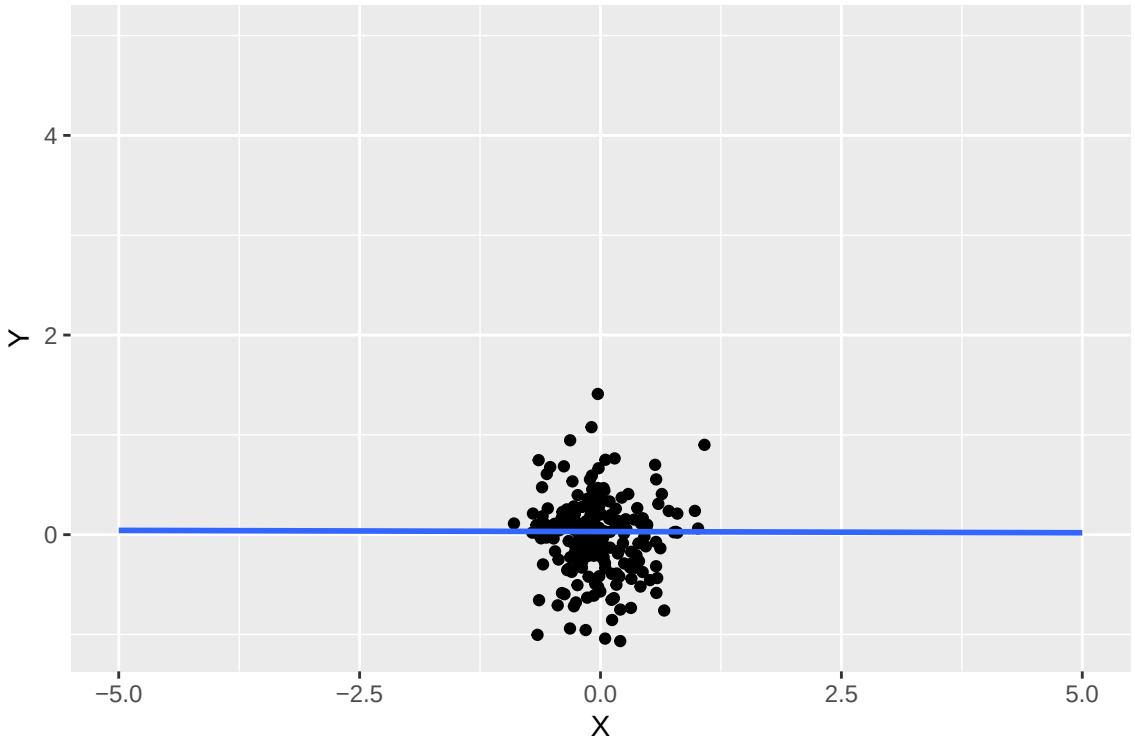
Math 455 **Final Exam** **Spring 2025**

No books, no notes, no calculators.

Name: _____

Page	Points	Score
2	14	
3	16	
4	4	
5	24	
6	18	
7	13	
8	22	
9	16	
10	16	
11	8	
12	20	
13	18	
14	16	
15	20	
16	12	
Total:	237	

1. You fit a linear regression in R with `lm(Y~ X)`, involving two variables, `X` and `Y`, and 197 rows of data. The regression model is $\texttt{Y} = \beta_0 + \beta_1 \texttt{X} + \epsilon$. The scatter plot with the regression line looks like this:



