Name: _____

This exam contains 24 questions, some of which have multiple parts. It is designed to take one hour and fifteen minutes, the entire class period. This means you do not have time to sit on one question for an extended period or else you will not be able to finish the exam. Please use your time wisely.

You are allowed to use an 8.5 inch x 11 inch sheet of notes (front and back). No other assistance is permitted during the exam including, but not limited to, discussion with friends and electronic devices.

This exam is worth 133 points (13.3%) of your grade.

Please move to the next page to start the exam.

1. (2 Points) When would a second degree polynomial regression model be useful?
    a. Our outcome is initially an increasing function of a covariate, but after some point the outcome begins decreasing as function of that covariate.
    b. When we want to complicate our life by making derivatives harder.
2. (2 Points) When would a model with interaction terms be useful?
    a. When the effect of one covariate on our outcome depends on another covariate.
    b. When our outcome variable is inversely related to one covariate, while being positively related to another covariate.

(2 Points Each) Matching

a. Nested Models b. Non-nested Models c. Omitted Variable Bias (OVB)
d. Homoskedastic Errors e. Heteroskedastic Errors f. Simultaneity g. Binary Outcome y_i
h. Latent Variable

3. Neither model is a special case of one another _____
4. Reverse causality _____
5. Excluding a relevant explanatory variable from the model _____
6. Unobserved _____
7. $\mathbb{V}[\boldsymbol{u} \mid X] = \text{diag}\left(\sigma_i^2\right)$ (non-constant error variance) _____
8. $\mathbb{E}[y_i] = \mathbb{P}(y_i = 1)$ _____
9. One model is a subset of another model's covariates _____
10. $\mathbb{V}[\boldsymbol{u} \mid X] = \sigma^2 I_n$ (constant error variance) _____

11. (4 Points) Suppose the true model is $y_i = \beta_0 + \beta_1 x_{i1} + u_i$, but we estimate $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$. What is this an example of? Are there any consequences to estimating the latter model instead of the former? If so, why? If not, why not?

12. (4 Points) Suppose the true model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, but we estimate $y_i = \beta_0 + \beta_1 x_{i1} + u_i$. Under what assumptions would the estimator $\widehat{\beta}_1$ be biased? State the four possible scenarios determining the direction of this bias.

13. (4 Points) Describe the Difference-in-Differences (DiD) estimator and its assumptions. Write the model mathematically while controlling for a vector of covariates denoted by $\boldsymbol{x}_i$.

14. (4 Points) Describe maximum likelihood estimation (MLE) and its intuition. What is the problem MLE solves to obtain the estimated parameters? What are some of its properties under mild assumptions?

15. (4 Points) Discuss the difference between a pooled-cross sectional data set and a panel data set. Why do we prefer a panel data set especially with regard to omitted variable bias (OVB)? What are the panel data estimators we discussed that allow us to alleviate the OVB concern?

16. (4 Points) Explain endogeneity and discuss why it is a problem in empirical econometrics. There are two main ways endogeneity can arise. What are they? Why is endogeneity not a concern in the natural sciences?

17. (8 Points) What is a counterfactual? Feel free to use our discussion on difference-in-differences (DiD) to illustrate this.

18. (6 Points) Let us examine heteroskedasticity.

    (a) (2 Points) Why is heteroskedasticity relevant regarding statistical inference/hypothesis testing?

    (b) (2 Points) Given your answer to (a), should we always use a heteroskedasticity correction even if we do not know heteroskedasticity is present? Why or why not?

(c) (2 Points) The heteroskedasticity corrections we discussed in class are based on asymptotic theory. What does this mean?

19. (6 Points) Consider the econometric model $colgpa_i = \beta_0 + \beta_1 male_i + \beta_2 hours_i + \beta_3 male_i \times hours_i + u_i$ where $colgpa_i$ is individual i's college GPA, $male_i$ is an indicator variable equaling one if an individual i is male, and $hours_i$ is the number of weekly hours individual i spends studying on average.

(a) (3 Points) Give the estimated regression equations for both males and females.

(b) (3 Points) Suppose that $\widehat{\beta}_1 < 0$ and $\widehat{\beta}_3 > 0$. What does this mean?

20. (20 Points) Consider the econometric linear probability model $cancer_i = \beta_0 + \beta_1 tobacco_i + \beta_2 alcohol_i + \beta_3 tobacco_i \times alcohol_i + u_i$, where $cancer_i$ is an indicator variable equaling one if an individual i has cancer, $tobacco_i$ is an indicator variable equaling one if an individual i is a tobacco user, and $alcohol_i$ is an indicator variable equaling one if an individual i is an alcohol user.

    Upon estimation you find

$$\begin{aligned}\widehat{cancer}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 tobacco_i + \widehat{\beta}_2 alcohol_i + \widehat{\beta}_3 tobacco_i \times alcohol_i \\ &= 0.1 + 0.3 tobacco_i + 0.2 alcohol_i + 0.35 tobacco_i \times alcohol_i.\end{aligned}$$

    (a) (4 Points) What is the interpretation of $\widehat{\beta}_2$?

    (b) (4 Points) What is the interpretation of $\widehat{\beta}_3$?

    (c) (4 Points) If someone consumes both tobacco and alcohol, what is the predicted probability that they will get cancer?

