1. Estimate the model:

$$wage = \beta_0 + \beta_1 education + \beta_2 experience + \beta_3 male + \epsilon$$

What is the estimated return to an additional year of education?

```
> summary(lm(wage ~ education + experience + male))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -5.13303    2.02262  -2.538   0.0113 *  
education     3.03576    0.13544  22.415  < 2e-16 ***
experience    0.21019    0.04473   4.699 2.99e-06 ***
male        -12.39477    0.75894 -16.332  < 2e-16 ***
---
```

The estimated returns to education are $3.04. That is, an additional year of education is estimated to increase earnings by $3.04 per hour.

2. Using the same variables, estimate a log-lin model. What are the estimated returns to education?

```
> summary(lm(log(wage) ~ education + experience + male))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.927209   0.074628  25.824  < 2e-16 ***
education    0.110600   0.004997  22.133  < 2e-16 ***
experience   0.008980   0.001651   5.441 6.67e-08 ***
male        -0.406692   0.028002 -14.523  < 2e-16 ***
---
```

The coefficient of 0.110600 is interpreted as: for a 1 year increase in education, wage is estimated to increase by 11.06%. (This is the estimated returns to education).

3. Estimate a polynomial regression model, which allows for education to have a non-linear effect on wage. Determine the appropriate degree, *r*, for the polynomial regression model. Report the results of any relevant *t*-tests.

Include newly created variables into the regression model (it is now a polynomial regression model of degree $r = 4$):

```
> educ2 <- education ^ 2
> educ3 <- education ^ 3
> educ4 <- education ^ 4
> summary(lm(log(wage) ~ education + educ2 + educ3 + educ4 + experi
ence + male))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.608e+00  4.519e-01   5.772 1.05e-08 ***
education   -3.366e-02  2.258e-01  -0.149    0.882    
educ2        3.499e-03  3.696e-02   0.095    0.925    
educ3        6.395e-04  2.453e-03   0.261    0.794    
educ4       -2.815e-05  5.715e-05  -0.492    0.623    
experience   9.045e-03  1.645e-03   5.500 4.84e-08 ***
male        -4.090e-01  2.786e-02 -14.681  < 2e-16 ***
---
```

We fail to reject that `educ4` is statistically insignificant (notice that the p-value is 0.623), this suggests that `educ4` is not needed. We drop it from the model and re-estimate with $r = 3$:

```
> summary(lm(log(wage) ~ education + educ2 + educ3 + experience + m
ale))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.7904451  0.2590991  10.770  < 2e-16 ***
education   -0.1370906  0.0829224  -1.653   0.0986 .
educ2        0.0212451  0.0082321   2.581   0.0100 *
educ3       -0.0005622  0.0002504  -2.245   0.0250 *
experience   0.0090181  0.0016431   5.488 5.15e-08 ***
male        -0.4087601  0.0278430 -14.681  < 2e-16 ***
---
```

Now, we reject the null that `educ3` is statistically insignificant. It should not be dropped from the model. The appropriate degree of polynomial is $r = 3$.

4. Building on your model from question 3, estimate a model that allows education to have a different effect on wages, depending on whether the worker is male or female.

In order to allow for education to have a different effect depending on gender, we must create some interaction terms and estimate the model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 educ^2 + \beta_3 educ^3 + \beta_4 male \times educ + \beta_5 male \times educ^2 + \beta_6 male \times educ^3 + \beta_7 experience + \beta_8 male + \epsilon \qquad (4.1)$$

We create the three new variables by multiplying `male` by all instances of the "education" variable:

```
> male_educ <- male * education
> male_educ2 <- male * educ2
> male_educ3 <- male * educ3
```