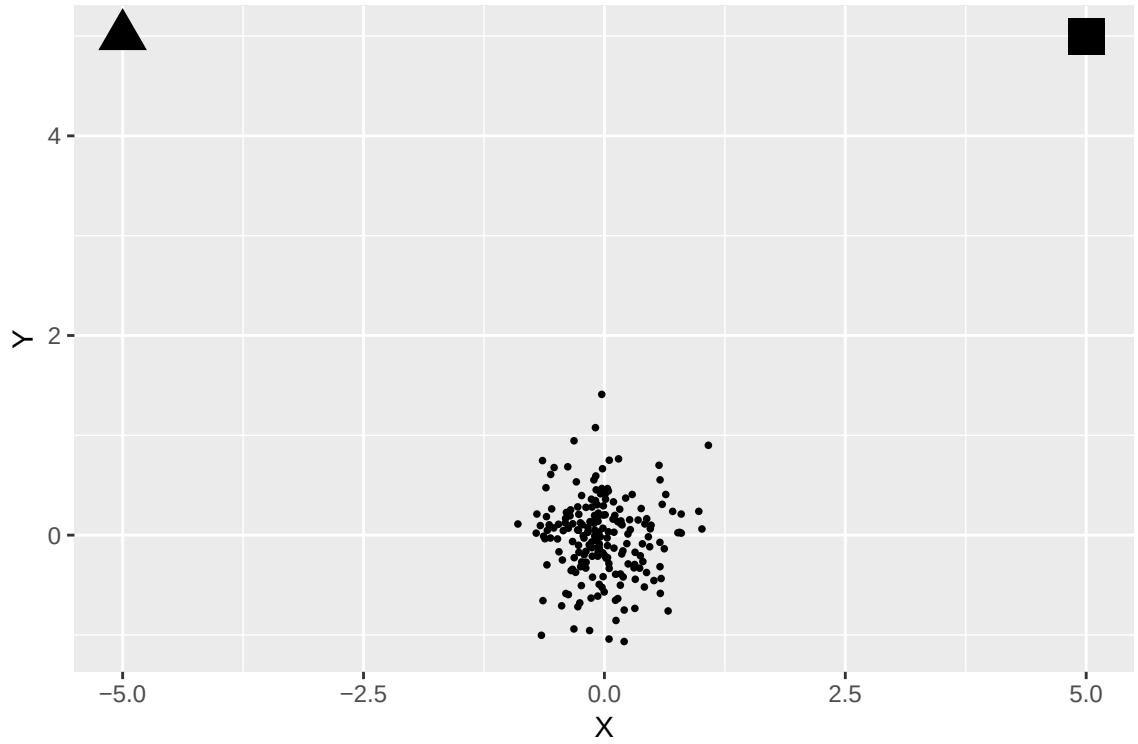
Answer the following questions about this regression:

(a) (4 points) What approximate values would you expect for the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ in the regression output, based on the graph above?

(b) (4 points) What approximate value would you expect for R^2 in the regression output?

(c) (6 points) Would you expect the t-statistic for $\hat{\beta}_0$ to be marked for significance at the 0.05 level (or better)? Why or why not?

Two points, rows 198 and 199 in our data (marked with a square and triangle in the scatter plot) are added to the data set, so that the scatter plot looks like this:

