Linear Regression using Matrix Notation

jdt

10/23/2021

Contents

Theory

In this document, I want to present the least squares problem for the linear regression model using matrix notation. The data in this example are randomly generated (hence you can modify the companion R markdown file and create your own example).

The R code uses matrix notation as described in the theory part of this document. I HAVE NOT included any SAS code or output as this part of the presentation has not changed from the recent examples that you have seen.

There is considerable background mathematical results that you probably have not seen before. These are included in the Appendix. I elected to place that material here rather than in the course notes, else you would have balked or dropped the course the first week! I will spend some time on this material in the upcoming weeks. Don't panic!

Matrix Approach to Linear Regression

In this chapter the results for the linear least squares problem are reproduced using matrix notation. The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

for $i = 1, 2, \ldots, n$ where ϵ_i represents the unobserved error (or distance) that the observed data value y_i is from its mean $\mu_{y|x_i} = \beta_0 + \beta_1 x_i$ when $x = x_i$. This model can be written as

$$\mathbf{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon} \tag{2}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \mathbf{j_n} & \mathbf{x} \end{pmatrix} \qquad \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

where $\mathbf{j_n}' = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$ and $\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$. I will use y to denote the vector rather than $\mathbf{y}$ when the context is clear.

Least Squares

The least squares problem becomes finding $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \ \hat{\beta}_1 \end{pmatrix}'$ that minimizes

$$\begin{aligned} Q(\beta) &= \epsilon'\epsilon \\ &= (y - X\beta)'(y - X\beta) \\ &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - 2\beta'(X'y) + \beta'(X'X)\beta \end{aligned}$$

since $Q(\beta)$ is a scalar, one can use the properties of matrix differentiation to find

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0.$$

From which one obtains the normal equations given by

$$X'X\beta = X'y. \tag{3}$$

If $rank[X] = 2$, then the normal equation has a unique solution given by

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{4}$$

Which can be written as

$$\begin{aligned} \hat{\beta} &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\ &= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\ &= \frac{1}{\sum (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}. \end{aligned}$$

By letting $L = (X'X)^{-1}X'$, it follows that the least squares estimate for β is a linear function of y, given by $\hat{\beta} = Ly$. From which we have that the predicted value for y can be written as $\hat{y} = X\hat{\beta} = XLy = X(X'X)^{-1}X'y = Hy$, where

$$H = X(X'X)^{-1}X', \tag{5}$$

is a $n \times n$ matrix called the *Hat Matrix*. The computed or estimated residuals are given by $\hat{e} = y - \hat{y} = y - Hy = (I - H)y$ and the residual sum of squares (SS_E) can be written as

$$Q(\hat{\beta}) = \hat{e}'\hat{e} = y'(I - H)'(I - H)y = y'(I - H)y = y'y - \hat{\beta}'X'y.$$

Inference

As in the previous chapter, assumptions are needed in order for the mathematical least squares problem to become a statistical problem. In this chapter assumptions are given on the error structure (rather than on the dependent variable y as in the previous chapter). That is, assume that the unobserved errors given by $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are i.i.d. normal variables with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. This assumption is written in matrix notation as,

$$\vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim N_n(0, \sigma^2 I_n).$$

Since $\mathbf{y}$ is an affine transformation of normal variables ($X\beta$ is an unknown constant), it follows that

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \sim N_n(X\beta, \sigma^2 I_n).$$

Likewise, since $\hat{\beta} = \mathbf{L}\mathbf{y}$, $\hat{\mathbf{y}} = H\mathbf{y}$, and $\hat{\epsilon} = (I - H)\mathbf{y}$ we have[1]

1. $\hat{\beta} \sim N_2(\beta, \sigma^2 (X'X)^{-1})$.
   (a) $\hat{\beta}$ is an unbiased estimate of β.
   (b) $var(\hat{\beta}_i) = \sigma^2((X'X)^{-1})_{ii}$.
   (c) $cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2((X'X)^{-1})_{ij}$.
   (d) $corr(\hat{\beta}_i, \hat{\beta}_j) = ((X'X)^{-1})_{ij}/[((X'X)^{-1})_{ii}((X'X)^{-1})_{jj}]^{1/2}$.
2. $\hat{y} \sim N_n(X\beta, \sigma^2 H)$.
   (a) $var(\hat{y}_i) = \sigma^2 h_{ii}$.
   (b) $cov(\hat{y}_i, \hat{y}_j) = \sigma^2 h_{ij}$, where $H = (h_{ij})$. Notice that the $\hat{y}_i's$ are not independent of one another unless $h_{ij} = 0$.
   (c) $corr(\hat{y}_i, \hat{y}_j) = h_{ij}/[h_{ii}h_{jj}]^{1/2}$.
3. $\hat{\epsilon} \sim N_n(0, \sigma^2(I - H))$.
   (a) $var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$.
   (b) $cov(\hat{\epsilon}_i, \hat{\epsilon}_j) = -\sigma^2 h_{ij}$.
   (c) $corr(\hat{\epsilon}_i, \hat{\epsilon}_j) = -h_{ij}/[(1 - h_{ii})(1 - h_{jj})]^{1/2}$.

Since the sum of squares terms as given in the ANOVA table are quadratic forms, one can derive their properties from the results given for quadratic forms.

[1] see the results given in the first chapter on linear transformations of normal variates.

Expected Values of the Sum of Squares

From the appendix, we know that the expected value of quadratic form given by $Q = y'Ay$ is $E(Q) = tr[AV] + \mu' A\mu$ when $\mathbf{y} = (y_1, y_2, \ldots, y_n)' \sim N_n(\mu, V)$. If $V = \sigma^2 I_n$, one has

- $E(y'Hy) = \sigma^2 tr[H] + \beta' X' H X\beta = 2\sigma^2 + \beta' X' X\beta$.
- $E(y'(I-H)y) = \sigma^2 tr[(I-H)] + \beta' X'(I-H)X\beta = (n-2)\sigma^2$.

The above expression can be used as a moment estimator (least squares methods) as

$$\hat{\sigma}^2 = y'(I-H)y/(n-2) = SS_E/(n-2) = MSE.$$

Distribution of the Sum of Squares

The following results follow from the properties of the distribution of quadratic forms (You should be able to derive the properties given in red). Note, it is necessary to divide the Sum of Square terms by σ^2 else the results do not follow as given in the appendix.

- $SS(\beta)/\sigma^2 \sim \chi^2(df = 2, \lambda = 1/2\beta' X' X\beta)$.
- $SS_E/\sigma^2 \sim \chi^2(df = (n-2))$.
- $SS(\beta)$ and SS_E are independent.
- $F = MS(\beta)/MS_E \sim F(df_1 = 2, df_2 = (n-2), \lambda = 1/2\beta' X' X\beta)$.
- When $\beta = 0$ it follows that $SS(\beta)/\sigma^2 \sim \chi^2(df = 2)$ and $F = MS(\beta)/MS_E \sim F(df_1 = 2, df_2 = (n-2))$.

Since there is little reason to believe that $\beta = 0$ when $\bar{y} \neq 0$, we make the following adjustments (applying the correction factor to the total sum of squares and using the corrected or adjusted total sum of squares).

