R Output for Linear Models using functions lm(), gls() & glm()

Different kinds of output related to linear models can be obtained in R using function lm() {stats} in the base installation as well as gls() & glm() in package {nlme} among others. The latter functions are primarily designed to extend linear models in various ways, but default to more-or-less the same analysis with differences in the format of the output. The objective of this worksheet is to show output from each function on the same data set, along with basic definitions. Since linear models may include a combination of numeric covariates and factors, simple ANCOVA data is utilized here from Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition. Specifics of the t & F tests can be found in Worksheet LM 03. Matrix algebra calculations are described in *Biostatistics* Worksheet 380.

Example in R:

```
#COMPARISON OF OUTPUT FROM lm(), gls() & glm():
library(nlme)
op=options(digits=15)
setwd("c:/DATA/Models")

#KNNL TABLE 22.1
K=read.table("CH22TA01.txt",header=TRUE)
K$T=factor(K$treatment)
K
```

```
> K
    Y  X treatment store T
1  38 21         1     1 1
2  39 26         1     2 1
3  36 22         1     3 1
4  45 28         1     4 1
5  33 19         1     5 1
6  43 34         2     1 2
7  38 26         2     2 2
8  38 29         2     3 2
9  27 18         2     4 2
10 34 25         2     5 2
11 24 23         3     1 3
12 32 29         3     2 3
13 31 30         3     3 3
14 21 16         3     4 3
15 28 29         3     5 3
```

Variables:
Y = dependent variable
X = numeric covariate
T = factor

Using lm():

```
#LINEAR MODEL lm():
FM=lm(Y~X+T,data=K) #FULL MODEL
summary(FM)
anova(FM)
drop1(FM)
```

```
> summary(FM)

Call:
lm(formula = Y ~ X + T, data = K)

Residuals:
             Min               1Q           Median               3Q
-2.434813925570 -1.273889555822 -0.336254501801  1.671008403361
             Max
 2.486914765906

Coefficients:
                    Estimate       Std. Error   t value   Pr(>|t|)
(Intercept)  17.353421368547   2.523004120596   6.87808 2.6635e-05 ***
X             0.898559423770   0.102584878233   8.75918 2.7310e-06 ***
T2           -5.075390156062   1.228965128624  -4.12981  0.0016727 **
T3          -12.976830732293   1.205623298658 -10.76359 3.5272e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.87256043775 on 11 degrees of freedom
Multiple R-squared: 0.940329040032,     Adjusted R-squared: 0.924055141
F-statistic: 57.7814258162 on 3 and 11 DF,  p-value: 5.08171867379e-07
```

Marginal output for covariate & each level of the factor

```
> anova(FM)
Analysis of Variance Table

Response: Y
          Df         Sum Sq         Mean Sq  F value     Pr(>F)
X          1 190.6777777778 190.67777777778 54.37865 1.4048e-05 ***
T          2 417.1509136988 208.57545684941 59.48282 1.2636e-06 ***
Residuals 11  38.5713085234   3.50648259304
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Serial output for covariate & combined factor levels in order specified by the formula

Marginal output for covariate & factor

```
> drop1(FM)
Single term deletions

Model:
Y ~ X + T
       Df      Sum of Sq            RSS            AIC
<none>                    38.5713085234 22.16687744720
X       1 269.0286914766 307.6000000000 51.31125054794
T       2 417.1509136988 455.7222222222 55.20749891507
```

Using gls():

```
#GAUSSIAN GENERALIZED LEAST SQUARES gls():
FM2=gls(Y~X+T,data=K) #FULL MODEL
summary(FM2)
anova(FM2)
drop1(FM2)
```

```
> summary(FM2)
Generalized least squares fit by REML
  Model: Y ~ X + T
  Data: K
               AIC              BIC            logLik
  65.6544520507195 67.6439284147113 -27.8272260253597

Coefficients:
                           Value         Std.Error             t-value
p-value
(Intercept)  17.35342136854739 2.523004120596327   6.87807888496266
0.0000
X             0.89855942376951 0.102584878232654   8.75918009798337
0.0000
T2           -5.07539015606242 1.228965128623691  -4.12980811078530
0.0017
T3          -12.97683073229291 1.205623298657670 -10.76358655870469
0.0000

 Correlation:
   (Intr) X      T2
X  -0.943
T2  0.026 -0.267
T3 -0.054 -0.187  0.523

Standardized residuals:
               Min                  Q1                 Med
Q3
-1.300259194032781 -0.680292892095387 -0.179569371979373
0.892365538474492
               Max
 1.328082509751125

Residual standard error: 1.872584377528
Degrees of freedom: 15 total; 11 residual
```

