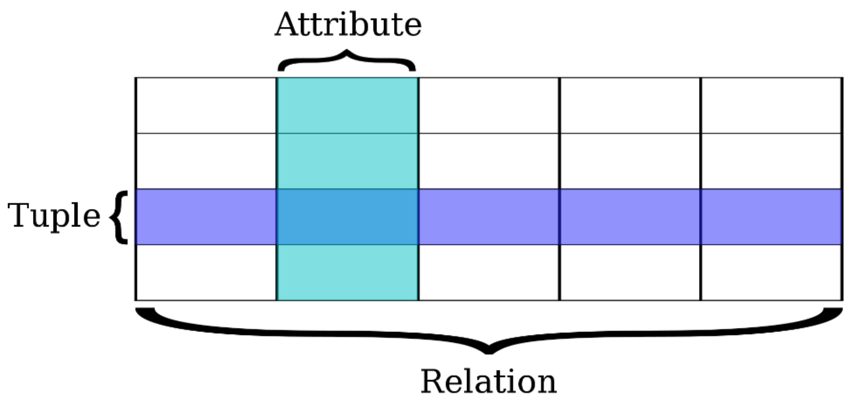
Multiple Linear Regression

Praphul Chandra

From 1 to many



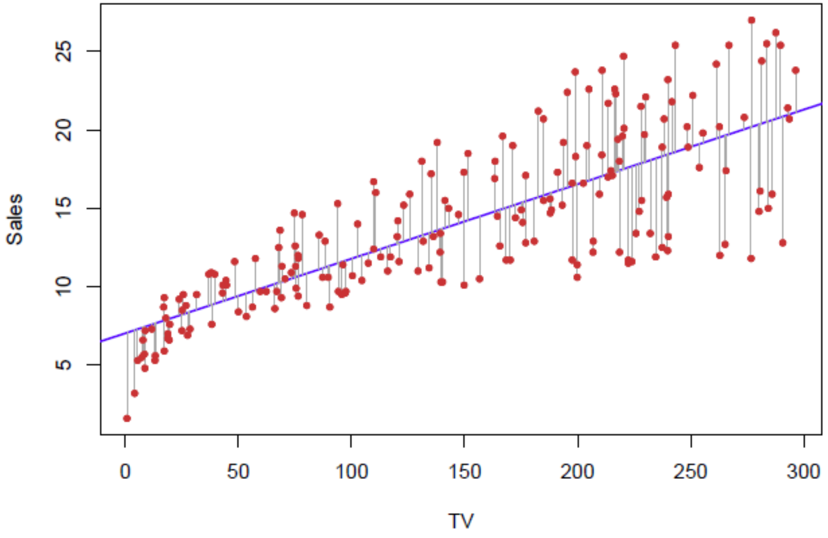
- Independent random variable.
  - Predictor variables, explanatory variables, feature, dimension, attributes

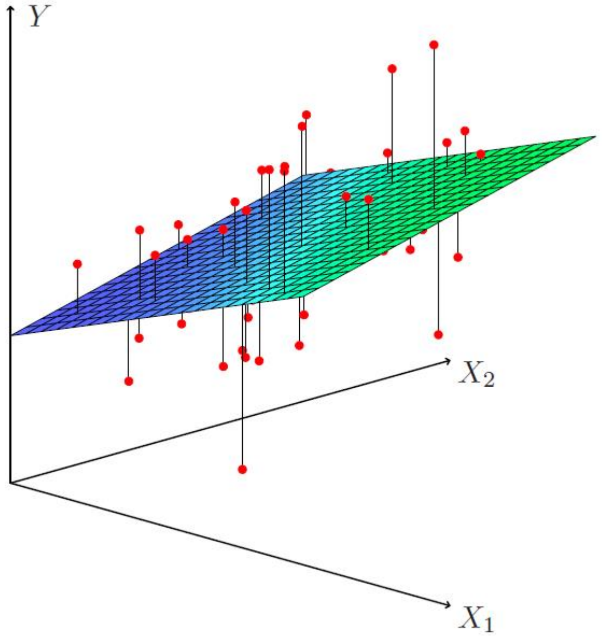
- Simple Linear Regression → Multiple Linear Regression
  - 1 independent r.v. → Multiple independent r.v.
  - Models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, y
  - The different x variables are combined in a linear way and each has its own regression coefficient
  - Same assumptions: Linearity, Noise is i.i.d. Normal with mean 0 and fixed variance

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_p x_p + \varepsilon$$

- Independence assumption (new!)
  - Independence among predictor variables : hence the name.
  - β parameters reflect the independent contribution of each variable, x, on the value of the dependent variable, y.
  - A coefficient is the slope of the linear relationship between the dependent variable (DV) and the independent contribution of the independent variable (IV),
    - i.e., that part of the IV that is independent of (or uncorrelated with) all other IVs.

Multiple Linear Regression : Visualization





p=1

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

p=2

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

p > 2 ?

Simple LR → Multiple LR

- Multiple Linear Regression
  - Dependent variable is numeric
  - Fit a ~~line~~ plane / hyper-plane

- ~~Line~~ Hyperplane Fitting
  - More than p+1 data points → Over-specified problem
  - Criteria : Minimize error (sum of squared residuals)

- Optimization problem
  - Solve (using matrix manipulation)
  - Find coefficients (line) which minimizes the Residual Sum of Squares

- Use estimated coefficients (“model”) to make predictions

Multiple Linear Regression : Math

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\min_{\beta} RSS$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$y = \beta_0 + \beta_1 x_1 + + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$\min_{\beta} RSS$$

$$\beta = (X^T X)^{-1} (X^T y)$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

Estimating the coefficients

$$\frac{\partial}{\partial \hat{\beta}}\left(\mathrm{SSE}\left(\hat{\alpha}, \hat{\beta}\right)\right) = -2\sum_{i=1}^{n}\left[(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\right](x_i - \bar{x}) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

$$minimize \sum (y - y')^2$$

Covariance matrix between Xs and Y

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \mathbf{b} = \mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{yx}}$$

Covariance matrix of Xs

$$\mathbf{Y}' = \mathbf{b}\mathbf{X}$$

$$\mathbf{Y} = \mathbf{Y}' + \varepsilon$$

More is better …. Or is it?

