

UNIVERSITY OF FORT HARE

STM 505

DEGREE EXAMINATIONS
JUNE 2017

.....

Time : 3 hours
Subject : FIXED EFFECTS MODELLING
Marks : 100

This paper consists of 13 printed pages including the cover page

Internal Examiners

Mr. I. Jubane
Prof Y. Qin

External Examiners

Prof DJ. De Waal

Instructions

Answer all questions

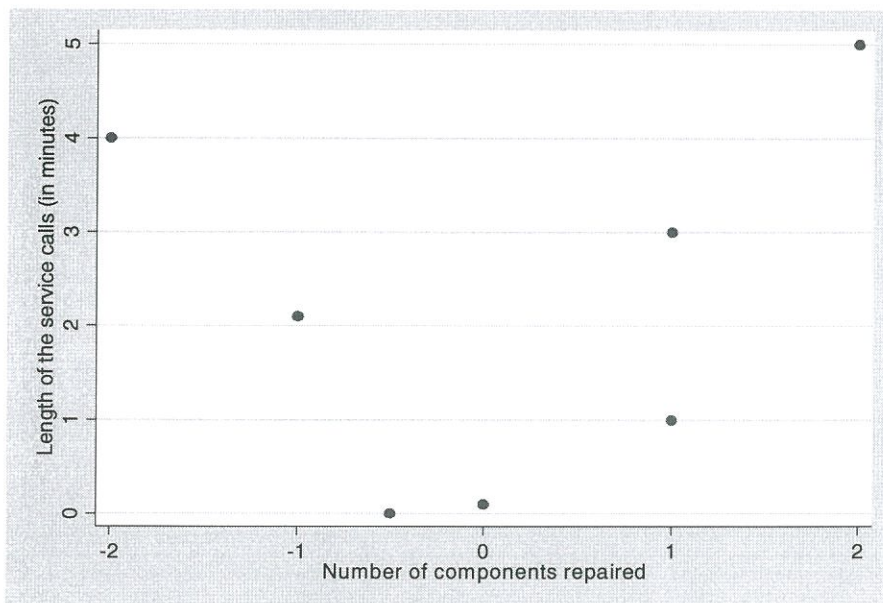
Statistical Tables will be provided

QUESTION ONE

[26 MARKS]

1.1 Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ($i = 1, \dots, n$) where $\mathbb{E}[\varepsilon] = 0$ and $\text{var}[\varepsilon] = \sigma^2 \mathbf{I}$. Write down the model using matrix notation and show that the least squares estimates of β_0 and β_1 are uncorrelated if and only if $\bar{x} = 0$. [6]

1.2 A company which markets and repairs small computers needs to forecast the number of service engineers required over the next few years. This requires consideration of the length of service calls, which in turn depends on the number of components that need to be repaired or replaced. The scatterplot given below relates the number of components repaired (*explanatory variable*) and the length of the service calls (in minutes) for a random sample of 7 calls.



We propose fitting a linear regression model

$$\mathbf{Y} = \beta \mathbf{X} + \varepsilon$$

(a) Comment on the STATA's simple regression analysis output and make recommendations. [3]

Source	SS	df	MS	Number of obs =	7
Model	.706260288	1	.706260288	F(1, 5) =	0.16
Residual	21.7080254	5	4.34160509	Prob > F =	0.7034
Total	22.4142857	6	3.73571429	R-squared =	0.0315
				Adj R-squared =	-0.1622
				Root MSE =	2.0837

(b) Explain how you would test the hypothesis $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$. [2]

- (c) Using your recommendations in (a), fit an appropriate model which deals with the nonlinearity indicated in the scatter diagram. The *data* is: [7]

X	-1	-0.5	1	1	-2	0	2
Y	2.1	0	3	1	4	0.1	5

- (d) Derive the F test for testing $H : 4\beta_1 = \beta_2$ and draw conclusions. [8]

QUESTION TWO

[13 MARKS]

