Mat 3375

Regression Analysis

Mayer Alvo

November 28, 2023

Contents

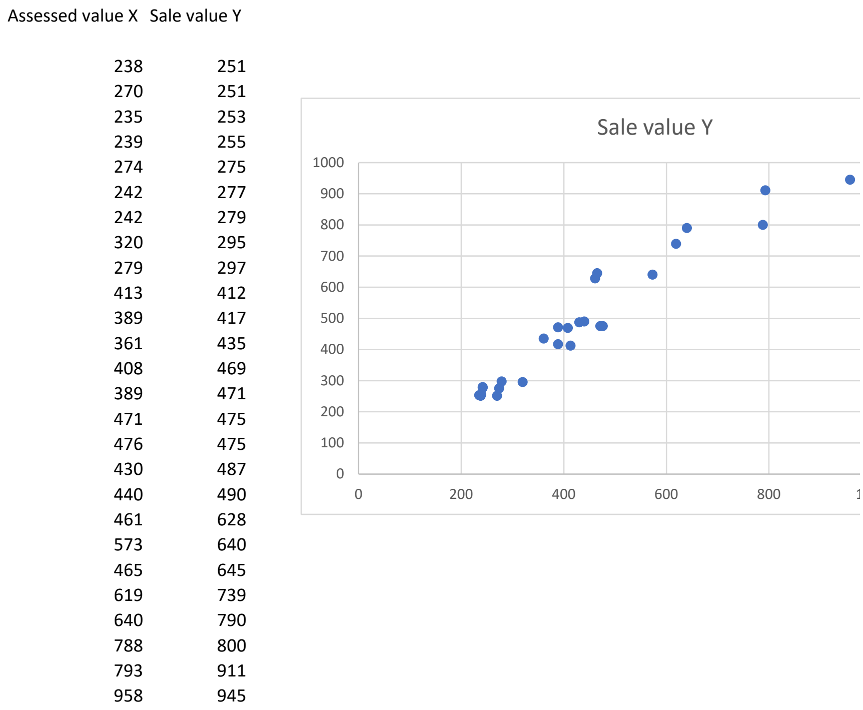
1 Introduction

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

$$\begin{aligned} b_0 &= \bar{Y} - b_1\bar{X} \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1.2) \\ b_1 &= \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2} \qquad (1.3) \\ &= \frac{\sum\left(X_i - \bar{X}\right)Y_i}{\sum\left(X_i - \bar{X}\right)^2} \qquad (1.4) \\ &= \sum k_i Y_i \end{aligned}$$

where

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

Then it can be shown

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

The equation of the fitted line is

$$\hat{Y} = b_0 + b_1 X \tag{1.5}$$

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, $\hat{\beta}_1 = \sum c_i Y_i$ which must satisfy

$$\begin{aligned} \beta_1 &= E[\hat{\beta}_1] \\ &= \sum c_i E[Y_i] \\ &= \sum c_i[\beta_0 + \beta_1 X_i] \end{aligned}$$

Hence, $\sum c_i = 0, \sum c_i X_i = 1$ 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 $Y_i=\beta_0+\beta_1X_i+\epsilon_i$ 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 $b_1=\sum k_iY_i$ 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 $Y_i\sim N(\beta_0+\beta_1X_i,\sigma^2)$, it follows that

$$\begin{aligned} b_1 &= \sum k_iY_i\sim N(\sum k_i(\beta_0+\beta_1X_i),\sigma^2\sum k_i^2) \\ &\sim N\left(\beta_1,\frac{\sigma^2}{\sum\left(X_i-\bar{X}\right)^2}\right)\end{aligned}$$

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

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

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

The ANOVA table indicates how one can test the null hypothesis

$$H_0 \;:\; \beta_1 = 0$$

$$H_1 \;:\; \beta_1 \neq 0$$

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* $ACT = 3.5$

1.5 Calculations using R

Computations for regression can be conveniently performed with the free software R. Here is a list of the most useful commands.

a) **Load the data**

Suppose we have data in a file in directory, usually in csv or text format . It can be read into R using the command

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

R will then open a window to browse for the document. To offset the default, the command header=TRUE indicates that we will name the columns.

finally, the command sep='t' is to use tabs to delimit the columns.

To verify that data is a data frame

```
is.data.frame(data)
[1] TRUE
```

We may display the names of the columns

```
names(data)
[1] ”length” ”Width”
```

To change names

```
names(data[1]=”volume”)
```

b) **Accessing the data**

To access the data length

```
data$length
```

c) **Descriptive measures**

```
mean(data$length)
```

We may assign

```
x=data$length
y=data$width
plot(x,y,ylab='width in inches',xlab='length in inches')
cor(x,y)
cor.test(x,y)
summary(data$length)
boxplot(data$length, data$width,names=c(”length”,”width”)
```

We can determine the number of rows and columns

```
nrow(data)
ncol(data)
```

d) **Graphics**

We can draw a histogram

```
hist(data$length,prob=TRUE,xlab='length',main='Density histogram of length')
```

To superimpose a normal density

```
curve(dnorm(x,mean(data$length),sd(data$length),add=TRUE)
```

e) **Fitting the model**

```
fit=lm(y~x)
fit
```

Alternatively, we may us the original names

```
fit=lm(width~length,data=data)
```

Regression without intercept

```
fit=lm(y~0+x,data=data) or
fit=lm(x-1,data=data)
```

f) **Confidence intervals**

The construction of confidence intervals can easily be done using the R commands once the model has been fitted.

Suppose that the data is labeled ACT

```
summary (fit) #provides the coefficients and their standard errors
```

For confidence intervals for the intercept

```
confint(fit,level=0.95)
```

g) For prediction of the mean for a new value say X_0

new.dat=data.frame(ACT=X_0)

```
predict(fit, newdata=new.dat, interval=”confidence”) # this provides a 95% confidence interval
```

For a prediction interval

```
predict(fit, newdata=new.dat, interval=”prediction”) # this provides a 95% prediction confidence interval
```

h) **Confidence bands using ggplot2**

```
ggplot(data,aes(x=ACT,y=GPA))+
geom_point()+
geom_smooth(method=lm,se=TRUE) #yields 95% confidence interval
temp_var=predict(fit,interval=’prediction’)
new_df=cbind(data,temp_var)
ggplot(new_df,aes(ACT,GPA))+
geom_point()+
geom_line(aes(y=lwr),color=’red’,linetype=’dashed’)+
geom_line(aes(y=upr),color=’red’,linetype=’dashed’)+
geom_smooth(method=lm,se=TRUE) #95% confidence and prediction intervals
```

1.6 R Session

1.6.1 Rocket Propellant data

To read the data

```
Rocket=read.table(file.choose(),header=TRUE,sep='\t')
Rocket #prints out the data
Shear.strength Age.of.Propellant
y=Rocket$Shear.strength
x=Rocket$Age.of.Propellant
plot(x,y)
hist(y,prob=TRUE,main='Density histogram of Shear Strength')
```

To conert data from table.b1 in R to csv format

```
write.csv(table.b1,'table.csv')# this will write the new table in the working directory of your computer
```

Regression model for Rocket data

```
fit=lm(y~x)
fit
Call: lm(formula = y ~ x)
Coefficients: (Intercept) x 2627.82 -37.15
summary(fit)
Call: lm(formula = y ~ x)
Residuals:
```

Min	1Q	Median	3Q	Max
-215.98	-50.68	28.74	66.61	106.76

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2627.822	44.184	59.48	< 2e-16 ***
x	37.154	2.889	-12.86	1.64e-10 ***

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared: 0.9018,
Adjusted R-squared: 0.8964
F-statistic: 165.4 on 1 and 18 DF, p-value: 1.643e-10
cor(x,y)
[1] -0.9496533
plot(Rocket$Age.of.Propellant, Rocket$hear.strength, xlab='Age',ylab='Shear Strength',mai
Propellant')
abline(Rocket,col='lightblue')
```

Regression without intercept

```
fit=lm(y~x-1,data=Rocket)
summary(fit)
Call: lm(formula = y ~ x - 1, data = Rocket)
Residuals:
```

Min	1Q	Median	3Q	Max
-1044.7	-497.6	742.3	1529.4	2428.2

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
x	112.98	19.22	5.878	1.16e-05 ***

```
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1315 on 19 degrees of freedom
Multiple R-squared: 0.6452,
Adjusted R-squared: 0.6265
F-statistic: 34.55 on 1 and 19 DF, p-value: 1.165e-05
```

plots using standard R routines

plots using ggplot

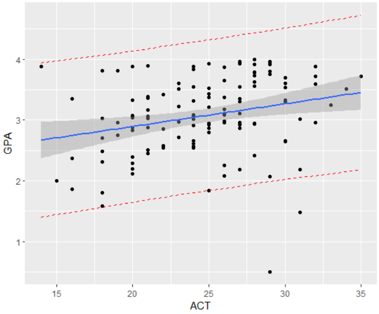
```
ggplot(Rocket,aes(x,y))+
+ geom_point()+
+ geom_smooth(method=lm,se=TRUE)
```

1.6.2 R Session for Plumbing Supplies data

Use this data to fit a regression without intercept

1.6.3 R Session for the GPA data

```
data=read.table(file.choose(),header=TRUE,sep='\t')
    names(data)
    [1] "GPA" "ACT"
    fit=lm(GPA~ACT,data=data)
    fit
    Call: lm(formula = GPA ~ ACT, data = data)
    Coefficients: (Intercept) ACT 2.14596 0.03735
    ggplot(data,aes(x=ACT,y=GPA))+
    geom_point()+
    geom_smooth(method=lm,se=TRUE) #yields 95% confidence interval
    temp_var=predict(fit,interval='prediction')
    new_df=cbind(data,temp_var)
    ggplot(new_df,aes(ACT,GPA))+
    geom_point()+
    geom_line(aes(y=lwr),color='red',linetype='dashed')+
    geom_line(aes(y=upr),color='red',linetype='dashed')+
    geom_smooth(method=lm,se=TRUE) #95% confidence and prediction intervals
```



1.7 Other DATA SETS

Rocket Propellant Data
Delivery Time Data
Patient Satisfaction Data
ACT Scores
Airfreight Data
Copier Maintenance Data
Crime Data
Toluca Refrigeration Data
Grocery Retailer Data
Plumbing Supples Data

1.7.1 Homework

Problems 2.1, 2.10, 2.22

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

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

where

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

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

A fundamental result is

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

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

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

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

Hence the solutions to the normal equations are

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

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

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

But

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

and

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

Hence,

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

The fitted line is then

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

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

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

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 $N\left(\mu,\sigma^2\right)$. 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,

$$A = P'\varLambda P.$$

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

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

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

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

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

$$\begin{aligned} E\left[Y'AY\right] &= Trace\left[A\Sigma\right] + (EY)'\,(EY) \\ &= \sigma^2 TraceA + \mu^2 1'A1 \\ &= \sigma^2\left(n-1\right) + 0 \end{aligned}$$

□

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

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

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,

$$Y'HY \sim \chi_p^2, Y'(I - H)Y \sim \chi_{n-p}^2$$

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

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

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

$$\begin{aligned} H_0 &: E[Y] = \beta_0 + \beta_1 X \\ H_1 &: E[Y] \neq \beta_0 + \beta_1 X \end{aligned}$$

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	$SSR\left(X_1,X_2,X_3\right)$	3	$MSR\left(X_1,X_2,X_3\right)$
X_1	$SSR\left(X_1\right)$	1	$MSR\left(X_1\right)$
$X_2\vert X_1$	$SSR\left(X_2\vert X_1\right)$	1	$MSR\left(X_2\vert X_1\right)$
$X_3\vert X_1,X_2$	$SSR\left(X_3\vert X_1,X_2\right)$	1	$MSR\left(X_3\vert X_1,X_2\right)$
Error	$SSE\left(X_1,X_2,X_3\right)$	n-4	$MSE\left(X_1,X_2,X_3\right)$
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

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

In general, suppose that

$$S_1 = SS\left(b_0\left(1\right),...,b_p\left(1\right)\right)$$

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

$$S_2 = SS\left(b_0\left(2\right),...,b_q\left(2\right)\right)$$

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

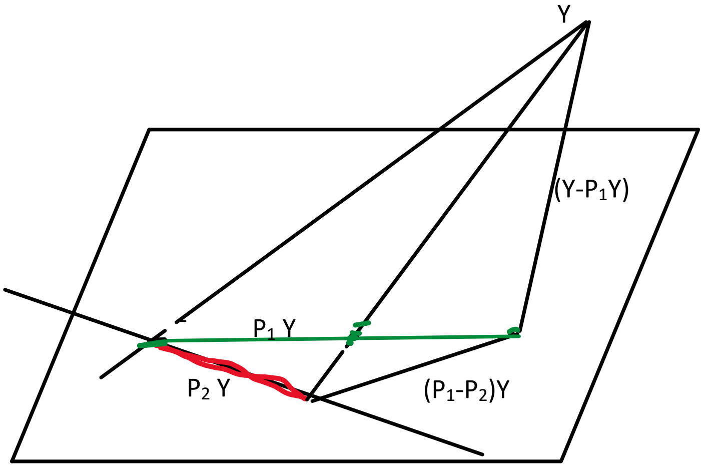
$$P_2 = H_2$$

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

$$\left\|P_1Y - P_2Y\right\|^2 = C$$

Pictorially we have

May 17, 2023 1:08 PM



New Section 1 Page 1

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

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

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

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

Consequently, using DeMorgan's identity

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

Suppose now that the events are such that

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

and hence

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

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

$$\begin{aligned} A_1' &: b_0 \pm t\,(1-\alpha/2; n-2)\, s[b_0] \\ A_2' &: b_1 \pm t\,(1-\alpha/2; n-2)\, s[b_1] \end{aligned}$$

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

$$\begin{aligned} 1-2\alpha &= 0.95 \\ \alpha &= 0.025 \end{aligned}$$

which implies we need to compute

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

In general, if p parameters are involved, then

$$\begin{aligned} P\left(\cap_i A_i'\right) &\geq 1 - p\alpha^* \\ &= 1-\alpha \end{aligned}$$

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

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.341231	1.096730	2.135	0.044170 *
X1	1.615907	0.170735	9.464	3.25e-09 ***
X2	0.014385	0.003613	3.981	0.000631 ***

--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 22 degrees of freedom

Multiple R-squared: 0.9596,

Adjusted R-squared: 0.9559

F-statistic: 261.2 on 2 and 22 DF,

p-value: 4.687e-16

c) **ANOVA**

In R, the Default function anova in R provides sequential sum of squares (type I) sum of squares. That is to say, each sum of squares for each variable is the sum of squares conditional on the previous variables is.

To obtain the regula analysis of variance, Use package car to get type II sum of square. The output looks similar to Minitab output.

