Three situations to use a log transformation

1. To accommodate nonlinearity in the regression relationship
2. To reduce right skewness in the error distribution
3. To eliminate heteroskedasticity of the form in which the conditional variance is proportional to the conditional mean squared.

Three Key Characteristics of a Good Linear Model:

- Mean of Y is linear in X
- Error terms (deviations from line) are normally distributed (few deviations > 3 sd away from line)
- Error terms have constant variance

F Test

$$f = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{SSR/k}{SSE/(N-k-1)}$$

Probability that all slopes are equal to zero
High F-stat means at least one slope is not equal to zero
If F-stat is greater than critical F – then reject null

Partial F Test

$$f = \frac{\Delta R^2/k_2}{(1-R^2_{full})/(N-(k_1+k_2)-1)} \sim F_{k_2, N-k_1-k_2-1} \quad \text{under } H_0$$

Your partial f test is a hypothesis test that all the variables you've thrown out have coeffs equal to zero

Adjusted R-squared

$$\overline{R}^2 = 1 - \frac{SSE/(N-k-1)}{SST/(N-1)} = 1 - \frac{s^2}{s_y^2}$$

Because R-squared always goes up as you add vars

t-stat = (estimate – hypo value) / standard error
standard error = (estimate – hypo value) / t-stat

$$t = \frac{b_1 - \beta_1^*}{s_{b_1}} = \frac{\text{estimate-hypo value}}{\text{std err}}$$

In Regression Output, hypo value = 0

Optimal GM = 1 / | optimal price elasticity |

Residuals vs. Fitted Plot Analysis

Look for curvature / non-linearity
Constant Variance
Normally distributed variance

Anderson Darling Statistic

A-D p-value > 0.05, Plot is assumed to be normal
A-D p-value < 0.05, Reject normality

Multiple regression coefficient interpretation

An Increase of x variable by one unit will lead to a (coefficient) increase of the y-variable, on average controlling for the other variables
Multiple Regression Residuals – residuals are deviations from your group mean – (ie men vs. women)
Calculates the **independent variation of x1, x2, and x3** that drives Y. If X3 = x1 + X2, then there is no independent variation for x3, so no coefficient

When we do a multiple regression, we begin looking at the **"pure" or partial effects of X1 and X2**. It could be the case that, while both variables are correlated, X1 is simply a better predictor and results in a significant coefficient.

Basically, the SLR model **has the influences of all of the other variables in the error term**. With simple linear regression, the coefficient explains the effect of X1 on Y **taking into account the co-movement of X1 with other variables**. Because it is averaged over the effects of other variables, the coefficient is therefore higher.

If there is correlation between two X variables, and you only regress on X1, X1 is **serving as a proxy for both variables** and thus the coefficient is higher

Simple Regression to get MR Coefficient

- X1 and X2 drive Y
- Regress X1 on X2 to purge relationship
- Residuals are independent variation of X1
- Regress Y on Residuals
    - The coefficient for this simple regression is the same as the coefficient X1 of the multiple regression

Simple Linear Regression Coefficient of P1=
b1 of P1 in **MR** + b1 of P2 in **MR** *(b1 of P1 of **simple** regression of P2 on P1)

Decomposing a Variable (for X1 and X2)
Direct Effect =MR coefficient of that Variable
Indirect Effect = MR coefficient of second variable
*coefficient of X1 in auxiliary regression of X2 on X1

Degrees of Freedom

N – 1 – k, where k = number of independent variables

Time Series - Exploit the Dependence in the series

Autocorrelation Analysis Process

1. Look at Autocorrelation Output to see which lags may make sense for an AR model

2. Fit the simplest model first
3. If there's still Autocorrelation in the residuals, try another variable
4. Repeat until residuals are normal and you have significant t-stats for the coefficients

Stationarity

The time series varies about a fixed mean and has constant variance; The dependence between successive observations does not change over time

