Regression model - Problem of Inference

2.1 Nature of Regression Analysis

Introduction: Regression Analysis is considered indispensible as a data analysis technique in a variety of disciplines. Regression analysis has been described as a study of the relationships between one variable, the response variable, and one or more other variables the predictor variables.

Regression analysis is viewed in the context of data analysis rather than strictly in the classical model formation. Regression analysis consists of graphic and analytic methods for exploring relationships between one-variable referred to as a response variable(Y) and one or more other variables called regressors or predictor variables .

Regression analysis addresses how the predictor variables influence, describe, or control the response. The relationship is intended to be directional in that one ignores whether the response variable affects the predictors. (Relationships among variables in a correlation analysis are generally not directional. One usually does not desire to study how some variables responds to others, but rather how they are mutually associated).

The goal of regression analysis is to express the response variable as a function of predictive variables. Once such an expression is obtained the relationship can be utilised to predict values of the response variable, identify which predictor variables are significant or most affect the response variable

2.2: Distribution of OLSE and Testing the significance of the regressor:

For testing various hypothesis on the regression model need to know the distribution of the estimators used for fitting the model and also distribution of the test statistics.

One of the basic assumption in the regression model is that ϵ_i's are independent and identically distributed $N(0,\sigma^2)$. From equation (10), it is clear that $\hat{\beta}_{OLS}$ is a linear combination of ϵ_i's. A linear combination of normal random variables is itself a normal random variable.

Hence $\hat{\beta}_{OLS}$ is $N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$.

Similarly $\hat{\alpha}_{OLS}\ is\ N(\alpha,\ \frac{\sigma^2\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2})$

$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}, is \quad \chi^2\ with \quad (n-2)$ degrees of freedom. (13)

Therefore $Z = \frac{(\hat{\beta} - \beta)}{(\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2})^{1/2}}\ follows \quad N(0,1)$ and $V = (n-2)\hat{\sigma^2}/\sigma^2\ follows \quad \chi^2_{(n-2)}$

Then $t = \frac{Z}{V/(n-2)}\ follows \quad t - distribution with (n-2) df.$ (14)

Under H_0: β=0, which means there is no relationship between Y_i and X_i or the regressor is insignificant.

Under H_0: $t_{obs} = \frac{\hat{\beta}}{\sqrt{(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2})}} = \frac{\hat{\beta}}{SE(\hat{\beta})}$ (15)

If $|t_{obs}| > t_{\alpha/2,(n-2)}$, then reject H_0 at α% significance level.

2.3 Prediction:

Prediction: One of the major objective of regression analysis is prediction. Prediction is forecasting the value of the response variable for specified values of the regressors. For example consider the need for a municipality to be able to predict energy consumption. The primary focus in developing a prediction equation for energy consumption is that the resulting equation is able to predict accurately. This requires correct specification of the model and the accuracy of individual parameter estimates.

Best Linear Unbiased Predictor (BLUP):

Consider a simple linear regression model $Y_i = \alpha + \beta X_i + \varepsilon_i, i = 1,2,...n$ satisfying all the basic ideal conditions.

Prediction problem involves forecasting or predicting a value of the endogenous variable Y correcponding to a specified values of the regressor X say X_0 .

Definition: Let $\hat{Y}_0$ be a linear function of sample observations $Y_1, Y_2, ...,Y_n$. $\hat{Y}_0$ is said to be best linear unbiased predictor of Y at $X=X_0$, if it satisfies the following conditions

(i) $E(\hat{Y}_0) = E(Y_0/X_0)$ and

(ii) $V(\hat{Y}_0)$ is minimum.

Let us derive BLUP for Y.

Let the linear function be defined as $\hat{Y}_0 = \sum_{i=1}^{n} C_i Y_i$.

We find the constants C_i's i=1, 2,....., n such that $\hat{Y}_0$ is Best Linear Unbiased Predictor of Y_0.

