Contents

1 Introduction

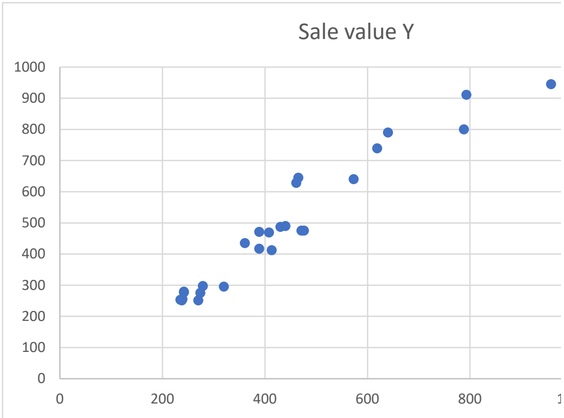
At the start, there are measurements on explanatory variables, denoted $X_1, ..., X_p$ as well on a response variable Y. Regression analysis then proceeds to describe the behavior of the response variable in terms of explanatory variables. Specifically, it seeks to establish a relationship between the response and the explanatory variables in order to monitor how changes in the latter affect the former. The relationship can also be used for predicting the value of a response given new values of the explanatory variables.

In all instances, the primary goal in regression is to develop a model that relates the response to the explanatory variables, to test it and ultimately to use it for inference and prediction.

Example 1.1. Suppose we have $Y = \textit{sale}$ values for $n = 25$ houses and $X = \textit{Assessed}$ values. Hence the given data consists of the pairs

$$\{(X_i, Y_i), i = 1, ..., n\}$$

Assessed value X	Sale value Y
238	251
270	251
235	253
239	255
274	275
242	277
242	279
320	295
279	297
413	412
389	417
361	435
408	469
389	471
471	475
476	475
430	487
440	490
461	628
573	640
465	645
619	739
640	790
788	800
793	911
958	945



We first plot the n paired data Y_i vs X_i. If it seems reasonable to fit a straight line to the points, we then postulate the following simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.1}$$

Here, ϵ represents an unobserved random error term, β_0 is the intercept whereas β_1 represents the slope of the line. Both β_0, β_1 are labeled parameters. They are unknown and would need to be estimated in some way from the observed data.

Alternatively, the model may be expressed in terms of $\left(X_i - \bar{X}\right)$

$$Y_i = \left(\beta_0 + \beta_1 \bar{X}\right) + \beta_1 (X_i - \bar{X}) + \epsilon_i$$

where $\bar{X}$ represents the average of the X_i .

The proposed model (1.1) is linear in the parameters β_0, β_1. The model would still be referred to as linear if instead we had X_i^2 instead of X_i. It is common practice to make the following assumption:

Assumption: The random error terms are uncorrelated, have mean equal to 0 and common variance equal to σ^2 .

Under this assumption

$$\begin{aligned} E[Y_i] &= \beta_0 + \beta_1 X_i \\ \sigma^2[Y_i] &= \sigma^2 \end{aligned}$$

CAUTION: We emphasize that a well fitting regression model does not imply causation. One can relate stock market prices in N.Y. to the price of bananas in an offshore island. This does not mean there is a causal relationship.

1.1 The method of least squares

The method of least squares due to Gauss-Legendre is the most popular approach to fitting a regression model.

Set Q as the sum of square errors

$$\begin{aligned} Q &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2 \end{aligned}$$

Then minimize Q with respect to the parameters by differentiating with respect to β_0, β_1.

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] = 0 \\ \frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] X_i = 0 \end{aligned}$$

The linearity assumption leads to two linear equations in two unknowns whose solutions denoted b_0, b_1 are

$$b_0 = \bar{Y} - b_1\bar{X} \tag{1.2}$$

$$b_1 = \frac{\sum\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum\left(X_i-\bar{X}\right)^2} \tag{1.3}$$

$$= \frac{\sum\left(X_i-\bar{X}\right)Y_i}{\sum\left(X_i-\bar{X}\right)^2} \tag{1.4}$$

$$= \sum k_iY_i$$

where

$$k_i = \frac{\left(X_i-\bar{X}\right)}{\sum\left(X_i-\bar{X}\right)^2}$$

Then it can be shown

$$\sum k_i = 0, \sum k_iX_i = 1, \sum k_i^2 = \frac{1}{\sum\left(X_i-\bar{X}\right)^2}.$$

The equation of the fitted line is

$$\hat{Y} = b_0 + b_1X \tag{1.5}$$

Alternatively,

$$\hat{Y} = \left(b_0 + b_1\bar{X}\right) + b_1(X-\bar{X}) \tag{1.6}$$

Theorem 1.1. *Gauss Markov) The least square estimators b_0, b_1 are unbiased and have minimum variance among all unbiased linear estimators.*

Proof. Consider an unbiased estimator for β_1 say, $\hat{\beta}_1 = \sum c_iY_i$ which must satisfy

$$\beta_1 = E[\hat{\beta}_1]$$

$$= \sum c_iE[Y_i]$$

$$= \sum c_i[\beta_0 + \beta_1X_i]$$

Hence, $\sum c_i = 0, \sum c_iX_i = 1$ and $\sigma^2[\hat{\beta}_1] = \sigma^2\sum c_i^2$. □

Consider setting $c_i = k_i + d_i$ where d_i is arbitrary. Then substituting

$$\begin{aligned}
\sum k_i d_i &= \sum k_i (c_i - k_i) \\
&= \sum c_i \frac{\left(X_i - \bar{X}\right)}{\sum \left(X_i - \bar{X}\right)^2} - \frac{1}{\sum \left(X_i - \bar{X}\right)^2} \\
&= 0
\end{aligned}$$

on using the properties of c_i. Hence $\{k_i\}$ and $\{d_i\}$ are uncorrelated and we have by the Pythagorean theorem

$$\begin{aligned}
\sigma^2[\hat{\beta}_1] &= \sigma^2 \sum c_i^2 \\
&= \sigma^2 \left\{\sum k_i^2 + \sum d_i^2\right\}
\end{aligned}$$

showing that the variance is minimized when d_i are all 0.

We may write $\hat{Y} = b_0 + b_1 X$ for the estimated or fitted line, $e_i = Y_i - \hat{Y}_i$ for the estimated i^{th} residual and $\hat{\sigma^2} = \frac{\sum e_i^2}{n-2}$ for the estimate of the variance σ^2.

