Biostatistics 208 HW #1, Due 1/28/20

INSTRUCTIONS

The first part of the assignment is to complete Problems 2.2-2.6 at the end of chapter 2 in the course textbook, *Regression Methods in Biostatistics*. The problems are also posted on the course website. For problems 2.2 and 2.3, you can control the number of observations using the `set obs` command before you generate the pseudo-random data, as shown in the first lab. Also, note that the `rnormal()` function used in the first lab is a simpler option to the `invnorm(uniform())` function specified in problem 2.2. For example, to generate a variable `x` containing random observations drawn from the standard normal distribution: `generate x=rnormal()`

The datasets for Problems 2.4-2.6 can be found on the course website.

The second part of the assignment involves a linear regression model with a categorical predictor, the four-level behavioral pattern variable `behpat` in the WCGS dataset used in Problem 2.6.

1. Compute average systolic blood pressure (variable `sbp`) for each level of behavioral pattern. Stata's `tab, sum()` or `tabstat` commands can be used.

2. Using the `i.` prefix (to be discussed in the second lecture and lab) for the categorical predictor variable `behpat`, run an unadjusted linear regression with outcome `sbp`.

3. Write down the values of the indicator variables required to represent `behpat` in the model, using a table similar to table 4.3 in the textbook.

4. Using the `lincom` command, and the appropriate `i.` prefixes to specify the various behavior patterns, obtain the average SBP in the four behavior pattern groups.

5. Use the `testparm` command to obtain an *F*-test of the null hypothesis that mean SBP is the same in all four behavioral pattern groups. Write a brief description of the results.

6. Fit the same model as in part 2, but use the `ib` operator with the prefix "`ibn.`" rather than "`i.`" for the behavior pattern variable. This tells Stata to omit a reference/base level allowing a mean outcome for each level of `behpat` to be estimated. Also, include the `hascons` option in the `regress` command so that Stata will know that the model including means for all levels is equivalent to the corresponding model that includes a constant. Note that the `noconstant` option will produce identical coefficient estimates, but will lead to different sums of squares, *F* test, R^2, etc. Explain differences between resulting coefficient estimates compared to those from the previously fitted model. Finally use `testparm` again to obtain an *F*-test of the null hypothesis that mean SBP is the same in all four behavioral pattern groups. (Note that you may need to consult the Stata help file and manual for `testparm` and section 11.4.3.2 of the manual on base levels for factor variables to complete this problem.)

The problems on the next page are *optional*, and are generally more technical than those above. You will receive extra credit for completing the optional exercises.

Please submit your assignment to the class website by 10:30 AM on Tuesday, Jan 28, 2020.

Optional problems (for extra credit)

O1. Do exercise 3.4 from section 3.9 of the textbook (reproduced below). *Hint*: The Stata function `runiform()` can be used to generate variables distributed uniformly on the interval [-10, 10].

> **Problem 3.4**. The correlation coefficient is a measure of linear association. Suppose x takes on values evenly over the range from –10 to 10, and that $E[y|x] = x^2$. In this case the correlation of x and y is zero, even though there is clearly a systematic relationship. What does this suggest about the need to test model assumptions? Using a statistical package, generate a random sample of 100 values of x uniformly distributed on [–10, 10], compute $E[y|x]$ for each value of x, add randomly generated standard normal errors to get the 100 values of y, and check the sample correlation of x and y.

O2. Following up on the discussion at the start of lecture 2, use simulation to investigate power of the Shapiro-Wilk test for normality (implemented by the `swilk` command in Stata) to detect non-normal distributions for a range of sample sizes. Recall that power for a hypothesis test is the probability of correct rejection of the (false) null hypothesis in favor of a true alternative. A simulation approach to investigating power requires that alternatives to the null hypothesis be specified. For this exercise, we'll assume the actual distribution is the gamma, with a range of shapes representing departures from symmetry. In Stata, a gamma random variable is generated using the `rgamma` function, which takes shape and scale parameters as arguments. (The mean is the product of the shape and scale.) As an example, the Stata code below generates and plots 3 gamma random variables (for 1000 observations) with varying degrees of asymmetry but the same mean, along with a normal random variable specified to have the same mean (10):

```
set obs 1000
gen r1 = rgamma(9,1.1)
gen r2 = rgamma(5,2)
gen r3 = rgamma(2.5,4)
gen r4 = rnormal(10,3.5)
twoway (kdensity r1) (kdensity r2) (kdensity r3) (kdensity r4)
```

Construct a simulation program to estimate the power of the `swilk` test to detect a gamma distributed random variable as non-normal for a range of sample sizes (25, 50 & 100). Use the three gamma distributions plotted above. You can modify the simulation code introduced at the end of lab 2 to get started. Note that the nominal *p*-value provided by the `swilk` test output represents the true-positive rejection probability if the actual distribution is non-normal. *Hint*: Save the *p*-values resulting from repeated applications of the test in your simulations (*p*-values are returned by `swilk` in the scalar `r(p)`); the proportion of these less than the nominal 0.05 level estimate power.

O3. Use algebra to verify that TSS = MSS + RSS, as stated in the last sentence of section 3.3.6 of the textbook. Use the notation presented in equations 3.7-3.9 of that section. Please show your work (using the insert equations feature of Word, or including a picture/scan of hand written work). *Hint*: You can use the following identity: $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$, and the fact that the residuals $(y_i - \hat{y}_i)$ must sum to zero. (Note that although you can likely find solutions online, it's worth trying this on your own if you really want to understand the derivation.)