LECTURE 1 Linear Modelling

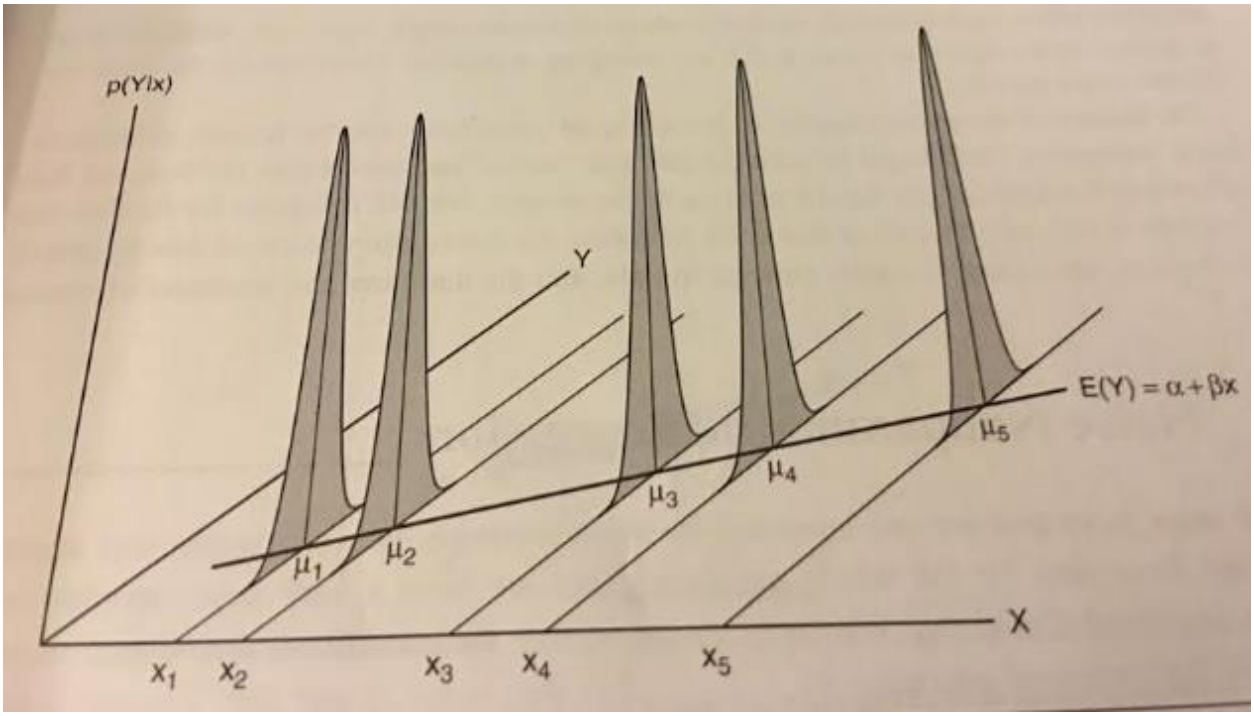
Objectives:

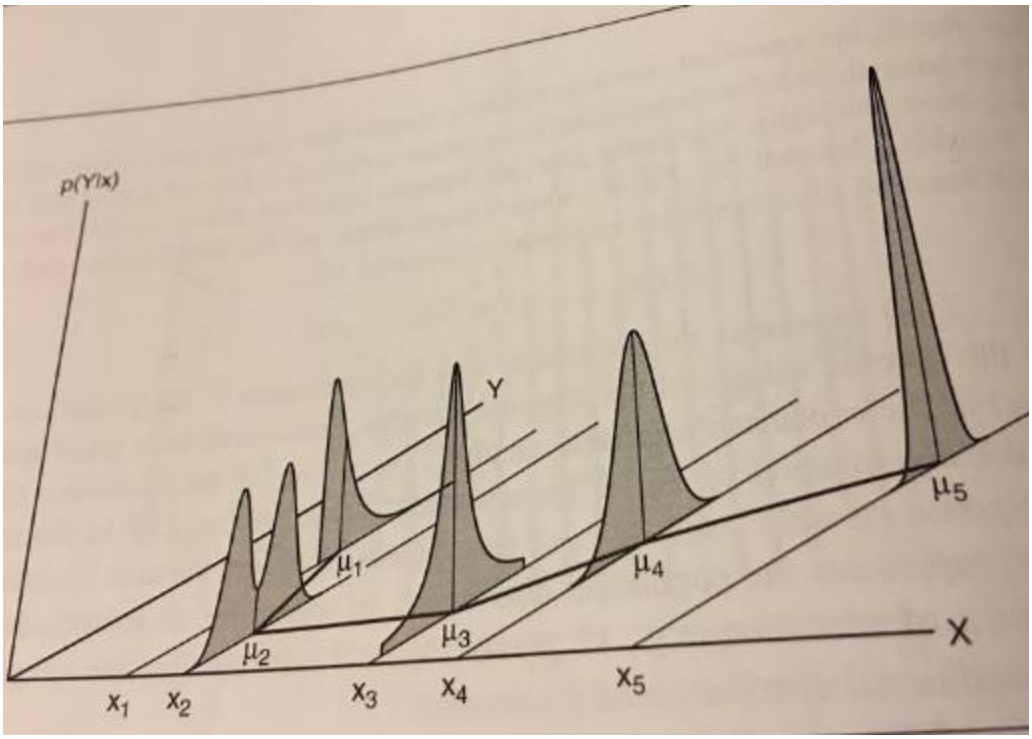
- *Regression definition and assumptions*
- *Naïve parametric regression, local averaging*
- *lowess regression.*

Introduction to Regression Analysis

- Regression is ascertaining the distribution of response variables as a function of one or more explanatory variables or
- Regression is ascertaining the distribution of response variables conditional on a set of one or more explanatory variables
- $p(y \mid x_1, x_2 \ldots\ldots\ldots x_k)$ Probability of observing a certain value of the response variable conditional to a set of explanatory variables $(x_1, x_2 \ldots\ldots\ldots x_k)$.
- $p(Y \mid x_1, x_2 \ldots\ldots\ldots x_k)$ Probability Distribution of Y conditional to a set of explanatory variables $(x_1, x_2 \ldots\ldots\ldots x_k)$.
- Regression is the relationship in which both the response and explanatory variables are quantitative.
- Qualitative variables can also be manipulated to apply regression techniques.

Regression Assumptions :





See Figure 2.3 and 2.4 Page 16 Example Income (Continuous response variable) earned for different Educational levels (Discrete explanatory variable).

- Normality: Conditional distribution of response variable depicted by $p(Y \mid x_1, x_2 \ldots\ldots\ldots x_k)$ is a Normal distribution. Skewness will violate this assumption for any particular value of the explanatory variable. Sample Mean of Y distribution will not be an appropriate measure of central tendency for the particular explanatory value.
  Multi modal data cannot be represented by sample mean as a measure of central tendency.
  Heavy Tails which is a non normal will not allow the sample mean to be a good estimator of the center of the Y distribution for that value of explanatory value even though the distribution is symmetric.

- Equality of Variance: Variances of the conditional distribution are the same regardless of the values of $(x_1, x_2 \ldots\ldots\ldots x_k)$. If $p(Y \mid x_1)$ as compared to $p(Y \mid x_2)$ has a different spread then the efficiency of least squares will be compromised.
  The expected value of Y is a linear function of Ys

