Question 1.

Effects of Physical Attractiveness on Wage

Hamermesh and Biddle (1994) used measures of physical attractiveness in a wage equation. Each person in the sample was ranked by an interviewer for physical attractiveness using five categories (homely, quite plain, average, good looking, and strikingly beautiful or handsome). Because there are so few people at the two extremes, the authors put people into one of three groups for the regression analysis: "average", "below average", and "above average", where **the base or reference group is "average"**. Using data from the 1977 Quality of Employment Survey, after controlling for the usual productivity characteristics, the following two regressions were estimated using data on n = 1,260:

Estimate the model (1.1) reports in the Table 1.1

$$\log(wage_i) = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 expersq_i + \beta_5 union_i + \beta_6 female_i + u_i \qquad \textbf{(1.1)}$$

Table 1.1

Source	SS	df	MS	Number of obs	=	1,260
				F(5, 1254)	=	149.25
Model	166.011417	5	33.2022834	Prob > F	=	0.0000
Residual	278.96855	1,254	.222462959	R-squared	=	0.3731
				Adj R-squared	=	0.3706
Total	444.979967	1,259	.353439211	Root MSE	=	.47166

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]
educ	.0708503	.0052325			
exper	.0389808	.0043524			
expersq	-.0005986	.0000975			
union	.1924593	.0301994			
female	-.4421609	.0289766			
_cons	.443611	.078859			

Omitted for the purpose of this exam

Estimate the model (1.2) reports in the Table 1.2

$$\log(wage_i) = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 expersq_i + \beta_5 union_i + \beta_6 female_i + \beta_7 belavg_i + \beta_8 abvavg_i + u_i \qquad \textbf{(1.2)}$$

where $\log(wage_i)$ or $lwage$ = logarithm of hourly wage (in USD)

$educ_i$ = years of schooling

$exper_i$ = years of workforce experience

$expersq_i$ = years of workforce experience squared

$union_i$ = 1 if union member

$female_i$ = 1 if female

$belavg_i$ = 1 if in below average physical attractiveness

$abvavg_i$ = 1 if in above average physical attractiveness

Table 1.2

Source	SS	df	MS	Number of obs	=	1,260
				F(7, 1252)	=	109.21
Model	168.697151	7	24.099593	Prob > F	=	0.0000
Residual	276.282816	1,252	.220673176	R-squared	=	0.3791
				Adj R-squared	=	0.3756
Total	444.979967	1,259	.353439211	Root MSE	=	.46976

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]
educ	.0691306	.00525			
exper	.0395785	.0043428			
expersq	-.0006081	.0000971			
union	.1884632	.0301843			
female	-.4388235	.028877			
belavg	-.1388291	.0417749			
abvavg	.0070104	.0302809			
_cons	.4737302	.0795614			

Omitted for the purpose of this exam

Answer the following questions.

1.a) Based on the regression results provided, write out the estimated coefficients in the form of regression equation (1.1). Interpret the estimated coefficients associated with $educ_i$. Based on Model (1.1), test whether education has an impact on logarithm of hourly wage. Show your work. (Use α = 0.05)

- First, write the estimated coefficients in linear form, using only 4 decimal points.

$$\log(\widehat{wage}_i) = 0.4436 + 0.0709educ_i + 0.039exper_i - 0.0006expersq_i + 0.1925union_i - 0.4422female_i$$

- Then, test the education parameter.

$t_{cal} = \frac{0.0709-0}{0.0052} = 13.6346$. The critical value when the degrees of freedom is 1,260-6 = 1,254 is ± 1.96. Calculated stat exceeds the critical value which means that we can reject the null hypothesis. In other words, we can make sure that education parameter is not zero 95 percent.

1.b) What is the overall significance of the regression from Model (1.2)? What test do you use? (Use α = 0.05)

- Calculate the $F_{cal} = \frac{R^2/(k-1)}{1-R^2/(n-k)} = \frac{0.3791/(8-1)}{1-0.3791/(1,260-8)} = 109.204$. Meanwhile, we can actually see this value from F (7,1252) above.