- 2.1 The link between use of an oral contraceptives and the incidence of myocardial infarction was investigated, let Y denote the response variable taking the value 1 if the woman has a myocardial infarction, and 0 if she does not, and let Z denote the indicator variable taking the value 1 if oral contraceptives were used, and 0 if it were not. The fitted logistic regression model relating $\Pr(Y = 1)$ to Z , is given by

$$\text{logit}(\pi) = -1.3564 + 0.9366Z$$

- (a) Give a general formula or expression for the derivation of the model parameters β of the above model. [4]
 (b) Compute the log odds ratio. [1]
 (c) What are the odds of having myocardial infarction for women using the pill with the odds of having a myocardial infarction for women not using the pill? [2]

- 2.2 The output below relates a multiple regression model for the presence or absence of evidence of significant coronary heart disease (CHD) for 100 subjects selected to participate in a study versus AGE, Weight at Last Menstrual Period (LWT), and number of First Trimester Physician Visits (FTV) from the Low Birth Weight Study.

Variable	Coeff.	Std. err.	z	P> z
Constant	1.295	1.0714	1.21	0.227
AGE	-0.024	0.0337	-0.71	0.480
LWT	-0.014	0.0065	-2.18	0.029
FVT	-0.049	0.1672	-0.30	0.768

Log likelihood = -111.286

A formulation of the information matrix which is useful when discussing model fitting and assessment of fit is $I(\beta) = \mathbf{X}'\mathbf{W}\mathbf{X}$

- (a) What is $\mathbf{W}$? [1]
 (b) Explain how the log-likelihood ratio can be used to choose between two logistic regression models. [3]
 (c) Comment on the model fit if the log likelihood for the constant model is $L = -117.336$. [2]

QUESTION THREE**[15 MARKS]**

3.1 The figure below shows the data from a regression study with FIVE predictors.

In each question, the correct answer may be more than one of A; B or C.

- (a) Give a definition of leverage. Which of the labelled points (A; B or C) would you think have a high leverage value? Explain. **[3]**
- (b) Which of the labelled points (A; B or C) would you think would be labelled as an outlier by an outlier detection test? Explain. **[2]**
- (c) What does Cook's Distance try to measure for a multiple linear regression model? Which of the labelled points (A; B or C) would you think would have a large Cook's distance? Explain. **[3]**

