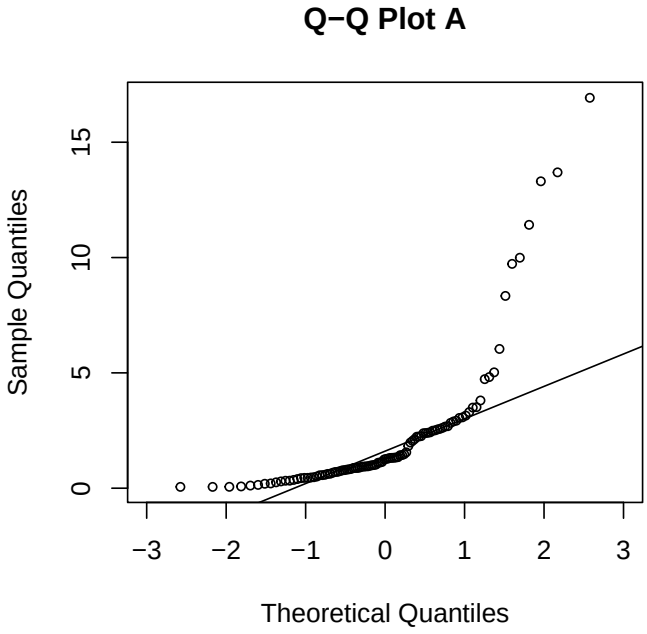
Math 455 **Midterm 2** **Spring 2026**

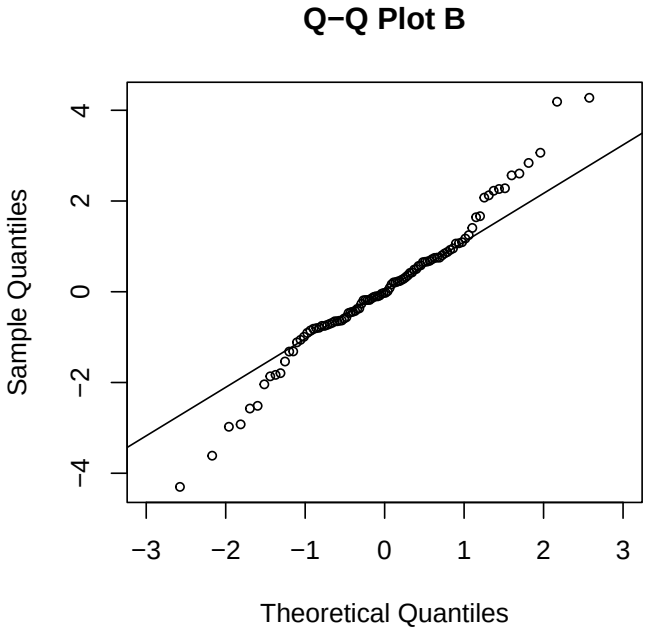
No books, no notes, no calculators.

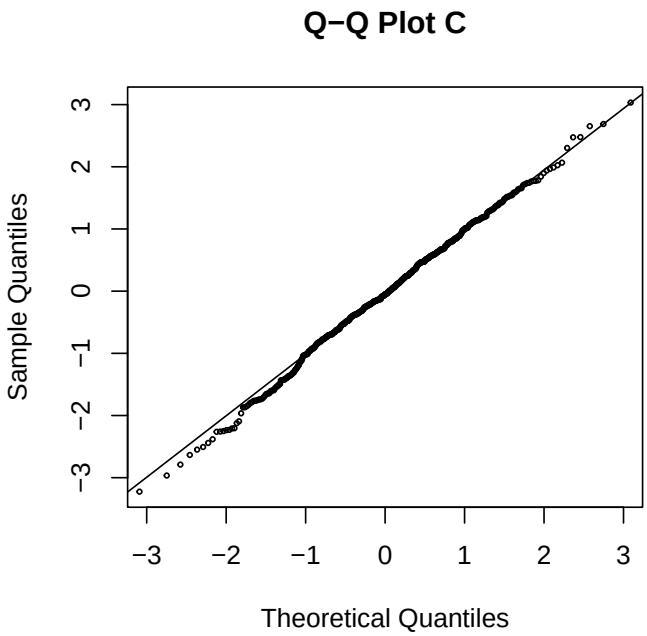
Name: _____

Page	Points	Score
2	4	
3	16	
4	8	
5	16	
6	8	
7	4	
8	8	
9	12	
10	10	
11	14	
12	18	
13	10	
14	26	
15	20	
Total:	174	

1. Each of the 3 Q-Q plots below is constructed by taking independent random samples from a probability distribution and applying the usual procedure to construct a Q-Q plot.







