Data Analysis Course

Multiple Linear Regression(Version-1)

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- **Multiple liner regression analysis**
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

Contents

- Why multiple regression?
- The multiple regression model
- Meaning of beta
- Variance explained
- Goodness of fit R sqared and Adj-R squared
- The F value
- Multicollinearity
- Prediction

Why Multiple Regression

- Our real world is multivariable
- Multivariable analysis is a tool to determine the relative contribution of all factors

- How do you estimate a country’s GDP? Single or multiple predictors
- Health is just dependent on smoking or drinking?
  - Diet, exercise, genetics, age, job, sleeping habits also play an important roll in deciding one’s health
  - We often want to describe the effect of smoking over and above these other variables.

Some Economics Models

Most of the real time models are multivariate

Wage determination model

$$w = f\left(S, A, A^2, T, G, L, \varepsilon\right)$$

Profit

$$\pi = \pi\left(p, a, a^2, P_c, y, i, w, \varepsilon\right)$$

Investment

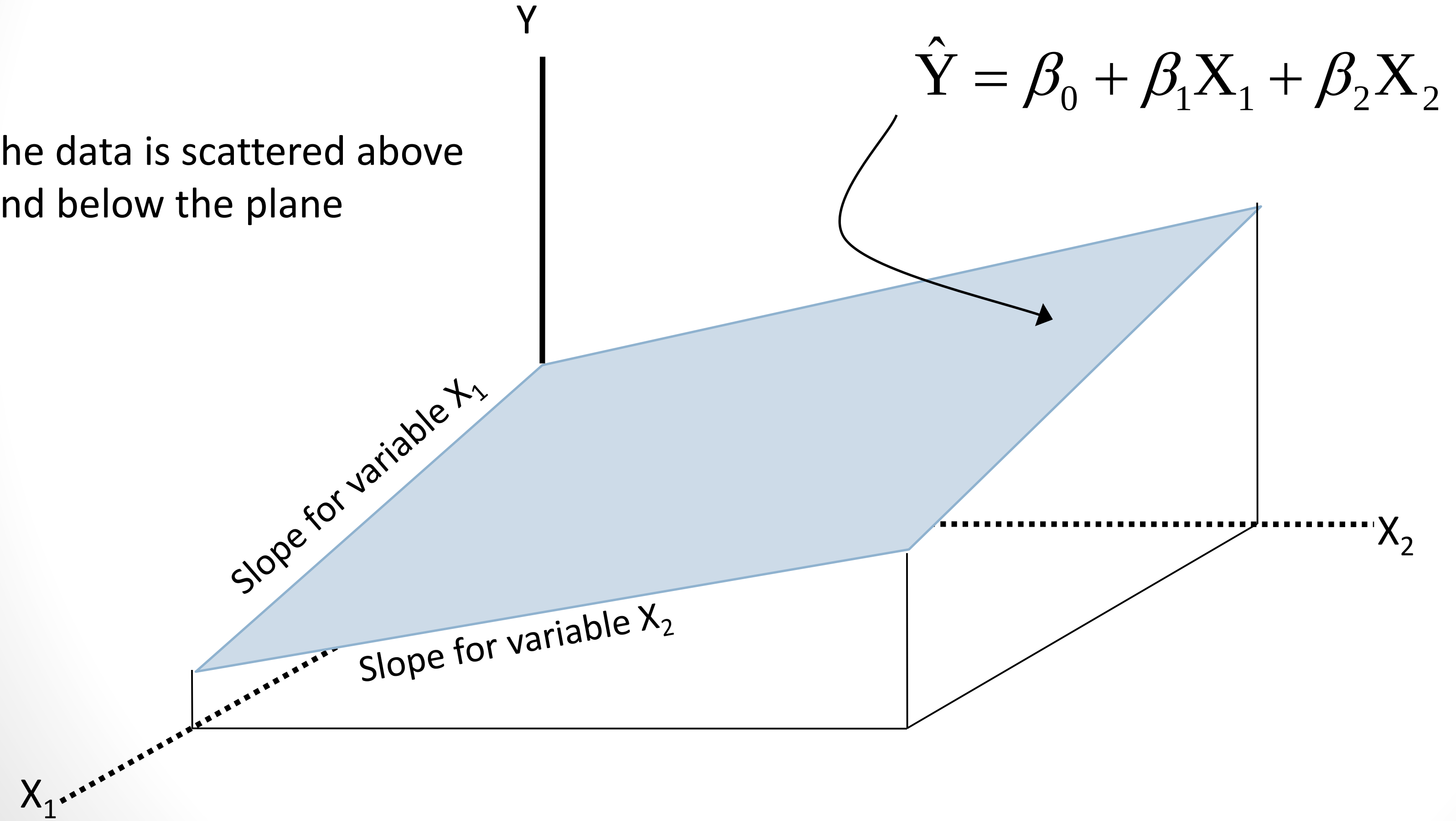
$$I = I\left(r, w, sk, t, D, P, \varepsilon\right)$$

