assign the children to the two parent categories (smoking or nonsmoking). Therefore, the study is an example of an observational study. Note that statistical inference of cause-and-effect relationship (or casual inference) can be drawn from randomized experiment, but not from an observational study. Therefore we cannot assess the effect of second-hand smoking on the subjects' FEV readings. Many factors (or confounding variables) may be responsible for any observed differences in the FEV readings. In other words, the observed differences in FEV readings (if any) may be due to other variables. Some of the confounding variables in this study are: the level of pollution in the neighborhood the child lives in, lifestyle (participation in sports), race, and the time period the child was exposed to second-hand smoke. It is clear that the effect of exposure to second-hand smoke may have a cumulative effect over a long time period. As the sample obtained is not random, the results cannot be generalized to any well defined population.

(b) The Q-Q plots of data for smoker and nonsmoker parents are given below:

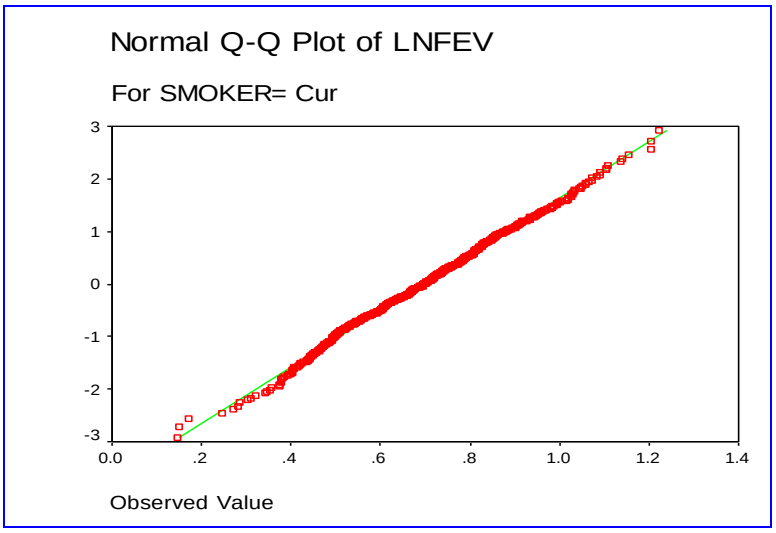


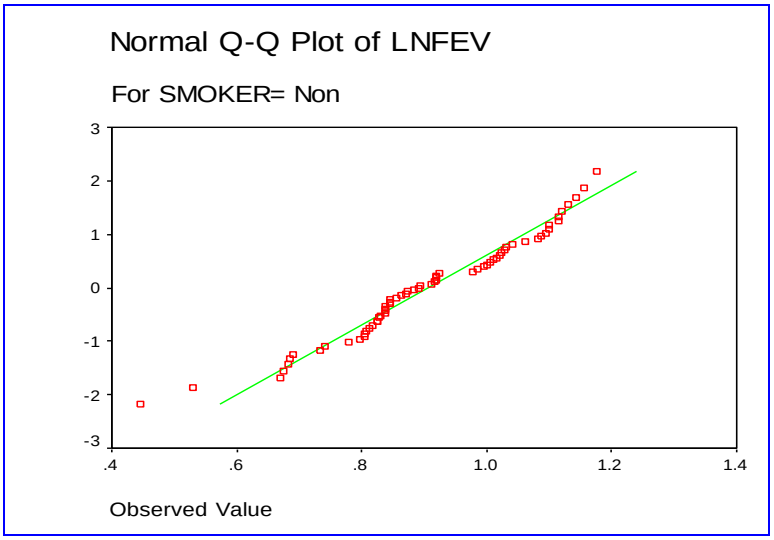




The Q-Q plot for non-smoker parents does not provide any evidence that the data are non-normal: there are no substantial and systematic departures of the points in the plot from the straight-line pattern. The two smallest observations separated from the other observations are outliers.

The points in the Q-Q plot for smoker parents do not follow a straight-line pattern; the concave curvature indicates that the data are right skewed. The log-transformation may be helpful to make the data approximately normal.

