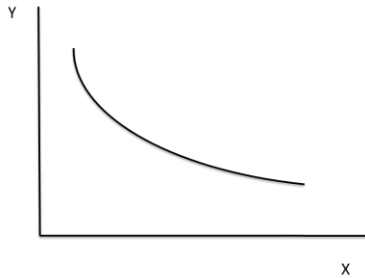


ECON 453 - Econometrics
Fall 2023
Exam 1 Practice Problems

These problems are intended to provide you with some additional examples of the types of questions I may ask on the exam, some of them are based on problems I have given on previous exams and assignments. **Note that these are not intended to cover every possible topic/question that could be included on the exam.** Please review the Exam 1 Study Guide for a more comprehensive list of topics.

1. Suppose we are running a regression with four explanatory variables. We want to perform a test to determine if the impact of the third variable (X_3) on Y is different from the impact of the fourth variable (X_4) on Y . Which **null** hypothesis should we use for this test?
 - a. $H_0 : \beta_3 = \beta_4 = 0$
 - b. $H_0 : \beta_3 \neq \beta_4 \neq 0$
 - c. $H_0 : \beta_3 = \beta_4$
 - d. $H_0 : \beta_3 \neq \beta_4$
2. Suppose we are estimating a simple linear regression where our y-variable is median household income in a state (in 1000s of \$) and our x-variable is the percentage of the adult population that has a college degree (on a scale between 0 and 100, so 45% = 45). If we estimate an equation of $\hat{y} = 16 + 2.2x$, how should we interpret the slope coefficient?
 - a. If the share of the population with a college degree increases by 1 percent, median income increases by \$2.20.
 - b. If the share of the population with a college degree increases by 1 percent, median income increases by \$2,200.
 - c. If the share of the population with a college degree increases by 1 percentage point, median income increases by \$2.20.
 - d. If the share of the population with a college degree increases by 1 percentage point, median income increases by \$2,200.
3. Which of the following is the most accurate definition of the exogeneity assumption in ordinary least squares regression?
 - a. We assume the explanatory variables are uncorrelated with the error term
 - b. We assume constant variance in the error term across the values of our explanatory variables
 - c. We assume there is independent variation in each of the explanatory variables
 - d. We assume that we know what we are doing, even when we clearly have no idea
4. If the classic assumptions of ordinary least squares regression are true, then we can expect our estimated coefficients to be unbiased. Which of these demonstrates the idea of an unbiased coefficient?
 - a. $E(\hat{\beta}_1) = 0$
 - b. $E(\beta_1) = 0$
 - c. $E(\beta_1) = \hat{\beta}_1$
 - d. $E(\hat{\beta}_1) = \beta_1$

5. According to our discussion, standardized coefficients can be useful because:
- They eliminate outliers from the sample
 - They allow us to compare the relative impact of variables measured on different scales
 - They convert negative coefficients to a positive equivalent
 - They can lessen the influence of multicollinearity on our coefficients
 - All of the above are reasons standardized coefficients can be useful
6. Suppose we believe that the relationship between X and Y is as shown in the graph below. If we estimate a model as: $Y = \beta_0 + \beta_1 X + \beta_2 (X^2)$, what signs do you expect for each of the coefficients?



- $\hat{\beta}_1 < 0, \hat{\beta}_2 > 0$
- $\hat{\beta}_1 < 0, \hat{\beta}_2 < 0$
- $\hat{\beta}_1 > 0, \hat{\beta}_2 > 0$
- $\hat{\beta}_1 > 0, \hat{\beta}_2 < 0$

For questions 7 and 8, suppose we use a Census sample of 25–35-year-olds working at least 30 hours a week to estimate the impact of gender and having a college degree on income. We measure Y (income) in thousands of dollars. We then include explanatory dummy variables for Bachelors (X_1 , equals 1 if college degree, 0 if not), Male (X_2 , equals 1 if male, 0 if not) and an interaction term. Using this, we come up with the following equation:

$$\hat{Y} = 29.0 + 17.8X_1 + 6.4X_2 + 12.1(X_1X_2)$$

7. In the equation above what does the coefficient 17.8 tell us?
- The impact of having a bachelor's degree on income for all individuals
 - The impact of having a bachelor's degree on income for males specifically
 - The impact of having a bachelor's degree on income for females specifically
 - The coefficient is not meant to represent any of these impacts
8. In the model above, if we want to test whether the effect of a bachelor's degree is different for men and women, which null hypotheses should we use?
- $H_0 : \beta_1 = 0$
 - $H_0 : \beta_2 = 0$
 - $H_0 : \beta_3 = 0$
 - $H_0 : \beta_1 = \beta_3$
 - $H_0 : \beta_1 = \beta_3 = 0$
9. The best way to deal with a severe multicollinearity problem in our regression is to use our statistical package (gretl, Excel, r, etc.) to correct the standard errors.
- True
 - False

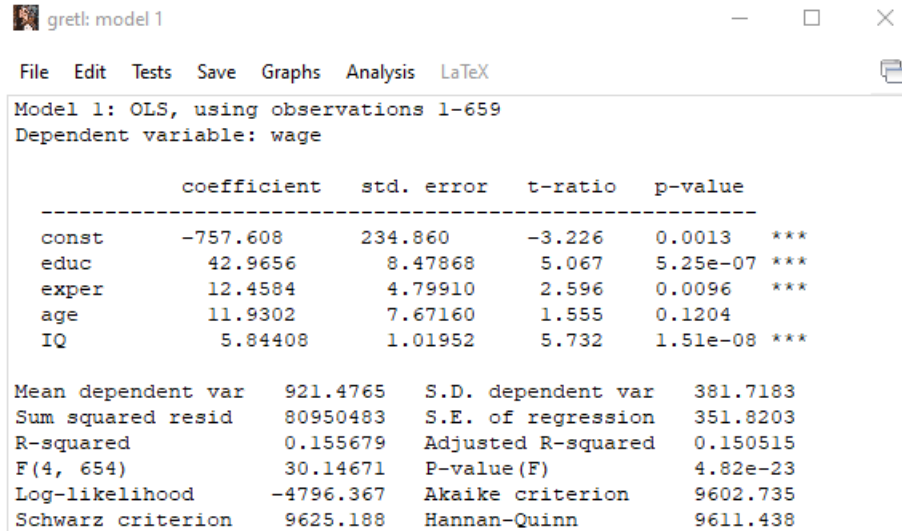
10. Suppose we are interested in estimating the relationship between the annual income in a household and the annual expenditures on clothing for a household. We collect a dataset that includes this information for a random sample of 1,000 households in 2022. Our explanatory variable, annual household income, is reported in 1,000s of US\$. Our dependent variable, annual household spending on clothing, is also measured in 1000s of US\$.
- We estimate this simple linear regression model and come up with the following prediction equation: $\hat{y}_i = -0.5 + 0.025x_i$. Interpret the coefficients in this model specifically. Does this result match your expectations?
 - Suppose you run into the Davidson family. They are old acquaintances that you lost touch with over the years. You ask them about little Donnie, and if he ever regained the feeling in his left hand. They do not want to speak about this topic. However, they do want to discuss how well they have been doing financially. They tell you their annual income for 2022 was \$110,000. Based on the model in part b, what do you predict the Davidsons spent on clothing in 2022? As the conversation progresses, you determine they spent \$5,000 (exactly) on clothes in 2022. What is the residual from our estimate?
 - One problem we are likely to have in a simple regression model like this one is confounding variables. Provide examples of two more explanatory variables you feel should be added to this model to help reduce this problem. Explain what relationship you expect each to have with the response variable.
 - This relationship (between income and clothing expenditures) is often used as an example of a regression that will produce a heteroskedasticity problem in our residuals. With the use of an example residual plot diagram, explain why there is likely to be a problem with our residuals that will violate our ordinary least squares regression assumptions.

11. Consider the dataset of males between the ages of 28 and 35 that contains information on weekly earnings and a number of other characteristics. We have worked with this dataset in class. Suppose you want to develop a model of the following form:

$$y_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 age_i + \beta_4 IQ_i + \mu_i$$

Where y = weekly earnings (in \$), $educ$ = years of education, $exper$ = years of work experience (overall), age = years of age, and IQ = IQ.