Money demand

$$\frac{M}{P} = \frac{M}{P}\left(i, Y, c, E\right)$$

Multiple Regression Model

The data is scattered above and below the plane



Assumptions

Same as earlier

- The errors are normally distributed
- Errors have a constant variance
- The model errors are independent
- Errors (residuals) from the regression model:

  $ei = (Yi - Yi)$

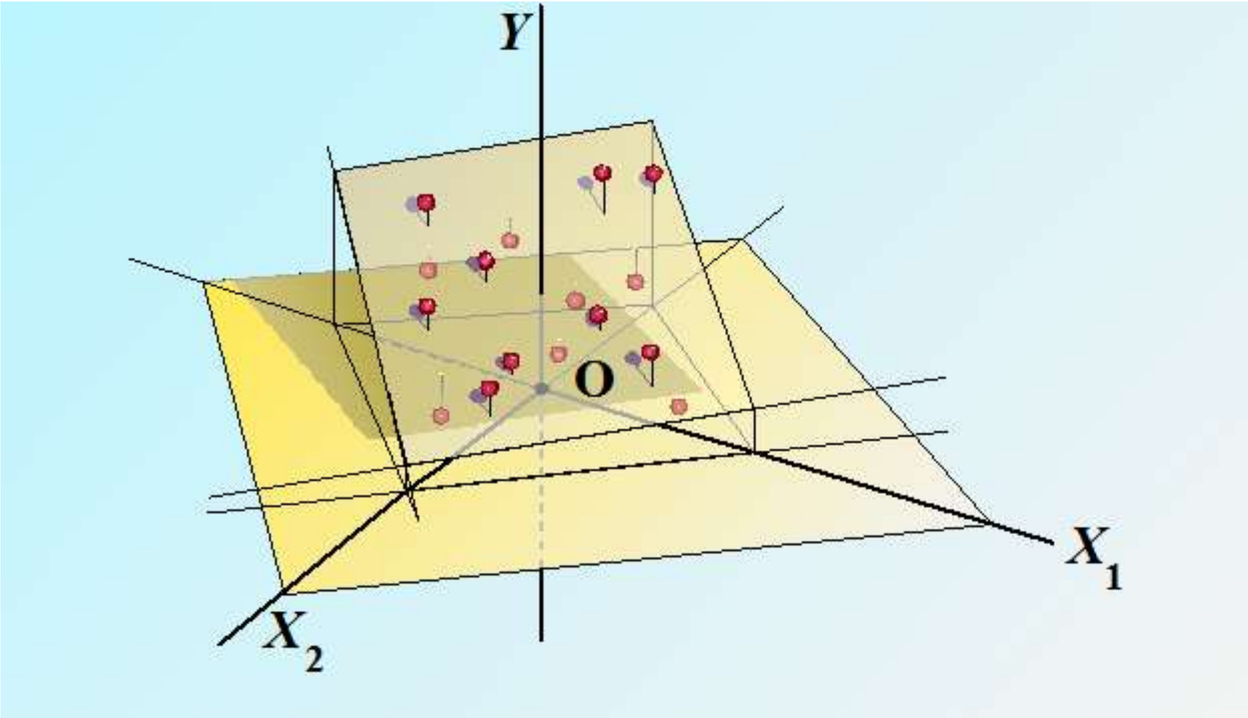
Least Squares Estimation

- The constant and parameters are derived in the same way as with the bi-variate model.
- Remember … “Minimizing sum of squares of deviation”?