- $SS(\beta_1 \mid \beta_0)/\sigma^2 \sim \chi^2(df = 1, \lambda = 1/2\beta_1^2 \mathbf{x}'\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$
- $SS(\beta_1 \mid \beta_0)$ and SS_E are independent.
- $F = MS(\beta_1 \mid \beta_0)/MS_E \sim F(df_1 = 1, df_2 = (n-2), \lambda = 1/2\beta_1^2 \mathbf{x}'\mathbf{x})$.
- When $\beta_1 = 0$ it follows that $SS(\beta_1 \mid \beta_0)/\sigma^2 \sim \chi^2(df = 1)$ and $F = MS(\beta_1 \mid \beta_0)/MS_E \sim F(df_1 = 1, df_2 = (n-2))$.

ANOVA Table

As in the previous chapter an analysis of variance table can be written as[2]

Source	Sum of Squares	Degrees of Freedom	Mean Square
due to β	$SS(\beta) = \hat{\beta}' X' y = y' H y$	2	$MS(\beta) = SS(\beta)/2$
Residual	$SS_E = y'y - \hat{\beta}' X' y = y'(I-H)y$	n-2	$MS_E = SS_E/(n-2)$
Total	$y'y$	n	

[2] you will never see this table as a part of standard software output.

Since β_1 is the only parameter that is needed in determining if the independent variable x explains significant variation for the dependent variable y, the sum of squares term is adjusted for β_0, that is

$$y'Hy = y'(H - \frac{1}{n}\mathbf{jj}')y + \frac{1}{n}y'\mathbf{jj}'y$$

or

$$SS(\beta) = SS(\beta_0, \beta_1) = SS(\beta_1 \mid \beta_0) + SS(\beta_0)$$

where $\frac{1}{n}y'\mathbf{jj}'y = n\bar{y}^2$ is the correction factor or the sum of squares due to β_0. In which case the ANOVA table become

Source	Sum of Squares	Degrees of Freedom	Mean Square
due to $\beta_1 \mid \beta_0$	$SS(\beta_1 \mid \beta_0) = y'(H - \frac{1}{n}\mathbf{jj}')y$	1	$MS(\beta_1 \mid \beta_0) = SS(\beta_1 \mid \beta_0)/1$
Residual	$SS_E = y'y - \hat{\beta}'X'y = y'(I - H)y$	n-2	$MS_E = SS_E/(n-2)$
Corrected Total	$y'y - \frac{1}{n}y'\mathbf{jj}'y$	n-1	
due to β_0	$SS(\beta_0) = \frac{1}{n}y'\mathbf{jj}'y$	1	
Total	$y'y$	n	

R

Generate the Data

```
#set parameters
set.seed(5431)                #remove if you want a new data set
beta_0 = 3
beta_1 = 1.4
sigma = 4
n = 15
x = runif(n,5,20)
y = beta_0 + beta_1*x + sigma*rnorm(n)
ds = data.frame(x,y)
#previous example data
#x=c(15.6,26.8,37.8,36.4,35.5,18.6,15.3,7.9,0)
#y=c(5.2,6.1,8.7,8.5,8.8,4.9,4.5,2.5,1.1)
```

R code Using the Matrix approach

```
x=ds$x
y=ds$y
n<-length(x)
X<-cbind(array(1,c(n,1)),x)
X
```

This is the design matrix X in the model $y = X\beta + \epsilon$

```
##                 x
##  [1,] 1  8.555576
##  [2,] 1 10.869475
##  [3,] 1 10.083765
##  [4,] 1 12.273666
##  [5,] 1 10.515211
##  [6,] 1  9.943090
##  [7,] 1 19.402795
##  [8,] 1 14.788449
##  [9,] 1  7.992549
## [10,] 1  9.410424
## [11,] 1 19.320474
## [12,] 1  8.541136
## [13,] 1 14.854014
## [14,] 1 18.211345
## [15,] 1 15.416878
```

```
XTXinv<-solve(t(X)%*%X)
XTXinv
```

This is $(X'X)^{-1}$

```
##                           x
##    0.77883648 -0.056171059
## x -0.05617106  0.004430387
```

```
B<-XTXinv%*%t(X)%*%y          #beta_hat
B
```

This is $\hat{\beta} = (X'X)^{-1}X'y$

```
##       [,1]
##   1.443340
## x 1.491419
```

```
yhat<-X%*%B                   #y_hat predicted values for y
yhat
```

These are the predicted values (on the line) given by $\hat{y} = X\hat{\beta}$

```
##           [,1]
##  [1,] 14.20329
##  [2,] 17.65428
##  [3,] 16.48246
##  [4,] 19.74852
##  [5,] 17.12593
##  [6,] 16.27266
##  [7,] 30.38104
##  [8,] 23.49912
##  [9,] 13.36358
## [10,] 15.47823
## [11,] 30.25827
## [12,] 14.18176
## [13,] 23.59690
## [14,] 28.60409
## [15,] 24.43637
```

```
resid<-y-yhat                 #resid = y - y_hat
resid
```

These are the residuals given by $\hat{\epsilon} = y - \hat{y}$

```
##              [,1]
##  [1,]  0.03661731
##  [2,]  4.33164552
##  [3,]  0.74739412
##  [4,] -0.69216252
##  [5,] -1.49850925
##  [6,] -3.53377898
##  [7,]  2.51868401
##  [8,]  2.04076459
##  [9,]  3.80686416
```

```
## [10,] -1.39098297
## [11,] -1.69553484
## [12,] -3.00954539
## [13,] -0.28165907
## [14,] -2.86638136
## [15,]  1.48658467
```

```
sse<-sum(resid^2)              #SSE
sse
```

This is the residual sum of squares given by $SS_E = y'(I - H)y$

```
## [1] 83.90845
```

```
n<-dim(X)[1]
p<-dim(X)[2]
mse<-sse/(n-p)                 #MSE
mse
```

```
## [1] 6.454496
```

```
root.mse = sqrt(mse)           #RMSE
root.mse
```

```
## [1] 2.54057
```

```
cssy<-sum((y-mean(y))^2)
rsquare<-(cssy-sse)/cssy       #R-square
rsquare
```

```
## [1] 0.8568044
```

```
cov.b<-XTXinv*mse              #cov(beta) is a 2x2 matrix
cov.b
```

This is $cov(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$ where $\hat{\sigma}^2$ is the MSE.

```
##                       x
##    5.0269972 -0.36255589
## x -0.3625559  0.02859592
```

```
stdb = sqrt(diag(cov.b))       #se for Beta_hat
stdb
```

These are the standard errors for $\hat{\beta}$

```
##                  x
## 2.2420966 0.1691033
```

```
t<-B/stdb                      #t statistic for beta_hat
t
```