We run the regression for a sample of 659 individuals and obtain the following results:



gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

Model 1: OLS, using observations 1-659
Dependent variable: wage

	coefficient	std. error	t-ratio	p-value	
const	-757.608	234.860	-3.226	0.0013	***
educ	42.9656	8.47868	5.067	5.25e-07	***
exper	12.4584	4.79910	2.596	0.0096	***
age	11.9302	7.67160	1.555	0.1204	
IQ	5.84408	1.01952	5.732	1.51e-08	***

Mean dependent var	921.4765	S.D. dependent var	381.7183
Sum squared resid	80950483	S.E. of regression	351.8203
R-squared	0.155679	Adjusted R-squared	0.150515
F(4, 654)	30.14671	P-value(F)	4.82e-23
Log-likelihood	-4796.367	Akaike criterion	9602.735
Schwarz criterion	9625.188	Hannan-Quinn	9611.438

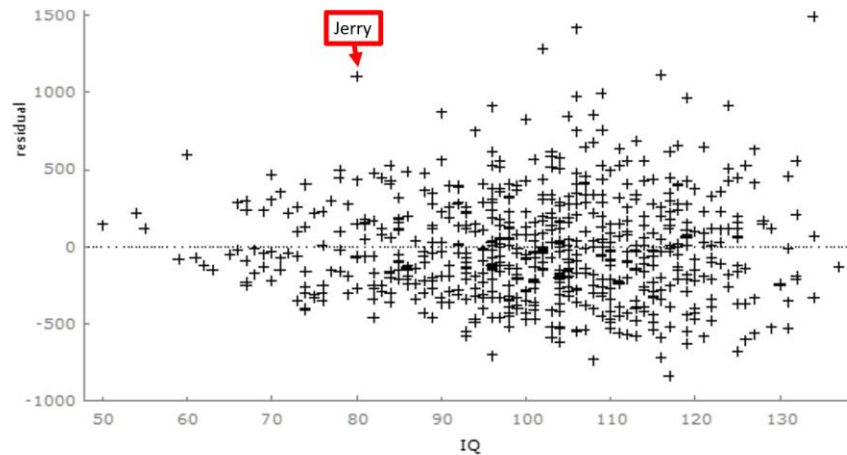
- Use this information to write out the regression equation and discuss the coefficients. Do these results match your expectations?
- What is the F-statistic (30.14671) and associated p-value ("P-value(F)") testing in this model? Discuss the hypotheses that are being tested, and what our conclusion should be in this case.
- How concerned should we be about multicollinearity in this model? Which variables (if any) in particular might be cause for concern? Consider the correlation table below and discuss how big of a problem this is. How else could we test for this issue?

	Wage	Educ	Exper	Age	IQ
Wage	1				
Educ	0.305	1			
Exper	-0.017	-0.454	1		
Age	0.117	0.062	0.399	1	
IQ	0.325	0.489	-0.224	-0.039	1

- Use the correlation table from part c. Explain, using the concept of omitted variable bias, what you expect would happen to the coefficient on the education variable if we dropped the IQ variable from our model.

CONTINUED ON NEXT PAGE

- e. Consider the residual plot for the “IQ” variable from the regression discussed in part a. Explain whether or not this residual plot is cause for concern. What do we know about Jerry based on the residual plot?



- f. Suppose I decide to try and simplify my model by removing some of the variables we used in Model 1. The model table below shows the original model and my new model (Model 2). Discuss how the remaining coefficients have changed from model 1 to model 2 and explain the intuition. Based on the result, is Model 1 or Model 2? What else should be done to improve the model going forward?

Dependent variable: wage		
	Model 1	Model 2
const	-757.61*** (234.86)	-154.25 (127.73)
educ	42.97*** (8.48)	67.20*** (7.40)
exper	12.46*** (4.80)	15.99*** (4.31)
age	11.93 (7.67)	
IQ	5.84*** (1.02)	
n	659	659
Adj. R ²	0.15	0.11
lnL	-4.8e+03	-4.8e+03