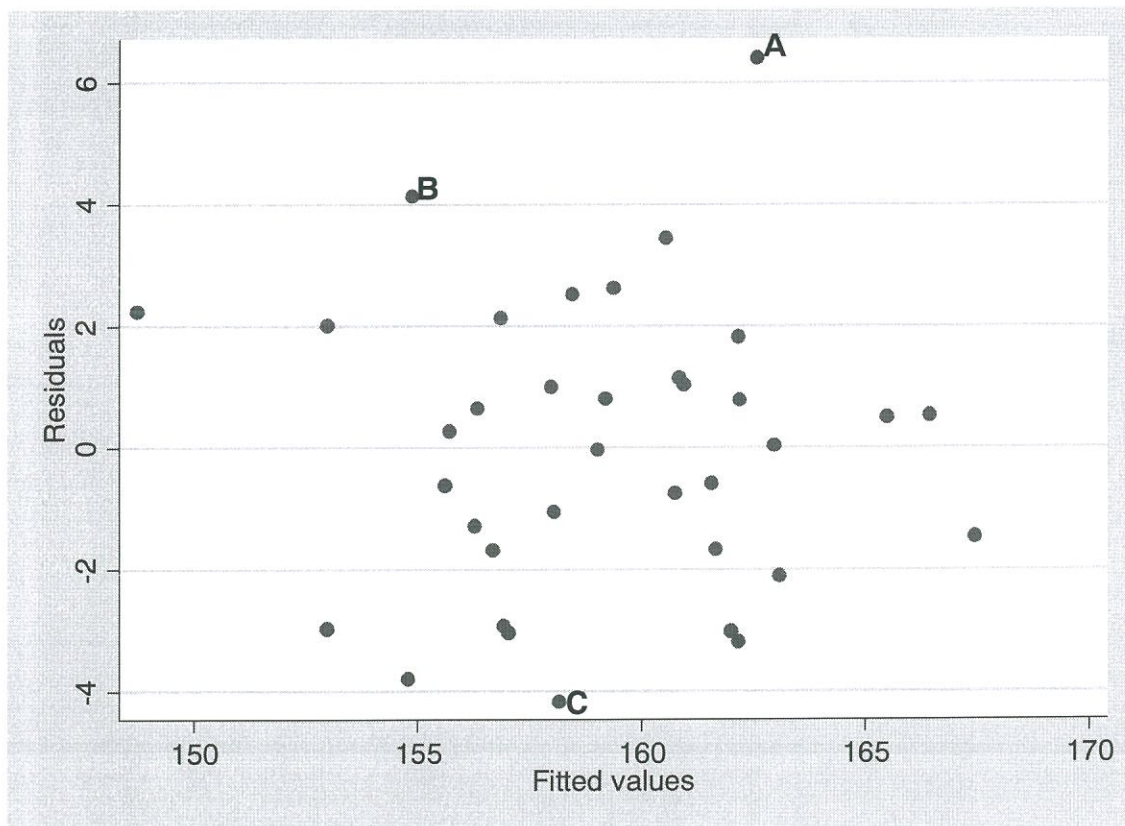


Figure 3(a)

3.2 The figures below shows the diagnostic plots from a model fit predict $\hat{Y}$ (Per capita expenditure) on five explanatory variables.

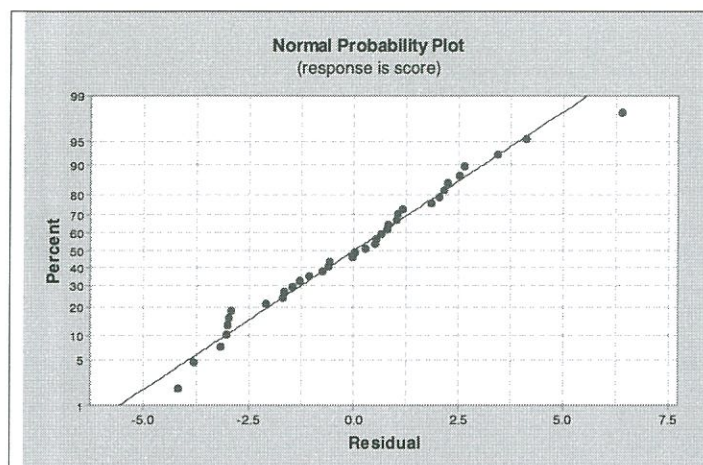


Figure 3(b)

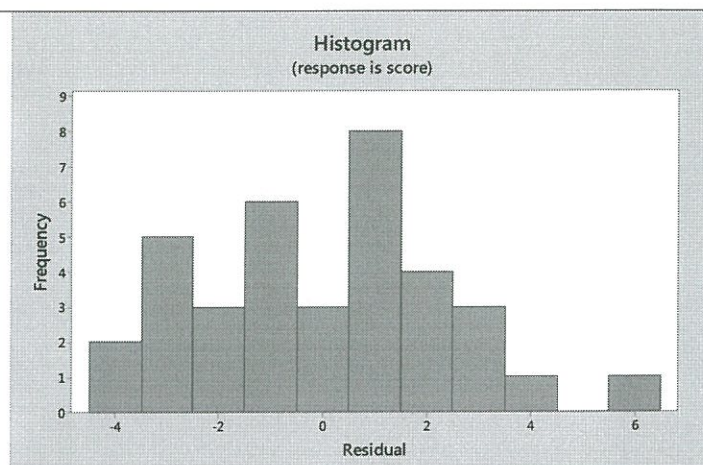


Figure 3(c)

Briefly describe the assumptions used in fitting this model. Do any of the assumptions seem to be violated based on the figures above? If any seem to be violated, suggest some possible fixes. [7]

QUESTION FOUR**[36 MARKS]**

It is of interest to understand the relationship, if any, between the analytical skills of young gifted children and a number of explanatory variables. The analytical skills are evaluated using a standard testing procedure, and the score in this test is used as the response variable.

