Extra Credit Problem

Econ 103

Regression with R

1. This question is based on the dataset on child test scores and mother characteristics. You can download the data with the following code:

```
data.url <- 'http://www.ditraglia.com/econ103/child_test_data.csv'
data <- read.csv(data.url)
head(data)

##   kid.score mom.hs    mom.iq mom.age
## 1        65      1 121.11753      27
## 2        98      1  89.36188      25
## 3        85      1 115.44316      27
## 4        83      1  99.44964      25
## 5       115      1  92.74571      27
## 6        98      0 107.90184      18

attach(data)
```

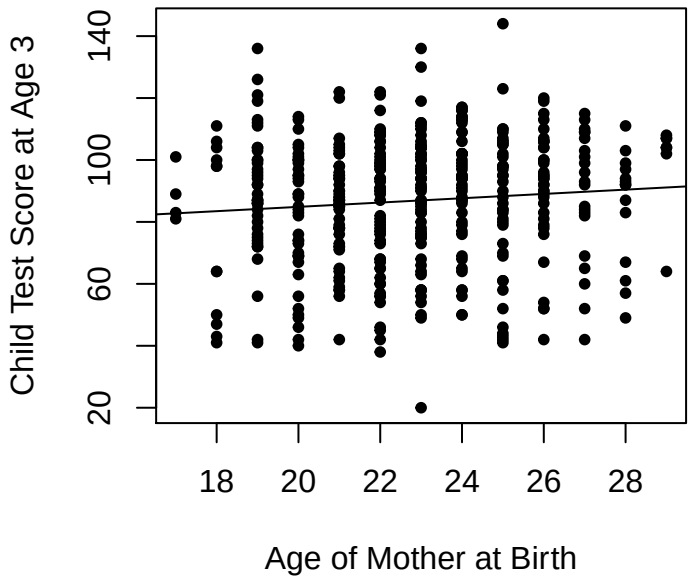
The columns contained in this dataset are as follows:

Variable Name	Description
`kid.score`	Child's Test Score at Age 3
`mom.age`	Age of Mother at Birth of Child
`mom.hs`	Mother Completed High School? (1 = Yes)
`mom.iq`	Mother's IQ Score

(a) Run a regression of `kid.score` on `mom.age`. Plot both the data and the fitted regression line, making sure to label the axes (Check R tutorial 3). Interpret the results.

Solution:

```
reg1 <- lm(kid.score ~ mom.age)
summary(reg1)
##
## Call:
## lm(formula = kid.score ~ mom.age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.946 -11.925   3.097  14.694  55.663
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.9569     8.3065   8.542 2.28e-16 ***
## mom.age       0.6952     0.3620   1.920   0.0555 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.35 on 432 degrees of freedom
## Multiple R-squared:  0.008464,Adjusted R-squared:  0.006168
## F-statistic: 3.688 on 1 and 432 DF,  p-value: 0.05548
plot(mom.age, kid.score, pch = 20, xlab = 'Age of Mother at Birth',
     ylab = 'Child Test Score at Age 3')
coefficients(reg1)
## (Intercept)     mom.age
##  70.9569209   0.6951862
intercept <- coef(reg1)[1]
slope <- coef(reg1)[2]
abline(a = intercept, b = slope)
```



Our model suggests that the children of mothers who were older when they gave birth tend to score higher. In particular, comparing two children whose mothers' age at birth differed by one year, we would predict that the child of the older mother will score, on average, 0.7 points higher. The standard error associated with the estimate, however, is fairly large. An approximate 95% CI would just barely include zero. Nevertheless, this result is suggestive that the children of older mothers do better on the test. This would seem to suggest that women should wait to have children until they are as old as possible. However, for this advice to truly be valid, it would have to be the case that being older when you give birth *caused* your child to have higher test scores. This seems unlikely, since there are many possible confounders here.

(b) Augment your model from part (a) by allowing a different intercept for children whose mother completed high school. Interpret your results and compare them to those you got in part (a).