(c) The Q-Q plots of LNFEV for smoker and nonsmoker parents are given below:




For smoker parents, all points in the Q-Q plot for the log-transformed observations lie on a straight line or are reasonably close to the line. Therefore, the natural logarithm transformation was effective in making the data approximately normal. There is no evidence of any violation of the normality assumption.

For non-smoker parents, there are some minor deviations of the points from straight-line pattern; but they are not serious enough to question the normality assumption for the data. Notice that the log-transformation was not very effective to push the two outliers closer towards the body of the data.

(d) The Independent sample test applied to LNFEV produced the following output:

**Group Statistics**

| | SMOKER | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| LNFEV | Cur | 589 | .6946 | .18610 | .00767 |
| | Non | 65 | .9065 | .15386 | .01908 |

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| LNFEV | Equal variances assumed | 3.392 | .066 | -8.853 | 652 | .000 | -.2120 | .02394 | -.25897 | -.16495 |
| | Equal variances not assumed | | | -10.306 | 86.090 | .000 | -.2120 | .02057 | -.25285 | -.17108 |

Since the p-value for Levene's test for equality of variance is 0.066, the test assuming equal variances should be applied to the data. Moreover, notice that the ratio of the sample standard deviations for the two groups is less than two which also supports the assumption of equal variances. We define the null and alternative hypotheses as follows:

$$H_0: \mu_{SMOKER} - \mu_{NONSMOKER} = 0 \text{ versus } H_A: \mu_{SMOKER} - \mu_{NONSMOKER} < 0.$$

The value of the test statistic is –8.853 and the p-value of the one-sided test is 0.000/2 = 0.000. Therefore, we strongly reject the null hypothesis. There is strong evidence that the mean of LNFEV of children of smoker parents is less than that of children of nonsmoker parent(s). Equivalently, there is strong evidence that the mean FEV of children of smoker parents is smaller than the mean FEV of children of non-smoker parents (the logarithm is a non-decreasing function).

(e) The related SPSS output is displayed below:

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | .907 | .023 | | 39.898 | .000 |
| | Z | -.212 | .024 | -.328 | -8.853 | .000 |

a. Dependent Variable: LNFEV

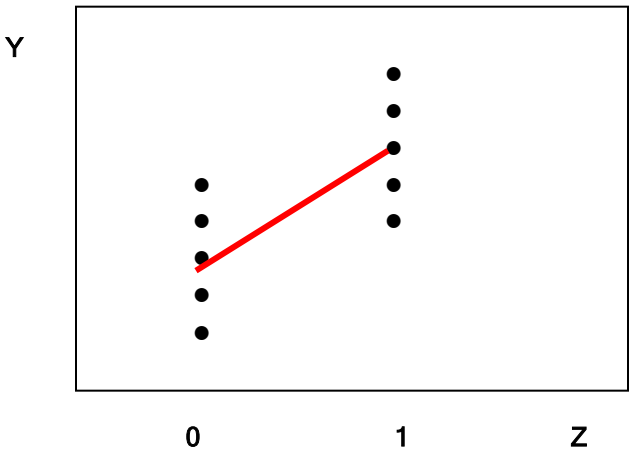
i) The equation of the least squares regression line is: $\hat{Y} = 0.907 - 0.212*Z$. Based on the results in part d., the mean of LNFEV for nonsmoker is 0.907 which exactly is the same as the value of intercept in the least square regression model. The mean of LNFEV for smoker parent(s) is 0.695 which equal to 0.907 -0.212= 0.695, the value of the intercept plus the value of the slope of the estimated regression line.

ii) The null and null hypotheses are: $H_0$: $\beta_1 = 0$ and $H_0$: $\beta_1 < 0$ and. The value of the test statistic for testing the slope of population regression line equal zero is -8.853 and this is equal to the value of the test statistic in part d. The p-value is 0.000/2=0.000 and it is equal to the p-value obtained in part d. The values of the test statistic and p-values are identical for the two tests.

iii) First we will explain the observed relationship in part i. Notice that the simple linear regression model has the following form: $\mu_Y = \beta_0 + \beta_1 * Z$, where Y is LNFEV. If we substitute Z = 0 (nonsmoker parents), we have $\mu_{NONSMOKER} = \beta_0 + \beta_1 * 0 = \beta_0$. On the other hand, if we substitute Z =1 (smoker parents), we have $\mu_{SMOKER} = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1$. Since the estimate of population mean is sample mean, the estimate of intercept is the sample mean of FEV for nonsmoker parents. Similarly, the estimate of intercept plus the estimate of slope of the regression line is sample mean of FEV for smoker parent(s).

