Statistical Data Mining
MATH 4720
Lecture 4
Multivariate Analysis

Gaurav Gupta

GEH A401

Types of Analysis

- **Univariate Analysis:**
  The examination of the distribution of cases on only **one variable** at a time (e.g. weight of college students)

- **Bivariate Analysis:**
  The examination of **two variables** simultaneously (e.g. the relation between gender, race and weight of college students)

- **Multivariate Analysis:**
  The examination of **more than two variables** simultaneously (e.g. the relationship between gender, race and weight of college students)

Simple Regression Analysis

- The Simple Regression model, relates one predictor and one response.
- Let n observations $(x_1,\ y_1),\ (x_2,\ y_2) \ldots \ldots (x_n,\ y_n)$ pairs of predictors and responses such that $e_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d (independent and identically distributed). For fixed numbers β_0 and β_1 (parameters), the model as follows:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- The fitted model is as follows: $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$

- The estimated parameter are $\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ and $\widehat{\beta_0} = \hat{y} - \widehat{\beta_1}\ \bar{x}$, such that $\bar{x}$ and $\bar{y}$ are the sample averages.

Multivariate Regression Analysis

- **Multivariate Regression** is a method used to measure the degree at which more than one independent variable (**predictors**) and more than one dependent variable (**responses**), are linearly related.

- The method is broadly used to predict the behavior of the response variables associated to changes in the predictor variables, once a desired degree of relation has been established.

Exploratory Question: Can a supermarket owner maintain stock of water, ice cream, frozen foods, canned foods and meat as a function of temperature, tornado chance and gas price during tornado season in June?

Multivariable and Multivariate Regression

While the multivariable model is used for the analysis with one outcome (dependent) and multiple independent (predictor or explanatory) variables, multivariate is used for the analysis with more than 1 outcomes (eg, repeated measures) and multiple independent variables.

However, the terms are sometimes used interchangeably in the literature as not many researchers are attentive to the distinction. The difference between these two terms was brought to attention by Hidalgo and Goodman in 2013.

Hosmer Jr DW , Lemeshow S, , Sturdivant RX. *Applied Logistic Regression* . Hoboken, NJ : John Wiley & Sons ; 2013

.

Multivariate Regression Model

Let $\boldsymbol{Y}$ be the $n \times 1$ response vector, $\boldsymbol{X}$ be an $n \times$ (q+1) matrix such that all entries of the first column are 1's, and q predictors. Let $\boldsymbol{e}$ be an n×1 vector such that $e_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d (independent and identically distributed), and β be an (q+1)×1 vector of fixed parameters. The model is as follows

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$$

- $\boldsymbol{e}$: nx1 Error term
- $\boldsymbol{Y}$: nx1
- $\boldsymbol{X}$: nx(px1) Data Matrix/ Design Matrix
- β: (px1)x1 Regression Coefficients

In detail notation we have:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & . & . & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & . & . & . \\ 1 & X_{3,1} & X_{3,2} & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & X_{n,1} & X_{n,2} & . & . & X_{n,p} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \qquad \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + \cdots + \beta_p X_{1,p} + \epsilon_1 \\ \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + \cdots + \beta_p X_{2,p} + \epsilon_2 \\ \beta_0 + \beta_1 X_{3,1} + \beta_2 X_{3,2} + \cdots + \beta_p X_{3,p} + \epsilon_3 \\ . \\ . \\ . \\ \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + \cdots + \beta_p X_{n,p} + \epsilon_n \end{bmatrix}$$

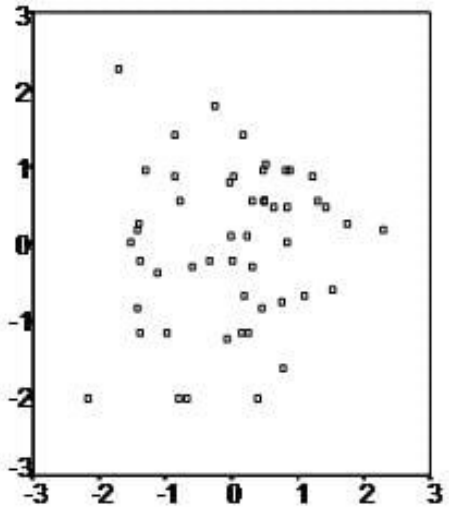
Assumptions

1. Linearity and additivity
2. Homoscedasticity or equal y-variance across the values of x
3. Uncorrelated error terms
4. Normality of the error