- Look for the critical value of $F_{7,1252} = 2.01$ when $\alpha = 0.05$.

- Since F_{cal} far exceeds the critical value, we can reject our null hypothesis. In other words, we can make sure that, overall, the independent variables in this model provide significant explanation in variation of the dependent variable.

1.c) If we are interested in testing whether "physical attractiveness" has an impact on logarithm of hourly wage at all, what kind of null/alternative hypothesis would we be testing? Perform the test and discuss your finding. (Use α = 0.05)

- Here we can rely on either t-test of $belavg_i$ and $abvavg_i$ coefficients or the F-test for marginal contribution of these two variables added in the model 1.2.

- Let's setup the F-test: calculating the marginal contribution using R^2 since the dependent variable is the same.

$$F_{cal} = \frac{R^2_{new} - R^2_{old}/(number\ of\ new\ regressors)}{1 - R^2_{new}/(n - k_{new})} = \frac{0.3791 - 0.3731/(2)}{1 - 0.3791/(1{,}260 - 8)} = 6.0493$$

- Look for the critical value of $F_{2,1252} = 3$ when $\alpha = 0.05$.

- Since F_{cal} far exceeds the critical value, we can reject our null hypothesis. In other words, we can make sure two variables added for physical attractiveness have marginal contribution to the model.

1.d) Is there convincing evidence that women with above average looks earn more than women with average looks? Explain.

- Actually, this is a trick question. To fully answer the question, we do need to have an interaction term between $female_i$ and $abvavg_i$. In the result table, both $female_i$ and $abvavg_i$ are independent of each other which means that when we interpret, for instance, $abvavg_i$ coefficient even when it is significantly different from zero, it does not take $female_i$ into account.

- If the interaction term is added, then we can say for sure whether being female and having above average attractiveness has the additional effect compared to just being female.

Question 2.

A household expenditure model is given by

$$hhexp_i = \beta_1 + \beta_2 area_i + \beta_3 child_i + u_i$$

where $hhexp_i$ = household expenditure per month
$area_i$ = a dummy variable for household location:
(0 if in a municipal area and 1 if otherwise)
$child_i$ = number of children in household *i*, aged under 15

Using socio-economic dataset collected in 2018 with 14,908 households, the result is given below with **t value in parentheses**. Answer the following questions.

$$\widehat{hhexp}_i = \underset{(43.83)}{9{,}736} - \underset{(-15.8)}{2{,}835} area_i + \underset{(6.82)}{881} child_i + \hat{u}_i$$

2.a) Do all the signs for each coefficient make economic sense? Explain.

- Yes. For $area_i$, a dummy variable to represent household locale. Not living in a municipal area may lessen household expenditure compared to living in a city. Therefore, the negative sign makes sense. For $child_i$, the more number of children in a household, the more expenditure since there might be many additional expenses, such as educational expense, involved. Hence, the positive sign makes sense as well.

2.b) Test each parameter separately if they are significantly different from zero or not. (Use α = 0.01)

- The degrees of freedom for t-test is n-k: 14,908 – 3 = 14,905.

- The critical value is ± 2.576 when $\alpha = 0.01$.

- t-values are already given in parentheses, so, we can see that all of them exceeds either positive or negative critical value.

- For all tests, we can reject the null hypothesis. In other words, we can make sure that 99% all of the parameters, separately not simultaneously, are significantly different from zero.

2.c) Find the expected value of a household expenditure not living in a municipal area with 3 children aged under 15.

