University of Texas at Austin

STA 235 Mastery Exam B 2025-05-01

Personal Data

Family Name:

Given Name:

Signature:

checked

Registration Number

0	☐	☐	☐	☐	☐	☐	☐	0
1	☐	☐	☐	☐	☐	☐	☐	1
2	☐	☐	☐	☐	☐	☐	☐	2
3	☐	☐	☐	☐	☐	☐	☐	3
4	☐	☐	☐	☐	☐	☐	☐	4
5	☐	☐	☐	☐	☐	☐	☐	5
6	☐	☐	☐	☐	☐	☐	☐	6
7	☐	☐	☐	☐	☐	☐	☐	7
8	☐	☐	☐	☐	☐	☐	☐	8
9	☐	☐	☐	☐	☐	☐	☐	9

In this section **no** changes or modifications must be made!

Scrambling

0 0

Type: 036

Exam ID: 25050100001

Please mark the boxes carefully: ☒ Not marked: ☐ or ■

This document is scanned automatically. Please keep clean and do not bend or fold. For filling in the document please use a **blue or black pen**.

Only clearly marked and positionally accurate crosses will be processed!

Answers 1 - 15

	a	b	c	d	e
1	☐	☐			
2	☐	☐	☐	☐	☐
3	☐	☐	☐	☐	☐
4	☐	☐	☐	☐	☐
5	☐	☐	☐	☐	☐
6	☐	☐	☐	☐	
7	☐	☐	☐	☐	☐
8	☐	☐	☐	☐	
9	☐	☐	☐	☐	
10	☐	☐	☐	☐	☐
11	☐	☐	☐	☐	☐
12	☐	☐	☐	☐	
13	☐	☐	☐	☐	☐
14	☐	☐	☐	☐	
15	☐	☐	☐	☐	☐

Answers 16 - 30

	a	b	c	d	e
16	☐	☐	☐	☐	☐
17	☐	☐	☐	☐	☐
18	☐	☐	☐	☐	☐
19	☐	☐	☐	☐	☐
20	☐	☐	☐	☐	
21	☐	☐	☐	☐	☐
22	☐	☐	☐	☐	☐
23	☐	☐	☐	☐	
24	☐	☐	☐	☐	☐
25	☐	☐	☐	☐	☐
26	☐	☐	☐	☐	☐
27	☐	☐	☐	☐	☐
28	☐	☐	☐	☐	
29	☐	☐	☐	☐	☐
30	☐	☐	☐	☐	☐

Answers 31 - 36

	a	b	c	d	e
31	☐	☐	☐	☐	☐
32	☐	☐	☐	☐	☐
33	☐	☐	☐	☐	☐
34	☐	☐	☐	☐	☐
35	☐	☐	☐	☐	☐
36	☐	☐	☐	☐	☐

For each question, select the best answer. (For numeric questions, this is the answer that is closest.)

1. *(0 points)* On this exam you may use one 8.5 × 11 sheet of notes (that you have personally prepared) and a graphing or scientific calculator (without a keyboard or internet access)—but no other resources. You can use the back of each page as scratch paper.

   Mark the box for this question to indicate that you understand this and will complete this exam on your own, without assistance from anyone else.

   a. I agree
   b. I do not agree (your exam will not be graded)

2. *(1 point)* You're modeling a time series of Target's quarterly revenue (in billions of dollars) over time. Your first attempt is **Model T1** with the following estimated ceofficients:

```
T1 <- lm(log(Revenue) ~ Quarter + Period, data = tgt)
```

```
##             Estimate CI.Lower CI.Upper p.value
## (Intercept)    3.009    2.944    3.073 <0.0001
## QuarterQ2     -0.227   -0.300   -0.154 <0.0001
## QuarterQ3     -0.209   -0.281   -0.137 <0.0001
## QuarterQ4     -0.251   -0.323   -0.179 <0.0001
## Period         0.053    0.048    0.057 <0.0001
```

   What is the appropriate interpretation of the trend in Model T1?

