Task 1. The following model allows the return to education to depend upon the total amount of both parents' education, called pareduc:

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \times pareduc + \beta_3 exper + \beta_4 tenure + u$$

(1) Show that, in decimal form, the return to another year of education in this model is

$$\frac{\Delta \ln(wage)}{\Delta educ} = \beta_1 + \beta_2 pareduc$$

What sign do you expect for β_2? Why?

(2) Using the data in `wage2.csv`, interpret the coefficient on the interaction term. It might help to choose two specific values for pareduc – for example, $pareduc = 32$ if both parents have a college education, or $pareduc = 24$ if both parents have a high school education – and to compare the estimated return to *educ*.

(3) Does the estimated return to education now depend positively on parent education? Test the null hypothesis that the return to education does not depend on parent education.

Solution:

(1)

$$\ln(wage) + \Delta \ln(wage) = \beta_0 + \beta_1 (educ + \Delta educ) + \beta_2 (educ + \Delta educ) \times pareduc + \beta_3 exper + \beta_4 tenure + u$$

$$\Delta \ln(wage) = \beta_1 \Delta educ + \beta_2 \Delta educ \times pareduc \implies$$

$$\frac{\Delta \ln(wage)}{\Delta educ} = \beta_1 + \beta_2 pareduc$$

I expect $\hat{\beta}_2$ to be positive since it is natural that weekly earnings increases faster with each additional year of education, for those, whose parents studied more.

(2) Lets compare effect of education, assuming that parents have college education versus parents having high school education:

$$\frac{\Delta \ln(wage)}{\Delta educ} = 0.00078 \times (32 - 24) = 0.00624 = 0.62\%$$

Interpretation says the following: each additional year of education is associated with 0.62% higher earnings if parents get college degree versus parents get only higher school degree.

(3) After estimation we got the following regression:

$$\ln(wage) = \underset{(0.38)}{4.94} + \underset{(0.027)}{0.097} educ + \underset{(0.017)}{0.033} pareduc + \underset{(0.0012)}{0.0016} educ \times pareduc + \underset{(0.004)}{0.020} exper + \underset{(0.003)}{0.010} tenure$$

Return on education will not depend on parents' education as long as $\hat{\beta}_2$ is not statisticallly different from zero. We can use t-statiscs for this coefficient which is:

$$t_{\hat{\beta}_2} = \frac{0.0016}{0.0012} = 1.33 < 1.96 \implies \beta_2 \text{ is insignificant}$$

Thus estimated return on education doesnt depend on parents' education.

Task 2. A model to explain the standardized outcome on a final exam (*stndfnl*) in terms of percentage of classes attended (*atndrte*), prior college grade point average (*priGPA*), and American College Testing score (*ACT*) is

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 GPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \times atndrte + u$$

Using the 680 observations in `attend.csv`, for students in microeconomic principles, the esti- mated equation is

$$\begin{aligned} stndfnl = 2.05 - 0.0067 atndrte - 1.63 priGPA - 0.128 ACT + \\ 0.296 priGPA^2 + 0.0045 ACT^2 + 0.0056 priGPA \times atndrte \end{aligned}$$

When $atndrte^2$ and *ACT* times atndrte are added to this equation, the R-squared becomes 0.232. Assuming the error term in the population regression to be homoscedastic, are these additional terms jointly significant at the 10% level? Would you include them in the model?

Solution: After using linear hypothesis testing, namely after testing that coefficients before both *atndrte2* and *atndrte_x_ACR* equal to zero we obtained $F = 1.2543$ and p-value of 0.2859 which implies that at 10% significance level those terms are not jointly significant. I would not include them in the model, since the adjusted r-squared metrics has lowered.

Task 3. Consider the standard linear multivariate regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

The OLS estimation is performed. The sample means of x_1 and x_2 are zero. The standard error of the regression is 3.0. The standard error for β_1 is 2.0; and the standard error for β_2 is 1.0. The number of observations is 100. Assume that the error term in the population regression is homoscedastic.