- Plugging all the value into the SRF, we get,

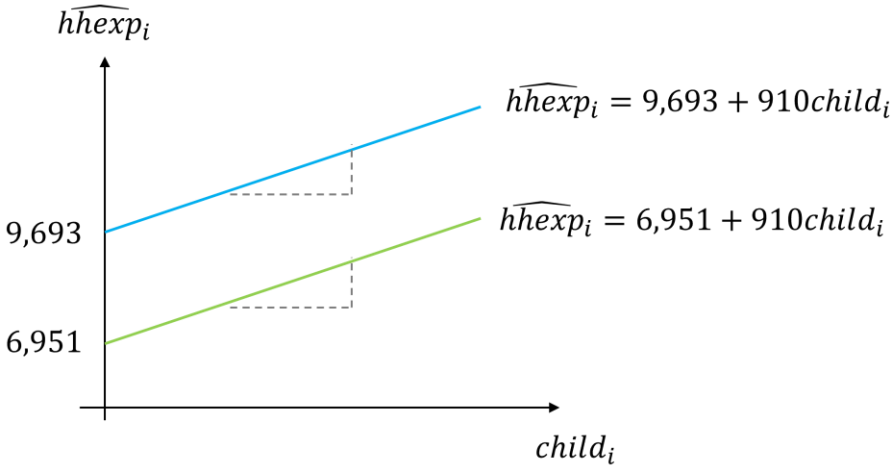
- $\widehat{hhexp}_i = 9{,}736 - 2{,}835(1) + 881(3) = 8{,}544.$

2.d) When an interaction term is included in this model, the result becomes with **t value in parentheses**.

$$\widehat{hhexp}_i = \underset{(34.38)}{9{,}693} - \underset{(-6.55)}{2{,}742}area_i + \underset{(5.17)}{910}child_i - \underset{(-0.25)}{64}(area_i * child_i) + \hat{u}_i$$

Draw a diagram for this model displaying sampled regression functions (SRF) with expected value of household expenditure on the vertical axis and number of children on the horizontal axis, taking **only significant parameter(s)** into account. Indicate the intercept and slope for each SRF where applicable. Testing of significance can be shortened.

- If we are using $\alpha = 0.05$, then the critical value is ± 1.96. Every parameter **but the interaction term's** is significant. Therefore, our diagram **will not include** the different slope between two groups, household living and not living in a municipal area.



- The blue SRF refers to household living in a municipal area while the green SRF refers to household not living in one. The difference between the two is from the dummy $area_i$ so the intercept of the latter group is 9,693 – 2,742 = 6,951.

- Both SRF slope is taken from $child_i$ coefficient which is 910 equally since the interaction term is not significantly different from zero.

Question 3.

Assume a multiple linear regression model as

$$hours_i = \beta_1 + \beta_2 sex_i + \beta_3 age_i + \beta_4 agesq_i + \beta_5 weekot_i + u_i$$

where $hours_i$ is hours worked in a week
sex_i is a dummy variable: 0 = male and 1 = otherwise
age_i is age of observation *i*
$agesq_i$ is age square observation *i*
$weekot_i$ is nominal overtime paid per week

Answer the following questions.

3.a) A VIF and tolerance table (postestimation) is given below

```
    Variable |       VIF       1/VIF
-------------+----------------------
       2.sex |      1.02    0.979129
         age |     50.61    0.019759
       agesq |     50.68    0.019731
      weekot |      1.01    0.985618
-------------+----------------------
    Mean VIF |     25.83
```