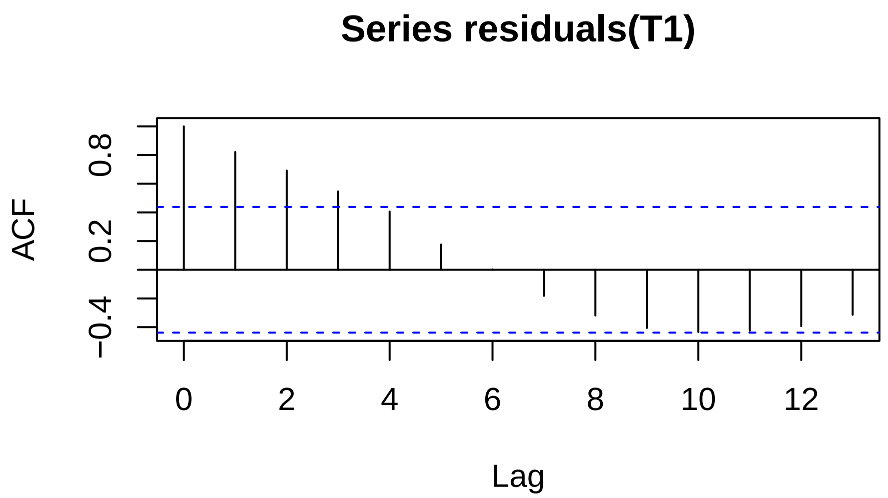
   a. Revenue increases by about 5.4% every quarter
   b. Revenue increases by about $0.212 billion every year
   c. Revenue increases by about $0.053 billion every quarter
   d. Revenue increases by about 23.6% every year
   e. Revenue increases by about 5.4% every year

3. *(1 point)* According to Model T1, after accounting for the trend in revenue, which quarter has the *lowest* expected revenue?

   a. Q3
   b. Q1
   c. Impossible to say without more information
   d. Q4
   e. Q2

4. *(1 point)* Model T1 uses log transform of revenue. What is the *most likely* reason why?

   a. The series exhibits strong seasonality
   b. The variation around the trend in the series gets larger over time
   c. Diagnostic plots suggested the residuals were not normally distributed
   d. Diagnostic plots suggested the residuals were not independent over time
   e. Quadratic or cubic trend models are necessary to satisfy the linearity assumption



5. *(1 point)*

   The autocorrelation plot of the residuals for Model T1 above indicates that

   a. The lag-1 autocorrelation is large but not statistically significantly different from zero
   b. The residuals are dependent over time
   c. The residuals don't have equal variance
   d. The residuals are (approximately) independent over time
   e. We should use a linear model instead of log-transforming Revenue

6. *(1 point)* You also considered **Model T2** below:

```
T2 <- lm(log(Revenue) ~ Quarter + Period + lag(log(Revenue)), data = tgt)
```

```
##                   Estimate CI.Lower CI.Upper p.value
## (Intercept)          1.613   0.5425    2.683   0.006
## QuarterQ2           -0.210  -0.2641   -0.156 <0.0001
## QuarterQ3           -0.170  -0.2248   -0.115 <0.0001
## QuarterQ4           -0.334  -0.4366   -0.232 <0.0001
## Period               0.028   0.0072    0.049   0.012
## lag(log(Revenue))    0.495   0.1033    0.886   0.017
```

   The last observation in the time series is given below:

```
##        Date Period Quarter Revenue
## 20 10/31/22     20      Q2    45.5
```

   Predict the value for Target's revenue in the **next** quarter.

   a. 3.89
   b. 50.44
   c. 3.92
   d. 49.05

7. *(1 point)* The recruiting group at an HR department of a large corporation is trying to advise the company on how to set salary offers so that candidates who are offered mid-level roles at the company will be more likely to accept the offer and join the company. They have a large data set with information about job offers made to candidates for various mid-level jobs at the company, including these variables:

- `offer`, the salary offered, in *thousands of dollars* (in other words, `offer` is 50 when the offered salary was $50,000)
- `prevsalary`, the salary the candidate had at their previous job, again in thousands of dollars
- `accept`, whether the candidates accepted the offer (1) or not (0)

Consider Model L1:

```
L1 <- glm(accept ~ offer, family = binomial, data = joboffers)
```

```
##             Estimate CI.Lower CI.Upper p.value
## (Intercept)   -11.53    -12.4   -10.64 <0.0001
## offer           0.22      0.2     0.23 <0.0001
```