Solution:

```
reg2 <- lm(kid.score ~ mom.hs + mom.age)
summary(reg2)
##
## Call:
## lm(formula = kid.score ~ mom.hs + mom.age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.980 -12.545   2.057  14.709  59.325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.4787     8.1068   8.694  < 2e-16 ***
## mom.hs       11.3112     2.3783   4.756  2.7e-06 ***
## mom.age       0.3261     0.3617   0.902    0.368
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.86 on 431 degrees of freedom
## Multiple R-squared:  0.05791,Adjusted R-squared:  0.05353
## F-statistic: 13.25 on 2 and 431 DF,  p-value: 2.614e-06
coef(reg2)
## (Intercept)      mom.hs     mom.age
##  70.4786610  11.3112315   0.3261332
```

By adding a dummy variable that equals one if a child's mother completed high school, we have controlled for one of the possible confounders from above: mother's level of education. We have done this by allowing the regression line to have a different intercept depending on mother's education. Comparing two children whose mothers are of the same age but only one whom attended high school, we predict that the child of the better educated mother will score, on average, 11.3 points higher. The standard error associated with this estimate is quite small, yielding a 95% CI that is is nowhere near zero. We have strong evidence of a large effect from mother's education level. In contrast, once we've controlled from mother's education, the estimated effect of `mom.age` falls substantially while the associated standard error stays the same. This results in an approximate 95% CI that includes many negative values. After controlling for

mother's education, there is much less evidence to suggest that older mothers have higher-scoring children.

(c) Now allow different slopes as well as intercepts for each group (those whose mother completed high school and those whose mother did not). Interpret your results.

Solution:

```
reg3 <- lm(kid.score ~ mom.hs + mom.age + mom.hs:mom.age)
summary(reg3)
##
## Call:
## lm(formula = kid.score ~ mom.hs + mom.age + mom.hs:mom.age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.535 -12.734   2.414  14.150  54.377
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    110.5417    16.4538   6.718 5.85e-11 ***
## mom.hs         -41.2875    18.9920  -2.174  0.03025 *
## mom.age         -1.5220     0.7532  -2.021  0.04391 *
## mom.hs:mom.age   2.3911     0.8567   2.791  0.00549 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.7 on 430 degrees of freedom
## Multiple R-squared:  0.07467,Adjusted R-squared:  0.06822
## F-statistic: 11.57 on 3 and 430 DF,  p-value: 2.64e-07
coef(reg3)
##    (Intercept)         mom.hs        mom.age mom.hs:mom.age
##     110.541718     -41.287465      -1.522014       2.391098
intercept.no.hs <- coef(reg3)[1]
intercept.hs <- coef(reg3)[1] + coef(reg3)[2]
slope.no.hs <- coef(reg3)[3]
slope.hs <- coef(reg3)[3] + coef(reg3)[4]
```

When we allow for different slopes as well as intercepts, by adding an *interaction* between `mom.hs` and `mom.hs`, namely `mom.hs:mom.age`, we find very different results depending on mother's education. (There is strong evidence that we should allow for different slopes, since the approximate 95% CI for the interaction does not include zero.) For children whose mothers attended high school, there is a *positive* relationship between mother's age at birth and child's test score. For children whose mothers did not attend high school, the relationship is *negative*. For children whose mothers were 18 then they gave birth, there is essentially *no* impact from mother's education level. As age of mother at birth increases, the impact of mother's education widens.

(d) Lastly, include mother's IQ as additional variable to the regression done in (c). Would you prefer to include mother's IQ instead of leaving it out of the regression? Compare regression results in (c) and (d).

Solution: Mother's IQ variable is highly statistically significant. Given everything else as fixed, we predict that the child of a mother whose IQ is 1 point higher will score, on average, 1.57 points higher. After controlling for mother's IQ, mother's high school dummy and mother's age variable are not statistically significant. However, the interaction between the high school dummy and age variable still matters in the regression. Given that the mother's IQ variable has explanatory power, I would prefer to include this variable in the regression. Also, comparing adjusted R-squared and residual standard error of two regressions, regression in (d) fits the data better.

```
reg4 <- lm(kid.score ~ mom.hs + mom.age + mom.hs:mom.age + mom.iq)
summary(reg4)
##
## Call:
## lm(formula = kid.score ~ mom.hs + mom.age + mom.hs:mom.age +
##     mom.iq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.838 -12.050   2.685  11.429  47.150
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     48.47926   16.59809   2.921  0.00368 **
## mom.hs         -28.70540   17.48948  -1.641  0.10147
## mom.age         -0.98470    0.69392  -1.419  0.15662
## mom.iq           0.54865    0.06085   9.017  < 2e-16 ***
## mom.hs:mom.age   1.56800    0.79166   1.981  0.04827 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.09 on 429 degrees of freedom
## Multiple R-squared:  0.2221,Adjusted R-squared:  0.2148
## F-statistic: 30.62 on 4 and 429 DF,  p-value: < 2.2e-16
```