12. Suppose we are running a bakery and are interested in determining how many pies we will sell each week. We are going to estimate the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, where y =number of pies sold in the week, x_1 = price of pies (in dollars), and x_2 is a dummy variable equal to 1 if a holiday occurred during the week and 0 if there was no holiday.
- If we collected data and estimated this equation as: $\hat{y} = 300 - 30x_1 + 15x_2$, interpret the coefficients on the explanatory variables. Predict the number of pies sold when pies cost \$5 each for both a holiday and non-holiday week.
 - We then added the interaction term and estimated the equation as: $\hat{y} = 320 - 40x_1 + 20x_2 + 30(x_1x_2)$. What equation would you use to predict sales based on price in a holiday week? In a non-holiday week? What is the economic explanation for what the interaction term specifically tells you, and does this seem reasonable? Predict the number of pies sold when pies cost \$5 each for both a holiday and non-holiday week, and compare to that from part a.
13. Suppose we are looking at our cross-sectional data of health and economic variables across countries from the World Bank indicators. We will estimate a model where our dependent variable is the average life expectancy in each country (in years). Our explanatory variables are: GDP per capita (in 1,000s of US\$), % of the country with access to improved water, and % of young children that have been immunized for measles. Note that the percentage variables are measured on a 0 to 100 scale (instead of 0 to 1). This means if you are predicting for a country that has 50% access to clean water, you would plug in a value of 50 (instead of 0.50).

gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

Model 1: OLS, using observations 1-214 (n = 161)
Missing or incomplete observations dropped: 53
Dependent variable: life_exp

	coefficient	std. error	t-ratio	p-value	
const	26.2705	3.23818	8.113	1.34e-013	***
gdp_1000	0.140135	0.0248180	5.647	7.48e-08	***
water_access	0.361894	0.0385956	9.377	7.26e-017	***
imm_measles	0.108479	0.0400575	2.708	0.0075	***
Mean dependent var	69.19317	S.D. dependent var	9.997073		
Sum squared resid	5065.784	S.E. of regression	5.680329		
R-squared	0.683203	Adjusted R-squared	0.677150		
F(3, 157)	112.8619	P-value(F)	5.39e-39		
Log-likelihood	-506.0823	Akaike criterion	1020.165		
Schwarz criterion	1032.490	Hannan-Quinn	1025.169		

- Interpret the coefficients and discuss how reasonable they seem to you. Discuss how strong the model is overall, both based on statistics and in terms of how good of a job we have done at identifying the true relationships with our model.
- Predict the life expectancy for a country with GDP per capita of \$12,000, 95% with access to clean water, and 60% immunized for measles.
- Suppose we run White's test on the regression above and come up with a p-value of 0.000. What does this mean about our regression and what (if anything) needs to be done?
- I computed the standardized coefficients in gretl (see below). What do these tell us?

	std. dev.	standardized coeff.
const	0	2.6696
gdp_1000	23.115	0.32917
water_access	15.579	0.57293
imm measles	13.828	0.15243

The next thing we try is to look for non-linear effects. Specifically, we are interested in whether or not there are decreasing or increasing returns to improving GDP per capita. We run the same regression as before, but now include GDP per capita (in 1000s) squared as an explanatory variable.

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

Model 2: OLS, using observations 1-214 (n = 161)
Missing or incomplete observations dropped: 53
Dependent variable: life_exp

	coefficient	std. error	t-ratio	p-value	
const	29.3788	3.22194	9.118	3.61e-016	***
gdp_1000	0.359841	0.0635047	5.666	6.86e-08	***
water_access	0.313995	0.0392538	7.999	2.65e-013	***
imm_measles	0.105004	0.0385139	2.726	0.0071	***
gdp_sqrd	-0.00267814	0.000717411	-3.733	0.0003	***
Mean dependent var	69.19317	S.D. dependent var	9.997073		
Sum squared resid	4650.360	S.E. of regression	5.459853		
R-squared	0.709182	Adjusted R-squared	0.701725		
F(4, 156)	95.10463	P-value(F)	8.19e-41		
Log-likelihood	-499.1944	Akaike criterion	1008.389		
Schwarz criterion	1023.796	Hannan-Quinn	1014.645		

- Discuss what this model predicts about the relationship between GDP per capita and life expectancy in a country. Do the results make sense?
- Predict the life expectancy for a country that has GDP per capita of \$10,000, 95% water access and 60% immunized for measles. Then predict the life expectancy for the same country if GDP per capita increases to \$20,000, and again for an increase to \$30,000. Discuss the findings.