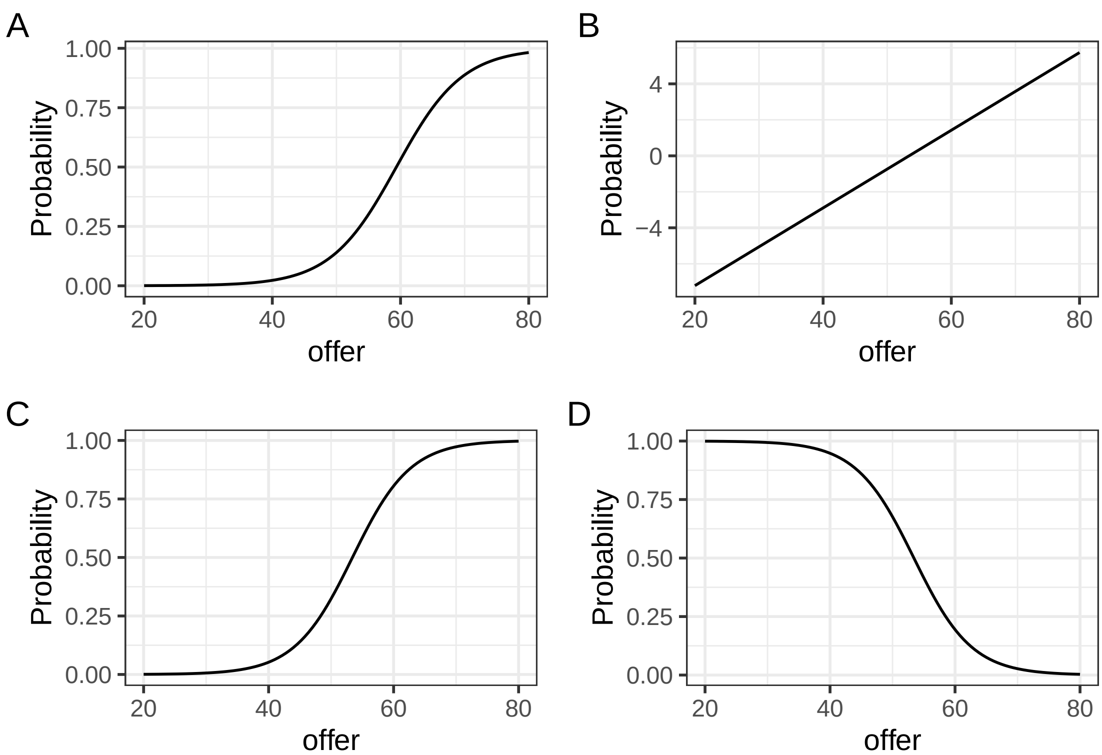
Model L1 indicates that increasing the offer by $1,000 increases the odds of acceptance by:

a. -11.53
b. $<$ 0.0001
c. 0.22
d. 24%
e. 124%

8. *(1 point)* Use Model L1 to predict the odds that a candidate offered a salary of $62,000 will accept the offer.

a. 2.11
b. 0.89
c. 8.25
d. None of these

9. *(1 point)* Which graph depicts the predicted probabilities from Model L1?



a. Graph A
b. Graph B
c. Graph C
d. Graph D

10. *(1 point)* Consider the following output:

```
predicted <- ifelse(predict(L1, type = "response") >= 0.5, 1, 0)
xtabs(~ predicted + accept, data = joboffers)

##          accept
## predicted    0    1
##         0 2209  824
##         1  376  591
```

Which of the following statements is/are TRUE?

I. The accuracy is 70%.
II. The accuracy of a “null” or “no-brainer” model is 35%.
III. This classification procedure allows us to predict whether a candidate will accept a job offer better than guessing.

a. I only
b. II only
c. III only
d. I and III only
e. I, II, and III

11. *(1 point)* Using the same confusion matrix as above, what is the **true positive rate** of this model’s classifications?

a. 15%
b. 42%
c. 61%
d. 85%
e. None of these

12. *(1 point)* Consider the following output, also from Model L1:

```
predicted <- ifelse(predict(L1, type = "response") >= ___, 1, 0)
xtabs(~ predicted + accept, data = joboffers)

##          accept
## predicted    0    1
##         0 2521 1204
##         1   64  211
```

What can you conclude about the missing number in the blank above?

a. It must be less than 0.5.
b. It must be equal to 0.5.
c. It must be greater than 0.5.
d. Nothing can be concluded about the missing number.

