**Question 1**

> psy = read.csv(choose.files(),header=TRUE)

> attach(psy)

> # part (a)

> model1 = lm (MRI~Gender+FSIQ+VIQ+PIQ+Weight+Height)

> summary(model1)

Call:

lm(formula = MRI ~ Gender + FSIQ + VIQ + PIQ + Weight + Height)

Residuals:

Min 1Q Median 3Q Max

-81324 -26361 -8034 20883 110071

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 164450.64 218056.88 0.754 0.4564

GenderMale 42368.74 24529.59 1.727 0.0941 .

FSIQ -9389.38 4651.64 -2.019 0.0523 .

VIQ 5388.76 2761.43 1.951 0.0601 .

PIQ 6287.51 2526.27 2.489 0.0184 \*

Weight 87.02 485.55 0.179 0.8589

Height 6883.32 3207.98 2.146 0.0398 \*

---

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

Residual standard error: 46760 on 31 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.6521, Adjusted R-squared: 0.5847

F-statistic: 9.684 on 6 and 31 DF, p-value: 5.117e-06

>

If someone‘s gender is Male, their MRI is 42368.74 higher than someone who is female , holding the other x variables constant.

># part (b)

> rstand = rstandard(model1)

> rstand[order(rstand)]

23 14 26 27 1 22

-1.85757235 -1.60697022 -1.45849096 -1.40239586 -0.99362626 -0.92638445

10 34 37 32 25 17

-0.74202080 -0.71998774 -0.69060653 -0.66405111 -0.65820067 -0.50183653

40 20 11 8 35 24

-0.47138535 -0.47024671 -0.46046100 -0.41800638 -0.35212024 -0.29970704

19 31 30 18 29 39

-0.18984614 -0.18266124 -0.16974008 -0.06419124 0.22552886 0.27027844

33 16 3 4 13 6

0.29793449 0.33714401 0.36610100 0.43091001 0.58763115 0.68572537

9 15 38 5 36 28

0.91969454 0.98882091 1.03073516 1.33745947 1.38625694 1.59623478

12 7

2.15253763 2.46881021

>

Observation 12 and 7 are outliers as they are > |2|

># part ©

> leverage = hatvalues(model1)

> leverage[order(leverage)]

7 15 18 5 36 1 31

0.09084357 0.09145738 0.09632154 0.09817127 0.10282631 0.10553738 0.10616428

11 38 23 4 27 40 30

0.11734245 0.11777923 0.12337840 0.12659539 0.13100495 0.13213810 0.13521482

35 6 33 26 34 28 17

0.14223438 0.14610916 0.14694700 0.14835618 0.17740084 0.18179366 0.18493743

16 19 20 39 37 24 12

0.18691410 0.18909928 0.19391029 0.21067015 0.21113841 0.21502278 0.22608661

29 8 22 25 10 9 14

0.23229684 0.23807204 0.24586393 0.25449739 0.26786624 0.28498836 0.29385518

32 3 13

0.31378618 0.34950630 0.38387220

> dim(psy)

[1] 40 7

> (3\*(7+1))/40

[1] 0.6

>

3 (k+1) / n = 3(7+1)/40 = 0.6

There is no points > 0.6 thus, there are no high leverages points.

> # part (d)

> model2 = lm (MRI~Gender+FSIQ+VIQ+PIQ+Weight+Height,subset=-c(12,7))

> summary(model2)

Call:

lm(formula = MRI ~ Gender + FSIQ + VIQ + PIQ + Weight + Height,

subset = -c(12, 7))

Residuals:

Min 1Q Median 3Q Max

-66740 -23461 -856 17091 74656

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 249919.6 182053.6 1.373 0.1803

GenderMale 48232.4 20355.4 2.370 0.0247 \*

FSIQ -5765.7 4034.4 -1.429 0.1636

VIQ 2752.1 2433.4 1.131 0.2673

PIQ 4545.1 2168.2 2.096 0.0449 \*

Weight 125.6 402.6 0.312 0.7573

Height 6555.1 2654.1 2.470 0.0196 \*

---

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

Residual standard error: 38500 on 29 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.7244, Adjusted R-squared: 0.6674

F-statistic: 12.71 on 6 and 29 DF, p-value: 5.386e-07

>

Answer on paper

Based on the test of goodness of fit statistic Model 2 is better than model 1 because the R^2, R^2 adjusted, standard error and the p-value of model 2 is better than model 1. So we prefer model 2.