The seven explanatory variables considered are father's IQ, mother's IQ, age in months when the child first said 'mommy' or 'daddy', age in months when the child first counted to 10 successfully, average number of hours per week the child's mother or father reads to the child, average number of hours per week the child watched an educational program on TV during the past three months, average number of hours per week the child watched cartoons on TV during the past three months. It is not expected that any of the explanatory variables interact.

The data are collected from schools in a large US city on a set of thirty-six children who were identified as gifted children soon after they reached the age of four.

Explanatory variables as follows:

x_1 : father's IQ

x_2 : mother's IQ

x_3 : age in months when the child first said 'mommy' or 'daddy'

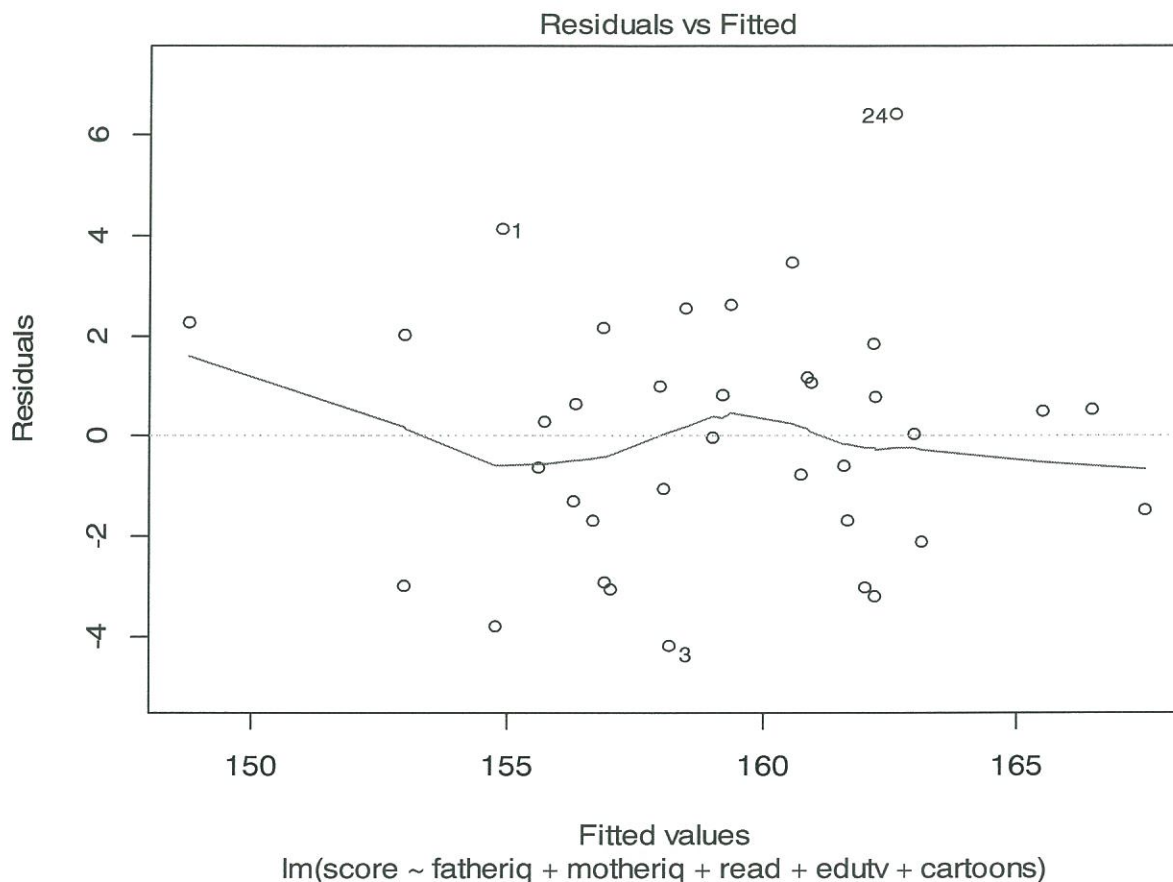
x_4 : age in months when the child first counted to 10 successfully

x_5 : average number of hours per week the child's mother or father reads to the child

x_6 : average number of hours per week the child watched an educational program on TV during the past three months

x_7 : average number of hours per week the child watched cartoons on TV during the past three months