13. *(1 point)* Now consider Model L2:

```
L2 <- glm(accept ~ offer * prevsalary, family = binomial, data = joboffers)
```

```
##                    Estimate CI.Lower CI.Upper p.value
## (Intercept)        -11.6686 -23.6843    0.426   0.058
## offer               -0.0771  -0.3141    0.157   0.521
## prevsalary          -0.0807  -0.3182    0.154   0.502
## offer:prevsalary     0.0074   0.0028    0.012   0.002
```

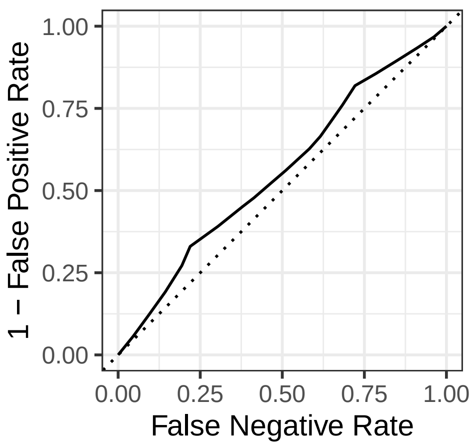
Model L2 indicates that

a. a candidate's previous salary impacts how important the salary offer is in determining how likely a candidate is to accept the offer; the salary offer has a bigger impact on likelihood of acceptance for candidates who made more money at their last job.
b. a candidate's previous salary impacts how important the salary offer is in determining how likely a candidate is to accept the offer; the salary offer has a bigger impact on likelihood of acceptance for candidates who made less money at their last job.
c. there is a negative relationship between how much money a candidate made at their last job and the salary offer they receive.
d. there is a negative relationship between how much money a candidate made at their last job and their likelihood of accepting the salary offer.
e. there is a negative relationship between the salary offer the candidate receives and their likelihood of accepting the offer.

14. *(1 point)* Model L2 indicates that

a. increasing the offer by $1,000 is associated with a decrease of 0.08 in the odds of acceptance.
b. increasing the offer by $1,000 is associated with a decrease of 93% in the odds of acceptance.
c. increasing the offer by $1,000 is associated with a decrease of 7% in the odds of acceptance.
d. None of these

15. *(1 point)* Consider the following ROC curve from a logistic regression model:



The AUC for this model could be

a. 0.38
b. 0.44

c. 0.5
d. 0.56
e. 1

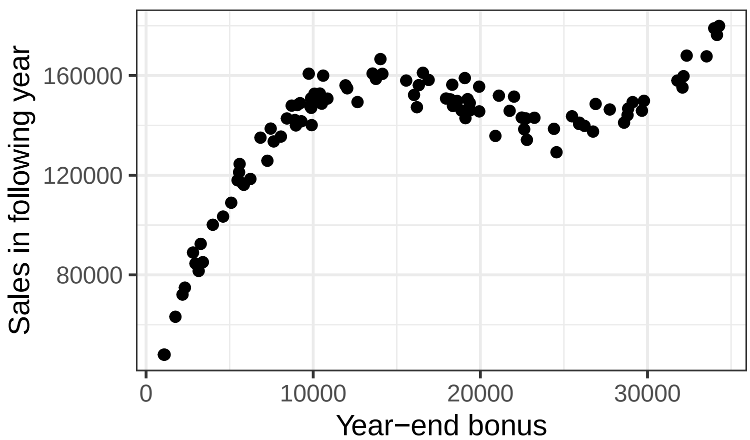
16. *(1 point)* Two models were built with the goal of predicting user engagement on social media posts. Model A had 5 predictors, while Model B had 15 predictors. The RSE and the average RMSPE from 5-fold cross validation are shown in the table below.

Model	RSE	5-fold CV
Model A	0.771	0.831
Model B	0.630	0.831

Which statement below is correct?

a. Since both models have the same 5-fold CV RMSPE, it doesn't matter which model we choose.
b. We would prefer Model A to Model B because while they have the same 5-fold CV RMSPE, Model A has fewer predictors and is less likely to overfit.
c. We would prefer Model B to Model A because while they have the same 5-fold CV RMSPE, Model B has a lower RSE.
d. We would prefer Model B to Model A because while they have the same 5-fold CV RMSPE, Model A has a higher RSE.
e. We shouldn't use either model as the test error is higher than the training RSE for both models, indicating we could do better with an even simpler model.