The condition $E(\hat{Y}_0) = E(Y_0/X_0)$ implies that

$$\sum_{i=1}^{n} C_i = 1 \text{ and } \sum_{i=1}^{n} C_i X_i = X_0.$$

Now
$$\begin{aligned} V(\hat{Y}_0) &= V\left(\sum_{i=1}^{n} C_i Y_i\right) \\ &= \sum_{i=1}^{n} C_i^2 V(Y_i) \\ &= \sum_{i=1}^{n} C_i^2 \sigma^2 \end{aligned}$$

We minimise $\frac{1}{2}\sum_{i=1}^{n} C_i^2 \sigma^2$ with respect to C_i subject to the restrictios

$$\sum_{i=1}^{n} C_i = 1 \text{ and } \sum_{i=1}^{n} C_i X_i = X_0.$$

i.e. minimise $\phi = \frac{1}{2}\sum_{i=1}^{n} C_i^2 \sigma^2 - \lambda(\sum_{i=1}^{n} C_i - 1) - \mu(\sum_{i=1}^{n} C_i X_i - X_0)$ (16)

The first order conditions for a minimum are:

$$\frac{\partial \phi}{\partial C_i} = 0 \Rightarrow \quad C_i = \lambda - \mu X_i \; and \quad \sum C_i = 1 \Rightarrow \lambda = \frac{1}{n} - \mu \overline{X}$$

Substituting value of λ in C_i and using $\sum_{i=1}^{n} C_i X_i = X_0$ we get

$$\mu = \frac{X_0 - \overline{X}}{\sum_{i=1}^{n} x_i X_i}, \quad and \quad C_i = \frac{1}{n} + \frac{X_0 - \overline{X}}{\sum_{i=1}^{n} x_i X_i} x_i$$

Thus we get BLUP of Y_0 as

$$\hat{Y}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{X_0 - \overline{X}}{\sum_{i=1}^{n} x_i X_i} x_i \right) Y_i$$

$$= \overline{Y} - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \overline{X} + \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} X_0 \qquad (17)$$

$$= \hat{\alpha}_{OLS} + \hat{\beta}_{OLS} X_0$$

There fore BLUP of Y_0 is given by

$\hat{Y}_0 = \hat{\alpha}_{OLS} + \hat{\beta}_{OLS} X_0$, $where \hat{\alpha} \quad and \; \hat{\beta}_{OLS}$ are OLS estimators of α and β.

Thus BLUP for the response variable for given value of the regressor is obtained by replacing the regression parameters with their BLUE estimators.

$$V\left(\hat{Y}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2} \right)$$

One can construct 95% confidence interval to these predictions for every value of X_0 as

$$\hat{Y}_0 \pm t_{0.025,n-2} \left[s \left(1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2} \right)^{1/2} \right] \qquad (18)$$

2.4 Numerical Example: The following table gives the annual consumption of 10 households each selected randomly from a group of households with fixed personal disposable income. Both income and consumption are measured in $10,000.

Observations	1	2	3	4	5	6	7	8	9	10
Consumption Y_i	4.6	3.6	4.6	6.6	7.6	5.6	5.6	8.6	8.6	9.6
Income X_i	5	4	6	8	8	7	6	9	10	12

Fitting a simple linear regression of consumption(Y) on income(X)

$$Y_i = \alpha + \beta X_i + \varepsilon_i, i = 1,2,...n$$

gives the least squares estimators of intercept and slope coefficient as:

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ =6.5-(0.8095)7.5=0.4286. This is the expected consumption at zero personal income.

And slope

$\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ =0.8095 which is the marginal propensity to consume. This is the extra consumption brought about by an extra dollar of disposable income. The estimate of error variance is

$\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{(n-2)}$ =0.311905 $\hat{\sigma} = 0.558$ this the standard error of the regression.

$$S.E.(\hat{\alpha}_{OLS}) = \sqrt{0.365374} = 0.60446 \quad and\ S.E(\hat{\beta}_{OLS}) = 0.077078$$

For the null hypothesis H_0: β=0, the observed t- values is

$$t_{obs} = \frac{\hat{\beta}}{\sqrt{(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2})}} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$ =(0.8095/0.077078)=10.50.

$P\left[t_8 \ \right| > 10.5\right] < 0.0001$. This probability(LHS of this enequality) is the p-value. We reject the null hypothesis since p<0.0001. Therefore the regressor income is highly significant with respect consumption expenditure. Hypothesis for the intercept H_0 :α=0, gives the t_{obs}=0.709 which is not significant since p-value $P\left[t_8 \ \right| > 0.709\right] < 0.498$. Therefore we donnot reject H_o:α=0.

Total sum of squares=36.9, regression sum of squares=34.404762.

The coefficient of determination R^2=SSR/SST =0.9324.

This means that personal disposable income explains 93.24% of the variation in consumption.

2.5 **Extension of Two variable regression model**:

So far we have considered only one regressor X besides the constant in the regression equation. Real life situations include more than one regressor. For example, a demand equation for a product depends on real price of that product, real income of the consumer, price of a competitive product and the advertising expenditures on this product. In this case we use a multiple linear regression model to explain the relationship between them.

