Regression Analysis (Spring, 2000)

By Wonjae

- **Purposes**: a. Explaining *the relationship between Y and X variables* with a model
  (Explain a variable Y in terms of Xs)
  b. Estimating and testing *the intensity of their relationship*
  c. Given a fixed x value, we can *predict y value*.
  (How does a change of in X affect Y, ceteris paribus?)
  (By constructing SRF, we can estimate PRF.)

- **OLS** (ordinary least squares) method: A method to choose the SRF in such a way that the sum of the residuals is as small as possible.

  Cf. Think of 'trigonometrical function' and 'the use of differentiation'

- **Steps of regression analysis**:

  1. Determine **independent and dependent variables**: Stare one dimension function model!
  2. Look that **the assumptions** for dependent variables are satisfied: Residuals analysis!

     a. *Linearity* (assumption 1)

     b. *Normality* (assumption 3)— draw histogram for residuals (dependent variable) or normal P-P plot

     (Spss→statistics→regression→linear→plots→'Histogram', 'Normal P-P plot of regression standardized')

     c. *Equal variance* (homoscedasticity: assumption 4)—draw scatter plot for residuals

     (Spss→statistics→regression→linear→plots: Y = *ZRESID, X =*ZPRED)

     ➔ Its form should be rectangular!

     If there were no symmetry form in the scatter plot, we should suspect the linearity.

     d. *Independence* (assumption 5,6: no autocorrelation between the disturbances, zero covariance between error term and X)—each individual should be independent

  3. Look at the **correlation between two variables** by drawing scatter graph:

     (Spss→graph→scatter→simple)

     a. Is there any correlation?

     b. Is there a linear relation?

     c. Are there outliers? → If yes, clarify the reason and modify it!

     (We should make outliers dummy as a new variable, and do regression analysis again.)

     d. Are there separated groups? → If yes, it means those data came from different populations

4. **Obtain a proper model** by using statistical packages (SPSS)

5. **Test the model**:

   a. Test *the significance* of the model (the significance of *slope*): **F-Test**

   In the ANOVA table, find the f-value and **p-value**(sig.)

   ➔ If p-value is smaller than alpha, the model is significant.

   b. Test *the goodness of fit* of the model→ In the 'Model Summary', look at **R-square**.

   R-square(coefficient of determination)—It measures the proportion or percentage of the total *variation* in Y explained by the regression model.

   (If the model is significant but R-square is small, it means that observed values are widely spread around the regression line.)

6. **Test that the slope is significantly different from zero**:

   a. Look at t-value in the 'Coefficients' table and find p-vlaue.

   b. T-square should be equal to F-value.

7. If there is the significance of the model, **Show the model and interpret it**!

- steps: a. Show the SRF

  b. In "Model Summary"→ Interpret R-square!

  c. In "ANOVA" table→ Show the table, interpret F-value and the null hypothesis!

  d. In "Coefficients" table→ Show the table and interpret beta values!

  e. Show the residuals statistics and residuals' scatter plot!

If there is no significance of the model, interpret it like this:

"X variable is little helpful for explaining Y variable." or

"There is no linear relationship between X variable and Y variable."

8. **Mean estimation** (prediction) and **individual prediction** :

   We can predict the mean, individuals and their confidence intervals.

   (Spss→statistics→regression→linear→save→predicted values: unstandardized)

Testing a model

Wonjae

- ***Before*** setting up a model

1. Identify the linear relationship between each independent variable and dependent variable.

   → Create scatter plot for each X and Y.

   (STATA: **plot Y X_1** , **plot Y X_2**)

   ovtest, rhs

   graph Y X1 X2 X3, matrix

   avplots)

2. Check partial correlation for each X and Y.

   (STATA: **pcorr Y X_1 X_2** , **pcorr X_1 Y X_2** , **pcorr X_2 Y X_1**)

- ***After*** setting up a model

1. Testing whether two different variables have same coefficients.

   → The null hypothesis is that "X_1" and "X_2" variables have the same impact on Y.

   (STATA: **test X_1 = X_2**)

2. Testing **Multicollinearity** (Gujarati, p.345)

   1) Detection

   ■ High R^2 but few significant t-ratios.

   ■ High pair-wise (zero-order) correlations among regressors

   (STATA: **regress Y X1 X2 X3**

   graph Y X1 X2 X3, matrix

   avplots)

   ■ Examination of partial correlations

   ■ Auxiliary regressions

   ■ Eigen-values and condition index

   ■ Tolerance and variance inflation factor

   (STATA: **regress Y X1 X2 X3**

   vif)

   # Interpretation: If a VIF is in excess of 20, or a tolerance (1/VIF) is .05 or less,
   There might be a problem of multicollinearity.