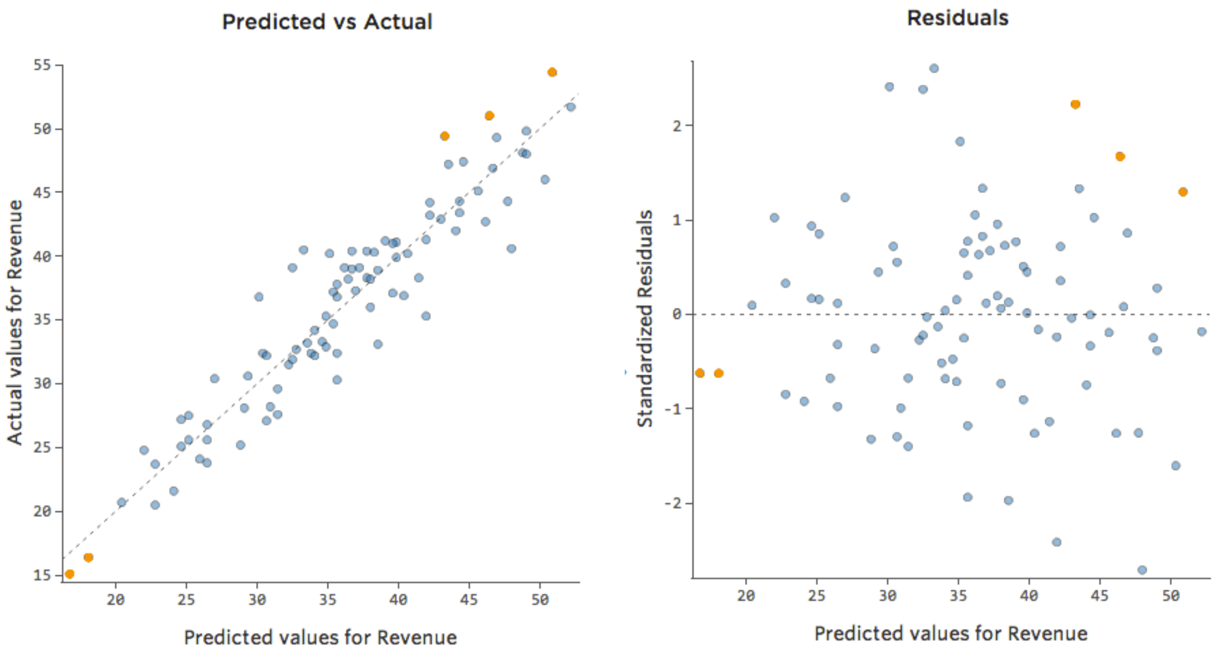
- More is better
    - More explanatory variables → More (potential) explanation → Higher R^2 (Better Fit)

- But
    - Multi-collinearity
    - Model comparison
    - Feature selection

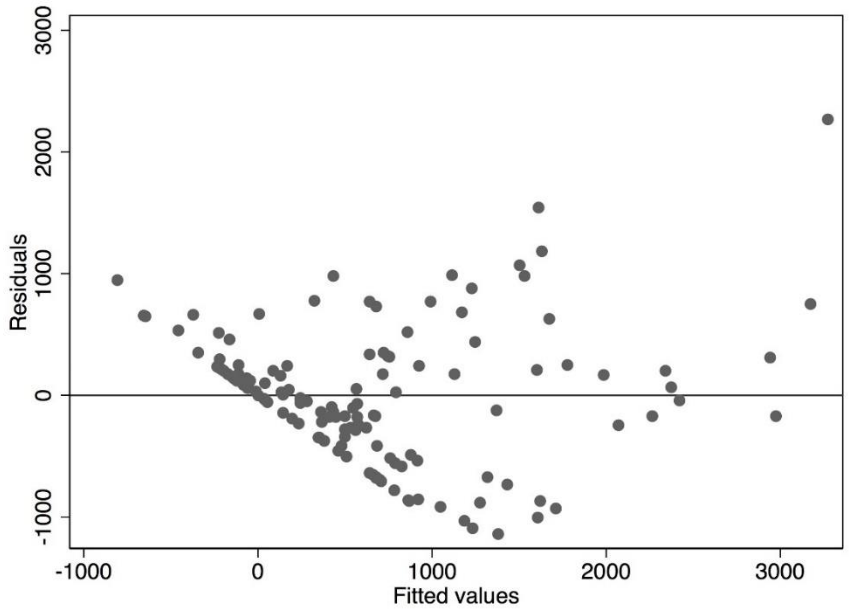
- Assumptions (as in simple linear regression)
    - Linearity
    - Homoscedasticity (constant variance)
    - Independence of errors
    - Normality of errors

Linear correlation between x and y?

- Is there a non-linear relationship?
  - Linear ➔ Plot between y & $(\beta_0 + \beta_1 x)$ would be linear
  - Linear ➔ Errors (Residuals) will not show any pattern
- Residual Plots
  - Graphical tool for identifying non-linearity
  - Plot residuals vs. fitted values
- Interpretation
  - No discernible pattern ➔ Linearity
  - U shape ➔ Non-linearity
- What next?
  - Feature Transformations (later)