Theorem 1.2. *The variances of the least squares estimators are*

$$\begin{aligned}
\sigma^2[b_0] &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum \left(X_i - \bar{X}\right)^2} \right) \\
\sigma^2[b_1] &= \sigma^2 \left(\frac{1}{\sum \left(X_i - \bar{X}\right)^2} \right)
\end{aligned}$$

These may be estimated by replacing σ^2 by

$$\hat{\sigma^2} = \frac{\sum e_i^2}{n-2} \tag{1.7}$$

also known as the mean square error and denoted MSE.

Properties of the fitted Regression line

1. $\sum e_i = 0$
2. $\sum Y_i = \sum \hat{Y}_i$
3. $\sum X_i e_i = 0$

4. $\sum\left(Y_i - \bar{Y}\right)^2 = b_1^2 \sum\left(X_i - \bar{X}\right)^2 + \sum\left(Y_i - \hat{Y}_i\right)^2$

5. The point $(\bar{X}, \bar{Y})$ is on the fitted line. This can be seen from (1.5)

6. Under the normality assumption $\{\epsilon_i\} \sim i.i.d.N\left(0, \sigma^2\right)$, the method of maximum likelihood leads to the method of least squares.

1.2 Inference in regression

The method of least squares was used to obtain the equation of the fitted regression line. For the purpose of drawing inference, it is necessary to make some assumptions on the distribution of the error terms, the most common of which is that the errors $\{\epsilon_i\} \sim i.i.d.N\left(0, \sigma^2\right)$.

Theorem 1.3. *Suppose that we have the model* $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ *where* $\{\epsilon_i\} \sim i.i.d.N\left(0, \sigma^2\right)$ *for* $i = 1, ..., n$ *. Then*

a) $\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$ *where* $s^2(b_1) = \frac{MSE}{\sum\left(X_i - \bar{X}\right)^2}$

b) $\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$ *where* $s^2(b_0) = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum\left(X_i - \bar{X}\right)^2}\right]$

c) MSE is an unbiased estimate of σ^2 *and* $\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$ *and independent of* b_0, b_1

Proof. a) We see that $b_1 = \sum k_i Y_i$ where $k_i = \frac{\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2}$. Hence, b_1 is unbiased in view of the properties of the $\{k_i\}$

Since $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, it follows that

$$\begin{aligned} b_1 &= \sum k_i Y_i \sim N(\sum k_i(\beta_0 + \beta_1 X_i), \sigma^2 \sum k_i^2) \\ &\sim N\left(\beta_1, \frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}\right) \end{aligned}$$

b) As well, $b_0 = \bar{Y} - b_1\bar{X} = \frac{1}{n}\sum Y_i - \sum k_i Y_i \bar{X} = \sum\left(\frac{1}{n} - k_i\bar{X}\right) Y_i$

The result follows from properties of the k_i

c) We shall demonstrate this result using the matrix approach in subsequent sections. □

This theorem can be used to test hypotheses about the parameters and to construct confidence intervals.

1.3 Analysis of Variance (ANOVA) table

It is customary and revealing to summarize the statistical analysis in the form of a table. We illustrate this for the case $p = 2$ exhibited in the table below.

Source	Sum of Squares (SS)	df	MS=SS/df	F statistic	E[MS]
Regression	SSR $=b_1^2 S_{XX}$	$p-1$	MSR	MSR/MSE	$\sigma^2 + \beta_1^2 \sum \left(X_i - \bar{X}\right)^2$
Error	SSE= $\sum \left(Y_i - \hat{Y}_i\right)^2$	$n-p$	MSE		σ^2
Total	SSTO= $\sum \left(Y_i - \bar{Y}\right)^2$	$n-1$			

$\sum \left(Y_i - \bar{Y}\right)^2$ has $n-1$ degrees of freedom because of the constraints that $\sum \left(Y_i - \bar{Y}\right) = 0$

$b_1^2 \sum \left(X_i - \bar{X}\right)^2$ has one degree of freedom because it is a function of b_1

$\sum \left(Y_i - \hat{Y}_i\right)^2$ has n-2 degrees of freedom because it is a function of two parametersEach of the sums of squares is a quadratic form where the rank of the corresponding matrix is the degrees of freedom indicated.

Cochran's theorem applies and we conclude that the quadratic forms are independent and have chi square distributions. It is well known that the ratio of two independent chi square divided by their degrees of freedom has a F-distribution

$$
\begin{aligned}
F &= \frac{\left[SSR/\left(\sigma^2 (p-1)\right)\right]}{\left[SSE/\left(\sigma^2 (n-p)\right)\right]} \\
&= \frac{MSR}{MSE} \sim F_{p,(n-p)}
\end{aligned}
$$

The ANOVA table indicates how one can test the null hypothesis

$$
\begin{aligned}
H_0 &: \beta_1 = 0 \\
H_1 &: \beta_1 \neq 0
\end{aligned}
$$

The null hypothesis is that the slope of the line is equal to 0. Under the null hypothesis, the expected mean square for regression and the expected mean square error are separate independent estimates of the variance σ^2. Hence if the null hypothesis is true, the F ratio should be small. On the other hand, if the alternative hypothesis H_1 is true, then the numerator of the F ratio will be expected to be large. Consequently, large values of the F statistic are consistent with the alternative. We reject the null hypothesis for large values of F.

Example 1.2. We consider the following example on grade point averages at the end of of the freshman year (Y) as a function of the ACT test scores (X). .

a) We plot the data
b) We obtain the least squares estimates
c) We plot the estimated regression function and estimate Y when $X = 30$
d) Compute the ANOVA table
e) Compute confidence intervals for the parameters

Exercise 1.1. Consider the following data on airfreight breakage (Y) as a function of shipment route (X). CH01PR21

i	1	2	3	4	5	6	7	8	9	10
X_i	1	0	2	0	3	1	0	1	2	0
Y_i	16	9	17	12	22	13	8	15	19	11

a) Compute the ANOVA table
b) Compute confidence intervals for the parameters
c) Compute a confidence interval for the average response when $X = 1$

1.4 Confidence Intervals

It is of interest to construct confidence intervals for

a) the average $E[Y] = \beta_0 + \beta_1 X$ for a new observation X
b) the prediction of a new value of Y for a given X

The point estimate $\hat{Y} = b_0 + b_1 X$ is used as the point estimate in both cases a) and b).