- When a new variable is added it affects the coefficients of the existing variables

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3))^2\end{aligned}$$

Meaning of beta



- The equation $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ has the following interpretation.
- Again, β_0 is the intercept (the value of Y when both X_1 and X_2 are 0).
- β_1 is the slope for X_1, so each unit increase in X_1 increases Y on AVERAGE by β_1 units.
- β_2 is the slope for X_2, so each unit increase in X_2 increases Y on AVERAGE by β_2 units.

How good is my regression line?-Recap

- Take a regression line; Estimate y by substituting xi from data; Is it exactly same as yi?
- Remember no line is perfect
- There is always some error in the estimation
- Unless there is comprehensive dependency between predictor and response, there is always some part of response(Y) that can't be explained by predictor (x)
- So, total variance in Y is divided into two parts,
  - Variance that can be explained by x, using regression
  - Variance that can't be explained by x

Explained and Unexplained Variation-Recap

- Total variation is made up of two parts:

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

- Total sum of Squares

Sum of Squares Error

Sum of Squares Regression

$$\mathrm{SST} = \sum (y - \bar{y})^2$$

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

$$\mathrm{SSR} = \sum (\hat{y} - \bar{y})^2$$

SST :Measures the variation of the y_i values around their mean y

SSE : Variation attributable to factors other than the relationship between x and y

SSR: Explained variation attributable to the relationship between x and y

Coefficient of determination-Recap

- A good fit will have
  - SSE (Minimum or Maximum?)
  - SSR (Minimum or Maximum?)
  - SSR/SSE(Minimum or Maximum?)
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

$$\boxed{R^2 = \frac{SSR}{SST}} \quad \text{where} \quad \boxed{0 \le R^2 \le 1}$$

In the single independent variable case, the coefficient of determination is equal to square of simple correlation coefficient

Data Analysis Course
Venkat Reddy

Lab

- Download death rate data from here
    - The data (X1, X2, X3, X4, X5) are by city. Total 50 cities data
    - X1 = death rate per 1000 residents
    - X2 = doctor availability per 100,000 residents
    - X3 = hospital availability per 100,000 residents
    - X4 = annual per capita income in thousands of dollars
    - X5 = population density people per square mile
    - X6 = Number of cars per 500

SAS Data Set

- What is the mathematical equation between death rate & other variables
- What are the coefficient signs? Are they intuitively correct?
- How good is the regression line?
- For a city Doctor availability per 100,000 residents is 112, hospital availability per 100,000 residents is 316,annual per capita income in thousands of dollars is 10.39 , population density people per square mile is 106. What is the expected death rate?
- What happens to death rate when doctor availability increases?
- Download SAT score data from here
- What is the mathematical relation between SAT score and College GPA, High school GPA, Quality recommendation
- How good is the regression line?

SAS Data Set

Lab :R squared & Adj R squared

- Download the sample data from here
- Estimate x1 using x2 & x3 what is R squared
- Estimate x1 using x2, x3, x4, x5 what is R squared
- Estimate x1 using x2, x3, x4, x5,x6,x7 what is R squared

SAS Data Set

- X4, x5, x6,x7 are some random variables in this dataset

R squared & Adj R squared

- The proportion of total variation in Y explained by all X variables taken together (the model)

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

- R-squared never decreases when a new X variable is added to the model – True?

Adjusted R squared

- Its value depends on the number of explanatory variables
- imposes a penalty for adding additional explanatory variables
- It is usually written as (R-bar squared)

$$\overline{R}^2 = R^2 - \frac{k-1}{n-k}(1-R^2)$$

n-number of observations, k-number of parameters

R-squared vs adj R squared

- 18 variables
- N=20
- R-squared=.95
- What is adjusted R-sauared?
- What is your conclusion?

$$R_a^2 = 1 - \left[\frac{n-1}{n-(k+1)} \right] \frac{SSE}{SS_{yy}}$$

$$= 1 - \left[\frac{n-1}{n-(k+1)} \right] \left(1 - R^2\right) = 1 - \frac{19}{1}(.05) = .05$$

The F statistics

- Is the Overall Model Significant?
- F-Test for Overall Significance of the Model: Shows if there is a relationship between all of the X variables considered together and Y
- Use F test statistic; Hypotheses:
  - H0: β1 = β2 = … = βk = 0 (no relationship)
  - H1: at least one βi ≠ 0 (at least one independent variable affects Y)
- Test statistic:

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

Details Later

Multiple variables…good or bad?