2) Correction:

A. Do nothing

- If the main purpose of modeling is predicting Y only, then don't worry.
  (since ESS is left the same)
- "Don't worry about multicollinearity if the R-squared from the regression exceeds the R-squared of any independent variable regressed on the other independent variables."
- "Don't worry about it if the t-statistics are all greater than 2."
  (Kennedy, Peter. 1998. A Guide to Econometrics: 187)

B. Incorporate additional information

- After examining correlations between all variables, find the most strongly related variable with the others. And simply *omit* it.
  (STATA: **corr X1 X2 X3**)

→ Be careful of the specification error, unless the true coefficient of that variable is zero.

- *Increase the number of data*
- *Formalize relationships among regressors*: for example, create interaction term(s)
  → If it is believed that the multicollinearity arises from an actual approximate linear relationship among some of the regressors, this relationship could be formalized and the estimation could then proceed in the context of a simultaneous equation estimation problem.
- *Specify a relationship among some parameters*: If it is well-known that there exists a specific relationship among some of the parameters in the estimating equation, incorporate this information. The variances of the estimates will reduce.
- *Form a principal component*: Form a composite index variable capable of representing this group of variables by itself, only if the variables included in the composite have some useful combined economic interpretation.
- Incorporate estimates from other studies: See Kennedy (1998, 188-189).
- Shrink the OLS estimates: See Kennedy.

3. Heteroscedasticity

1) Detection

- Create scatter plot for 'residual squares' and Y (p.368)
- Create scatter plot for each X and Y residuals (standardized)
  ("Partial Regression Plot" in SPSS)
  (STATA: **predict rstan**
  plot X1 rstan

plot X2 rstan)

- ***White's test*** (p.379)

  Step 1: regress your model (STATA: **reg Y X1 X2…**)

  Step 2: obtain the residuals and the squared residuals

  (STATA: **predict resi** / **gen resi2 = resi^2**)

  Step 3: generate the fitted values yhat and the squared fitted values yhat

  (STATA: **predict yhat** / **gen yhat2 = yhat^2**)

  Step 4: run the auxiliary regression and get the R^2

  (STATA: **reg resi2 yhat yhat2**)

  Step 5: 1) By using f-statistic and its p-value, evaluate the null hypothesis.

  or 2) By comparing $\chi^2_{\text{calculated}}$ (n times R^2) with χ^2_{critical} , evaluate it again.
  If the calculated value is greater than the critical value (reject the null),
  there might be 'heteroscedasticity' or 'specification bias' or both.

- Cook & Weisberg test

  (STATA: **regress Y X1 X2 X3**
  hettest)

- **The Breusch-Pagan test**

  (STATA**: reg Y X1 X2 …. / predict resi / gen resi2 = resi^2 / reg res2 X1 X2…**)

2) Remedial measures

- when variance is known: use WLS method

  (STATA: **reg Y* X_0 X_1* noconstant**)

  cf. Y* = Y/δ , X* = X/δ

- when variance is not known: use white' method

  (STATA: **gen X2r =sqrt(X) , gen dX2r = 1/X2r**
  gen Y* = Y/X2r
  reg Y* dX2r X2r, noconstant)

4. Autocorrelation

1) Detection

- Create plot

  (STATA: **predict resi, resi**
  gen lagged resi = resi[_n-1]
  plot resi lagged resi)

- Durbin-Watson d test

  (Run the OLS regression and obtain the residuals → compute 'd' →
  find $d_{\text{Lcritical}}$ and d_{Uvalues}, given the N and K → decide according to the decision rules)

```
( STATA:    regress Y X1 X2 X3
            dwstat                    )
```

- Runs test

```
( STATA:    regress Y X1 X2 X3
            predict resi, resi
            runtest resi              )
```

2) Remedial measures (pp.426-433)

- Estimate ρ: $\rho = 1 - d/2$ (D-W) or $\rho = n^2(1- d/2) + k^2/n^2 - k^2$ (Theil-Nagar)
- Regress with transformed variables and get the new d statistic.
- Compare it with $d_{Lcritical}$ and $d_{Uvalues}$

5. Testing Normality of residuals

→ obtain 'normal probability plot'

(With 'ZY' and 'ZX', choose 'Normal probability plot' in SPSS)

```
( STATA:   predict resi, resi       egen zr = std(resi)
           pnorm zr                                        )
```

6. Testing Outliers

- Detection

(STATA: avplot x1 / cprplot X1 / rvpplot X1)

- Cooksd test: Cook's distance, which measures the aggregate change in the estimated coefficients, when each observation is left out of the estimation

```
(STATA:   regress Y X1 X2
          predict c, cooksd
          display 4/d.f.          → remember this cut point value!
          list X1 X2 c            → compare the values in c with the cut point value!
          list c if c > 4/d.f.    → identify which observations are outliers!
          drop if c > 4/d.f.      → If you want to drop outliers!                    )
```