WWS 509 Generalized Linear Models: Precept 2
Section 2.6 through 2.10.4

Kristin E. Bietsch

Office of Population Research, Princeton University

October 2012

1 Introducing the Data

This precept uses the same data from last week. I have rescaled HDI to range between 1 and 100. Also, I have created categorical variables for TFR and HDI.

- TFR:
  - Low: (0,3.5]
  - Medium: (3.5, 5]
  - High: (5,6.5]
  - Very High: (6.5, ∞)
- HDI:
  - Low: (0,35]
  - Medium: (35, 50]
  - High: (50,100]

2 Categorical Variables

When using a reference cell:

- μ becomes the expected value of the reference cell
- a_i becomes the effect of level i of the factor, compared to the reference level

2.1 One Variable

```
. reg propurb05 tfr_med tfr_high tfr_vhigh

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) =    8.59
       Model |  5969.04464     3  1989.68155           Prob > F      =  0.0001
    Residual |   10657.018    46  231.674304           R-squared     =  0.3590
-------------+------------------------------           Adj R-squared =  0.3172
       Total |  16626.0626    49    339.3074           Root MSE      =  15.221

------------------------------------------------------------------------------
   propurb05 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tfr_med |  -16.59167   6.213886    -2.67   0.010    -29.09957   -4.083761
    tfr_high |  -27.27337   6.135721    -4.45   0.000    -39.62394   -14.92281
   tfr_vhigh |  -30.44344   6.993498    -4.35   0.000    -44.52063   -16.36626
       _cons |     60.249   4.813256    12.52   0.000     50.56042    69.93758
------------------------------------------------------------------------------
```

Figure 1: Regression of TFR on Proportion Urban

1. Interpret the coefficients of the TFR categorical variables
2. Discuss the significance levels of these variables
3. Create an anova table

Source of Variation	Sum of Squares	D.F.	Mean Squared	F-Ratio
TFR				
Residual				
Total				

2.2 Two Variables

```
.
. reg propurb05 tfr_med tfr_high tfr_vhigh hdi_med hdi_high

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  5,    44) =    6.07
       Model |  6786.64319     5  1357.32864           Prob > F      =  0.0002
    Residual |  9839.41941    44  223.623168           R-squared     =  0.4082
-------------+------------------------------           Adj R-squared =  0.3409
       Total |  16626.0626    49    339.3074           Root MSE      =  14.954

------------------------------------------------------------------------------
   propurb05 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tfr_med |  -11.09761   8.607562    -1.29   0.204    -28.44501    6.249795
    tfr_high |  -18.53574   9.088703    -2.04   0.047    -36.85282   -.2186635
   tfr_vhigh |  -18.16984   10.45492    -1.74   0.089    -39.24035    2.900671
     hdi_med |   9.374642   5.617085     1.67   0.102    -1.945849    20.69513
    hdi_high |   14.91044   9.108216     1.64   0.109    -3.445968    33.26684
       _cons |   45.89214   9.717739     4.72   0.000     26.30733    65.47696
------------------------------------------------------------------------------
```

Figure 2: Regression of TFR and HDI on Proportion Urban

Fill in this hierarchical anova table for the above model

Source of Variation	Sum of Squares	D.F.	Mean Squared	F-Ratio
TFR				
HDI|TFR				
Residual				
Total				

