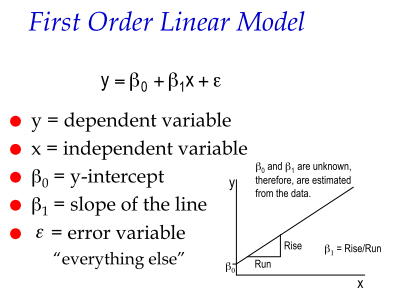
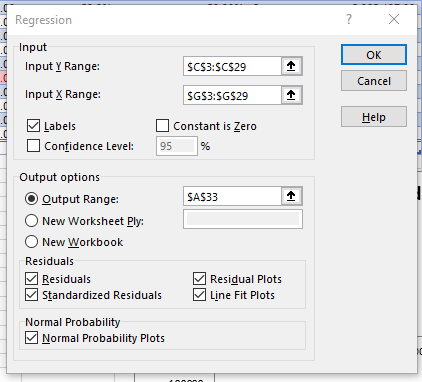
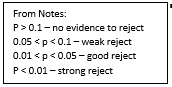
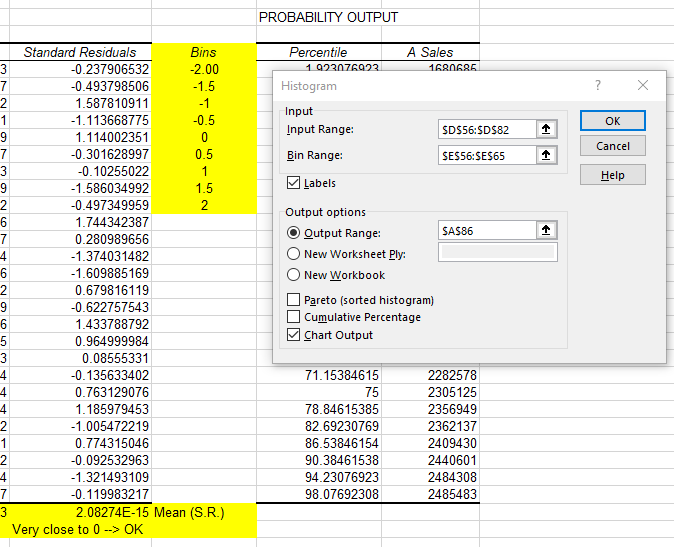
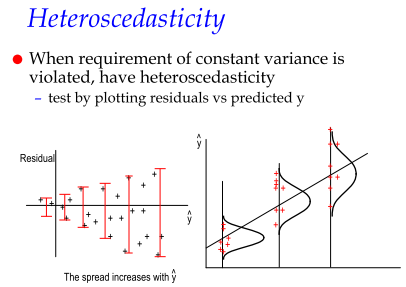
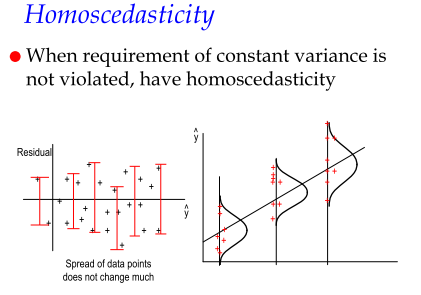
# Class 3 – Linear and Multiple Regression – 2017-08-28

**Linear Regression**

* **Linear Regression**: Modeling of relationship between dependent and independent variable
* **“Capital Asset Pricing Model”**:impact of market on stock price (Company’s Beta)

