```
> setwd("~/Desktop")
> #Problem 3
> #part(a)
> library("foreign")
> iq.data<-read.dta("kidiq.dta")
> ## part a
> library("foreign")
> iq.data <- read.dta("kidiq.dta")
> r <- lm(kid_score ~ mom_iq + mom_work + mom_age,iq.data)
> summary(r)

Call:
lm(formula = kid_score ~ mom_iq + mom_work + mom_age, data = iq.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.533 -12.786   2.011  12.111  47.695 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17.16064    9.11068   1.884   0.0603 .  
mom_iq       0.59928    0.05909  10.141   <2e-16 ***
mom_work     0.52736    0.75411   0.699   0.4847    
mom_age      0.35903    0.32904   1.091   0.2758    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.27 on 430 degrees of freedom
Multiple R-squared:  0.2045,	Adjusted R-squared:  0.1989 
F-statistic: 36.84 on 3 and 430 DF,  p-value: < 2.2e-16

> #part (b)
> #The F-statistic: 36.84 on 3, dof=430. This suggests that none of the coefficients 
are linearly
> #associated with the response, i.e. unable to predict the response.
> #Additionally, the low Multiple and Adjusted R-squared values suggest that there is 
a weak
> #correlation between the predictor variables mom.iq, mom_work, mom_age and the 
response.
> #However, the marginal p-value for mom.iq, 2e-16, is statistically significant at 
all levels
> #of significance.
> #This hints at the multicollinearity phenonemnon.
> 
> # part (c)
> r2 <- lm(kid_score ~ mom_iq,iq.data)
> summary(r2)

Call:
lm(formula = kid_score ~ mom_iq, data = iq.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.753 -12.074   2.217  11.710  47.691 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.79978    5.91741    4.36 1.63e-05 ***
mom_iq       0.60997    0.05852   10.42  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.27 on 432 degrees of freedom
Multiple R-squared:  0.201,	Adjusted R-squared:  0.1991
F-statistic: 108.6 on 1 and 432 DF,  p-value: < 2.2e-16

> #To test the hypothesis, we that coefficients for mom_work and mom_age are 0.
>
> anova(r2,r)
Analysis of Variance Table

Model 1: kid_score ~ mom_iq
Model 2: kid_score ~ mom_iq + mom_work + mom_age
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    432 144137
2    430 143502  2     635.1 0.9515  0.387
>
> #The ANOVA Table gives a p-value = 0.387, so we accept the alternative hypothesis
that at least
> #one of the predictor variables, mom_work or mom_age, from the first model has
predicitive power,
> #i.e. is statistically significant.
>
> #If we assume a simple linear model between mom_age and kid_score we can make
inferences
> #regarding the influence of mother's age on the childs test scores, i.e. how the
predictor
> #variable mom_age influences the response variable, kid_score.
>
> #The coefficient  for mom_age = 0.35. If we compare any two children whose mothers'
age at birth
> #differed by 1 year, it can be predicted that there will be an approximately 0.35
increase in
> #the test score.
> #This suggests that children born from older mothers do better on these exams, so
from this
> #analysis, it is tempting to advise mothers to have children at very old ages.
However, this is
> #obviously not a great recommendation because there are other factors to consider,
such as fertility
> #and birth defects at older ages. Therefore, this recommendation assumes that
children born
> #from older mothers do better on tests, no matter how old. It has already been
stated that this is
> #not necessarily the case.
> #If we furthmore assume that mom_age is the (or one of) variable with predictive
power
> #(one of the variables the hypothesis test picked up on), the recommendation is
valid.
>
> #part (d)
> r3 <- lm(kid_score ~ .,iq.data)
```

```
> summary(r3)

Call:
lm(formula = kid_score ~ ., data = iq.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.134 -12.624   2.293  11.250  50.206 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.82261    9.18765   2.266   0.0239 *  
mom_hs       5.56118    2.31345   2.404   0.0166 *  
mom_iq       0.56208    0.06077   9.249   <2e-16 ***
mom_work     0.13373    0.76763   0.174   0.8618    
mom_age      0.21986    0.33231   0.662   0.5086    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.17 on 429 degrees of freedom
Multiple R-squared:  0.215,	Adjusted R-squared:  0.2077 
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16

> 
> #With the predictor variable mom_hs added, the coefficients for mom_work and mom_age changed quite
> #a bit, while mom_iq remained relatively stable. The F-statistic is around the same value, though
> #a bit smaller: 29.38 on 4, dof=429. Therefore, we have the same results as in part (a), where only
> #mom_iq has a significant p-value.
> #Although mom_age coefficient decreased a bit, it remains positive, so our recommendation for
> #mothers to have children at an older age must remain with the same assumptions from part (b).