2.3 Interactions

Sometimes we might believe that our model should not be completely additive. For example, we might expect that the effect of marriage on health differs by sex. To test this, we would want to add an interaction effect into our model. We then can use an F-test to see if this improves the model. **The key feature of this model is that the effect of a factor now depends on the level of the other.**

```
.
. reg propurb05 tfr_med tfr_high tfr_vhigh hdi_med hdi_high ///
>         t_med_h_med t_med_h_high t_high_h_med t_high_h_high t_vhigh_h_med t_vhigh_h_high
note: t_vhigh_h_med omitted because of collinearity
note: t_vhigh_h_high omitted because of collinearity

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  9,    40) =    3.27
       Model |  7048.43192     9  783.159102           Prob > F      =  0.0045
    Residual |  9577.63068    40  239.440767           R-squared     =  0.4239
-------------+------------------------------           Adj R-squared =  0.2943
       Total |  16626.0626    49   339.3074            Root MSE      =  15.474

------------------------------------------------------------------------------
   propurb05 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tfr_med |  -3.960716   22.65143    -0.17   0.862    -49.74097    41.81953
    tfr_high |  -15.36643   20.67783    -0.74   0.462    -57.15789    26.42503
   tfr_vhigh |    -12.015   18.95155    -0.63   0.530    -50.31751     26.2875
     hdi_med |   4.049286    12.4067     0.33   0.746    -21.02559    29.12416
    hdi_high |   21.02595   20.49321     1.03   0.311    -20.39236    62.44427
 t_med_h_med |   3.399806   17.18761     0.20   0.844    -31.33765    38.13726
t_med_h_high |  -11.76595   25.67902    -0.46   0.649    -63.66519    40.13329
t_high_h_med |   9.120179   14.76693     0.62   0.540    -20.72489    38.96525
t_high_h_h~h |  -7.640237   26.33663    -0.29   0.773    -60.86855    45.58808
t_vhigh_h_~d |  (omitted)
t_vhigh_h_~h |  (omitted)
       _cons |   40.92072   19.83348     2.06   0.046     .8357547    81.00568
------------------------------------------------------------------------------
```

Figure 3: Regression of TFR and HDI with Interactions on Proportion Urban

1. Comment on the coefficients and significance of the iteration terms
2. Why are two omitted?
3. Fill in the anova table below and decide if interactions should be included in this model.

Source of Variation	Sum of Squares	D.F.	Mean Squared	F-Ratio
TFR				
HDI|TFR				
Interaction				
Residual				
Total				

2.4 Analysis of Covariance

Of course, you can have models where some variables are linear and some are categorical.

When looking at an additive model, each category has the same slope, but a different intercept.
For example, here is Figure 2.5 from Germàn's notes.

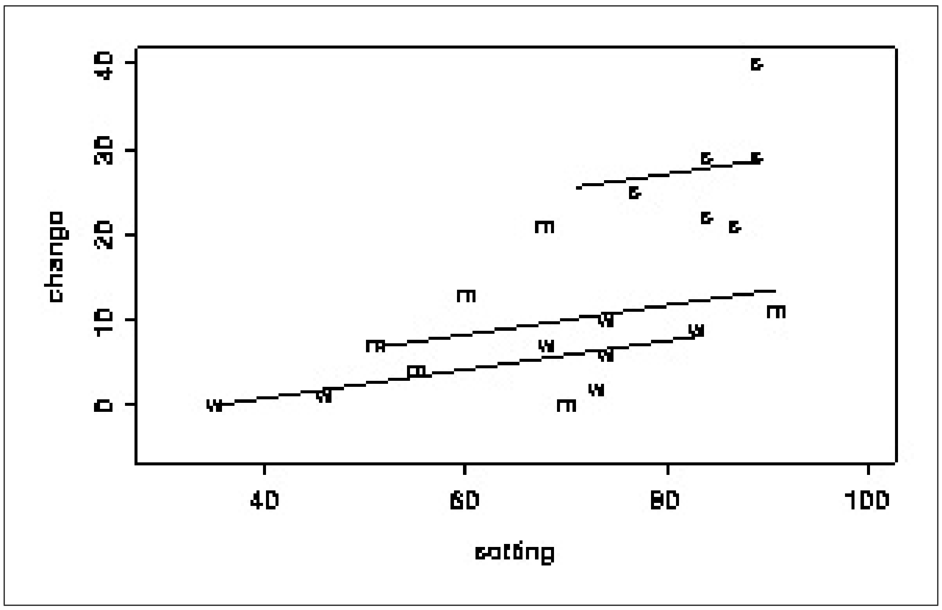

Figure 4: Analysis of Covariance Model for CBR Decline by Social Setting Score and Level of Program Effort

The slope is the same regardless of the program level, but the regression lines have different intercepts,

1. How would you interpret this?

But what if you didn't believe that countries (controlling for effort) had the same unit change for each additional point of setting? If you are suspicious, you will want to check the assumption of parallelism. It is time to add an interaction

term.
Here is Table 2.7 from the notes.

