# EDA-Proc Univariate, Proc Means (some additional options include median mode var Q1 Q3 RANGE QRANGE, maxdec=#, fw=# (field width for all columns), printalltypes, obs=#)

## Screen for unusual values

## Inspect the spread and shape

### Spread/Variability

#### range

#### IQR

#### Standard deviation = sqrt(Variance)

#### Coefficient of variation

### Proc Means-default options

#### Minimum

#### Maximum

#### Mean

#### Standard Deviation

#### N (non missing values)

## Characterize Central Tendency

### Mean

### Median/Quartiles

### Mode

## Draw Preliminary Conclusions

## Visualize Your Data-Proc sgscatter, proc sgplot, proc sgpanel, proc sgrender

### Histogram of the data

#### Normal?

##### Visual inspection

###### QQ plot

###### Box plot

##### Statistical summaries

###### Skewness (left or right)

###### Kurtosis (leptokurtic or platykurtic)

# Regression

## ANOVA: categorical predictor, continuous response (comparing means)

## Regression: two continuous variables (understanding relationship two continuous variables

## PROCS

### Scatter plots: proc scatter

### Correlations: proc corr

### Linear regression: proc reg or proc glm

### Automated model selection: proc reg or proc glmselect

## EDA

### Scatter plots (proc sgscatter-i.e. look for heteroskedasticity, outliers, linear vs nonlinear)

## Correlations statistics (Proc corr-plots=matrix or scatter, plots(only),nvar=all, ellipse=none)

### Measures linear relationship between two continuous variables-Pearson correlation from -1 to +1

### Analyze predictor vs outcome & predictor vs predictor (check for collinearity)

### Null hypothesis rho=0

### Avoid common errors

#### Outliers

##### Why an outlier?

###### Valid

Collect data

Replicate data

###### Error

###### Compute two correlation coefficients (one with outlier, one without, report both)

#### Concluding cause and effect relationship

#### Misinterpreting the type of relationship

## Multiple Linear Regression (proc reg, proc glm)

### Null hypothesis B1=B2=….=Bk=0, Alt hyp: at least one of the slopes (Bk) !=0

### Assumptions:

#### Mean of Y is linearly related to X

#### Errors are normally distributed with mean=0

##### QQ plot

##### Residual look normal on histogram?

#### Equal variance

##### Fit diagnostics-random spread around 0?

#### Errors are independent

##### residual vs predictors-random?

### Comparing models

#### Use adjusted R square (higher better)

## Model building & interpretation

### Automate approach to models (proc reg or proc glmselect)

#### All possible regression method (best=#)

#### Stepwise selection method (proc glmselect, selection= , select=sl, slentry=.1, slstay=.2)

##### Forward

##### Backward

##### Stepwise

### Compare models

#### AIC-lower better

#### AICC-lower better

#### SBC-lower better

#### Adj R squared-higher better

#### Cp (Mallow’s Cp statistic-model bias-over or underfitting)(plots(only)=(cp)

##### Prediction-cp <=p (# parameters in model+intercept)

##### Parameter estimation: Hocking’s criteria: cp <= 2p – pfull +1

### Honest assessment

#### Training & testing

#### Cross-validation

#### Bootstrapping

## Information Criteria (proc glmselect)-search for most parsimonious model-smaller better

### Akaike’s information criteria (AIC)-(2p+n+2)

### Corrected Akaike’s information criteria (AICC)-(n(n+p)/n-p-2)

### Sawa Bayesian information criteria (BIC)-(2(p+2)q – 2q^2)

### Schwarz Bayesian information criteria (SBC)-(plog(n))

## Model post-fitting (proc reg-qq, residualbypredicted residuals)

### Examining residuals (also look for outliers)

#### Plot residuals vs predicted values

##### Equal variance

##### Linearity

##### Independence

#### Plot residuals vs independent variables

##### Equal variance

#### Histogram or normal probability plot of residuals

##### Normal distribution

### Outliers

#### Outliers

##### Student (residuals/standard error ~z score; |student|>2, larger data sets |student| > 3)

#### Influential observations (proc reg)

##### Cookd (>4/n)(parameter estimation)

##### Rstudent (if student != rstudent; |rstudent|>2)

##### Dffits (>2; 2\*sqrt(p/n)(p=# terms in model incl intercept)(predictive model)

##### Dfbetas (>2, 2\*sqrt(1/n)

#### Handling influential observations

##### Data entry errors?

##### Adequate model?

##### Influential observations/observations > 5% model is probably wrong

##### Valid but unusual

##### If exclude then document type of observation excluded & why

### Detecting collinearity (proc reg, vif, collin, collinoint)

#### Two or more predictor variables are highly correlated with each other

#### Clue: p-value for overall model significant by predictor variable p-values not significant

#### Vif > 10 for any predictors in the model

# Categorical Data

## EDA-Proc freq (tables, plot=freqplot)

### One way frequency table: variable variable

### Two way frequency table or Cross-tabulation table: variable\*variable

## Tests of Association-Proc freq (tables / chisq expected cellchi2 relrisk measures cl)

### Null hypothesis: no association

### predictor -> response: Pearson chi square (effected by sample size)

### Cramer’s V Statistic (2 by 2:-1 to 1, larger tables: 0 to 1)

### Ordinal associations: Mantel-Haenszel (depend on sample size, yes/no association)

## Logistic Regression (proc logistic-(clodds=PL|WALD|BOTH))

### 3 types

#### Binary

#### Nominal

#### Ordinal

### Effect coding-dummy coding

### Reference cell coding (param=reference|ref)

## Goodness of fit

### AIC (explanatory)

### SC (predictive)

## Odds ratio=the increase or decrease in odds associated with a one unit difference of the predictor variable by default

## Association of predicted probabilities and Observed responses table (proc logistic)

### Comparing pairs-higher percent concordant & lower percent discordant

### Somer’s D-higher level better

### Gamma-higher level better

### Tan-a-higher level better

### c-higher level better

# Multiple Logistic Regression (proc logistic, selection=, slstay=, sls=, clodds=pl, oddsratio; store then proc plm)

## Automated model selection

### Backward

### Forward

### None

### Stepwise

### score

## Adjusted odds ratio

## Fit statistics

### AIC

### SC

### Percent concordant, discordant

### C statistic-measures predictive ability (higher better)

## Can include interactions between independent variables

### with interactions (Variable | Variable)