- Multiple regression is wonderful in that it allows you to consider the effect of multiple variables simultaneously.
- Multiple regression is extremely unpleasant because it allows you to consider the effect of multiple variables simultaneously.
- The relationships between the explanatory variables are the key to understanding multiple regression.

Significance of coefficients?

- Check list
  - F-test for model is significant or not –F value
  - Does the model have the best available predictors for y? –Adj R
  - Are all the terms in the model important for predicting y? P value
  - Are all the predictor variables significant? P value
  - Note that when there is just one predictor, F-test test reduces to the F-test for testing in simple linear regression whether or not Beta1= 0

Significance testing of individual variables Beta

To test

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

Test statistic:

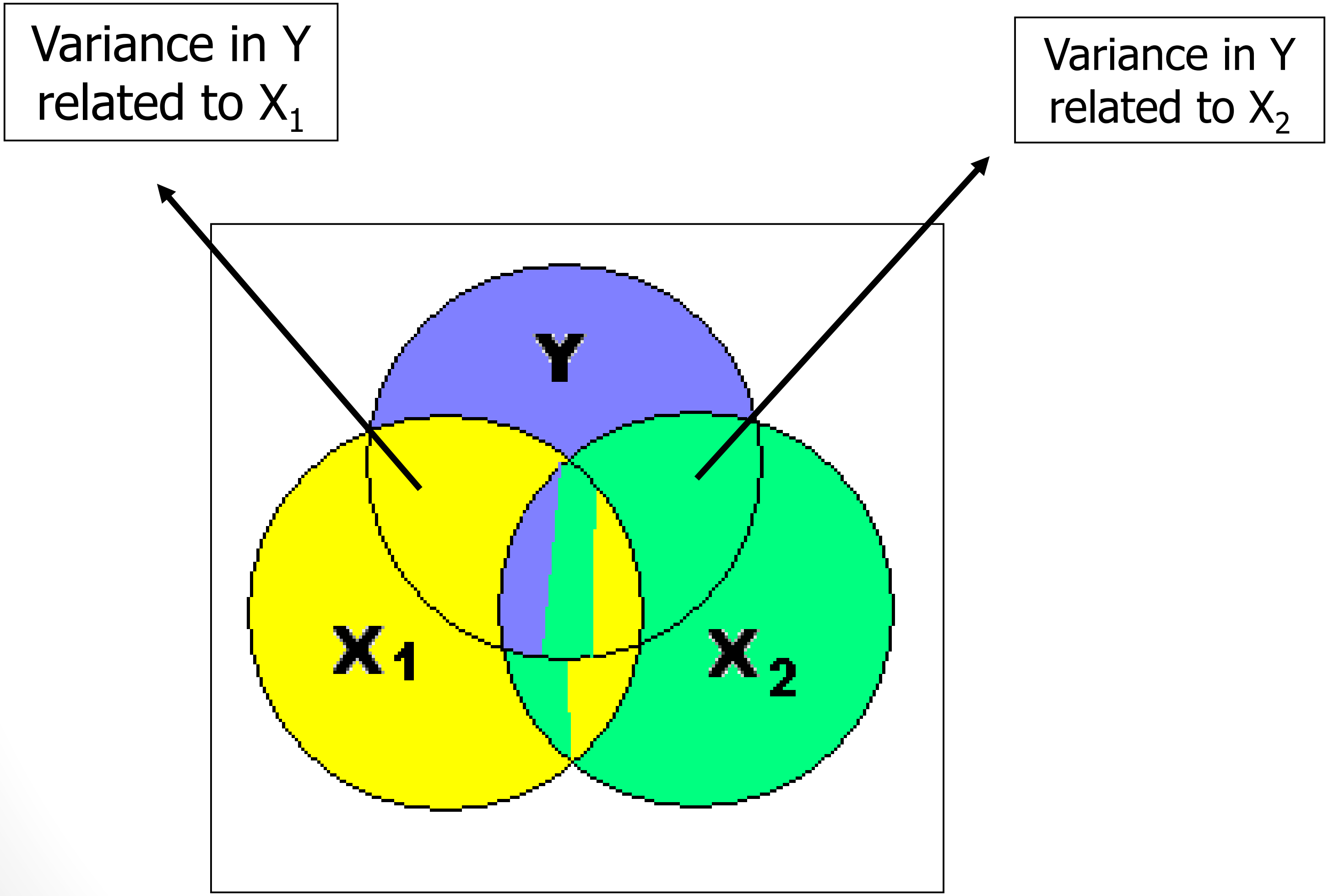
$$t = \frac{b_i}{s(b_i)}$$

Reject H0 if

$$t > t(\frac{\alpha}{2}; n - k - 1) \quad or$$

$$t < -t(\frac{\alpha}{2}; n - k - 1)$$

Variance explained by individual var



Variance in Y related to X_1

Variance in Y related to X_2

What is P-values of a coefficient

- P value gives us an idea about the significance of each variable
- The p-values for the individual variables are computed AFTER accounting for the effects of all the other variables.
- Thus, the p-value shown for the variable is the p-value “after accounting to everything else”.