It is unbiased and has a normal distribution as seen from

$$\begin{aligned}\hat{Y} &= b_0 + b_1 X \\ &= \sum \left[\frac{1}{n} + k_i \left(X - \bar{X}\right)\right] Y_i\end{aligned}$$

Moreover,

$$\begin{aligned}\sigma^2 \left[\hat{Y}\right] &= \sigma^2 \sum \left[\frac{1}{n} + k_i \left(X - \bar{X}\right)\right]^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\left(X - \bar{X}\right)^2}{\sum \left(X_i - \bar{X}\right)}\right]\end{aligned}$$

We note that the variance increases with the distance of X from $\bar{X}$. The variance $\sigma^2 \left[\hat{Y}\right]$

is estimated by

$$s^2\left[\hat{Y}\right] = MSE\left[\frac{1}{n} + \frac{\left(X - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)}\right]$$

Hence inference in the form of confidence interval an hypothesis testing for the average E[Y] is conducted using the fact that

$$\frac{\hat{Y} - E[Y]}{s[\hat{Y}]} \sim t_{n-2}$$

a Student t distribution with $n - 2$ degrees of freedom.

For the prediction problem, note that

$$Y_{new} = \beta_0 + \beta_1 X + \epsilon_{new}$$

and

$$\hat{Y}_{new} = \hat{Y} + \epsilon_{new}$$

$$\begin{aligned} \sigma^2\left[\hat{Y}_{new}\right] &= \sigma^2\left[\hat{Y}\right] + \sigma^2 \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)}\right] \end{aligned}$$

The variance $\sigma^2\left[\hat{Y}_{new}\right]$ is estimated by

$$s^2\left[\hat{Y}_{new}\right] = MSE\left[1 + \frac{1}{n} + \frac{\left(X - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)}\right]$$

Hence inference in the form of confidence interval an hypothesis testing for the prediction of a new value is conducted using the fact that

$$\frac{\hat{Y}_{new} - Y_{new}}{s[\hat{Y}_{new}]} \sim t_{n-2}$$

Example 1.3. Consider the grade point average data (ACT).

a) Compute a confidence interval for the average response $when\ ACT = 3.5$

b) Compute a prediction interval for the average response $when\ ACT = 3.5$

2 Matrix Approach to Regression

We will preamble the matrix presentation by describing some distributional results.

2.1 Distributional Results

Let $Y = [Y_1, ..., Y_n]'$ be the transpose of the column data vector.

Define the expectation

$$E[Y] = [EY_1, ..., EY_n]'$$

Proposition If $Z = AY + B$ for some matrix of constants A, B, then

$$E[Z] = AE[Y] + B$$

Proof: $(EZ_i) = E\left\{\left[\sum_j a_{ij} Y_j\right] + b_i\right\} = \left[\sum_j a_{ij} EY_j\right] + b_i$

Definition 2.1. The covariance $COV\,[Y] = E\left\{[Y - EY]\,[Y - EY]'\right\} \equiv \Sigma$

Proposition $COV\,[AY] = A\Sigma A'$

Definition 2.2. A random vector Y has a multivariate normal distribution if its density is given by

$$f(y_1, ..., y_n) = \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} exp - \frac{1}{2}(y - \mu)' \, \Sigma^{-1}(y - \mu)$$

where

$$y' = (y_1, ..., y_n), \mu' = (\mu_1, ..., \mu_n), \Sigma = COV\,[Y]$$

denoted $Y \sim N_n(\mu, \Sigma)$.

A fundamental result is

Theorem 2.1. *Let $Y \sim N_n(\mu, \Sigma)$. Let A be an arbitrary $p \times n$ matrix of constants. Then*

$$Z = AY + B \sim N_n(A\mu + B, A\Sigma A')$$

This theorem implies that any linear combination of normal variates has a normal distribution. We do not prove this theorem here.

Example 2.1. Let $Y \sim N_n(\mu, \Sigma)$. Let $A = (1, ..., 1)$.Then

$$AY \sim N_1(A\mu, A\Sigma A')$$

where

$$A\mu = \sum_{i=1}^{n} \mu_i, A\Sigma A' = \sum \sigma_j^2 + 2\sum_{i \neq j} \sigma_{ij}$$

The matrix representation of regression makes it easy to generalize to fitting several independent variables.

Let $Y = [Y_1, ..., Y_n]'$ be the transpose of the column data vector.

Let $\beta = [\beta_0, \beta_1, ...\beta_{p-1}]'$ be the transpose of the coefficients

Let $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_n]'$ be the transpose of the random error terms

Let $X = \begin{pmatrix} 1 & X_{11} & .. & X_{1p} \\ 1 & X_{21} & .. & X_{2p} \\ . & . & .. & .. \\ 1 & X_{n1} & .. & X_{np} \end{pmatrix}$ be the matrix which incorporates the p explanatory variables

If $\epsilon \sim N_n(0, \sigma^2 I_n)$, then the regression model may be expressed as

$$Y = X\beta + \epsilon \sim N_n(X\beta, \sigma^2 I_n)$$

where I_n is the $n \times n$ identity matrix and N_n is the multivariate normal distribution.

Derivatives If $z = a'y$,then

$$\frac{\partial z}{\partial y} = a$$

If $z = y'y$,

$$\frac{\partial z}{\partial y} = 2y$$

If $z = a'Ay$,

$$\frac{\partial z}{\partial y} = A'a$$

If $z = y'Ay$,

$$\frac{\partial z}{\partial y} = A'y + Ay$$

If $z = y'Ay$, and A is symmetric

$$\frac{\partial z}{\partial y} = 2A'y$$

The sum of squares is given by

$$Q = (Y - X\beta)' (Y - X\beta)$$

Differentiating with respect to the vector β

$$\begin{aligned} \frac{\partial Q}{\partial \beta} &= -2X' ((Y - X\beta)) \qquad (2.1) \\ &= -2 (X'Y - X'X\beta) = 0 \end{aligned}$$

Hence the solutions to the normal equations are

$$\begin{aligned} b &= (X'X)^{-1} X'Y \\ &= AY \end{aligned}$$

where $A = (X'X)^{-1} X'$ provided the inverse of $(X'X)$ exists. It follows that

$$b \sim N_p \left(AX\beta, \sigma^2 AA'\right)$$

But

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

and

$$\begin{aligned} AA' &= (X'X)^{-1} X'X (X'X)^{-1} \\ &= (X'X)^{-1} \end{aligned}$$

Hence,

$$b \sim N_p \left(\beta, \sigma^2 (X'X)^{-1}\right)$$

The fitted line is then

$$\begin{aligned} \hat{Y} &= Xb \\ &= X (X'X)^{-1} X'Y \\ &= HY \end{aligned}$$

where the "hat" matrix H (because it puts a hat on Y) is given by

$$H = X\left(X'X\right)^{-1}X' \tag{2.2}$$

2.2 Properties of the hat matrix H

The hat matrix has some nice properties.

