Statistics and Probability in Decision Modeling

Linear Regression

Multiple Linear Regression

- Linear regression models the effect of one independent variable, x, on one dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_n x_n + \varepsilon$$

- The β parameters reflect the **independent contribution** of each independent variable, x, to the value of the dependent variable, y.

Assumptions of Multiple Linear Regression

- Same as simple linear regression
  - Linearity
  - Independence of errors
  - Homoscedasticity (constant variance)
  - Normality of errors

- Methods of checking assumptions are also the same

Determining the Multiple Regression Equation

- *k+1* equations to solve for *k* independent variables and the intercept.

Determining the Multiple Regression Equation - Excel

In a real estate study, multiple variables were explored to determine the price of a house.

- # of bedrooms
- # of bathrooms
- Age of the house
- # of square feet of living space
- Total # of square feet of space
- # of garages

Find the equation if you want to predict the price of the house by total square feet and age of the house.

SSE and Standard Error of the Estimate, *SE*

$$SSE = \sum (y - \hat{y})^2$$

$$SE = \sqrt{\frac{SSE}{n - k - 1}}$$

Coefficient of Multiple Determination, R^2

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Adjusted R^2

As additional independent variables are added to the regression model, the value of R^2 increases.

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

However, sometimes these variables are insignificant and add no real value, yet inflating the R^2 value.

Adjusted R^2 takes into consideration both the additional information and the changed degrees of freedom.

$$Adjusted\ R^2 = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SS_{yy}}{n-1}} = R^2 - (1 - R^2)\frac{k}{n-k-1}$$

Nonlinear Models – Polynomial Regression

For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \epsilon$
How is this a special case of the general linear model?
Replace ${x_1}^2$ with x_2, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Multiple linear regression assumes a linear fit of the regression coefficients and regression constant, but not necessarily a linear relationship of the independent variable values.

Tukey's Ladder of Transformations

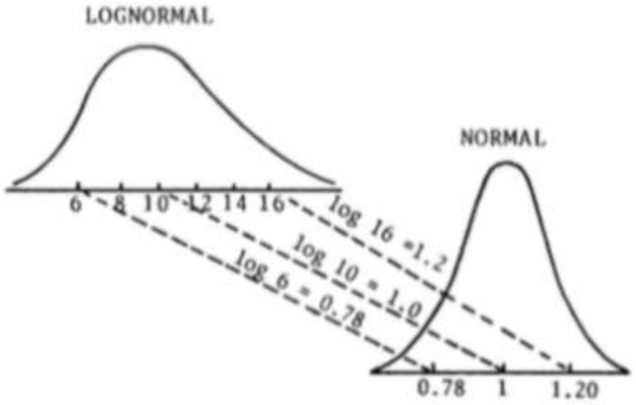
Ladder for x		
Up ladder	Neutral	Down ladder
$\ldots, x^4, x^3, x^2, x$	$\sqrt{x}, x, logx$	$-\frac{1}{\sqrt{x}}, -\frac{1}{x}, -\frac{1}{x^2}, -\frac{1}{x^3}, \ldots$
Ladder for y		
Up ladder	Neutral	Down ladder
$\ldots, y^4, y^3, y^2, y$	$\sqrt{y}, y, logy$	$-\frac{1}{\sqrt{y}}, -\frac{1}{y}, -\frac{1}{y^2}, -\frac{1}{y^3}, \ldots$

More thoughts on Transformations

DATA TRANSFORMATION

As suggested by Tabachnick and Fidell (2007) and Howell (2007), the following guidelines (including SPSS compute commands) should be used when transforming data.

If your data distribution is…	Use this transformation method.
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)



C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

Source: http://oak.ucc.nau.edu/rh232/courses/eps625/handouts/data%20transformation%20handout.pdf

Approach to determine whether to transform X or Y to achieve **linearity**, **homoscedasticity** and **normality**:

1. Often, a transformation that fixes one, fixes all.
2. In general, transforming both is not required, although sometimes it is.
3. A general rule of thumb:
    1. Transform Y first to remove heteroscedasticity.
    2. Then transform X to remove non-linearity.