> 
> #part (e)
> par(mfrow = c(3, 2))
> plot(r3, which = c(1:6))
> summary(influence.measures(r3))
Potentially influential observations of
	 lm(formula = kid_score ~ ., data = iq.data) :

    dfb.1_ dfb.mm_h dfb.mm_q dfb.mm_w dfb.mm_g dffit   cov.r   cook.d hat  
7    0.07   0.02    -0.28    -0.09     0.15    -0.35_*  0.98    0.02   0.02
32   0.01   0.08     0.00    -0.10     0.00     0.16    0.96_*  0.01   0.00
72   0.00  -0.03     0.04    -0.02    -0.01     0.06    1.04_*  0.00   0.03
73   0.01   0.02    -0.02    -0.01     0.01    -0.03    1.04_*  0.00   0.03
87   0.06   0.14    -0.17    -0.22     0.11     0.33_*  0.96_*  0.02   0.02
96  -0.01  -0.02     0.00     0.01     0.02     0.03    1.04_*  0.00   0.03
111  0.23   0.26    -0.16     0.16    -0.28    -0.43_*  0.98    0.04   0.03
118 -0.06   0.01     0.09     0.08    -0.01     0.17    0.96_*  0.01   0.01
152  0.27   0.00    -0.19     0.23    -0.26    -0.40_*  0.98    0.03   0.03
213  0.08  -0.24     0.16    -0.03    -0.14     0.36_*  0.94_*  0.03   0.02
255  0.00   0.00     0.00     0.00     0.00     0.00    1.04_*  0.00   0.03
273  0.12  -0.05    -0.10     0.15    -0.11    -0.26    0.92_*  0.01   0.01
```

```
286 -0.06   0.23     0.08     0.04    -0.08    -0.33_*  0.93_*  0.02   0.01
307 -0.07  -0.05    -0.05    -0.11     0.16    -0.24    0.95_*  0.01   0.01
312  0.13  -0.04    -0.04     0.12    -0.17    -0.24    0.96_*  0.01   0.01
368 -0.11  -0.08     0.15    -0.10     0.05    -0.22    0.96_*  0.01   0.01
403  0.03   0.05     0.01     0.04    -0.07    -0.10    1.04_*  0.00   0.03
> #The Residual plot against the fitted values suggest a constant variance. There
doesn't appear to be
> #any systemic departure dependent on the fitted values. Thus, showing signs of
constant variance.
>
> #The Quantile-Quantile plot appears to be approximately linear with standardized
residuals
> #against the theoretical quantiles; however, the lower and upper tails of the QQ-
Plot are
> #a little skewed suggesting partial drift from normality. Overall, though, the
residuals are
> #approximately normally distributed.
>
> #In the Residual plot, it appears that there are at least 6 outliers: 3 on each side
of the line about y=0.
> #about y=0. Because it appears that the error distribution is approximately normal,
it is possible
> #that these potential outliers are in fact outliers of the data set.
>
> #The points 7, 87, 111, 152, 213, and 286, have asteriks next to them in the dffit
column
> #drawing attention to their high influence. Thus, these are potential influential
points.
> #The hat values, or leverage scores, in the hat column with higher values, those
with a
> #leverage score of 0.03 likely have an impact on the data. Here, 0.03 is a high hat
value since
> #our sample size is relatively large, n = 434.
>
> #part (f)
> r4 <- lm(formula = kid_score ~ . + mom_hs:mom_age, data = iq.data)
> summary(r4)

Call:
lm(formula = kid_score ~ . + mom_hs:mom_age, data = iq.data)

Residuals:
    Min      1Q  Median      3Q     Max
-53.686 -12.185   2.798  11.475  47.187

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     48.31617   16.64424   2.903  0.00389 **
mom_hs         -28.78386   17.51531  -1.643  0.10104
mom_iq           0.54820    0.06097   8.991  < 2e-16 ***
mom_work         0.13085    0.76504   0.171  0.86428
mom_age         -0.98928    0.69523  -1.423  0.15547
mom_hs:mom_age   1.56774    0.79256   1.978  0.04856 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 18.11 on 428 degrees of freedom
Multiple R-squared:  0.2222,     Adjusted R-squared:  0.2131
F-statistic: 24.45 on 5 and 428 DF,  p-value: < 2.2e-16

> #The resulting model from augmenting the data with an interaction between mom_age
and mom_hs is
> #kid_score = 48.32 - 28.78*mom_hs + 0.54*mom_iq + 0.13*mom_work - 0.99*mom_age
+1.57*mom_hs:mom_age
>
> #The model from part (d) was:
> #kid_score = 20.82 + 5.56*mom_hs +0.56*mom_iq + 0.13*mom_work + 0.22*mom_age
>
> #There are both differences and similarities between the models, which are actually
very interesting.
> #First, the intercepts for each model has changed quite significantly. While they
are both positive,
> #there is still a significant increased from model d to model f.
> #Even more interesting, adding the interaction between both mom_hs and mom_age into
the model caused
> #a significant decrease in their coefficients while the other variables remained
exactly constant!
> #This suggests that for children whose mother went to highschool, there is a
positive relationship
> #between mother's age at birth and the child's test score. The resulting model also
suggests that
> #there is a negative relationship between mother's age of child's birth and the
child's test
> #score. The addition of the interaction had no effect on the coefficient for
mom_work and very
> #little effect on mom_iq which suggests that the influence of mom_work and mom_iq
are independent
> #of the interaction between mom_hs and mom_age.
>
> anova(r3,r4)
Analysis of Variance Table

Model 1: kid_score ~ mom_hs + mom_iq + mom_work + mom_age
Model 2: kid_score ~ mom_hs + mom_iq + mom_work + mom_age + mom_hs:mom_age
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    429 141595
2    428 140312  1    1282.7 3.9128 0.04856 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #Our p-value from resulting from the F-statistic is 0.04856 where we are analyzing
on the
> #alpha = 0.05 significance level. Because the p-value < alpha, we reject the null
hypothesis
> #that the reduced model (in part (d)) is correct. In other words, we accept the
alternative
> #hypothesis that our full model, which includes the interaction variable
mom_hs*mom_age, is
> #correct. In particular, at least one of our new coefficients is non-zero. Because
there is only
> #one additional coefficient, that coefficient is significant. Hence, the coefficient
for
> #the interaction is statistically significant.
```