a) It is a projection matrix, idempotent and symmetric

$$\begin{aligned} HH &= H \\ H' &= H \end{aligned}$$

b) The matrix H is orthogonal to the matrix $I - H$

$$(I - H)\,H = H - HH = 0$$

Moreover, $(I - H)$ is idempotent and is a projection matrix as well.

c) The residual vector is expressible as

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - HY \\ &= (I - H)\,Y \end{aligned}$$

d) Properties b) and c) imply that the observation vector Y is projected onto a space spanned by the columns of H and the residuals are in a space orthogonal to it

$$Y = HY + (I - H)\,Y$$

By the Pythagorean theorem

$$\|Y\|^2 = \|HY\|^2 + \|(I - H)\,Y\|^2 \tag{2.3}$$

We note that

$$\begin{aligned}
\sigma^2[e] &= Variance\,[(I-H)\,Y] \\
&= (I-H)\,\sigma^2\,[Y]\,(I-H)' \\
&= \sigma^2\,(I-H)
\end{aligned}$$

which is estimated by

$$s^2[e] = (MSE)\,(I-H)$$

Moreover,

$$\begin{aligned}
\sigma^2[b] &= \left[(X'X)^{-1}\,X'\right]\sigma^2\left[X\,(X'X)^{-1}\right] \\
&= \sigma^2\,(X'X)^{-1}
\end{aligned}$$

Exercise 2.1. a) For the case $p = 2$, obtain the hat matrix. Show that rank H= Trace H =2

b) Show the relationship

$$\begin{aligned}
\sum\left(Y_i-\bar{Y}\right)^2 &= b_1^2\sum\left(X_i-\bar{X}\right)^2+\sum\left(Y_i-\hat{Y}_i\right)^2 \\
Total\ Sum\ of\ Squares &= Regression\ Sum\ of\ Squares + Error\ Sum\ of\ Squares
\end{aligned}$$

Definition 2.3. Let $Y_1, ..., Y_n$ be a random sample from $N\,(\mu, \sigma^2)$. A quadratic form in the $Y's$ is defined to be the real quantity

$$Q = Y'AY$$

where A is a symmetric positive definite matrix.

Our next results permit us to compute the expectation of quadratic forms.

Let A be a symmetric matrix and let Y be a random vector. Then the singular value decomposition of A implies that there exists an orthogonal matrix P such that if $\Lambda = (\lambda_i)$ is the diagonal matrix of eigenvalues of A,

$$A = P'\Lambda P.$$

Proposition $E\,[Y'AY] = Trace\,[A\Sigma] + (EY)'\,(EY)$

Proof: $Y'AY = Y'P'\Lambda PY = (PY)'\,\Lambda\,(PY) = \sum\lambda_i\,\|(PY)_i\|^2$

where $(PY)_i$ indicates the i^{th} element in PY.

$(PY)_i$ is a random variable and its second moment is

$$\begin{aligned} E\left\|(PY)_i\right\|^2 &= Var\left\|(PY)_i\right\| + \left[E\left(PY\right)_i\right]^2 \\ &= \left(P\Sigma P'\right)_{ii} + \left[(PEY)_i\right]^2 \end{aligned}$$

Hence

$$\begin{aligned} E\sum\lambda_i\left\|(PY)_i\right\|^2 &= \sum\lambda_i\left(P\Sigma P'\right)_{ii} + \sum\lambda_i\left[(PEY)_i\right]^2 \\ &= Trace\left(\Lambda P\Sigma P'\right) + \mu' A\mu \\ &= Trace\left(P'\Lambda P\Sigma\right) + \mu' A\mu \end{aligned}$$

Lemma 2.1. *The sample variance S_n^2 is an unbiased estimate of the population variance.*

Proof. Suppose $Y_1, ..., Y_n \underset{\sim}{i.i.d.} N\left(\mu, \sigma^2\right)$. Let $Y = [Y_1, ..., Y_n]'$ and

$$\begin{aligned} A &= \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & ... & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & ... & -\frac{1}{n} \\ ... & ... & ... & ... \\ -\frac{1}{n} & -\frac{1}{n} & ... & 1-\frac{1}{n} \end{pmatrix} \\ &= I - \frac{11'}{n} \end{aligned}$$

Then $(n-1)\,S_n^2 = \sum\left(Y_i - \bar{Y}\right)^2 = Y'AY$ and

$$\begin{aligned} E\left[Y'AY\right] &= Trace\left[A\Sigma\right] + (EY)'\,(EY) \\ &= \sigma^2 Trace A + \mu^2 1' A 1 \\ &= \sigma^2\left(n-1\right) + 0 \end{aligned}$$

□

In the regression model,

$$Y - \hat{Y} = (I - H)\,Y$$

Since $I-H$ is idempotent,

$$\left(Y-\hat{Y}\right)'\left(Y-\hat{Y}\right)=Y'\left(I-H\right)Y$$

and

$$\begin{aligned} E\left[Y'\left(I-H\right)Y\right] &= Trace\left(I-H\right)\Sigma+\mu'\left(I-H\right)\mu \\ &= \sigma^2 Trace\left(I-H\right)+\left(X\beta\right)'\left(I-H\right)\left(X\beta\right) \\ &= \sigma^2\left(n-p\right)+0 \end{aligned}$$

Definition 2.4. (a) A random variable U is said to have a χ^2_ν distribution with ν degrees of freedom if its density is given by

$$f\left(u;\nu\right)=\frac{1}{2^{\nu/2}\Gamma\left(\nu/2\right)}u^{(\nu/2)-1}e^{-u/2}, u>0, \nu>0$$

The mean and variance of U are respectively $\nu, 2\nu$.

(b) A random variable U is said to have a non-central $\chi^2_\nu\left(\lambda\right)$ distribution with ν degrees of freedom and non centrality parameter λ if its density is given by

$$f\left(u;\nu,\lambda\right)=\sum_{i=0}^{\infty}e^{-\lambda/2}\frac{\left(\lambda/2\right)^i}{i!}f\left(u;\nu+2i\right), u>0, \nu>0$$

The non-central chi square distribution is a Poisson weighted mixture of central chi square distributions.The mean and variance are respectively $(\nu+\lambda)$ and $(2\nu+4\lambda)$.

(c) We include here the fact that the distribution of the ratio of two independent central chi square distributions divided by their respective degrees of freedom

$$F_{\nu_1,\nu_2}=\frac{\left(\chi^2_{\nu_1}/\nu_1\right)}{\left(\chi^2_{\nu_2}/\nu_2\right)}$$

is an F distribution with ν_1 and ν_2 degrees of freedom. If the numerator is a non central chi square distribution, then the F becomes a non central F distribution.