(1) Is it possible that the sample variance of x_1 is larger than the sample variance of x_2? Discuss.
(2) What is the smallest possible value of R^2?
(3) What is the smallest possible value of the standard error for $\hat{\beta}_1 + \hat{\beta}_2$?
(4) Define $x_i = (1, x_{1i}, x_{2i})'$. Suppose that $\sum_{i=1}^n x_i x_i'$ is diagonal. Find the diagonal of the matrix $\sum_{i=1}^n x_i x_i'$

Solution:

(1) It is possible if the following holds:

$$\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 > \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2$$
$$\sum_{i=1}^n x_{1i}^2 - 2\bar{x}_1 \sum_{i=1}^n x_{1i} + \bar{x}_1^2 > \sum_{i=1}^n x_{2i}^2 - 2\bar{x}_2 \sum_{i=1}^n x_{2i} + \bar{x}_2^2$$
$$\sum_{i=1}^n x_{1i}^2 > \sum_{i=1}^n x_{2i}^2$$

Lets remember that:

$$\hat{\beta}^{\text{OLS}} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u$$

$$\text{Var}(\hat{\beta}|X) = (X'X)^{-1}X'\text{Var}(u|X)((X'X)^{-1}X')' = (X'X)^{-1}X'\text{Var}(u|X)X(X'X)^{-1} =$$

$$= \sigma_u^2 \left(X'X\right)^{-1} = \sigma_u^2 \begin{pmatrix} \sum_{i=1}^n 1^2 & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{pmatrix}^{-1}$$

Then it follows that:

$$\text{Var}(\hat{\beta}_1|X) = \frac{\sigma_u^2}{n}\frac{1}{\det(X'X)}\det\begin{bmatrix} 1 & \frac{1}{n}\sum_{i=1}^n x_{2i} \\ \frac{1}{n}\sum_{i=1}^n x_{2i} & \frac{1}{n}\sum_{i=1}^n x_{2i}^2 \end{bmatrix} = \frac{\sigma_u^2}{n^2}\frac{\sum_{i=1}^n x_{2i}^2}{\det(X'X)} = 4$$

and

$$\text{Var}(\hat{\beta}_2|X) = \frac{\sigma_u^2}{n}\frac{1}{\det(X'X)}\det\begin{bmatrix} 1 & \frac{1}{n}\sum_{i=1}^n x_{1i} \\ \frac{1}{n}\sum_{i=1}^n x_{1i} & \frac{1}{n}\sum_{i=1}^n x_{1i}^2 \end{bmatrix} = \frac{\sigma_u^2}{n^2}\frac{\sum_{i=1}^n x_{1i}^2}{\det(X'X)} = 1$$

Thus

$$\frac{\text{Var}(\hat{\beta}_2|X)}{\text{Var}(\hat{\beta}_1|X)} = \frac{\sum_{i=1}^n x_{1i}^2}{\sum_{i=1}^n x_{2i}^2} = \frac{4}{1}$$

and consequently sample variance of x_1 is higher than sample variance of x_2.

(2) By definition:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

(3) $\text{Var}(\hat{\beta}_1 + \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) + \text{Var}(\hat{\beta}_2) = 5 + \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$

From the first point it follows that:

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\sigma_u^2}{n}\frac{1}{\det(X'X)}\det\begin{bmatrix} 1 & \frac{1}{n}\sum_{i=1}^n x_{1i} \\ \frac{1}{n}\sum_{i=1}^n x_{2i} & \frac{1}{n}\sum_{i=1}^n x_{2i}x_{1i} \end{bmatrix} = \frac{\sigma_u^2}{n^2}\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\det(X'X)}$$

the latter term can be rewritten as:

$$\frac{\sigma_u^2}{n^2}\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\det(X'X)} = \frac{\sigma_u^2}{n^2}\frac{\sum_{i=1}^n x_{1i}x_{2i}}{n(\sum_{i=1}^n x_{1i}^2 \sum_{i=1}^n x_{2i}^2 - (\sum_{i=1}^n x_{1i}x_{2i})^2)}$$

Task 4. Use the housing price data in `hprice1.csv` for this exercise.

(1) Estimate the model

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

where price is the price of a house in thousands of U.S. dollars, *lotsize* is the size of a lot in square feet, *sqrft* is the size of a house in square feet, *bdrms* is the number of bedrooms. Report the results in the usual form, including the standard error of the regression. Obtain predicted price, when we plug in $lotsize = 10{,}000$, $sqrft = 2{,}300$, $bdrms = 4$; round this price to the nearest dollar.

(2) Run a regression that allows you to put a 95% confidence interval around the predicted value in part (i). Note that your prediction will differ somewhat due to rounding error.

(3) Let $price^0$ be the unknown future selling price of the house with the characteristics used in parts (i) and (ii). Find a 95% CI for $price^0$ and comment on the width of this confidence interval.