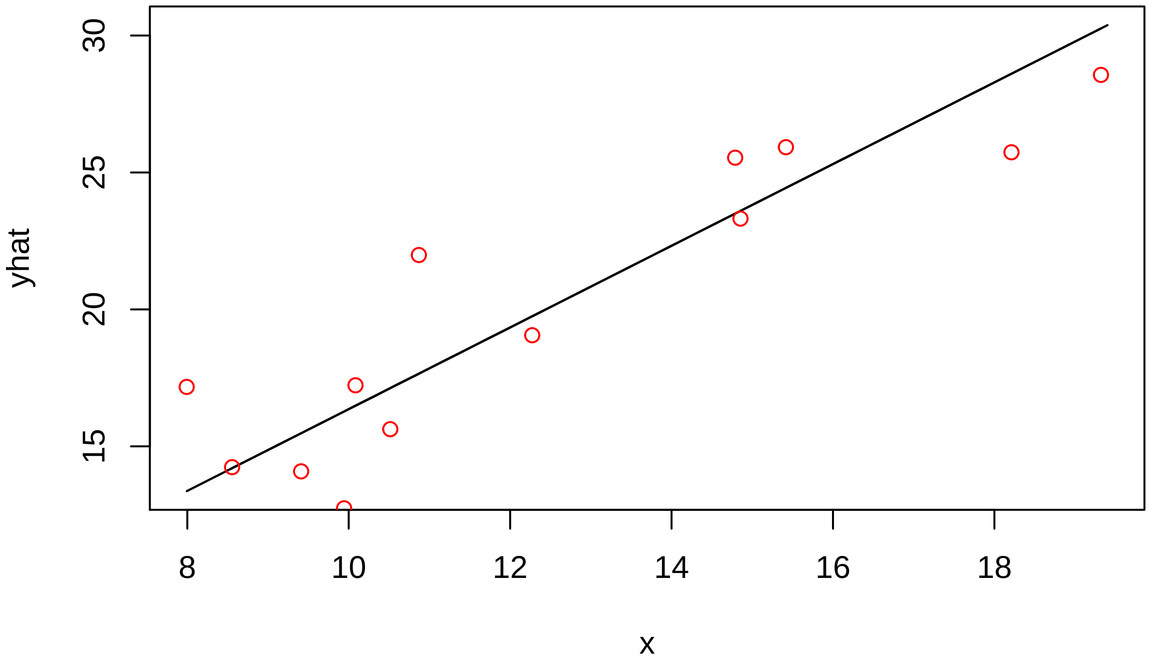
```
##          [,1]
##   0.6437458
```

```
## x 8.8195765
```

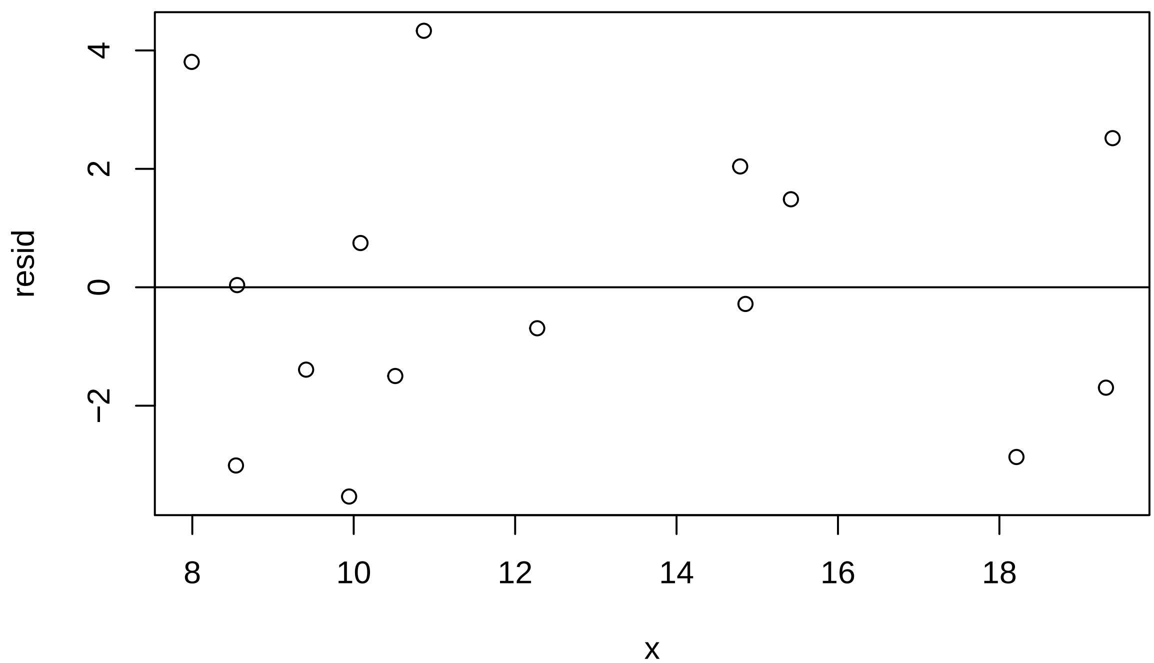
```r
p.value<-2*(1 - pt(abs(t),n-p))
p.value                              #p-value for H0:beta = 0
```

```
##           [,1]
##   5.309337e-01
## x 7.568304e-07
```

```r
#par(mfrow=c(2,1))
plot(yhat~x,type="l")
points(y~x,col="red")
```




```r
plot(x,resid)
abline(h=0)
```



#Regular R Approach to Regression

```
result<-lm(y~x)
summary(result)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5338 -1.5970 -0.2817  1.7637  4.3316
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.4433     2.2421   0.644    0.531
## x             1.4914     0.1691   8.820 7.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.541 on 13 degrees of freedom
## Multiple R-squared:  0.8568, Adjusted R-squared:  0.8458
## F-statistic: 77.78 on 1 and 13 DF,  p-value: 7.568e-07
```

```
aov(result)
```

```
## Call:
```

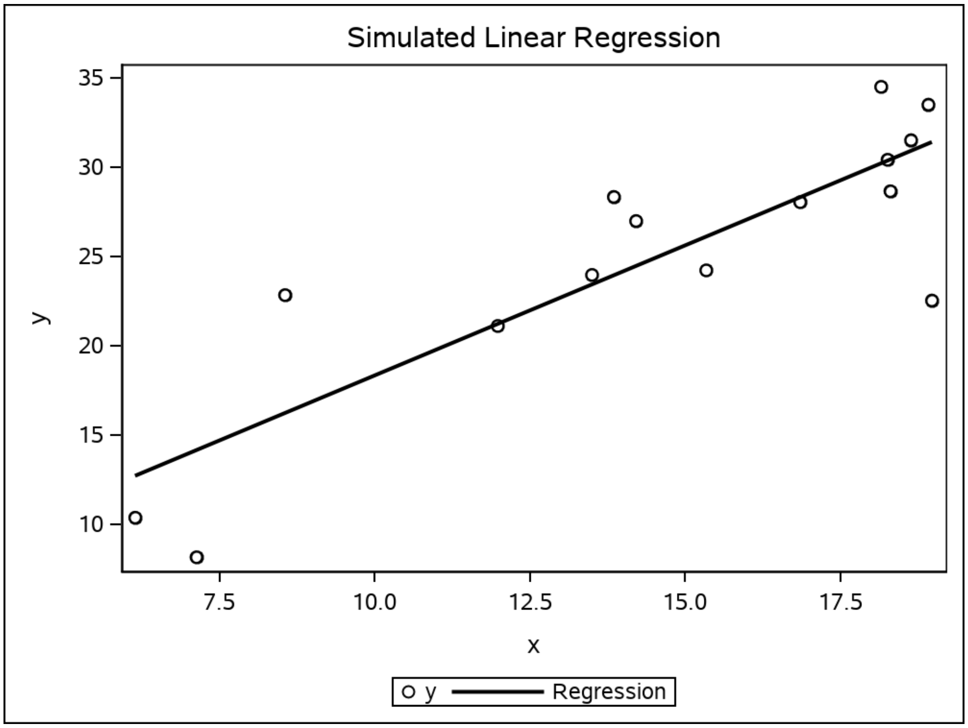
```
##    aov(formula = result)
##
## Terms:
##                        x Residuals
## Sum of Squares  502.0625   83.9085
## Deg. of Freedom        1        13
##
## Residual standard error: 2.54057
## Estimated effects may be unbalanced
```