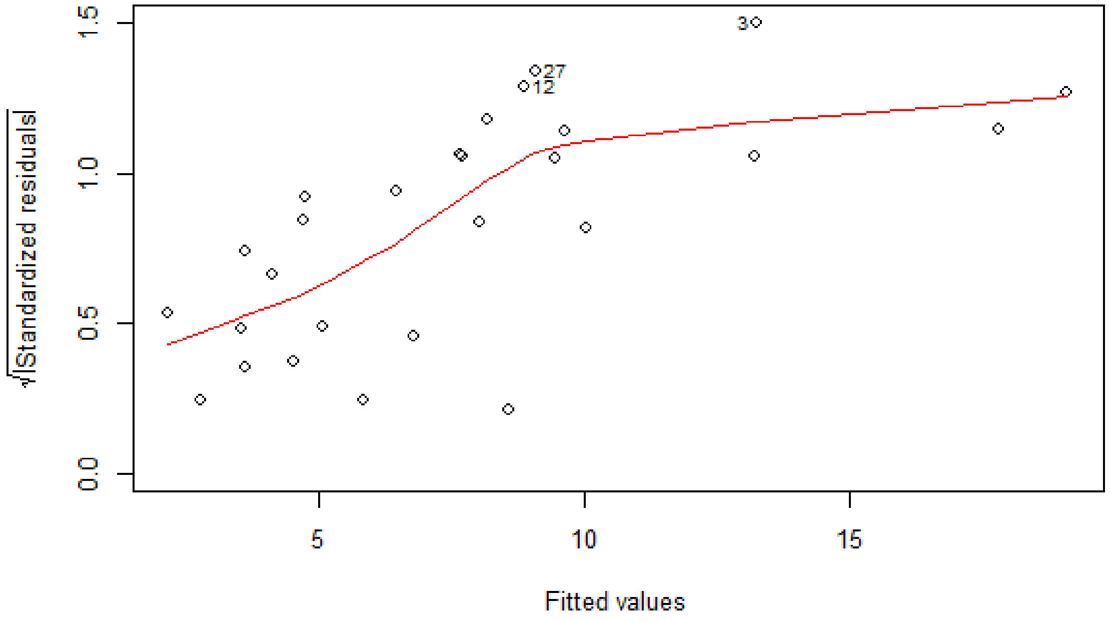


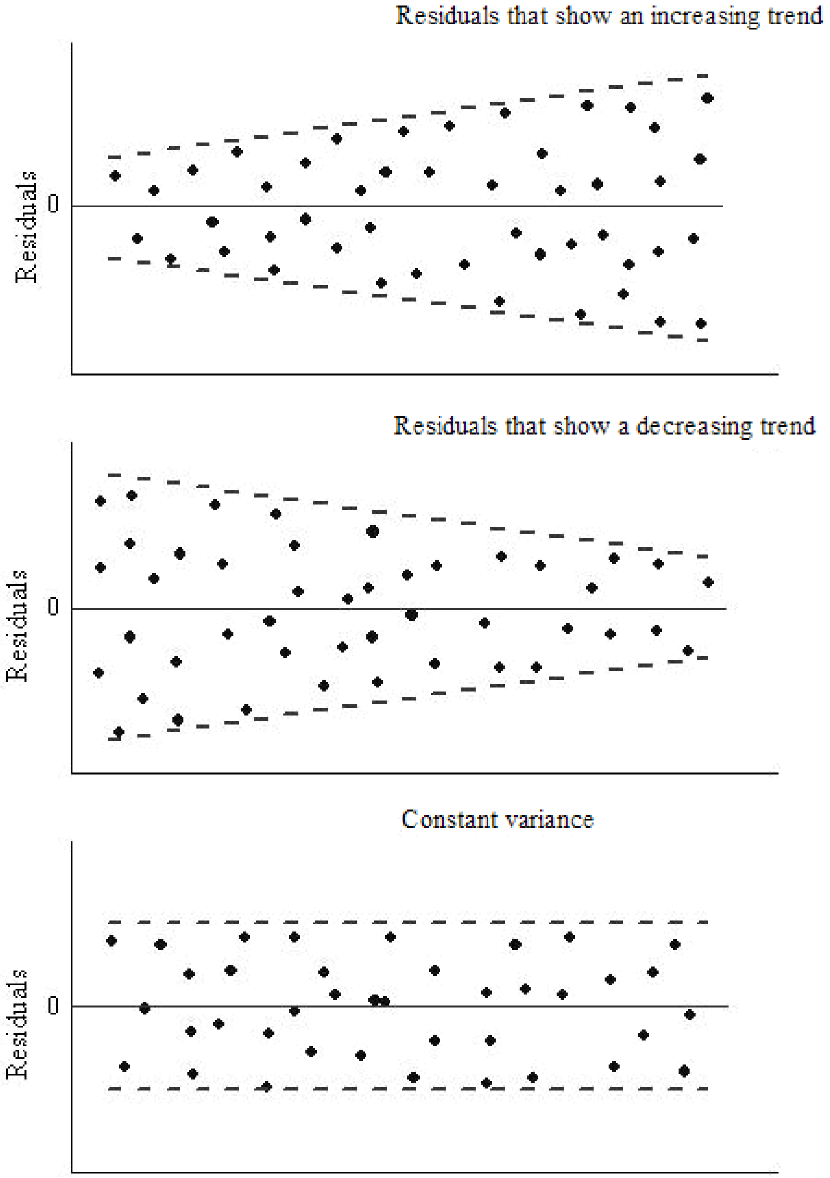
http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/



Noise has fixed variance

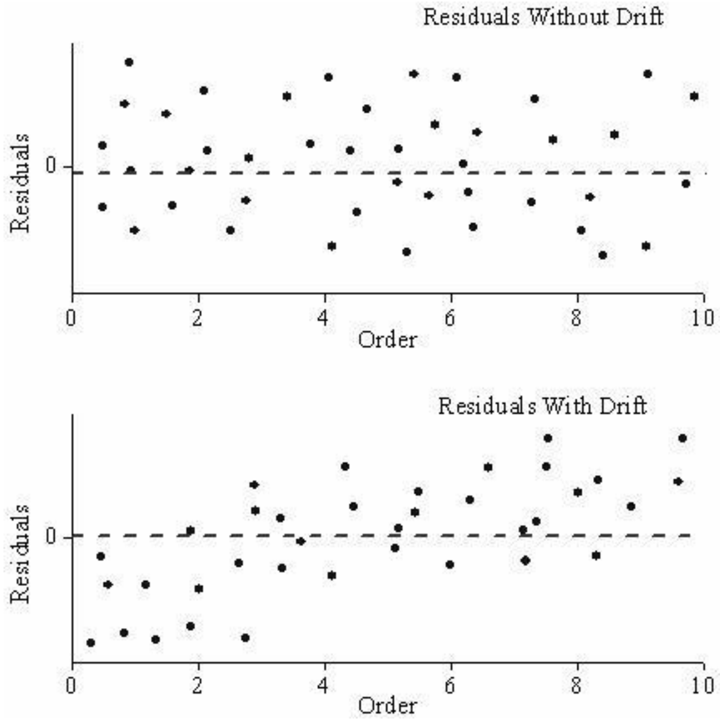
- The error terms have constant variances : homoscedasticity
  - What if variance depends on the predictor variable?
  - Need to check for heteroscedasticity
- What next?
  - Feature Transformations (later)

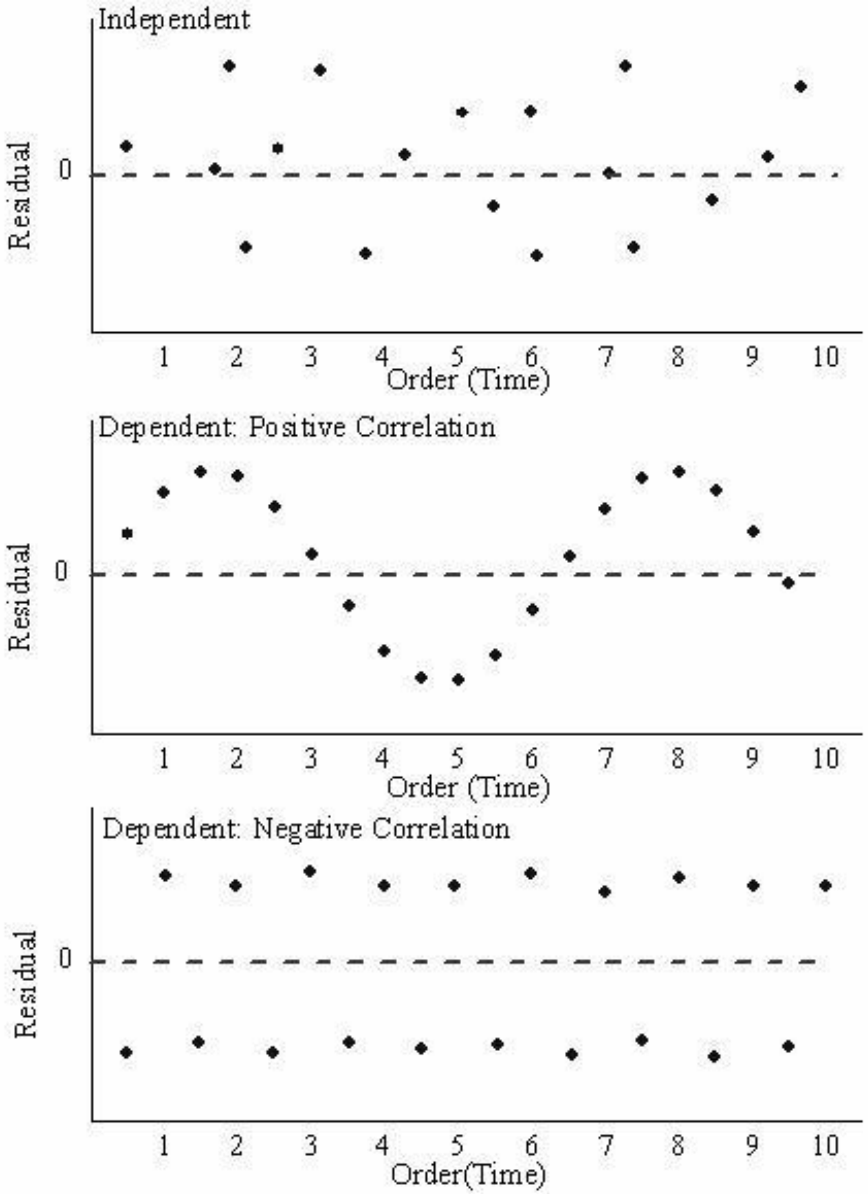




Noise (residuals) independent (i.i.d.) ?