```
>library(car)
>Anova(mylm, type="II")
anova(model)
```

Analysis of Variance Table

Response: Y

	DF	Sum Sq	MeanSq	F value	Pr(>F)			
X1	1	5382.4	5382.4	506.619	< 2.2e-16	***		
X2	1	168.4	168.4	15.851	0.0006312	***		
Residuals	22	233.7	10.6					

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

d) **Extra Sum of Squares**

```
Full =lm(Y~X1+X2,data=Delivery)
Reduced=lm(Y~X1)
anova(Reduced, Full)
```

Analysis of Variance Table

Model 1: Y ~ X1

Model 2: Y ~ X1 + X2

Res	Df	RSS	DF	Sum of Sq	F	Pr(>F)
1	23	402.13				
2	22	233.73	1	168.4	15.851	0.0006312 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

3.4 DATA SETS

Patient Satisfaction Data
    Grocer Retailer Data
    Delivery Time Data

3.5 Suggested Problems

Problem 3.7

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

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

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

$$\begin{aligned} \sigma^2\left[e_i\right] &= \sigma^2\left[1-h_{ii}\right] \\ Cov\left[e_i,e_j\right] &= \sigma^2\left[1-h_{ij}\right] \end{aligned}$$

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

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

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 boxcox(fit) # determines the value of λ so that $Z = Y^\lambda$ is normally distributed

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

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

4.8 R Session

We will use the Delivery time data

```
boxplot(X1,X2)
fit=lm(Y~X1+X2,data=Delivery)
plot(fit,2) #normal QQ plot of sqrt(standardized residuals) vs theoretical quantiles
library(MASS)
sresid=studres(fit)
hist(sresid,freq=FALSE)
sfit=seq(min(sresid),max(sresid),length=25)
yfit=dnorm(sfit)
lines(sfit,yfit) #superimposes a normal density on the histogram
```

Alternatively

```
x=fit$residuals
curve(dnorm(x,mean(x),sd(x)),add=T)
boxcox(fit) # BoxCox transformation for normality
```

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

```
plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot: Hit <Return> to see next plot:
Hit <Return> to see next plot:
```

To see all 4 plots in a single page

```
par(mfrow=c(2,2))
plot(fit)
z=fit$residuals
qqnorm(z) # qq plot
qqline(z) # normal line superimposed
```

4.9 DATA SETS

Restaurant Data

Rocket Propellant Data

Electric Utility Data

Windmill Data

5 Regression Diagnostics

We begin by considering transformations to linearize the model and weighting corrections for violation of the variance assumption.

5.1 Transformations and weighting

5.1.1 Variance stabilizing transformations

It may happen that the variance in the general linear model is not constant. In those cases, it may be useful to transform the data. Here are some examples of transformations.

(a) Poisson In the case of the Poisson, the variance is equal to the mean. Bartlett (Anscombe) showed that if Y is distributed as a Poisson variable with mean λ, then $\sqrt{Y}$ is distributed more nearly normally with variance approximately 1/4 if λ is large. (The demonstration is through a Taylor series expansion of $\sqrt{Y}$). In that case, $\sqrt{Y}$ may be used and regressed against X.

(b) The similar transformation for a binomial variable $Y \sim B(p,n)$, with mean $m = np$ is

$$sin^{-1}\sqrt{\left(\frac{Y+c}{n+2c}\right)}$$

The optimal value of c is 3/8 if m and $n-m$ are large. The variance is approximately $\frac{1}{4}\left(n+\frac{1}{2}\right)^{-1}$.

5.1.2 Transformations to linearize the model

It may happen that a plot of Y against X does not appear linear. In those cases, it may be useful to transform the dependent variable. Here are some examples of transformations.

(a) The exponential model

$$Y = \beta_0 e^{\beta_1 X}\epsilon$$

may be transformed by taking logs

$$lnY = ln\beta_0 + \beta_1 X + ln\epsilon$$

The usual assumptions would then have to be made and verified on the transformed model.

(b) The model

$$Y = \beta_0 + \beta_1 X^{-1} + \epsilon$$

can be linearized using the reciprocal transformation $X^* = X^{-1}$.

(c) The model

$$\frac{1}{Y} = \beta_0 + \beta_1 X + \epsilon$$

can be linearized using the reciprocal transformation $Y^* = Y^{-1}$.

(d) The model

$$Y = \frac{X}{\beta_0 + \beta_1 X}$$

can be linearized using the reciprocal transformation in two steps. First

$$Y^* = Y^{-1}$$

and then

$$X^* = X^{-1}$$

to obtain

$$Y^* = \beta_1 + \beta_0 X^*$$

5.1.3 Box-Cox transformations

At times, the data may not appear to be normally distributed. Box-Cox suggested a power transformation of the type

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \\ \dot{Y} lnY & \lambda = 0 \end{cases} \tag{5.1}$$

where

$$\dot{Y} = ln^{-1} \left[\frac{\sum lnY_i}{n} \right]$$

The value of λ is usually determined by trial and error whereby a model is fitted to $Y^{(\lambda)}$ for various values of λ and selecting the one which minimizes the residual sum of squares from a graphic plot.

We note as well that a confidence interval can be constructed for λ . This is useful in that one may select a simple value of λ which is in the interval such $\lambda = 0.5$. (see 5.4.1 p.189 in Montgomery et al.)

The theory behind the transformation is as follows.

The original model was $Y \sim N_n\left(X\beta, \sigma^2 I\right)$

The transformed model is $Y^{(\lambda)}$ therefore has likelihood function given by

$$\frac{1}{(2\pi)^{n/2}\sigma^n} exp\left\{-\frac{\left(y^{(\lambda)} - X\theta\right)'\left(y^{(\lambda)} - X\theta\right)}{2\sigma^2}\right\} J\left(\lambda; y\right)$$

with the new parameters θ and Jacobian for the transformation

$$J\left(\lambda; y\right) = \Pi_{i=1}^{n}\left|\frac{dy_i^{(\lambda)}}{dy_i}\right|$$

The maximum likelihood estimator of the variance is given by

$$\begin{aligned}\hat{\sigma}^2 &= \frac{Y^{(\lambda)\prime}\left[I - X\left(X'X\right)^{-1}X'\right]Y^{(\lambda)}}{n} \\ &\equiv \frac{S\left(\lambda\right)}{n}\end{aligned}$$

The maximized log likelihood for fixed λ is

$$\begin{aligned}L_{max}\left(\lambda\right) &= -\frac{n}{2}log\hat{\sigma}^2 + logJ\left(\lambda; y\right) \\ &= -\frac{n}{2}log\hat{\sigma}^2 + \left(\lambda - 1\right)\sum logy_i\end{aligned}$$

under the proposed (5.1).

We may then plot $L_{max}\left(\lambda\right)$ vs λ to find the value of λ which yields the maximum.

The exact value of the maximizing λ can be determined by differentiating the above $L_{max}\left(\lambda\right)$ and then solving numerically for λ.

5.2 Weighted least squares

Suppose that the error terms are such that $\sigma^2\left[\epsilon_i\right] = \sigma_i^2$. Instead of minimizing the sum of the square errors, we may minimize the sum of the weighted squared errors

$$\sum w_i\epsilon_i^2$$

where the weights satisfy

$$\sigma^2\left[\sqrt{w_i}\epsilon_i\right] = \sigma^2$$

Different weights may be chosen as for example, $w_i = \sqrt{X_i}$ or $w = \sqrt{Y}$.

The theory proceeds as follows.

Define the matrix W of weights

$$W = \begin{pmatrix} w_1 & 0 & & 0 \\ 0 & w_2 & & \\ & & & \\ & & & \\ 0 & & & w_n \end{pmatrix}$$

Then the original model goes from

$$Y = X\beta + \varepsilon$$

to

$$\begin{aligned} W^{1/2}Y &= W^{1/2}X\beta + W^{1/2}\varepsilon \\ Y_W &\equiv X_W\beta + \varepsilon_W \end{aligned}$$

The least squares approach leads to

$$\begin{aligned} b_W &= \left(X_W' X_W\right)^{-1} X_W' Y_W \\ &= \left(X'WX\right)^{-1} X'WY \end{aligned}$$

The MSE$_W$ becomes

$$\begin{aligned} MSE_W &= \frac{\sum w_i \left(Y_i - \hat{Y}_i\right)^2}{n-p} \\ &= \frac{\sum w_i \left(e_i\right)^2}{n-p} \end{aligned}$$

One way to proceed is to perform the usual regression. Then, group the data using the X variable. Estimate the variances s_i^2 the Y_i for each group. Then fit the variances against the averages of the X$_i$ of the groups. We illustrate this approach with the Turkey data.

The Breusch-Pagen test for constancy of variance assumes the model

$$Log\ \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

and we wish to test

$$H_0 : \gamma_1 = 0$$

We regress the squared residuals e_i^2 against X_i in the usual way and calculate an error sum of squares SSR^* . We reject th null hypothesis that $\gamma_1 = 0$ whenever

$$\chi_1^2 \equiv \frac{\frac{SSR^*}{2}}{\frac{SSE}{n}}$$

is larger than the chi square critical value. This test can be conducted in R using

```
library(lmtest)

bptest(fit) #Breusch-Pagen test
```

Example 5.1. Commercial properties

a) Obtain boxplots for each variable

b) scatter plot matrix

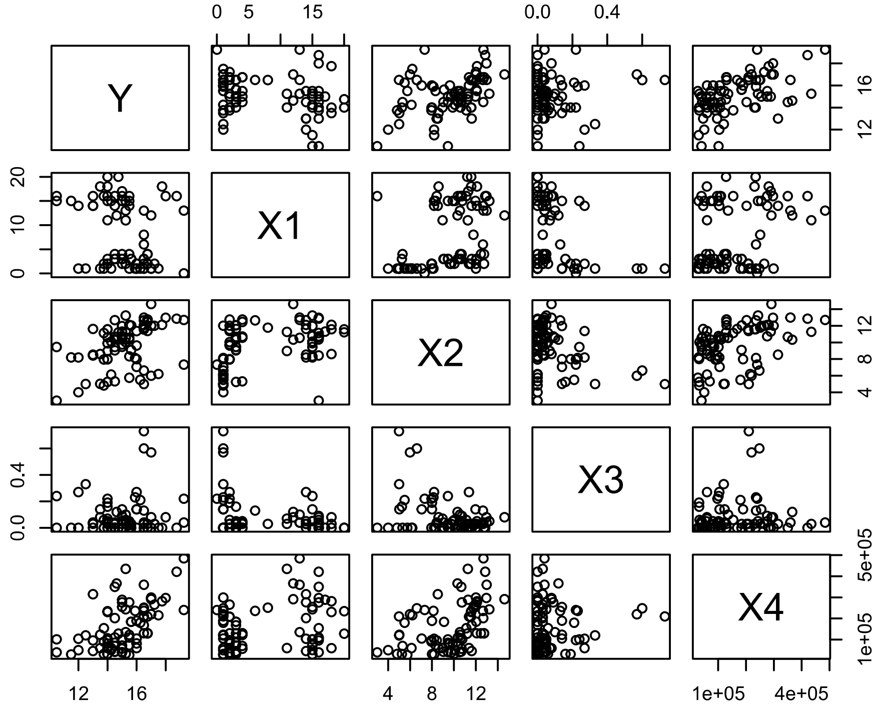
c) Fit the regression model

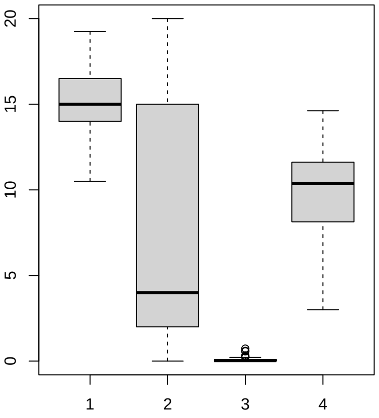
d) Obtain the residuals and compute a boxplot

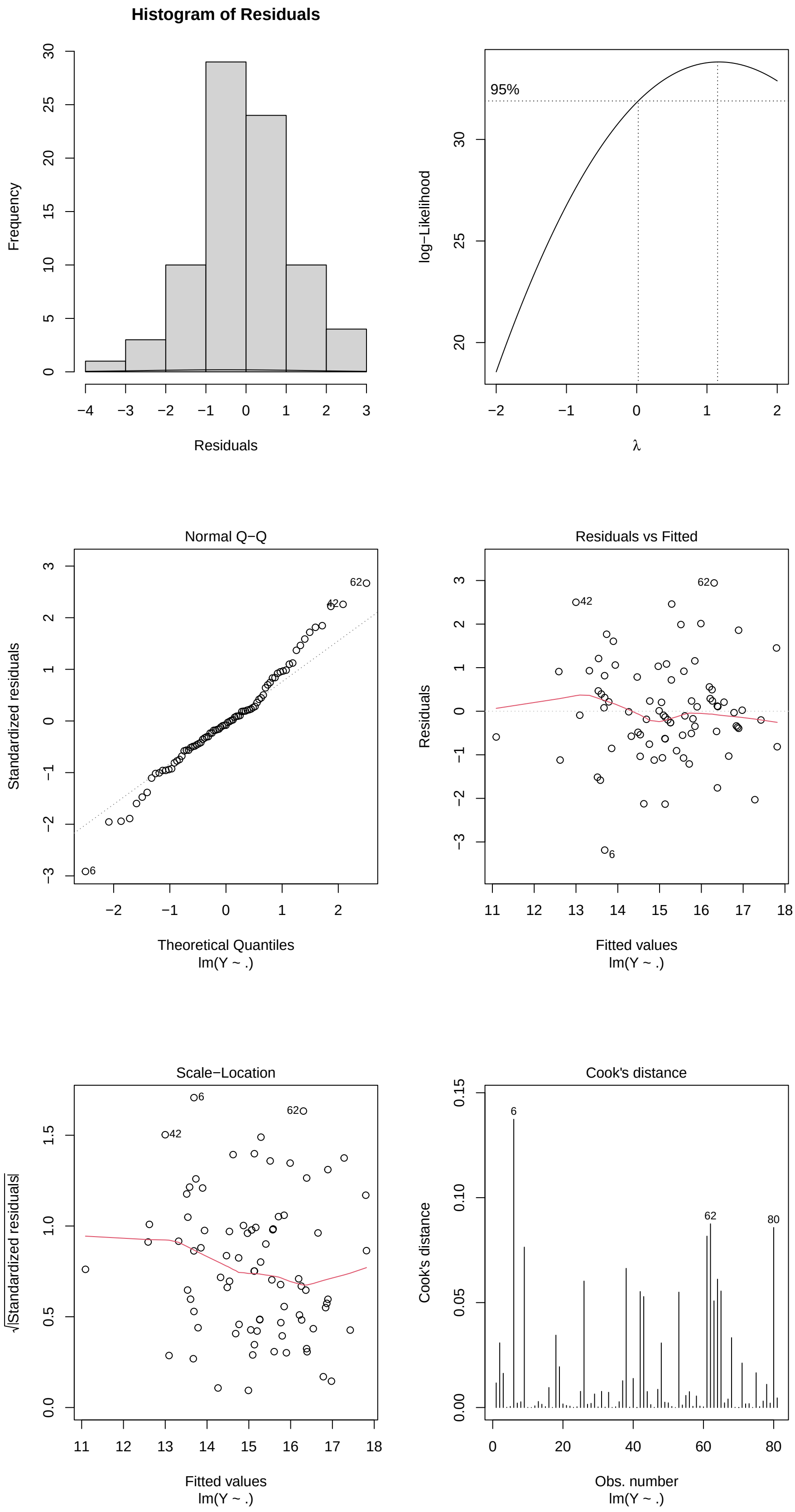
e) Plot residuals against each predictor variable

f) constancy of variance test