TABLE 1. Task 4 regression

	Dependent variable:
	price
lotsize	0.002
	(0.001)
sqrft	0.123***
	(0.018)
bdrms	13.853
	(8.479)
Constant	−21.770
	(37.138)
Note:	*p<0.1; **p<0.05; ***p<0.01 ; $\hat{\sigma}_u = 59.83$

Solution:

(1) After performing OLS estimation assuming heteroscedastic standard errors we got the following results: Prediction for $lotsize = 10000$, $sqrft = 2300$, $bdrms = 4$ is \$337

(2) We will use the following:

$$\beta_0 + 10000\beta_1 + 2300\beta_2 + 4\beta_3 = 337 \implies \beta_0 = 337 - 10000\beta_1 - 2300\beta_2 - 4\beta_3$$

and estimate the regression:

$$price = \underbrace{337}_{\tilde{\beta}_0} + \beta_1(lotsize - 10000) + \beta_2(sqrft - 2300) + \beta_3(bdrms - 4) + u$$

where $\tilde{\beta}_0$ will represent the predicted value of interest. Estimated value for intercept is \$336.7 $\sim$ \$337

(3) Since standard deviation of intercept is ~ 7.4 we get that estimated house price is within $\pm 1.96 \times 7.4$ interval, or more precisely:

TABLE 2. Confidence interval for value of interest

	2.5 %	97.5 %
(Intercept)	322.042	351.372

Task 5. Estimate a regression of $\ln(ahe)$ on *yrseduc*, *female*, and $female \times yrseduc$. (Note: you will need to create a new variable for the interaction term $female \times yrseduc$)

(1) Explain the meaning of the coefficients on yrseduc and $female \times yrseduc$.

(2) Plot, on the same graph, the two estimated regression lines describing the relation between $\ln(ahe)$ and *yrseduc* for men and for women. (You may do this in R, in EXCEL, by hand using graph paper, or by any other method.)
(3) Test (at the 5% significance level) the null hypothesis that the value of an additional year of school is the same for men and women, against the two-sided alternative that it differs.
(4) Test (at the 5% significance level) the null hypothesis that the regression line is the same for men and women, that is, that the population slope and intercept for women are the same as the population slope and intercept for men.
(5) What do you conclude about any differences between men and women in the value (as measured by log hourly earnings) of an additional year of school? Are these differences, if any, significant or important in a real-world sense?

Solution:

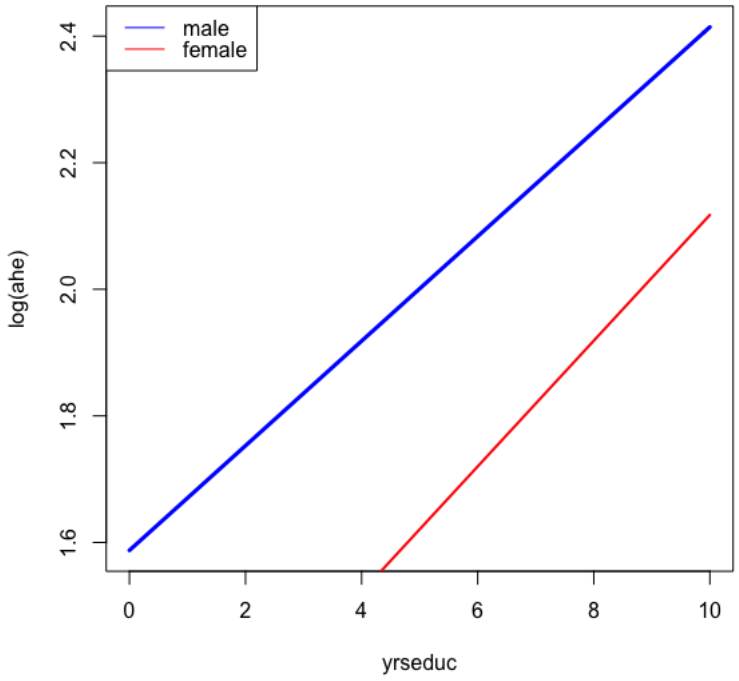
(1) After estimating the model, we obtained following results:

TABLE 3. Task 5 regression

	Dependent variable:
	log(ahe)
yrseduc	0.083^{***}
	(0.001)
female	-0.464^{***}
	(0.027)
female_x_yrseduc	0.017^{***}
	(0.002)
Constant	1.588^{***}
	(0.017)
Observations	37,810
R^2	0.218
Adjusted R^2	0.218
Residual Std. Error	0.475 (df = 37806)
F Statistic	$3{,}508.517^{***}$ (df = 3; 37806)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that coefficients with *yrseduc* are positive and significant. That means that each additional year of education was associated with $8.3\% + 1.7\% = 10\%$ increase of earnings for women and 8.3% increase for non-women.

