Name: _____

This exam contains 32 questions. 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) Not studying for an exam being inversely related to college GPA is a statement of
   a. Causation
   b. Correlation
2. (2 Points) Attending a private high school raises ones income by 10% is a statement of
   a. Causation
   b. Correlation

(2 Points Each) Matching

a. Explanatory Variables b. Outcome Variable c. $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$ d. $\widehat{\boldsymbol{\beta}}$ e. $\boldsymbol{y} - X\boldsymbol{\beta}$
f. Parameter Vector

3. $\boldsymbol{\beta}$ _____
4. Error Vector $\boldsymbol{u}$ _____
5. Residual Vector $\widehat{\boldsymbol{u}}$ _____
6. $\boldsymbol{y}$ _____
7. Feature Matrix X _____
8. $(X'X)^{-1}X'\boldsymbol{y}$ _____

9. (2 Points) Describe a cross-sectional dataset.

10. (4 Points) State MLR Assumption 1. Does this assumption restrict us from examining a non-linear relationship between our outcome and our covariate(s)?

11. (4 Points) State MLR Assumption 2. What exactly does this assumption mean?

12. (4 Points) State MLR Assumption 3. What happens when this assumption is violated and what would this mean?

13. (4 Points) State MLR Assumption 4 in matrix form. What does this assumption mean in non-math terms?

14. (4 Points) What does the adjusted R-squared measure and why do we prefer it over the regular R-squared?

15. (4 Points) Is higher variance in our regressors a good thing? If so, why? If not, why not?

16. (4 Points) What does the Guass-Markov Theorem say? Under what assumptions does the Gauss-Markov Theorem hold? How does this relate to our estimator being BLUE and what does this acronym mean?

17. (4 Points) State MLR Assumption 4 in matrix form and explain what it means in non-math terms.

18. (4 Points) State the assumptions needed to guarantee the OLS estimator is unbiased. What does the OLS estimator being unbiased mean in non-math terms?

19. (4 Points) Why can research scientists conduct an experimental study, run an OLS regression of

$$got\_covid_i = \beta_0 + \beta_1 got\_vaccine_i + u_i,$$

and interpret the estimate of β_1 as the causal effect of the vaccine on whether or not one gets COVID? What type of dataset must be used here compared to the kind of datasets economists typically have available to them?

20. (4 Points) What assumption must be made to conduct inference in finite samples? What does this assumption imply about our OLS estimators?

21. (4 Points) What is the term given for the smallest level of significance such that the null hypothesis can be rejected?

22. (4 Points) What are asymptotics and why is it important in statistical inference? What assumption do we get to avoid making and still have reliable statistical inferences with sufficiently large samples?

23. (7 Points) Given the linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$, what is the formula for the OLS estimator and its estimated variance-covariance matrix? What do the square roots of the elements along the diagonal of this estimated variance-covariance matrix represent?

Use the following information to answer Problems 24 through 26. You collect data from a survey on yearly income $income_i$, education in years $educ_i$, experience in years $exper_i$, age age_i, gender $gender_i$, and average number of times weed is smoked per week $weed_i$.

24. (4 Points) Write out the equation that would allow you to estimate the effects of weed usage on yearly income, while controlling for other factors. You should be able to make the statement "smoking weed five more times per week would lead to a predicted change in yearly income of x%."

25. (4 Points) Suppose upon estimating this model you see that the p-value on gender is well above 0.05. What does this imply for a T-test of $H_0 : \beta_{gender} = 0$ versus $H_A : \beta_{gender} \neq 0$ with $\alpha = 0.05$?

26. (4 Points) Suppose upon estimating this model we conduct an F-test where the null hypothesis is that each included slope coefficient is equal to zero. After carrying out this test, we fail to reject this null hypothesis and drop these regressors from our model so that our updated model is now $income_i = \beta_0 + u_i$. What does this conclusion imply about the predictions of our model?