```
> anova(FM2)
Denom. DF: 11
            numDF         F-value p-value
(Intercept)     1 4887.11965490090  <.0001
X               1   54.37864661197  <.0001
T               2   59.48281541838  <.0001
```

```
> drop1(FM2)
Error in terms.default(object) : no terms component nor attribute
```

function not defined for gls()

Using glm():

```
#GAUSSIAN GENERALIZED LINEAR MODEL glm():
FM3=glm(Y~X+T,data=K) #FULL MODEL
summary(FM3)
anova(FM3)
drop1(FM3)
```

> summary(FM3)

```
Call:
glm(formula = Y ~ X + T, data = K)

Deviance Residuals:
           Min              1Q          Median              3Q
-2.434813925570  -1.273889555822  -0.336254501801   1.671008403361
           Max
 2.486914765906

Coefficients:
                    Estimate       Std. Error   t value   Pr(>|t|)
(Intercept)  17.353421368547   2.523004120596   6.87808 2.6635e-05 ***
X             0.898559423770   0.102584878233   8.75918 2.7310e-06 ***
T2           -5.075390156062   1.228965128624  -4.12981  0.0016727 **
T3          -12.976830732293   1.205623298658 -10.76359 3.5272e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 3.50648259303722)

    Null deviance: 646.40000000000  on 14  degrees of freedom
Residual deviance:  38.57130852341  on 11  degrees of freedom
AIC: 66.73503344334

Number of Fisher Scoring iterations: 2
```

> anova(FM3)

```
Analysis of Deviance Table

Model: gaussian, link: identity

Response: Y

Terms added sequentially (first to last)


     Df       Deviance Resid. Df     Resid. Dev
NULL                          14 646.4000000000
X     1 190.6777777778        13 455.7222222222
T     2 417.1509136988        11  38.5713085234
```

> drop1(FM3)

```
Single term deletions

Model:
Y ~ X + T
       Df       Deviance            AIC
<none>     38.5713085234 66.73503344334
X       1 307.6000000000 95.87940654408
T       2 455.7222222222 99.77565491121
```

Model Matrix, Hat Matrix & Fitted Values:

#MODEL MATRIX, HAT MATRIX & FITTED VALUES:
X=model.matrix(FM)
X #INDEPENDENT VARIABLES IN MODEL FORM
H=X%*%solve(t(X)%*%X)%*%t(X)
H # HAT MATRIX

^ NxN matrix

> X #INDEPENDENT VARIABLES IN MODEL FORM

```
   (Intercept)  X T2 T3
1            1 21  0  0
2            1 26  0  0
3            1 22  0  0
4            1 28  0  0
5            1 19  0  0
6            1 34  1  0
7            1 26  1  0
8            1 29  1  0
9            1 18  1  0
10           1 25  1  0
11           1 23  0  1
12           1 29  0  1
13           1 30  0  1
14           1 16  0  1
15           1 29  0  1
attr(,"assign")
[1] 0 1 2 2
attr(,"contrasts")
attr(,"contrasts")$T
[1] "contr.treatment"
```

```
H%*%K$Y  #FITTED VALUES CALCULATED USING HAT MATRIX
fitted(FM)
```

$$H = X\cdot(X^TX)^{-1}\cdot X^T$$

$$Fitted = H\cdot Y$$

```
> H%*%K$Y  #FITTED VALUES CALCULATED USING HAT MATRIX
                [,1]
1  36.2231692677070
2  40.7159663865545
3  37.1217286914766
4  42.5130852340937
5  34.4260504201680
6  42.8290516206482
7  35.6405762304921
8  38.3362545018006
9  28.4521008403361
10 34.7420168067227
11 25.0434573829531
12 30.4348139255701
13 31.3333733493397
14 18.7535414165666
15 30.4348139255701

> fitted(FM)
                1                 2                 3                 4
36.2231692677071 40.7159663865546 37.1217286914766 42.5130852340936
                5                 6                 7                 8
34.4260504201681 42.8290516206483 35.6405762304922 38.3362545018007
                9                10                11                12
28.4521008403361 34.7420168067227 25.0434573829532 30.4348139255702
               13                14                15
31.3333733493397 18.7535414165666 30.4348139255702
```

Regression Parameter Estimates:

```
#REGRESSION PARAMETER ESTIMATES:
b=solve(t(X)%*%X)%*%t(X)%*%K$Y  #Coefficients
b
```

$$b = X^TX\cdot(X^TX)^{-1}$$

```
> b
                    [,1]
(Intercept)  17.353421368547416
X             0.898559423769506
T2           -5.075390156062418
T3          -12.976830732292926
```

Sum of Squares Error & Mean Squares Error:

```
#SUM OF SQUARES ERROR(SSE) = RESIDUAL SUM OF SQUARES (RSS):
SSE=sum(residuals(FM)^2)
SSE
RSS=deviance(FM) # deviance is the same as RSS=SSE in Linear Models
RSS
```

deviance() calculates SSE = RSS for Linear Models

```
#RESIDUAL DEGREES OF FREEDOM:
n=length(K$T) # number of objects
n
k=4 # number of parameters/factor levels including intercept
dfR=df.residual(FM)
dfR  # = n-k
```

df.residual() gives residual degrees of freedom

```
#MEAN SQUARES ERROR:
MSE=SSE/dfR
MSE
```

```
> SSE
[1] 38.5713085234092

> RSS
[1] 38.5713085234092

> n
[1] 15

> dfR  # = n-k
[1] 11

> MSE
[1] 3.5064825930372
```

Residual Standard Error = Error Standard Deviation:

```
#RESIDUAL STANDARD ERROR:
RSE=sqrt(MSE)
RSE
```

```
> RSE
[1] 1.87256043775286
```

Standard Error of Regression Estimates:

```
#STANDARD ERROR OF REGRESSION ESTMATES:
sb=sqrt(diag(MSE*solve(t(X)%*%X)))
sb
```

$sb = sqrt(diag(MSE \cdot (X^T X)^{-1}))$

```
> sb
      (Intercept)                 X                T2
  T3
2.523004120596323 0.102584878232654 1.228965128623691
1.205623298657670
```

Saturated Model:

```
#SATURATED MODEL:
A=factor(1:length(K$Y))
FMS=glm(Y~A,data=K)
summary(FMS)
anova(FMS)
anova(FMS,FM3)
```

The saturated model contains a parmeter for each data point and as a result has zero error variance.

zero SSE

anova() fails in here

```
> summary(FMS)

Call:
glm(formula = Y ~ A, data = K)

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.80000000000e+01         NA      NA       NA
A2           1.00000000000e+00         NA      NA       NA
A3          -2.00000000000e+00         NA      NA       NA
A4           7.00000000000e+00         NA      NA       NA
A5          -5.00000000000e+00         NA      NA       NA
A6           5.00000000000e+00         NA      NA       NA
A7           5.03060675734e-14         NA      NA       NA
A8           5.25195927430e-14         NA      NA       NA
A9          -1.10000000000e+01         NA      NA       NA
A10         -4.00000000000e+00         NA      NA       NA
A11         -1.40000000000e+01         NA      NA       NA
A12         -6.00000000000e+00         NA      NA       NA
A13         -7.00000000000e+00         NA      NA       NA
A14         -1.70000000000e+01         NA      NA       NA
A15         -1.00000000000e+01         NA      NA       NA

(Dispersion parameter for gaussian family taken to be NaN)

    Null deviance: 6.464000000000e+02  on 14  degrees of freedo
Residual deviance: 9.504196186103e-27  on  0  degrees of freedo
AIC: -864.8235586097

Number of Fisher Scoring iterations: 1
```

```
> anova(FMS)
Error in if (dispersion == 1) Inf else object$df.residual :
  missing value where TRUE/FALSE needed
```

```
> anova(FMS,FM3)
Analysis of Deviance Table

Model 1: Y ~ A
Model 2: Y ~ X + T
  Resid. Df     Resid. Dev  Df        Deviance
1         0  0.00000000000
2        11 38.57130852341 -11 -38.57130852341
```

Null Model:

```
#NULL MODEL:
NM=glm(Y~1,data=K)
summary(NM)
anova(NM)
anova(NM,FM3)
```

SSE for the Null model = Total SS

> summary(NM)

```
Call:
glm(formula = Y ~ 1, data = K)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -12.8    -4.3     0.2     4.2    11.2

Coefficients:
                 Estimate     Std. Error t value   Pr(>|t|)
(Intercept) 33.80000000000  1.75445012414 19.2653 1.7849e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 46.171428571428

    Null deviance: 646.4  on 14  degrees of freedom
Residual deviance: 646.4  on 14  degrees of freedom
AIC: 103.0186805877

Number of Fisher Scoring iterations: 2
```

> anova(NM)

```
Analysis of Deviance Table

Model: gaussian, link: identity

Response: Y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    14      646.4
```

> anova(NM,FM3)

```
Analysis of Deviance Table

Model 1: Y ~ 1
Model 2: Y ~ X + T
  Resid. Df     Resid. Dev Df       Deviance
1        14 646.4000000000
2        11  38.5713085234  3 607.8286914766
```

```
›m
›m
```

:6)