Theorem 2.2. *Cochran's Theorem Let Y be a random vector with distribution $N_n\left(\mu,\sigma^2 I\right)$. Suppose that we have the decomposition*

$$Y'Y=Q_1+...+Q_k$$

where $Q_i=Y'A_iY$ rank $\left(A_i\right)=n_i$. Then $\left\{\frac{Q_i}{\sigma^2}\right\}$ are independent and have $\left\{\chi^2_{n_i}\left(\lambda_i\right)\right\}$

distributions if and only if

$$\sum n_i = n$$

The ranks n_i are referred to as degrees of freedom. Here, $\lambda_i = \mu' A_i \mu$.

Applications Suppose $Y_1, ..., Y_n \underset{\sim}{i.i.d.} N(\mu, \sigma^2)$. Let $Y = [Y_1, ..., Y_n]'$ and

$$\begin{aligned} A &= \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & ... & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & ... & -\frac{1}{n} \\ ... & ... & ... & ... \\ -\frac{1}{n} & -\frac{1}{n} & ... & 1-\frac{1}{n} \end{pmatrix} \\ &= I - \frac{11'}{n} \end{aligned}$$

Then

$$\begin{aligned} \sum_{i=1}^{n} Y_i^2 &= Y'Y = Y'AY + Y'\left(\frac{11'}{n}\right)Y \\ n &= rank A + rank\left(\frac{11'}{n}\right) \\ &= (n-1) + 1 \end{aligned}$$

From Cochran's theorem, $Q_1 = \frac{Y'AY}{\sigma^2} \sim \chi^2_{n-1}$ and $Q_2 = \frac{Y'\left(\frac{11'}{n}\right)Y}{\sigma^2} \sim \chi^2_1$ are independent. But

$$Q_1 = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{\sigma^2}, Q_2 = \frac{n\bar{Y}^2}{\sigma^2}$$

and hence the ratio

$$\begin{aligned} F_{1,n-1} &= \frac{Q_2/1}{Q_1/(n-1)} \\ &= \frac{n\bar{Y}^2}{S_n^2} \end{aligned}$$

has an F distribution with degrees of freedom $1, (n-1)$. Equivalently,

$$T_{n-1} = \frac{\sqrt{n}\bar{Y}}{S_n}$$

has a Student distribution with $(n-1)$ degrees of freedom.

Application In linear regression,

$$\begin{aligned} Y &= Xb + (Y - Xb) \\ \|Y\|^2 &= \|Xb\|^2 + \|Y - Xb\|^2 \\ &= Y'HY + Y'(I - H)Y \end{aligned}$$

By Cochran's theorem,

$$Y'HY \sim \chi^2_p, Y'(I - H)Y \sim \chi^2_{n-p}$$

and are independent. The first term is the sum of squares due to the regression whereas the second represents the error sum of squares. We summarize this in the next section in the analysis of variance table.

3 Multiple Linear Regression

In practice, one is often presented with several predictor variables. For two predictors, the linear regression model becomes

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

with the assumptions that $\{\epsilon_i\}$ are i.i.d. $N(0, \sigma^2)$. This model describes a plane in three dimensions. It is an additive model where β_1 represents the rate of change in a unit increase in X_1 when X_2 is held fixed. An analogous interpretation can be made for β_2.

In general, we may have the linear regression model involving $(p-1)$ explanatory variables

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i$$

The predictor variables may be qualitative taking values 0 or 1 as for example if one wishes to take into account gender. So here

$$X = \begin{cases} 0 & \text{if the subject is male} \\ \\ 1 & \text{if the subject is female} \end{cases}$$

We may also have a second degree polynomial

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

a transformed response

$$ln\ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

interaction effects

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$$

In all cases, it is instructive to make use of the matrix approach to unify the development.

We recall from Chapter 2

Let $Y = [Y_1, ..., Y_n]'$ be the transpose of the column data vector.

Let $\beta = [\beta_0, \beta_1, ...\beta_{p-1}]'$ be the transpose of the coefficients

Let $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_n]'$ be the transpose of the random error terms

Let $X = \begin{pmatrix} 1 & X_{11} & .. & X_{1p-1} \\ 1 & X_{21} & .. & X_{2p-1} \\ . & . & .. & .. \\ 1 & X_{n1} & .. & X_{np-1} \end{pmatrix}$ be the matrix which incorporates the p explanatory variables

Then the regression model may be expressed as

$$Y = X\beta + \epsilon, \epsilon \sim N_n(0, \sigma^2 I_n)$$

where I_n is the $n \times n$ identity matrix and N_n is the multivariate normal distribution.

Letting

$$b' = [b_0, b_1, ..., b_{p-1}]$$

be the least squares estimate of β we have

$$b = (X'X)^{-1} X'Y$$

The fitted values are

$$\begin{aligned} \hat{Y} &= Xb \\ &= HY \end{aligned}$$

where the hat matrix $H = X(X'X)^{-1} X$. The variance -covariance matrix of the residuals $e = (I - H) Y$ is

$$\sigma^2 [e] = \sigma^2 (I - H)$$

which is estimated by

$$s^2 [e] = (MSE)(I - H)$$

Also

$$s^2 [b] = (MSE)(X'X)^{-1}$$

We may summarize the results in an ANOVA table

Source	SS	df	MS
Regression	SSR	p-1	MSR=SSR/(p-1)
Error	SSE	n-p	MSE=SSE/(n-p)
Total	SSTO	n-1	

where

$$\begin{aligned} SSTO &= Y'Y - \left(\frac{1}{n}\right) Y'JY \\ &= Y'\left[I - \left(\frac{1}{n}\right) J\right] Y \end{aligned}$$

$$SSE = e'e = Y'(I - H)\,Y$$

$$\begin{aligned} SSTR &= b'X'Y - \left(\frac{1}{n}\right) Y'JY \\ &= Y'\left[H - \left(\frac{1}{n}\right) J\right] Y \end{aligned}$$

To test the hypothesis

$$H_0 \quad : \quad \beta_1 = \beta_2 = ... = \beta_{p-1}$$

$$H_1 : \text{ not all } \beta_k = 0$$

we use the test statistic

$$F = \frac{MSR}{MSE} \sim F\,(p-1, n-p)$$

and reject H_0 for large values.