- Are error terms correlated?
  - Are “successive” residuals correlated?
  - ϵ_i is positive provides no information about the sign of ϵ_{i+1}
- Successive?
  - Temporal
  - Any sequence… (e.g. spatial)
- Source
  - Sampling error!
  - Design of Experiment error!

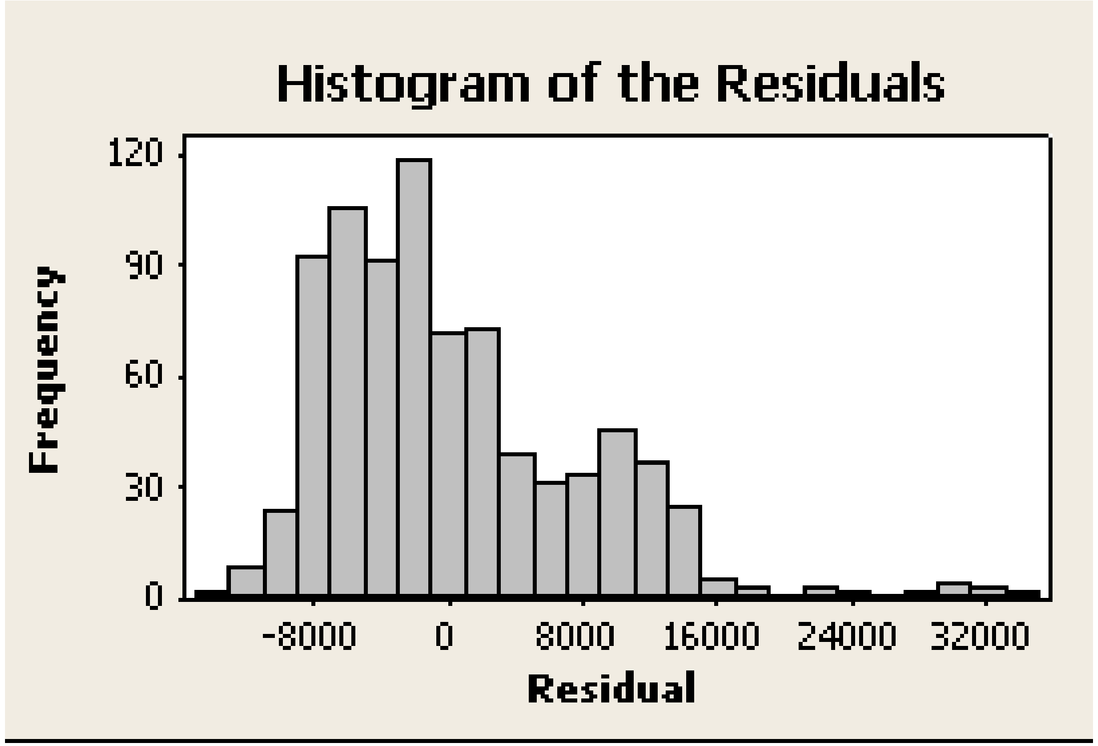




http://www.originlab.com/doc/en/UserGuide/images/Graphic_Residual_Analysis/Graphic_Residual_Analysis-4.jpg

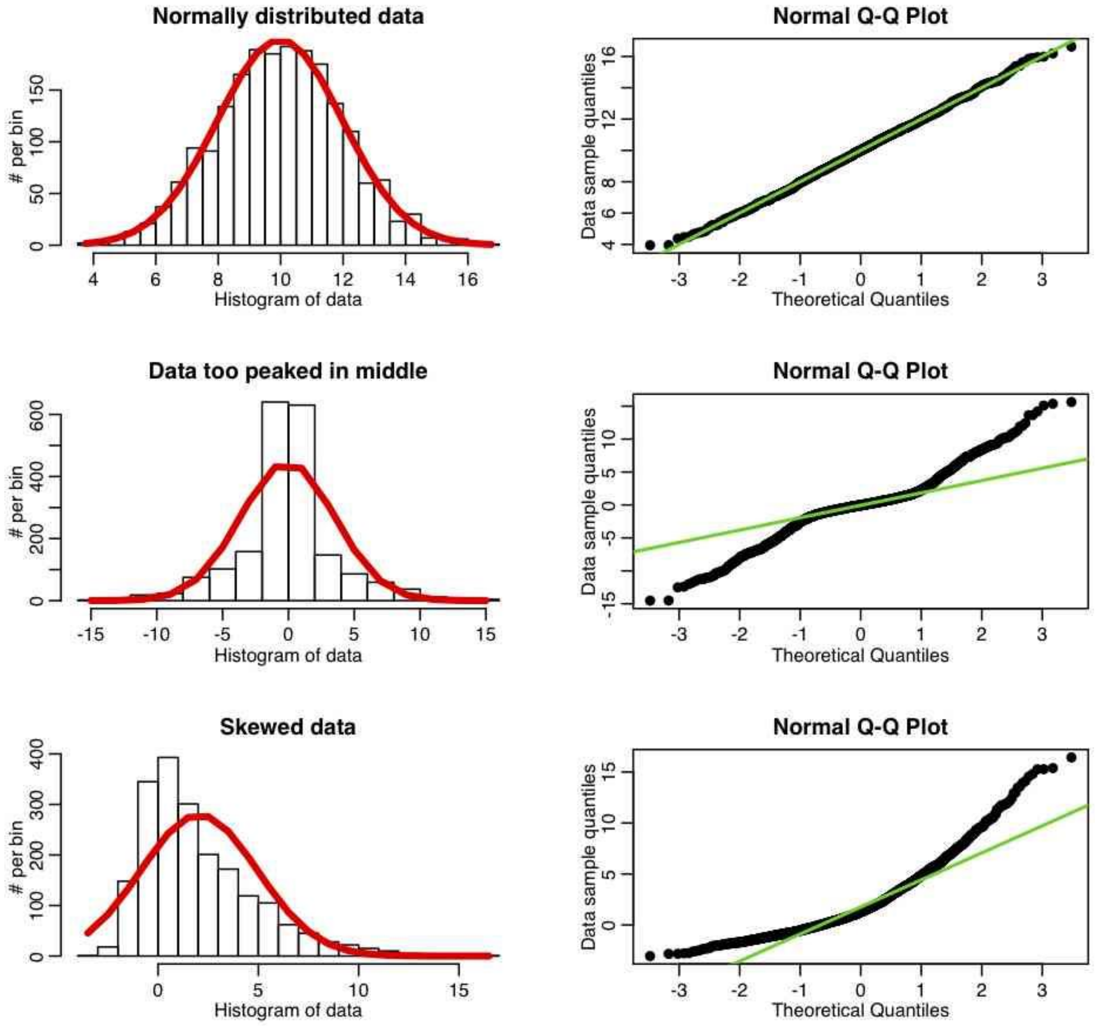
Noise (residuals) normally distributed?