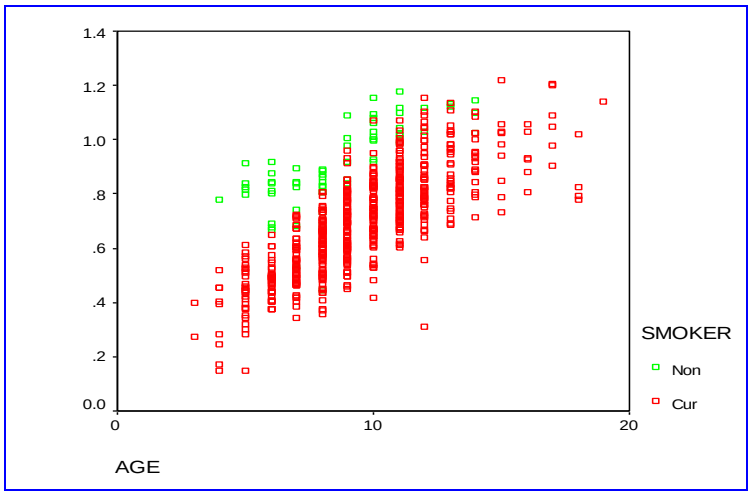
In terms of the above simple regression model, $\mu_{SMOKER} = \beta_0 + \beta_1$ and $\mu_{NONSMOKER} = \beta_0$. Therefore, testing the null hypothesis $H_0$: $\mu_{SMOKER} - \mu_{NONSMOKER} = 0$ is equivalent to testing the null hypothesis $H_0$: $\beta_1 = 0$. In general, the t-test for comparing two population means using equal variances assumption is equivalent to a simple linear regression model when the explanatory variable as an indicator variable. The corresponding t statistics must be equal in magnitude (the signs will be the same if the t statistic is calculated by subtracting the mean of group 0 from the mean of group 1).

Moreover, the assumptions of normality and equal variances in the t test are equivalent to the assumptions of normality and equal variance in the simple linear regression model. Indeed, the linear regression model says that data are normally distributed about the regression line with constant standard deviation σ. On the other hand, the predictor variable Z takes on only two values: 0 or 1. Therefore, there are only two locations along the regression line where there are data (0 and 1 on the Z axis). Thus constant spread and normality about the regression line is equivalent to the assumptions of normality and equal variances for the two populations.



2)

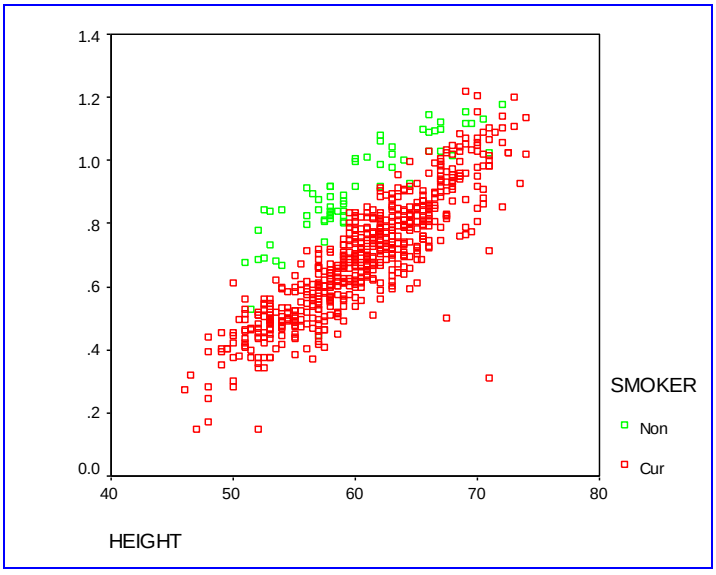
(a) The scatter plot of LNFEV versus Age with different markers for nonsmoker and smoker parents is given below:




For smoker parents, it is clear that there is strong positive linear relationship between the mean of LNFEV and Age. The values of LNFEV spread approximately equally among values of Age: there is no evidence of any departure from the equality of variances assumption. There is one outlier. Notice that the scatter flattens out for the higher values of AGE; there is no or very little change in LNFEV for older children subjected to second-hand smoking.