17. *(1 point)* A Vice President of Sales is trying to determine how a year-end bonus can predict a salesperson's sales in the following year. Graphing the data, it looks like this:



Suppose they split up their into a training and test set and fit models P1, P2, P3, P4, etc. (all the way up to P10) using the training set:

```
P1 <- lm(sales ~ bonus, data = training)
P2 <- lm(sales ~ bonus + I(bonus^2), data = training)
P3 <- lm(sales ~ bonus + I(bonus^2) + I(bonus^3), data = training)
P4 <- lm(sales ~ bonus + I(bonus^2) + I(bonus^3) + I(bonus^4), data = training)
...
```

What do you expect will happen to the training error and the test error as the degree of the polynomial is increased?

a. Training error increases, but test error decreases.

b. Training error decreases; test error decreases and then increases.
c. Both training and test error increase.
d. Both training and test error decrease.
e. None of the above.

18. *(1 point)* A data set was split into 80% training set, 20% test set as below. The model was built using the training set. What is the out of sample performance of this model on the test set, as measured by RMSPE?

Obs	Set	Y	$\hat{Y}$
1	Training	5	5.5
2	Training	2	3
3	Test	6	10
4	Training	3	4.5
5	Training	2.5	4

a. 2.1
b. 10
c. -4
d. 4
e. 2

19. *(1 point)* Which of the following questions is not suitable for answering with a traditional randomized controlled trial?

a. Does a behavioral therapy program reduce symptoms of depression compared to a placebo?
b. What is the effect of a new drug on reducing blood pressure in adults?
c. What is the impact of long-term exposure to environmental pollution on the development of chronic diseases over several decades?
d. Does a specific surgical technique improve patient survival rates compared to the standard technique?
e. All questions are suitable for an RCT.

20. *(1 point)* Suppose we want to test the efficacy of a new way of teaching STA 235 using a flipped classroom model. There are 450 students in total, and based on performance in STA 301, students are grouped into one of three prior performance blocks (Low, Medium, High). Within each block, 75 were randomly assigned to the flipped model (Treatment) and 75 to the standard method (Control). Our outcome variable is the final grade in STA 235 (all assessments were kept the same). The results of a model regressing final score on treatment, block, *AND* an interaction between the two can be seen below:

```
Call:
lm(formula = FinalScore ~ Treatment * Block, data = your_data)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             65.200      1.500    43.47  < 2e-16 ***
Treatment                5.500      2.121     2.59   0.0106 *
BlockMedium              8.100      2.121     3.82   0.0002 ***
BlockHigh               15.300      2.121     7.22  1.8e-11 ***
Treatment:BlockMedium   -2.300      3.000    -0.77   0.4420
Treatment:BlockHigh      4.100      3.000     1.37   0.1729
```

Which of the following statements are TRUE?

I. The Average Treatment Effect of the flipped classroom on final grades for all students is 5.5 points.
II. We cannot be confident that there is a difference in the impact the flipped model has on final grades between the high performance block and the low performance block.
III. On average, we expect students in the high performance block to benefit more from the flipped model than those in the low or medium blocks.

a. I and II
b. I and III
c. II and III
d. I, II, and III

21. *(1 point)* To determine the effect of "None of the above" as an answer choice on student psychology and performance on exams, two variants of the exam were generated, one with 5 multiple choice distractors (Treatment=0), and another with 4 distractors and "None of the above" (Treatment=1). The answers were identical across exams (meaning neither the fifth distractor nor "None of the above" were ever the correct answers). What is the average treatment effect of using "None of the above" as a distractor on exam performance, based on the data given below?