(a) (4 points) Which, if any, of these Q-Q plots shows a distribution with a long (or heavy) right tail?

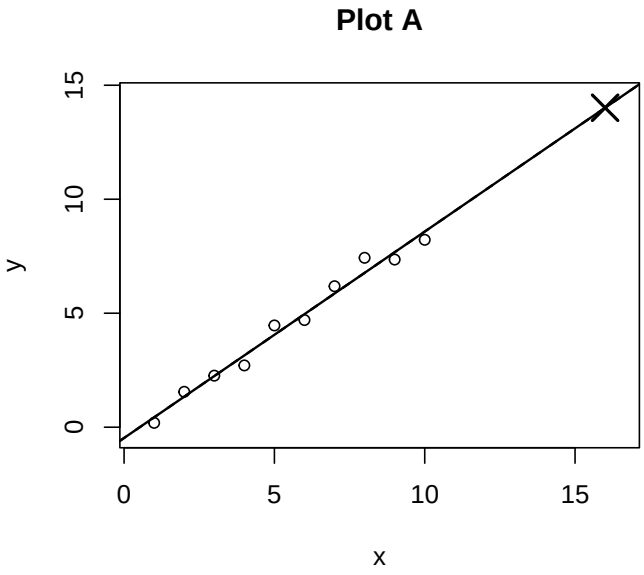
(b) (4 points) Which, if any, of these Q-Q plots shows a distribution with long (or heavy) tails on *both* sides?

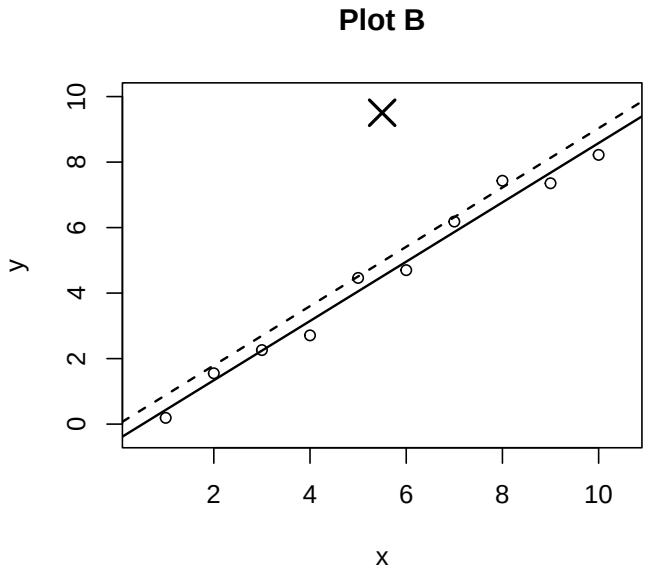
(c) (4 points) Which, if any, of these Q-Q plots is most likely to be constructed by taking samples from a normal distribution?

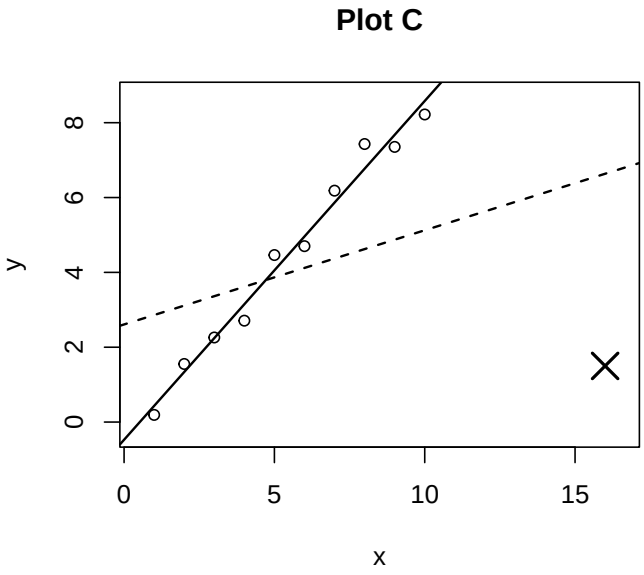
(d) (4 points) Which, if any, of these Q-Q plots is most likely to be constructed by taking samples from an exponential distribution?