```
library(lmtest)

bptest(fit)

studentized Breusch-Pagan test

data: fit BP = 12.978, df = 4, p-value = 0.01139
```

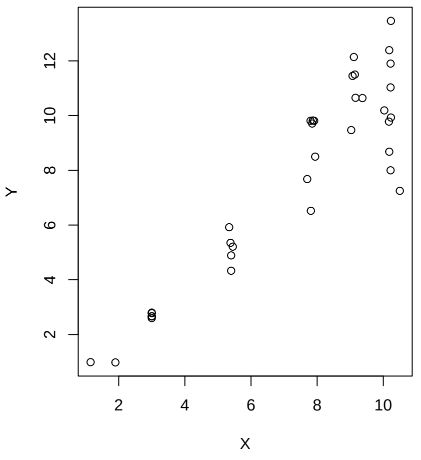

Y
X1
X2
X3
X4
0
5
15
0.0
0.4
12
16
20
10
0
4
8
12
0.4
0.0
5e+05
1e+05
12
16
4
8
12
1e+05
4e+05


20
15
10
5
0
1
2
3
4


Histogram of Residuals
Frequency
0
5
10
15
20
25
30
−4
−3
−2
−1
0
1
2
3
Residuals
log−Likelihood
95%
20
25
30
−2
−1
0
1
2
λ
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ .)
62
42
6
Residuals vs Fitted
Residuals
Fitted values
lm(Y ~ .)
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
Fitted values
lm(Y ~ .)
Cook's distance
0.00
0.05
0.10
0.15
0
20
40
60
80
Obs. number
lm(Y ~ .)
80

Example 5.2. Weighted least squares data

We consider the weighted least squares turkey data.



It can be seen that there is a telescoping effect.

Next we computed averages and variances for subsets of the data and then fitted the variances against the averages

$\bar{X}_i$	3.0	5.4	7.8	9.1	10.2
s_j^2	0.0072	0.3440	1.7404	0.8683	3.8964

$\hat{s}^2 = 1.5329 - 0.7334\bar{X} + 0.0883\bar{X}^2$

The weights were then computed as inverses of the variances. The unweighted regression was

$$Y = -0.579 + 1.14X$$

The weighted regression was

$$Y = -0.892 + 1.16X$$

Other R commands for comparisons

```
fit=lm(Y~X, data=Weighted)
wls_model <- lm(Y ~ X, data = Weighted, weights=W)
plot(fit,1)
plot(wls_model,1)
plot(fit,2)
plot(wls_model,2)
plot(fit,3)
plot(wls_model,3)
```

5.3 Checking on the linear relationship assumption

5.3.1 Descriptive Measures of Linear Association

We may define the coefficient of determination

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

This coefficient may be interpreted as the proportion of explained variation by the regression model. The larger the proportion the better the model is.

Care must be taken however when using this measure because a large value of R^2 may arise if the points lie on a quadratic.

5.4 Calculations Using R

a) **Plots**

A plot of the residuals vs fitted values can be used to check on the linearity assumption. The residuals should hover around 0 and appear random. A plot that exhibits a pattern presents a flag that perhaps one or more terms are missing from the fit.

plot(fit,1) # plots the residuals vs $\hat{Y}$

R can be used to plot 4 plots on a 2 x 2 layout as follows

par(mfrow=c(2,2))

plot(fit) # 4 plots will appear on a single page. It will include fits 1,2,3.

(fit,1) plots e vs $\hat{Y}$ is used to check the linear assumptions

A horizontal line with no pattern is indicative of a good fit

(fit,2) this is a qq normal plot used to examine if the residuals are normally distributed. Normality is accepted if we see a ”straight” line

(fit,3) This is a scale-location plot: $\sqrt{Standardizedresidual}$ vs $\hat{Y}$ the fitted values. It is used to check homogeneity of variance of the residuals. A horizontal line with equally spread points is a good indication that the variance is constant (homoscedasticity)

(fit,4) shows the standardized residuals vs leverage which is used to flag influential observations

plot(fit,5) shows Cook’s distance for the 3 most extreme values. If you want the top 5 extreme values

plot(fit,4,id.n=5)

model.diag.metrics%,% top_n(3,wt=cooksd)

A plot of the residuals against the time ordered sequence may reveal dependencies. A formal test using the Durbin-Watson statistic can be used to check for autocorrelated errors.

b) Box Cox transformation

Using R

boxcox(model) # determines the value of λ so that $Z = Y^{\lambda}$ is normally distributed

```
lambda <- b$x[which.max(b$y)] # Exact lambda
lambda
```

c) **Durbin Watson test**

```
durbinWatsonTest(fit) # tests for autocorrelated errors
```

d) **Test on constancy of variance**

```
library(lmtest)
bptest(fit) #Breusch-Pagen test
```

e) Installing olsrr

```
install.packages("olsrr")
library(olsrr)
ols_plot_dfbetas(fit) #plot of dfbetas
ols_plot_dffits(fit)# plot of dffits
```

5.5 R Session

We use the Delivery data

a) **Breusch Pagen test**

```
bptest(fit)
studentized Breusch-Pagan test
data: fit BP = 11.988, df = 2, p-value = 0.002493
```

b) **Weighted regression**

```
wls_model <- lm(Y ~ X1+X2, data = Delivery, weights=1/Y)
summary(wls_model )
Call: lm(formula = Y ~ X1 + X2, data = Delivery, weights = 1/Y)
Weighted Residuals:
```

Min	1Q	Median	3Q	Max
-1.04397	-0.26730	0.00011	0.23581	1.33217

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.622929	0.903842	4.008	0.000591 ***	
X1	1.583045	0.163823	9.663	2.24e-09 ***	
X2	0.011142	0.003123	3.567	0.001722 **	

```
--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.6156 on 22 degrees of freedom
Multiple R-squared: 0.9369, Adjusted R-squared: 0.9312
F-statistic: 163.3 on 2 and 22 DF, p-value: 6.321e-14
```

c) **Cook’s distance**

```
plot(fit,5)
plot(fit,4,id.n=5)
```

d) **Installing olsrr**

```
install.packages("olsrr")
library(olsrr)
ols_plot_dfbetas(fit) #plot of dfbetas
ols_plot_dffits(fit)# plot of dffits
```

5.6 Data Sets

Commercial Properties

Weighted least squares data

Restaurant Data

Electric Utility Data

Windmill Data

Plumbing Supplies Data

5.7 R session commercial properties