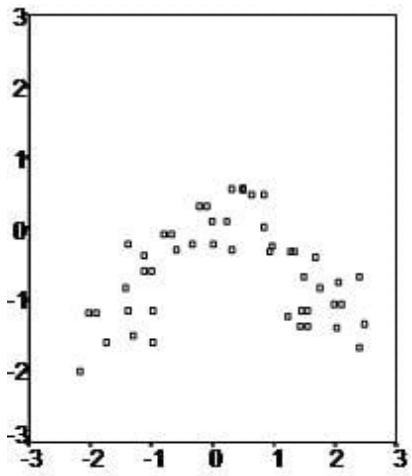
1. Linearity & additivity

The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.

The effects of different independent variables on the expected value of the dependent variable are additive.



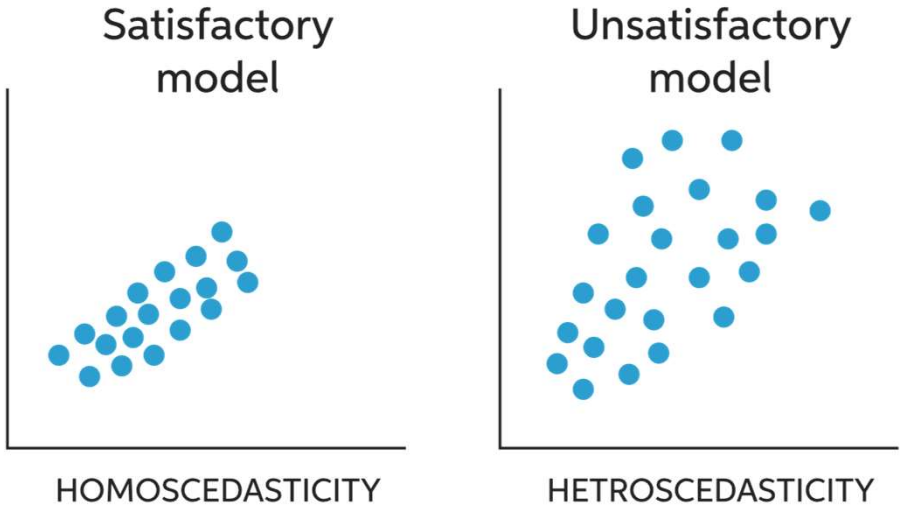
Curvilinear relationship



Linear relationship

2. Homoscedasticity (Constant variance)

This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.



$$\sigma_{y_1}^2 = \sigma_{y_2}^2 = \sigma_{y_3}^2 = \ldots\ldots\ldots = \sigma_{y_n}^2 = \sigma^2$$

3. Uncorrelated error terms

- The model must have no serial correlation (also known as autocorrelation) in the error terms.
- The problem of serial correlation in the error terms is that estimated standard errors will be wrong.

- For example, with a positive serial correlation in the error terms, standard errors will be too low, which means you will tend to reject the null hypothesis too often.
-

  Definition
  Conditional on X, the error term (ϵ) in two different time periods are uncorrelated, for all t different froms: Corr[ϵ_i, ϵ_k|X] = 0

4. Normality of the error

- Each error is random. Multiple regression assumes that the errors are normally distributed.
- The error term ϵ is independent of the independent variables and is normally distributed with zero mean and variance
- $\epsilon_i \sim N(0, \sigma^2)$

MSE in Matrix Form

MSE is calculated by summing the squares of $\boldsymbol{e}$ from all observations and dividing the sum by number of observations in the data table

$$\boldsymbol{where\ e = Y - X\beta}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{i=n} e_i^2$$

$$\sum_{i=1}^{i=n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + e_4^2 + ... + e_n^2 = [e_1\ e_2\ e_3\ e_4\ \dots\ e_n] \times \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ . \\ . \\ . \\ e_n \end{bmatrix} = e^T e$$