The scatter for nonsmoker parent(s) shows a strong positive relationship between LNFEV and Age. There is no obvious evidence of any serious departure from the equality of variances assumption, since the values of LNEFV spread approximately equally among values of Age. Notice smaller spread of LNFEV values over the AGE range compared to the range for smoker parents. Note that the range of values for Age for nonsmoker parent(s) is smaller than the range of data for Age for smoker parents. There are no outliers in this group.

(b) The scatter plot of LNFEV versus Height with different markers for nonsmoker and smoker parents is given below:

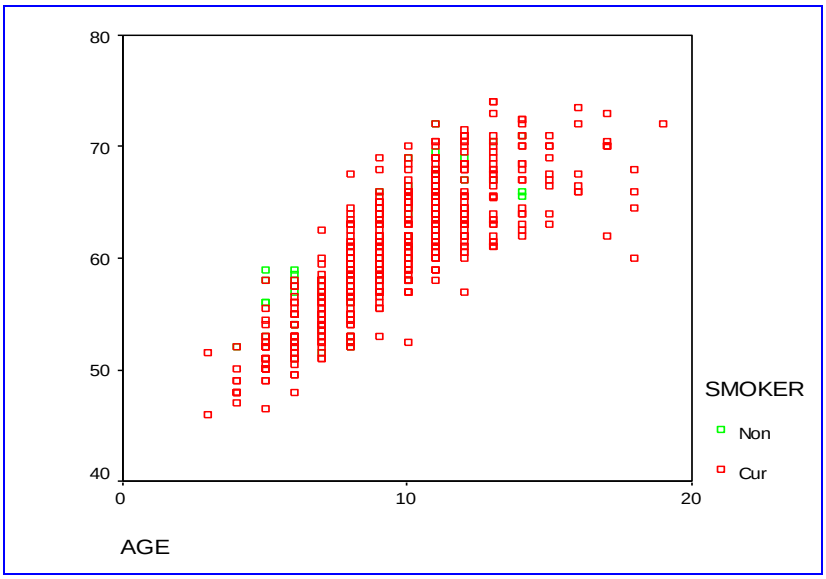


For smoker parents, there is a strong positive linear relationship between LNFEV and Height. There is one outlier. Except the outlier, the values of LNFEV spread approximately equally for all values of Height. Therefore, the scatterplot supports the assumption of equal variances.

For nonsmoker parents, there is strong positive linear relationship between LNFEV and height. In addition, there is no evidence of violation of equal variances assumption, since the values of LNFEV spread approximately equally along the possible values of Height.

The plot clearly shows that LNFEV values tend to be higher for nonsmoker parents. Note larger differences in LNFEV values between the two parent groups for younger children. Overall, the observations for the two groups form two separate and approximately parallel bands; it indicates that two separate regression lines (one for each parent group) are necessary to fit the data.

c) The scatter plot of Height versus Age with different markers for nonsmoker and smoker parents is given below:




According to the above scatterplot, there is a strong approximately linear relationship between Age and Height. The scatter seems to flatten out for higher values of AGE; it reflects the fact that the variable height increases only up to a specific age and then remains the same (approximately the same) for the rest of life for the subject. There is no difference between the height of children of nonsmoker parents and the height of children of smoker parent(s). If we have to choose one explanatory variable, either Age or Height, it seems that we should choose Height. Indeed, a quick glance at the two scatterplots obtained above shows that Height has a stronger linear relationship with LNFEV.