Tests

$$H_0 \ : \ \beta_k = 0$$

$$H_1 \ : \ \beta_k \neq 0$$

for individual coefficients may be conducted using the fact that the standardized coefficient has a Student t distribution

$$\frac{b_k}{s\,[b_k]} \sim t_{n-p}$$

3.1 Extra sum of squares principle

A more general approach to regression, labeled the extra sum of squares principle, which will be useful for more complex models consists of the following steps illustrated here for $p = 2$.

Step 1 Specify the Full (F) model $Y = \beta_0 + \beta_1 X + \epsilon$ and obtain the error sum of squares

$$SSE(F) = \sum \left(Y_i - \hat{Y}_i\right)^2$$

Step 2 Consider the Reduced (R) model whereby $\beta_1 = 0$

$$Y = \beta_0 + \epsilon$$

and obtain the corresponding error sum of squares

$$SSE(R) = \sum \left(Y_i - \bar{Y}\right)^2$$

The logic now is to compare the two error sum of squares. With more parameters in the model, we expect that

$$SSE(F) \leq SSE(R)$$

If we have equality above, we may conclude the model is not of much help. As a result, we may test the benefit of the model be computing the test statistic

$$F^* = \frac{\left[\frac{SSE(R)-SSE(F)}{df_R - df_F}\right]}{\left[\frac{SSE(F)}{df_F}\right]} \tag{3.1}$$

and rejecting the null hypothesis $H_0 : \beta_1 = 0$ for large values of F^* which has an F distribution $F(df_R - df_F, df_F)$.

Application An immediate application of this approach is to the situation where there are repeat observations at the same values of X. Suppose that the full model is given by

$$Y_{ij} = \mu_j + \epsilon_{ij}, i = 1, ..., n_j; j = 1, ..., c$$

and $\{\epsilon_{ij}\}$ are i.i.d. $N\left(0, \sigma^2\right)$.

The $\{\mu_{ij}\}$ are unrestricted parameters when $X = X_j$.Their least squares estimates are

$$\bar{Y}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j}$$

The error sum of squares for this full unrestricted model is

$$SSE(F) = \sum_{ij} \left(Y_{ij} - \bar{Y}_j\right)^2$$

The corresponding degrees of freedom are

$$\begin{aligned} df_F &= \sum_{j=1}^{c} (n_j - 1) \\ &\equiv n - c \end{aligned}$$

Note that if all $n_j = 1$, then $df_F = 0$, $SSE(F) = 0$ and the analysis does not proceed any further.

Consider now the reduced model which specifies the linear model

$$Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$$

which has error sum of squares equal to

$$SSE(R) = \sum_{ij} \left(Y_{ij} - \hat{Y}_{ij} \right)^2$$

where

$$\hat{Y}_{ij} = b_0 + b_1 X_j \tag{3.2}$$

The degrees of freedom are $df_R = (n - 2)$. Hence, we may test

$$H_0 \ : \ E[Y] = \beta_0 + \beta_1 X$$

$$H_1 \ : \ E[Y] \neq \beta_0 + \beta_1 X$$

by computing the ratio

$$F^* = \frac{\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right]}{\left[\frac{SSE(F)}{df_F}\right]}$$

So the test here is on whether a linear model is justified at all. This is different from just testing that the slope is zero.

We may gain some insight into the components of the F^* ratio. Note that

$$\left(Y_{ij} - \hat{Y}_{ij} \right) = \left(Y_{ij} - \bar{Y}_j \right) - \left(\bar{Y}_j - \hat{Y}_{ij} \right)$$

and

$$\sum_{ij} \left(Y_{ij} - \hat{Y}_{ij} \right)^2 = \sum_{ij} \left(Y_{ij} - \bar{Y}_j \right)^2 + \sum_{ij} \left(\bar{Y}_j - \hat{Y}_{ij} \right)^2$$

The corresponding degrees of freedom are $df_R = (n-2)$, $df_{PE} = (n-c)$, $df_{LF} = (c-2)$

We label these sums of squares as follows:

Definition 3.1. SSE(R) $= \sum_{ij} \left(Y_{ij} - \hat{Y}_{ij} \right)^2$ Error sum of squares for the reduced model

SSPE $= \sum_{ij} \left(Y_{ij} - \bar{Y}_j \right)^2$ Pure error sum of squares

SSLF $= \sum_{ij} \left(\bar{Y}_j - \hat{Y}_{ij} \right)^2$ Error sum of squares due to lack of fit which in view of (3.2) is independent of i

An ANOVA table summarizes the analysis.

Source	SS	df	MS	F	E(MS)
Regression	SSR= $\sum_{ij} \left(\hat{Y}_{ij} - \bar{Y} \right)^2$	1	MSR=SSR/1	F=MSR/MSE	
Residual error	SSE(R)=$\sum_{ij} \left(Y_{ij} - \hat{Y}_{ij} \right)^2$	n-2	MSE=SSER/(n-2)		
Lack of fit	SSLF=$\sum_{ij} \left(\bar{Y}_i - \hat{Y}_{ij} \right)^2$	c-2	MSLF=SSLF/(c-2)	F^*	$\sigma^2 + \frac{\sum n_i(\mu_i-\beta_0}{c-2}$
Pure error	SSPE=$\sum_{ij} \left(Y_{ij} - \bar{Y}_i \right)^2$	n-c	MSPE=SSPE/(n-c)		σ^2
Total	$\sum_{ij} \left(Y_{ij} - \bar{Y} \right)^2$	n-1			

We note

$$SSE\,(R) = SSLF + SSPE$$

The approach can be extended to multiple regression. We define

$$SSR\,(X_2|X_1) = SSE\,(X_1) - SSE\,(X_1, X_2) \tag{3.3}$$

to be the reduction in the error sum of squares when after X_1 in included, an additional variable X_2 is added to the model. Since

$$SSTO = SSR + SSE$$

we may re-express (3.3) as

$$SSR\,(X_2|X_1) = SSR\,(X_1, X_2) - SSR\,(X_1)$$

Similarly, when three variables are involved, we may breakdown the sum of squares due to the regression as

$$SSR\,(X_1, X_2, X_3) = SSR\,(X_1) + SSR\,(X_2|X_1) + SSR\,(X_3|X_1, X_2)$$

This decomposition enables us to judge the effect an added variable has on the sum of squares due to the regression. An ANOVA table would be decomposed as follows

Source	SS	df	MS
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \| X_1$	$SSR(X_2 \| X_1)$	1	$MSR(X_2 \| X_1)$
$X_3 \| X_1, X_2$	$SSR(X_3 \| X_1, X_2)$	1	$MSR(X_3 \| X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	n-4	$MSE(X_1, X_2, X_3)$
Total	$SSTO$	n-1	