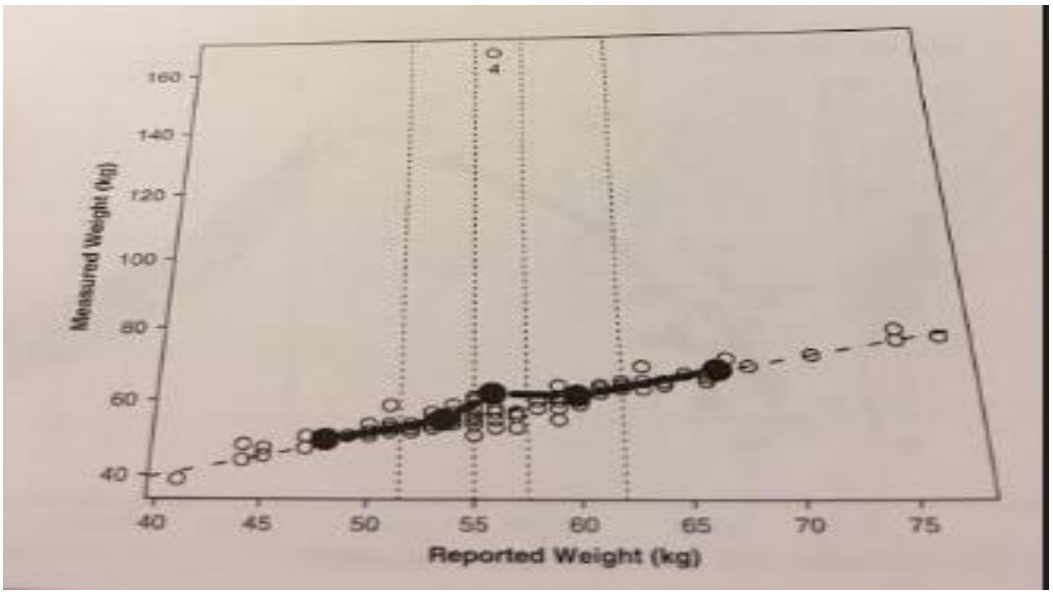
  $$\mu = E(Y \mid x_1, x_2 \ldots\ldots\ldots x_k) = \alpha + \beta_1 x_1 + \ldots\ldots\ldots \beta_k x_k$$

- Linearity: Non linearity assumption should be checked especially when there are multiple explanatory variables.
- The ideal condition of normality, equality of variance, linearity is the basis of Linear Least Square Estimation.
- Practically knowing these assumptions for Linear least square estimation statisticians transform data to conform to these assumptions in order to obtain credible analysis.

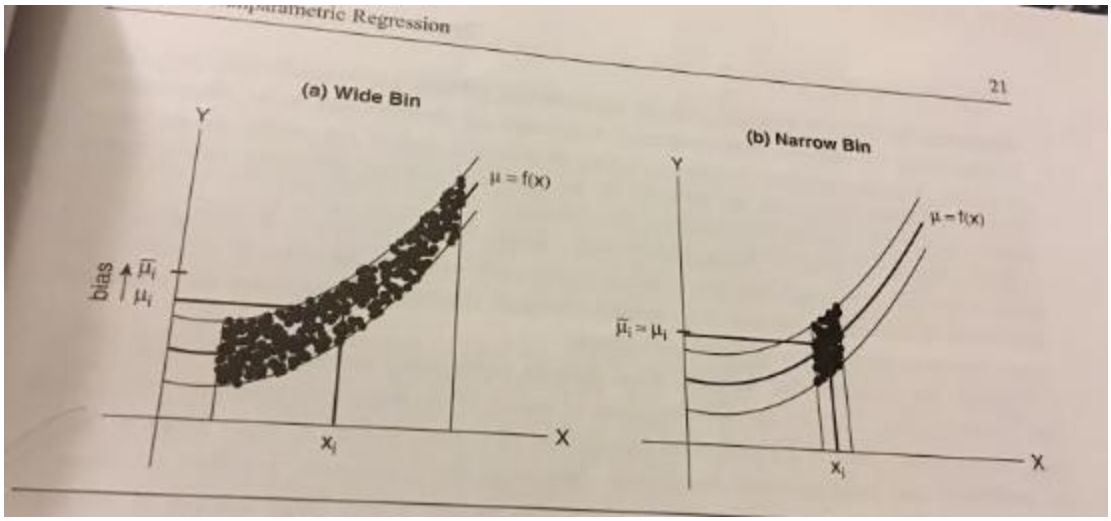
- In order to not deal with the difficulties of Linear regressions and its restrictions non parametric regression methods can be employed.

Naïve Non Parametric Regression:

See Figure 2.5 and 2.6 Page 18-19 Example Measured Weight (Continuous response variable) for different reported weight (Continuous explanatory variable). Objective here will be to predict the measured weight variable given a reported weight.



