Introductory Computing for Statistics

Lecture 5:Correlation & Regression

Xiao Li

October 28, 2017

Highlights from Lecture 4

- t-test: PROC MEANS, PROC TTEST
- ANOVA: PROC GLM, PROC ANOVA

Today's topic

1. Correlation: PROC CORR
2. Regression: PROC REG, PROC GLM
3. Summary

Pearson's Correlation

Definition: Pearson's correlation

Pearson's correlation is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between $+1$ and -1 inclusive.

- For population:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

- For sample:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

PROC CORR

PROC CORR is used to compute correlation coefficients of two random variables, which are listed in matrix form.

Syntax

```
PROC CORR data = dataset;
    BY variables;
    VAR variables;
RUN;
```

Example5.1(open code file)

Regression

Model: Regression

Regression is a statistical methodology that studies the relationship between two or more variables so that DEPENDENT variable can be predicted from other INDEPENDENT variable(s).

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$. y is the response variable and $x_1, ..., x_p$ are independent explanatory variables.

Regression

The basic regression procedures in SAS are PROC REG and PROC GLM.

- PROC REG can only be used to analyze numerical variables.
- PROC GLM can handle a broader class of models, i.e., fit models with categorical type dependent variable(s).
- PROC GLM procedure uses more computing time and memory space and has less options, which is a trade of for its border models.

PROC REG

PROC REG is the fundamental SAS procedure that performs regression analysis for quantitative variables. With various options and statements, it can be mainly used to

- Fit the least-square regression line
- Perform step-wise model selections(NOT available in PROC GLM)
- Produce prediction and residual values

PROC REG-Syntax

Syntax

```
PROC REG data = dataset / <options>;
    MODEL dep_var = indep_var(s) / <options>;
    OUTPUT out=newdata <options>;
RUN;
```

PROC REG-options

Options following the PROC statement

```
PROC REG data = dataset / <options>;
```

- **simple** : prints descriptive statistics for variables in the MODEL statement: sum, mean, variance, standard deviation(NOT available in PROC GLM)
- **noprint** : it suppresses the printed output

PROC REG-options

Options following the MODEL statement

```
MODEL dep var = indep var(s) / < options >;
```

- **p** : prints the observed value, the predicted value and the residual for each observations in the data set.
- **r** : prints everything the P-option prints plus the standard errors of the predicted values and residuals, the studentized residuals, and Cook's D-statistics. (could be used to detect outliers, NOT available in PROC GLM).
- **cli** : prints 95% prediction intervals.
- **clm** : prints 95% confidence intervals.

Example5.2(open)

PROC REG-Model checking

Options following the MODEL statement

```
MODEL dep var = indep var(s) / < options >;
```

- **collin/collinoint**: check the collinearity or multicollinearity
- **influence/covratio/dffits/dfbetas**: identify influential observations
- **vif**: print the variance inflation factors
- **tol**: the tolerance values for parameter estimations

PROC REG-Model selection

Options following the MODEL statement

```
MODEL dep var = indep var(s) / < options >;
```

- **selection=forward**: perform forward selection for regression model
- **selection=backward:**: perform backward selection for regression model
- **selection=stepwise**: perform stepwise selection for regression model
- **selection=rsquare**: compute R-square for each independent variable

Example5.3(open)

PROC GLM

- PROC GLM is a more general procedure than PROC REG. It can additionally handle categorical type of independent variables, where a categorical variable can be, for examples, gender (Male and Female), food taste (bad, ok, good, and excellent) and blood types (O, A, B, AB).
- Although PROC GLM fits more types of models than PROC REG, it in many cases requires more computing resource/space and provide less output, and many options available in PROC REG are unavailable in PROC GLM

PROC GLM-Syntax

Syntax

```
PROC GLM data = dataset / <options>;
    CLASS variable(s); /* specify the categorical var(s) */
    MODEL dep_var = indep_var(s) / <options>;
    OUTPUT out=newdata <options>;
    MEANS effects / < options >;
RUN;
```

PROC GLM

- The CLASS statement can be used to define categorical variables(i.e. indep-vars can include one or more categorical variables)
- The MEANS statement can split the numerical variables according to the levels of categorical variable(s) and make comparison for the means and stardard devations.Options “tukey” or “scheffe” can be used
- PROC GLM can handle both categorical and numerical variables in independent variables(PROC REG and PROC ANOVA can each handle one type)

Example5.4(open)

Summary -1. Descriptive statistics

Table: Descriptive statistics

PROCEDURES	Type of variable applied
PROC UNIVARIATE	Numerical
PROC MEANS	Numerical
PROC FREQ	Character (Categorical)

Summary -2. T-test & ANOVA

Table: Difference between PROC TTEST and PROC ANOVA(GLM)

PROCEDURES	# of CLASS vars	levels in CLASS vars	Equal variance?
PROC TTEST	1	2	NOT REQUIRED
PROC ANOVA(GLM)	$>= 1$	$>= 2$	YES

Summary -3. Regression

Table: Comparison of usages between three regression procedures

PROCEDURES	dependent vars	independent vars
PROC ANOVA	continuous	all categorical
PROC GLM	categorical or continuous	categorical or continuous
PROC REG	continuous	all continuous

Others

- PROC PRINT
- PROC SORT
- PROC PLOT
- PROC CHART