> # part (e)

> cor(cbind(MRI,FSIQ,VIQ,PIQ,Weight,Height))

MRI FSIQ VIQ PIQ Weight Height

MRI 1.0000000 0.3576410 0.3374777 0.3868173 NA NA

FSIQ 0.3576410 1.0000000 0.9466388 0.9341251 NA NA

VIQ 0.3374777 0.9466388 1.0000000 0.7781351 NA NA

PIQ 0.3868173 0.9341251 0.7781351 1.0000000 NA NA

Weight NA NA NA NA 1 NA

Height NA NA NA NA NA 1

>

Most of the variables has a strong positive relationship with other x variables except MRI with FSCIQ , MRI with VIQ and MRI with PIQ.

># part (f)

> model3 = lm (MRI~Gender+Weight+Height)

> anova(model1,model3)

Analysis of Variance Table

Model 1: MRI ~ Gender + FSIQ + VIQ + PIQ + Weight + Height

Model 2: MRI ~ Gender + Weight + Height

Res.Df RSS Df Sum of Sq F Pr(>F)

1 31 6.7779e+10

2 34 1.0570e+11 -3 -3.7921e+10 5.7813 0.002924 \*\*

---

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

>

> qf(0.95,3,34)

[1] 2.882604

Hypothesis on paper

As F-statistic of 5.7813 is > F-statistic cutoff point which is 2.882604, so we reject the hypothesis. This means the FSIQ, VIQ and PIQ are jointly significant.

**Question 2**

> ability = read.csv(choose.files(),header=TRUE)

> attach(ability)

># part (a)

> model1 = lm (ECAB ~ EX+MET+GROW+YOUNG+OLD+factor(WEST))

> rstand = rstandard(model1)

> rstand[order(rstand)]

24 7 43 34 48 11

-2.26603694 -1.95324349 -1.22247551 -1.21199374 -1.19506390 -1.09855788

12 38 39 36 5 23

-1.07530753 -1.06503882 -0.85296108 -0.80260157 -0.74309975 -0.68031712

25 20 26 10 45 46

-0.57132375 -0.53424856 -0.51202966 -0.49090554 -0.39455583 -0.35615074

2 42 30 3 31 9

-0.27626321 -0.23053012 -0.21831230 -0.20247137 -0.07198533 -0.04394357

40 4 35 8 33 21

-0.02187347 0.05828816 0.16337626 0.17507364 0.20970100 0.25096611

6 29 37 32 27 13

0.28591296 0.36701076 0.38457760 0.39093600 0.40176538 0.42078004

22 19 16 17 28 44

0.51021688 0.54300608 0.68835161 0.72486622 0.78724974 0.83442359

41 14 18 15 1 47

0.83830766 0.83903396 1.21105017 1.41569901 1.43102633 4.88750812

>

Points 24 and 47 are outliers as the standard residuals > |2|. But point 47 has the highest outlier

> modelA = lm (ECAB ~ EX+MET+GROW+YOUNG+OLD+factor(WEST),subset=-c(47))

>part (b)

> ability1 = data.frame(WEST,EX,MET,GROW,YOUNG,OLD,ECAB)

> abline(ability1)

> pairs(ability1,upper.panel=NULL



> young2 = YOUNG\*YOUNG

> modelA1 = lm (ECAB ~ EX+MET+GROW+YOUNG+OLD+factor(WEST)+young2,subset=-c(47))

> summary(modelA)

Call:

lm(formula = ECAB ~ EX + MET + GROW + YOUNG + OLD + factor(WEST),

subset = -c(47))

Residuals:

Min 1Q Median 3Q Max

-18.3910 -5.4497 -0.1732 6.5224 17.7069

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 235.99198 57.11603 4.132 0.000178 \*\*\*

EX 0.14940 0.02932 5.096 8.71e-06 \*\*\*

MET 0.02828 0.07858 0.360 0.720801

GROW -0.17506 0.09362 -1.870 0.068828 .

YOUNG -5.37363 1.40182 -3.833 0.000438 \*\*\*

OLD -3.28409 1.42102 -2.311 0.026064 \*

factor(WEST)1 -0.72104 3.16455 -0.228 0.820925

---

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

Residual standard error: 8.413 on 40 degrees of freedom

Multiple R-squared: 0.7496, Adjusted R-squared: 0.712

F-statistic: 19.96 on 6 and 40 DF, p-value: 1.258e-10

> summary(modelA1)

Call:

lm(formula = ECAB ~ EX + MET + GROW + YOUNG + OLD + factor(WEST) +

young2, subset = -c(47))

Residuals:

Min 1Q Median 3Q Max

-15.2218 -5.7018 -0.8379 6.2107 16.6521

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.94851 216.71937 0.157 0.8763

EX 0.14888 0.02935 5.073 9.93e-06 \*\*\*

MET 0.04371 0.08025 0.545 0.5891

GROW -0.18398 0.09415 -1.954 0.0579 .