(e) (4 points) Which, if any, of these Q-Q plots shows a distribution with a short (or thin) left tail?

2. The three graphs below each show ten data points (circles) and one additional point marked with an X. In each plot, the **solid line** is the regression line fit to the ten circle points *only*, and the **dashed line** is the regression line fit to *all eleven points* (circles plus the X).







(a) (4 points) In which plot or plots is the point marked X an *outlier*? (Only an answer is necessary.)

(b) (4 points) In which plot or plots is the point marked X a *high-leverage* point? (Only an answer is necessary.)

(c) (4 points) In which plot or plots is the point marked X an *influential* point? (Only an answer is necessary.)

(d) (4 points) In plain language, comprehensible to a non-statistician, explain what it means for a data point to be *influential* in a regression.

3. A hospital administrator wants to understand factors that affect patient length of stay (in days). Data are collected on a random sample of $n = 3{,}800$ patients with the following variables:

   - Y: Length of hospital stay (in days)
   - S: Severity score (1 = Low severity, 0 = High severity)
   - I: Insurance type (1 = Private insurance, 0 = Public/no insurance)

   Three models are fit:

   - (Model 1) $Y = \beta_0 + \beta_1 S$
   - (Model 2) $Y = \beta_0 + \beta_1 S + \beta_2 I$
   - (Model 3) $Y = \beta_0 + \beta_1 S + \beta_2 I + \beta_3 SI$

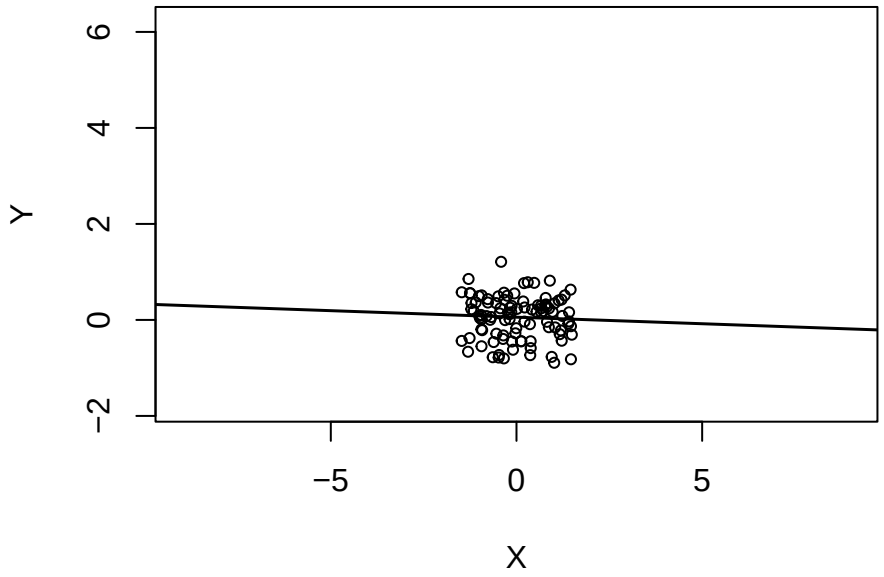
   (a) (4 points) For Model 1, suppose $\hat{\beta}_1 = -2$. Write a sentence interpreting this estimate in plain language, without using the word "coefficient" or "regression."

   (b) (4 points) For Model 2, suppose $\hat{\beta}_1 = -2$ and $\hat{\beta}_2 = -1.5$. What does Model 2 predict as the difference in average length of stay between a low-severity patient with private insurance and a low-severity patient with public/no insurance?

(c) (4 points) For Model 3, suppose $\hat{\beta}_1 = -2$, $\hat{\beta}_2 = -1.5$, and $\hat{\beta}_3 = 1$. What is the predicted change in length of stay associated with switching from high to low severity for a patient *with* private insurance?