Now, we include all of these interaction terms in our regression:

```
> summary(lm(log(wage) ~ education + educ2 + educ3 + male_educ + ma
le_educ2 + male_educ3 + experience + male))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.4995757  0.4178220   5.982 3.07e-09 ***
education   -0.1745586  0.1313903  -1.329   0.1843    
educ2        0.0299745  0.0127863   2.344   0.0193 *  
educ3       -0.0008696  0.0003833  -2.269   0.0235 *  
male_educ    0.0955622  0.1681905   0.568   0.5700    
male_educ2  -0.0179645  0.0166043  -1.082   0.2796    
male_educ3   0.0006089  0.0005030   1.210   0.2264    
experience   0.0090942  0.0016205   5.612 2.60e-08 ***
male         0.0043265  0.5230657   0.008   0.9934    
---
```

5. Using your models from question 3 and 4, test the hypothesis that the returns to education do not depend on gender. Report any relevant test results.

The appropriate null hypothesis for this question is:

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0$$

This is a multiple hypothesis (it involves multiple *betas*), and we should use an *F*-test. This null hypothesis suggests a *restricted* model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 educ^2 + \beta_3 educ^3 + \beta_7 experience + \beta_8 male + \epsilon \qquad (5.1)$$

This restricted model has been obtained by substituting the values for the *betas* in the null hypothesis ($\beta_4 = \beta_5 = \beta_6 = 0$) into equation (4.1). Note that this model has already been estimated in Question 3.

The null hypothesis may now be tested by comparing model 4.1 (as the *unrestricted* model) to model (5.1) (as the *restricted* model). A version of the *F*-test statistic formula (available on your formula sheet) is:

$$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(n - k_U - 1)}$$

The number of restrictions is 3, so that $q = 3$. The number of *betas* in the unrestricted model (4.1) is 8, so that $k_U = 8$. The sample size is 1000, so that $n = 1000$. The (unadjusted) R-square from the unrestricted model is $R_U^2 = 0.4422$. The R-square from the restricted model is $R_R^2 = 0.4245$. Substituting thes values into the *F*-statistic formula we get:

$$F = \frac{(0.4422 - 0.4245)/3}{(1 - 0.4422)/(1000 - 8 - 1)} = 9.8333$$

Comparing this *F*-statistic of 9.83 to the 5% critical value of 2.60 (see Table 7.1 on page 94 of the text book) we reject the null hypothesis that there is no difference in the effect of education on earnings for men and for women. Note that the *t*-statistics on the *betas* involved in the null (0.568, -1.082, 1.210) tell quite a different story (they suggest we should fail to reject).

6. Use your model from question 4. What is the estimated difference in the returns to education, between men and women?

To interpret the effect of changes in education on wage, we need to consider different starting values for education (since we have estimated a polynomial regression model). Let's begin by obtaining the predicted effect of a 1 year increase for males, with 12 years of education:

$$\widehat{wage}|_{educ=13,male=1} - \widehat{wage}|_{educ=12,male=1} = -\,0.1745586(13) + 0.0299745(13^2) - 0.0008696(13^3) + 0.0955622(13) - 0.0179645(13^2) + 0.0006089(13^3) + 0.1745586(12) - 0.0299745(12^2) + 0.0008696(12^3) - 0.0955622(12) + 0.0179645(12^2) - 0.0006089(12^3) = 0.0989853$$

Now, we get the same predicted effect, but for women:

$$\widehat{wage}|_{educ=13,male=0} - \widehat{wage}|_{educ=12,male=0} = -0.1745586(13) + 0.0299745(13^2) - 0.0008696(13^3) + 0.1745586(12) - 0.0299745(12^2) + 0.0008696(12^3) = 0.1669615$$

So, we see that the estimated difference in the effect of an extra year of education for men and women is an extra 0.1669615 - 0.0989853 = \$0.07 / hour for women. However, since this is a polynomial regression model, the effect of an extra year of education depends on the starting value for education. For example:

$$\widehat{wage}|_{educ=9,male=1} - \widehat{wage}|_{educ=8,male=1} = 0.0686017$$

$$\widehat{wage}|_{educ=9,male=0} - \widehat{wage}|_{educ=8,male=0} = 0.1463047$$