The extra sum of squares principle described in Chapter 3 considers a full model and a reduced model. It then makes use of the statistic below to determine the usefulness of the reduced model

$$F^* = \frac{\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right]}{\left[\frac{SSE(F)}{df_F}\right]} \sim F(df_R - df_F, df_F)$$

In general, suppose that

$$S_1 = SS(b_0(1), ..., b_p(1))$$

represents the sum of squares residual when p variables are included and

$$S_2 = SS(b_0(2), ..., b_q(2))$$

represents the sum of squares residual when q variables are included, with $p > q$. Then the difference $S_1 - S_2$is defined to be the extra sum of squares. It will be used to test the hypothesis that

$$H_0 : \beta_{q+1} = ... = \beta_p = 0$$

It can be shown that under that hypothesis, $\frac{S_1 - S_2}{p-q}$ is an unbiased estimate of σ^2 independent of MSE and hence their ratio will have an F distribution. Define

$$P_1 = H_1$$

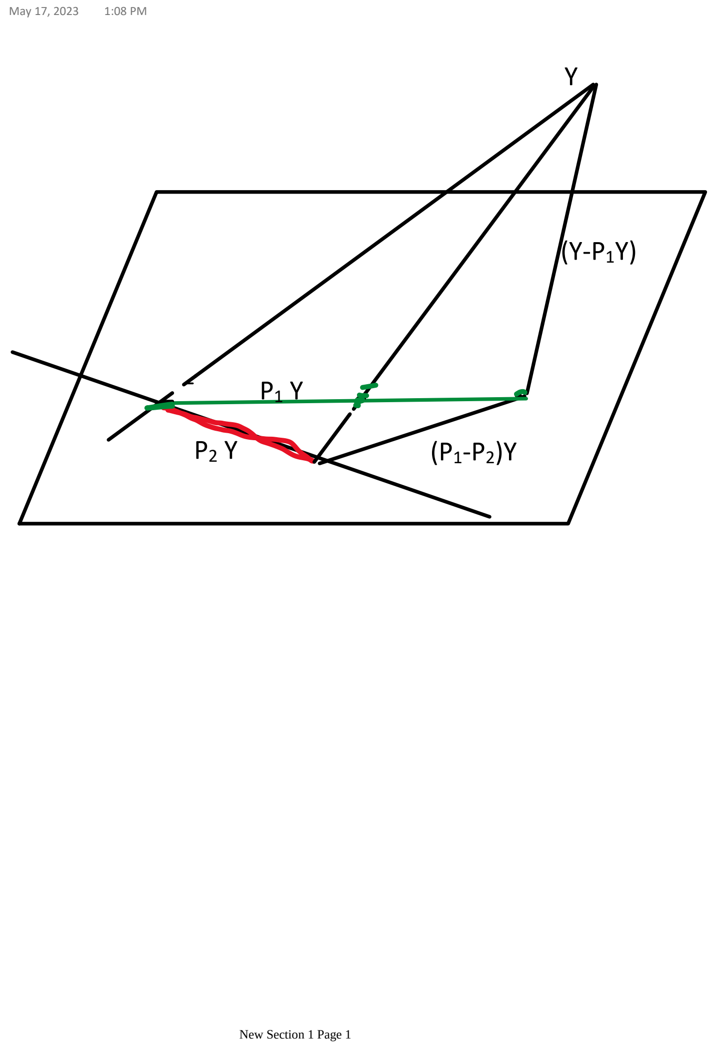
the projection matrix of Y on the $p+1$ dimensional space and let

$$P_2 = H_2$$

be the projection matrix of Y under H_0 on the $q+1$ dimensional space. The difference is used to construct the extra sum of squares. In fact

$$\|P_1 Y - P_2 Y\|^2 = C$$

Pictorially we have



By construction, $P_2'(P_1 - P_2) = 0$ since $P_2(P_1 Y) = P_2 Y$. This can be seen in the simple case when $p = 2$. The vectors 1 and $X - \bar{X}1$ are orthogonal and span P_1 The projection P_2 is spanned by 1.

We may compute

$$\begin{aligned} E(S_1 - S_2) &= E\{Y'(P_1 - P_2) Y\} \\ &= Trace(P_1 - P_2) + \mu'(P_1 - P_2)\mu \\ &= (p - q)\sigma^2 + 0 \end{aligned}$$

By repeated application of this principle, we can successively obtain for any regression model

$SS(b_0), SS(b_1|b_0), SS(b_2|b_1, b_0), ..., SS(b_p|b_{p-1}, ..., b_0)$

All these sums of squares are distributed as chi square with one degree of freedom independent of MSE. The tests are conducted using t tests.

3.2 Simultaneous confidence intervals

There are occasions when we require simultaneous or joint confidence intervals for the entire set of parameters. As an example, suppose we wish to obtain confidence intervals for both the intercept and the slope of a simple linear regression. Computed separately, we may obtain 95% confidence interval for each. If the statements are independent, then the probability that both statements are correct is given by $(0.95)^2 = 0.9025$. Moreover, the intervals make use of the same data and consequently, the events are not independent.

One approach that is frequently used begins with the Bonferroni inequality. For two events A_1, A_2

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &\leq P(A_1) + P(A_2) \end{aligned}$$

Consequently, using DeMorgan's identity

$$\begin{aligned} P(A_1' \cap A_2') &= 1 - P(A_1 \cup A_2) \\ &\geq 1 - P(A_1) - P(A_2) \end{aligned}$$

Suppose now that the events are such that

$$P(A_1) = P(A_2) = \alpha$$

and hence

$$\begin{aligned} P(A_1' \cap A_2') &\geq 1 - P(A_1) - P(A_2) \\ &\geq 1 - 2\alpha \end{aligned}$$

Now the event $(A_1' \cap A_2')$ is the event that the intervals

$$A_1' \quad : \quad b_0 \pm t\,(1-\alpha/2; n-2)\, s[b_0]$$

$$A_2' \quad : \quad b_1 \pm t\,(1-\alpha/2; n-2)\, s[b_1]$$

simultaneously cover β_0, β_1. If $\alpha = 0.05$, $1-2\alpha = 0.90$.

On the other hand, if we wish to have a confidence of 0.95 for the two intervals, then we should choose

$$1 - 2\alpha \;=\; 0.95$$

$$\alpha \;=\; 0.025$$

which implies we need to compute

$$t\,(0.9875; n-2)$$

In general, if p parameters are involved, then

$$P\left(\cap_i A_i'\right) \;\geq\; 1 - p\alpha^*$$

$$= \; 1 - \alpha$$

so that $\alpha^* = \frac{\alpha}{p}$ and each confidence interval has confidence $1 - \frac{\alpha}{p}$.