- Replace $\boldsymbol{e}$ with $\boldsymbol{Y - X\beta}$

$$MSE = \frac{1}{n}(Y - X\beta)^T(Y - X\beta)$$

$$= \frac{1}{n}(Y^T - \beta^T X^T)(Y - X\beta)$$

$$= \frac{1}{n}(Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta)$$

Since, $\beta^T X^T Y = (Y^T X\beta)^T$ and $Y^T X\beta$ being a 1×1 matrix, $Y^T X\beta$ can be replaced with $\beta^T X^T Y$

$$MSE = \frac{1}{n}(Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta)$$

MSE equation is used as cost function (objective function in optimization problem) which needs to be minimized to estimate best fit parameters in our regression model.

Gradient needs to be estimated by taking derivative of ***MSE*** function with respect to parameter vector β and to be used in gradient descent optimization.

Gradient of MSE

As mentioned above, gradient is expressed as

$$\nabla MSE = \frac{1}{n}(\nabla Y^T Y - 2\nabla \beta^T X^T Y + \nabla \beta^T X^T X\beta)$$

Where, ∇ is the differential operator used for gradient. Using matrix. differentiation rules, we get following equations.

$$= \frac{1}{\mathrm{n}}\left(0 - 2x^T Y + 2X^T X\beta\right)$$

$$= \frac{2}{\mathrm{n}}(x^T X\beta + X^T Y)$$

The above matrix is called ***Jacobian*** which is used in gradient descent optimization along with learning rate (***lr***) to update model parameters.

$$\boldsymbol{J(\beta) = \frac{2}{n}(X^T X\beta - X^T Y)}$$

Gradient Descent Method

- The formula for gradient descent method to update model parameter is shown below.

$$\beta_{new}=\beta_{old} - lr \times J(\beta)$$

- β_{old} is the initialized parameter vector which gets updated in each iteration and at the end of each iteration β_{old} is equated with β_{new}. ***lr*** is the learning rate which represents step size and helps preventing overshooting the lowest point in the error surface. The iteration process continues till ***MSE*** value gets reduced and becomes flat.

Exercise:

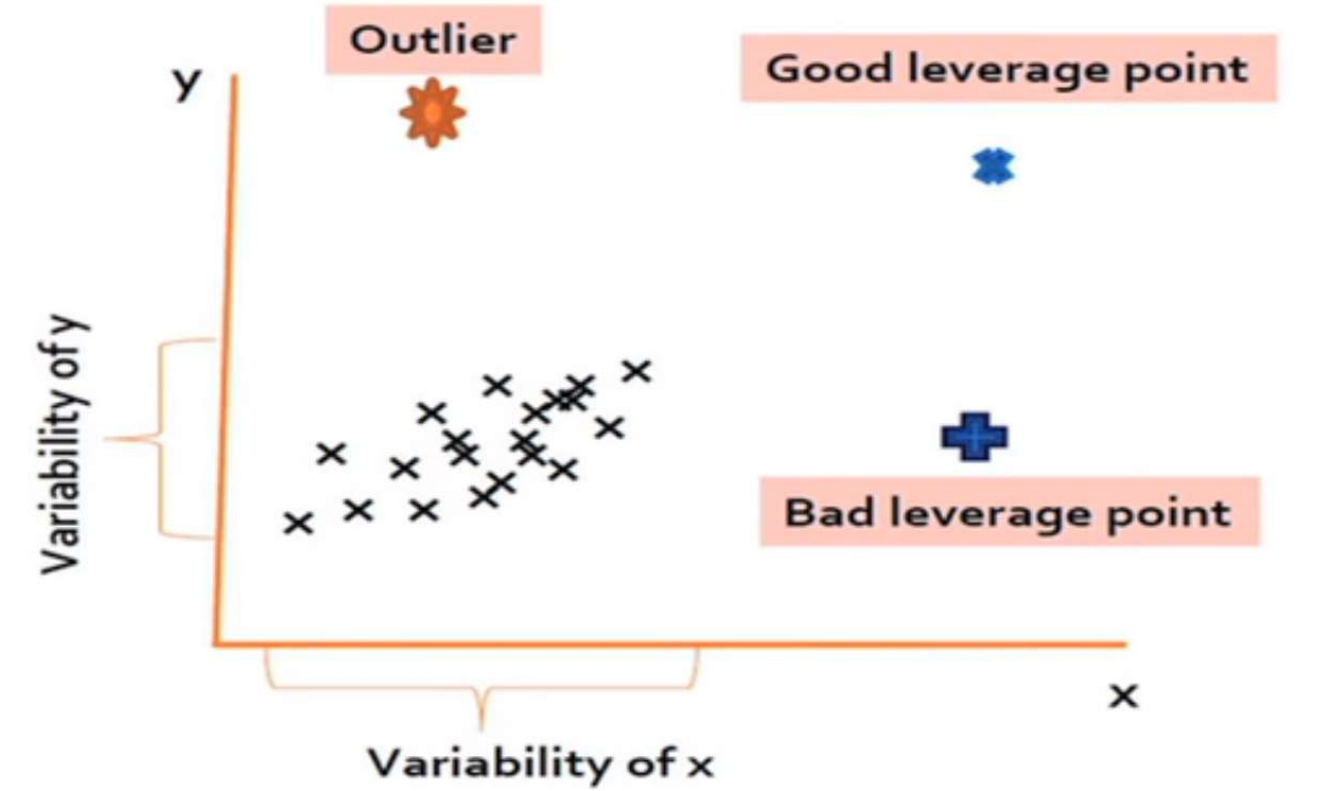
- You can develop a Multivariate regression model using example data set.
- Use Gradient descent method to estimate model parameters ***a***, ***b***, ***c*** and ***d***.
- The values of the matrices ***X*** and ***Y*** are known from the data whereas β vector is unknown which needs to be estimated.

- Initially, ***MSE*** and gradient of ***MSE*** will be computed followed by applying gradient descent method to minimize ***MSE***.

Leverage points

Diagnostics issues include

1. Finding out leverages
2. Finding out influential observations
3. Detecting and remedying multicollinearity



Model diagnostic of MLR

$H = X(X^TX)^{-1}X^T$ ← X-space

$$= \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & & & \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}$$

i	h_{ii}
1	h_{11}
2	h_{22}
3	⋮
⋮	
n	h_{nn}

$h_{ii} = 1, 2, \cdots n$ measure the leverage values of observation $i = 1, 2, \cdots n$.

$\sum_{i=1}^{n} h_{ii} = p+1$

$h_{ii} \cong (p+1)/n$ if each obvs contributes equally

$$\frac{(h_{ii} - 1/n)/p}{(1-h_{ii})/n-p-1}$$ follows $F_{p, n-p-1}$

~~$F^{\alpha=0}$~~:

$$F^{\alpha=0.05}_{p>10,\ n-p-1>50} < 2$$

So Cut off for leverage point $> \frac{2(p+1)}{n}$

Identification of leverages: Cook's distance

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p s_e^2}, \; i = 1, 2, \ldots, n$$

$$D_i = \frac{r_i}{p} \frac{h_{ii}}{1 - h_{ii}}, \; i = 1, 2, \ldots, n$$

$$r_i = \frac{e_i}{\sqrt{s_e^2 (1 - h_{ii})}}, \; i = 1, 2, \ldots, n$$

$$D_i \sim F_{p, n-p-1}$$

Cut off: Di > 1

$$D_{10} = 0.52 < F_{2,9}(0.25) = 1.62$$

COOK's D
0.000877
0.131387
0.147671
0.025554
0.030221
0.022519
0.012249
0.002592
0.000014
0.520644
0.007862
0.102077

Conclusions: No influential observations

Hessian Matrix:

- The key to the standard errors is the Hessian matrix. The variance-covariance-matrix of the coefficients is the inverse of the Hessian matrix. So the standard errors are the square root of the values on the diagonal of the inverse Hessian matrix

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix},$$