SAS

```
libname LDATA '/home/jacktubbs/my_shared_file_links/jacktubbs/LaTeX/';

/* Simplified LaTeX output that uses plain LaTeX tables  */
ods latex path='/home/jacktubbs/my_shared_file_links/jacktubbs/LaTeX/Output'
 file='lin_reg_sim.tex' style=journal
stylesheet="sas.sty"(url="sas");

/*
http://support.sas.com/rnd/base/ods/odsmarkup/latex.html
*/
ods graphics / reset width=5in  outputfmt=png
  antialias=on;
*/
*ods graphics on;
title1 'Simulated Linear Regression';

/**********************************************************************
 Simple Linear Regression Models
 *********************************************************************/

%let N = 15;                          /* size of each sample      */
%let beta_0 = 3;                      /* true y-intercept         */
%let beta_1 = 1.4;                    /* true slope               */
%let sigma=4;                         /* true sigma               */
data Reg1(keep=x y);
call streaminit(1);
do i = 1 to &N;
   x = 15*rand("Uniform") + 5;        /* explanatory variable     */
   eps = rand("Normal", 0, &sigma);   /* error term N(0,sigma)    */
   y = &beta_0 + &beta_1*x + eps;
   output;
end;
run;

proc sgplot data=Reg1;
scatter y=y x=x;
```

```
reg y=y x=x;
run;

proc reg data=Reg1 plots=(FITPLOT diagnostics);
   model y = x;
*   ods exclude NObs;
   run;


/*******************************************************************
 Matric Notation Using PROC IML;
 ******************************************************************/
 proc iml;

    use Reg1;
    read all into tc;
    z=tc[,1];  *Used z since SAS is not case specific;
    y=tc[,2];
    n=nrow(tc);
    one=j(n,1,1);
X = onez;
create temp from X;
  append from X;

run;
b = inv(x`*x) * x`*y;
print b;

yhat = x*b;
r = y-yhat;

sse = ssq(r);
dfe = nrow(x)-ncol(x);
mse = sse/dfe;
print sse dfe mse;

H = x*inv(x`*x)*x`;
leverage = vecdiag(H);
print y yhat r leverage;

/*******************************************************************
 Define a Function within PROC IML;
 ******************************************************************/
start Regress;                    /* begin module        */
  xpxi = inv(x`*x);               /* inverse of X'X      */
  beta = xpxi * (x`*y);           /* parameter estimate  */
  yhat = x*beta;                  /* predicted values    */
  resid = y-yhat;                 /* residuals           */

  sse = ssq(resid);               /* SSE                 */
```

```
  n = nrow(x);                    /* sample size         */
  dfe = nrow(x)-ncol(x);          /* error DF            */
  mse = sse/dfe;                  /* MSE                 */
  cssy = ssq(y-sum(y)/n);         /* corrected total SS  */
  rsquare = (cssy-sse)/cssy;      /* RSQUARE             */
  results = sse  dfe  mse  rsquare;
  print results[c={"SSE" "DFE" "MSE" "RSquare"}
              L="Regression Results"];

  stdb = sqrt(vecdiag(xpxi)*mse); /* std of estimates    */
  t = beta/stdb;                  /* parameter t tests   */
  prob = 1-probf(t#t,1,dfe);      /* p-values            */
  paramest = beta  stdb  t  prob;
  print paramest[c={"Estimate" "StdErr" "t" "Pr>t"}
               L="Parameter Estimates" f=Best6.];
  print y yhat resid;
finish Regress;                   /* end module          */

run Regress;                     /* run module          */

quit;


ods latex close;

quit;
```


Simulated Linear Regression
35
30
25
20
15
10
y
7.5
10.0
12.5
15.0
17.5
x
y
Regression

Simulated Linear Regression

The REG Procedure

Model: MODEL1

Dependent Variable: y

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance					
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	1	578.35456	578.35456	34.11	<.0001
Error	13	220.40676	16.95437		
Corrected Total	14	798.76132			

Root MSE	4.11757	*R-Square*	0.7241
Dependent Mean	25.00630	*Adj R-Sq*	0.7028
Coeff Var	16.46612		

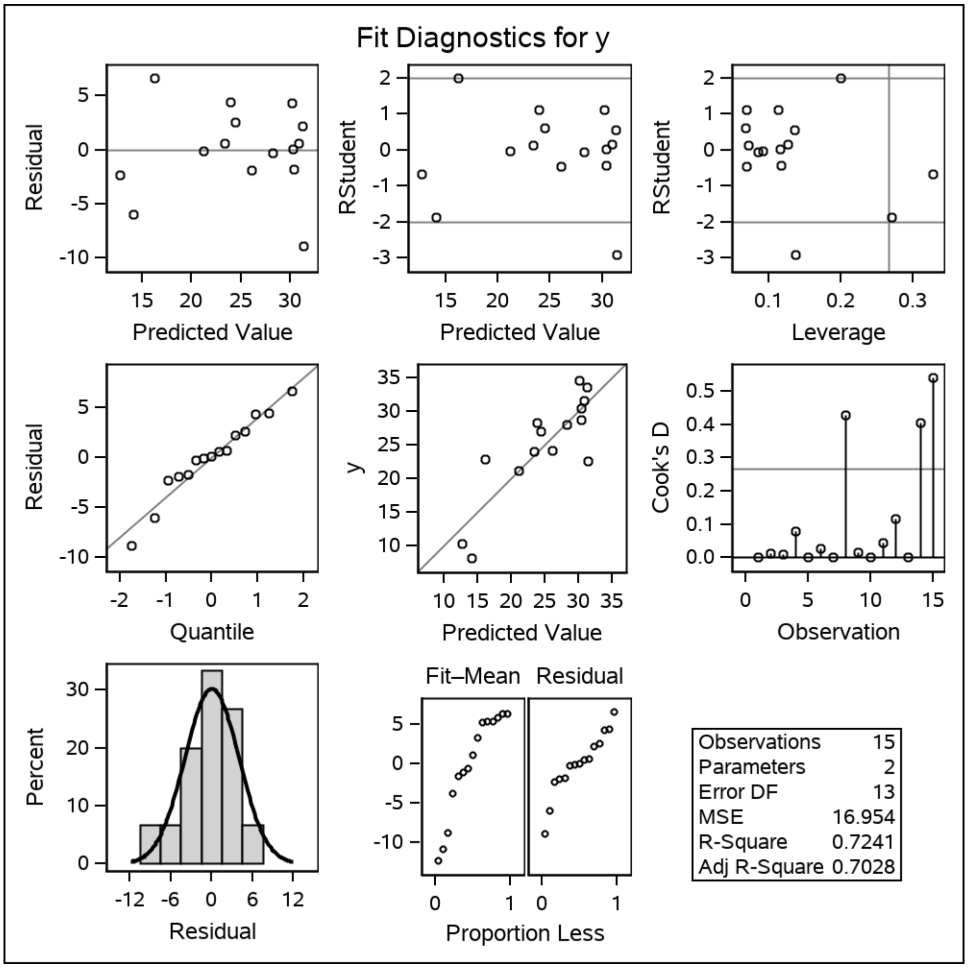
Parameter Estimates					
Variable	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	1	3.73908	3.79331	0.99	0.3423
x	1	1.45783	0.24960	5.84	<.0001

