WWS 509 Generalized Linear Models: Precept 1
Section 2.1 through 2.5.6

Kristin E. Bietsch

Office of Population Research, Princeton University

September 2012

1 Introducing the data

This precept uses data from the United Nations about urbanization and fertility. Today we will focus on 50 Africa countries, the percentage of a population in each country classified as "urban," the county's total fertility rate (ask the demographer sitting closest to you for a definition), and the Human Development Index score in 2005. Please do not cite this data or the regressions.

2 Null and Saturated Models

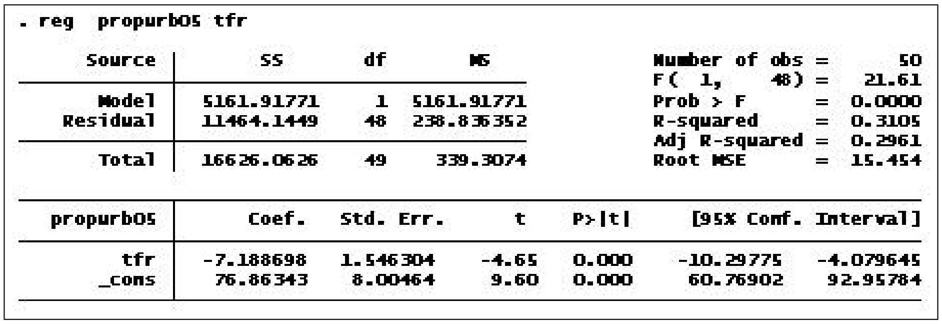
- A null model postulates no systematic differences between units
  - In this week's example, the null model would be that all countries have the same percentage of population living in urban areas
- A saturated model has as many parameters in the linear predictor as it has observation.
  - In this week's example, imagine dummy variables for Ghana, Nigeria, South Africa, etc.

Neither of these models are very informative. Let's look in the middle!

3 Simple Linear Regression

3.1 Fertility and Urbanization

Presented below is the Stata output from a regression of TFR on urbanization:



```
. reg  propurb05 tfr

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   21.61
       Model |  5161.91771     1  5161.91771           Prob > F      =  0.0000
    Residual |  11464.1449    48  238.836352           R-squared     =  0.3105
-------------+------------------------------           Adj R-squared =  0.2961
       Total |  16626.0626    49    339.3074           Root MSE      =  15.454

------------------------------------------------------------------------------
   propurb05 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tfr |  -7.188698   1.546304    -4.65   0.000    -10.29775   -4.079645
       _cons |   76.86343    8.00464     9.60   0.000     60.76902    92.95784
------------------------------------------------------------------------------
```

Figure 1: Regression of TFR on Proportion Urban

1. Interpret the coefficient for the constant and TFR

   (a) The constant is the expected response when x equals 0. This means that if TFR=0 (meaning no one was having children) the expected urbanization is 76.9%. However, this is meaningless because TFR never equals 0. For more interpretability of the constant, it would be a good idea to center TFR.

   (b) The coefficient on TFR represents the expected increment of change in the response per unit in change x. Therefore, a increase in 1 in the TFR is associated with a -7.2 percentage point change in urbanization level.

2. Construct the Likelihood ratio where this model is nested in the null

Source of variation	Degrees of freedom	Sum of squares	Mean squared	F-ratio
TFR	1	5161.91771	5161.91771	21.61
Residual	48	11464.1449	238.836352	
Total	49	16626.0626		

