BIOSTATISTICS LABORATORY PART 3:

MULTIVARIATE ANALYSIS:
INTERPRETING LINEAR AND LOGISTIC REGRESSION OUTPUT

Learning objectives:
1) Understand that linear regression is used with continuous dependent variables and logistic regression with dichotomous dependent variables.
2) Learn the commands for linear and logistic regression.
3) Interpret the output from a linear regression and logistic regression models.
4) Understand that regression models can give wrong answers if not used correctly. Help should be obtained from a statistician unless you are comfortable with these methods.

OVERVIEW OF THE LAB:

Multivariate analyses simultaneously consider the relationship of multiple independent variables to a single dependent variable. For this lab, we will show you the STATA commands for the two techniques most often used for multivariate analysis: linear regression and logistic regression. However, we will not go into the assumptions or calculations underlying these methods. You should learn to correctly interpret the output from the models, much as you would in the real world where you will be working with data analysts or statisticians.

LINEAR REGRESSION:

Linear regression is used with a continuous dependent variable and one or more independent variables. The independent variables can be continuous, categorical, or dichotomous.

Simple linear regression: Simple linear regression involves the dependent variable (outcome variable) and only one independent variable. Remember that a variable must be normally distributed to be the dependent variable in linear regression.

The general command structure is as follows:

```
regress varname1 varname2
```

Note: the first variable is always the dependent variable and any following this are the independent variables. Above, varname1 is the dependent variable and varname2 is the only independent variable.

Open the Maryland CABG dataset used in the previous labs. As an example, we will consider the relationship of old age (greater than 80 years) to length of stay. (Use the variable “age80” created during lab 2.) Since length of stay is skewed to the right (from lab 2), we will need to log-transform the variable.
STATA Command:

```
generate loglos = log(los)
```

This command will generate a new variable ("loglos") that will follow a normal distribution. Let's graph los before and after log-transformation to make sure:

STATA Commands:

```
histogram los, frequency normal
histogram loglos, frequency normal
```

STATA graphs:
Prior to log-transformation:

![Histogram of los before log-transformation](image1)

After log-transformation:

![Histogram of loglos after log-transformation](image2)

The distribution looks much closer to the normal distribution after transformation. Now we will perform the simple linear regression.

STATA Command:

```
regress loglos age80
```

STATA output:

```
Source |       SS       df       MS              Number of obs =    4666
-------------+------------------------------   F(  1,  4664) =  107.59
Model |  32.3829564     1  32.3829564           Prob > F      =  0.0000
Residual | 1403.79186  4664  .300984533           R-squared     =  0.0225
          |------------------------------   Adj R-squared =  0.0223
Total | 1436.17482  4665  .307861698           Root MSE      =  .54862

loglos |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+-------------------------------------------------------------
age80 |   .2910824   .0280627    10.37   0.000     .2360662    .3460986
_cons |   1.915681   .0084194   227.53   0.000     1.899175    1.932187
```

The regression "coefficient" for age80

P-value for the relationship of age80 and loglos

95% confidence interval around the "coefficient" for age80
The coefficient tells us how much the dependent variable ("loglos") changes with each unit of the independent variable ("age80"). Since "age80" is a dichotomous variable, the coefficient is the difference between those older and younger than 80 years old.

(For those who care about the math: linear regression assumes a linear relationship between the dependent and independent variables of the form $y=mx+b$. In the STATA output, the slope or "m" is the coefficient—the change in $y$ associated with a change in one unit of $x$—and "b" or the intercept is shown in the output as "_cons".)

With this example, the units of "loglos" have no inherent meaning so we will transform the coefficient back into units that we can understand. To accomplish this conversion, the anti-log of the coefficient must be calculated. We can perform this directly in STATA (or you can cut-and-paste the whole table into Excel).

Within STATA, type the following in the command window:

```stata
display exp(0.2910824)
```

STATA output: 1.3378748

Typing "display" in STATA allows it to function like a calculator. The "exp()" function takes the antilog of the number in parentheses (the coefficient for age80 in the above command). The antilog of the coefficient should be interpreted as the percent change—rather than the absolute change—associated with a change in one unit of the independent variable. Thus, patients greater than 80 years old have length of stay 34% greater (1.34 times) than those younger than 80 years old.

Multiple linear regression: To demonstrate multiple linear regression, we will use the same example but add another independent variable. Adding additional independent variables allows for the coefficient for one variable to be "adjusted" for the other variables in the analysis. In this case, we will add gender into our model. Thus, the effect of age older than 80 on length of stay will be adjusted for gender.