Table 2.27: Parameter Estimates for Ancova Model with Different Slopes for CBR Decline by Social Setting and Family Planning Effort (Social setting centered around its mean)

Parameter		Symbol	Estimate	Std.Error	t-ratio
Constant		μ	6.356	2.477	2.57
Effort	moderate	α_2	3.584	3.662	0.98
	strong	α_3	13.333	8.209	1.62
Setting	(linear)	β	0.1836	0.1397	1.31
Setting × Effort	moderate	γ_2	-0.0868	0.2326	-0.37
	strong	γ_3	0.4567	0.6039	0.46

1. What can you say about the interactions?
2. Does it look as if there is an argument against parallelism?

3 Regression Diagnostics

3.1 Residuals

You will spend a lot of time looking at residuals in this class. What is a residual?

$$r_i = y_i - \hat{y}_i$$

- Where y_i is the observed response and $\hat{y}_i = x'_i\hat{\beta}$ is the fitted value for the i-th unit

It can be hard to identify from the residuals which ones are outliers, because the variance of residuals is greatest near the mean and decreases as one moves towards either extreme. How can we deal with this difference in variance?

- Standardized Residuals:
  - Divides the raw residual by an estimation of its standard deviation
  - $s_i = \frac{r_i}{\sqrt{1-h_{ii}}\hat{\sigma}}$
    - Where $\hat{\sigma}$ is the estimate of the standard deviation based on the residual sum of squares
      - The problem is that the standard deviation may be affected by outliers
  - Useful for detecting outliers
    - Standardized residuals greater than 2 deserve greater scrutiny

- Jack-Knife Residuals
  - This address the problem with standardized residuals (that the standardized residual may be influenced by an outlier)
  - Estimates the error variance by omitting the i-th observation
  - $t_i = \frac{r_i}{\sqrt{1-h_{ii}}\hat{\sigma}_{(i)}}$
    - Where $\hat{\sigma}_{(i)}$ is the estimate of the standard deviation obtained by fitting the model without the i-th observation
- The jacked-knife residual is a function of the standardized residual
  - $t_i = s_i\sqrt{\frac{n-p-1}{n-p-s_i^2}}$
  - t_i is a monotonic function of s_i, so ranking observations by their standardized residuals is equivalent to ordering them by their jack-knifed residuals

Don't worry about doing this by hand, Stata makes it easy using the predict command and different options.

You can also make a residual plot for a nice visual to look for outliers. This is easily done by predicting the fitted values, and graphing them with the residuals.

Another graphic is to make a probability (Q-Q) plot, which is a graph of residuals versus the expected order statistics of the standard normal distribution.

- Germàn recommends using jack-knifed residuals.
- The plot should come very close to a straight line
  - Curvature in the Q-Q plot indicates skew distributions
    - Downward concavity corresponds to negative skewness (long tail to the left)
    - Upward concavity indicates positive skewness
    - S-shaped indicates heavy tails, or an excess of extreme values
  - To decide if your distribution or residuals is normal, refer to Filliben's table.

4 Transforming the Data

- The most popular method for a linerarizing transformation is the logarithm (naturally, pun intended!)
  - Useful when you expect the effects to be proportional to the response
  - The linear transformation looks like this: $log(Y) = \alpha + \beta x + \epsilon$

  - If $|\beta|$ is small, (less than 0.10) then you can assume $e^{\beta} - 1 = \beta$
    * So how would a coefficent of 0.07 be intrepreted?

- Another option is to use the Box Cox transformation
  - This can only be used for non-negative responses (and if you have responses of 0, you should add 0.5 to each response).
  - You can use the boxcox command in Stata to find the appropriate transformation
  - Usually you want to choose something easy, such as -1 (reciprocal), 0 (logarithim), 1/2 (square root), 1 (identity), or 2 (square)
  - Use power transformation

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases}$$