(2) Chart is below (sorry latex skills dont allow me to put picture in there)
(3) We can see that coefficient on interaction term *female_x_yrseduc* is significant, thus additional year of school differs for men and women



(4) Once again, the model was:

$$\ln(ahe) = \beta_0 + \beta_1 yrseduc + \beta_2 female + \beta_3 female \times yrseduc + u$$

Regression lines for men and women will be the same if the following holds:

$$\begin{cases} \beta_0 + \beta_2 = \beta_0 \\ \beta_1 + \beta_3 = \beta_1 \end{cases} \implies \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \end{cases}$$

We will use F statistics for testing restricted model. Results are: $F = 1192.8$ and p-value is zero, thus regression lines differ for men and women.

(5) So each additional year of education was associated with earnings by 1.7% higher for women than for men. This could be the consequence of women being more diligent and focused during the studies. This increase can hardly be named significant in a real-world sense, since it is only 10% of education contribution.

Task 6. Create the variable *male* that $= 1$ if the worker is male, and $= 0$ if female, then create the interaction term $male \times yrseduc$. Estimate a regression of $\ln(ahe)$ on $yrseduc$, $female$, $female \times yrseduc$, and $male \times yrseduc$. What happens? Why?

Solution: One interaction term has been omitted by R because together with another interaction term thet created perfect multicollinearity. Namely:

$$yrseduc = male \times yrseduc + female \times yrseduc$$

and R had to drop on of the factors in order to obtain regression estimates, which depends on rank of the data matrix. That dependency arises because $\hat{\beta} = (X'X)^{-1}X'y$ and in order to obtain $(X'X)^{-1}$ matrix X should have non-zero determinant (i.e no perfect multicollinearity).

Task 7. Norms about women working have evolved greatly over the past thirty years. Women who are 60 years old in the 1999 CPS started their working lives in very different circumstances than women who are 30 years old in the 1999 CPS. Might it be, then, that the gender gap and the gender difference in the value of an additional year of school depend on the age (that is, generation) of the worker? To find out, estimate two regressions: (i) ln(*ahe*) on *yrseduc*, *female*, $female \times yrseduc$, *age*, age^2, and age^3; and (ii) ln(*ahe*) on *yrseduc*, *female*, $female \times yrseduc$, *age*, age^2, age^3, $female \times age$, $female \times age^2$, and $female \times age^3$.

(1) Does allowing for the additional interaction in specification (ii) make any difference, in a real-world sense, to your estimate of the value of an additional year of education for men? for women?
(2) Based on the results for specification (ii), estimate the gender gap in earnings for 25 year old workers with 16 years of education.
(3) Based on the results for specification (ii), estimate the gender gap in earnings for 55 year old workers with 16 years of education.
(4) Test (at the 5% significance level) the hypothesis that the population gender gap in earnings does not depend on age.
(5) In words, is there evidence that the gender gap is less for younger than older workers?

Solution:

(1) Results of estimation are presented in tables 4 and 5. We can see that allowing for additional factors, they are independently insignificant. In a real-world sense those variables dont make much sense, since in general everybody receives education at the same age and not a lot of people are trying to pursue higher education in their mid-ages.
(2) Let gender gap be calculated as follows:

$$\text{gap} = \ln(ahe|female = 1) - \ln(ahe|female = 0)$$

Using predict function in R we obtain that difference in log earnings is 0.084
(3) Using predict function in R we obtain that difference in log earnings is 0.226.

TABLE 4. (I) specification

	Dependent variable:
	log(ahe)
yrseduc	0.081^{***}
	(0.001)
female	-0.519^{***}
	(0.027)
age	0.103^{***}
	(0.013)
age2	-0.002^{***}
	(0.0003)
age3	0.00001^{***}
	(0.00000)
yrseduc:female	0.021^{***}
	(0.002)
Constant	-0.230
	(0.175)
Observations	37,810
R^2	0.245
Adjusted R^2	0.245
Residual Std. Error	0.467 (df = 37803)
F Statistic	$2{,}041.220^{***}$ (df = 6; 37803)
Note:	*p<0.1; **p<0.05; ***p<0.01

TABLE 5. (II) specification

	Dependent variable:
	log(ahe)
yrseduc	0.080^{***}
	(0.001)
female	0.042
	(0.353)
age	0.113^{***}
	(0.017)
age2	-0.002^{***}
	(0.0004)
age3	0.00001^{***}
	(0.00000)
yrseduc:female	0.020^{***}
	(0.002)
female:age	-0.025
	(0.026)
female:age2	0.0003
	(0.001)
female:age3	-0.00000
	(0.00000)
Constant	-0.452^{**}
	(0.230)
Observations	37,810
R^2	0.246
Adjusted R^2	0.246
Residual Std. Error	0.467 (df = 37800)
F Statistic	$1{,}371.828^{***}$ (df = 9; 37800)
Note:	*p<0.1; **p<0.05; ***p<0.01