Subject	Treatment	Score
A	1	7.5
B	1	10.5
C	1	14.5
D	0	18

a. 3.62
b. -7.17
c. 14.5
d. 12.62
e. 10.83

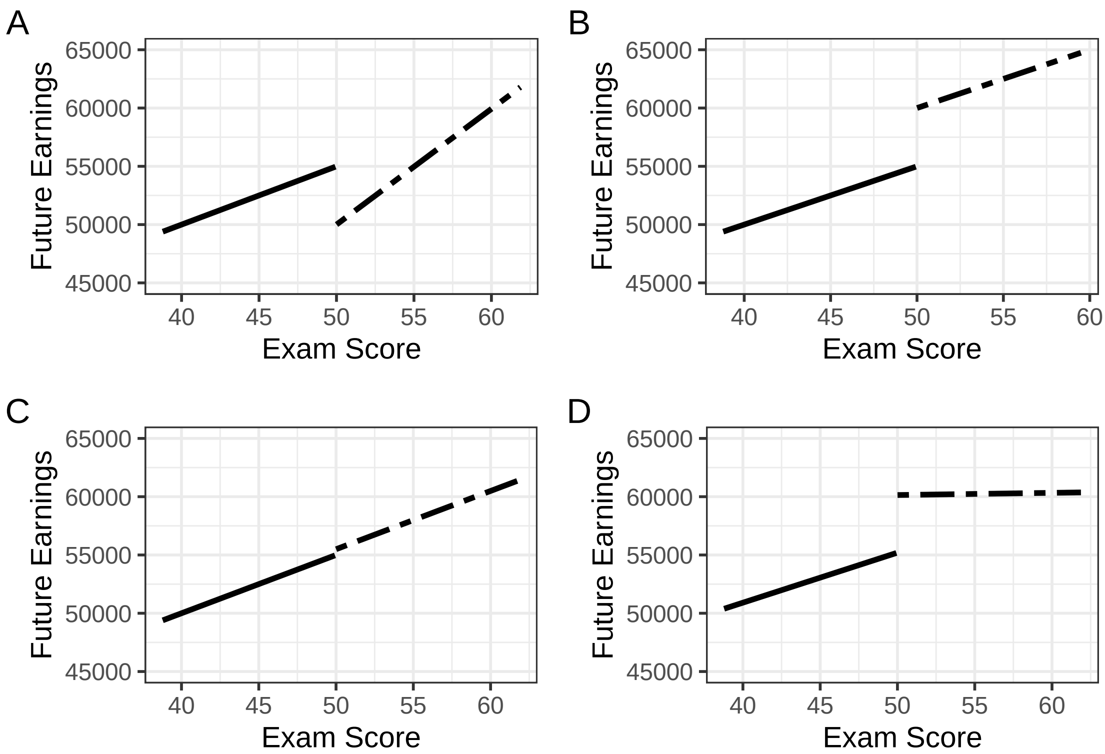
22. *(1 point)* Researchers want to analyze the impact of the Stats Slugger scholarship program on future earnings of STA 235 students. Students who scored above a 50 on the final exam received the scholarship. Below is the output of a regression discontinuity design run on data collected from 200 past students.

```
##                             Estimate CI.Lower CI.Upper p.value
## (Intercept)                    55347    54368    56326 <0.0001
## Scholarship                     5239     3908     6570 <0.0001
## ScoreCentered                    492      283      702 <0.0001
## Scholarship:ScoreCentered       -493     -784     -202   0.001
```

What is the estimated impact of the Stats Slugger Scholarship program on future earnings?

a. $-1
b. $492
c. $4746
d. $5239
e. $-493

23. *(1 point)* Which graph depicts the RDD model fit above?



   a. Graph A
   b. Graph B
   c. Graph C
   d. Graph D

24. *(1 point)* Which of the following is a key assumption required for a valid Regression Discontinuity Design (RDD) analysis?

   a. The scholarship must have been randomly assigned across all students.
   b. Students around the cutoff are comparable.
   c. Exam scores must have a strong correlation with future earnings.
   d. Future earnings must be normally distributed.
   e. The true effect of the scholarship on future earnings cannot be zero.

25. *(1 point)* Which of the following statements about RDD and DiD is true?

   a. DiD requires a running variable with a strict cutoff, while RDD does not.
   b. RDD compares trends over time across groups, while DiD compares individuals at a cutoff.
   c. RDD requires panel data over multiple periods, while DiD does not.
   d. DiD assumes individuals are randomly assigned to treatment based on a cutoff.
   e. None of the above.