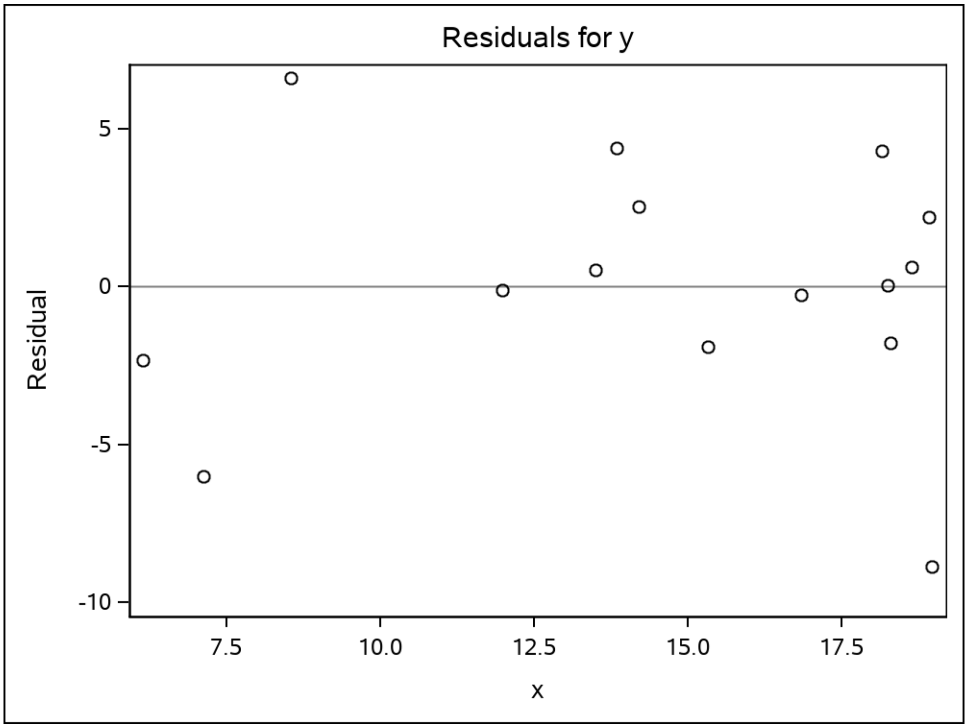
Simulated Linear Regression

The REG Procedure

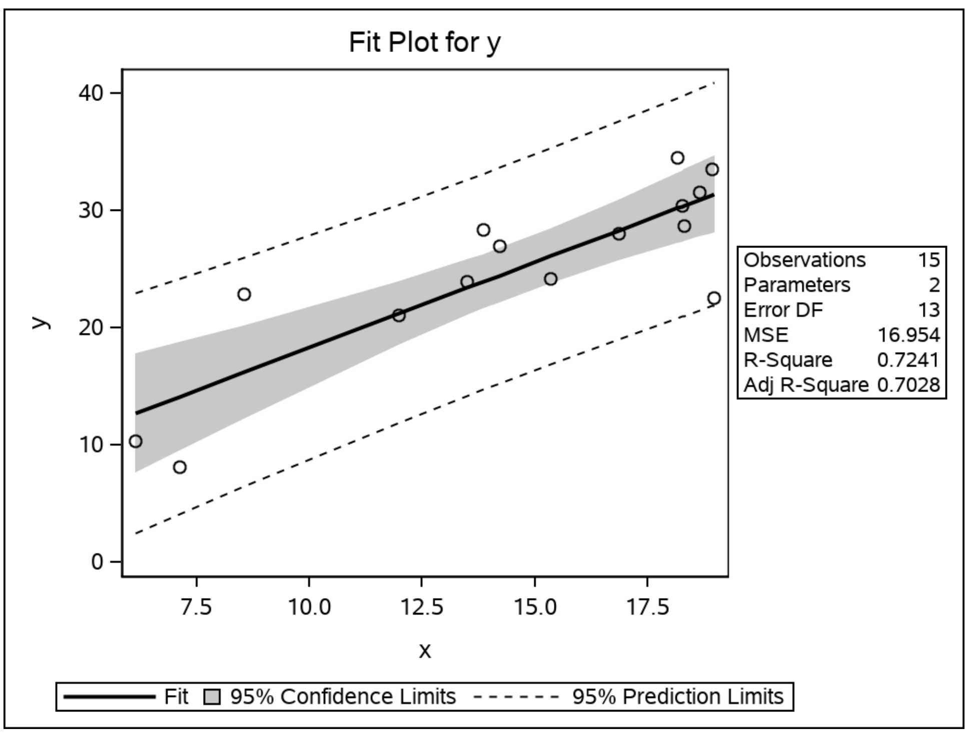
Model: MODEL1

Dependent Variable: y





Residuals for y
Residual
5
0
-5
-10
7.5
10.0
12.5
15.0
17.5
x



Fit Plot for y
y
40
30
20
10
0
7.5
10.0
12.5
15.0
17.5
x
Observations 15
Parameters 2
Error DF 13
MSE 16.954
R-Square 0.7241
Adj R-Square 0.7028
Fit
95% Confidence Limits
95% Prediction Limits

SAS Output with PROC IML

Simulated Linear Regression

b
3.7390774
1.457832

sse	dfe	mse
220.40676	13	16.954366

y	yhat	r	leverage
30.413948	30.356157	0.0577907	0.1161535
28.655401	30.423392	−1.767991	0.1174052
24.208531	26.10288	−1.894349	0.0687458
34.502726	30.209046	4.2936801	0.1134693
23.963334	23.423474	0.5398592	0.0709985
33.511479	31.321803	2.1896765	0.1356305
28.034008	28.305693	−0.271686	0.085489
22.533157	31.401699	−8.868542	0.1373864
26.98826	24.456288	2.531972	0.0671897
21.097187	21.20914	−0.111953	0.0915968
28.339971	23.932137	4.4078341	0.0686617
10.360527	12.695395	−2.334868	0.3287178
31.522593	30.914241	0.6083519	0.1270167
22.828668	16.210786	6.6178825	0.2004274
8.1347446	14.132403	−5.997659	0.2711116