- 4.1 Define multicollinearity and comment on its presence in this data. **[2]**
- 4.2 Investigate whether the analytical skills of young gifted children are related to any of the seven explanatory variables considered. In particular, use the provided computer output to find the model which gives the best fit to the data, and find out which one of the explanatory variables in itself provides most useful information about the analytical skills of young gifted children. **[7]**
- 4.3 Consider the 5 best predictor model found by R's step procedure. What happens to the score as the number of hours per week the child watched cartoons increases? How would you test whether this coefficient was greater than 2? **[1, 3]**
- 4.4 How would you test if the coefficient on the mother's IQ was the same as the coefficient on the father's IQ? **[6]**
- 4.5 Define an influential observation. How do we use COVRATIO to identify influential observations? **[3]**
- 4.6 Why do you think observations **1**, **3** and **24** are flagged in the diagnostic graph below? Show all working. **[14]**



QUESTION FIVE

[10 MARKS]

Description: For a 82 chondrites, the age and type (Type I or Type II) were noted, and the rust component was measured. Initially, the research question concerned the dependence of rustiness (Y) on age (X). A further question is whether there is a difference between the two types of chondrites, Type I and Type II.

- 5.1 We seek to verify a linear relationship between rustiness and age of chondrite. How do we incorporate the effect of different types of chondrites into the model? [2]
- 5.2 What hypothesis would you test to determine whether there is a difference in mean rustiness from Type I and Type II chondrites? [1]
- 5.3 Compare whether the effect of age on the rustiness of chondrites is the same for each type. [7]

***** END *****

COMPUTER OUTPUT

```
> cor(gifted)
      score fatheriq motheriq   speak   count   read   edutv cartoons
score    1.0000  0.18808  0.57124  0.26789  0.54421  0.52520 -0.3703  0.2451
fatheriq 0.1881  1.00000 -0.02481 -0.03054 -0.07502 -0.06822  0.1162 -0.2484
motheriq 0.5712 -0.02481  1.00000  0.07219  0.02426 -0.04303 -0.3300  0.3384
speak    0.2679 -0.03054  0.07219  1.00000  0.05954  0.18507 -0.1545  0.1094
count    0.5442 -0.07502  0.02426  0.05954  1.00000  0.91025 -0.2157  0.1549
read     0.5252 -0.06822 -0.04303  0.18507  0.91025  1.00000 -0.1666  0.1257
edutv    -0.3703  0.11622 -0.33000 -0.15452 -0.21568 -0.16656  1.0000 -0.9234
cartoons 0.2451 -0.24835  0.33842  0.10936  0.15490  0.12573 -0.9234  1.0000
```

```
> XX <- cor(gifted[,-7])
```

```
> eigen(XX, only.values = TRUE)
```

```
$values
```

```
[1] 2.43834 1.71913 0.99426 0.94763 0.75911 0.08555 0.05598
```

```
$vectors
```

```
NULL
```

```
f2 = step(lm(formula = score ~ fatheriq + motheriq + speak + count + read +
  edutv + cartoons, data = gifted))
```

```
> f2
```

```
Residuals:
```

```
      Min      1Q   Median      3Q      Max
-4.16242 -1.67298  0.16074  1.32635  6.41215
```

```
Coefficients:
```

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 79.046776   23.739610  3.3297  0.002313 **
fatheriq     0.237361    0.136390  1.7403  0.092055 .
motheriq     0.412112    0.071767  5.7424 2.871e-06 ***
read        11.901319    2.172310  5.4786 6.031e-06 ***
edutv        -4.888141    2.161545 -2.2614  0.031143 *
cartoons     -3.820233    1.967591 -1.9416  0.061632 .
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.5805 on 30 degrees of freedom
```