Suppose a sample of n observations on the dependent variable Y and k independent variables X_1, X_2 ,...,X_k are available. To allow intercept term in the model the first regressor X_1 is assumed to be a constant at unity. Let Y_i denotes the ith observation on Y and X_{pi} denotes ith observation on the p^{th} regressor.

A multiple linear regression model corresponding to the i^{th} observation has the representation

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ji}, \quad i = 1,2,...,n \qquad with\ X_{1i} = 1 \forall i = 1,..n \qquad (19)$$

In matrix notation we can write equation (19) for all the n observations as

$$Y = X\beta + \varepsilon \qquad (20)$$

Where Y is nx1 observational vector, X is the nxk data matrix , β is kx1 vector of regression parameters and ϵ is unobservable error term given by

Y= $\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix}$, $X = \begin{bmatrix} X_{11} & X_{21} & . & . & X_{k1} \\ X_{12} & X_{22} & . & . & X_{k2} \\ . & . & . & . & . \\ . & . & . & . & . \\ X_{1n} & X_{2n} & . & . & X_{kn} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{bmatrix}$, and $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$

Basic ideal Conditions:

We make the following assumptions on the multiple linear regression model

A-1) X is non-stochastic. That is the regressors are not random variables.

A-2) Rank(X) =k. No multicollinearity problem.

A-3) E(ϵ)=0, No specification error

A-4) D(ϵ) =V=σ^2I , where I is the nxn identity matrix. This assumptions indicates absence of autocorrelation and heteroscedasticity)

A-5) ϵ is multivariate normal random vector with mean vector zero and variance-covariance matrix V=$\sigma^2 I_n$

A-6) $\lim_{n\to\infty}\left(\frac{X'X}{n}\right) = Q$ a finite and non-singular matrix. That is the model do not contain too big or too small regressors such as X_t =t, or $X_t = \lambda^t$, $0<\lambda<1$, t= 1,2,...

Violations of the above assumptions leads to various problems in the regression model.

For example Violation of A-2 raises the problem of multicollinearity while violation of A-4 indicates the presence of Heteroscedasticity problem.

2.6: Estimation of Model Parameters:

We use least squares method for estimating the regression coefficient vector β and error variance σ^2

The OLS estimator is obtained by minimizing the error sum of squares $\epsilon'\epsilon$.

That is minimise $\phi = (Y-X\beta)'(Y-X\beta)$

i.e minimise $ESS = Y'Y - 2X'Y\beta + \beta'X'X\beta$

The first order conditions for minimising ESS are

$$\frac{\partial(ESS)}{\partial\beta} = 0$$

Applying the rules of matrix differentiation we get

$-X'Y + X'X\beta = 0$

Thus OLS estimator of β is given by

$$\hat{\beta} = (X'X)^{-1}X'Y$$

provided $(X'X)^{-1}$ exists which is guaranteed from assumption Rank(X) =k.

The equation of best fitting is $\hat{Y} = X\hat{\beta}$.

Analogy with the 2-variable case we may define

$$TSS = Y'Y - n\bar{Y}^2, \quad SSR = \hat{Y}'\hat{Y} - n\bar{Y}^2 \text{ and } \quad SSE = e'e ,$$

Where 'e' is the OLS residual given by

$$\begin{aligned} e &= Y - \hat{Y} \\ &= MY, \quad M = I - H, \quad H = X(X'X)^{-1}X' \\ &= M\varepsilon \end{aligned}$$

M is symmetric and idempotent matrix and MX=0.

TSS denotes total sum of squares, SSR denotes regession sum of squares and SSE denotes the residual sum of squares. It can be shown that

TSS = SSR +SSE

For testing the overall significance of the model we test the hypothesis $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$.

The test statistic for testing this hypothesis is $F = [\frac{R^2}{1-R^2}][\frac{n-k}{k-1}]$

Where the coefficient of determination R^2 is R^2= SSR/TSS.

The F-statistic defined above follows F distribution with (k-1, n-k) degrees of freedom. Insignificant values of F indicate a poor model fit.

Adjusted coefficient of determination: In case of multiple linear regression , as we increase the total number of regressors by including redundant and irrelevant variables in the model there is a possibility of increase in R^2, misguiding overall goodness of fit of the model. Hence a supplementary measure $\overline{R}^2$ (or adjusted R^2) is defined as

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

Adjusted R^2 cannot be increased by simply increasing k.