STATA Command:

```
regress loglos age80 female
```

STATA Output:

```
Source |       SS       df       MS              Number of obs =    4666
-------------+------------------------------   F(  2,  4663) =   76.69
Model |  45.7332702     2  22.8666351           Prob > F      =  0.0000
Residual |  1390.44155  4663   .29818605           R-squared     =  0.0318
-------------+------------------------------   Adj R-squared =  0.0314
Total |  1436.17482  4665  .307861698           Root MSE      =  .54606

         loglos |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+-------------------------------   -----------------------------
   age80 |   .2747812   .028038       9.80   0.000     .2198135     .329749
   female |   .1169733   .0174817     6.69   0.000     .0827009    .1512458
        _cons |   1.88185   .0097872   192.28   0.000     1.862663    1.901038

The regression "coefficient" for age80

P-value for the relationship of age80 and loglos

95% confidence interval around the "coefficient" for age80

.2747812 ± 0.028038

.2198135 to .329749

95% confidence interval around the "coefficient" for female

.1169733 ± 0.0174817

.0827009 to .1512458

95% confidence interval around the "coefficient" for _cons

1.88185 ± 0.0097872

1.862663 to 1.901038
**Task:** Using the “display” command shown above, take the anti-log of the regression coefficient for “age80”. Did the magnitude of the effect change much after adjusting for gender?

**LOGISTIC REGRESSION:**

Logistic regression is used with a dichotomous (0 or 1) dependent variable and one or more independent variables.

Simple logistic regression: Simple logistic regression involves the dichotomous dependent variable and only one independent variable. The output from the simple logistic regression analysis contains an odds ratio (similar but not identical to a relative risk).

The generic command structure for logistic regression is the following (where varname1 is the dependent variable and varname2 is the independent variable):

```
logistic varname1 varname2
```

For example, if the dependent variable is death after CABG and the independent variable is the presence or absence of a risk factor for death (i.e., age older than 80 years) the output will give you the odds of death for patients with the characteristic compared to the odds of death for patients without the characteristic.

**STATA Command:**

```
logistic died age80
```

**STATA output:**

```
Logistic regression                               Number of obs   =       4661
LR chi2(1)      =       6.81
Prob > chi2     =     0.0091
Log likelihood = -593.64042                       Pseudo R2       =     0.0057
------------------------------------------------------------------------------
died |   Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
age80 |   1.982001   .4833179     2.81   0.005     1.228953    3.196483
------------------------------------------------------------------------------
```

The odds ratio for death for patients older than 80 years (compared to those younger than 80) is 1.98.

The P-value for the comparison is 0.005, indicating that the relationship is statistically significant.

The 95% confidence interval for the odds ratio is (1.228953, 3.196483), which does not overlap 1.0, further supporting the statistical significance.

This output can be interpreted in the following way: the odds of death is two times higher (odds ratio of 1.98) for patients older than 80 years old compared to those younger than 80 years old. Note that the 95% confidence interval does not overlap “1.0”. If it did, the p-value would be greater than 0.05 and the relationship would not be statistically significant.
Multiple logistic regression: Now we will add additional independent variables in our logistic regression analysis. Like linear regression, the odds ratios will be “adjusted” for the other variables. For example, if we add gender the odds ratio for age80 will be adjusted for gender:

STATA Command:

```
Logistic died age80 female
```

STATA Output:

```
Logistic regression

Number of obs = 4661
LR chi2(2) = 8.18
Prob > chi2 = 0.0168
Log likelihood = -592.9583
Pseudo R2 = 0.0068

------------------------------------------------------------------------------
died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
age80 | 1.920856   .4713761     2.66   0.008     1.187438    3.107269
female | 1.245949   .2321497     1.18   0.238     .8647718    1.795144
------------------------------------------------------------------------------
```

The odds ratio for death is slightly changed after adding “female”
P-value for the comparison
95% confidence interval for the odds ratio

After adding the variable “female” the odds ratio doesn’t change very much. But it is “adjusted” for differences in gender between the age groups. Often, conventional risk-adjustment is done using multiple logistic regression analyses.

Also note that female has an odds ratio associated with it in the STATA output. The odds ratio for “female” (1.25) indicates that females are at slightly higher risk of death after CABG. However, the P-value is high (and the confidence interval overlaps “1.0”), so this relationship does not meet statistical significance.

Task: Try adding other independent variables to this analysis [hint: use the same command above and add other variables after “female”]. Can you interpret the odds ratios for these other variables?