```
Multiple R-squared:  0.73376, Adjusted R-squared:  0.68938
```

```
F-statistic: 16.536 on 5 and 30 DF, p-value: 7.6691e-08
```

```
> anova(f2)
```

```
Analysis of Variance Table
```

```
Response: score
```

```
      Df   Sum Sq Mean Sq F value    Pr(>F)
fatheriq  1  26.5417  26.5417  3.98593  0.055027 .
motheriq  1 249.0076 249.0076 37.39507 1.0110e-06 ***
read     1 240.0982 240.0982 36.05708 1.3759e-06 ***
edutv    1   9.7910   9.7910  1.47038  0.234749
cartoons  1 25.1020 25.1020  3.76973  0.061632 .
Residuals 30 199.7651   6.6588
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> data.frame(Residuals=resid(f2),leverage=hatvalues(f2),covratio=covratio(f2))
```

```
      Residuals   leverage   covratio
```

1	4.132011213	0.153005048	0.76433816
2	1.830059822	0.157313200	1.28916609
3	-4.162415028	0.138490687	0.75263873
4	-1.049545495	0.384245307	1.88578222
5	0.289325589	0.216573035	1.55936177
6	-2.965211728	0.260423601	1.14676500
7	-1.282071805	0.125818297	1.32463120
8	-2.095563280	0.102275442	1.17650643
9	0.032144942	0.185124036	1.50394340
10	1.052953782	0.174057958	1.42501986
11	-2.913709174	0.154415216	1.06368472
12	-3.175735338	0.133802246	0.98682741
13	0.545313308	0.252637506	1.62036537
14	2.027851289	0.229540331	1.35214100
15	-0.032641788	0.198690560	1.52940300
16	2.144326184	0.092895697	1.15799791
17	0.810265973	0.091104088	1.31942896
18	-3.033621819	0.066737682	0.96922521
19	-0.753778235	0.402698216	1.99392067
20	-3.784192715	0.143647187	0.84700548
21	0.510963511	0.161207228	1.44750994
22	2.533396694	0.167881681	1.16292469
23	1.158440654	0.060344663	1.24932322
24	6.412152847	0.139872674	0.27611494
25	-1.671090709	0.214574379	1.40100248
26	-0.583258417	0.062106723	1.29255929
27	-1.458078657	0.162649633	1.35550636
28	0.793557954	0.098374847	1.33102726
29	1.010248803	0.146571793	1.38524422
30	-1.678635587	0.078518527	1.21242948
31	-0.613014734	0.125763773	1.38387773
32	0.654459413	0.121325312	1.37450136
33	2.250038796	0.328669629	1.44920249
34	2.626852929	0.195640184	1.17088306
35	3.446780477	0.223375051	0.97843803
36	-3.008579670	0.049628565	0.96194769

MINITAB OUTPUT

Response is score
Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	75.5	24.0	(26.3, 124.7)	3.14	0.004	
fatheriq	0.252	0.138	(-0.029, 0.534)	1.84	0.077	1.20
motheriq	0.4001	0.0729	(0.2507, 0.5494)	5.49	0.000	1.17
speak	0.188	0.148	(-0.115, 0.490)	1.27	0.214	1.18
count	0.206	0.266	(-0.339, 0.752)	0.78	0.445	6.89
read	7.54	5.59	(-3.90, 18.99)	1.35	0.188	6.90
edutv	-4.20	2.25	(-8.80, 0.40)	-1.87	0.072	8.21
cartoons	-3.34	2.02	(-7.47, 0.79)	-1.65	0.109	8.37

Best Subsets Regression: score versus fatheriq, motheriq, ...

Response is score

							f	m	c
							a	o	a
							t	t	r
							h	h	s
							c	e	t
							e	e	p
							r	r	e
							i	i	a
							n	a	n
							t	n	t
							d	v	s
Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	q	q	k
1	32.6	30.7	572.9	23.6	43.3	3.8557	X		
1	29.6	27.5	579.8	22.7	46.7	3.9411			X
2	62.9	60.7	318.1	57.6	11.5	2.9038	X		X
2	60.8	58.4	344.2	54.1	13.9	2.9864	X	X	
3	68.7	65.8	275.6	63.3	7.0	2.7080	X	X	X
3	66.7	63.5	311.4	58.5	9.3	2.7961	X	X	X
4	70.8	67.0	277.1	63.1	6.7	2.6598	X	X	X
4	70.7	66.9	299.4	60.1	6.8	2.6636	X		X
5	73.4	68.9	283.9	62.2	5.8	2.5805	X	X	X
5	71.8	67.1	276.9	63.1	7.5	2.6562	X	X	X
6	74.4	69.1	284.5	62.1	6.6	2.5726	X	X	X
6	73.5	68.0	299.5	60.1	7.6	2.6178	X	X	X
7	75.0	68.7	299.4	60.1	8.0	2.5905	X	X	X

SAS OUTPUT

Number in Model	R-Square	C(p)	AIC	Variables in Model
1	0.3263	43.3217	99.1108	motheriq
1	0.2962	46.6934	100.6872	count
1	0.2758	48.9662	101.7123	read
1	0.1371	64.4782	108.0225	edutv

2	0.6291	11.4650	79.6217	motheriq read
2	0.6078	13.8557	81.6396	motheriq count
2	0.3780	39.5478	98.2396	motheriq speak
2	0.3672	40.7452	98.8542	fatheriq motheriq

3	0.6872	6.9673	75.4860	fatheriq motheriq read
3	0.6666	9.2811	77.7926	fatheriq motheriq count
3	0.6465	11.5240	79.8958	motheriq speak count
3	0.6449	11.7015	80.0571	motheriq speak read

4	0.7077	6.6799	75.0505	fatheriq motheriq speak count
4	0.7069	6.7730	75.1529	motheriq read edutv cartoons
4	0.7040	7.0929	75.5026	fatheriq motheriq speak read
4	0.7003	7.5083	75.9517	fatheriq motheriq read edutv

5	0.7338	5.7678	73.6904	fatheriq motheriq read edutv cartoons
5	0.7179	7.5400	75.7724	fatheriq motheriq speak count read
5	0.7160	7.7482	76.0091	motheriq speak read edutv cartoons
5	0.7142	7.9490	76.2362	fatheriq motheriq speak read edutv

6	0.7442	6.6012	74.2512	fatheriq motheriq speak read edutv cartoons
6	0.7351	7.6147	75.5049	fatheriq motheriq count read edutv cartoons
6	0.7333	7.8237	75.7580	fatheriq motheriq speak count edutv cartoons
6	0.7251	8.7375	76.8445	fatheriq motheriq speak count read edutv

7	0.7496	8.0000	75.4865	fatheriq motheriq speak count read edutv cartoons

R OUTPUT