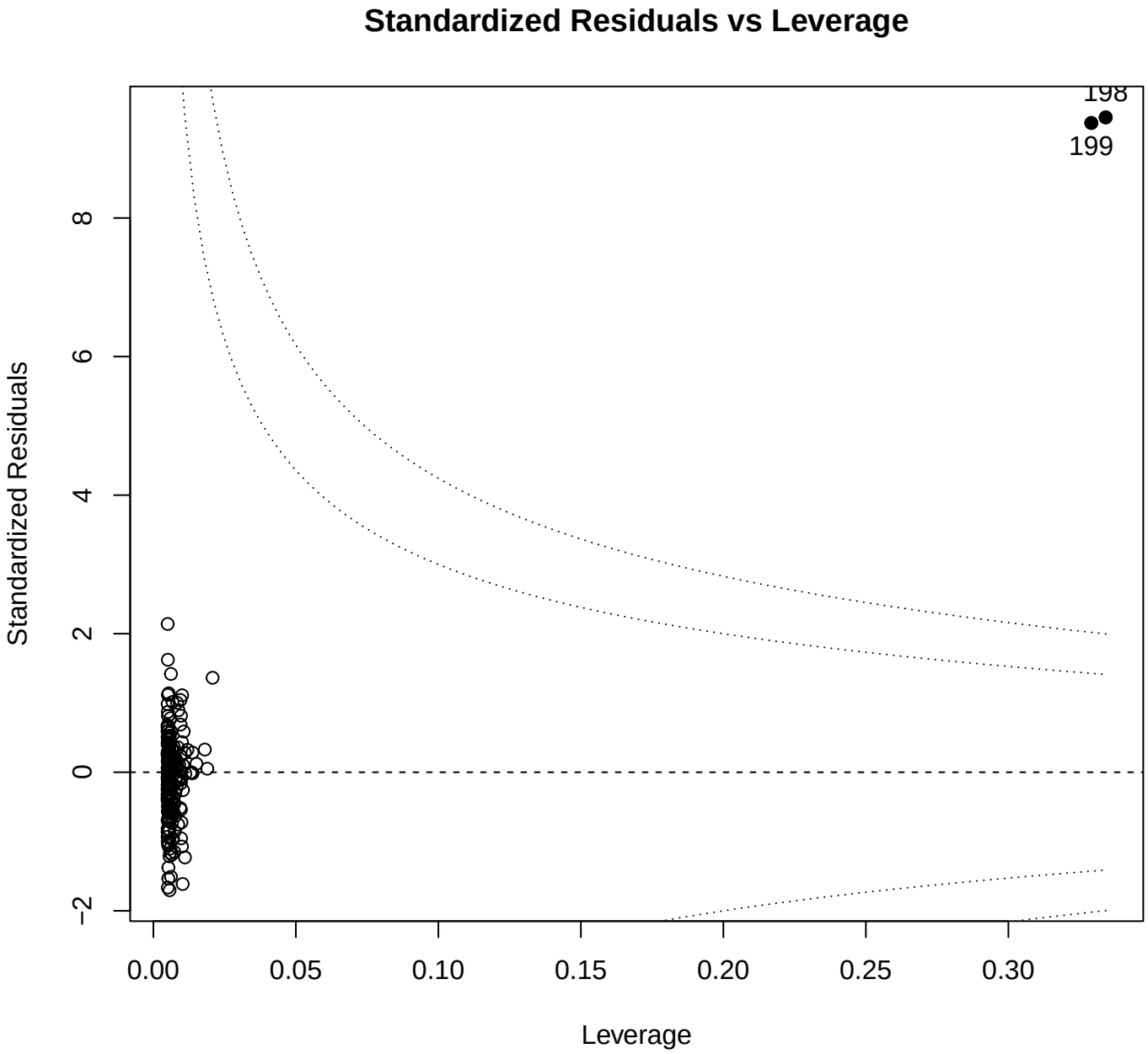


(d) (4 points) If we redo the regression with the point marked with a triangle (but *not* the point marked by the square) added to the data set, what approximate values would you expect for the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ in the regression output, based on the graph above?

(e) (6 points) In this new regression, with the point marked by the triangle (but *not* the point marked by the square) added, would you expect the t-statistic for $\hat{\beta}_1$ to be marked for significance at the 0.05 level (or better)? Why or why not?

(f) (6 points) If we redo the regression *again* with *both* the square *and* triangle points added to the data set, what approximate values would you expect for the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ in the regression output, based on the graph above?

The fourth diagnostic plot for the regression *with the two points added* (both the square and the triangle) is shown below:



(g) (2 points) For purposes of this part, we will say that a point is *influential* if it has Cook's distance greater than 0.5. Is the point marked by the square influential? How about the point marked by the triangle? (Only answers are necessary.)

(h) (2 points) Has the addition of both these points (square and triangle) meaningfully changed the regression line? (Only a yes or no answer is necessary; you may refer to your answer to (f).)

(i) (10 points) Explain any conflict or connection between your answers to the two previous parts.