```
library(olsrr)
    Properties=read.table(file.choose(),header=TRUE,sep=’\t’)
    plot(Properties)
    fit=lm(Y~.,data=Properties)
    RentalRates=Properties$Y
    Age=Properties$X1
    Vacancyrates=Properties$X2
    Vacancyrates=Properties$X3
    Expenses=Properties$X2
    Footage=Properties$X4
    boxplot(RentalRates,Age,Vacancyrates,Expenses)
    boxplot(Vacancyrates)
    boxplot(Footage)
```

6 Diagnostics for Leverage and Measures of Influence

A single observation may unduly influence the results of a regression analysis. Hence, the detection of such influential observations is important. In this connection the hat matrix plays a very important role. We begin with the minimized sum of squares

$$\begin{aligned} R\left(\hat{\beta}\right) &= \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) \\ &= Y'Y - Y'X\left(X'X\right)^{-1}X'Y \\ &= Y'\left(I - X\left(X'X\right)^{-1}X'\right)Y \\ &= Y'\left(I - H\right)Y \end{aligned}$$

where $H = X\left(X'X\right)^{-1}X'$. Let x_i' being the i^{th} row of X. Then the i^{th} diagonal of H is for $i = 1, ..., n$

$$h_{ii} = x_i'\left(X'X\right)^{-1}x_i$$

In the simple linear regression with $p = 2$, $x_i' = [1, X_i]$

$$\begin{aligned} h_{ii} &= \frac{x_i'\begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}x_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ &= \frac{\sum X_i^2 - 2X_i\sum X_i + nX_i^2}{n\sum X_i^2 - \left(\sum X_i\right)^2} \qquad (6.1) \\ &= \frac{\sum X_i^2 - n\bar{X}^2 + n\bar{X}^2 - 2nX_i\bar{X} + nX_i^2}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ &= \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)^2} \end{aligned}$$

Therefore, h_{ii} is a measure of how far the i^{th} observation is from the mean. If $X_i = \bar{X}$, then $h_{ii} = \frac{1}{n}$ which is the minimum value.

Definition 6.1. The quantity h_{ii} is called the leverage of the i^{th} observation.

A further insight is gained by writing the mean $\bar{X}$ in terms of the mean $\bar{X}_{(i)}$ when the i^{th} observation is deleted. We can show

$$\bar{X} = \frac{1}{n}\left(X_i + (n-1)\,\bar{X}_{(i)}\right)$$

so that

$$\begin{aligned} X_i - \bar{X} &= X_i - \frac{1}{n}\left[X_i + (n-1)\,\bar{X}_{(i)}\right] \\ &= \frac{n-1}{n}\left[X_i - \bar{X}_{(i)}\right] \end{aligned}$$

Hence,

$$\begin{aligned} h_{ii} &= \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)^2} \\ &= \frac{1}{n} + \left(\frac{n-1}{n}\right)^2 \frac{\left(X_i - \bar{X}_{(i)}\right)^2}{\sum\left(X_i - \bar{X}\right)^2} \end{aligned}$$

This shows that the leverage of the i^{th} observation will be large if X_i is far from the mean of the other observations. So the leverage is concerned with the location of points in the space of the independent variables which may be influential.

6.1 Properties of the leverage

The leverage from (6.1) can be utilized to flag influential observations. This follows from the fact that

$$\begin{aligned} Trace\ H &= Tr\left[X\left(X'X\right)^{-1}X'\right] \\ &= Tr\left[\left(X'X\right)^{-1}X'X\right] \qquad (6.2) \\ &= Tr\left[I_p\right] = p \end{aligned}$$

and hence the average

$$\frac{\sum h_{ii}}{n} = \frac{p}{n}$$

Consequently, observations with a value of the leverage greater than twice the average should be flagged i.e.

$$h_{ii} > 2\left(\frac{p}{n}\right)$$

We note that not all points with high leverage will be influential. A point with high leverage may lie close to the regression line and hence will not be influential on the fit. On the other hand, a point with high leverage may be quite far from the fitted line and may be quite influential.

It is usually important to look at the studentized residuals in conjunction with the leverage. Observations with large leverage and large residuals are likely to be influential.

6.1.1 DFFITS

A useful measure of the influence that case i has on the fitted value $\hat{Y}_i$ is given by

$$\begin{aligned} DFFITS_i &= \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} \qquad (6.3) \\ &= t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} \end{aligned}$$

where

$$t_i = e_i \left[\frac{n-p-1}{SSE\,(1-h_{ii}) - e_i^2}\right]^{\frac{1}{2}}$$

represents the Studentized residual. This shows that it can be calculated from the original residuals, the error sum of squares and the hat matrix values. The value of $DFFITS_i$ represents the number of estimated standard deviations of $\hat{Y}_i$ that the fitted value increases or decreases with the inclusion of the i^{th} case in fitting the regression model.

If case i is an X outlier and has high leverage, then $\left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} > 1$ and $DFFITS$ will be large in absolute value. As a guideline, influential cases are flagged if

$$|DFFITS_i| > 1$$

for small to medium data sets and

$$|DFFITS_i| > 2\sqrt{\frac{p}{n}}$$

for large data sets.

6.1.2 Cook's Distance

Unlike the previous section which considered the influence of the i^{th} case on the fitted value $\hat{Y}_i$. Cook's distance considers the influence of the i^{th} case on the entire collection of n fitted values

$$\begin{aligned} D_i &= \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE} \qquad (6.4) \\ &= \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \end{aligned}$$

Cook's distance is a function of the residual e_i and the leverage h_{ii}. It can be large if either the residual is large and the leverage moderate, or if the residual is moderate and the leverage is large, or both are large. It can be shown that approximately

$$D_i \cong F(p, n-p)$$

Since $F_{0.50}\,(p, n-p)\cong 1$, we consider points for which $D_i > 1$ to be influential. Ideally, we want the estimated $\hat{\beta}_{(i)}$to be within the boundary of the $10-20\%$ confidence region. In R these regions are indicated in red.

6.1.3 DFBETAS

DFBETAS are a measure for the influence that case i has on each of the regression coefficients $b_k, k = 0, 1, ..., p-1$.

$$DFBETAS_{(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{ii}}}$$

where c_{ii} is the i^{th} diagonal element of $(X'X)^{-1}$. The $MSE_{(i)}$ may be computed from the relationship

$$(n-p)\,MSE = (n-p-1)\,MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

A large value of $DFBETAS_{(i)}$ indicates a large impact of the i^{th} case on the k^{th} regression coefficient. As a guideline

$$DFBETAS_{(i)} > \begin{cases} \frac{2}{\sqrt{n}} & large\, n \\ 1 & small\, n \end{cases}$$

6.1.4 Deletion of Observations: theoretical developments

In this subsection we present the theoretical developments underlying the calculations of the diagnostic measures above. This development is due to A.C. Atkinson Plots, transformations and regression (1987).

We begin with a matrix identity. Let A be a $p \times p$ square matrix and let U, V be $p \times m$ matrices. Then

$$(A - UV')^{-1} = A^{-1} + A^{-1}U\left(I - V'A^{-1}U\right)^{-1} V'A^{-1} \tag{6.5}$$

where I is the $m \times m$ identity matrix. The matrix identity is verified by multiplying the right hand side by $(A - UV')$.

Consider the partition of the X matrix

$$X = \begin{pmatrix} X_{(M)} \\ \\ X_M \end{pmatrix}$$

where $X_{(M)}$ is the reduced matrix when M rows are deleted and X_M is the matrix containing the deleted rows. Similarly, let

$$Y = \begin{pmatrix} Y_{(M)} \\ \\ Y_M \end{pmatrix}$$

Then,

$$\begin{aligned} X'X &= X'_{(M)}X_{(M)} + X'_M X_M \\ Y'Y &= Y'_{(M)}Y_{(M)} + Y'_M Y_M \\ X'Y &= X'_{(M)}Y_{(M)} + X'_M Y_M \end{aligned}$$

Setting $A = X'X$ and $U' = V' = X_M$ in (6.5) we have

$$\left(X'_{(M)}X_M\right)^{-1} = (X'X)^{-1} + (X'X)^{-1} X'_M (I - H_M)^{-1} X_M (X'X)^{-1}$$

where

$$H_M = X_M (X'X)^{-1} X'_M$$

is the hat matrix for the observations left out.

If M observations are left out and the model is refitted, the least squares estimate becomes

$$\begin{aligned}\hat{\beta}_{(M)} &= \left(X'_{(M)}X_{(M)}\right)^{-1} X'_{(M)}Y_{(M)} && (6.6)\\ & && (6.7)\\ &= \left(X'_{(M)}X_{(M)}\right)^{-1}\left(X'Y - X'_M Y_M\right) && (6.8)\end{aligned}$$

Substituting the expression for $\left(X'_{(M)}X_M\right)^{-1}$ above, we have

$$\hat{\beta}_{(M)} = \left[(X'X)^{-1} + (X'X)^{-1}X'_M\left(I - H_M\right)^{-1}X_M\left(X'X\right)^{-1}\right]\left(X'Y - X'_M Y_M\right)$$

Let $\hat{Y}_M = X_M\hat{\beta}$ be the estimate of response for the M values left out and set $e_M = Y_M - \hat{Y}_M$ be the corresponding vector of residuals.

A little algebra leads to the expression

$$\hat{\beta}_{(M)} - \hat{\beta} = -\left(X'X\right)^{-1}X'_M\left(I - H_M\right)^{-1}e_M \qquad (6.9)$$

The expression (6.9) shows the change in the parameter estimate when M data are deleted. When $M = 1$,

$$\hat{\beta}_{(i)} - \hat{\beta} = \frac{-\left(X'X\right)^{-1}x_i e_i}{1 - h_{ii}}$$

We may also compute the effect of deletion on the error sum of squares. The error sum of squares for the full model is

$$SSE = (n-p)\,S^2 = Y'Y - \hat{\beta}'X'Y$$

After deleting M observations we have

$$\begin{aligned}(n-p-M)\,S^2_{(M)} &= Y'_{(M)}Y_{(M)} - \hat{\beta}'_{(M)}X'_{(M)}Y_{(M)}\\ &= Y'Y - Y'_M Y_M - -\hat{\beta}'_{(M)}\left(X'Y - X'_M Y_M\right)\\ &= (n-p)\,S^2 + \hat{\beta}'X'Y - \hat{\beta}'_{(M)}X'Y - Y'_M Y_M + \hat{\beta}'_{(M)}X'_M Y_M\\ &\equiv (n-p)\,S^2 + A + B\end{aligned}$$

where $A = \left(\hat{\beta}' - \hat{\beta}'_{(M)}\right)X'Y$ and $B = -Y'_M Y_M + \hat{\beta}'_{(M)}X'_M Y_M$.

From (6.9) we have that since $\hat{Y}_M = X_M \hat{\beta}$,

$$\begin{aligned} A &= \left(X'X\right)^{-1} X'_M \left(I - H_M\right)^{-1} e_M X'Y \\ &= \hat{Y'}_M \left(I - H_M\right)^{-1} e_M \end{aligned}$$

Similarly,

$$\begin{aligned} B &= -Y'_M Y_M + \hat{\beta}'_{(M)} X'_M Y_M \\ &= -Y'_M Y_M + + \left[\hat{\beta} - e'_M \left(I - H_M\right)^{-1} X_M \left(X'X\right)^{-1}\right] X'_M Y_M \\ &= -Y'_M Y_M + \hat{\beta}' X'_M Y_M - e'_M \left(I - H_M\right)^{-1} H_M Y_M \\ &= -e'_M Y_M - Y'_M H_M \left(I - H_M\right)^{-1} e_M \end{aligned}$$

since it is all scalar. It follows

$$\begin{aligned} \left(n-p-M\right) S^2_{(M)} &= \left(n-p\right) S^2 + A + B \\ &= \left(n-p\right) S^2 + \hat{Y'}_M \left(I - H_M\right)^{-1} e_M + -e'_M Y_M - Y'_M H_M \left(I - H_M\right)^{-1} e_M \end{aligned}$$

Since

$$\left(I - H_M\right)^{-1} - H_M \left(I - H_M\right)^{-1} = I$$

we have

$$\hat{Y'}_M \left(I - H_M\right)^{-1} e_M - Y'_M H_M \left(I - H_M\right)^{-1} e_M = e'_M Y_M$$

and hence

$$\begin{aligned} \left(n-p-M\right) S^2_{(M)} &= \left(n-p\right) S^2 + e'_M \left(I - H_M\right)^{-1} \hat{Y}_M - e'_M \left(I - H_M\right)^{-1} Y_M \\ &= \left(n-p\right) S^2 - e'_M \left(I - H_M\right)^{-1} e_M \end{aligned}$$

This shows that the residual sum of squares when M observations are deleted is reduced by an amount that depends on e_M and on the inverse of a matrix containing elements of the hat matrix.

When $M = 1$,

$$\left(n-p-1\right) S^2_{(i)} = \left(n-p\right) S^2 - \frac{e_i^2}{1 - h_{ii}}$$

Finally, we may compare the value of the i^{th}observation Y_i with the prediction $\hat{Y}_{(i)}$ when that observation is not used in the fitting.

Let

$$\begin{aligned} d_i &= Y_i - \hat{Y}_{(i)} \\ &= Y_i - x_i' \hat{\beta}_{(i)} \\ &= \frac{e_i}{1-h_{ii}} \end{aligned}$$

.

We note the variance is equal to

$$\begin{aligned} \sigma^2 \left[d_i \right] &= \sigma^2 \left[1 + x_i' \left(X_{(i)}' X_{(i)} \right)^{-1} x_i \right] \\ &= \frac{\sigma^2}{1-h_{ii}} \end{aligned}$$

where σ^2 is estimated by $MSE_{(i)} \equiv S_{(i)}^2$ which is independent of Y_i.

The above expression follows from

$$\begin{aligned} x_i' \left(X_{(i)}' X_{(i)} \right)^{-1} x_i &= x_i' \left(X'X \right)^{-1} x_i + \frac{x_i' \left(X'X \right)^{-1} x_i x_i' \left(X'X \right)^{-1} x_i}{1-h_{ii}} \\ &= h_{ii} + \frac{h_{ii}^2}{1-h_{ii}} \end{aligned}$$

and

$$\left[1 + x_i' \left(X_{(i)}' X_{(i)} \right)^{-1} x_i \right] = \frac{1}{1-h_{ii}}$$

Consequently, the studentized residuals are

$$\begin{aligned} \frac{d_i}{S_{(i)} \sqrt{\left[1 + x_i' \left(X_{(i)}' X_{(i)} \right)^{-1} x_i \right]}} &= \frac{e_i}{S_{(i)}} \frac{\sqrt{1-h_{ii}}}{1-h_{ii}} \\ &= \frac{e_i}{S_{(i)}} \frac{1}{\sqrt{1-h_{ii}}} \sim t_{(n-p-1)} \end{aligned}$$

We may also compare

$$\begin{aligned} DFFITS &= \frac{\hat{Y}_i - \hat{Y}_{(i)}}{S_i\sqrt{h_{ii}}} \\ &= e_i\left[\frac{n-p-1}{SSE\left(1-h_{ii}\right)-e_i^2}\right]^{1/2}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \\ &= t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \end{aligned}$$

since

$$(n-p-1)\,S_{(i)}^2 = (n-p)\,S^2 - \frac{e_i^2}{1-h_{ii}}$$

We flag influential cases when

$$DFFITS\begin{cases} > 1 & smal/medium\ data \\ \\ > 2\sqrt{\frac{p}{n}} & large\ data \end{cases}$$

Example 6.1. For the GPA data example,

a) Obtain the residuals

b) Obtain a histogram plot of the residuals and a qq-plot to check for normality

c) Obtain confidence and prediction bands using ggplot2

d) Plot of residuals against predictor variable

e) Plot of absolute residuals against predictor variable

f) Plot of residuals against fitted value

g) Plot of residuals against time

h) Calculate test for the non constancy of variance

Example 6.2. Consider the crime data for a city where

X= % of individuals having at least a high school diploma

Y= # crimes reported per 100,000 residents last year

Example 6.3. a) Illustrate lack of fit test using bank data Table 3.4 where

X= size of minimum deposit

Y=# of new accounts

Branch	1	2	3	4	5	6	7	8	9	10	11
X	125	100	200	75	150	175	75	175	125	200	100
Y	160	112	124	28	152	156	42	124	150	104	136

b) Use the following to create diagnostic plots of residuals vs fitted (to check linearity), normal qq plots, scale-location vs fitted values and residuals vs leverage

6.2 Calculations Using R

Another package that plots diagnostic measures

```
install.packages("ggfortify")
library(ggfortify)
autoplot(fit)
```

6.3 R Session

Textbook 4.1

In R

the diagnostic plots are exhibited using the package

```
library(olsrr)
ols_plot_cooksd_bar(fit) # yields a plot of Cook's distance vs the observations
ols_plot_cooksd_bchrt(fit) # also yields a plot of Cook's distance vs the observations
ols_plot_dfbetas(fit) # yields a plot of DFBETAS vs the observations
ols_plot_dffits(fit) # yields a plot of DFFITS vs the observations
```

Also the following will display the Hat matrix diagonal elements;

under coefficients, a matrix whose i-th row contains the change in the estimated coefficients which results when the i-th case is dropped from the regression;

under sigma, a vector whose i-th element contains the estimate of the residual standard deviation obtained when the i-th case is dropped from the regression

under wt.res. a vector of weighted (or for class glm rather deviance) residuals

```
diag=lm.influence(model)
diag
```

To see values of Y X .fitted .resid .hat .sigma .cooksd .std.resid type in R

```
library(broom)
K=model.diag.metrics=augment(model)
head(model.diag.metrics, 20) or
model.diag.metrics(model) # shows values of fit,hat,sigma,Cook,std.resid influence.measures(
summary(influence.measures(model)) #this exhibits the potentially influential observations
```

Alternatively use

```
hatvalues(model)
dfbetas(model,~)
dffits(model)
cooks.distance(model)
```

6.4 Data Sets

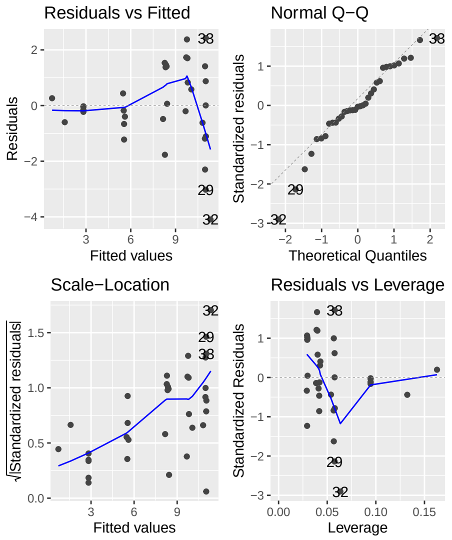
Weighted Data

- Crime Data
- Delivery Time Data
- Bank Data
- Housing Data
- Grades Data

6.5 R session

Using Weighted data