Regression Results			
SSE	*DFE*	*MSE*	*RSquare*
220.40676	13	16.954366	0.7240643

Parameter Estimates			
Estimate	*StdErr*	*t*	*Pr>|t|*
3.7391	3.7933	0.9857	0.3423
1.4578	0.2496	5.8406	577E−7

y	*yhat*	*resid*
30.413948	30.356157	0.0577907
28.655401	30.423392	−1.767991
24.208531	26.10288	−1.894349
34.502726	30.209046	4.2936801
23.963334	23.423474	0.5398592
33.511479	31.321803	2.1896765
28.034008	28.305693	−0.271686
22.533157	31.401699	−8.868542
26.98826	24.456288	2.531972
21.097187	21.20914	−0.111953
28.339971	23.932137	4.4078341
10.360527	12.695395	−2.334868
31.522593	30.914241	0.6083519
22.828668	16.210786	6.6178825
8.1347446	14.132403	−5.997659

Appendix

Some Matrix Properties

The material in this section is intended as a review of material that you should have seen in an undergraduate math course in linear algebra.

Matrix Algebra - Review

A matrix $\mathbf{A} = (a_{ij}), i = 1, 2, \ldots, r, j = 1, 2, \ldots, c$ is said to be an $r \times c$ matrix given by

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1c} \\ a_{21} & a_{22} & \ldots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \ldots & a_{rc} \end{pmatrix}$$

A vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is said to be a $n \times 1$ row vector, x' is a $1 \times n$ column vector given by

$$\mathbf{x}' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Special Matrices

1. $\mathbf{D} = diag(A)$ is the diagonal of the $r \times r$ matrix A given by

$$\mathbf{D} = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{rr} \end{pmatrix}.$$

2. $\mathbf{I}_n$ is called the $n \times n$ identity matrix given by

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

3. $\mathbf{J}_n$ is an $n \times n$ matrix with each element equal to one.

4. $\mathbf{j}$ is a $n \times 1$ vector with each element equal to one where $\mathbf{J} = \mathbf{j}\mathbf{j}'$.

Addition

$C = A \pm B$ is defined as $c_{ij} = a_{ij} \pm b_{ij}$ provided both A and B have the same number of rows and columns. It can easily be shown that $(A \pm B) \pm C = A \pm (B \pm C)$ and $A + B = B + A$.

Multiplication

$C = AB$ is defined as $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$ provided A and B are conformable matrices (A is $r \times p$ and B is $p \times c$). Note: Even if both AB and BA are defined they are not necessarily equal. It follows that $A(B \pm C) = AB \pm AC$). Two vectors a and b are said to be orthogonal, denoted by $a \perp b = 0$, if $ab = \sum_{i=1}^{n} a_i b_i = 0$.

Kronecker or Direct Product

If A is $m \times n$ and B is $s \times t$, the *direct or Kronecker product* of A and B, denoted by $A \otimes B$, is an $ms \times nt$ matrix given by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}.$$

Properties are given as

1. $(A \otimes B)(C \otimes D) = (AC \otimes BD)$.
2. $((A + B) \otimes (C + D)) = (A \otimes C) + (A \otimes D) + (B \otimes C) + (B \otimes D)$.
3. $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.

Inverse

A $n \times n$ matrix A is said to be nonsingular if there exists a matrix B satisfying $AB = BA = I_n$. B is called the inverse of A is denoted by A^{-1}.

Transpose

If A is $r \times c$ then the transpose of A, denoted by A', is a $c \times r$ matrix. It follows that

1. $(A')' = A$
2. $(A \pm B)' = A' \pm B'$
3. $(AB)' = B'A'$
4. If $A = A'$ then A is said to be symmetric.
5. $A'A$ and AA' are symmetric.
6. $(A \otimes B)' = (A' \otimes B')$.

Trace

Suppose that the matrix $A = (a_{ij}), i = 1, \dots, n, j = 1, \dots, n$ then the trace of A given by $tr[A] = \sum_{i=1}^{n} a_{ii}$. Provided the matrices are conformable

1. $tr[A] = tr[A']$.
2. $tr[A \pm B] = tr[A] \pm tr[B]$.
3. $tr[AB] = tr[BA]$.
4. $tr[ABC] = tr[CAB] = tr[BCA]$.
5. $tr[A \otimes B] = tr[A]tr[B]$.

   For a square matrix A, one can write $Ax = \lambda x$ for some non-null vector x, then λ is called a *characteristic or eigenvalue or latent root* of A. x is called the corresponding characteristic vector (eigenvector or latent vector).

   If A is a symmetric $n \times n$ matrix with eigenvalues λ_i for $i = 1, 2, \dots, n$, then

6. $tr[A] = \sum_{i=1}^{n} \lambda_i$
7. $tr[A^s] = \sum_{i=1}^{n} \lambda_i^s$
8. $tr[A^{-1}] = \sum_{i=1}^{n} \lambda_i^{-1}$, A nonsingular.

Rank

Suppose that A is a $r \times c$ matrix with r rows $a_1, a_2, \dots, a_c$ are said to be linearly independent if no a_i can be expressed as a linear combination of the remaining $a_i's$, that is, there does not exist a non-null vector $c = (c_1, c_2, \dots, c_r)$ such that $\sum_{i=1}^{r} c_i a_i = 0$. It can be shown that the number of linearly independent rows is equal to the number of linearly independent columns of any matrix A and that number is the rank of the matrix. If the rank of A is r then the matrix A is said to be full row rank. If the rank of A is c then A is said to be full column rank.

1. $rank[A] = 0$ if and only if $A = 0$.
2. $rank[A] = rank[A']$.
3. $rank[A] = rank[A'A] = rank[AA']$.
4. $rank[AB] \leq min\{rank[A], rank[B]\}$
5. If A is any matrix, and P and Q are any conformable nonsingular matrices then $rank[PAQ] = rank[A]$.
6. If A is $r \times c$ with rank r then AA' is nonsingular ($(AA')^{-1}$ exists and $rank[AA'] = r$). If the rank of A is c then $A'A$ is nonsingular ($(A'A)^{-1}$ exists and $rank[A'A] = c$).
7. If A is symmetric, then $rank[A]$ is equal to the number of nonzero eigenvalues.

Quadratic Forms

Let $\mathbf{A}$ be a symmetric $n \times n$ matrix and $\mathbf{x} = (x_1, x_2, \ldots, n)$ be a vector. Then $q = \mathbf{x}'\mathbf{A}\mathbf{x}$, is called a quadratic form of A. The quadratic form is a second degree polynomial in the $x_i's$, since $q = \mathbf{z}'A\mathbf{z} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}z_i z_j$ where $A = (a_{ij})$. In this definition I have assumed that $\mathbf{A}$ is symmetric. This follows since any non-symmetric matrix $\mathbf{B}$ can be written as $1/2[\mathbf{B} + \mathbf{B}']$ and $1/2[\mathbf{B} + \mathbf{B}']$ is symmetric. Furthermore, $\mathbf{x}'\mathbf{B}\mathbf{x} = 1/2[\mathbf{x}'\mathbf{B}\mathbf{x} + \mathbf{x}'\mathbf{B}'\mathbf{x}]$.