3)

(a) The following table gives the Pearson's correlation coefficients. The Pearson's correlation coefficient between LNFEV and Age is 0.703. Therefore, there is a positive linear relationship between the two variables. The Pearson's correlation coefficient between LNFEV and Height is 0.821, and hence there is a strong positive linear relationship between LNFEV and Height. The relationship between LNFEV and Height is stronger than the relationship between LNFEV and Age. The Pearson's correlation coefficient between Age and Height is 0.806. Hence, Age and Height are strongly correlated. If we have to choose between Height and Age, we should choose Height since its correlation coefficient with LNFEV is

higher. Therefore, the coefficients confirm our conclusion in Question2.

**Correlations**

| | | LNFEV | HEIGHT | AGE |
|---|---|---|---|---|
| LNFEV | Pearson Correlation | 1.000 | .821** | .703** |
| | Sig. (2-tailed) | . | .000 | .000 |
| | N | 654 | 654 | 654 |
| HEIGHT | Pearson Correlation | .821** | 1.000 | .806** |
| | Sig. (2-tailed) | .000 | . | .000 |
| | N | 654 | 654 | 654 |
| AGE | Pearson Correlation | .703** | .806** | 1.000 |
| | Sig. (2-tailed) | .000 | .000 | . |
| | N | 654 | 654 | 654 |

**. Correlation is significant at the 0.01 level (2-tailed).

(b) The partial correlation between LNFEV and Age controlling for Height is 0.1221, while the partial correlation coefficient between LNFEV and Height controlling for Age is 0.6034. Therefore, including Height in the model will explain most of the variation in LNFEV. Based on the p-values for the partial correlations, we should use both Height and Age to predict LNFEV. Based on the magnitudes of the partial correlations, we should use Height to predict LNFEV if we have to choose between Age and Height in the model.

4)

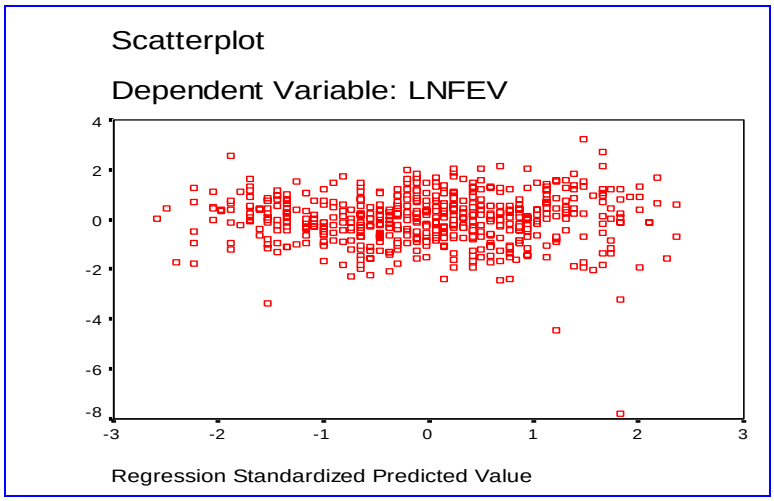
(a) The least-square regression lines for smoker and nonsmoker parents are:

$$\textbf{Smoker}: \ \mu\{LNFEV \mid Height\} = -1.062 + 0.02897 * Height$$

$$\textbf{Nonsmoker}: \ \mu\{LNFEV \mid Height\} = -0.562 + 0.02441 * Height$$

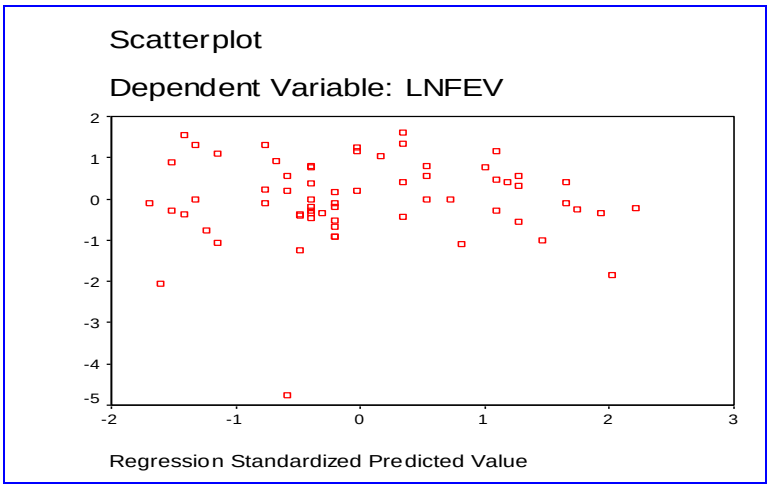
The value of the slope 0.02897 shows the estimated mean increase in LNFEV of children of smoker parents as Height increases by one inch. The value of the slope 0.02441 shows the estimated mean increase in LNFEV of children of nonsmoker parents as height increases by one inch.
For nonsmoker parents, the coefficient of determination for the regression model has $R^2 = 0.776$. Thus 77.6% of the variation in LNFEV is explained by Height. For smoker parents, the coefficient of determination is $R^2 = 0.728$ and therefore 72.8% of the variation in LNFEV is explained by Height.