2. Assume that you have a linear regression model for the resale price of used laptops in terms of age in years, and battery wear level (measured as a percentage, where 0% means a new battery and 100% means completely worn out). So, a brand new laptop is 0 years old with 0% battery wear, while a 2-year-old laptop with 40% battery wear has values of 2 and 40, respectively. The model is given by the equation:

$$\text{price} = \alpha_0 + \alpha_1 \cdot \text{wear} + \alpha_2 \cdot \text{age} + \epsilon,$$

where ϵ is a normally distributed error. You fit this model and obtain coefficients $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2$.

You know two common-sense facts about used laptop prices:

Laptops that are older and/or have higher battery wear tend to have lower prices.

Older laptops tend to have more battery wear, and vice versa.

(a) (6 points) What sign(s) do you expect $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2$ to have? (Only answers are necessary.)

(b) (4 points) You now fit a new model with age omitted, i.e.:

$$\text{price} = \beta_0 + \beta_1 \cdot \text{wear} + \epsilon,$$

and obtain coefficients $\widehat{\beta}_0, \widehat{\beta}_1$.
Do you expect that $\widehat{\beta}_1 \approx \widehat{\alpha}_1$, $\widehat{\beta}_1 < \widehat{\alpha}_1$, or $\widehat{\beta}_1 > \widehat{\alpha}_1$? (Only an answer is necessary.)

(c) (4 points) Give a reason for your answer in the previous part.

3. Suppose we have a data set with five predictors, X_1 = Age, X_2 = DrivingExperience (in years), X_3 = HasAccidents (1 if the person has had any accidents in the past 5 years, 0 otherwise), X_4 = the product of Age and DrivingExperience, and X_5 = the product of Age and HasAccidents. The response Y is the person's monthly car insurance premium (in dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 270$, $\hat{\beta}_1 = -1.5$, $\hat{\beta}_2 = -2.0$, $\hat{\beta}_3 = 60$, $\hat{\beta}_4 = 0.05$, and $\hat{\beta}_5 = 1.2$.

   (a) (6 points) Predict the monthly premium (in dollars) for a person who is 30 years old, has 10 years of driving experience, and has had at least one accident in the past 5 years. (You need only write out the equation; it is not required that you do the calculations.)

   (b) (6 points) Which statement is correct, and why?

      i. For a fixed value of Age and DrivingExperience, people with accidents pay less on average than those without.

      ii. For a fixed value of Age and DrivingExperience, people with accidents pay more on average than those without.

      iii. For a fixed value of Age and DrivingExperience, people with accidents pay more on average than those without, provided that the Age is high enough.

      iv. For a fixed value of Age and DrivingExperience, people with accidents pay less on average than those without, provided that the Age is high enough.

   (c) (6 points) True or false: Since the coefficient for the Age-DrivingExperience product term is small, there is very little evidence of an interaction effect. Justify your answer.

4. A sample of data from 100 retail stores across the Midwest and Southeast has the following variables:

   - Y = average customer satisfaction score (out of 100)
   - X_1 = average manager salary per employee (in thousands of dollars per year)
   - X_2 = percentage of employees with at least 5 years of experience
   - X_3 = neighborhood economic index (higher is wealthier)
   - X_4 = average communication skills score of store staff
   - X_5 = average years of formal training of employees

   We run the regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

   and obtain the least-squares estimates

$$\hat{\beta}_0 = 62.5 \quad \hat{\beta}_1 = -1.85 \quad \hat{\beta}_2 = 0.067 \quad \hat{\beta}_3 = 0.112 \quad \hat{\beta}_4 = 0.89 \quad \hat{\beta}_5 = -1.47$$

   (a) (4 points) Why might the sign of $\hat{\beta}_1$ and $\hat{\beta}_5$ be regarded as surprising?

   (b) (6 points) How would you explain the signs of $\hat{\beta}_1$ and $\hat{\beta}_5$ to a non-statistican?

   (c) (3 points) What does VIF stand for?