Positive Semidefinite Matrices

A symmetric matrix A is said to be positive semidefinite (p.s.d.) if and only if $q = x'Ax \geq 0$ for all x.

1. The eigenvalues of p.s.d. matrices are nonnegative.
2. If A is p.s.d. then $tr[A] \geq 0$.
3. A is p.s.d. of rank r if and only if there exists an $n \times n$ matrix R of rank r such that $A = RR'$.
4. If A is an $n \times n$ p.s.d. matrix of rank r, then there exists an $n \times r$ matrix S of rank r such that $S'AS = I_r$.
5. If A is p.s.d., then $X'AX = 0 \Rightarrow AX = 0$.

Positive Definite Matrices

A symmetric matrix A is said to be positive definite (p.d.) if and only if $q = x'Ax > 0$ for all $x, x \neq 0$.

1. The eigenvalues of p.d. matrices are positive.
2. A is p.d. if and only if there exists an nonsingular matrix R such that $A = RR'$.
3. If A is p.d. then so is A^{-1}.
4. If A is p.d. then $rank[CAC'] = rank[C]$.
5. If A is $n \times n$ p.d. matrix and C is a $p \times n$ matrix of rank p, then CAC' is p.d.
6. If X is $n \times p$ of rank p then $X'X$ is p.d.
7. If A is p.d. if and only if all the leading minor determinants of A are positive.
8. The diagonal elements os a p.d. matrix are all positive.
9. (Cholesky decomposition). Is A is p.d. there exists a unique upper triangular matrix U with positive diagonal elements such that $A = U'U$.

Idempotent Matrices

A matrix P is said to be idempotent if $P^2 = P$. A symmetric idempotent matrix is called a projection matrix.

1. If P is symmetric, then P is idempotent and of rank r if and only if it has r eigenvalues equal to unity and $n - r$ eigenvalues equal to zero.

2. If P is a projection matrix then the $tr[P] = rank[P]$.
3. If P is idempotent, so is $I - P$.
4. Projection matrices are positive semidefinite.

Orthogonal Matrices

An $n \times n$ matrix A is said to be orthogonal if and only $A^{-1} = A'$. If A is orthogonal then

1. $-1 \leq a_i \leq 1$.
2. $AA' = A'A = I_n$.
3. $\mid A \mid = 1$.

Vector Differentiation

Let X be an $n \times m$ matrix with elements x_{ij}, then if $f(X)$ is a function of the elements of X, we define

$$\frac{df}{dX} = [(\frac{df}{dx_{ij}})]$$

then

1. $\frac{d(\beta' a)}{d\beta} = a$.
2. $\frac{d(\beta' A\beta)}{d\beta} = 2A\beta$. ($A$ symmetric).
3. if $f(X) = a'Xb$, then $\frac{df}{dX} = ab'$.
4. if $f(X) = tr[AXB]$, then $\frac{df}{dX} = A'B'$.
5. if X is symmetric and $f(X) = a'Xb$, then $\frac{df}{dX} = ab' + b'a - diag(ab')$.
6. if X is symmetric and $f(X) = tr[AXB]$, then $\frac{df}{dX} = A'B' + BA - diag(BA)$.
7. if X and A are symmetric and $f(X) = tr[AXAX]$, then $\frac{df}{dX} = 2AXA$.

The Generalized Inverse

A matrix B is said to be the generalized inverse of A if it satisfies $ABA = A$. The generalized inverse of A is denoted by A^-. If A is nonsingular then $A^{-1} = A^-$. If A is singular then A^- exists but is not unique.

1. If A is an $r \times c$ matrix of rank c. Then the generalized inverse of A is $A^- = (A'A)^{-1}A'$.
2. If A is an $r \times c$ matrix of rank r. Then the generalized inverse of A is $A^- = A(AA')^{-1}$.
3. If A is an $r \times c$ matrix of rank c. Then $A(A'A)^- A'$ is symmetric, idempotent, of rank A, and unique.

Generalized Inverse of $X'X$

Let G denote the generalized inverse of $X'X$, that is

$$X'XGX'X = X'X.$$

Clearly, $X'X$ is symmetric although G may not be. However, it follows that G' is also the generalized inverse of $X'X$, or

$$X'XG'X'X = X'X,$$

and that

$$(X'X)^{-} = GX'XG'$$

which is symmetric.

Other properties of G are;

- G' is also the generalized inverse of $X'X$.
- $XGX'X = X$ or GX' is the generalized inverse of X.
- $X'XG'X' = X'$ or XG' is the generalized inverse of X'.
- XGX' is invariant to the choice of G.
- XGX' is symmetric for any choice of G.
- For V being symmetric and positive definite (i.e. a covariance matrix) then

  $$X(X'V^{-1}X)^{-}X'V^{-1} \text{ is invariant to } (X'V^{-1}X)^{-}$$

  and

  $$X(X'V^{-1}X)^{-}X'V^{-1}X = X.$$

Solution of Linear Equations

A system of linear equations given by $Ax = b$ is said to be consistent and has a solution which can be expressed as $\tilde{x} = A^{-}b$. If A is nonsingular then $\tilde{x}$ is unique.

Random Variables and Vectors

Expectations

Let U denote a random variable with expectation $E(U)$ and $Var(U) = E(U - E(U))^2)$. Let a and b denote any constants, then we have

1. $E(aU \pm b) = aE(U) \pm b$.
2. $Var(aU \pm b) = a^2 Var(U)$.

Suppose that $t(x)$ is a statistic that is used to estimate a parameter θ. If $E(t(x)) = \theta$, the statistic is said to be an unbiased estimate for θ. If $E(t(x)) = \eta \neq \theta$ then $t(x)$ is biased and the bias is given by $Bias = (\theta - \eta)$, in which case the mean square error is given by

$$MSE(t) = E(t(x) - \theta)^2 = Var(t(x)) + Bias^2.$$

Covariance

Let U and V denote two random variables with respective means, μ_u and μ_v. The covariance between the two random variables is defined by

$$Cov(U, V) = E[(U - \mu_u)(V - \mu_v)] = E(UV) - \mu_u\mu_v.$$

If U and V are independent then $Cov(U, V) = 0$, one has the following:

1. $Cov(aU \pm b, cV \pm d) = acCov(U, V)$.
2. $-1 \leq Corr(U, V) = \rho = \frac{Cov(U,V)}{[Var(U)Var(V)]^{1/2}} \leq 1$.

Linear Combinations

Suppose that one has n r.v. given by $u_1, u_2, \ldots, u_n$ and one defines

$$u = \sum_{i=1}^{n} a_i u_i$$

where $E(u_i) = \mu_i$, $Var(u_i) = \sigma_i^2$, and $cov(u_i, u_j) = \sigma_{ij}$ when $i \neq j$. Then

1. $E(u) = \sum_{i=1}^{n} a_i \mu_i$,
2. $Var(u) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \sum \sum a_i a_j \sigma_{ij}$.