Calculations using R

a) **Model fitting**

Suppose we wish to fit a regression of a response against 4 variables X_1, X_2, X_3, X_4.

model=lm($Y \sim X_1 + X_2 + X_3 + X_4, data = CH$)

A more efficient command is

model=lm($Y \sim ., data = CH$)

If it is desired to exclude a specific variable from a long list of other variables to be included

model=lm($Y \sim .-X_1, data = CH$)

b) **ANOVA table**

after fitting a model, say Retailer, you may obtain an ANOVA table using the command

anova(RETAILER)

c) **Extra sum of squares**

The following R commands carry out the analysis for the extra sum of squares

Reduced=lm(y~x) # fits the reduced model

Full=lm(y~0+as.factor(x)) #fits the full model

anova(Reduced, Full) # gets the lack of fit test

d) **Simultaneous confidence intervals**

to obtain the t distribution cutoff

confint(fit, level=1-0.05/2)

Alternatively we can obtain the quantile,

qt(0.9875,n-2)

The Bonferroni intervals are often used because they provide shorter length confidence intervals than some other methods such as Scheffe.

3.3 R Session

We will use the Delivery Time data

a) **Graphic**

Delivery=read.table(file.choose(),header=TRUE,sep='\t')

names(Delivery)

[1] "Delivery.Time" "Number.of.Cases" "Distance"

plot(Delivery) #two-dimensional scatter plot

install.packages("plot3D") #install three-dimensional plot routine

library("plot3D")

x=Delivery$Number.of.Cases #define the variables

y=Delivery$Distance

z=Delivery$Delivery.Time

scatter3D(x,y,z,theta=15,phi=20,xlim=c(1,30),ylim=c(30,150)) #plot in 3D; many options are available

b) **Model fitting**

X1=Delivery$Number.of.Cases

X2=Delivery$Distance

Y=Delivery$Delivery.Time

model=lm(Y~X1+X2,data=Delivery)

summary(model)

Call: lm(formula = Y ~ X1 + X2, data = Delivery)

Residuals:

Min	1Q	Median	3Q	Max
-5.7880	-0.6629	0.4364	1.1566	7.4197

4 Model adequacy checking

After fitting a regression model, it is important to verify whether or not the assumptions that led to the analysis are satisfied.

The basic assumptions that were made were

1. $\{\epsilon_i\}$ are normally distributed

2. $E[\epsilon_i] = 0$ and $\sigma^2[\epsilon_i] = \sigma^2$

3. $\{\epsilon_i\}$ are independent

As well, we need to check for influential observations which may unduly influence the fitted model. Very large or very small values of the response may sometimes heavily alter the value of the estimated coefficients.

Definition 4.1. The basic tool that is used consists of analyzing the residuals

$$e_i = Y_i - \hat{Y}_i \tag{4.1}$$

4.1 Checking for normality

a) Box plots of residuals under normality should indicate a symmetric box around the median of 0

b) A histogram of the residuals provides a graphical check on normality

c) A qq- plot (i.e. quantile-quantile plot) consists of comparing the quantiles of the residual data with the quantiles from a normal distribution. This is a plot of the ranked residuals against the expected value under normality. Set

$$E_k = \sqrt{MSE}\Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right), k = 1, ..., n$$

Then plot $e_{(k)} vs E_k$ where $e_{(k)}$ is the residual with rank k. Under normality, one expects a straight line plot.

4.2 Checking for constancy of variance

We note that the variance and covariances of the residuals are respectively

$$\sigma^2[e_i] = \sigma^2[1 - h_{ii}]$$

$$Cov[e_i, e_j] = \sigma^2[1 - h_{ij}]$$

where h_{ij} is the ij^{th} element of the hat matrix. This demonstrates that the variances of the residuals are not equal. For this reason, we may define the Studentized or standardized residuals which have equal variance

$$e_i^* = \frac{e_i}{\sigma[e_i]} = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

where $s^2 = MSE$.

Definition 4.2. The semi studentized residuals are defined as

$$\frac{e_i}{\sqrt{1 - h_{ii}}}$$

A plot of the standardized residuals vs fitted values is a useful check for non constancy of variance. The plot should show a random distribution of the points. Alternatively, a non constancy of variance would appear as a telescoping increasing or decreasing collection of points.

plot(fit,3) #plots of standardized residuals e_i^* vs $\hat{Y}_i$

A scale-location plot can also be used to examine the homogeneity of the variance of the residuals. This is a plot of

$$\sqrt{|e_i^*|} vs \hat{Y}_i$$

Definition 4.3. Press residuals p.139

4.3 Residual plots against fitted values

If the residuals lie in a narrow band around 0 then there are no obvious needs for corrections.

If the residuals show a telescopting pattern, either increasing or decreasing, this is a sign that the variance is non constant.

A double-bow pattern is a sign that the variance in the middle is larger than the variance at the extremes as is the case for binomial data.

If the residuals exhibit a quadratic relationship, we may have a nonlinear relationship that has not been accounted for.

See p.144 for the graphics.

If some residuals are very large, they may arise from an outlier. Of course, they may be due to a non constant variance or a missing term.

4.4 Residual plots against the regressor

As in section 4.3 above, plots of the residuals against the independent variables are similarly interpreted.

Residuals may also be plotted against independent variables not in the model.

See p.146 in Montgomery et al for the graphics.

4.5 Residuals in time sequence plot

Such plots would reveal a time dependence is they appear as in section 4.3 above

p.148

4.6 Lack of fit of the regression model

See section 3.1 of these notes where the topic was addressed. See also section 4.5 of Montgomery et al

4.7 Calculations Using R

Suppose the data $(Y, X_1, ..., X_p)$ is stored in "file"

plot(file) # provides a scatter plot matrix

boxplot(y~x) # creates side by side boxplots

cor(file) # computes a correlation matrix

cor.test(x,y) # test plus confidence interval for rho

plot(fit,2) #qq plot of sqrt(standardized residuals) vs theoretical quantiles

Other plots may be obtained using

library(MASS)

sresid=studres(fit) #provides the Studentized residuals

hist(sresid,freq=False, main="Distribution of Studentized residuals")

sfit=seq(min(sresid),max(sresid),length=n)

yfit=dnorm(sfit)

lines(sfit,yfit)# superimposes a normal density on the histogram

Transformations

We may also try different transformations of Y to obtain a better fit