(d) (8 Points) Types of cancer are known to be hereditary meaning individuals who are related to others that have these types of cancer are more likely to have it themselves. Also, individuals who are related to those who use alcohol and tobacco are more likely to use it themselves. Given this information, are you concerned with the parameter estimates of β_2 and β_3? If so, make a case for why this econometric model is concerning, make a clear statement about what is wrong with these estimates, and provide a solution to the problem. If you are not concerned, make an argument for why this econometric model is appropriate to identify the true causal effects of tobacco and alcohol use on cancer probability.

21. (28 Points) Suppose $y_i^* = \boldsymbol{x}_i'\boldsymbol{\theta} + \epsilon_i$ is latent and we only observe $y_i = \mathbb{1}(y_i^* > 0)$. Recall that for binary response models we have the conditional distribution of our outcome y_i given our covariate vector $\boldsymbol{x}_i$ is

$$f_{Y|X}(y_i \mid \boldsymbol{x_i}; \boldsymbol{\theta}) = G\left(\boldsymbol{x}_i'\boldsymbol{\theta}\right)^{y_i} \left(1 - G\left(\boldsymbol{x}_i'\boldsymbol{\theta}\right)\right)^{1-y_i}.$$

(a) (4 Points) What is the conditional probability that $y_i = 1$? What is the conditional probability that $y_i = 0$?

(b) (8 Points) State the maximum likelihood problem given we assume the error term follows that logistic distribution meaning

$$G\left(\boldsymbol{x}_i'\boldsymbol{\theta}\right) = \Lambda\left(\boldsymbol{x}_i'\boldsymbol{\theta}\right) = \frac{\exp\left(\boldsymbol{x}_i'\boldsymbol{\theta}\right)}{1 + \exp\left(\boldsymbol{x}_i'\boldsymbol{\theta}\right)}.$$

(c) (4 Points) Given the assumption of part (b), what model are we estimating?

(d) (4 Points) Why is this model preferable over the linear probability model (LPM)?

(e) (4 Points) Suppose you estimate this model in R where your outcome variable is y and your regressor is x. Write the R code to estimate this model as well as to obtain the summary output assuming your data is contained in the data table "dt". Do the estimates you see represent the partial effects? If so, why? If not, write the R code to obtain the average partial effects.

(f) (4 Points) Explain why we need typically need a numerical solver like Newton's Method to iteratively solve the parameter estimates when estimating binary response models rather than simply taking first order conditions like we did with OLS.

Hint: Think about what kind of function Λ is and how that differs from MLR Assumption 1 which we used to obtain the OLS solution by taking first order conditions.

You will use the following R output to answer Problems 22 through 24.

Suppose we estimate a fixed effects model using a panel data set tracking wage and a rich set of covariates across 545 individuals from 1980 to 1987. Upon estimation we obtain:

```
Oneway (individual) effect Within Model

Call:
plm(formula = lwage ~ married + union + d81 + d82 + d83 + d84 +
    d85 + d86 + d87 + I(educ * d81) + I(educ * d82) + I(educ *
    d83) + I(educ * d84) + I(educ * d85) + I(educ * d86) + I(educ *
    d87), data = dt, model = "within", index = c("nr", "year"))

Balanced Panel: n = 545, T = 8, N = 4360

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-4.152111 -0.125630  0.010897  0.160800  1.483401

Coefficients:
                 Estimate Std. Error t-value  Pr(>|t|)
married        0.0548205  0.0184126  2.9773  0.002926 **
union          0.0829785  0.0194461  4.2671 2.029e-05 ***
d81           -0.0224158  0.1458885 -0.1537  0.877893
d82           -0.0057611  0.1458558 -0.0395  0.968495
d83            0.0104297  0.1458579  0.0715  0.942999
d84            0.0843743  0.1458518  0.5785  0.562965
d85            0.0497253  0.1458602  0.3409  0.733190
d86            0.0656064  0.1458917  0.4497  0.652958
d87            0.0904448  0.1458505  0.6201  0.535216
I(educ * d81)  0.0115854  0.0122625  0.9448  0.344827
I(educ * d82)  0.0147905  0.0122635  1.2061  0.227872
I(educ * d83)  0.0171182  0.0122633  1.3959  0.162830
I(educ * d84)  0.0165839  0.0122657  1.3521  0.176437
I(educ * d85)  0.0237085  0.0122738  1.9316  0.053479 .
I(educ * d86)  0.0274123  0.0122740  2.2334  0.025583 *
I(educ * d87)  0.0304332  0.0122723  2.4798  0.013188 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 474.35
R-Squared:      0.1708
Adj. R-Squared: 0.048567
F-statistic: 48.9069 on 16 and 3799 DF, p-value: < 2.22e-16
```

Here, *lwage* represents the natural log of hourly wage, *married* indicates if an individual is married or not, *union* indicates if an individual is a member of a union or not, $d81$ to $d87$ are time indicator variables, and $I(educ*d81)$ to $I(educ*d81)$ are interaction terms of education level and the time indicator variables.

22. (6 Points) By including individual fixed effects, give an example of something this model controls for that would have been difficult to control for otherwise leading to less biased parameter estimates.

23. (6 Points) The model says "Oneway (individual) effect Within Model." Does this model also control for time effects? Why or why not? Give an example of such a time effect this model could be controlling for leading to less biased parameter estimates.

24. (9 Points) What is the interpretation of the estimate of $I(d87 * educ)$? What does the evolution of the estimates of $I(educ * d81)$ to $I(educ * d87)$ tell you about returns to education over the 1980 to 1987 time period?