26. *(1 point)* **The remaining questions concern the "Loyalty Program Impact Analysis" handout.**

   The average difference in spending post-enrollment between customers who enrolled in the loyalty program and those who did not is about

a. \$385
b. \$134
c. \$53
d. \$234
e. \$887

27. *(1 point)* The difference in the estimated effect of enrollment between models M1 and M2 is *most likely* due to

    a. Collider bias, since we're only looking at data on customers who were eligible for enrollment
    b. Customers who bought more in the past are more likely to buy in the future and to enroll in the loyalty program
    c. Overfitting, since this model includes several additional variables
    d. Random variability, since the dataset doesn't contain all customers
    e. Multicollinearity between past spend and past purchases

28. *(1 point)* Which statement about `discount_code` is best supported by the problem description and model output?

    a. We shouldn't include it in our model because simpler models are better
    b. We should omit it from our models because we don't have statistically significant evidence that it predicts spending above and beyond the other variables
    c. Since receiving the discount code could also drive engagement we should only include one or the other in any model
    d. It predicts spending above and beyond the other variables, so we should include it in the model since it isn't influenced by enrollment

29. *(1 point)* Which statement comparing models M2 and M3 is best supported by the model output and problem statement?

    a. We can't make an educated choice between the two models without estimates of their out-of-sample prediction error, e.g. from cross-validation.
    b. We should prefer M2 to M3 because it includes fewer variables and less subject to multicollinearity/overfitting concerns
    c. We should prefer M3 to M2 because it controls for more potential confounders
    d. We should prefer M2 to M3 because the loyalty program might improve spending by increasing engagement
    e. We should prefer M3 to M2 because it has higher predictive accuracy

30. *(1 point)* Your boss asks for an analysis of the effect of the loyalty program on engagement among customers who made at least one purchase any time after the enrollment period. What's the *biggest* problem with their request?

    a. The sample is smaller after we restrict to post-enrollment purchasers
    b. In this analysis, purchasing behavior after the enrollment period would be a confounder
    c. There is multicollinearity between enrollment and engagement
    d. In this analysis, purchasing behavior after the enrollment period would be a collider
    e. Your boss loves busywork

31. *(1 point)* The company is considering revising what it charges for the loyalty program. Based on the analysis presented, which of the options below is the *most* the company should spend on the loyalty program (per customer) if it wants to at least break even?

a. \$0
b. \$100
c. \$225
d. \$350
e. \$50

32. *(1 point)* The company decides to do a randomized controlled trial of the loyalty program instead, by offering a randomly selected subset of customers free enrollment in the loyalty program. Consider the following potential problems with this study:

    I. Since customers in the trial didn't pay for the loyalty program they might use it differently than customers who pay
    II. The customers who are chosen from the program might be systematically different than customers who aren't
    III. Some customers might not enroll in the program even if it's free

    Which of these are a threat to the **external validity** of *this* RCT?

    a. I
    b. II
    c. III
    d. I and III
    e. None of these are threats to external validity

33. *(1 point)* Which of items I, II, and III are are a threat to the **internal validity** of *this* RCT?

    a. I
    b. II
    c. III
    d. I and III
    e. None of these are threats to internal validity

34. *(1 point)* Which of the following would be an example of *blocking* in this RCT?

    a. Grouping participants by a characteristic like prior spending before randomizing within each group to reduce variability.
    b. Using a larger sample size to reduce variability.
    c. Offering enrollment in the loyalty program to the groups most likely to enroll
    d. Including confounders when analyzing the experiment with a regression model
    e. Selecting customers at random to be offered

35. *(1 point)* The company also has historical data from the rollout of the loyalty program, which was introduced in different regions at different times during a period when the region-specific discount codes weren't active. They want to assess the effect of the loyalty program on revenue using a DiD approach comparing the North region to the South region across two time periods, where the South got the loyalty program in the second period and the North didn't get the loyalty program. What is the key assumption behind this analysis?

    a. Time-varying factors influencing the company's revenue were the same in the North and South regions.
    b. Revenue in the North is roughly the same in both periods.
    c. The South region was randomly selected to receive the loyalty program.
    d. In the "before" period, revenue is roughly the same in the North and South.

e. The North and South regions are comparable on factors that drive the company's revenue.

36. *(1 point)* The company found that before the program was introduced in the South, revenue per store was \$8609 in the North and \$9386 in the South. After the program was introduced in the South, revenue per store was \$10227 in the South; during the same time period (i.e., after the program was introduced in the South), revenue per store was \$8953 in the North.

    What is your best estimate of the causal impact of the loyalty program on revenue?

    a. 841
    b. 1274
    c. 777
    d. 497
    e. The causal impact cannot be calculated in this scenario.