YOUNG 9.01509 14.95328 0.603 0.5501

OLD -3.44117 1.43144 -2.404 0.0211 \*

factor(WEST)1 -0.96590 3.17727 -0.304 0.7627

young2 -0.25326 0.26204 -0.967 0.3397

---

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

Residual standard error: 8.42 on 39 degrees of freedom

Multiple R-squared: 0.7555, Adjusted R-squared: 0.7116

F-statistic: 17.21 on 7 and 39 DF, p-value: 3.818e-10

Table in paper.

Model A is statistically better based on all the criteria except R^2. There for it is insignificant to make the quadratic term on Young.

> # part (d)

> step(modelA)

Start: AIC=206.62

ECAB ~ EX + MET + GROW + YOUNG + OLD + factor(WEST)

Df Sum of Sq RSS AIC

- factor(WEST) 1 3.67 2835.0 204.68

- MET 1 9.17 2840.5 204.77

<none> 2831.3 206.62

- GROW 1 247.50 3078.8 208.56

- OLD 1 378.06 3209.4 210.51

- YOUNG 1 1040.11 3871.4 219.33

- EX 1 1838.20 4669.5 228.14

Step: AIC=204.68

ECAB ~ EX + MET + GROW + YOUNG + OLD

Df Sum of Sq RSS AIC

- MET 1 8.32 2843.3 202.82

<none> 2835.0 204.68

- GROW 1 259.10 3094.1 206.79

- OLD 1 460.77 3295.8 209.76

- YOUNG 1 1358.16 4193.2 221.08

- EX 1 2274.18 5109.2 230.37

Step: AIC=202.82

ECAB ~ EX + GROW + YOUNG + OLD

Df Sum of Sq RSS AIC

<none> 2843.3 202.82

- GROW 1 250.8 3094.1 204.79

- OLD 1 706.9 3550.2 211.26

- EX 1 2536.9 5380.2 230.80

- YOUNG 1 3575.6 6418.9 239.09

Call:

lm(formula = ECAB ~ EX + GROW + YOUNG + OLD, subset = -c(47))

Coefficients:

(Intercept) EX GROW YOUNG OLD

255.0847 0.1429 -0.1716 -5.8412 -3.6284

> > summary(modelA)

Call:

lm(formula = ECAB ~ EX + MET + GROW + YOUNG + OLD + factor(WEST),

subset = -c(47))

Residuals:

Min 1Q Median 3Q Max

-18.3910 -5.4497 -0.1732 6.5224 17.7069

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 235.99198 57.11603 4.132 0.000178 \*\*\*

EX 0.14940 0.02932 5.096 8.71e-06 \*\*\*

MET 0.02828 0.07858 0.360 0.720801

GROW -0.17506 0.09362 -1.870 0.068828 .

YOUNG -5.37363 1.40182 -3.833 0.000438 \*\*\*

OLD -3.28409 1.42102 -2.311 0.026064 \*

factor(WEST)1 -0.72104 3.16455 -0.228 0.820925

---

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

Residual standard error: 8.413 on 40 degrees of freedom

Multiple R-squared: 0.7496, Adjusted R-squared: 0.712

F-statistic: 19.96 on 6 and 40 DF, p-value: 1.258e-10

> modelA2= lm (ECAB ~ EX + MET + GROW + YOUNG + OLD + factor(WEST),

+ subset = -c(47))

> modelA2= lm (ECAB ~ EX + MET + GROW + YOUNG + OLD, subset=-c(47))

> summary(modelA2)

Call:

lm(formula = ECAB ~ EX + MET + GROW + YOUNG + OLD, subset = -c(47))

Residuals:

Min 1Q Median 3Q Max

-18.1996 -5.5473 0.0557 6.5744 17.9205

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 241.65814 50.82010 4.755 2.46e-05 \*\*\*

EX 0.14623 0.02550 5.735 1.03e-06 \*\*\*

MET 0.02686 0.07742 0.347 0.7304

GROW -0.17772 0.09181 -1.936 0.0598 .