(d) (4 points) If we find that the VIF of $\hat{\beta}_i$ is $10.3, 11.9, 9.4, 12.1$ for $i = 1, 2, 3, 5$ respectively, what does this indicate?

(e) (6 points) What changes might we make to the model as a result of this information?

5. I collect a set of data ($n = 200$ observations) containing a single predictor and a quantitative response. I divide this data into two parts: the first 100 observations, which I call the training data, and the last 100 observations, which I call the test data. I then fit a linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ to the *training* data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$ to the *training* data.

   (a) (4 points) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the *training* residual sum of squares (RSS) for the linear regression, and also the *training* RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   (b) (4 points) Suppose now that we use the coefficient estimates obtained from the training data to predict the response on the *test* data. Now answer (a) for the RSS obtained from the *test* data.

   (c) (4 points) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the *training* RSS for the linear regression, and also the *training* RSS

for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) (4 points) Suppose now that we use the coefficient estimates obtained from the training data to predict the response on the *test* data. Now answer (c) for the RSS obtained from the *test* data.

6. (8 points) A marketing analyst is studying customer satisfaction using survey data from $n = 50$ customers. Each customer is classified according to 8 different categorical characteristics (such as product type, store region, customer age group, etc.), and each categorical variable has 8 levels.

   (a) How many dummy variables are needed to represent all of these categorical variables in a regression model?

   (b) Why might this be a problem for a linear regression if we only have 50 observations?

7. Suppose the average patient recovery rate at hospitals (`recovery`) depends on two factors: average number of continuing education hours for nurses (`nurseedu`) and average experience level of doctors (`docexp`):

$$\texttt{recovery} = \beta_0 + \beta_1 \texttt{nurseedu} + \beta_2 \texttt{docexp} + \epsilon$$

   Assume that higher values of `recovery` are associated with higher values of `nurseedu` and higher values of `docexp`, and that the standard assumptions of linear regression are satisfied.

   (a) (4 points) What signs do you expect for $\hat{\beta}_1$ and $\hat{\beta}_2$?

(b) (4 points) If additional training programs for nurses have been targeted toward hospitals with less experienced doctors, so that `nurseedu` and `docexp` are negatively correlated, how do you expect the estimate $\hat{\alpha}_1$ obtained from the simple regression of `recovery` on `nurseedu` (`recovery` $= \alpha_0 + \alpha_1$`nurseedu` $+ \epsilon$) to compare with the estimate $\hat{\beta}_1$ from the previous regression? Your answer should be that $\widehat{\beta}_1 \approx \widehat{\alpha}_1$, $\widehat{\beta}_1 < \widehat{\alpha}_1$, or $\widehat{\beta}_1 > \widehat{\alpha}_1$.

8. A researcher fits the following regression model using data on the earnings of individuals:

$$\log(\texttt{wage}) = \alpha + \beta \cdot \texttt{educ} + \gamma \cdot \texttt{exper} + \epsilon$$

Here: `wage` is the hourly wage (in dollars), `educ` is years of education, `exper` is years of labor market experience, and ϵ is the error term.

The estimated coefficients are:

$$\hat{\alpha} = 3 \qquad \hat{\beta} = 0.08 \qquad \hat{\gamma} = 0.02$$

(a) (4 points) Write a sentence interpreting the estimated coefficient of `educ` in the context of this model, comprehensible to a non-statistican.

(b) (4 points) Write a sentence interpreting the estimated coefficient of `exper` in the context of this model, comprehensible to a non-statistican.

(c) (4 points) What is the predicted wage for someone with 12 years of education and 10 years of experience? You need not do the calculation, just write out the formula.