Random Vectors

Let $u = (u_1, u_2, \ldots, u_n)'$ denote a n-dimensional vector of random variables. Then the expected value of u is given by $E(u) = (E(u_1), E(u_2), \ldots, E(u_n))'$. The covariance matrix is an $n \times n$ matrix given by

$$cov(u) = E[(u - E(u))(u - E(u))'] = \Sigma = (\sigma_{ij})$$

where $\sigma_{ij} = cov(u_i, u_j)$. There are several properties for Σ

1. Σ is symmetric and at least a p.s.d. $n \times n$ matrix.
2. $E(uu') = \Sigma - E(u)E(u)'$.
3. $cov(u + d) = cov(u)$.
4. $tr[cov(u)] = trE[(u - E(u))(u - E(u))'] = E[(u - E(u))'(u - E(u))] = \sum_{i=1}^{n} \sigma_{ii}$ is the total variance of u.

   Suppose that A is a $r \times n$ matrix and one defines $v = Au \pm b$, then

5. $E(v) = AE(u) \pm b$.
6. $cov(v) = Acov(u)A' = A\Sigma A'$. Note $cov(v)$ is an $r \times r$ symmetric and at least p.s.d. matrix.

   Suppose that B is a $s \times n$ matrix and one defines $w = Bu \pm d$, then

7. $cov(v, w) = A\Sigma B'$. Note $cov(v, w)$ is a $r \times s$ matrix.

Distributions

Multivariate Normal

Recall in the univariate case, the normal density function for y is given by

$$f_y(y) = k\ exp[-1/2\sigma^2(y-\mu)^2]$$

where $E(y) = \mu$, $var(y) = \sigma^2$ and k is the normalizing constant given by

$$k = (2\pi\sigma^2)^{-1/2}.$$

Let $y = (y_1, y_2, \ldots, y_n)'$ denote an n-dimensional vector with density function given by

$$f(y_1, y_2, \ldots, y_n) = k\ exp[-1/2(y - E(y))'\Sigma^{-1}(y - E(y))],$$

where,

1. $k = (2\pi)^{-n/2} \mid \Sigma \mid^{-1/2}$ is the normalizing constant and $\mid \Sigma \mid$ is the determinate of Σ.
2. $E(y) = \mu = (\mu_1, \mu_2, \ldots, \mu_n)'$ and $cov(y) = \Sigma$.
3. $Q = (y - E(y))'\Sigma^{-1}(y - E(y)) \sim \chi^2_n$, where χ^2_n is a Chi-square with n degrees of freedom.
4. y is said to have an n-dimensional multivariate normal distribution with mean $= \mu$ and covariance matrix $= \Sigma$ provided Σ is nonsingular. This is denoted by $y \sim N_n(\mu, \Sigma)$.

Suppose that $y \sim N_n(\mu, \Sigma)$ and A is a $r \times n$. Define $u = Ay \pm b$ then $u \sim N_r(\mu_u = A\mu \pm b, \Sigma_u = A\Sigma A')$ provided $A\Sigma A'$ is nonsingular (i.e. $rank(A) = r$).

Chi-Square, T and F Distributions

Recall from univariate statistics that if $z_i \sim N(0,1)$ for $i = 1, 2, \ldots, n$ then

1. $z_i^2 \sim \chi^2(1)$ and $\sum_{i=1}^n z_i^2 \sim \chi^2(n)$.
2. $(n-1)s_z^2 = \sum_{i=1}^n (z_i - \bar{z})^2 \sim \chi^2(n-1)$.
3. $\bar{z}$ and s_z^2 are independent.
4. If $z \sim N(0,1)$ and $u \sim \chi^2(n)$ then $\frac{z}{\sqrt{u/n}} \sim$ t-dist(n).
5. If $u \sim \chi^2(n)$ and $v \sim \chi^2(m)$ then $\frac{u/n}{v/m} \sim$ F-dist(n,m).
6. $z = (z_1, z_2, \ldots, z_n)'$ then $z'z = \sum_{i=1}^n z_i^2 \sim \chi^2(n)$.
7. If $x \sim N(\mu, 1)$ then $x^2 \sim \chi^2(df = 1, \lambda = \mu^2)$. x^2 is said to have a noncentral Chi-square distribution with non-centrality parameter λ.

Quadratic Forms of Normal Variables

1. Let $z = (z_1, z_2, \ldots, z_n)' \sim N_n(0, I_n)$. Define the quadratic form $q = z'Az$ then
   (a) The expected value of q is $E(q) = tr[A]$.
   (b) The variance of q is $Var(q) = 2\ tr[A^2]$.
   (c) $q \sim \chi^2(a)$ if and only if $A^2 = A$ (A is idempotent) where $a = rank[A] = tr[A]$.
2. Let $x = (x_1, x_2, \ldots, x_n)' \sim N_n(\mu, I_n)$. Define the quadratic form $q = x'Ax$ then
   (a) The expected value of q is $E(q) = tr[A] + \mu' A\mu$.
   (b) The variance of q is $Var(q) = 2\ tr[A^2] + 4\mu' A^2 \mu$.
   (c) $q \sim \chi^2(a, \lambda)$ if and only if $A^2 = A$ (A is idempotent) where $a = rank[A] = tr[A]$ and $\lambda = 1/2\mu' A\mu$.
   (d) If $x \sim N_n(\mu, \sigma^2 I_n)$ then $(x-\mu)'A(x-\mu)/\sigma^2 \sim \chi^2(a)$ if and only if A is idempotent and $a = tr[A]$.
3. Let $x = (x_1, x_2, \ldots, x_n)' \sim N_n(\mu, V)$ (This means that the $x_i's$ are not independent of one another). Define the quadratic form $q_1 = x'Ax$ then
   (a) The expected value of q_1 is $E(q_1) = tr[AV] + \mu' A\mu$.
   (b) The variance of q_1 is $Var(q_1) = 2\ tr[AVAV] + 4\mu' AV\mu$.
   (c) $q_1 \sim \chi^2(a, \lambda)$ if and only if $(AV)^2 = AV$ (AV is idempotent) where $a = rank[A]$ and $\lambda = 1/2\mu' A\mu$.

   Suppose that $q_2 = x'Bx$ and $t = Cx$ where C is an $c \times n$ matrix. Then

   (d) $cov(q_1, q_2) = 2\ tr[AVBA] + 4\mu' AVB\mu$.
   (e) $cov(x, q_1) = 2\ VA\mu$.
   (f) $cov(t, q_1) = 2\ CVA\mu$.
   (g) q_1 and q_2 are independent if and only if $AVB = BVA = 0$.
   (h) q_1 and t are independent if and only if $CVA = 0$.
4. (Cochran's Theorem) Let $x \sim N_n(\mu, V), A_i, i = 1, 2, \ldots, m$ be symmetric, $rank[A_i] = r_i$, and

$$A = \sum A_i$$

   with $rank[A] = r$. If AV is idempotent, and $r = \sum r_i$ then $q_i = x'A_i x$ are mutually independent with $q_i \sim \chi^2(df = r_i, \lambda_i = \mu' A_i \mu/2)$.