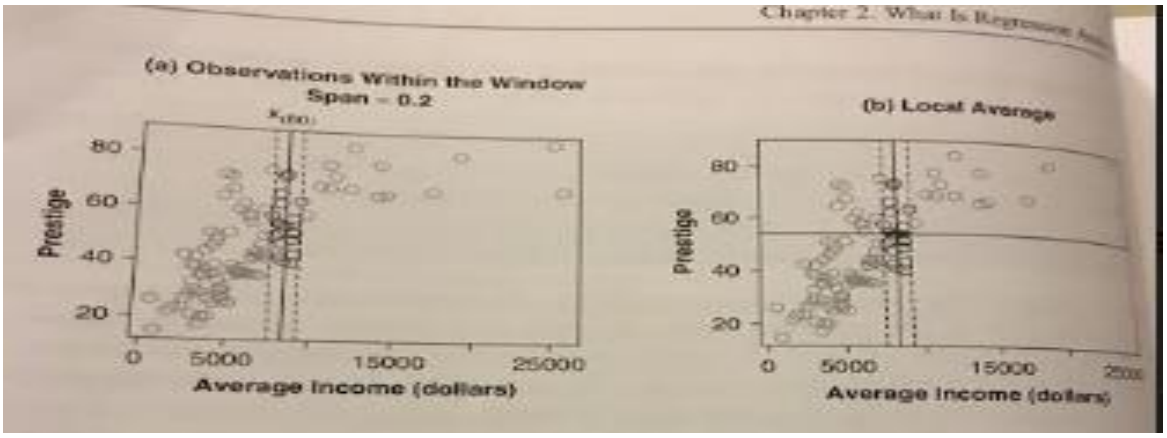
- The prediction of Y given X requires ascertaining the mean value of Y as a function X for the random sample collected.
  $\mu = E(Y \mid X) = f(x)$
- The explanatory variable is continuous therefore even for a large sample it is difficult to obtain replicated values of X (explanatory variable) therefore the conditional distribution of Y (response variable) given different values of X cannot be explored. The conditional means for response variable Y cannot be ascertained for specific values of explanatory variables X.
- The ranges of X (explanatory variable) can be divided into bins. For each bin the conditional mean of Y (response variable) can be ascertained.
- Large sample entails larger number of narrower bins wheras smaller sample requires smaller number of bins with broader bins. The bins need not have equal width.
- The non parametric regression line drawn in each bin is created by the conditional response means $\overline{Y}$ and explanatory variable $\overline{X}$



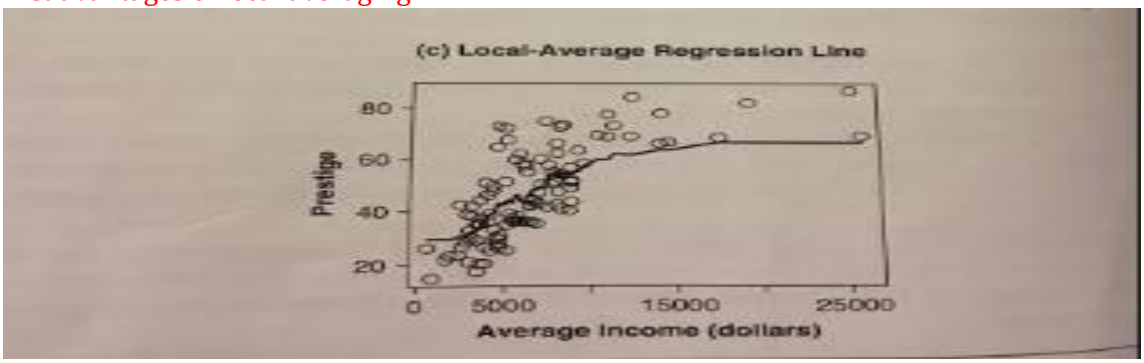
- The sources of errors that can occur in binning and averaging are: sampling error and bias.
  Sampling error or variance: The conditional means of the response variable $\overline{Y}$ will change if for different samples. Variance will be minimized if small number of wide bins are used. If you have a narrow bins the small variations will be captured and magnified wheras for a small number of wider bin the variance will be smaller. The fluctuation of the data will be seen for narrower bins. Too much of the noise is seen in the data. This is called overfitting of the model onto the data.

