**Problem 1**

> Reg<-lm(log(wage)~educ+exper+iq+meduc+feduc,data=x)
> summary(Reg)

Call:
`lm(formula = log(wage) ~ educ + exper + iq + meduc + feduc, data = x)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.94854</td>
<td>-0.22473</td>
<td>0.01827</td>
<td>0.26189</td>
<td>1.26116</td>
</tr>
</tbody>
</table>

Coefficients:

|                  | Estimate  | Std. Error | t value | Pr(>|t|)  |
|------------------|-----------|------------|---------|-----------|
| (Intercept)      | 5.158075  | 0.136448   | 37.803  | < 2e-16   |
| educ             | 0.049805  | 0.008513   | 5.850   | 7.46e-09  |
| exper            | 0.022801  | 0.003737   | 6.102   | 1.72e-09  |
| iq               | 0.004996  | 0.001161   | 4.304   | 1.91e-05  |
| meduc            | 0.008662  | 0.006276   | 1.380   | 0.168     |
| feduc            | 0.009663  | 0.005511   | 1.753   | 0.080     |

---

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

Residual standard error: 0.3813 on 716 degrees of freedom
(213 observations deleted due to missingness)
Multiple R-squared: 0.1792,  Adjusted R-squared: 0.1735
F-statistic: 31.26 on 5 and 716 DF,  p-value: < 2.2e-16

**a.** The $R^2$ for this regression is 0.1792. This means that 17.9% of variation in log(wage) is captured by the model.

**b.** The t-stat for iq is 4.304. Given a null hypothesis that the coefficient on iq is equal to zero, the probability of rejecting this hypothesis in favor of a two sided alternative and being wrong is 1.91x10-05. This is an extremely small probability of false rejection, much smaller of what is required to reject the null using a 95% level of confidence.

Confidence intervals and t-tests are also acceptable for this answer.

**c.** At the 99% level of confidence, the t critical value is 2.575. Thus, the confidence level is written as:

$0.022801-0.003737*2.575 < B_{exper} < 0.022801+0.003737*2.575$

$0.01317822 < B_{exper} < 0.03242378$

With 99% confidence, the effect of a one year increase in experience is to increase the wage between 1.3% and 3.2%.
d. This means that my estimate is far enough away to suggest that it is unlikely to be random. The probability that I’m wrong is the p-value. According to the results, this probability is 0.168. This is pretty low, but generally not low enough to conclude that meduc has an effect on the wage.

e. To do this, call theta=B_m-B_f. The hypothesis is that this is equal to zero. Writing B_m=theta+B_f, we substitute into our regression equation and rearrange:

\[
\text{low(wage)} = B_0 + B_{educ} \cdot \text{Educ} + B_{exp} \cdot \text{Exper} + B_{iq} \cdot \text{IQ} + B_m \cdot \text{MEduc} + B_f \cdot \text{FEduc} + u
\]

Running this regression, we get:

```r
> Reg <- lm(log(wage) ~ educ + exper + iq + meduc + I(meduc + feduc), data = x)
> summary(Reg)
```

Call:

```
lm(formula = log(wage) ~ educ + exper + iq + meduc + I(meduc + feduc), data = x)
```

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.94854</td>
<td>-0.22473</td>
<td>0.01827</td>
<td>0.26189</td>
<td>1.26116</td>
</tr>
</tbody>
</table>

Coefficients:

|              | Estimate | Std. Error |    t value |  Pr(>|t|) |
|--------------|----------|------------|------------|---------|
| (Intercept)  | 5.158075 | 0.136448   |  37.803    | < 2e-16 *** |
| educ         | 0.049805 | 0.008513   |   5.850    | 7.46e-09 *** |
| exper        | 0.022801 | 0.003737   |   6.102    | 1.72e-09 *** |
| iq           | 0.004996 | 0.001161   |   4.304    | 1.91e-05 *** |
| meduc        | -0.001001| 0.010163   |  -0.098    | 0.922   |
| I(meduc + feduc) | 0.009663 | 0.005511   |   1.753    | 0.080 . |

---

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

Residual standard error: 0.3813 on 716 degrees of freedom

(213 observations deleted due to missingness)

Multiple R-squared: 0.1792, Adjusted R-squared: 0.1735

F-statistic: 31.26 on 5 and 716 DF, p-value: < 2.2e-16

Evaluating the coefficient on meduc, which is the estimate of theta, we see that it is small and very insignificant. That is, if we reject the null that theta is zero in favor of it being non-zero, we would be
wrong 92% of the time. Not good! The effects of mother’s and father’s wages are insignificantly different from one another.

f. The new regression equation is written as:

\[ \text{wage} = B_0 + B_{\text{educ}} \times \text{Educ} + B_{\text{exp}} \times \text{Exper} + B_{\text{iq}} \times \text{IQ} \]

The prediction is written as:

\[ \theta = B_0 + B_{\text{educ}} \times 10 + B_{\text{exp}} \times 5 + B_{\text{iq}} \times 140 \]