(b) The plot of standardized residuals versus standardized predicted values for the smoker parents is displayed below:



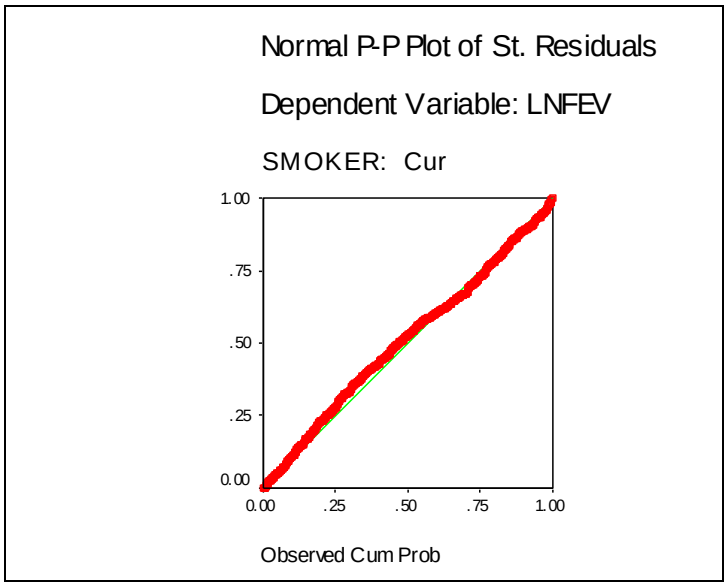
The above plot is used to check the assumption of equal standard deviations. The residuals in the plot are scattered randomly about a horizontal line at zero. The spread of residuals is approximately uniform over the range of standardized predicted value. Although the plot contains some outliers, there is no evidence of a serious violation of the assumption of equal standard deviations.

The plot of standardized residuals versus standardized predicted values for the nonsmoker parents is displayed below:

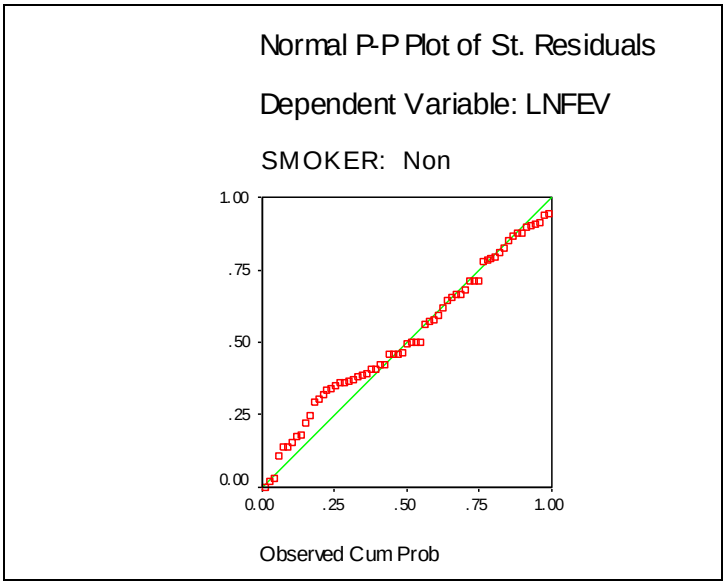



The residuals in the plot scattered randomly about a horizontal line at zero. However, the spread of residuals is not equal for all values of the standardized predicted values. This indicates that the assumption of equal standard deviations may be violated in this case.