9. Suppose we fit the following linear regression model in R:

```
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
```

and obtain the following output:

```
>summary(lmod)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Now suppose we modify the model by rescaling the response variable and run:

```
lm(formula = I(sr/100) ~ pop15 + pop75 + dpi + ddpi, data = savings)
```

For each of the following output quantities, state whether it changes or does not change, by circling the appropriate answer. If it changes, give the new numerical value (three significant figures are enough).

(a) (2 points) The R^2 **changes** | **does not change**

(b) (2 points) The intercept **changes** | **does not change**

(c) (2 points) The standard error of `pop15` **changes** | **does not change**

(d) (2 points) The p-value of the coefficient of `pop15` **changes** | **does not change**

(e) (2 points) The residual standard error **changes** | **does not change**

Now suppose instead that we keep the original regression equation, but rescale *one of the predictors*:

```
lm(formula = sr ~ pop15 + pop75 + I(dpi/1000) + ddpi, data = savings)
```

For each of the following output quantities, state whether it changes or does not change, by circling the appropriate answer. If it changes, give the new numerical value (three significant figures are enough).

(f) (2 points) The R^2 **changes** | **does not change**

(g) (2 points) The coefficient of `dpi` **changes** | **does not change**

(h) (2 points) The standard error of `dpi` **changes** | **does not change**

(i) (2 points) The p-value of the coefficient of `dpi` **changes** | **does not change**

(j) (2 points) The residual standard error **changes** | **does not change**

10. (8 points) Which of the standard assumptions of linear regression do Generalized Least Squares (GLS) and Weighted Least Squares (WLS) methods allow us to relax? For each item A-D, say whether the corresponding assumption is relaxed by GLS and WLS.

A. The assumption of an approximate linear relationship between the response Y and the predictors $X_1, \ldots, X_p$.

B. The assumption of homoscedasticity.

C. The assumption of independence of errors.

D. The assumption of normality of the errors.

To answer, fill in the table below with "Yes" if the method in the column relaxes the assumption in the the row.

	GLS	WLS
A		
B		
C		
D		

11. (8 points) The following equation represents the effects of departmental budget allocation on subsequent student performance `perf` across universities in the United States:

$$\texttt{perf} = \beta_0 + \beta_T \texttt{share}_T + \beta_R \texttt{share}_R + \beta_S \texttt{share}_S + \epsilon$$

where `perf` is the average standardized student score improvement from freshman to senior year, $\texttt{share}_T$ is the share of teaching expenditures in the total academic budget, $\texttt{share}_R$ is the share of research spending, $\texttt{share}_S$ is the share of student services spending, and $\texttt{share}_A$ is all other academic expenses. All variables are measured in the same fiscal year.

By definition, the four shares add up to one. Other factors affecting `perf`, included in ϵ, might be student demographics, faculty quality, and campus resources.

Notice that $\texttt{share}_A$ was omitted from the regression equation displayed above. Was this a mistake, and why or why not?

12. (10 points) We perform best subset, forward stepwise, and backward stepwise selection on a regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$. For each approach, we obtain $p+1$ models, containing $0, 1, 2, \ldots, p$ predictors. Mark each statement below true or false.

A. The predictors in the k-variable model identified by forward stepwise selection are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

B. The predictors in the k-variable model identified by backward stepwise selection are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

C. The predictors in the k-variable model identified by backward stepwise selection are a subset of the predictors in the $(k+1)$- variable model identified by forward stepwise selection.

D. The predictors in the k-variable model identified by forward stepwise selectionare a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

E. The predictors in the k-variable model identified by best subset selection are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

For the following multiple choice questions, circle the letter corresponding to the most accurate answer. There is no penalty for guessing.

13. (4 points) Which is the best description of the Akaike Information Criterion (AIC) and its use in the context of linear regression model selection? Circle the letter corresponding to the most accurate answer.

    A. AIC is a technique for handling outliers in the data.

    B. AIC is a measure of variable importance in regression; more important variables have lower AIC.