You will use the following R output to answer Problems 27 through 31. Suppose we estimate a linear regression model with college GPA as the outcome and get this output from R:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.80612 -0.30055  0.02035  0.33471  1.18006

Coefficients:
                Estimate   Std. Error t value             Pr(>|t|)
(Intercept) -0.023759108  0.174500199  -0.136                0.892
hsgpa        0.648049922  0.053154305  12.192 < 0.0000000000000002 ***
act          0.026520295  0.005488096   4.832              0.00000161 ***
study        0.003921578  0.006589506   0.595                0.552
study_sq    -0.000002247  0.000168500  -0.013                0.989
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4653 on 809 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.2591,    Adjusted R-squared:  0.2555
F-statistic: 70.73 on 4 and 809 DF,  p-value: < 0.00000000000000022
```

where hsgpa is high school GPA, act is a student's ACT score, and study represents the hours spent studying per week.

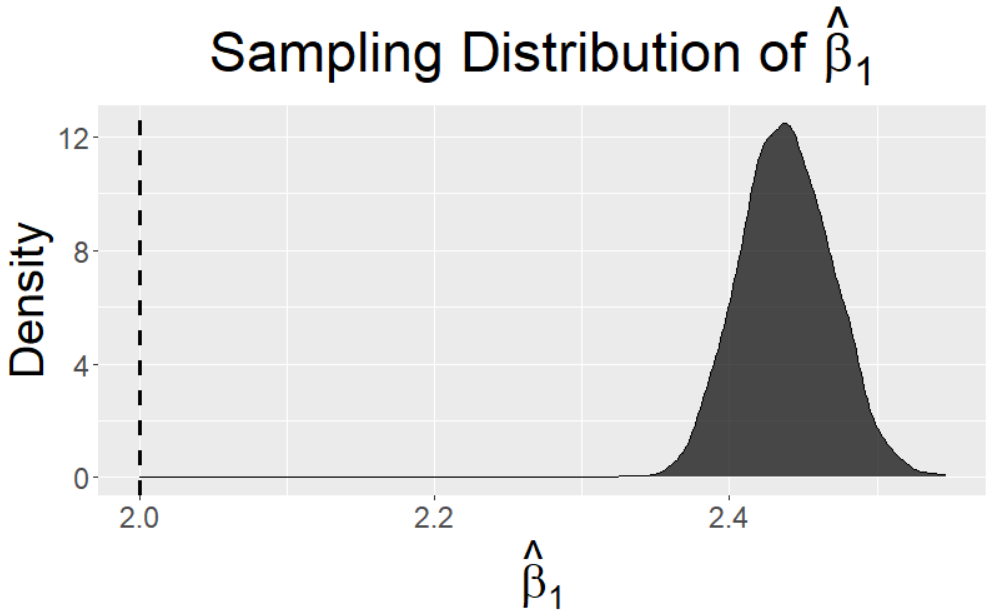
27. (6 Points) Write the R code to run this regression as well as to get the summary output table above. You can assume the data is contained in the data table "dt" and that the college GPA variable is defined as "colgpa".

28. (8 Points) Mathematically, what is the estimated partial effect of high school GPA on college GPA (think about the derivative)? How much does gaining an additional point in one's high school GPA change their college GPA?

29. (8 Points) Mathematically, what is the estimated partial effect of hours spent studying on college GPA (think about the derivative)? How much does studying for additional five hours per week alter one's college GPA?

30. (8 Points) Construct a 95% confidence interval around $\widehat{\beta}_{act}$. The critical value here is 1.96. Don't worry about finishing the computation as you don't have a calculator. Suppose 0.03 was not included in this interval. What could you say about a hypothesis test of $H_0 : \beta_{act} = 0.03$ versus $H_A : \beta_{act} \neq 0.03$?

31. (8 Points) Suppose we wish to conduct a hypothesis test of $H_0 : \beta_{hsgpa} = 1$ versus $H_A : \beta_{hsgpa} \neq 1$ at the $\alpha = 0.1$ significance level. What is the R command to obtain the critical value for this specific test? What is the test statistic corresponding to this test? State the condition under which we would reject the null hypothesis. If this condition holds true, does this necessarily mean the null hypothesis is false?



32. (6 Points) Suppose we gather 200 samples and estimate $\widehat{\beta}_1$ for each sample. We magically know the true parameter is $\beta_1 = 2$. What does the above graph represent? Can we conclude $\widehat{\beta}_1$ is biased? What could cause this bias? If we cannot conclude $\widehat{\beta}_1$ is biased, how would the graph look if this estimator were biased?