Stat 525 Regression Analysis

Lecture 1 : Linear Regression with One Predictor Variable

Zheni Utic
Department of Mathematics and Statistics, University of Massachusetts Amherst

Outline

- Relations between Two Variables
- Regression Models and Their Uses
- Simple Linear Regression Model with Distribution of Error Terms Unspecified
- Overview of Steps in Regression Analysis
- Point estimation of $E(Y) = \beta_0 + \beta_1 X$
- Point estimation of $\sigma^2(Y) = \sigma^2$
- Simple Linear Regression Model with Normal Distribution Error Terms

1.1 Variable Types and Relations between Two Variables

Dependent vs. Independent

- Independent variable(X) : predictor, explanatory variable

  · manipulated or changed by the experimenter

  · influences the dependent variable

- Dependent variable(Y) : response variable, outcome variable

  · observed result of the independent variable being manipulated

  · we want to predict

  (e.g.) A call center where the number of customers serviced per hour, depends on the number of agents, and average service time per customer.

Quantitative vs. Qualitative

- Quantitative variable

  · naturally measured as a number for which meaningful arithmetic operations make sense.

  · discrete variable and continuous variable

- Qualitative variable : categorical Variable

  · have no numerical meaning and take a value that is one of several possible categories

If X is an independent and quantitative variable and Y is a dependent and quantitative variable,

- Functional Relation : $Y = f(X)$
- Statistical Relation : $Y = f(X) + \epsilon$ where ϵ is an (random) error term

  · variation in Y that is not associated with X and that is considered to be of a random nature

  · all data points do not fall directly on the line of relationship

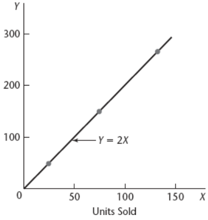


Figure 1.1(KNN)

Y(dollar sales of a product)

X(# of units sold)

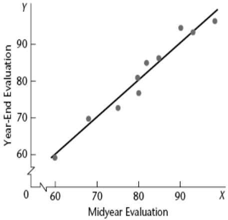


Figure 1.2(KNN)

Y(Year-end evaluation)

X(midyear evaluation)

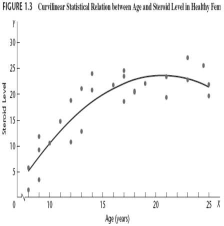


Figure 1.3(KNN)

Y(level of a steroid in plasma)

X(age)

1.2 Regression Models and Their Uses

1) Purpose of regression models

- determine the magnitude of the (typically imperfect) relationship between Y and a set of Xs
- predict Y from a set of Xs

2) Basic concepts

- A tendency of Y to vary with X in a systematic fashion
- A scattering of points around the curve of statistical relationship

  · Probability distribution of Y for each level of X : $f(Y \mid X = x)$

  · Regression function of Y on X, $E(Y \mid X) \equiv \int y\, f(Y \mid X)dy$: the means of these probability distributions of Y vary in some systematic fashion with X and it is a function of X

  · $Y = f(X) + \epsilon = E(Y \mid X) + \epsilon$

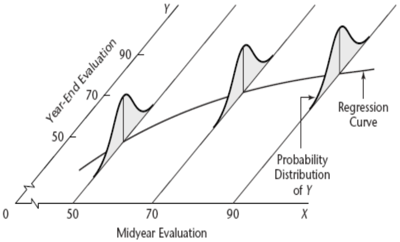


Figure 1.4(KNN)

3) Construction of Regression Models

- Selection of a set of "good" Xs
- Functional form of regression relation
- Scope of regression model
- Regression and Causality

4) Data for regression analysis

- Observational data from nonexperimental studies that do not control Xs of interest

  · no adequate information about cause-and-effect relationships

- Experimental data from experimental studies that control Xs of interest through randomization

  · stronger information about cause-and-effect relationships

  · randomization balancing out the effects of other predictors that might affect Y

1.3 Simple Linear Regression Model with Distribution of Error Terms Unspecified

$$Y_i = E(Y_i \mid X_i) + \epsilon_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, \ldots, n \tag{1}$$

- Assumptions

  · Y_i is the i-th value of the response variable

  · X_i is the i-th known value of the predictor variable (constant)

  · β_0 and β_1 are parameters (unknown constant) (regression coefficients)

  · ϵ_i is an uncorrelated random error term with $E(\epsilon_i) = 0$, $\sigma^2(\epsilon_i) = \sigma^2$ and $\sigma(\epsilon_i, \epsilon_j) = 0$

  So, $E(Y_i) =$, $\sigma^2(Y_i) =$, $\sigma(Y_i, Y_j) =$

  · simple : there is only one X (multiple : # of X in the model >1)

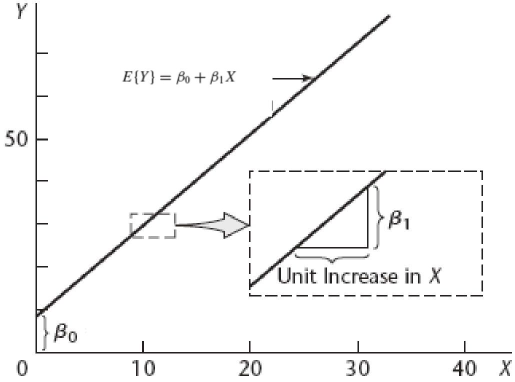
  · linear in the parameters

  · β_0, β_1 and σ^2 are the unknown parameters.