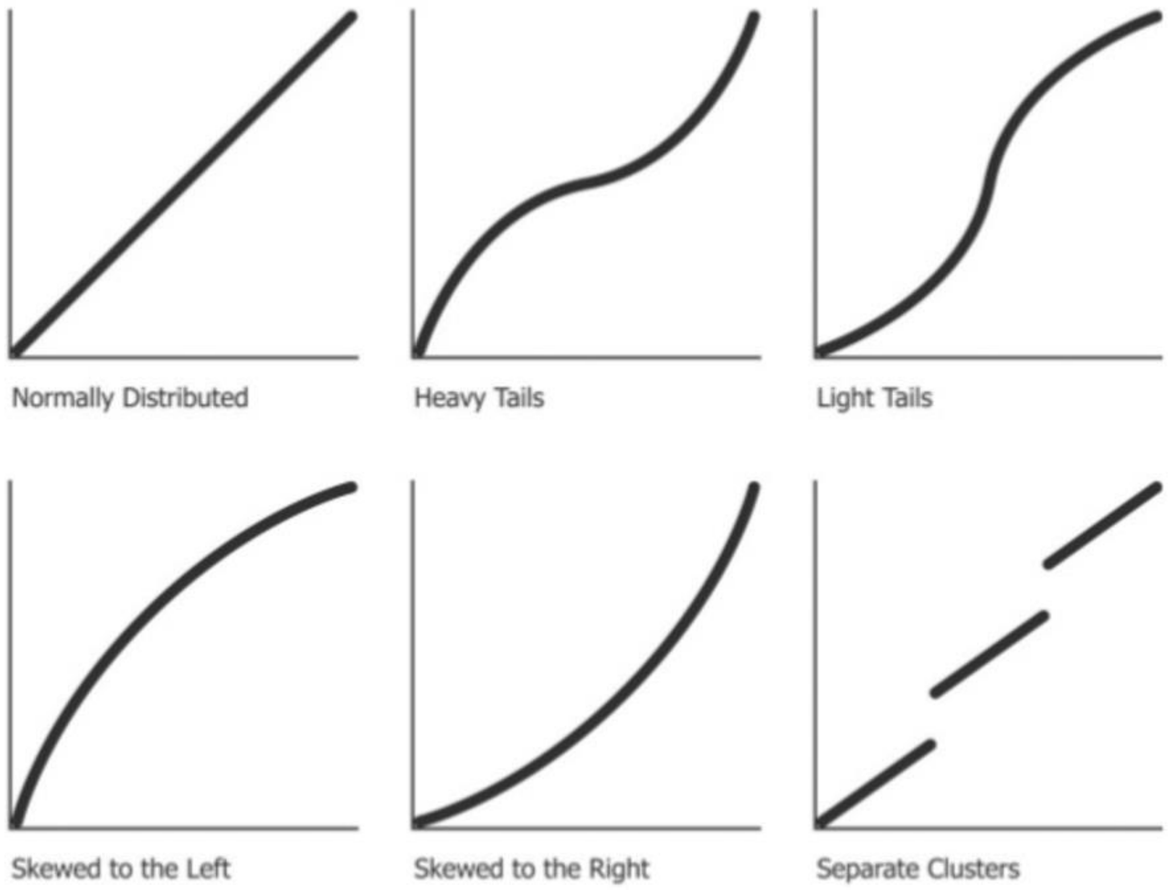
- Residual distribution is normal?
  - Plot & check
- Q-Q plot
  - Used to validate distributional assumptions of a data set.
  - Normality ➔ z-scores of the residuals should be equal to the expected z-scores at corresponding quantiles.



http://sherrytowers.com/wp-content/uploads/2013/08/qqplot_examples.jpg

Noise (residuals) normally distributed? (cont'd)





Example

- 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
- Predict the price of the house by total square feet and age of the house.

$$\hat{y} = 57.35 + 0.0177Area - 0.666Age$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57.35074586	10.00715186	5.73097587	1.31298E-05	36.47619286	78.22529885
Area (sq ft) (x1)	0.017718036	0.00314562	5.632605205	1.63535E-05	0.011156388	0.024279685
Age of House (years) (x2)	-0.666347946	0.227996703	-2.922620973	0.008417613	-1.141940734	-0.190755157

Inferential statistic : Caution!

- p-value for a regression coefficient
  - Probability of obtaining a t-statistic very far from 0 by chance
  - Expected number of coefficient for which this will happen by chance?
  - What if number of dimensions =10? 100? 1000?
  - p-value=0.05, number of dimensions = 100 ➔ 5 predictor variables may show up as "significant" by chance!

- F-statistic for k-dimensional multiple regression model
  - Tests that at least one of the regression coefficients is different from 0. (Null hypothesis : All coefficients 0)
  - $R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$
  - $F = \frac{ESS/k}{TSS/(n-k-1)} = \frac{(TSS-RSS)/k}{TSS/(n-k-1)}$
  - F-test: Comparing "the variance explained by the model" to "the variance not explained by the model"
  - When there is no relationship between the dependent variables and the predictors F is close to 1
  - If F is large, there is a relationship

(Multi)-collinearity

More isn't always better

(Multi)-collinearity

	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

- Violation of "independence" among predictor variables
  - Predictor variables are correlated
  - Which coefficient should be higher in the model?
- Impact
  - Impacts the interpretability of the model (not necessarily predictive power)
  - A variable which is in fact important may end with a lower coefficient (correlated variable gets a higher coefficient)
  - A regression coefficient which should be +ve ends up –ve (correlated variable gets a high +ve coefficient)
  - Removing one independent variable drastically changes the coefficient of others
- For example, fuel rate and coal production are highly correlated (0.968).
  - $\hat{y}=44.869+0.7838(fuel\ rate)$
  - $\hat{y}=45.072+0.0157(coal)$
  - $\hat{y}=45.806+0.0277(coal)-0.3934(fuel\ rate)$

http://circabook.com/multicollinearity-in-regression-models-penn-state-eberly-/

(Multi)-collinearity

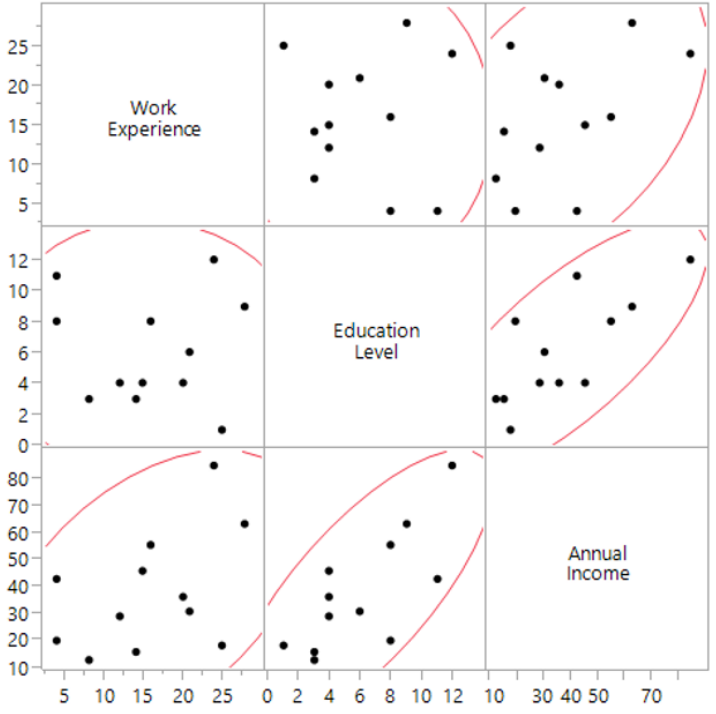
- What next?
  - Check for correlation among predictor variables.
  - Ideally before building the model
  - Drop correlated predictor variables
  - Feature transformation: PCA, PLS