    C. AIC is a metric for assessing multicollinearity; models with higher AIC have more problems with multicollinearity.

    D. AIC is a measure of goodness-of-fit that penalizes model complexity.

14. (4 points) In the context of linear regression, if we are using AIC as the criterion for model selection, is it possible to prefer a model with a lower R^2 value? Circle the letter corresponding to the most accurate answer.

    A. No, a lower R^2 always indicates an inferior model regardless of the AIC.

    B. Yes, AIC always prefers a simpler model, even if more complex models have higher R^2 values.

    C. No, AIC and R^2 are always in agreement about the preferred model.

    D. Yes, only if the model with the lower R^2 has fewer parameters.

15. (4 points) How does Weighted Least Squares (WLS) handle observations with higher variability differently from those with lower variability? Circle the letter corresponding to the most accurate answer.

    A. It assigns lower weights to observations with lower variability.

    B. It assigns lower weights to observations with higher variability.

    C. WLS treats all observations equally.

    D. WLS excludes observations with high variability.

16. (4 points) Which of the following statements is true regarding Ridge Regression?

    A. It performs variable selection by shrinking some coefficients to exactly zero.

    B. It adds a penalty term to the ordinary least squares (OLS) objective function proportional to the sum of the absolute values of the coefficients.

    C. It is particularly useful when dealing with multicollinearity.

    D. It typically results in a model with fewer predictors than OLS.

17. (4 points) When considering the trade-off between bias and variance in model selection, which of the following is true?

    A. Models with more predictors tend to have higher bias and lower variance.

    B. Ridge and Lasso regression introduce bias to reduce variance.

    C. Stepwise selection methods always result in the model with the lowest bias.

    D. A model with low R^2 always has high bias.

18. (4 points) A limitation of the standard Box-Cox procedure is that it:

    A. Cannot be applied if the response variable includes zero or negative values.

    B. Requires the predictors to be normally distributed.

    C. Only considers linear transformations of the response.

    D. Is insensitive to violations of the constant variance assumption.

19. (4 points) In Ridge and Lasso regression, the tuning parameter (λ in the textbook) controls:

    A. The number of observations used in the model fitting.

    B. The strength of the penalty applied to the coefficients.

    C. The significance level for including predictors.

    D. The functional form of the relationship between predictors and response.

20. (4 points) Which of the following statements correctly describes a key difference in the outcomes of Ridge versus Lasso regression when dealing with many predictors?

    A. Ridge tends to shrink coefficients towards zero but keeps most predictors, while Lasso tends to set some coefficients exactly to zero, effectively performing variable selection.

    B. Lasso tends to shrink coefficients towards zero but keeps most predictors, while Ridge tends to set some coefficients exactly to zero.

    C. Ridge always produces a model with fewer predictors than Lasso.

    D. Lasso is computationally more expensive than Ridge.

21. (4 points) When using AIC or BIC for model selection, the criterion includes a term that penalizes the number of parameters (p) in the model. How do the penalties in AIC and BIC typically compare?

    A. BIC applies a larger penalty for each additional parameter than AIC, especially for larger sample sizes (n).

    B. AIC applies a larger penalty for each additional parameter than BIC, especially for larger sample sizes (n).

    C. AIC and BIC apply the same penalty per parameter.

    D. AIC penalizes the number of parameters, while BIC penalizes the sample size.

22. (4 points) The primary purpose of the penalty term in both Ridge and Lasso regression is to:

    A. Increase the bias of the coefficient estimates.

    B. Reduce the variance of the coefficient estimates and prevent overfitting.

    C. Ensure that all predictors have statistically significant coefficients.

    D. Change the functional form of the relationship between variables.

23. (4 points) Which of the following is *not* one of the commonly accepted Hill Criteria for assessing causality?

    A. Strength of association

    B. Consistency

    C. Linearity

    D. Temporality

24. (4 points) How does Robust Regression typically handle outliers or influential observations compared to Ordinary Least Squares (OLS)?

    A. It assigns them a higher weight in the fitting process.

    B. It assigns them a lower weight in the fitting process.

    C. It removes them from the dataset before fitting the model.

    D. It only considers them if their leverage is low.