  Bias: Is the how far the mean of the bin from the central value mean(mean of the whole data set). The wider the bin the far removed these values will be. To minimize bias the bins should be large in number and as narrow as possible. If we have very wide bins then the mean of each bin will be further from the mean of the data. This is called underfitting of the model onto the data.
- Tradeoff between Variances and Bias have to be balanced by statisticians. http://scott.fortmann-roe.com/docs/BiasVariance.html
- The weakness in Naïve nonparametric is that the regression lines are crude and regression is evaluated at a very small number of values.
- Ideally a large sample which is divided into large number of bins and each bin consisting of large number of values will minimize bias and variance and this naïve regression curve will be a consistent estimator of the population regression curve.
- The problem of binning and averaging increases when there are more than one explanatory variables. If we have four explanatory variables and each variable has 10 discrete values then there are 10^4 combinations a very large sample will be required to calculate conditional mean of Y.

Local Averaging:



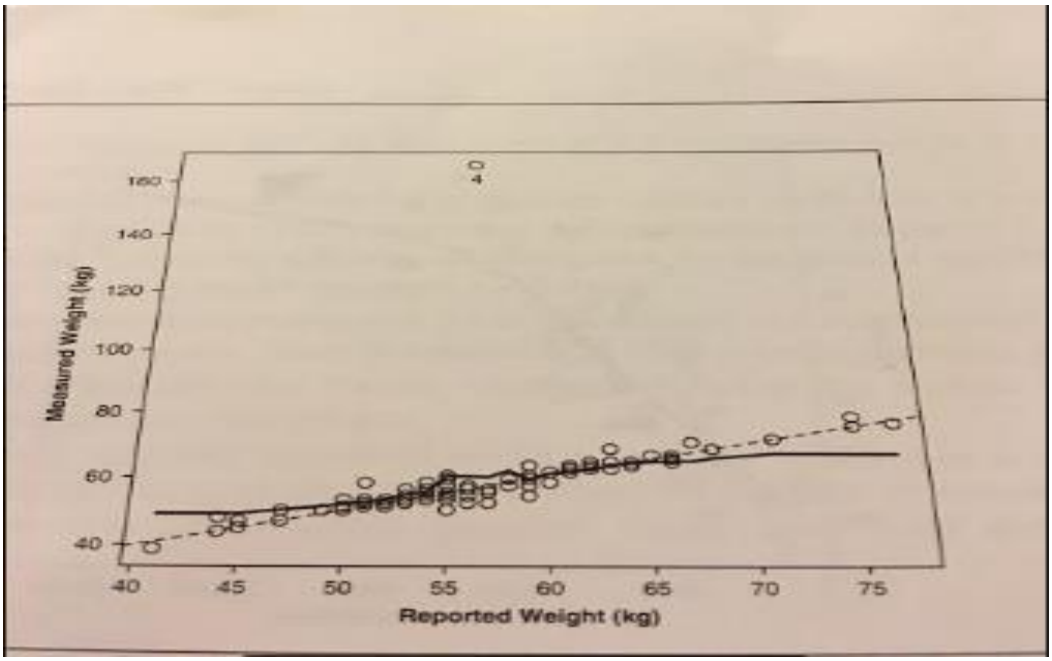
- Local averaging by employing computer software can be made smoother by using overlapping bins or calculating the conditional means for response variable Y across moving bin windows of fixed length with the focus as the centered x value. The variable length windows of x can be used for Nearest neighbours of x. The span of the local averaging s=m/n (Fixed number of values an NN/Total values).
- For small samples the local average can determine the mean Y values for each x value. For a larger data we can find local averages at representative x values.
- Disadvantages of local averaging



Boundary Bias: If the first few local /and or the last few local averages are similar it flattens the estimated regression line which creates boundary bias.
Rough Jumps: If the relationships between variable is weak then the average jumps up and down as the bin window progresses for different x focal values.
Outliers: Unusual data values effect the averages if they are within the current bin window.

LOWESS:

- Lowess (Locally weighted Regression) is a local averaging smoothing techniques to overcome the disadvantages of local averaging. Lowess uses the fitted value computed by locally weighted least square line giving more weightage to neighboring values from the focal x as compared to far off values.
- The bias and variance tradeoff is still valid for Lowess. The smaller the span the smaller the bias and larger the variance and viseversa.
- Larger spans creates smoother regression curves.



- Lowess as compared to Local averaging results in smoother regression curves, reduces boundary bias and also reduces the effect of outliers.