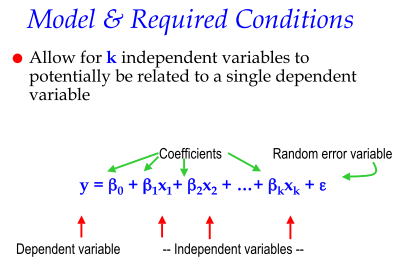


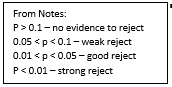
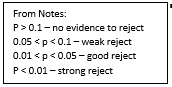
* + **Definition:** Producing a straight line that minimizes sum of squared vertical different between all points on the line (line of best fit)
  + **Analysis**: There will be a line produced whether or not there is a relationship; therefore the model needs to be analyzed
* Step 1: Scatter Plot
  + Y is Dependent; X is Independent
* Step 2: Setting up Analysis
  + Go to: Data 🡪 Data Analysis 🡪 Regression
  + *Input*: No need to force constant to 0; Confidence level is automatically 95%
  + *Output*: Select area to plot results
  + *Residuals (difference between the line of best fit and the actual observations)*:
    - Residuals and residual plots: illustrate data and plot for X vs residual
    - Line fit plot: same as a scatter plot with trend line (X vs Y)
    - Normal probability plot: for Y, closer the points to a straight line 🡪 normal distribution
* Step 3: Interpreting Regression Analysis
  + **Sample Correlation (Multiple R, r)**: ranges -1 (negative) 🡪 0 (no) 🡪 1 (positive) correlation.
  + **Coefficient of Determination (R2)**: measure of fit ranges 0 (no relationship) to 1 (perfect fit)
    - A “good” fit depends on application (in come cases 1 is expect, others 0.3 is acceptable)
  + **\*Adjusted R2**: modified R2 to incorporate sample size and dependent variables (\*for Multi-LR)
  + **Standard Error:** Variability of Y values from Predicted Values.
    - If Standard Error value < STDEV of Y values 🡪 Good estimate
  + **Intercept** = Bo ; **Slope of the Line** is noted as the variable name = B1
    - Hypothesis test: H0: B1 = 0; Ha B1 ≠ 0
    - T-Test: the critical t-value for comparison is not provided; so we use the p-value to draw conclusions on the test
      * If p-value of b1 < α < 0.05 (assume standard confidence level of 95%), we confirm statistically B1 ≠ 0 & reject H0 & conclude there is a significant relationship between X and Y; Otherwise we fail to reject it.
* \*ANOVA Table: not important linear regression because it duplicates other tests (\*only for Multi-LR)
* Step 4: Regression Diagnosis – Confirm conditions are valid
  + “RESIDUAL OUTPUT” Chart:
    - **Predicted**: Predicted Values of Y
    - **Residuals**: Predicted Value of Y – Actual Value of Y for X
    - **Standard Residuals:** Residuals / Standard deviation of residuals
  + Conditions to confirm: (1) Error variable is normally distributed with mean 0; (2) Error variance is constant for all values of x; (3) Errors are independent of each other
    - 1. If Standard Residuals are normal with Mean close to 0 🡪 Valid
      * **Outliers**: Identified in scatter diagram or if an SR is ±2 or 3 🡪 remove
      * Histogram of SRs are in a normal dist. (*or straight line on Normal Prob. Plot*)
    - 2. SR vs Predicted Values is constant 🡪 Valid
      * (1) Scatter of SR vs Predicted Values Plot & (2) Residual Plot has “Homoscedasticity” vs “Heteroscedasticity”
    - 3. For **time series** (Y versus time, X): plot residual versus time to see if there is a pattern 🡪 error is “autocorrelated”



**Multiple Linear Regression**

* **MLR**: modeling relationship between dependant variable and multiple independent variables. Similar to Linear Regression. Additional points of importance in **BLUE**. “*Why the Y goes up and down?”*