YOUNG -5.51416 1.24420 -4.432 6.82e-05 \*\*\*

OLD -3.39719 1.31602 -2.581 0.0135 \*

---

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

Residual standard error: 8.315 on 41 degrees of freedom

Multiple R-squared: 0.7493, Adjusted R-squared: 0.7187

F-statistic: 24.51 on 5 and 41 DF, p-value: 2.443e-11

> modelA3= lm (ECAB ~ EX+ GROW + YOUNG + OLD, subset=-c(47))

> summary(modelA3)

Call:

lm(formula = ECAB ~ EX + GROW + YOUNG + OLD, subset = -c(47))

Residuals:

Min 1Q Median 3Q Max

-17.9675 -5.8356 0.0426 6.5748 17.7473

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 255.08472 32.59720 7.825 9.83e-10 \*\*\*

EX 0.14287 0.02334 6.122 2.65e-07 \*\*\*

GROW -0.17162 0.08916 -1.925 0.0610 .

YOUNG -5.84117 0.80374 -7.268 6.04e-09 \*\*\*

OLD -3.62841 1.12287 -3.231 0.0024 \*\*

---

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

Residual standard error: 8.228 on 42 degrees of freedom

Multiple R-squared: 0.7485, Adjusted R-squared: 0.7246

F-statistic: 31.26 on 4 and 42 DF, p-value: 4.295e-12

>

From the stepwise regression and the backwards regression, we got both reduced model are the same which ECAB is determined by EX, GROW, YOUNG and OLD

> # part(e)

> OG = OLD\*GROW

> modelA3 = lm (ECAB ~ EX+MET+GROW+YOUNG+OLD+factor(WEST)+OG,subset=-c(47))

> summary(A3)

Error in summary(A3) : object 'A3' not found

> summary(modelA3)

Call:

lm(formula = ECAB ~ EX + MET + GROW + YOUNG + OLD + factor(WEST) +

OG, subset = -c(47))

Residuals:

Min 1Q Median 3Q Max

-18.299 -5.758 -1.313 5.838 17.608

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 228.79906 57.86396 3.954 0.000314 \*\*\*

EX 0.14853 0.02942 5.049 1.07e-05 \*\*\*

MET 0.01945 0.07945 0.245 0.807909

GROW 0.18887 0.42514 0.444 0.659313

YOUNG -5.38226 1.40590 -3.828 0.000456 \*\*\*

OLD -2.42232 1.73059 -1.400 0.169506

factor(WEST)1 -1.50935 3.29831 -0.458 0.649770

OG -0.03846 0.04382 -0.878 0.385480

---

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

Residual standard error: 8.438 on 39 degrees of freedom

Multiple R-squared: 0.7545, Adjusted R-squared: 0.7104

F-statistic: 17.12 on 7 and 39 DF, p-value: 4.123e-10

Table on paper

The interaction term between Old and Grow is not significant as the initial model A has better statistical characteristic compared to the new model except its R^2 is slightly lower than the new model.

**Question 3**

> US =read.csv(choose.files(),header=TRUE)

> attach(US)

> model1 = lm(MPG~WT)

> anova(model1)

Analysis of Variance Table

Response: MPG

Df Sum Sq Mean Sq F value Pr(>F)

WT 1 1333.49 1333.49 74.846 5.784e-11 \*\*\*

Residuals 43 766.11 17.82

---

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

>

> # part (a-i)

> summary(model1)

Call:

lm(formula = MPG ~ WT)

Residuals:

Min 1Q Median 3Q Max

-5.8488 -1.6113 -0.6239 0.3385 15.1512

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 80.6110 4.7688 16.904 < 2e-16 \*\*\*

WT -1.6250 0.1878 -8.651 5.78e-11 \*\*\*

---

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

Residual standard error: 4.221 on 43 degrees of freedom

Multiple R-squared: 0.6351, Adjusted R-squared: 0.6266

F-statistic: 74.85 on 1 and 43 DF, p-value: 5.784e-11

> # part (a-11)

> newdata= data.frame(WT=25)

Calculation on paper

The MPG is 39.9863 when WT is 25.

> predict(model1,newdata,interval="prediction",level=.95)

fit lwr upr

1 39.98639 31.37971 48.59306

> qt(0.975,43)

[1] 2.016692

>

The 95% confidence interval is between (31.37971,48.59306)

* Part (a-iv)

This means we are 95% confidence that the MPG has average value between (31.37971,48.59306).