   (a) With 1 and 48 degrees of freedom

3. Test the significance for the coefficent for TFR

   (a) $\frac{-7.188698}{1.546304} = -4.65$

   (b) Significant at the 1% level

4. How would you calculate the R^2 by hand?

   (a) $\frac{sumofsquares}{total}$

   (b) $\frac{5161.91771}{16626.0626} = 0.31047$

5. Calculate Pearson's r

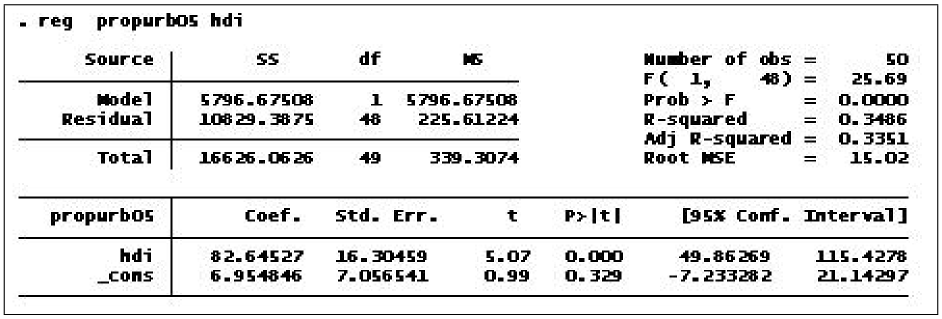
   (a) The square root of the proportion of variance explained in a simple linear regression model, with the same sign as the regression coefficent, is Pearson's linear correlation coefficient.

(b) $\sqrt{\frac{5161.91771}{16626.0626}}$

(c) In this example, each standard deviation increase in the total fertility rate is associated with an addition decrease in the proportion urban of 0.557 standard deviations.

3.2 HDI and Urbanization

Presented below is the Stata output from a regression of HDI on urbanization:



```
. reg  propurb05 hdi

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   25.69
       Model |  5796.67508     1  5796.67508           Prob > F      =  0.0000
    Residual |  10829.3875    48   225.61224           R-squared     =  0.3486
-------------+------------------------------           Adj R-squared =  0.3351
       Total |  16626.0626    49    339.3074           Root MSE      =   15.02

------------------------------------------------------------------------------
   propurb05 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hdi |   82.64527   16.30459     5.07   0.000     49.86269    115.4278
       _cons |   6.954846   7.056541     0.99   0.329    -7.233282    21.14297
------------------------------------------------------------------------------
```

Figure 2: Regression of HDI on Proportion Urban

1. Interpret the coefficient for the constant and HDI

   (a) The constant is the expected response when x equals 0. This means that if HDI=0 (wow that would suck) the expected urbanization is 6.95%. However, this is meaningless because HDI never equals 0. For more interpretability of the constant, it would be a good idea to center TFR.

   (b) The coefficient on HDI represents the expected increment of change in the response per unit in change x. However, HDI is only measured on a scale from 0 to 1. Therefore, a increase in 0.1 in the HDI is associated with a 8.26 percentage point increase in urbanization level.

2. Construct the Likelihood ratio where this model is nested in the null

Source of variation	Degrees of freedom	Sum of squares	Mean squared	F-ratio
TFR	1	5796.67508	5796.67508	25.69
Residual	48	10829.3875	225.61224	
Total	49	16626.0626		

   (a) With 1 and 48 degrees of freedom

3. Test the significance for the coefficent for HDI

   (a) $\frac{82.64527}{16.30459} = 5.07$

(b) Significant at the 1% level

4. How would you calculate the R^2 by hand?

   (a) $\frac{sum of squares}{total}$

   (b) $\frac{5796.67508}{16626.0626} = 0.3486$

5. Calculate Pearson's r

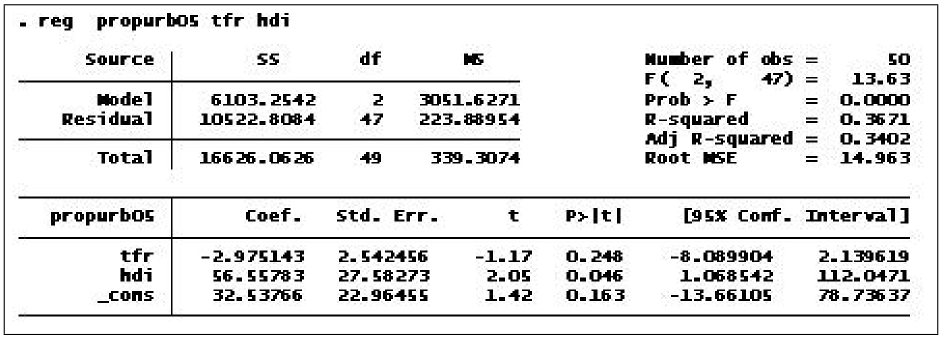
   (a) The square root of the proportion of variance explained in a simple linear regression model, with the same sign as the regression coefficient, is Pearson's linear correlation coefficient.

   (b) $\sqrt{\frac{5796.67508}{16626.0626}} = 0.59$

   (c) In this example, each standard deviation increase in the HDI is associated with an addition increase in the proportion urban of 0.59 standard deviations.