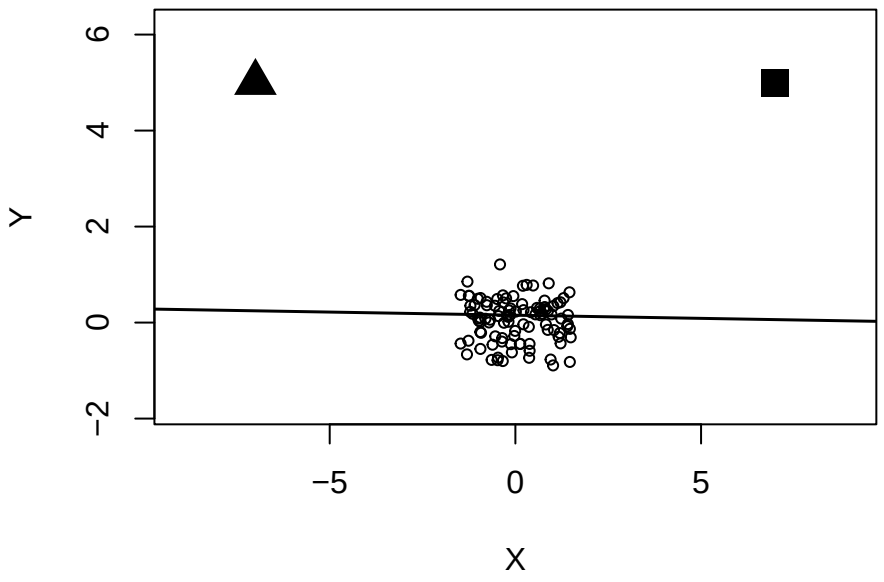
(d) (4 points) Suppose, for Model 3 we have the same coefficient estimates as in the previous part. What is the predicted change in length of stay associated with switching from high to low severity for a patient *without* private insurance?

4. You fit a linear regression in R with `lm(Y~X)`, using 100 rows of data. The scatter plot with the regression line looks like this:



(a) (4 points) Based on the graph above, what approximate values would you expect for the slope $\hat{\beta}_1$ and the intercept $\hat{\beta}_0$? What approximate value would you expect for R^2?

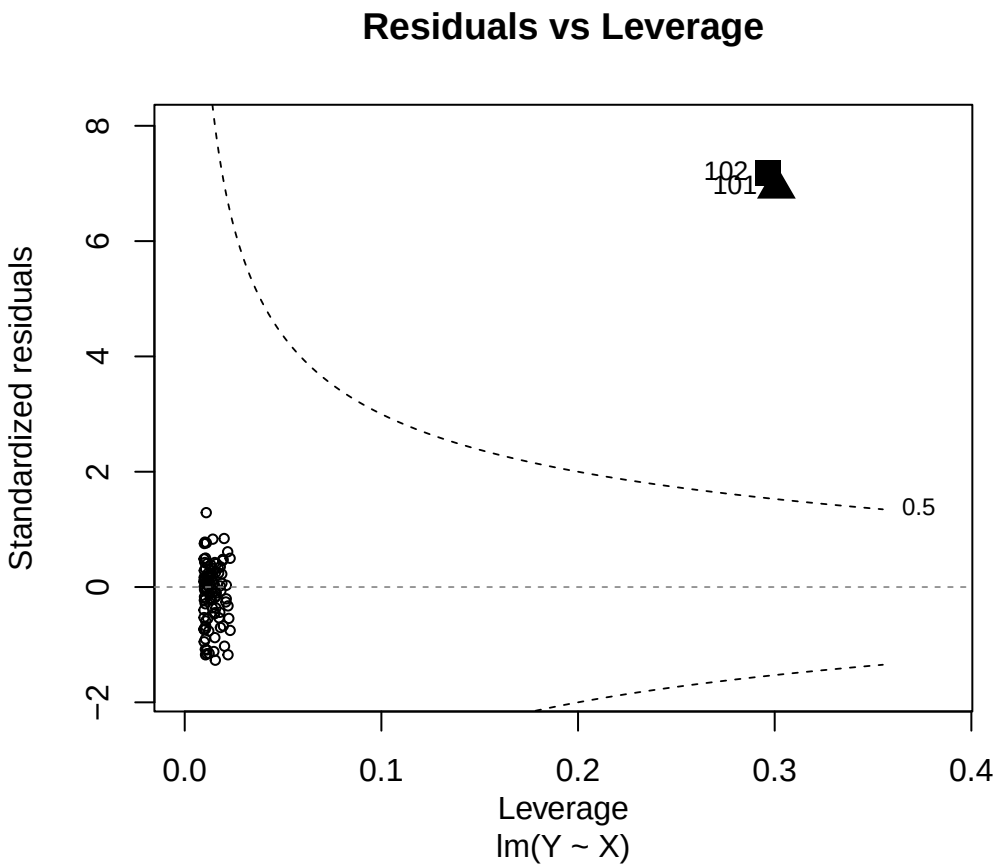
Two additional rows of data (rows 101 and 102) are added to the dataset. Row 101 is marked with a triangle (▲) and row 102 with a square (■) in the scatter plot below:



(b) (4 points) If we refit the regression using only the triangle point added to the original 100 rows (but *not* the square), what approximate values would you expect for $\hat{\beta}_1$ and $\hat{\beta}_0$?

(c) (4 points) If we refit the regression with *both* the triangle and the square added to the original 100 rows, what approximate values would you expect for $\hat{\beta}_1$ and $\hat{\beta}_0$?

The fourth diagnostic plot for the regression *with both points added* (all 102 rows) is shown below:



(d) (4 points) Suppose we say (just for purposes of this question) that a point is *influential* if its Cook's distance exceeds 0.5. Are either or both of the two added points influential? How can you tell from the plot?

(e) (8 points) Is there a conflict between your answers to parts (c) and (d)? Why or why not?

5. Assume the linear regression model

$$Y = 2 + X_1 + 3X_2 + \epsilon$$

with the standard assumptions is exactly true, and that $\text{Cor}(X_1, X_2) = \rho$. Additionally, assume that ϵ is independent of X_1 and X_2, $\text{Var}(\epsilon) = \sigma^2$, and $\text{Var}(X_i) = \sigma_i^2$.

Suppose we do not observe X_2 and instead fit the misspecified simple regression model

$$Y = \beta_0 + \beta_1 X_1 + \epsilon'$$

In the big-data limit the estimated coefficient satisfies

$$\hat{\beta}_1 = \frac{\text{Cov}(Y,\, X_1)}{\text{Var}(X_1)}.$$

(a) (6 points) Compute $\text{Cov}(Y, X_1)$ in terms of the quantities defined above. Show your work.

(b) (4 points) Use your answer from part (a) to express $\hat{\beta}_1$ in terms of ρ, σ_1, and σ_2.

(c) (4 points) Suppose $\rho = -1/2$ and $\sigma_1^2 = \sigma_2^2 = 1$. What is the numerical value of $\hat{\beta}_1$?

(d) (6 points) Is this value the same as the coefficient of X_1 in the true model? Explain in plain language why or why not, in terms comprehensible to a non-statistician.

(e) (4 points) What is the name of this phenomenon?