Substituting the prediction into the regression and simplifying

\[ \text{wage} = B_0 + B_{\text{educ}} \times \text{Educ} + B_{\text{exp}} \times \text{Exper} + B_{\text{iq}} \times \text{IQ} \]

\[ \text{wage} = \theta - B_{\text{educ}} \times 10 - B_{\text{exp}} \times 5 - B_{\text{iq}} \times 140 + B_{\text{educ}} \times \text{Educ} + B_{\text{exp}} \times \text{Exper} + B_{\text{iq}} \times \text{IQ} \]

\[ \text{wage} = \theta + B_{\text{educ}} \times (\text{Educ} - 10) + B_{\text{exp}} \times (\text{Exper} - 5) + B_{\text{iq}} \times (\text{IQ} - 140) \]

Running this regression, we get:

```r
> Reg<-lm(wage~I(educ-10)+I(exper-5)+I(iq-140),data=x)
> summary(Reg)
```

Call:

```
  lm(formula = wage ~ I(educ - 10) + I(exper - 5) + I(iq - 140),
     data = x)
```

Residuals:

```
  Min      1Q  Median      3Q     Max
-922.10 -240.77  -44.92  191.56 2128.11
```

Coefficients:

```
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)   838.3487    61.3762  13.6599  < 2e-16 ***
I(educ - 10)   58.1038     7.0562   8.2345  6.07e-16 ***
I(exper - 5)   17.4171     3.1155   5.5904  2.98e-08 ***
I(iq - 140)     5.0688     0.9408   5.3879  9.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 370.8 on 931 degrees of freedom
Multiple R-squared: 0.162, Adjusted R-squared: 0.1593
F-statistic: 59.99 on 3 and 931 DF,  p-value: < 2.2e-16

A person with 10 years of education, 5 years of experience, and a 140 iq makes, on average, $838.3 per month. Given the 90% critical value of 1.64, the 90% confidence interval for this prediction is
838.3487-1.64*61.3762<theta<838.3487+1.64*61.3762

737.6917<theta<939.0057

Problem 2

> Reg_R<-lm(log(wage)~educ+exper,data=x)
> Reg_UR<-lm(log(wage)~educ+exper+sibs+brthord,data=x)
>
> summary(Reg_R)

Call:
  lm(formula = log(wage) ~ educ + exper, data = x)

Residuals:
  Min     1Q  Median     3Q    Max
-1.86915 -0.24001  0.03564  0.26132  1.30062

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.502710   0.112037  49.115  < 2e-16 ***
edu           0.077782   0.006577  11.827  < 2e-16 ***
exer          0.019777   0.003303   5.988 3.02e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.393 on 932 degrees of freedom
Multiple R-squared: 0.1309, Adjusted R-squared: 0.129
F-statistic: 70.16 on 2 and 932 DF,  p-value: < 2.2e-16

> sum(Reg_R$residual^2)
[1] 143.9786

Call:
  lm(formula = log(wage) ~ educ + exper + sibs + brthord, data = x)

Residuals:
  Min     1Q  Median     3Q    Max
-1.82525 -0.23290  0.03236  0.25516  1.27452

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.645358   0.124370  45.392  < 2e-16 ***
edu           0.073912   0.006986  10.580  < 2e-16 ***
exer          0.017663   0.003422   5.162 3.05e-07 ***
sibs         -0.007871   0.007274  -1.082    0.279
brthord  -0.013620  0.010353  -1.316  0.189  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3859 on 847 degrees of freedom
(83 observations deleted due to missingness)
Multiple R-squared: 0.1407,  Adjusted R-squared: 0.1367
F-statistic: 34.68 on 4 and 847 DF,  p-value: < 2.2e-16
> sum(Reg_UR$residual^2)
[1] 126.1572

a. Since the base model is nested within the model in ‘a’, we use an F-Test. The F-statistic is written as:

\[
\frac{(\text{sum}(\text{Reg}_R\text{residual}^2)-\text{sum}(\text{Reg}_UR\text{residual}^2))/2/\text{sum}(\text{Reg}_UR\text{residual}^2)/\text{Reg}_UR\text{df})}{\text{sum}(\text{Reg}_UR\text{residual}^2)/\text{Reg}_UR\text{df}}
\]

[1] 59.82516

This is a very large f-statistic; clearly larger than the f-critical value of 3.00 (for two restrictions). Thus, the model in ‘a’ is preferred to the base model.

b. These models are not nested, and thus one must rely on adjusted R^2 to compare models. Running the new model in ‘b’, we get:

> summary(lm(wage~educ+exper+sibs+brthord,data=x))

Call:
lm(formula = wage ~ educ + exper + sibs + brthord, data = x)

Residuals:
   Min     1Q Median     3Q    Max
-893.76 -245.51  -35.79  199.00 2141.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -132.332    120.715  -1.096    0.273
educ         72.203      6.781  10.648  < 2e-16 ***
exper        15.713      3.321   4.731 2.62e-06 ***
sibs         -8.681      7.060  -1.230    0.219
brthord      -12.839     10.049  -1.278    0.202
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 374.6 on 847 degrees of freedom
(83 observations deleted due to missingness)
Multiple R-squared: 0.1436,  Adjusted R-squared: 0.1396
F-statistic: 35.52 on 4 and 847 DF,  p-value: < 2.2e-16
In the new model, the adjusted $R^2$ is 0.1396. In the base model, the $R^2$ is 0.129. Thus, via this crude technique, we prefer the model in ‘b’.

c. Here, we require an interaction between education and iq. In doing so, it’s important to remember that one must include the interaction term, and non-interaction terms for both variables. If not, we will be introducing an omitted variables bias. Thus, the new estimating equation will be:

$$\text{low(wage)} = B_0 + B_{\text{educ}} \cdot \text{Educ} + B_{\text{exp}} \cdot \text{Exper} + B_{\text{iq}} \cdot \text{IQ} + B_{\text{educ_iq}} \cdot \text{Educ} \cdot \text{IQ}$$

Regressing:

```
> summary(lm(log(wage)~educ+exper+iq+I(educ*iq),data=x))
```

Call:

```
lm(formula = log(wage) ~ educ + exper + iq + I(educ * iq), data = x)
```

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.94016</td>
<td>-0.21346</td>
<td>0.02131</td>
<td>0.26097</td>
<td>1.25665</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 5.103e+00 | 5.713e-01 | 8.932 | < 2e-16 *** |
| educ | 6.441e-02 | 4.333e-02 | 1.487 | 0.137 |
| exper | 1.957e-02 | 3.257e-03 | 6.009 | 2.68e-09 *** |
| iq | 6.693e-03 | 5.397e-03 | 1.240 | 0.215 |
| I(educ * iq) | -6.903e-05 | 4.037e-04 | -0.171 | 0.864 |

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

Residual standard error: 0.3863 on 930 degrees of freedom
Multiple R-squared: 0.1623, Adjusted R-squared: 0.1587
F-statistic: 45.04 on 4 and 930 DF,  p-value: < 2.2e-16

To test whether the returns to education depend on iq, we evaluate the coefficient on the interaction term:

```
> summary(lm(log(wage)~educ+exper+iq+I(educ*iq),data=x))
```

```
Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| I(educ * iq) | -6.903e-05 | 4.037e-04 | -0.171 | 0.864 |

This interaction term is small and insignificantly different from zero. Precisely, if I reject the hypothesis that the interaction is insignificant, I have an 86.4% probability of being wrong. This is
not sufficient to reject the hypothesis using a standard 95% confidence level.

d. For this question, we introduce a squared term to the base specification. Precisely, we regress:

\[ \text{low(wage)} = B_0 + B_{\text{educ}} \times \text{Educ} + B_{\text{exp}} \times \text{Exper} + B_{\text{educ}^2} \times \text{educ}^2 \]

Regressing:

```r
> summary(lm(log(wage) ~ educ + exper + I(educ^2), data = x))
```

Call:
```
  lm(formula = log(wage) ~ educ + exper + I(educ^2), data = x)
```

Residuals:
```
     Min       1Q   Median       3Q      Max
-1.87004 -0.24490  0.02913  0.26936  1.28050
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.472985   0.553205   8.086 1.92e-15 ***
educ         0.225587   0.078042   2.891  0.00393 **
exper        0.020753   0.003338   6.218 7.60e-10 ***
I(educ^2)   -0.005221   0.002747  -1.901  0.05766 .
```

---

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

Residual standard error: 0.3925 on 931 degrees of freedom
Multiple R-squared: 0.1342,     Adjusted R-squared: 0.1314
F-statistic: 48.11 on 3 and 931 DF,  p-value: < 2.2e-16

First, note that the coefficient on the squared term in education is less than zero. This guarantees that there will be a level of education above which the returns to education fall. The question is now where this threshold lies, and whether it is relevant. To answer this question, note that the returns to education equal zero if:

\[ B_{\text{educ}} + 2 \times B_{\text{educ}^2} \times \text{educ} = 0 \]

Solving for the level of education such that this is the case:

\[
\text{educ} = -\frac{(B_{\text{educ}}/2 \times B_{\text{educ}^2})}{(2 \times -0.005221)} = 21.60381
\]

Thus, once you receive 21 years of education, the returns to education begin to fall. While the relevance of this is number is arguable, note the following summary information regarding education attainment in our sample.
> summary(x$educ)

     Min. 1st Qu.  Median     Mean 3rd Qu.    Max.
      9.00    12.00    12.00   13.47    16.00    18.00

> sqrt(var(x$educ))

[1] 2.196654

Thus, 21 is outside of our sample, and roughly 4-5 standard deviations above the mean. For most purposes, this threshold is likely irrelevant.