- Inferential Statistics Explanation
  - Challenge: Which coefficient should be higher in the model?
  - ➔ Reduces the accuracy of the estimates of the model coefficients
  - ➔ Sampling distribution variance increases
  - ➔ Standard Errors of the coefficients increases
  - ➔ t-statistic decreases
  - ➔ p-value increases

Multivariate Correlations

	Work Experience	Education Level	Annual Income
Work Experience	1.0000	-0.0423	0.4628
Education Level	-0.0423	1.0000	0.7551
Annual Income	0.4628	0.7551	1.0000

Scatterplot Matrix



http://circabook.com/multicollinearity-in-regression-models-penn-state-eberly-/

Example

- A drug precursor molecule is extracted from a type of nut, which is commonly contaminated by a fungal toxin that is difficult to remove during the purification process. The suspected predictors of the amount of fungus are:
  - Rainfall (cm/week)
  - Noon temperature (oC)
  - Sunshine (h/day)
  - Wind speed (km/h)
  - The fungal toxin concentration is measured in µg/100 g.

```
> correlation
                Toxin         Rain    NoonTemp   Sunshine    WindSpeed
Toxin      1.00000000  0.868734134 -0.07319548 -0.05169949 -0.270555628
Rain       0.86873413  1.000000000  0.11691043  0.16841144 -0.002180167
NoonTemp  -0.07319548  0.116910426  1.00000000  0.50082303 -0.368972511
Sunshine  -0.05169949  0.168411437  0.50082303  1.00000000 -0.018439486
WindSpeed -0.27055563 -0.002180167 -0.36897251 -0.01843949  1.000000000
```

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$Sunshine + ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
      1       2       3       4       5       6       7       8
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813
      9      10
-1.1552  0.8429

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          31.6084     7.1051   4.449  0.00671 **
ToxinConc$Rain        7.0676     1.0031   7.046  0.00089 ***
ToxinConc$NoonTemp   -0.4201     0.2413  -1.741  0.14215
ToxinConc$Sunshine   -0.2375     0.5086  -0.467  0.66018
ToxinConc$WindSpeed  -0.7936     0.2977  -2.666  0.04458 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,	Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

Example (cont'd)

- Remove one of the correlated variables
- Rebuild model

- Business Implication
  - Toxin concentrations increase with increasing rainfall and decrease in drier climates characterized by higher temperatures and wind speeds.
  - The business can take a decision to rent farms in drier climates if the cost benefits of saved nuts versus higher rents are high.

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6394 -0.9308  0.1394  0.6545  2.0909

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          31.5651     6.6253   4.764  0.00311 **
ToxinConc$Rain        7.0108     0.9285   7.551  0.00028 ***
ToxinConc$NoonTemp   -0.4790     0.1919  -2.495  0.04682 *
ToxinConc$WindSpeed  -0.8218     0.2718  -3.023  0.02331 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.468 on 6 degrees of freedom
Multiple R-squared:  0.915,	Adjusted R-squared:  0.8726
F-statistic: 21.54 on 3 and 6 DF,  p-value: 0.001298
```

Multicollinearity

- Testing for pair-wise correlation not enough
  - A predictor variable may be correlated with two other variables taken together

- Variance Inflation Factor (VIF)
  - Intuition : Regress each predictor variable w.r.t. other predictors.
  - Predict an independent variable by the other independent variables.
  - The independent variable being predicted becomes the dependent variable in this analysis.
  - A “large” VIF (>> 10) indicates multicollinearity.

$$VIF(\hat{\beta}_j) = \frac{1}{1-R^2_{X_j|X_{-j}}}$$

- Stepwise regression prevents this problem to a great extent.

Checking for multi-collinearity in R

```
Call:
lm(formula = model0, data = regData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25843 -0.11727 -0.00533  0.07364  0.49503

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.85028    3.07946   3.523  0.00182 **
log(Miles)       0.42533    0.22528   1.888  0.07170 .
log(Speed)      -0.75004    0.73563  -1.020  0.31853
log(Hours)      -0.45601    0.18423  -2.475  0.02111 *
log(Population)  0.02401    0.04341   0.553  0.58559
LoadFactor      -5.82500    0.49084 -11.867 2.76e-11 ***
log(Capacity)   -1.80998    0.14851 -12.187 1.62e-11 ***
log(AdjAsset)    0.11555    0.07611   1.518  0.14259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1747 on 23 degrees of freedom
Multiple R-squared:  0.9898, Adjusted R-squared:  0.9868
F-statistic: 320.3 on 7 and 23 DF,  p-value: < 2.2e-16
```