Next, I remove the quadratic term, and instead include dummy variables for the continents.

gretl: model 4

File Edit Tests Save Graphs Analysis LaTeX

Model 4: OLS, using observations 1-214 (n = 161)
Missing or incomplete observations dropped: 53
Dependent variable: life_exp

	coefficient	std. error	t-ratio	p-value	
const	44.1141	3.56776	12.36	9.62e-025	***
gdp_1000	0.106366	0.0212445	5.007	1.52e-06	***
water_access	0.211842	0.0345919	6.124	7.46e-09	***
imm_measles	0.0929170	0.0327653	2.836	0.0052	***
africa	-9.63045	1.69226	-5.691	6.31e-08	***
europa	1.08485	1.67583	0.6473	0.5184	
asia	-1.72882	1.61546	-1.070	0.2862	
north_amer	0.872483	1.78763	0.4881	0.6262	
oceania	-0.0570011	2.18485	-0.02609	0.9792	
Mean dependent var	69.19317	S.D. dependent var	9.997073		
Sum squared resid	3109.770	S.E. of regression	4.523164		
R-squared	0.805526	Adjusted R-squared	0.795290		
F(8, 152)	78.69922	P-value(F)	3.75e-50		
Log-likelihood	-466.8015	Akaike criterion	951.6030		
Schwarz criterion	979.3357	Hannan-Quinn	962.8636		

- Discuss the interpretations of the coefficients on the dummy variables. Is this model an improvement over that in part a? What would you recommend doing to improve the model going forward?

14. Consider a sample of all 25 to 35-year-olds living in Idaho from the 2015 American Community Survey. We will estimate a model where we predict income (in thousands of dollars) based on various factors among those who report normally working at least 30 hours per week. We will use a logarithmic transformation of income in this model. The explanatory variables are dummy variables for female and having a bachelor's degree, and quantitative variables for age, the typical number of hours worked per week, and number of children (living in the household).

gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

Model 1: OLS, using observations 1-1602
Dependent variable: l_incwage

	coefficient	std. error	t-ratio	p-value	
const	9.54763	0.0933950	102.2	0.0000	***
age	0.00748796	0.00115833	6.464	1.35e-010	***
female	-0.314978	0.0273038	-11.54	1.25e-029	***
bach	0.432507	0.0293931	14.71	4.68e-046	***
nchild	0.0295276	0.0117116	2.521	0.0118	**
hours	0.0128612	0.00149032	8.630	1.46e-017	***
Mean dependent var	10.52092	S.D. dependent var	0.594729		
Sum squared resid	427.3471	S.E. of regression	0.517457		
R-squared	0.245339	Adjusted R-squared	0.242975		
F(5, 1596)	103.7714	P-value(F)	5.80e-95		
Log-likelihood	-1214.689	Akaike criterion	2441.378		
Schwarz criterion	2473.652	Hannan-Quinn	2453.361		

- Interpret the estimated coefficients in the model and discuss whether or not they match your expectations.
- Would it make sense to use a logarithmic transformation on any of the other variables in the model? Discuss.
- I tried a model where I included an interaction term between female and bachelor's degree. Discuss what this tells us about the gender gap and the benefits to having a college degree.

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

Model 2: OLS, using observations 1-1602
Dependent variable: l_incwage

	coefficient	std. error	t-ratio	p-value	
const	9.55384	0.0935540	102.1	0.0000	***
age	0.00753910	0.00115915	6.504	1.04e-010	***
female	-0.333442	0.0319351	-10.44	9.87e-025	***
bach	0.403858	0.0390459	10.34	2.59e-024	***
nchild	0.0304957	0.0117429	2.597	0.0095	***
hours	0.0127671	0.00149260	8.554	2.75e-017	***
fem_bach	0.0659868	0.0592070	1.115	0.2652	
Mean dependent var	10.52092	S.D. dependent var	0.594729		
Sum squared resid	427.0146	S.E. of regression	0.517417		
R-squared	0.245926	Adjusted R-squared	0.243090		
F(6, 1595)	86.69632	P-value(F)	3.36e-94		
Log-likelihood	-1214.065	Akaike criterion	2442.131		
Schwarz criterion	2479.784	Hannan-Quinn	2456.112		