- That p-value basically compares the
  - fit of the regression “with everything except to the variable” vs the fit of the regression “with everything including the variable”.
  - There will be no decremant or minimal change in adj r squared if we remove the variable
- Note it is possible all the variables in a regression to produce great individual fits, and yet have none of the variables be individually significant.

Contribution of a Single Independent Variable Xj

SSR(X_j | all variables except X_j)

= SSR (all variables) – SSR(all variables except X_j)

- Measures the contribution of X_j in explaining the total variation in Y (SST)

- Consider here a 3-variable model:

SSR(X_1 | X_2 and X_3)

$$= \underbrace{\text{SSR (all variablesX1-x3)}}_{\substack{SSR_{UR}\\ \text{Model}}} - \underbrace{\text{SSR}(X_2 \text{ and } X_3)}_{\substack{SSR_{R}\\ \text{Model}}}$$

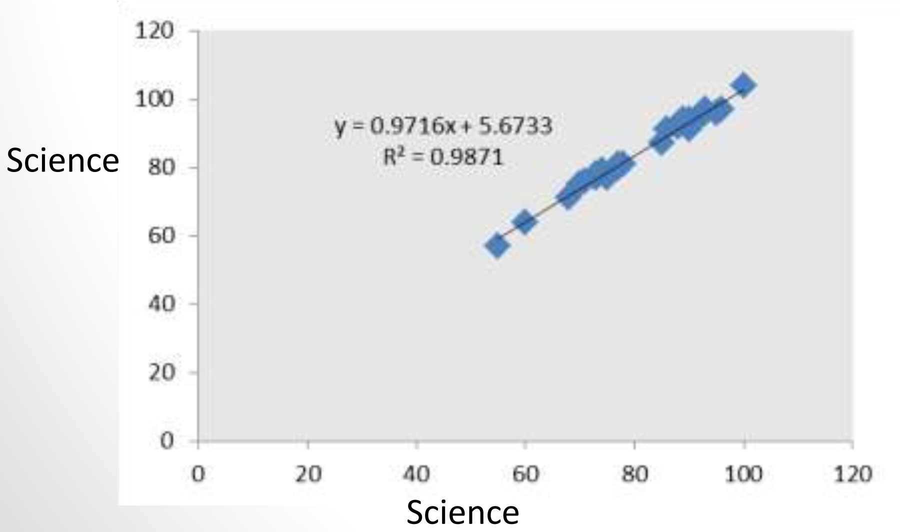
Lab

- In the death rate data what is the individual significance of effect of each variable?
- What is the least significant variable?
- With increase in number of cars does death rate increase or decrease?
- Remove the least significant variable and fit a new regression line
- What is new R squared & adjusted r squared?
- Download CAT exam data from here
- What is r square & adj R square
- What is the best predictor for CAT score?
- As mathematics score increases 10 units what happens to CAT score?
- Remove “scinece” variable from the model & see maths effect – Multicollinearity