- Meaning of regression coefficients, β_0 and β_1

  · $\beta_1 =$ the slope (the change in the mean of the probability distribution of Y per unit increase in X)

  · $\beta_0 =$ the intercept (the mean of the probability distribution of Y at $X = 0$)



- Features

  · Y_i is a random variable (why?)

  · mean response (regression function), $E(Y_i) = \beta_0 + \beta_1 X_i$

  · $\sigma^2(Y_i) = \sigma^2$: each probability distribution of Y has the same variance σ^2

  · $\sigma(Y_i, Y_j)$: Y_i and Y_j are uncorrelated

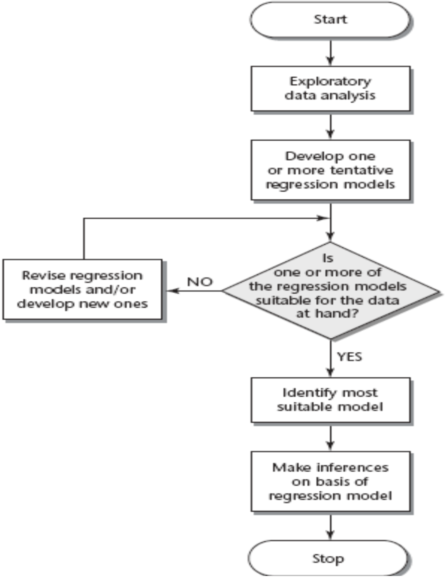
  $\rightarrow$ Y_i comes from probability distributions whose means are $\beta_0 + \beta_1 X_i$ and whose variances are σ^2, the same for all levels of X. In addition, Y_i and Y_j are uncorrelated.

- Alternative versions of $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

  · $Y_i = \beta_0 X_0 + \beta_1 X_i$ where $\beta_0 \equiv 1$

  · $Y_i = \beta_0^\star + \beta_1 (X_i - \bar{X})$ where $\beta_0^\star = \beta_0 + \beta_1 \bar{X}$

1.5 Overview of Steps in Regression Analysis



Simple linear regression Model

(e.g.) Plot of Y vs X

Point estimation for the model (Chapter 1)

Check the assumptions of the regression model (Chapter 3)

Confidence interval, Testing and Prediction (Chapter 2 and 4)

1.6 Estimation of β_0, β_1 and σ^2

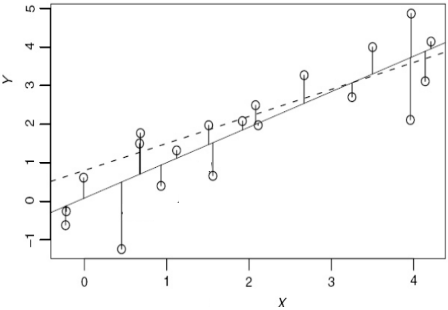
For the observations $(X_1, Y_1), \ldots, (X_i, Y_i), \ldots, (X_n, Y_n)$,

- Use the **method of least squares** to obtain estimators of β_0 and β_1

  [Idea] the estimators of β_0 and β_1 are those values b_0 and b_1, respectively, minimizing Q

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - E(Y_i))^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

  where $Y_i - \beta_0 - \beta_1 X_i$ is the deviation of Y_i from its expected value



- Least Squares estimators for β_0 and β_1 are

$$b_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})Y_i}{\sum_i (X_i - \bar{X})^2}, \quad b_0 = \bar{Y} - b_1\bar{X}$$

· How? solve $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$ simultaneously

- Meaning of b_1 and b_0

  (Study example) Suppose one is interested in the relationship between the number of hours (X) given for study and score on a test (Y). Given 20 observations (X_i, Y_i), a simple linear regression was applied, and β_1 and β_0 using the method of least squares were calculated : $b_1 = 3.5$ and $b_0 = 15.05$

  · students score ___ on average when they did not study

  · adding an additional hour to your study time will result in an average score of __ point higher

- Properties of b_0 and b_1

  · b_0 and b_1 are BLUE(Best Linear Unbiased Estimator)

- Point estimation of $E(Y) = \beta_0 + \beta_1 X$

  · Given b_0 and b_1, the estimated regression function at X is

$$\hat{Y} = b_0 + b_1 X \tag{2}$$

  so, $\hat{Y}_i = b_0 + b_1 X_i$ where $i = 1, \ldots, n$ (called as the i-th fitted value)

  (e.g.) In our (Study example), $\hat{Y} = 15.05 + 3.5X$. For a student studying 4 hours, the expected score on the exam is   .