```
Weighted=read.table(file.choose(),header=TRUE,sep=’\t’)
names(Weighted) [1] "X" "Y" "W"
Diameter=Weighted$X
Area=Weighted$X
install.packages("ggfortify")
library(ggfortify)
fit=lm(Y~X, data=Weighted)
autoplot(fit)
```



Plots of residuals vs leverage exhibit residuals with non-linear patterns. In a good model, a plot of residuals vs fitted will show points randomly distributed. In a poor model, there will be some pattern.

A scale-location plot (or spread-location plot) will show if the residuals are spread equally along the range of predictors. It enbales us to check the assumption of equal variance. A plot is good if there is a horizontal line with eually randomly spread of points.

A plot of residuls vs leverage helps to locate influential cases if any. Some cases may be influential even if they appear to be in a reasonable range of the values. We watch for outlying values in the upper right or lower right plot. Look for values outside the dashed lines where Cook's distance scores are highest.

6.6 Suggested Problems

4.3, 4.21,

7 Different Models

The regression set up permits us to consider a variety of models which we will discuss here

7.1 Polynomial regression models

A k-order polynomial regression model in one variable takes the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + ... + \beta_k X^k + \epsilon$$

which may be fitted using the matrix approach. It is important to keep in mind that in considering such a model, the order k should be as low as possible. The inversion of the matrix $X'X$ will be inaccurate resulting in poor estimates of the parameters and their variances.

Often, orthogonal polynomials defined below, are used in the modeling because they simplify the fitting process

$$Y_i = \beta_0 P_0\left(X_i\right) + \beta_1 P_1\left(X_i\right) + \beta_2 P_2\left(X_i\right) + ... + \beta_k P_k\left(X_i\right) + \epsilon_i$$

where P_j is a j order orthogonal polynomial satisfying

$$\begin{aligned}\sum_{i=1}^{n} P_j\left(X_i\right) P_l\left(X_i\right) &= 0, j \neq l \\ P_0\left(X_i\right) &= 1\end{aligned}$$

Such polynomials have been tabulated (See Biometrika Tables for Statisticians). The least squares estimates are given by

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} P_j\left(X_i\right) Y_i}{\sum_{i=1}^{n} P_j^2\left(X_i\right)}, j = 0, 1, ..., k$$

The principal advantage of using orthogonal polynomials is that the model can be fitted sequentially. This specific advantage is less important today in the age of high speed computing compared to the times when much of the modeling was done using calculators.

Sometimes a low order polynomial does not fit the data well. This can be due to the fact that the function in question behaves differently in different parts of the range. In that case, it is common to use splines functions or piece wise polynomial fitting. We do not pursue this topic here. (Refer to the text book p.236-242)

When two or more variables are involved cross terms are included as in the following model involving two variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon$$

Such models are called surface response surfaces. Such models are often used in control theory problems to optimize the selection of control settings of the variables.

7.2 Indicator regression models

This subject is treated in Chapter 8 of Montgomery et al.

Regression analysis allows the use of indicator variables which are qualitative or categorical in nature. Such variables are labeled dummy variables. For example, to take into account gender we may define

$$X_2 = \begin{cases} 1 & male \\ \\ 0 & female \end{cases}$$

An interesting application is to the case where one wishes to fit a simple linear model as a function of gender. Set

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

In that case,

$$Y = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon & males \\ \\ \beta_0 + \beta_1 X_1 + \epsilon & female \end{cases}$$

So here the two lines are parallel. This can be generalized to 2 or more dummy variables

X_2	X_3	
0	0	observation from category 1
1	0	observation from category 2
0	1	observation from category 3

Example 7.1. a) Suppose that we have the following time data and we wish to fit two

lines when the abscissa is known to be in 1972.

	X_0	X_1	X_2	Y
1970	1	1	0	2.3
	1	2	0	3.8
1971	1	3	0	6.5
	1	4	0	7.4
1972	1	5	0	10.2
	1	5	1	10.5
1973	1	5	2	12.1
	1	5	3	13.2
1974	1	5	4	13.6

The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

b) Suppose that we have the same data as in a) but the abscissa is unknown. In that case we need an additional dummy variable

	X_0	X_1	X_2	X_3	Y
1970	1	1	0	0	2.3
	1	2	0	0	3.8
1971	1	3	0	0	6.5
	1	4	0	0	7.4
1972	1	5	0	1	10.2
	1	5	1	1	10.5
1973	1	5	2	1	12.1
	1	5	3	1	13.2
1974	1	5	4	1	13.6

The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

For other examples, see Example 8.5 p.280 of Montgomery et al.

We conclude by noting that analysis of variance models make extensive use of indicator variables.

7.3 R Session

Textbook 8.1

7.4 Suggested Problems

7.4; 7.6; 8.16

7.5 Data Sets

Hardwood Data

Voltage Drop Data

Windmill Data

Tool Life Data

Turkey data

8 Multicollinearity

When the variables are correlated among themselves multicollinearity is said to exist. This can cause serious problems among them being that the estimates become highly unstable. What are some of the symptoms of multicollinearity?

1. Large variation in the estimated coefficients when a new variable is either added or deleted.

2. Non significant results in individual tests on the coefficients of important variables.

3. large coefficients of simple correlation between pairs of variables.

4. Wide confidence interval for the regression coefficients of important variables.

The principal difficulty is that the matrix $(X'X)$ may not be invertible. As well, multicollinearity affects the interpretation of the coefficients in that they may vary in value. To illustrate, consider the case of two predictor variables X_1, X_2 . If the variables are standardized then the matrix

$$(X'X) = \frac{1}{1-r_{12}^2}\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

where r_{12} is the correlation between the two variables. Moreover the variance -covariance matrix is

$$\sigma^2 \left(X'X\right)^{-1} = \sigma^2 \frac{1}{1-r_{12}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

The two regression coefficients have the same variance and this increases as the correlation increases.

Consequently, as $|r_{12}| \to 1$, $Var\hat{\beta}_k \to \infty; Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) \to \pm\infty$ as $r_{12} \to \pm 1$

The estimates are

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

Hence, $\hat{\beta}_1 = \frac{r_{1Y} - r_{12}}{1-r_{12}^2}; \hat{\beta}_2 = \frac{r_{2Y} - r_{12}}{1-r_{12}^2}$

In general, the diagonal elements of $(X'X)^{-1}$are $C_{jj} = \frac{1}{1-R_j^2}$
where R_j^2 is the Rsquare value obtained from the regression
of X_j on the other $p-1$ variables.

If there is a strong multicollinearity between X_j and the other $p-1$ variables,
$R_j^2 \cong 1$ and $Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{1-R_j^2} \cong \infty$

As well, under multicollinearity, the values of the estimates will be large. Set

$$L = \left\|\hat{\beta} - \beta\right\|^2$$

Then,

$$\begin{aligned} E\sum_{j=1}^{p}\left(\hat{\beta}_j - \beta_j\right)^2 &= \sum_{j=1}^{p} Var\left(\hat{\beta}_j\right) \\ &= \sigma^2 Trace\left(X'X\right)^{-1} \\ &= \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \end{aligned}$$

where $\{\lambda_j\}$are the eigenvalues of $X'X$.

Under multicollinearity, some of these eigenvalues will be small

and hence their inverses will be large.

$$L = \hat{\beta}'\hat{\beta} - 2\hat{\beta}\beta + \beta'\beta$$

Taking the expectation, we see

$$\begin{aligned} E\left[L\right] &= E\left\|\hat{\beta}\right\|^2 - \left\|\beta\right\|^2 \\ &= \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \end{aligned}$$

Hence, $= E\left\|\hat{\beta}\right\|^2 = \left\|\beta\right\|^2 + \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$

The eigenvalues (λ_j) can also be used to measure the extent of multicollinearty in the system.

If one or more are small, then there are near linear dependencies in the columns of $X'X$.

The condition number κ and condition indices κ_j of $X'X$ are defined to be

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}, \kappa_j = \frac{\lambda_{max}}{\lambda_j}$$

$\kappa < 100$	no serious multicollinearity
$100 < \kappa < 1000$	moderate to strong
$1000 < \kappa$	severe

The starting point is to first standardize the variables as

$$\begin{aligned} Y_i^* &= \frac{1}{\sqrt{n-1}}\left(\frac{Y_i-\bar{Y}}{s_Y}\right) \\ X_{ik}^* &= \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik}-\bar{X}_k}{s_k}\right), k=1,...,p-1 \end{aligned}$$

When the standardized regression model with no intercept

$$Y_i^* = \sum_{k=1}^{p-1} \beta_i^* X_{ik}^* + \varepsilon_i^*$$

is fitted, we have the relationship

$$\begin{aligned} \beta_k &= \left(\frac{s_Y}{s_k}\right)\beta_k^* \\ \beta_0 &= \bar{Y} - \sum_{i=1}^{k-1} \beta_i \bar{X}_i \\ r_{XX} \quad b^* &= r_{YX} \end{aligned}$$

where r_{XX} is the correlation matrix

$$r_{XX} = \begin{pmatrix} 1 & r_{12} & ... & r_{1,p-1} \\ & 1 & ... & r_{2,p-1} \\ & & ... & ... \\ r_{1,p-1} & & & 1 \end{pmatrix}$$

and $r'_{YX} = (r_{Y1}...r_{rY,p-})$, $r_{ij} = cor\left(X_i, X_j\right), r_{Yi} = cor\left(Y, X_i\right).$

Mathematically, multicollinearity may be diagnosed using variance inflation factors. Specifically, suppose that the regression is fitted using the standardized predictor variables. Then,

$$\sigma^2\left[b^*\right] = \sigma^2 r_{XX}^{-1}$$

We define the variance inflation factor (VIF)

$$(VIF)_k = \left(1 - R_k^2\right)^{-1}$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p-2$ other X variables. Hence,

$$\sigma^2\left[b_k\right] = \sigma^2\left(1 - R_k^2\right)^{-1}$$

The $(VIF)_k = 1$ when $R_k^2 = 0$, i.e. whenever X_k is not linearly related to the other X

variables in the model. Under perfect correlation, i.e.$R_k^2 = 1$ the variance is unbounded. As a rule of thumb, a value $(VIF) > 10$ indicates that multicollinearity exists.

Table-3: VIF interpretation

VIF -value conclusion

VIF 1 Not correlated 1

VIF 5 Moderately correlated

VIF 5 Highly correlated

Tolerance is the amount of variability in one independent variable that is no explained by the other independent variables, and it is in fact 2 1 R .Tolerance values less than 0.10 indicate collinearity.

The eigenvalues of X'X are the squares of the singular values of X. The condition indices are the square roots of the ratio of the largest eigenvalue to each individual eigenvalue. The largest condition index is the condition number of the scaled X matrix.

Alternatively, as a diagnostic tool, we may compute the average

$$\overline{VIF} = \frac{\sum (VIF)_k}{p-1}$$

Mean values much greater than 1 point to serious multicollinearity.

Ridge regression is considered as a remedial measure to multicollinearity. The theory is as follows. The normal equation

$$(X'X)\, b = X'Y$$

is transformed by using the standardized variables so that it becomes

$$r_{XX} b = r_{YX}$$

We suppress the * symbol for b for ease of notation.

Consider solving instead the equation

$$(r_{XX} + cI)\, b^R = r_{YX}$$

where $c \geq 0$ is a constant and the superscript R indicates "ridge". The ridge standardized regression coefficients become

$$b^R = (r_{XX} + cI)^{-1}\, r_{YX} \tag{8.1}$$

The constant c reflects the fact that the ridge estimators will be biased but they tend to be more stable or less variable than the ordinary least squares estimators.

The constant c is usually chosen is such a way that the estimators of b_k^R are stable in value. Alternatively, whenever the $(VIF)_k$ are stable in value. A plot of the coefficients against c is called the ridge trace and this helps in the selection of c.

Finally, we note that ridge regression can also be obtained from the method of penalized regression. From (8.1), we have the following system of equations:

$$\begin{aligned} (1+c)b_1^R + r_{12}b_2^R + ..., + r_{1,p-1}b_{p-1}^R &= r_{Y1} \\ r_{21}b_1^R + (1+c)b_2^R + ..., + r_{2,p-1}b_{p-1}^R &= r_{Y2} \\ \\ r_{p-1,1}b_1^R + r_{p-1,2}b_2^R + ..., + (1+c)b_{p-1}^R &= r_{Y,p-1} \end{aligned} \tag{8.2}$$

On the hand, consider the penalized least squares

$$Q = \sum \left[Y_i - \beta_1 X_{i1} - ... - \beta_{p-1}X_{i,p-1}\right]^2 + c\sum_{j=1}^{p-1}\beta_j^2 \tag{8.3}$$

Differentiating (8.3) with respect to each of the parameters leads to (8.2).

The major drawback of ridge regression is that the ordinary inference procedures are no longer applicable. In that case we need to use bootstrapping methods to obtain the precision of the estimators.

8.1 Calculations Using R

We consider the Body fat Data which consists of

V1 Tricepts skinfold thickness

V2 Thigh circumference

V3 Midarm circumference

V4 Body fat

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

names(Bodyfat) [1] "Tricepts" "Thigh" "Midarm" "Fat"

V1=Bodyfat$Tricepts

V2=Bodyfat$Thigh

V3=Bodyfat$Midarm

V4=Bodyfat$Fat

a) **Collinearity**

diagnostics use the olsrr package

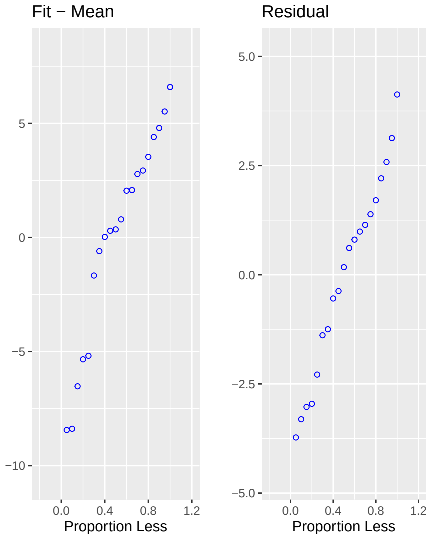
ols_coll_diag(model)

library(olsrr)

model=lm(V4~V1+V2+V3,data=Bodyfat)

ols_plot_resid_fit_spread(model)

We obtain plots showing model fit assessment. These plots are used to detect nonlinearity, influential observations and outliers. They consist of side-by-side quantile plots of the centered fit and the residuals. It shows how much variation in the data is explained by the fit and how much remains in the residuals. For inappropriate models, the spread of the residuals in such a plot is often greater than the spread of the centered fit.



Next we obtain correlations

ols_correlations(model)

Correlations - Variable Zero Order Partial Part -

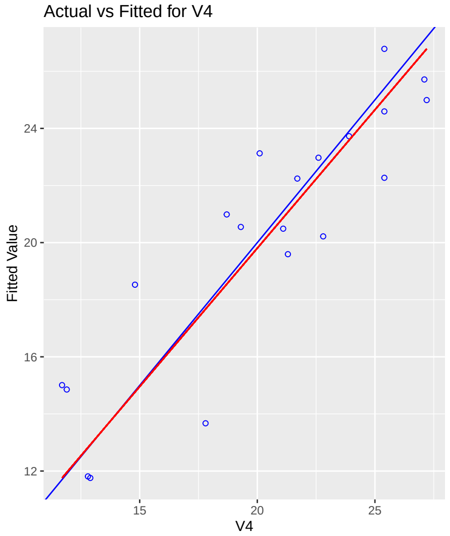
V1 0.843 0.338 0.160

V2 0.878 -0.267 -0.123

V3 0.142 -0.324 -0.153

Next we compare the observed vs predicted plot to assess the fit of the model. Ideally, all your points should be close to a regressed diagonal line. Draw such a diagonal line within your graph and check out where the points lie. If your model had a high R Square, all the points would be close to this diagonal line. The lower the R Square, the weaker the Goodness of fit of your model, the more foggy or dispersed your points are from this diagonal line.

ols_plot_obs_fit(model)



Next we diagnose