```
> vif(fit)
     log(Miles)      log(Speed)      log(Hours) log(Population)      LoadFactor   log(Capacity)   log(AdjAsset)
      15.437923       14.227428        2.600507        3.761584        4.586951        6.951357       18.006015
```

http://subhasis4analytics.blogspot.in/2014/09/linear-regression-analysis-with-r-and.html

Feature (Model) Selection

Feature engineering

Feature Selection

- Best subset selection
    - Brute force : Try all possible combinations from the available set of predictors
    - Number of models to try 2^p
    - Computational load?

- Forward subset selection
    - p simple linear regression models; Select the best one.
    - Greedy approach: May not find THE best model but often good enough

- Backward subset selection
    - Start will all variables in.
    - Remove insignificant variables one-by-one

- Hybrid subset selection
    - Grown & prune

Forward (Hybrid) subset selection

- 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 x1.
  - If no variable produces a significant t, the search stops with no model.
- Step 2
  - All possible two-predictor regression models with x1 as one variable.
  - Model with largest absolute t value in conjunction with x1 and one of the other k-1 variables denoted x2.
  - Occasionally, if x1 becomes insignificant, it is dropped and search continued with x2.
  - 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.

Example: Feature (Model) selection

- 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?
- Search procedures help choose the more attractive model.
  - If 3 variables can explain the variation nearly as well as 5 variables, the simpler model is better.
  - All variables used in all combinations ➔ search among 31 models
  - Tedious, Time-Consuming, Inefficient, Overwhelming.
  - Use Forward subset selection

Example (cont'd)

Dependent Variable	Independent Variable	*t* Ratio	*p*-value	R^2
Oil production	Energy consumption	11.77	1.86e-11	85.2%
Oil production	Nuclear	4.43	0.000176	45.0
Oil production	Coal	3.91	0.000662	38.9
Oil production	Dry gas	1.08	0.292870	4.6
Oil production	Fuel rate	3.54	0.00169	34.2

$$y = 13.075 + 0.580x_1$$

$$y = 7.14 + 0.772x_1 - 0.517x_2$$

Dependent Variable, *y*	Independent Variable, x_1	Independent Variable, x_2	*t* Ratio of x_2	*p*-value	R^2
Oil production	Energy consumption	Nuclear	-3.60	0.00152	90.6%
Oil production	Energy consumption	Coal	-2.44	0.0227	88.3
Oil production	Energy consumption	Dry gas	2.23	0.0357	87.9
Oil production	Energy consumption	Fuel rate	-3.75	0.00106	90.8

Example (cont'd)

Dependent Variable, *y*	Independent Variable, x_1	Independent Variable, x_2	Independent Variable, x_3	*t* Ratio of x_3	*p*-value
Oil production	Energy consumption	Fuel rate	Nuclear	-0.43	0.672
Oil production	Energy consumption	Fuel rate	Coal	1.71	0.102
Oil production	Energy consumption	Fuel rate	Dry gas	-0.46	0.650

- No t ratio is significant at $\alpha = 0.05$. No new variables are added to the model.

Categorical Predictor Variables

Dealing with categorical variables

- Type of r.v. (Till now: Assume all numeric)
  - If dependent r.v. categorical : Logistic regression
  - If independent r.v. categorical : One hot encoding
- Categorical Predictor variables
  - Gender, geographic region, occupation, marital status, level of education, economic class, buying/renting a home, etc
- Replace with Indicator (Dummy) random 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:
  - If there are *n* categories, *n-1* dummy variables need to be inserted into the regression analysis.