Nonlinear Models – With Interaction

Interaction can be examined as a separate independent variable in regression.

For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

Indicator (Dummy) Variables

Categorical variables such as gender, geographic region, occupation, marital status, level of education, economic class, religion, buying/renting a home, etc. can also be used in multiple regression analysis.

If there are n categories, n-1 dummy variables need to be inserted into the regression analysis.

Indicator (Dummy) Variables

If a survey question asks about the region of country your office is located in, with North, South, East and West as the options, the **recoding** can be done as follows:

Region
North
East
North
South
West
West
East

North	West	South
1	0	0
0	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	0

Model Building: Search Procedures

Suppose a model to predict the world crude oil production (barrels per day) is to be developed and the predictors used are:

- US energy consumption (BTUs)
- Gross US nuclear electricity generation (kWh)
- US coal production (short-tons)
- Total US dry gas (natural gas) production (cubic feet)
- Fuel rate of US-owned automobiles (miles per gallon)

What does your intuition say about how each of these variables would affect the oil production?

Model Building: Search Procedures

Two considerations in model building:

- Explaining most variation in dependent variable
- Keeping the model simple AND economical

Quite often, the above two considerations are in conflict of each other.

If 3 variables can explain the variation nearly as well as 5 variables, the simpler model is better. Search procedures help choose the more attractive model.

Search Procedures: All Possible Regressions

All variables used in all combinations. For a dataset containing *k* independent variables, 2^k-1 models are examined. In the example of the oil production, 31 models are examined.

Tedious, Time-Consuming, Inefficient, Overwhelming.

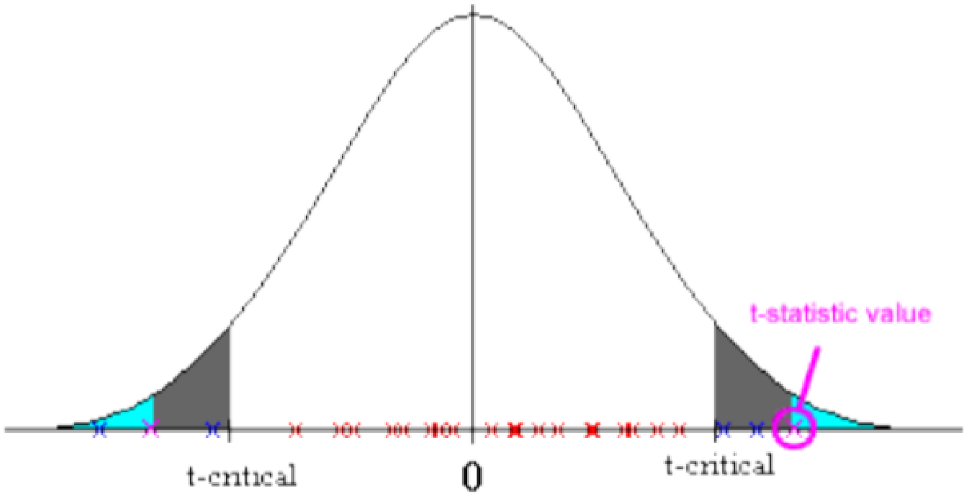
Search Procedures: Stepwise Regression

Starts a model with a single predictor and then adds or deletes predictors one step at a time.

- Step 1
  - Simple regression model for each of the independent variables one at a time.
  - Model with largest absolute value of t selected and the corresponding independent variable considered the best single predictor, denoted x_1.
  - If no variable produces a significant t, the search stops with no model.

Why LARGEST absolute t value and not the SMALLEST?

*Visualize the normal (or **t**) distribution, recall hypothesis testing, think of what the null hypothesis is and then understand what the largest and smallest absolute **t** values mean in terms of the distance from the null value.*



Search Procedures: Stepwise Regression

- Step 2
  - All possible two-predictor regression models with x_1 as one variable.
  - Model with largest absolute t value in conjunction with x_1 and one of the other k-1 variables denoted x_2.
  - Occasionally, if x_1 becomes insignificant, it is dropped and search continued with x_2.
  - If no other variables are significant, procedure stops.
- The above process continues with the 3rd variable added to the above 2 selected and so on.