```
> gifted$type
[1] 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 1 1
[39] 1 0 1 1 0 1 0 0 1 0 0 1 0 0 0 1 0 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1
[77] 1 1 0 1 0 1
Levels: 0 1
> m1=lm(rust~age+type+age:type,data=gifted)
> anova(m1)
Analysis of Variance Table

Response: rust
          Df Sum Sq Mean Sq F value    Pr(>F)
age         1  8305.1   8305.1  37.9007 3.013e-08 ***
type        1     1.5     1.5   0.0069   0.9340
age:type    1    60.2    60.2   0.2746   0.6018
Residuals  78 17091.9    219.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(m1)

Call:
lm(formula = rust ~ age + type + age:type, data = gifted)

Residuals:
    Min       1Q   Median       3Q      Max
-36.470  -9.132  -0.915   8.108  70.802

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.1977633   3.0299533    7.986 1e-11 ***
age          0.0008187   0.0002003    4.088 0.000105 ***
type1       -1.7757343   5.1844568   -0.343 0.732888
age:type1    0.0001556   0.0002969    0.524 0.601752
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.8 on 78 degrees of freedom
Multiple R-squared:  0.3286,    Adjusted R-squared:  0.3028
F-statistic: 12.73 on 3 and 78 DF,  p-value: 7.455e-07
```