```
ols_coll_diag(model) #Tolerance and Variance Inflation Factor -
```

Variables	Tolerance	VIF
V1	0.001410750	708.8429
V2	0.001771971	564.3434
V3	0.009559681	104.6060

Tolerance is the % of variation in the predictor not explained by the other predictors.

To calculate it, regress the k^{th}predictor on the other predictors in the model. Compute $R^2_{k.}$ Then

Tolerance $=1 - R^2_{k.}$

b) Diagnostics Panel Panel of plots for regression diagnostics

```
ols_plot_diagnostics(model)
```

c) **Ridge regression**

Next we load the package MASS for ridge regression

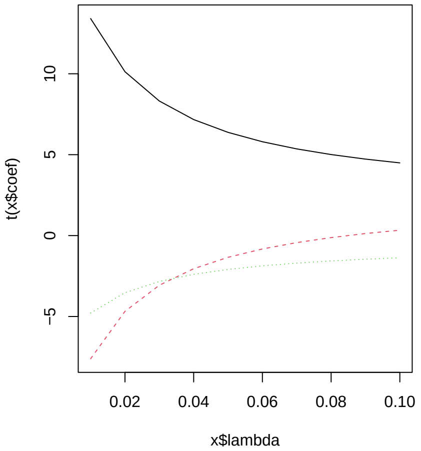
```
library(MASS)
y= c(0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09,0.10)
lm.ridge(V4~V1+V2+V3,Bodyfat,lambda=y)
```

```
y V1 V2 V3
0.01 64.389989 2.7395079 -1.49200145 -1.3458552
0.02 42.218200 2.0683914 -0.91772066 -0.9921824
0.03 30.007043 1.6986330 -0.60142746 -0.7972812
0.04 22.276858 1.4644460 -0.40119461 -0.6738065
0.05 16.944517 1.3028056 -0.26306759 -0.5885534
0.06 13.044863 1.1845111 -0.16204816 -0.5261373
0.07 10.069445 1.0941794 -0.08496703 -0.4784538
0.08 7.724954 1.0229365 -0.02422730 -0.4408273
```

```
0.09 5.830318 0.9653040 0.02486091 -0.4103718
0.10 4.267704 0.9177169 0.06534957 -0.3852088
plot(lm.ridge(V4~V1+V2+V3,Bodyfat,lambda=y))
```

Alternatively, instead of computing y, write

```
lm.ridge(V4~V1+V2+V3,Bodyfat,lambda=seq(0,0.1,0.01))
```



Here we see plots of the estimated coefficients with varying values of c

d) Condition Index

Most multivariate statistical approaches involve decomposing a correlation matrix into linear combinations of variables. The linear combinations are chosen so that the first combination has the largest possible variance (subject to some restrictions we won't discuss), the second combination has the next largest variance, subject to being uncorrelated with the first, the third has the largest possible variance, subject to being uncorrelated with the first and second, and so forth. The variance of each of these linear combinations is called an eigenvalue. Collinearity is spotted by finding 2 or more variables that have large proportions of variance (.50 or more) that correspond to large condition indices. A rule of thumb is to label as large those condition indices in the range of 30 or larger.

The R command is

```
model = lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_eigen_cindex(model)
```

e) Added Variable Plot

An added variable plot provides information about the marginal importance of a predictor variable Xk , given the other predictor variables already in the model. It shows the marginal importance of the variable in reducing the residual variability.

The added variable plot was introduced by Mosteller and Tukey (1977). It enables us to visualize the regression coefficient of a new variable being considered to be included in a model. The plot can be constructed for each predictor variable.

Let us assume we want to test the effect of adding/removing variable X from a model. Let the response variable of the model be Y

Steps to construct an added variable plot:

Regress Y on all variables other than X and store the residuals (Y residuals). Regress X on all the other variables included in the model (X residuals). Construct a scatter plot of Y residuals and X residuals. What do the Y and X residuals represent? The Y residuals represent the part of Y not explained by all the variables other than X. The X residuals represent the part of X not explained by other variables. The slope of the line fitted to the points in the added variable plot is equal to the regression coefficient when Y is regressed on all variables including X.

A strong linear relationship in the added variable plot indicates the increased importance of the contribution of X to the model already containing the other predictors.

```
model = lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_added_variable(model)
```

f) Residual Plus Component Plot

The residual plus component plot was introduced by Ezekeil (1924). It was called as Partial Residual Plot by Larsen and McCleary (1972). Hadi and Chatterjee (2012) called it the residual plus component plot.

Steps to construct the plot:

Regress Y on all variables including X and store the residuals (e). Multiply e with regression coefficient of X (eX). Construct scatter plot of eX and X The residual plus component plot indicates whether any non-linearity is present in the relationship between Y and X and can suggest possible transformations for linearizing the data.

```
model = lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
K=ols_plot_comp_plus_resid(model)
K
```

g) Now we consider fitting the ridge regression model.

```
install.packages("glmnet")
> library(glmnet)
> y=Bodyfat$Fat
> x=data.matrix(Bodyfat[,c('Tricepts','Thigh','Midarm')])
> model <- glmnet(x, y, alpha = 0) #fit ridge regression model
> cv_model=cv.glmnet(x,y,alpha=0) #perform k-fold cross-validation to find optimal lambda value
>plot(cv_model)
> best_lambda=cv_model$lambda.min
> best_lambda
[1] 0.4370159
> best_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)
> coef(best_model) #find coefficients of best model
> plot(model, xvar = "lambda") #produce Ridge trace plot
#use fitted best model to make predictions and obtain an R square
> y_predicted <- predict(model, s = best_lambda, newx = x)
> #find SST and SSE
> sst <- sum((y - mean(y))^2)
> sse <- sum((y_predicted - y)^2)
> #find R-Squared
> rsq <- 1 - sse/sst
> rsq
[1] 0.7789357
```

8.2 Data Sets

Body fat Data

Acetylene Data

9 Building the Regression Model

When several predictor variables are involved, the issue that comes up naturally is to determine how to select the variables as parsimoniously as possible and still produce a "good" model. If $p-1$ predictors are available, then there will be 2^p-1 possible models which can be constructed. An approach which may be misleading is one where all the predictors are initially included and then discarding the ones where the studentized coefficients are not significant. If multicollinearity exists, this approach can lead to error. The use of diagnostic procedures is important in the final selection of the model as outliers uncovered by a residual analysis can greatly influence the solution. Some criteria is essential for the ultimate selection of the model.

9.1 Criteria for model selection

When presented with a set of possible models it is important to develop some criteria for selection.

9.1.1 R^2

This criteria chooses the model with the largest value of explained variation. A plot of R^2 vs the number of variables in the model will appear as a parabola with the last entry "curving" up a bit. This is the one with all variables in. One may draw a horizontal line parallel to the x-axis. The point where it meets the parabola determines the best fitting model since it will be as good as when all the variables are included.

An adjusted R^2 takes into account the values of n, p

$$\begin{aligned} R_a^2 &= 1-\left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} \\ &= 1-\frac{MSE\,(p)}{SSTO/\,(n-1)} \end{aligned}$$

The criteria minimum $MSE\,(p)$ and maximum adjusted R^2 are equivalent.

9.1.2 Mallows C_p

To derive the Mallows criteria, Suppose that the true model has q predictor variables

$$Y = X_q\beta_q + \varepsilon$$

Suppose instead we fit a model using only p predictor variables. Let H_p be the hat matrix using only p variables. The bias for the i^{th} fitted value is

$$E\left(\hat{Y}_i\right) - \mu_i$$

where μ_i is the true mean. Consequently,

$$E\left(\hat{Y}_i - \mu_i\right)^2 = \left(E\left(\hat{Y}_i\right) - \mu_i\right)^2 + \sigma^2\left(\hat{Y}_i\right)$$

The total mean squared error for all the fitted values divided by σ^2 is

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_i\left(E\hat{Y}_i - \mu_i\right)^2 + \sum_i \sigma^2\left[\hat{Y}_i\right]\right\}$$

We may estimate σ^2 by the MSE when all the variables are included. The vector of residuals becomes

$$e_p = (I - H_p)\,Y$$

and the error sum of squares is

$$SSE_p = e_p'e_p$$

It follows that

$$\begin{aligned} bias = E\left(e_p\right) &= (I - H_p)\,EY \\ &= EY - E\hat{Y} \end{aligned}$$

since $E\left[H_pY\right] = H_pE\left[Y\right] = E\left[\hat{Y}\right]$.

When $p = q, bias = EY - E\hat{Y} = 0$. Now using the idempotency of $(I - H_p)$,

$$\begin{aligned} E[SSE_p] &= E\left[e_p' e_p\right] \\ &= E\left[Y'(I - H_p)Y\right] \\ &= E\left[Y'(I - H_p)(I - H_p)Y\right] \\ &= \sigma^2 Trace\,(I - H_p) + (bias)'(bias) \\ \\ &= \sigma^2(n-p) + \sum_i \left(E\hat{Y}_i - \mu_i\right)^2 \end{aligned}$$

The total mean squared error for all the fitted values divided by σ^2 is

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2} E\left[\sum_i \left(E\hat{Y}_i - \mu_i\right)^2 + \sum_i \sigma^2\left[\hat{Y}_i\right]\right] \\ &= \frac{1}{\sigma^2}\left[E[SSE_p] - \sigma^2(n-p) + \sum_i \sigma^2\left[\hat{Y}_i\right]\right] \\ &= \frac{1}{\sigma^2} E[SSE_p] - (n-p) + \frac{1}{\sigma^2}\sum_{i=1}^n \sigma^2\left[\hat{Y}_i\right] \\ &= \frac{1}{\sigma^2} E[SSE_p] - (n-p) + p \end{aligned}$$

The last term follows since $\hat{Y}_i = x_i'\hat{\beta}$ and

$$\begin{aligned} \sum_{i=1}^n \sigma^2\left[\hat{Y}_i\right] &= \sigma^2 \sum \left[x_i'(X'X)^{-1}x_i\right] \\ \\ &= \sigma^2 Trace\left[X'(X'X)^{-1}X\right] \\ \\ &= \sigma^2 TraceH = p\sigma^2 \end{aligned}$$

Hence,

$$\Gamma_p = \frac{1}{\sigma^2} E[SSE_p] - (n - 2p)$$

If we estimate σ^2 by MSE and $E[SSE_p]$ by $[SSE_p]$ then the Mallows criteria becomes

$$C_p = \frac{SSE_p}{MSE} - (n - 2p)$$

If the p-term model has negligible bias, then $E(SSE_p) \simeq (n-p)\sigma^2$ and $C_p \simeq p$.

Mallows proposed a graphical method for finding the best model. Plot C_p vs p. Models having little bias will be close to the line $C_p = p$. Models with substantial bias will be above the line. Sometimes a model may show some bias but contains fewer variables and as a result may be preferred.

9.1.3 Akaike information criterion

Akaike proposed a criteria based on minimizing the expected entropy of the model. which is essentially a penalized likelihood measure, In the case of ordinary least squares regression, it becomes

$$AIC_p = nln\left(SSE_p\right) - nln\, n + 2p$$

As more variables are included, AIC_p decreases and the issue becomes whether or not the decrease justifies the inclusion of more variables.

9.1.4 Schwartz's Bayesian criterion (SBC)

A Bayesian extension of the Akaike criterion was proposed by Schwartz

$$BIC_{Sch} = nln\left(SSE_p\right) - nln\, n + p\left(ln\, n\right)$$

This criterion places a greater penalty than the Akaike criterion and it is the one used by **R.**

9.1.5 Prediction sum of squares criterion(PRESS)

Sometimes regression equations are used to predict future values. A criteria that is used is to select the model which minimizes

$$PRESS_p = \sum_i \left[Y_i - \hat{Y}_{(i)}\right]^2$$

where $\hat{Y}_{(i)}$ is the fitted value when the i^{th} observation is deleted.

9.2 Model selections

We shall consider various methods for selection of a model

9.2.1 All possible models

This method is self explanatory. We consider all 2^{p-1}possible models.

In R

with the program olsr

K=ols_step_all_possible(model) #yields all possible subsets and computes R^2, R_a^2 and C_p

9.2.2 Forward ,Backward and Stepwise Regression

The evaluation of all possible regressions becomes computationally challenging when there are many predictor variables. Methods for evaluating a small number of subset regression models are the forward, backward and stepwise regression approaches. These are iterative procedures. We begin with the Forward Selection.

Forward

Step 1 Begin with no regressors in the model. Compute the standardize student t statistic for each variable and choose the one with the greatest absolute value to include in the model. This is also the variable that has the largest simple correlation with the response. A pre-selected critical F value, say F_{in} is chosen.

Step 2 With the variable in Step 1 in, choose the next variable using the same criteria as in step 1 after adjusting for the effect of the first variable selected. The criteria makes use of partial correlations which are computed between the residuals from Step1 and the residuals from the regressions of the other regressors on X_j that is residuals from $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ and residuals $\hat{X}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} X_1$ for $j = 2, ..., K$. If X_2 is selected, it implies that the largest partial F statistic is

$$F = \frac{SSR\left(X_2|X_1\right)}{MSE\left(X_1, X_2\right)}$$

If $F > F_{in}$, then X_2 is entered into the model.from Check to drop a variable already in the model if its t value is below a preset limit.

Step 3 Repeat the steps above until the largest partial F statistic no longer exceeds F_{in} or until all the variables are included.

Backward

Begin with all the regressors in the model. Compute the partial F statistic for each regressor as if it were the last one to enter the model. We compare the smallest partial F with the preselected F_{out}. If it is smaller, then that variable is removed from the model. The procedure is repeated until the smallest partial F statistic is not less than F_{out}. Backward elimination is often preferred to forward regression because it begins with all the variables in the model.

Stepwise

Stepwise regression combines the previous two approaches. It is a modification of forward regression in that it reassesses the each of the regressors already in the model to see if it has become redundant. Here we need to prespecify F_{in}, F_{out} . Usually, we choose $F_{in} > F_{out}$ so that it becomes more difficult to add a variable than to remove it.

9.2.3 LASSO and LAR regression

See original paper and see R output

Reference: Montgomery et al p.321.

9.3 Calculations Using R

We begin with the all possible model part