Random Walk

$$Y_t = \beta_o + Y_{t-1} + \varepsilon_t$$

High Autocorrelations = not a trend

Look at difference data Autocorrelation – if low AC, then it must be a Random Walk

AutoCorrelation Standard Error

$$\text{Std Err}(r_s) = \frac{1}{\sqrt{T}}$$

Forecasting Y T+2 given Y T+1, N(0,4)

Hint: $Y_{T+2} = 1 + .8Y_{T+1} + \varepsilon_{T+2}$

$Y_{T+2} = 1 + .8(1 + .8Y_T + \varepsilon_{T+1}) + \varepsilon_{T+2}$

$$Y_{T+2} = 1 + .8Y_{T+1} + \varepsilon_{T+2}$$
$$Y_{T+2} = 1 + .8(1 + .8Y_T + \varepsilon_{T+1}) + \varepsilon_{T+2}$$
$$\hat{Y}_{T+2} = 1 + .8(1 + .8Y_T + 0) + 0 = 1 + .8 + .64Y_T$$

95% PI Calculation for Y_{T+2}

$$\hat{Y}_{T+2} - Y_{T+2} = -\varepsilon_{T+2} - .8\varepsilon_{T+1}$$
$$\text{Var}\left(\hat{Y}_{T+2} - Y_{T+2}\right) = 4 + .64 \times 4$$
$$95\%\text{PI}: \hat{Y}_{T+2} \pm 2 \times \left(\sqrt{4 + .64 \times 4}\right)$$

Prediction Error

$$e_f = Y_f - \hat{Y}_f = Y_f - b_0 - b_1 X_f$$

$$\text{Var}\left(e_f = Y_f - \hat{Y}_f\right) = \text{Var}\left(e_f\right) + \text{Var}\left(\hat{Y}_f\right)$$

Logistic Regression

Binary dependent variable

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Log Manipulation

$$\log\left(S_{new,128}\right) - \log\left(S_{old,128}\right)$$
$$= b_0 + b_1 \log\left(.9P_{old,128}\right) + b_2 \log\left(P_{old,64}\right) - \left(b_0 + b_1 \log\left(P_{old,128}\right) + b_2 \log\left(P_{old,64}\right)\right)$$
$$= b_1\left(\log\left(.9P_{old,128}\right) - \log\left(P_{old,128}\right)\right)$$
$$= b_1 \log(.9) = -2.93 \log(.9) = .309$$

Non Linearity (Options)

- Add additional independent variables
  - Add x1^2 (if only 1 is non-linear)
  - Add Interaction X1 * X2 (if both look non-linear)
- Transform the dependent variable (log)
- Transform independent variable

Another trick is to convert numerical non continuous data into categories or dummy variables

Basic Slope Formula

$$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \times \frac{s_y}{s_x}$$

Predictive Standard Error

$$s_{pred} = s\left(1 + \frac{1}{N} + \frac{\left(X_f - \bar{X}\right)^2}{(N-1)s_X^2}\right)^{.5}$$

$$s = \sqrt{\frac{SSE}{N-2}} = \text{standard error of the regression}$$

Standard Error of Slope Coefficient

$$s_{b_1} = \sqrt{\frac{s^2}{(N-1)s_X^2}}$$

SST - total sum of squares = SSR + SSE

SSR - regression sum of squares
(Predicted – Average Y)^2

SSE - error sum of squares
(Predicted – Actual Y)^2

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

Use the t statistic. $t = \frac{b_S - b_W}{\text{std error}\left(b_S - b_W\right)}$

remembr std error = estimated standard deviation!!

$$\text{Var}\left(b_S - b_W\right) = \text{Var}\left(b_S\right) + \text{Var}\left(b_W\right) - 2\,\text{cov}\left(b_S, b_W\right)$$

Calculator Functions

STAT – edit

STAT – 1-Var Stats

2nd – List – Math – stdDev, use { }

$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab*\text{cov}(X,Y)$