(c) The P-P plot of residuals for smoker parents is displayed below. The plot clearly supports the assumption of normality for the residuals.



The P-P plot for nonsmoker parents shows some departure from the straight line pattern and indicates that the assumption of normality may be violated. The consequences of violating this assumption may be minor as the estimates of the coefficients and their standard errors are robust to non-normal distributions.




5)

a) The least-squares regression line has the following equation:

$$\mu\{LNFEV \mid Height, Z\} = -0.811 + 0.02856 * Height - 0.227 * Z$$

According to the SPSS output, the above multiple regression model explains 79.6 % of the variation in LNFEV.

b)

This if we substitute Z=0, we get

**Nonsmoker**: $\mu\{LNFEV \mid Height\} = -0.811 + 0.02856 * Height$

If we substitute Z=1, we get

**Smoker**: $\mu\{LNFEV \mid Height\} = -1.308 + 0.02856 * Height$

The above lines are different from the fitted lines obtained in Question 4 part b. Note that the above regression model implies that the two fitted regression lines for the two parent groups are parallel (since they have the same slope). On the other hand, the scatter plot obtained in Question 2 b indicated that the two bands corresponding to the two parent groups can be modeled by straight lines with slightly different slopes (not parallel). Thus, a better fit can be obtained by incorporating the interaction term Height*Z in the model.

Therefore the separate simple linear regression models for the two parent groups discussed in Question 4 may provide a better fit to the data than the multiple regression model discussed in Question 5.

# ASSIGNMENT 3 MARKING SCHEMA

Proper Header: 10 points

## Question 1

a) Study design: 2 point
Casual inferences: 2 points
Confounding variables: 2 points

b) Normality plots: 3 points each (6 points total)
Comments about normality: 2 points

c) Normality plots for transformed data: 3 points each (6 points total)
Comments about normality: 2 points

d) Null and alternative hypotheses: 2 points
The test statistic with equal variances: 2 points
P-value (one-sided): 2 points
Conclusions: 2 points

e)

i. The equation of estimated regression line: 2 points
Comparing intercept and slope with means: 2 points

ii. The test statistics: 2 points
Comparing the test statistics: 2 point
Comparing the p-values: 2 point
Conclusions: 2 point

iii. Explanation of the relationship between the means of LNFEV and regression coefficients: 3 points
Explanation of the relationship between the two tests: 3 points
Equivalence of the assumptions for the two tests: 3 points

## Question 2

a) Scatterplot (properly formatted, different markers for the two groups): 4 points
Pattern in the scatterplot (linearity, direction strength) for each parent group: 3 points each (6 points total)
Outlier(s): 1 point

b) Scatter plot (properly formatted, different markers for the two groups): 4 points
Pattern in the scatterplot (linearity, direction strength) for each parent group: 3 points each (6 points total)
Outlier(s): 1 point

c) Scatter plot (properly formatted with different markers for the two groups): 4 point
Comments about the relationship: 2 points
Which explanatory variable is a better predictor: 2 point

## Question 3

a) Pearson's correlation coefficients: 2 points
Comments about sign and magnitude of the coefficients: 2 points

b) Partial correlation coefficients: 2 points
Which of the two predictors is better?: 2 points

## Question 4

a) The equation of estimated regression line for nonsmoker and smoker parents: 2 points
The equation of estimated regression line for nonsmoker and nonsmoker parents: 2 points

b) Interpretation of the slopes: 2 points
$R^2$ Value: 2 points

c) Plots of standardized residuals versus standardized predicted values: 2 points each (4 points total)
Pattern of residuals: 3 points
Assumption tested: 1 point
Outliers: 1 point

d) P-P plots of residuals: 2 points each (4 points total)
Description of the pattern in the plots: 3 points

## Question 5

a) The equation of estimated regression line: 2 points
Percentage of the variation explained: 2 points

b) Estimated regression lines for each parent group: 2 points each (4 points total)
Comparing regression models in Question 4 and Question 5: 4 points

# TOTAL = 135

Created by Alizera Simchi
Revised by Henryk Kolacz
September 2012