Search Procedures: Stepwise Regression - R

AIC (Akaike's Information Criterion)

AIC = $2k + n\ln(RSS/n)$ where RSS is Residual Sum of Squares or SSE.

k is the number of parameters including intercept.

Sum of Sq is the additional reduction is SSE due to the addition of a variable or additional increase in SSE due to the removal of a variable.

```
> stepAICOil <- stepAIC(CrudeOilOutputlm, direction = "both")
Start:  AIC=15.29
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal + CrudeOilOutput$USDryGas

                                Df Sum of Sq    RSS    AIC
- CrudeOilOutput$USDryGas        1     0.151 29.661 13.425
- CrudeOilOutput$USNuclear       1     0.651 30.161 13.860
<none>                                       29.510 15.293
- CrudeOilOutput$USAutoFuelRate  1     2.640 32.150 15.521
- CrudeOilOutput$USCoal          1     2.683 32.193 15.555
- CrudeOilOutput$USEnergy        1    31.720 61.231 32.270

Step:  AIC=13.42
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal

                                Df Sum of Sq     RSS    AIC
- CrudeOilOutput$USNuclear       1     0.583  30.243 11.931
<none>                                        29.661 13.425
- CrudeOilOutput$USCoal          1     4.296  33.956 14.941
- CrudeOilOutput$USAutoFuelRate  1     4.575  34.236 15.154
+ CrudeOilOutput$USDryGas        1     0.151  29.510 15.293
- CrudeOilOutput$USEnergy        1   137.158 166.818 56.329

Step:  AIC=11.93
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USCoal

                                Df Sum of Sq     RSS    AIC
<none>                                        30.243 11.931
- CrudeOilOutput$USCoal          1     3.997  34.240 13.158
+ CrudeOilOutput$USNuclear       1     0.583  29.661 13.425
+ CrudeOilOutput$USDryGas        1     0.082  30.161 13.860
- CrudeOilOutput$USAutoFuelRate  1    13.531  43.774 19.545
- CrudeOilOutput$USEnergy        1   195.845 226.088 62.234
```

Multicollinearity - R

Two or more independent variables are highly correlated.

	Energy consumption	Nuclear	Coal	Dry gas	Fuel rate
Energy consumption	1				
Nuclear	0.856	1			
Coal	0.791	0.952	1		
Dry gas	0.057	-0.404	-0.448	1	
Fuel rate	0.791	0.972	0.968	-0.423	1

Multicollinearity

Sign of estimated regression coefficient when interacting may be opposite of the signs when used as individual predictors.

For example, fuel rate and coal production are highly correlated (0.968).

$$\hat{y} = 44.869 + 0.7838(fuel\ rate)$$

$$\hat{y} = 45.072 + 0.0157(coal)$$

$$\hat{y} = 45.806 + 0.0277(coal) - 0.3934(fuel\ rate)$$

Multicollinearity

Multicollinearity can lead to a model where the model (*F* value) is significant but all individual predictors (*t* values) are insignificant.

(Recall the with- and without-interaction example)

Correlation between stock 2 and stock 3 is 0.96

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.687213365
R Square	0.47226221
Adjusted R Square	0.384305911
Standard Error	4.570195728
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	224.2930654	112.1465327	5.369282452	0.021602756
Residual	12	250.6402679	20.88668899		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	50.85548009	3.790993168	13.41481713	1.38402E-08	42.59561554	59.11534464
Stock 2 ($)	-0.118999968	0.19308237	-0.616317112	0.54919854	-0.539690313	0.301690376
Stock 3 ($)	-0.07076195	0.198984841	-0.35561478	0.728301903	-0.504312675	0.362788775

Multicollinearity

- Stepwise regression prevents this problem to a great extent.

- Variance Inflation Factor (VIF): A regression analysis is conducted to predict an independent variable by the other independent variables. The independent variable being predicted becomes the dependent variable in this analysis.

$$VIF = \frac{1}{1 - {R_i}^2}$$

VIF > 10 or ${R_i}^2$>0.90 for the largest VIFs indicates a severe multicollinearity.

Thank You...