Final Exam Practice

Below, you will find practice questions that should help you to study for the PUAD 6080 final exam. The questions will closely resemble the types of prompts you will see on the final exam.

1. Suppose you study the age of female black bears and the number of cubs they have borne. For the 5 bears you assess, you find age as follows: $A = \{5, 10, 15, 20, 25\}$. Furthermore, you find that the number of cubs these bears have had is: $C = \{0, 6, 5, 12, 17\}$. Estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ by hand. Interpret these coefficients. What percent more cubs is a 23 year-old bear projected to have compared to a 7 year-old bear? Suppose that $\hat{\beta}_1$ has a standard error equal to 0.14. Given an α-threshold of 0.05 (one-tailed) and degrees of freedom equal to 3, is the relationship between age and cub-bearing statistically significant?

2. Suppose you want to study how different features in automobiles affect their fuel efficiency. You decide to operationalize your dependent variable as the miles a given car is estimated to be able to travel on a gallon of gasoline. For predictor variables, you examine the number of cylinders in a car's engine (Cylinders); engine displacement, which is measured in cubic inches (Displacement); engine horse power (Horse Power); rear axle ratio (Rear Axle Ratio); weight, which is measured in thousands of pounds (Weight); the time it takes a car to travel one-quarter of a mile, measured in seconds (1/4 Mile Time); and whether a car's transmission is an automatic, measured dichotomously (Automatic).

Table 1: Automobile fuel economy

Variable	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	p-value
Cylinders	-0.34	0.85	0.36
Displacement	0.01	0.01	0.18
Horse Power	-0.02	0.02	0.16
Rear Axle Ratio	0.82	1.48	0.34
Weight	-3.99	1.23	0.00
1/4 Mile Time	0.86	0.59	0.14
Automatic	2.72	1.79	0.12
Intercept	15.31	15.99	0.24

Notes: The dependent variable is a vehicle's estimated fuel efficiency (miles per gallon). $N = 32$. $F = 22.47$ ($p < 0.000$). $R^2 = 0.87$.

Using an α-threshold of 0.10, which variables are statistically significant? Of those that are, what is the direction of the relationship? Be able to interpret their partial slope coefficients as well. Is the model as a whole statistically significant? If so, what's it's goodness of fit?

3. What are the assumptions of the Classical Linear Regression Model (CLRM)? Be certain that you can explain what each of these assumptions requires in practice, and be aware of the types of data that are likely to result in violations of the CLRM. What is true about OLS if each of these assumptions holds?

4. Suppose you want to understand the effect that religious adherence has upon one's belief in an afterlife, $y \in [0, 100]$, with 100 indicating certain belief and 0 indicating certain disbelief. You ask respondents their religious affiliation and code a nominal variable, $x \in \{\text{Atheist}, \text{Christian}, \text{Jewish}, \text{Muslim}\}$. Write an OLS model that utilizes information in y and x, and explain how you will interpret OLS coefficients.

5. What are interaction effects in multiple regression analysis? How do we estimate them? How is failing to account for interactive effects that belong in the model contributing to bias among our partial slope coefficients and therefore a violation of the Gauss-Markov theorem?

6. Suppose you gather information on an individual's income in dollars, their race (measured dichotomously, "1" if white, "0" else), and their gender (measured dichotomously, "1" if male, "0" else). You specify the following linear regression model:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{White}_i + \hat{\beta}_2 \text{Male}_i + \hat{\beta}_3 (\text{White}_i \times \text{Male}_i) + \hat{u}_i.$$

The results of your linear regression model are presented in Table 2 and derive from 2018 responses to the General Social Survey. Using an α-threshold of 0.05, interpret the results from this regression. Which effects are statistically significant? What's the predicted income for white women, white men, nonwhite women, and nonwhite men?

Table 2: Individual demographics and income

Variable	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	p-value
White	5510.67	2277.82	0.02
Male	5186.16	2835.04	0.07
White × Male	7219.72	3362.82	0.04
Intercept	16151.64	1898.34	0.00

Notes: The dependent variable is an individual's real income. $N = 1,363$. $F = 26.85$ ($p < 0.000$). $R^2 = 0.06$.

7. An analysis of the residuals from the regression in the previous question returns a Shapiro-Wilk test statistic equal to 0.70 ($p < 0.000$). A Breusch-Pagan test returns a test statistic equal to 161.77 ($p < 0.000$). A RESET test returns a test statistic equal

to 4.61 ($p = 0.032$). Using an α-threshold of 0.05, what are each of these tests telling us about our regression model? What steps should we take to address them?

8. Explain the following concepts with respect to linear regression analysis: Outlier, influence point, leverage point. What kinds of effects do these types of observations have upon our OLS estimates? Is OLS still BLUE in the presence of any of these kinds of observations? How does one go about detecting "problematic" observations, and what should one do if they find any?

9. Suppose you estimate a linear regression model of individual income with the following independent variables: Individual Education, Individual Socioeconomic Status, Parents' Income, Parents' Education, Parents' Socioeconomic Status. You run your regression and find that none of these variables, despite your expectations, are statistically significant. Upon assessing variables' variance inflation factors, you find that each of your variables of interest have VIFs greater than 15.00. If you do nothing, absent other issues with your model, are your OLS estimates still BLUE? Why are you likely finding null results for each of your variables? What steps might you take to address issues related to multicollinearity?

10. Explain how missing data might be problematic (or not) for OLS. What does OLS do with missing data that can lead it to bias our coefficient estimates?

11. Suppose you analyze the end-of-year value of the Dow Jones Industrial Average between 2000 and 2020. As independent variables, you control for the yearly unemployment rate along with the annual gross domestic product (in billions). The results of your linear regression appear in Table 3. While the model appears to do a nice job at explaining outcomes in the dependent variable, a Breusch-Godfrey test returns a test-statistic of 5.44 ($p = 0.020$). If we use an α-threshold of 0.05, how are we in violation of the assumptions of the CLRM; what does that mean for the results of our model; and what steps can we take to address issues like this?

Table 3: The value of the DJIA

Variable	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	p-value
Unemployment	-755.62	309.98	0.03
GDP (billions)	1.73	0.17	0.00
Intercept	-7132.40	3221.66	0.04

Notes: The dependent variable is the end-of-year value of the DJIA. $N = 21$. $F = 57.07$ ($p < 0.000$). $R^2 = 0.85$.

12. What is the linear probability model? Explain why, even if it satisfies the assumptions of the CLRM, it is probably not well-suited for regression analysis. Explain how logit or probit address this deficiency.