```
Cement=read.table("C:\\Users\\malvo\\OneDrive - University of Ottawa\\Documents\\CC
3375\\2023\\Hald Cement data.txt",header=TRUE,sep="\t")
library(olsrr)
model = lm(Y ~., data = cement)
ols_step_all_possible(model)
k = ols_step_all_possible(model)
k
plot(k)
```

Next, the best subset selection

```
ols_step_best_subset(model)
k = ols_step_best_subset(model)
plot(k)
```

Next, forward regression

```
ols_step_forward_p(model)
k = ols_step_forward_p(model)
plot(k)
ols_step_forward_p(model, details = TRUE)
```

Next, backward regression

```
ols_step_backward_p(model)
k = ols_step_backward_p(model)
plot(k)
ols_step_backward_p(model, details = TRUE)
```

Next, stepwise regression

```
ols_step_both_p(model)
```

Next, stepwise using AIC

```
ols_step_forward_aic(model)
```

```
k = ols_step_backward_aic(model)
k
plot(k)
```

Next, both

```
ols_step_both_aic(model) or
k = ols_step_both_aic(model)
plot(k)
```

9.4 R Session

9.5 Data Sets

Hald Cement Data. The data relate to an engineering application that was concerned with the effect of the composition of cement on heat evolved during hardening. The response variable Y is the heat evolved in a cement mix.

$X1$ percentage weight in clinkers of 3CaO.Al2O3
$X2$ percentage weight in clinkers of 3CaO.SiO2
$X3$ percentage weight in clinkers of 4CaO.Al2O3.Fe2O3
$X4$ percentage weight in clinkers of 2CaO.SiO2
Y heat evolved (calories/gram)

9.6 Suggested Problems

10.6,10.12,10.14

10 Logistic Regression

Sometimes the response variable is discrete. For example, we may wish to model gender or to estimate the likelihood that a person is wearing a life jacket. Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

$$Y_i = \begin{cases} 1 & \text{with probability} \pi_i \\ \\ 0 & \text{with probability} 1 - \pi_i \end{cases}$$

Then

$$E[Y_i] = \pi_i$$

The usual least squares fitting approach is problematic for the following reasons

1. the variance of $Y_i = \pi_i (1 - \pi_i)$ which is not constant

2. the error terms ϵ_i are not normally distributed

3. There is no guarantee that the fitted model will force the estimate $\hat{Y}_i$ to be in the interval $(0, 1)$.

Definition 10.1. The logistic distribution has density

$$f(x) = \frac{e^x}{(1+e^x)^2}, -\infty < x < \infty \tag{10.1}$$

and cumulative density function

$$F(t) = \frac{e^t}{(1+e^t)}$$

We can show

$$EX = 0, \sigma^2[X] = \frac{\pi}{3}$$

Suppose that a random variable$Y$ is binary with

$$Y_i = \begin{cases} 1 & \text{if } \beta_0^* + \beta_1^* X_i + \epsilon_i^* < k \\ \\ 0 & \text{if } \beta_0^* + \beta_1^* X_i + \epsilon_i^* > k \end{cases}$$

for some constant k where ϵ_i^* has a logistic distribution. Then

$$\begin{aligned} \pi_i = P\left(Y_i = 1\right) &= P\left(\beta_0^* + \beta_1^* X_i + \epsilon_i^* < k\right) \\ \\ &= F\left(k - \beta_0^* - \beta_1^* X_i\right) \\ \\ &= F\left(\beta_0 + \beta_1 X_i\right) \\ \\ &= \frac{exp\left(\beta_0 + \beta_1 X_i\right)}{\left(1 + exp\left(\beta_0 + \beta_1 X_i\right)\right)} \end{aligned}$$

where

$$\beta_0 = k - \beta_0^*, \beta_1 = -\beta_1^*$$

It is common practice to model the logarithm of the odds

$$\begin{aligned} log\left(\frac{\pi_i}{1-\pi_i}\right) &= log\left(\frac{P\left(Y_i = 1\right)}{1 - P\left(Y_i = 1\right)}\right) \qquad (10.2) \\ \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

The estimation of the parameters is based on maximizing the likelihood.

$$\begin{aligned} \Pi_i f\left(y_i\right) &= \Pi_i \pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i} \\ \\ &= \Pi_i\left[\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}\left(1-\pi_i\right)\right] \\ \\ &= \sum y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum log\left(1-\pi_i\right) \end{aligned}$$

$log\left(\Pi_i f\left(y_i\right)\right) = \sum y_i\left(\beta_0 + \beta_1 X_i\right) - \sum log\left(1 + exp\left(\beta_0 + \beta_1 X_i\right)\right)$

There is no closed form solution.

Instead, iterative methods are used to obtain a solution b_0, b_1 and then

$$\hat{\pi}_i = \frac{exp\left(b_0 + b_1 X_i\right)}{\left(1 + exp\left(b_0 + b_1 X_i\right)\right)}$$

To interpret the parameters in the logistic regression model, let us consider the fitted value at a specific value of X, say X$_0$. Then the difference between the log odds at $X_0 + 1$ and the log odds at X_0 is

$$logodds\left(X_0 + 1\right) - logodds\left(X_0\right) = \hat{\beta}_1$$

Taking the antilogarithms, we obtain the odds ratio

$$\hat{O}_R = e^{\hat{\beta}_1}$$

The odds ratio is the estimated increase in the probability of successes associated with a one unit change in the value of the predictor variable. For a change of d units, the odds ratio becomes

$$\hat{O}_R = e^{d\hat{\beta}_1}$$

10.1 Repeat Observations

Suppose that we have repeat observations at each of the levels of the x variables and set Y_i to be the number of 1's observed for the i^{th} observation. Let n_i be the number of trials at each observation. Then $Y_i \sim Binomial(n_i, \pi_i)$. In that case, estimation is done by maximizing

$$\begin{aligned} logL\left(\beta_0, \beta_1\right) &= log \quad \mathit{\Pi}_{i=1}^{n} \binom{n_i}{Y_i} \pi_i^{Y_i} \left(1 - \pi_i\right)^{n_i - Y_i} \\ &= \quad \sum_{i=1}^{n} \left\{ log \binom{n_i}{Y_i} + Y_i \left(log \pi_i\right) + \left(n_i - Y_i\right) log\left(1 - \pi_i\right) \right\} \end{aligned}$$

Example 10.1. Snoring and heart failure

Snoring	Score X	Heart	Disease		$\hat{\pi}$
		Yes	No	n	
Never	0	24	1355	1379	0.021
Sometimes	2	35	603	638	0.044
Almost nightly	4	21	192	213	0.093
Every night	5	30	224	254	0.132

$b_0 = -3.866, b_1 = 0.397$

$\hat{\pi}' = -3.866 + 0.397X$

$O_R = e^{b_1(X_2 - X_1)}$
Comparing $X_1 = 2, X_2 = 5$, we have $O_R = 3.2904$

10.2 Multiple Logistic models

Multiple logistic models can also be fitted. Specifically, we replace we have $\beta_0 + \beta_1 X$ by

$$X_i'\beta = \beta_0 + \beta_1 X_{i1} + ... + \beta_{i,p-1} X_{i,p-1}$$

$$E[Y] = \frac{\beta' X}{1 + \beta' X}$$

so that

$$log \frac{\pi}{1-\pi} = \beta' X$$

10.3 Inference on model parameters

The maximum likelihood estimators are for large sample sizes approximately normally distributed with variances and covariances that are functions of the second order partial derivatives of the likelihood function.

Let

$$G = \left(\frac{\partial^2 L\left(\beta\right)}{\partial \beta_i \partial \beta_j} \right) \equiv \left(g_{ij}\right)$$

labeled the Hessian where

$$logL\left(\beta\right) = \sum_{i=1}^{n} Y_i \left(X_i'\beta\right) - \sum_{i=1}^{n} log\left(1 + e^{X_i'\beta}\right)$$

It can be shown that

$$E\left[b\right] = \beta$$

The variance estimate is given by

$$Var\left(b\right) = \left(X'VX\right)^{-1}$$

where $V = diag\left(n\hat{\pi}_i\left(1 - \hat{\pi}_i\right)\right)$. Moreover, we have

$$\frac{b_k - \beta_k}{s\left[b_k\right]} \sim N\left(0,1\right), k = 0, ..., p-1$$

which is used for testing and constructing confidence intervals.

To test whether several coefficients are 0, we make use of the likelihood ratio test. We consider likelihood ratio tests whereby we compare the full model (FM) with the reduced model (RM). Let

$$G^2 = -2log\left[\frac{L\left(RM\right)}{L\left(FM\right)}\right]$$

If the reduced model is correct, G^2 follows asymptotically a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the full and reduced-models, $df_{RM} - df_{FM} = (n-q) - (n-p)$. We reject for large values,i.e. $G^2 > \chi^2_{p-q}$.

In the present situation, for the simple logistic model, the Full Model is the one that has been fitted whereas the Reduced Model is the one with constant probability of success

$$E\left(Y\right) = \pi = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

Under RM, the mle of π is Y/n . Hence,

$$lnL\left(RM\right) = \left[ylny + \left(n-y\right)ln\left(n-y\right) - nlnn\right]$$

Hence the likelihood ratio statistics for testing significance of regression is

$$\begin{aligned} L &= 2\left\{\sum Y_i ln\hat{\pi}_i + \sum\left(n_i - Y_i\right)\ln(1-\hat{\pi}_i)\right\} \\ &- 2\left[YlnY + \left(n-Y\right)ln\left(n-Y\right) - nlnn\right] \end{aligned}$$

where Y is the total number of successes observed and n is the total number of observations. We reject the null hypothesis that the regression is non significant if L is large.

10.4 Test for Goodness of Fit

Before a logistic regression model is accepted, it needs to be examined. This is analogous to the usual lack of fit testing regression problem. In that context, we required repeat observations as we do here. We would like to test

$$H_0 \quad : \quad E\left[Y\right] = \left(1+e^{-X'\beta}\right)^{-1}$$

$$H_1 : \; E\left[Y\right] \; \neq \left(1+e^{-X'\beta}\right)^{-1}$$

Here we will make use of a Pearson chi-square goodness of fit test. The expected number of successes is $n_i\hat{\pi}_i$ and the expected number of failures is $n_i\left(1-\hat{\pi}_i\right)$. The Pearson Chi square test rejects the null hypothesis whenever

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{(n_i - Y_i - n_i(1-\hat{\pi}_i))^2}{n_i(1-\hat{\pi}_i)} \right] > \chi^2_{\alpha,n-p}$$

If the fitted model is correct, the HL statistics follows a chi square with $g-1$ degrees of freedom. We reject the null hypothesis for large values of the statistic HL.

10.4.1 Deviance Goodness of Fit Test

Another test for the model is based on the likelihood ratio test whereby we consider the reduced and the full models. Here we compare the current model to a saturated model whereby each observation (or group when $n_i > 1$) has its own probability of success estimated by Y_i/n_i.

Under the reduced model

$$E[Y_i] = \left(1 + e^{-X_i'\beta}\right)^{-1}$$

whereas under the full model (also called the saturated model)

$$E[Y_i] = \pi_i$$

The deviance goodness of fit statistic , also called Deviance) , is given by

$$\begin{aligned} DEV(X_0, X_1, ..., X_{p-1}) &= -2\left[logL(RM) - log(FM)\right] \\ &= -2\sum_{i=1}^{n}\left[Y_i log\left(\frac{Y_i}{n_i\hat{\pi}_i}\right) + (n_i - Y_i)\, log\left(\frac{n_i - Y_i}{n_i(1-\hat{\pi}_i)}\right)\right] \end{aligned}$$

We reject for large values i,e $DEV > \chi^2_{n-p}$. The deviance in logistic regression plays an analogous role to the residual mean squares in ordinary regression.

R computes the null deviance (the deviance of the worst model without any predictor) and the residual deviance. The quantity

$$1 - \frac{Deviance}{Null\ Deviance}$$

is equal to 1 for a perfect fit and equal to 0 if the predictors do not add anything to the model.

10.4.2 Hosmer-Lenshow Goodness of Fit Test

When there are no replicated on the regressor variables, observations may be grouped before performing a Pearson chi-square test. Generally about $g = 10$ groups are used. Let O_j and $N_j - O_j$ be the observed number of successes and failures respectively in group j where N_j is the total number of observations in the group. The estimated probability of success $\hat{\pi}_j$ in the j^{th} group is the average estimated success probability. Then the Hosmer-Lemenshow test statistic is

$$\begin{aligned} HL &= \sum_{j=1}^{g} \frac{\left(O_j - N_j\hat{\pi}_j\right)^2}{N_j\hat{\pi}_j} + \sum_{j=1}^{g} \frac{\left(N_j - O_j - N_j\left(1-\hat{\pi}_j\right)\right)^2}{N_j\left(1-\hat{\pi}_j\right)} \\ &= \sum_{j=1}^{g} \frac{\left(O_j - N_j\hat{\pi}_j\right)^2}{N_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)} \end{aligned}$$

10.5 Diagnostic Measures for Logistic Regression

We shall consider the ungrouped case only. Residuals in that case can be used to diagnose the adequacy of the fitted model. The ordinary residuals are defined as

$$e_i = Y_i - \hat{\pi}_i$$

These do not have constant variance. The deviance residual is for $i = 1, ..., n$

$$d_i = \pm \left\{ 2\left[Y_i log\left(\frac{Y_i}{n_i\hat{\pi}_i}\right) + \left(n_i - Y_i\right) log\left(\frac{n_i - Y_i}{n_i\left(1-\hat{\pi}_i\right)}\right)\right]\right\}^{1/2}$$

where the sign of d_i is the same as the sign of e_i

Similarly we may compute the standardized Pearson residuals

$$r_{Pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)}}$$

which do not have unit variance or the studentized Pearson residuals

$$sr_{Pi} = \frac{r_{Pi}}{\sqrt{1-h_{ii}}}$$

where h_{ii} is the i^{th} diagonal element of the hat matrix

$$H = V^{1/2}X\left(X'VX\right)^{-1}XV^{1/2}$$

and V is the diagonal matrix with $V_{ii} = n_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)$.The studentized Pearson residuals are useful to check for outliers.

The deviation residual is

$$dev_i = sgn\left(Y_i - \hat{\pi}_i\right)\sqrt{-2\left[Y_i ln\frac{Y_i}{\hat{\pi}_i} + \left(1_i - Y_i\right) ln\frac{1-Y_i}{\left(1-\hat{\pi}_i\right)}\right]}$$

so that

$$DEV = \sum_{i=1}^{n}\left(dev_i\right)^2$$

For a good model, $E\left[Y_i\right] = \hat{\pi}_i$ and plots of r_{SPi}vs $\hat{\pi}_i$and r_{SPi}vs linear predictor $X_i'\beta$ should show a smooth horizontal Lowess line through 0. Plots of the deviance and the studentized Pearson residuals are useful to check for outliers. A normal probability plot of the deviance residuals can be used to check for the fit of the model and for outliers. A plot of the deviance vs the estimated probability of success can be used to determine where the model is poorly fitted, at high or low probabilities.

Similarly for a plot of dev_i vs linear predictor $X_i'\beta$.

10.5.1 Detection of Influential Observations

In order to flag influential cases, we consider deleting one observation at a time and measuring its effect on both the χ^2 and DEV statistics. Plots of these vs i will show spikes for influential observations.

Similarly for plots vs $\hat{\pi}_i$

10.5.2 Influence on the Fitted Linear Predictor

Cook's distance here measures the standardized change in the linear predictor $X'\beta$ when the i^{th} case is deleted.

Indexed plots of Cook's distance identify cases that have a large influence on the fitted predictor.

Indexed plots of the leverage values h_{ii} help to identify outliers in the X space.

In all cases, visual assessment are needed because here is no actual rule of thumb for flagging outlier cases.

10.6 Calculations Using R

We will use the data on programming experience. Twenty five individuals were selected. They hadvarying amounts of experience in programming. All were given the same task and the results are given as a binary variable.

Data for logistic analysis may come in one of two forms: either Bernoulli or binomial. In this example it is binary.