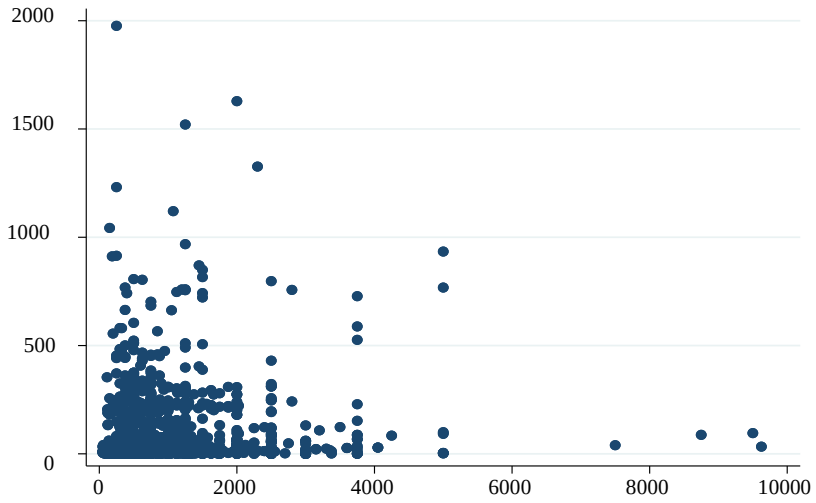
Given that you are exploring multicollinearity assumption, which pair of variables that you suspect they might be linearly correlated? Provide clear explanation what criteria (ion) that you rely on making that judgement.

age and agesq seems to be very much linearly correlated, considering the rule of thumb for both VIF that should not exceed 10 and the tolerance (1/VIF) that should tend to 1.

3.b) From **(3.a)**, do you consider removing one of the variables from the model? Why or why not and which one that you choose to remove, if that is the case?

Probably, I consider dropping the agesq if there is no explicit evidence or theory to back up. The squared term is generated to capture if there is any curvature of hours worked. In other words, it is in the model to capture any reduction of hours worked once we get older. If I am sure that it does not make any sense for a person to drop his/her working hour before retirement, I would drop the agesq term undoubtedly.

3.c) The graph provided below is a scatter plot between $\hat{u}_i^2$ (vertical axis) and $weekot_i$ (horizontal axis). Using the graphical method, do you conclude that heteroscedasticity is present in this model or not. Explain clearly to support your answer.



Heteroscedasticity can be detected when the predicted squared error term ($\hat{u}_i^2$) is correlated with any X or Y. From visual inspection, the larger weekot$_i$ it gets, $\hat{u}_i^2$ seems to be lower. Therefore, I highly suspect that heteroscedasticity is present in this model, but weakly due to unclear trend of $\hat{u}_i^2$ and weekot$_i$.

3.d) An auxiliary model here is estimated and the result is given in the table below.

$$\hat{u}_i^2 = \beta_1 + \beta_2 sex_i + \beta_3 age_i + \beta_4 agesq_i + \beta_5 weekot_i + v_i$$

```
      Source |       SS           df       MS      Number of obs   =     2,032
-------------+----------------------------------   F(4, 2027)      =      9.52
       Model |  829063.863         4  207265.966   Prob > F        =    0.0000
    Residual |    44148135     2,027   21780.037   R-squared       =    0.0184
-------------+----------------------------------   Adj R-squared   =    0.0165
       Total |  44977198.8     2,031  22145.3465   Root MSE        =    147.58
------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |  -5.648899   6.630832    -0.85   0.394    -18.65286    7.355058
         age |  -2.490434    2.37094    -1.05   0.294    -7.140168      2.1593
        age2 |    .044175   .0301279     1.47   0.143    -.0149098    .1032599
      weekot |   .0229916   .0043502     5.29   0.000     .0144603    .0315229
       _cons |    83.8484    44.4418     1.89   0.059    -3.307973    171.0048
------------------------------------------------------------------------------
```

From the table, setup the hypotheses and perform the Breusch-Pagan test to check that heteroscedasticity is present in the original model or not.

Setup BP test by using R^2,

- $F_{cal} = \frac{R^2_{\hat{u}_i^2}/(k)}{(1-R^2_{\hat{u}_i^2})/(n-k-1)} = \frac{0.0184/5}{(1-0.0184)/(2{,}032-5-1)} = 7.5954$

- Look for the critical value of $F_{5,2026} = 2.21$ when $\alpha = 0.05$.

- Since F_{cal} exceeds the critical value, we can reject the null hypothesis of homoscedasticity. In other words, we can make sure that heteroscedasticity is present in our model.

****