* Step 1: Perform regression analysis of multiple variables in Excel
  + Identify dependant variable Y & indep variables X1, X2… Xk (must be in continuous columns)
  + *No scatter plot required (cannot illustrate Y vs X1 vs X2…* Xk*)*
* Step 2: Interpret coefficients and predictions
  + **Multiple Correlation Coefficient (R)**: Relationship between Y and X1, X2… Xk
    - A lower R may suggests other indep. Variables not included influence Y
  + **Coefficient of Multiple Determination (R2):** % of variation of Y dependent on X1, X2… Xk
  + **Standard Error of Estimate**: If Standard Error value < STDEV of Y values 🡪 Good estimate
* Step 3: Assess model fit using the statistics
  + **Linear Relationship (ANOVA)** 🡪 is there a relationship between 1 indep. Variable?
    - Test significance for entire model
      * Ho: B1 = B2 = … = Bk vs H1: *at least* on B ≠0
      * Using F test, if Significance F < 0.01 (same scale as p-values), reject Ho 🡪 Conclude relationship with at least one indep. Variable; Otherwise we fail to reject it & no relationships at all.
    - Note: if we reject, we CANNOT say there is a relationship with all indep. Variables
  + **Sum of Squares Analysis** 🡪 test hypothesis for each individual indep. Variable to determine individual significance (not covered)
  + **Testing Coefficients (B1,2..k=0 hypothesis)**
    - If p-value of B1,2..k < α < 0.05 (assume standard confidence level of 95%), we confirm statistically B1,2…k ≠ 0 & reject H0 & conclude there is a significant relationship between X1,2…k and Y; Otherwise we fail to reject it.
* **Insignificent Coefficients:** 
  + **High P-value:** should not high P-value (ie rejected) variables in a regression model as they do not significantly contribute to R2.
  + **Small |t| stat:** If a variable |t| stat is <1; this indicates a negative impact on R2; if we remove we will achieve a stronger regression model
  + **~~However do not remove all at once~~**~~: Remove independently to determine overall impact. This can be evaluated by~~ **~~Adjusted R2~~**~~.~~  *Hendricks does not want any removed. If it’s a bad model, explain it or build better one.*
* Step 4: Diagnose violation of required conditions (Residuals) – Confirm Valid
  + Conditions to confirm: (1) Error variable is normally distributed with mean 0; (2) Error variance is constant for all values of x; (3) Errors are independent of each other; (4) **Multicollinearity**
    - 1. If Standard Residuals are normal with Mean close to 0 🡪 Valid
      * **Outliers**: Identified in scatter diagram or if an SR is ±2 or 3 🡪 remove
      * Histogram of SRs are in a normal dist. (*or straight line on Normal Prob. Plot*)
    - 2. SR vs Predicted Values is constant 🡪 Valid
      * (1) Scatter of SR vs Predicted Values Plot & (2) Residual Plot has “Homoscedasticity” vs “Heteroscedasticity”
    - 3. For **time series** (Y versus time, X): plot residual versus time to see if there is a pattern 🡪 error is “autocorrelated”
    - 4. **Multicollinearity**: a condition where independent variables have the same information; are correlated; and predict each other better than the dependant variable which negatively impacts the model: **inflates p-value** (ie not rejecting insig. variables) and **B can not be interpreted as a “slope” 🡪** Use **Variance Inflation Factor** 
      * Regress X’s against other X’s before putting them in the model
      * **If** , subsequently R2 > 80%, we assume Multicollinearity & do not use one of the Xs

**Remedying Violations to meet Conditions & Cautions**

* **Nonnormality or Hetroscedastity**: Transform Y variable
  + Positively skewed: y’ = Log y (y>0)
  + Negatively skewed: y’ = y^2
  + When Se^2 is proportional to E(y) (mean of Y) : y’ = sqrt(y)
  + When Se^2 increases as y increased: y’ = 1/y
* **R2 does not tell us:**
  + Indep. Variable X truly influences Y
  + If we are missing an important indep. Variable X
  + We used the best model/appropriate indep. Variable X
  + Presence of multi-collinearity/the model can be improved with a transformation

**Build your Model**

1. Identify logical model in mind before working
2. Identify dependent variable (Y)
3. Identify predictors (independent variables X1,2…k)
   1. Consider: multicollinearity; cost of data; min 50+8 observations/variable
4. ~~Consider “stepwise” approach of introducing 1 variable at a time~~ *Do not use stepwise – Hendricks call is “fishing”*

**Additional Links**

* <http://www.statisticshowto.com/excel-regression-analysis-output-explained/>
* <https://help.xlstat.com/customer/en/portal/articles/2062231-multiple-linear-regression-in-excel-tutorial?b_id=9283>