Properties of OLS estimators:

Result 1 (Gauss-Markov): OLS estimator of β is the best linear unbiased estimator(BLUE)

That is

i) OLS estimator of β is linear and unbiased.

ii) If β* is any other unbiased estimator , then OLS estimator of β is more efficient than β*

Proof;

We have $\hat{\beta} = (X'X)^{-1}X'Y$

$=\beta + (X'X)^{-1}X'\epsilon$

Now $E(\hat{\beta}) = \beta + E(X'X)^{-1}X'\varepsilon$
$= \beta$

Therefore $\hat{\beta} = (X'X)^{-1}X'Y$ is unbiased estimator of β.

Let β*=CY, where C is a constanct matrix.

Define another matrix D as $D = C-(X'X)^{-1}X'$

Now $E(\beta^*)=E(CY)$

$=E(D+(X'X)^{-1}X')(X\beta+\epsilon)$

$=E(\beta +(X'X)^{-1}X'\epsilon +D\epsilon)$ since $DX=0$

$=\beta$

Therefore β* is u.b. for β.

$$Var(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)'$$
$$= (X'X)^{-1} X' E\varepsilon\varepsilon' X (X'X)^{-1}$$
$$= (X'X)^{-1} \sigma^2$$

Var(β*)=E(β*-β)(β*-β)'

=E[(X'X)$^{-1}$X'ϵ +Dϵ][(X'X)$^{-1}$X'ϵ +Dϵ]'

=σ^2[(X'X)$^{-1}$ + DD']

$$= Var(\hat{\beta}) + DD''\sigma^2$$

We know that DD' is positive definite or p.s.d. for any D.

Therefore $Var(\beta_i^*) - Var(\hat{\beta}_i) \geq 0 \qquad \forall i = 1,..k.$

This implies that OLS estimator $\hat{\beta} = (X'X)^{-1} X'Y$ is more efficient than any other linear unbiased estimator.

Estimation of σ^2:

There is one more parameter in the model the error variance

An unbiased and consistent estimator for σ^2 is $\hat{\sigma}^2 = \frac{e'e}{n-k}, \quad e = Y - X\hat{\beta}$

Proof:

We have

e'e=Y'MY, M=I-H, H is the hat matrix.

=ϵ'Mϵ

E(e'e)=E(ϵ'Mϵ)

=E(Trace ϵ'Mϵ)

=Trace(M)E ϵϵ')

=σ^{2}TraceM

$=\sigma^2(n-k)$ Since Trace(M)=TraceI_n –Trace(H)=n-k.

hence(SSE)=$(n-k)\sigma^2$.

This implies that $\hat{\sigma}^2 = \frac{e'e}{n-k}, \quad e = Y - X\hat{\beta}$ is u.b. for σ^2.

Statistical properties:

i) OLSE of β is multivariate normal with mean vector β and var.cov. matrix $(X'X)^{-1}$

ii) $(n-k)\hat{\sigma}^2 / \sigma^2 \text{ follows } \chi^2 \text{ with } n-k df.$

iii) $\hat{\beta}_{OLS} \text{ and } \hat{\sigma}^2$ are independently distributed.

Conclusion:

1. The problem of testing hypothesis on the simple linear regression model is discussed. The distribution of the OLS estimator of regression parameter is obtained and found to be normally distributed. An unbiased and consistent estimator for error variance is proposed using OLS residuals and shown that this estimator follows chi-square distribution with (n-2) df. The distribution of the test statistics for testing the significance of the regressor follows t-distribution. Using R^2 we examine the goodness of fit of the model and ANOVA can be used to to test the overall significance of the model. Confidence interval for the mean response at a specified value of X is given. The expression for best linear unbiased predictor (BLUP) for the response at a specified value of the regressor is derived. It is shown that the problem of deriving BLUP reduces just to Obtain BLUE for the regression parameters and BLUP is obtained by replacing the slope and intercept coefficients with their BLUE estimators(i.e.OLS estimators) for specified value of the regressor. If any one of the assumption is violated then OLSE is not the appropriate method of estimation. The discussions were illustrated with a real life example.

2.Multiple linear regression model is specified. Stated with explanation, a set of basic ideal conditions to be satisfied by the model. Ordinary least squares estimators for the model parameters are derived and shown that these estimators are BLUE and consistent only when the model satisfies all the basic ideal conditions. If the model violates any of the basic assumptions the OLS estimator do not enjoy the properties discussed above and hence cannot be used for estimation and testing. Therefore care should be taken to verify the basic assumptions before choosing OLS procedure to estimate the model parameters and carry out testing various hypothesis on the regression model.