Region
North
East
South
West

→

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

Example

- Consider the issue of gender discrimination in the salary earnings of workers in some industries. If there is discrimination, how much is one gender earning more than the other?

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.732060612	0.235584356	7.352189	8.83E-06	1.218766395	2.245354829
Age (10 years)	0.111220164	0.072083424	1.542937	0.148796	-0.045836124	0.268276453
Gender (1=Male, 0=Female)	0.458684065	0.053458498	8.58019	1.82E-06	0.342208003	0.575160126

  - Interpret as two equations.

LR with categorical variables & ANOVA

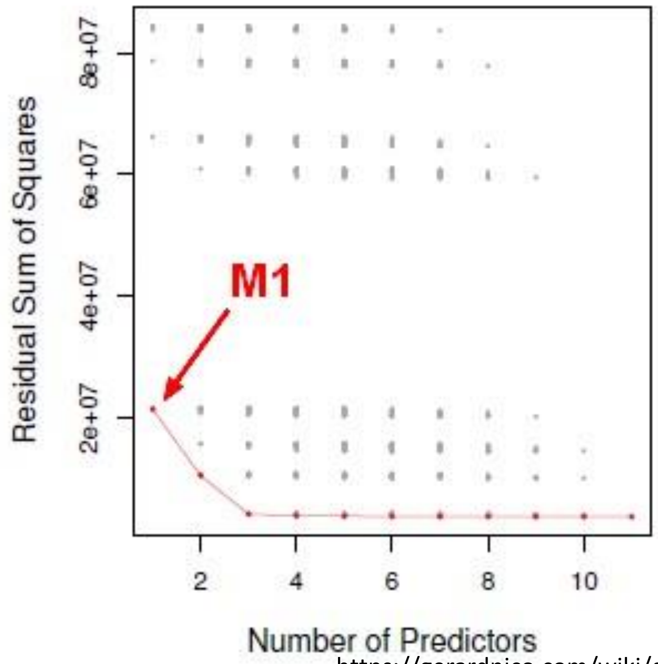
- Consider a LR with
  - Dependent variable : Numeric
  - Predictor variables: Categorical
- If the model is a good fit:
  - ➔ categorization from the input X is a good way to explain the output.
  - ➔ there is some significant difference between the groups which impacts the mean value of the dependent variable
- If the model is a good fit (F-statistic is large enough)
  - F-test, comparing “the variance explained by the model” to “the variance not explained by the model”.
  - This is exactly how one-way ANOVA works!
  - ANOVA is often expressed in terms of comparing variance within groups to variance between groups.
- Thus, ANOVA assumed
  - Identical variance between groups: the variance from the group mean for each group is the regression residual.
  - ANOVA is particularly sensitive to this assumption: if the data are heteroscedastic (i.e., the groups have different variances), ANOVA-based tests will often fail.

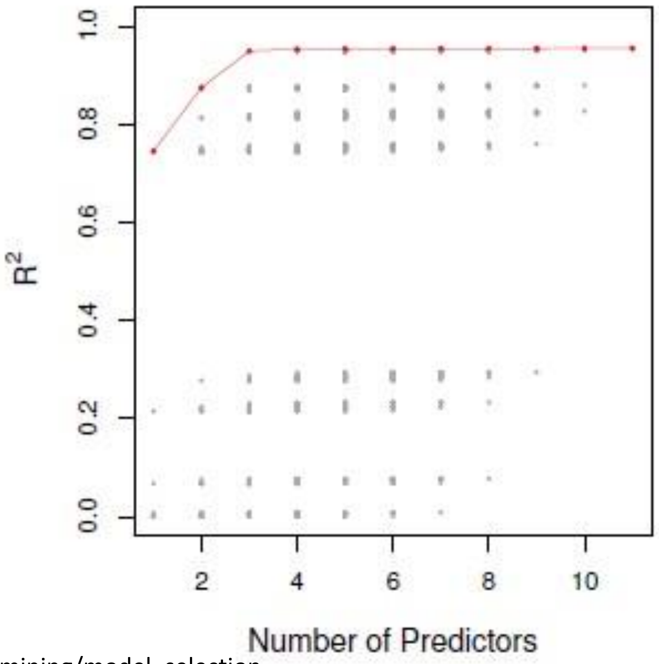
Model Comparison

Apples & Oranges

Model Comparison : Challenge

- Comparing two models with the same number of predictor variables
  - Higher R^2 better
- Comparing two models with different number of predictor variables
  - Apples & Oranges
  - More predictor variables ➔ More "flexibility" in the model
  - However, sometimes these variables are insignificant and add no real value, yet inflating the R2 value.





https://gerardnico.com/wiki/data_mining/model_selection

Model Comparison : Challenge

- Comparing two models with different number of predictor variables
  - Apples & Oranges
  - More predictor variables ➔ More "flexibility" in the model
  - Potential of overfitting

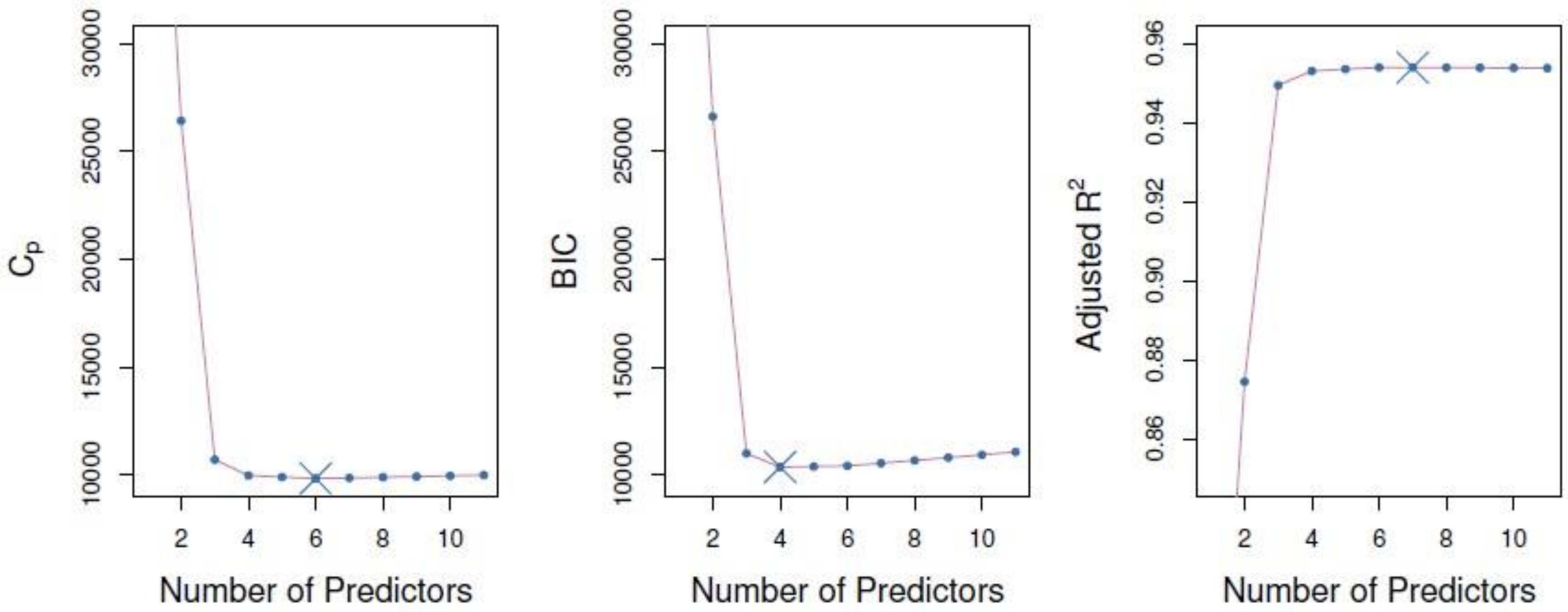
- Generalization Error (BIG Idea)
  - Sample vs. Population

- 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.

Model Comparison : Statistics

- Key Idea
  - Penalize models for using more parameters (predictor variables)
  - $Adjusted\ R^2 = 1 - \frac{\frac{RSS}{(n-d-1)}}{\frac{TSS}{n-1}}$
  - Cp, AIC, BIC
  - Aim to estimate the performance of the model learnt from sample on the population (train-test)

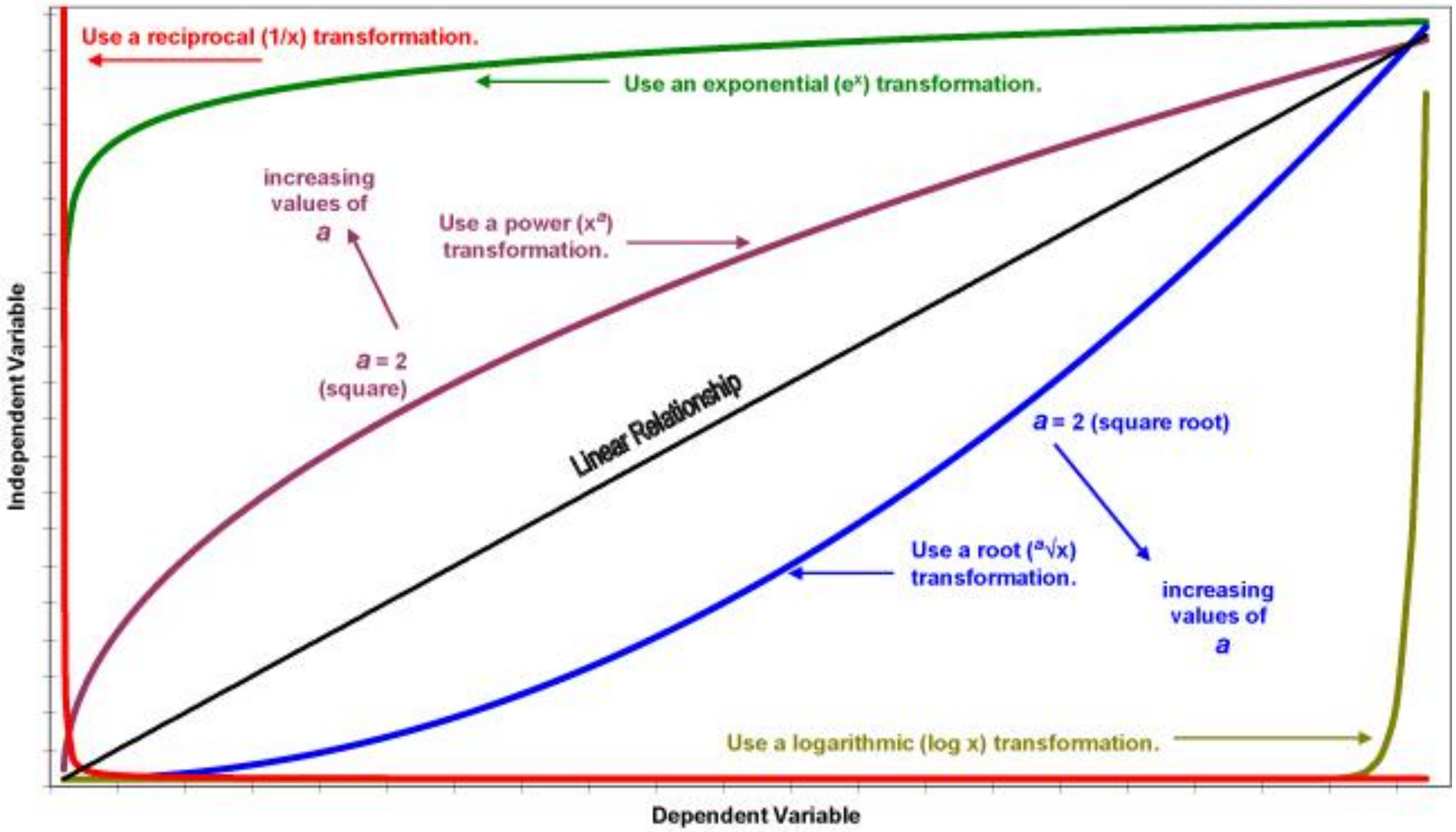
$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$
$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$
$$BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$$



Moving beyond Linearity

Feature engineering

Data Transformation cheat sheet



https://statswithcats.wordpress.com/2010/11/21/fifty-ways-to-fix-your-data/

Other tricks in Multiple Linear Regression

- Interaction Terms
- Interaction can be examined as a separate independent variable in regression.
- For example, $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2+\varepsilon$

	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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Q?

Praphul Chandra