4 Multiple Linear Regression

In this section we look at the additive model which includes TFR and HDI. In an additive model, the effect of each perdictor on the response is assumed to be the same for all values of the other predictors.
Presented below is the Stata output from a regression of HDI and TFR on urbanization:



```
. reg  propurb05 tfr hdi

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   13.63
       Model |  6103.2542     2  3051.6271           Prob > F      =  0.0000
    Residual | 10522.8084    47  223.88954           R-squared     =  0.3671
-------------+------------------------------           Adj R-squared =  0.3402
       Total | 16626.0626    49  339.3074            Root MSE      =  14.963

------------------------------------------------------------------------------
   propurb05 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tfr |  -2.975143   2.542456    -1.17   0.248    -8.089904    2.139619
         hdi |   56.55783   27.58273     2.05   0.046     1.068542    112.0471
       _cons |   32.53766   22.96455     1.42   0.163    -13.66105    78.73637
------------------------------------------------------------------------------
```

Figure 3: Regression of HDI and TFR on Proportion Urban

1. Interpret the coefficient for the constant, HDI, and TFR

   (a) Constant: the constant is the expected response if both HDI and TFR eqaul 0. This would mean that the expected level of urbanization was 32.5%.

   (b) HDI: A 0.1 increase in HDI, holding TFR constant, represents the expected change in the level of level of urbanization of 5.66 percetage points.

(c) TFR: A 1 increase in TFR, holding HDI constant, represents the expected change in the level of level of urbanization of -2.98 percetage points.

2. Test the significance for HDI and TFR

   (a) HDI

      i. $\frac{56.55783}{27.58273} = 2.05$
      ii. Significant at the 5% level

   (b) TFR

      i. $\frac{-2.975143}{2.542456} = -1.17$
      ii. Not significant

   - Has significance changed from the previous models?

4.1 Gross and Net Effects

- Gross effect: the change in the response that can be associated with a given predictor in a simple linear regression
  - What is the gross effect of TFR on proportion urban? (Fill in the box below)
  - What is the gross effect of HDI on proportion urban?
- Net effect: the change in the response that can be associated with a given predictor for fixed values of other predictors
  - What is the net effect of TFR on proportion urban?
  - What is the net effect of HDI on proportion urban?

Predictor	Gross	Net
TFR	-7.189	-2.975
HDI	82.645	56.558

4.2 ANOVA for Multiple Regression

Fill in the following table for the Analysis of Variance for Multiple Regression of Proportion Urban by TFR and HDI:

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Regression	6103.2542	2	3051.6271	13.63
Residual	10522.8084	47	223.88954	
Total	16626.0626	49		

4.2.1 Hierarchical Table 1

Fill in the following table for the Hierarchical Analysis of Variance for Multiple Regression of Proportion Urban by TFR and HDI:

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
TFR	5161.91771	1	5161.91771	23.0556
HDI|TFR	941.33649	1	941.33649	4.204468
Residual	10522.8084	47	223.88954	
Total	16626.0626	49		

1. TFR and HDI|TFR are significant
2. TFR at the 1% level
3. HDI at the 5% level

4.2.2 Hierarchical Table 2

Fill in the following table for the Hierarchical Analysis of Variance for Multiple Regression of Proportion Urban by TFR and HDI:

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
HDI	5796.67508	1	5796.67508	25.89078
TFR|HDI	306.57912	1	306.57912	1.36933
Residual	10522.8084	47	223.88954	
Total	16626.0626	49		

1. HDI is significant at the 1% level
2. TFR|HDI is not significant

4.3 Partial and Multiple Correlation

- Multiple Correlation Coefficient: square root of the proportion of variance explained
- Partial Correlation Coefficient: square root of the proportion of variation explained by the second variable out of the amount left unexplained by the first
  - What are the parital correlations for TFR and Urbanization?

Multiple Correlation Coefficient	$\sqrt{\frac{6103.2542}{16626.0626}}$	0.60587
Partial Correlation Coefficient: TFR	$\sqrt{\frac{(6103.2542-5796.67508)}{10829.3875}}$	0.16826
Partial Correlation Coefficient: HDI	$\sqrt{\frac{(6103.2542-5161.91771)}{11494.1449}}$	0.28618