IGP 304 (Spring 2006) Homework 3 Keys:

1. The hospital data (see here for variable explanations) are a sample from a larger data set collected on people discharged from a selected Pennsylvania hospital as part of a retrospective chart review of antibiotic usage in hospitals.
    - Find the best-fitting linear relationship between ln(duration of hospitalization) and age.

A look at the histogram of `dur_stay` will tell you it is right-skewed and a transformation may be needed. Suppose we have defined a new variable `lndur` as the logarithm of `dur_stay` (Stata command `gen lndur=log(dur_stay)`).

If we ignore all the other variables, we can fit a regression line: lndur = $\beta_0 + \beta_1 \times$age (Stata command `regress lndur age`). The best-fitting line is:

lndur = 1.578 + 0.01 × age.

- Test for the significance of this relationship. State any underlying assumptions you have used.

The p-value for age effect is 0.085, which is marginally significant and thus we have weak evidence for association between age and ln(duration of hospitalization). The assumptions we made are: (i) there is a linear relationship between ln(duration of hospitalization) and age, (ii) for a given age, the ln(duration of hospitalization) is normally distributed, and (iii) the variance of the normal distributions are the same for different ages.

- What is R^2 for this regression?

$R^2 = 0.12$ is relatively small.

- Assess the goodness of fit of the regression line.

We can examine the goodness of fit by checking the residual plots and inverse normal plot for the residuals and carrying out Shapiro-Wilk test on the residuals. The Stata commands are:

```
regress lndur age
predict residuals, resid
qnorm residuals
swilk residuals
```

The results indicate no violation of the normality assumption. There might be a tendency for the variance to increase as age increases, but the pattern is not strong.

2. The birthweight-estriol data are from a study to relate birthweight to the estriol level of pregnant women. If you draw a scatter plot of birthweight versus estriol level, you will see a linear relationship between them, although this relationship is not consistent and considerable scatter exists throughout the plot.
    - How can this relationship be quantified?

To find the best fit of a linear relationship between birthweight and the mother's estriol level, we fit a simple linear regression using birthweight as the outcome (Stata command `regress birthweight estriol`). The fitted line is

birthweight = 21.52 + 0.608 × estriol

This tells us for every one more milligram/24 hour estriol in the mother, the baby's birthweight increases by 60.8 gram on average. Note that the unit for birthweight is 100 g and the unit for estriol level is mg/24 hr.

Note: Here, we assumed equal variance, which apparently is not a good assumption if you look at the scatter plot or the residual plot. A more appropriate analysis will be weighted regression (Stata command `regress birthweight estriol [aweight=estriol]`), which leads to a different fitted line: birthweight = 22.99 + 0.529 × estriol.

- What is the estimated average birthweight if a pregnant woman has an estriol level of 15 mg/24 hr?

If a pregnant woman has an estriol level of 15 mg/24 hr, the average birthweight is estimated to be 3064 grams (21.52 + 0.608 × 15 = 30.64).

If you have fitted a weighted regression, the answer is 3092 grams (22.99 + 0.529 × 15 = 30.92).

- Low birthweight is defined here as ≤2500 g. For what estriol level would the predicted birthweight be 2500 g?

Let a be the estriol level at which the predicted birthweight is 2500 g. Then $21.52 + 0.608a = 25$, which leads to $a = (25 - 21.52)/0.608 = 5.7$. When the mother's estriol level is 5.7 mg/24 hr, the predicted birthweight will be 2500 g.

If you have fitted a weighted regression, the answer is $a = (25 - 22.99)/0.529 = 3.8$ mg/24 hr.

- Interpret the slope of the regression line.

The slope is 0.608. Noting that the unit for birthweight is 100 g and the unit for estriol level is mg/24 hr, the interpretation of the slope is: for every one more milligram/24 hour estriol in the mother, the baby's birthweight increases by 60.8 gram on average.

If you have fitted a weighted regression, replace 60.8 with 52.9 in the above answer.

3. The vital lung data (Stata format, text format) looks at mine workers' vital lung capacity (a continuous measure of lung health), exposure to cadmium, and age.
   - Let's ignore the age variable for this question. Carry out a one-way ANOVA analysis (or a t-test) to see if vital capacity differs between mine workers with >10 years of cadmium exposure and those without exposure. Is the effect of cadmium exposure on vital lung capacity significant? What assumptions do we make in this analysis? How can we check the validity of the assumptions?

An ANOVA (or two-sample t-test or simple linear regression) gives a p-value of 0.047, which is significant evidence for rejecting the null hypothesis of no vital capacity difference between the two groups of miners (or the null hypothesis of no effect of cadmium exposure on vital capacity). The Stata command is `oneway vital_c cadmium` (or `ttest vital_c, by(cadmium)` or `regress vital_c cadmium`).

The assumptions we make are (1) equal within-group variance (or equal residual variances) and (2) normal distribution of vital capacity within each group (or normality of the residuals). The first assumption can be checked by using the Bartlett's test in the output of the `oneway` command. The test result is not significant (p-value 0.076), thus we cannot reject the null of equal variance. The second assumption can be checked by examining the inverse normal plot of the residuals or carrying out the Shapiro-Wilk test on the residuals. The Stata command sequence is

```
regress vital_c cadmium
predict residuals, resid
qnorm residuals
swilk residuals
```

The results indicate no violation of the normality assumption.

- Do you think age should be taken into consideration? Provide the rationale for your answer.

Age should be taken into consideration. A scatter plot of `vital_c` versus `age` will show lung capacity decreases as a miner gets older (Stata command `scatter vital_c age`). Thus, age may be an important variable that contributes to the variation of lung capacity in addition to cadmium exposure.

- Suppose we think age should be taken into account, and we carry out a linear regression of vital lung capacity on age and cadmium. Is the effect of cadmium exposure on vital lung capacity significant? Compare the result with that of the first question. If the results are different, what are the reasons for the difference?

When we carry out a linear regression of vital lung capacity on age and cadmium (Stata command `regress vital_c age cadmium`), the results indicate that the effect of cadmium exposure on lung capacity is no longer significant after age is taken into account. This result seems to be different from above analysis in which only cadmium is considered. The reasons might be one of the following (but see my answer to the next question): (i) Cadmium exposure may be confounded with age, and thus, once age is taken into account, cadmium is no longer significant. Such confounding is possible because miners with cadmium exposure >10 years are older. (ii) After age is taken into account, cadmium exposure's effect may appear to be non-significant marginally but might become significant again if some other hidden patterns/variables are considered. (Remember Simpson's paradox?)

- Do you want to carry out further analyses? Provide the rationale for your answer. If your answer is yes, carry out the analyses you propose to do.

If we draw scatter plots of vital_c versus age separately for the two cadmium exposure groups (Stata command `scatter vital_c age, by(cadmium)`), we can see the decrease of lung capacity as a miner gets older is faster in the miners with >10 years of cadmium exposure than that in the miners without cadmium exposure. This indicates that the age effect on lung capacity may depend on cadmium exposure; in other words, there may be an interaction effect between age and cadmium exposure on lung capacity. Thus, we try to fit a regression model with age, cadmium exposure, and their interaction:

```
gen agebycad = age*cadmium
regress vital_c age cadmium agebycad
```

The results show strong interaction effect and strong cadmium effect even after age and age-cadmium interaction have been taken into account.