SAS Data Set

Redundancy in the variables

- Remember the individual p-values indicate the significance of the variable AFTER all the other variables have been accounted for.
- It is possible that science (x1) score and maths score(x2) basically are providing the same information about Y.
    - Thus, after X_1, X_2 conveys little extra information. Similarly, after X_2, X_1 conveys little extra information.
- We can see this by plotting X_1 against X_2

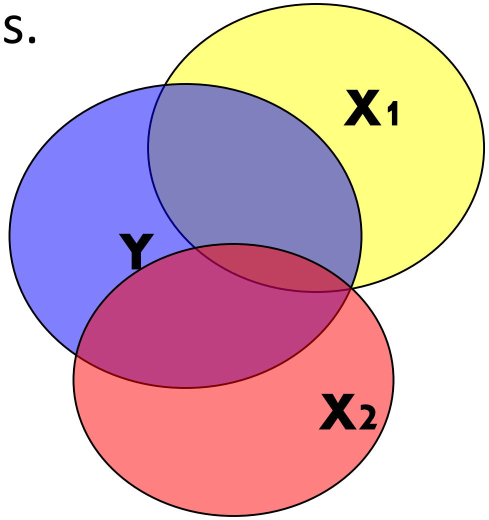


The explanatory variables are closely related to each other, hence each provides essentially the same information about Y.

Multicollinearity

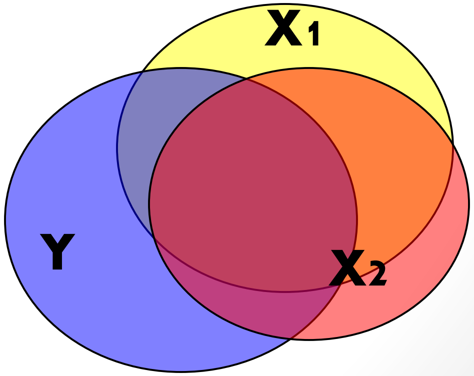
Multicollinearity (or inter correlation) exists when at least some of the predictor variables are correlated among themselves.

- When correlation among X's is low, OLS has lots of information to estimate b. This gives us confidence in our estimates of b. What is the definition of regression coefficient?



- When correlation among X's is high, OLS has very little information to estimate b. This makes us relatively uncertain about our estimate of b

- When the explanatory variables are closely related to each other, we say they are "collinear". The general problem is called multicollinearity.



Multicollinearity-Detection

- **VIF**=$1/(1-R_k2)$
- **High sample correlation** coefficients are sufficient but not necessary for multicollinearity.
- **A high F statistic or R squared** leads us to reject the joint hypothesis that all of the coefficients are zero, but the individual t-statistics are low. (why?)
- One can compute the **condition number**. That is, the ratio of the largest to the smallest root of the matrix x'x. This may not always be useful as the standard errors of the estimates depend on the ratios of elements of the characteristic vectors to the roots.

Multicollinearity-Effects

- **Effects**
  - Even in the presence of multicollinearity, OLS is BLUE and consistent.
  - Counter institutive coefficients
  - **Standard errors of the estimates tend to be large**: Large standard errors mean large confidence intervals. Large standard errors mean small observed test statistics. The researcher will accept too many null hypotheses. The probability of a type II error is large.(Any easy way to understand this?)
  - Estimates of standard errors and parameters tend to be sensitive to changes in the data and the specification of the model.

Multicollinearity-Redemption

- **Drop the troublesome RHS variables:** easy & most used method
- **Principal components estimator:** This involves using a weighted average of the regressors, rather than all of the regressors.
- **Ridge regression technique:** This involves putting extra weight on the main diagonal of x'x so that it produces more precise estimates. This is a biased estimator.
- **Use additional data sources**. This does not mean more of the same. It means pooling cross section and time series.
- **Transform the data.** For example, inversion or differencing.
- **Use prior information** or restrictions on the coefficients.

Lab

- Find the vif for each of the variables in Cat score data - use vif option in SAS
- Drop the troublesome variable and build a new line
- As mathematics score increases 10 units what happens to CAT score?
- Remove maths score and build a model.
- As science score increases 10 units what happens to CAT score?

Stepwise Regression

- Forward selection
- Backward elimination
- Stepwise regression

- Details later

Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

+91 9886 768879