- Residuals, $e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i$

  · vertical deviation of Y_i from the corresponding fitted value $\hat{Y}_i$

  · difference between $\epsilon_i = Y_i - E(Y_i)$ and $e_i = Y_i - \hat{Y}_i$

  · very very useful for studying an estimated regression model is suitable for the n observations (X_i, Y_i) (Chapter 3)

- Properties of e_i and $\hat{Y}_i$

  · $\sum_i e_i = 0$ and $\sum_i e_i^2$ is a minimum

  · mean of $\hat{Y}_i$ = $\bar{Y}$, i.e., $\frac{1}{n}\sum_i \hat{Y}_i = \frac{1}{n}\sum_i Y_i$

  · $\sum_i X_i e_i = 0$ and $\sum_i \hat{Y}_i e_i = 0$

  · the regression line goes through the point $(\bar{X}, \bar{Y})$

- Point Estimation of $\sigma^2(Y) = \sigma^2(\epsilon) = \sigma^2$

  · Y_i from different probability distributions with different means depending on X_i

  · deviation of Y_i from $\hat{Y}_i$: $e_i = Y_i - \hat{Y}_i$

  · $s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$ where MSE is residual mean square and SSE is residual sum of squares : an estimator for σ^2

  · $E(s^2) = \sigma^2$

  · $s = \sqrt{s^2}$ for the standard deviation $\sigma = \sqrt{\sigma^2}$

1.8 Simple Linear Regression Model with Normal Distribution Error Terms

- Method of least squares

  · only know $E(\epsilon_i) = 0$ and $\sigma^2(\epsilon_i) = \sigma^2$ (the distribution of the ϵ_i is unspecified)

  · b_1 and b_0 are BLUE for β_0 and β_1 in Eq. (1), and $s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$ is an unbiased estimator for σ^2

- One more assumption about the distribution of the ϵ_i in $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

  · need for interval estimators and hypothesis testing

  · $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma(\epsilon_i, \epsilon_j) = 0$ (uncorrelatedness implies independence between ϵ_i and ϵ_j.

  Then, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$,

  and $\sigma(Y_i, Y_j) = 0$

  (so, Y_i and Y_j are independent).

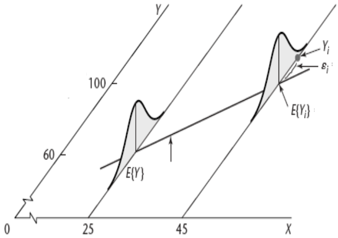


Figure 1.6 (KNN)

- Estimation of β_0, β_1 and σ^2 in $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ when $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma(\epsilon_i, \epsilon_j) = 0$

  · use Method of Maximum Likelihood

  [Idea] construct the likelihood function of β_0, β_1 and σ^2, $L(\beta_0, \beta_1, \sigma^2)$, and find values of the parameters maximizing the log of $L(\beta_0, \beta_1, \sigma^2)$, $\ell(\beta_0, \beta_1, \sigma^2)$

  : Since $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, $f(Y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$.
  Then the (log) likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(Y_i; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right]$$

$$\ell(\beta_0, \beta_1, \sigma^2) = \log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

  Then solve $\frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = 0$, $\frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = 0$ and $\frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = 0$ simultaneously.

- Estimators for β_0, β_1 and σ^2 and their properties

Parameter	Method of Least Squares	Method of Maximum Likelihood
β_1	$b_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$	$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$
β_0	$b_0 = \bar{Y} - b_1 \bar{X}$	$\hat{\beta}_0 = \bar{Y} - b_1 \bar{X}$
σ^2	$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}$	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}$

· $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased, sufficient and consistent

· $\hat{\beta}_1$ and $\hat{\beta}_0$ have minimum variance among all unbiased (linear or otherwise) estimators

· s^2 is unbiased, but $\hat{\sigma}^2$ is biased for finite n.

Calculate the regression coefficients and MSE with R

```
>x=c(-1,0,-2,-3)            #to introduce a variable
>y=c(-5,-4,2,-7)
>b_1=cov(x,y)/var(x)        #to calculatee b1
>b_1                        #to call the value of b1
[1] 0.2
> b_0=mean(y)-b_1*mean(x)   #to calculate b0
> b_0                       #to call the value of b0
[1] -3.2
>yhat=b_0+b_1*x             #to introduce the regression model
>yhat                       #to call the value of yhat
[1] -3.4 -3.2 -3.6 -3.8
> n=length(y)               #to introduce the number of observations
> n                         #to call the value of n
[1] 4
> MSE=sum(y-yhat)^2/(n-2)   #to introduce MSE=S^2 as a variable
> MSE                       #to call the value of MSE
[1] 9.860761e-32
```