```
library(ggplot2)
names(file)[1]=”experience’
names(file)[2]=”success’
mlogit=glm(success~experience,data=file,family=”binomial”)
summary(mlogit)
confint(mlogit) #confidence intervals
exp(coef(mlogit)) #odds ratio
exp(cbind(OR=coef(mlogit),confint(mlogit))) #odds ratio and 95% confidence interval
newdata=with(file,data.frame(experience=10))
predict(mlogit,newdata=newdata,se=TRUE)
```

We now proceed with the analysis for the experience data

```
mydata=read.table(file.choose(),header=TRUE,sep='\t')
head(mydata)
```

	experience	success	Fittedvalue	
1	14	0	0.310262	
2	29	0	0.835263	
3	6	0	0.109996	
4	25	1	0.726602	
5	18	1	0.461837	
6	4	0	0.082130	

```
summary(mydata)
```

	experience	success	Fittedvalue	
Min	4.00	0.00	0.08213	
1st Qu	9.00	0.00	0.16710	
Median	18.00	0.00	0.46184	
Mean	16.88	:0.44	0.44000	
3rd Qu	24.00	1.00	0.69338	
Max	32.00	1.00	0.89166	
standard deviations	9.0752410	0.5066228	0.2874901	

```
sapply(mydata,sd) *
mlogit=glm(success~experience, data=mydata,family="binomial")
summary(mlogit)
Call: glm(formula = success ~ experience, family = "binomial", data = mydata)
Deviance Residuals:
Min 1Q Median 3Q Max
-1.8992 -0.7509 -0.4140 0.7992 1.9624
```

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970 1.25935 -2.430 0.0151 *
experience 0.16149 0.06498 2.485 0.0129 *
--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 34.296 on 24 degrees of freedom
Residual deviance: 25.425 on 23 degrees of freedom
AIC: 29.425
Number of Fisher Scoring iterations: 4
confint(mlogit)
Waiting for profiling to be done
2.5 % 97.5 %
(Intercept) -6.03725238 -0.9160349
experience 0.05002505 0.3140397
exp(coef(mlogit)) (Intercept) experience
0.04690196 1.17525591
exp(cbind(OR=coef(mlogit),confint(mlogit)))
OR 2.5 % 97.5 %
(Intercept) 0.04690196 0.002388112 0.4001024
experience 1.17525591 1.051297434 1.3689441
newdata=with(mydata,data.frame(experience=10))
predict(mlogit,newdata=newdata,se=TRUE)
$fit 1 -1.444837
$se.fit [1] 0.7072129
$residual.scale [1] 1
```

Another package

```
# Installing the package install.packages("dplyr")
# Loading package library(dplyr)
# For Logistic regression install.packages("caTools")
# For ROC curve to evaluate model install.packages("ROCR")
# Loading package library(caTools)
#library(ROCR)
```

Sometimes the data can be split into a training set and a testing set

```
# Splitting dataset
split = sample.split(mtcars, SplitRatio = 0.8)
split train_reg = subset(mtcars, split == "TRUE")
test_reg = subset(mtcars, split == "FALSE")
# Training
model logistic_model = glm(vs ~ wt + disp, data = train_reg, family = "binomial")
```

```
logistic_model
# Summary
summary(logistic_model)
```

When the data is aggregated,

```
p=Y/n
mlogit=glm(p~X,data=Toxicity,weights=n,family="binomial")
mlogit
predict_reg =predict(mlogit, type = "response")
predict_reg
```

Stepwise logistic regression can also be done when several variables are involved

```
install.packages(“MASS”)
library(MASS)
stepAIC(model,trace=FALSE)
```

10.7 Data Sets

Program experience

Pneumocomiosis Data

10.8 R Session

Problem 13.1 in Montgomery et al.

11 Poisson Regression

In Poisson regression, we have counting data Y which follows a Poisson distribution with mean μ and hence variance μ.

$$f\left(y\right)=\frac{e^{-\mu}\mu^{y}}{y!},y=0,...$$

The model is then given as

$$Y_i=\mu_i+\varepsilon_i,i=1,...,n$$

We assume that there is a link function $g\left(\mu_i\right)$ that specifies the mean which may be one of the following

$$\begin{aligned} g\left(\mu_i\right) &= \mu_i = & X_i'\beta, identity\ link \\ g\left(\mu_i\right) &= & log\left(\mu_i\right)=X_i'\beta, log\ link \end{aligned}$$

The estimation of the parameters β is obtained using the method of maximum likelihood. As for the logistic case, there is no closed form for the solution.

The log likelihood is given by

$$logL\left(y,\beta\right)=\sum_{i=1}^{n}y_i log\mu_i-\sum_{i=1}^{n}\mu_i-\sum_{i=1}^{n}logy_i!$$

The fitted Poisson model is then

$$\begin{aligned} \hat{Y_i} &= g^{-1}\left(X_i'\hat{\beta}\right) \\ &= \begin{cases} X_i'\hat{\beta} & identity\ link \\ exp\left(X_i'\hat{\beta}\right) & log\ link \end{cases} \end{aligned}$$

Inference on the Poisson model is conducted as in the case of the logistic model.

Both the logistic and the Poisson models are particular examples of a more general linear model (GLM).

The response is assumed o have a distribution which is a member of the exponential family of distributions.

$$f\left(y_i,\theta_i,\phi\right) = exp\left\{\frac{\left[y_i\theta - b\left(\theta_i\right)\right]}{a\left(\phi\right)} + h\left(y_i,\phi\right)\right\}$$

Here, $\mu = E\left(Y\right) = \frac{db(\theta_i)}{d\theta_i}$ $Var\left(Y\right) = \frac{d^2b(\theta_i)}{d\theta_i^2}a\left(\phi\right)$

The basic idea is to develop a linear model for a function of the mean.

Set $\eta_i = g\left(\mu_i\right) = X_i'\beta$. The function g is called the link function.

For the logistic,

$$\begin{aligned} f\left(y_i,\theta_i,\phi\right) &= \binom{n_i}{y_i}p^{y_i}\left(1-p\right)^{n-y_i} \\ &= exp\left\{\left[y_ilog\left(\frac{p}{1-p}\right) + nlog\left(1-p\right)\right] + log\binom{n_i}{y_i}\right\} \end{aligned}$$

For the Poisson,

$$\begin{aligned} f\left(y_i,\theta_i,\phi\right) &= \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\ &= exp\left\{\left[y_ilog\lambda - \lambda\right] - logy_i!\right\} \end{aligned}$$

11.1 R Session

11.2 Data Sets

Aircraft Damage Data: This data refers to 30 strike missions involving two types of aircraft during the Vietnam war. $X1$ is an indicator variable for the type of aircraft used. $X2$ and $X3$ are bomb loads in tons and total months of aircrew experience respectively. The response variable Y is the number of locations where damage was inflicted.

```
Aircraft=read.table(file.choose(),header=TRUE,sep='\t')
names(Aircraft) [1] "Locationnumber" "Indicator" "load" "experience"
V1=Aircraft$Locationnumber
V2=Aircraft$Indicator
V3=Aircraft$load
V4=Aircraft$experience
summary(Aircraft)
```

	Locationnumber	Indicator	load	experience
Min	0.000	0.0	4.0	50.00
1st Qu	0.250	0.0	6.0	66.45
Median	1.000	0.5	7.5	80.25
Mean	1.533	0.5	8.1	80.77
3rd Qu	2.000	1.0	10.0	94.50
Max	7.000	1.0	14.0	120.0

```
mlogit=glm(V1~V2+V3+V4, data=Aircraft,family=poisson(link="log"))
summary(mlogit)
Call: glm(formula = V1 ~ V2 + V3 + V4, family = poisson(link = "log"), data = Aircraft)
Deviance Residuals:
Min 1Q Median 3Q Max
-1.6418 -1.0064 -0.0180 0.5581 1.9094
Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.406023 0.877489 -0.463 0.6436
V2 0.568772 0.504372 1.128 0.2595
V3 0.165425 0.067541 2.449 0.0143 *
V4 -0.013522 0.008281 -1.633 0.1025
--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 53.883 on 29 degrees of freedom
Residual deviance: 25.953 on 26 degrees of freedom
AIC: 87.649
Number of Fisher Scoring iterations: 5
```

Does the model has over-dispersion or under-dispersion?

If the Residual Deviance is greater than the degrees of freedom, then over-dispersion exists. This means that the estimates are correct, but the standard errors (standard deviation) are wrong and unaccounted for by the model.

The Null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean) whereas residual with the inclusion of independent variables.

So, to have a more correct standard error we can use a quasi-poisson model:

```
mlogit=glm(V1~V2+V3+V4, data=Aircraft,family=quasipoisson(link="log"))
summary(mlogit)
Call: glm(formula = V1 ~ V2 + V3 + V4, family = quasipoisson(link = "log"), data = Aircraft)
Deviance Residuals:
Min 1Q Median 3Q Max
```

```
-1.6418 -1.0064 -0.0180 0.5581 1.9094
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.406023 0.841982 -0.482 0.6337
V2 0.568772 0.483963 1.175 0.2505
V3 0.165425 0.064808 2.553 0.0169 *
V4 -0.013522 0.007946 -1.702 0.1007
--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for quasipoisson family taken to be 0.9207088)
Null deviance: 53.883 on 29 degrees of freedom
Residual deviance: 25.953 on 26 degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 5
```

For estimation and prediction we can use the same commands as for logistic regression.

12 References

The following references were used in the preparation of these notes.

[1] THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA BY F. J. ANSCOMBE. Biometrika, Volume 35, Issue 3-4, December 1948, Pages 246–254, https://doi-org.proxy.bib.uottawa.ca/10.1093/biomet/35.3-4.246

[2] Linear Models with R, second edition, 2014. Julian J. Faraway.

[3] Applied Regression Analysis and Other Multivariable Methods, fifth edition, 2014. David Kleinbaum, Larry Kupper, Azhar Nizam, Eli S. Rosenberg.

[4] Yu Guan Variance stabilizing transformations of Poisson, binomial and negative binomial distributions Statistics and Probability Letters 79 (2009) 1621–1629

[5] Applied Linear Statistical Models, fifth edition, 2005. Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li

[6] Introduction to Linear Regression Analysis, sixth edition, 2021. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining.