VHM 812/802: Linear Regression Exercises

Data

This exercise will help refresh some of the issues in building a linear model, assessing the extent of confounding, and fitting the model to the data. The data to be used for these exercises comes from work done by Wayne Martin and colleagues in Ireland. There are 3000 observations from 1797 herds.

The data concern the time between episodes of bovine tuberculosis in cattle herds in Ireland. What we have noted is that many herds never get bovine tuberculosis while a small subset seem to get tuberculosis repeatedly. One way of expressing this is to measure the time between episodes and we have done this over a period of 16 years—the data we are using here (called **btb-episodes**) is a subset of this larger data set. Our desire is to see no new episodes of tuberculosis but if it recurs we want long periods between episodes. This data set only contains herds that developed tuberculosis more than once in the 16 year period.

A major hypothesis is that herds that repeatedly develop tuberculosis may have had a more serious previous episode, and we measure this by the number of reactors in the previous episode---all these animals would have been removed from the herd but the test is imperfect and we think that some infected animals might remain in these herds although they test negatively. So our major exposure variable is the number of reactors in the previous episode.

Two potential confounders include the herd size (number of cattle over 6 months of age in the herd) and calendar year of the previous episode.

Variable	Description	Range
herd	A herd identifier (string type variable)	NA
hdsize	Herd size (# cattle > 6 mths of age)	1-472
intvl	The interval (in days) between outbreaks (to be log transformed into -intvl_ln-)	34-6684
prev_epi	The number of previous tuberculosis episodes	1-9
p_rct	The number of positive reactor cattle in the previous episode	0-102
p_year	The calendar year of the previous episode	1989-2007

Exercise #1
Introduction to Linear Regression

1. The first thing we will do is look at the distribution of the outcome of interest (-intvl-).
Note: Although the model assumptions will be about the distribution of the errors and not the outcome, we may get a good sense of whether transformation is needed from the distribution of the outcome. For example, if this distribution is strongly skewed, the error distribution will most likely have the same characteristic.

(a) Generate a histogram to depict the distribution. Does this look suitable for a linear regression model?
(b) Natural log transform the variable and depict that distribution. (This is not perfect, but gives better hopes of achieving normally distributed errors in our models; we will return to the question of transformation in Exercise #3).

2. Using the log-transformed outcome, fit simple linear regressions for each of: -p_rct-, -p_year-, and -hdsize-. For each simple regression model:

(a) Is the predictor significant? Interpret the coefficient.
(b) Interpret the intercept, and explain what the root MSE tells you.

3. Fit a multiple linear regression model with all three predictors.

(a) Compare the coefficients for -hdsize- from the simple and multiple regression models. Has the significance of the predictor changed?
(b) Interpret the coefficient (for -hdsize-) from the multiple linear regression model.

4. We will now turn our attention to predicting the -intvl_ln- based solely on the herd size. Compute and graph prediction intervals for the mean interval (for herds of a given size) and for an individual herd. For simple linear regression, the graphs can be built directly with the twoway command (available through the Graphics-Twoway graphs menu). Alternatively, and more generally, you can compute the quantities needed for the graphs and plot them (with the twoway command).

5. Fit a model with -p_year- and -hdsize- as predictors and then add -p_rct-. What happens to the R^2 when you add -p_rct-? Can you explain why?

Exercise #2
Linear Regression – Model building

1. Draw a causal diagram showing how you think the three predictors might relate to the outcome and each other.

2. Fit models with -p_year- as the only predictor. First fit it as a continuous variable and then as a categorical variable represented by a set of indicator variables.

(a) Interpret the two intercepts.
(b) How could you improve the "interpretability" of the intercept from the first model? Make the necessary change(s) to do this.
(c) Which is better, the original model (with -p_year- as a continuous predictor) or this new model?

3. We need to determine if the relationship between -p_year- and -intvl_ln- is linear.

(a) Evaluate this by fitting a scatter plot with a lowess smoothed curve through the points.(Note: There are two other commands in Stata that can help with this process: `-lintrend-` and `-lincheck-`. Call up the help files if you want to use either of those commands.)
(b) Also evaluate it by fitting both linear and quadratic terms to the model. (Do this be creating a new variable for -p_year- squared, e.g. as: **`generate pyear_sq=p_year^2`**)

4. Compute the VIFs from the model with the quadratic term. Do these signal a problem with the model? Explore how the VIFs change after centring -year- before squaring it. (Call the new variable -pyear_ct-.)

5. Is the year (fit as both linear and quadratic terms) a confounder for the effect of herd size (-hdsize-)?

6. We want to evaluate an interaction, but to simplify this process, we ar going to convert -p_year- into a dichotomous variable. There are many ways to do this; one possibility is:

```
egen pyear_c2=cut(p_year), at(0 1999 2999) icodes
```

(a) Is there interaction between -p_rct- and this new dichotomous variable for year?
(b) What was the effect of -p_rct- in the early years (1989-1998)?
(c) What was the effect of -p_rct- in the later years (1999-2007)?

7. To simplify things, we will carry on using the dichotomous version of -p_year- (-pyear_c2-).

(a) If your main predictor of interest is -p_rct-, should either of the other two predictors be included in a regression model? Why?
(b) What is your final model? What do you conclude about the relationship between -p_rct- and -intvl_ln-?

Exercise #3
Linear Regression – Diagnostics

This exercise is designed to help you recall how to check the overall and case-by-case fit of the model to the data. Remember we do this to validate our model but also to ensure that we learn about the associations, and what influences them, as we go.

Begin by running the model with -intvl_ln- as the outcome and -p_rct-, -hdsize-, -pyear_ct- and -pyear_sq- as predictors.

1. Evaluate the assumption of homoscedasticity both graphically and statistically. What do you conclude? Explain the pattern of observations.

2. Evaluate the assumption of normality both graphically (using both a histogram and a normal quantile/probability plot) and statistically. What is your interpretation?

3. Correcting problems of overall fit: It appears that we have problems with the overall fit. Often if we can correct one of the problems, that will correct both problems. So let's look and see what we might do to obtain somewhat more normally-distributed residuals.

   (a) First, start with the original outcome variable -intvl- and carry out a Box-Cox analysis using the `-boxcox-` command to see if there is a better transformation than "ln". Note that we have already transformed Y to lnY, so we should go back and rerun the model with -intvl- as the outcome for this question. What does this approach suggest?
   (b) Second, try applying the Box Cox method to the variable intvl_ln-. In this instance we should obtain a theta= 1.886 so in approximate terms we should model -intvl_ln-^2. Does this help?

4. Subject by subject analysis of fit. Go back to the model with -intvl_ln- as the outcome.S

   (a) We first look for poor fitting subjects (potential outliers). These will have standres values >3 or <-3. Are there any?
   (b) Let's now look for leverage subjects; these could have big impact on the model. Look for cases with extreme values. Are there many? What are their characteristics?
   (c) We now move to identifying subjects that actually have a large influence on the model (using either or both of Cook's D or Dfits). Are there many? Do any in particular stand out?
   (d) Finally, because -p_rct- is our main exposure of interest, we might see if any of the observations have a particularly large influence on the estimate of that coefficient. Again, what do you find?
   (e) Try refitting the model but leave out either the observations which have the biggest impact on overall fit, or the ones which most influence the estimate of the coefficient for -p_rct-. Does this make much difference?