6. A public health researcher fits the following regression model to data from 100 hospitals:

```
lm(formula = cost ~ beds + nurses + procedures, data = hospital)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.3416    8.1247    6.44   <2e-16 ***
beds           0.4812    0.0653    7.37   <2e-16 ***
nurses        -0.2103    0.1847   -1.14    0.258
procedures     1.8534    0.2419    7.66   <2e-16 ***
---
Residual standard error: 12.43 on 96 degrees of freedom
Multiple R-squared:  0.7621,    Adjusted R-squared:  0.7542
F-statistic: 102.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

Here `cost` is the average daily cost per patient (in hundreds of dollars), `beds` is the number of hospital

beds, `nurses` is the full-time nurse count, and `procedures` is the number of distinct medical procedures offered.

Suppose we replace `cost` by `I(cost * 100)`, converting the response to dollars, and refit. For each quantity below, circle **changes** or **does not change**. If it changes, give the new value to three significant figures.

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

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

(c) (2 points) The standard error of the intercept **changes** — **does not change**

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

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

Now suppose instead that, keeping the original response `cost`, we replace `procedures` by `I(procedures * 10)` and refit. For each quantity below, circle **changes** or **does not change**, and give the new value if it changes.

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

(g) (2 points) The intercept **changes** — **does not change**

(h) (2 points) The coefficient shown for `procedures` in the new output **changes** — **does not change**

(i) (2 points) The p-value shown for `procedures` in the new output **changes** — **does not change**

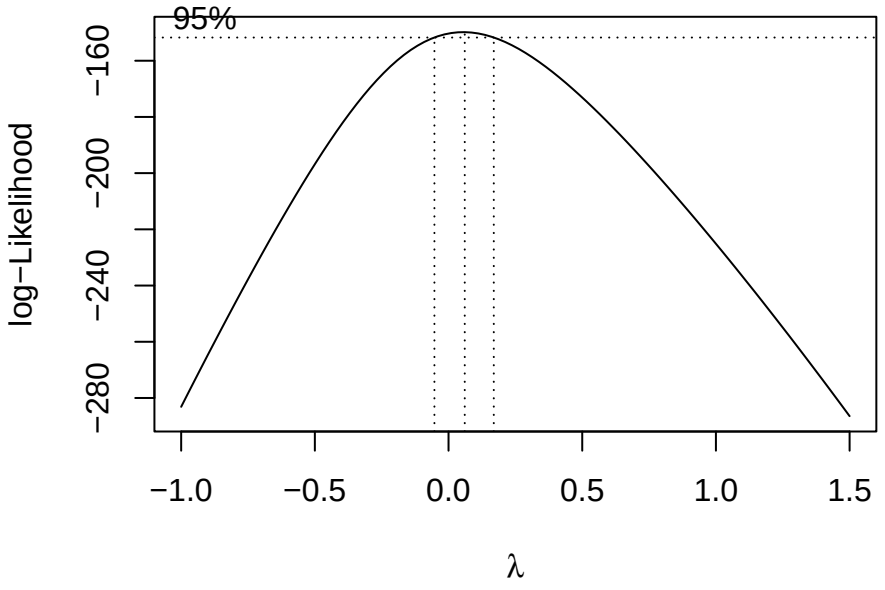
(j) (2 points) The residual standard error **changes** — **does not change**

7. Answer the following questions about the Box-Cox procedure for selecting a transformation in the regression $Y = \beta_0 + \beta_1 X + \epsilon$.

   (a) (1 point) Does the Box-Cox procedure select a transformation of the response Y, or of the predictor X?

   (b) (4 points) Describe the family of transformations considered by the Box-Cox procedure. What is the parameter being estimated, and what does it represent?

   (c) (3 points) What restriction on the data is required for the Box-Cox procedure to be applicable, and why?



(d) (6 points) The Box-Cox log-likelihood plot above has its maximum very close to $\lambda = 0$, and $\lambda = 0$ falls well within the 95% confidence interval. What transformation would you apply, and why?

Multiple Choice: Circle the item corresponding to the best answer. *There is no penalty for guessing.*

8. (4 points) When comparing two nested models, which of the following is the most appropriate formal test for deciding whether to include the additional predictors?

   a) Comparing their Adjusted R^2 values.

   b) An F-test (partial F-test / ANOVA).

   c) Examining each predictor's t-value in the larger model individually.

   d) Selecting the model with the smaller condition number.

9. (4 points) Variance Inflation Factors (VIFs) are used to diagnose:

   a) Non-normality of the residuals.

   b) Non-constant variance of the residuals.

   c) Multicollinearity among the predictors.

   d) Influential observations in the data.

10. (4 points) A data point has very *high leverage* but a *small residual*. This means:

    a) The point has a large Cook's distance and is definitely influential.

    b) The point lies far from the center of the predictor space and is poorly fitted.

    c) The point lies far from the center of the predictor space, but the fitted model passes close to it.

    d) The point is an outlier in Y but not in X.

11. (4 points) Omitting a relevant variable from a regression model will generally cause:

    a) The standard errors of all remaining coefficients to be underestimated.

    b) The remaining coefficient estimates to be biased if the omitted variable is correlated with any included predictor.

    c) The R^2 of the model to increase.

    d) No harm provided the sample size is large enough.

12. (4 points) When severe collinearity is present among the predictors in a regression, which of the following is most likely to occur?

a) The overall F-test becomes insignificant even when several predictors are genuinely related to Y.

b) Individual t-tests may fail to flag predictors as significant even though the predictors are collectively highly significant.

c) The residual standard error increases sharply.

d) The fitted values $\hat{Y}$ become unreliable.

13. (4 points) In a regression with three predictors, the condition number of the design matrix is computed to be $\kappa = 450$. Which conclusion is most appropriate?

a) There is no evidence of collinearity; condition numbers below 500 are always acceptable.

b) There is moderate-to-strong collinearity; the parameter estimates may be unstable.

c) The model has heteroscedastic errors and should be refit with WLS.

d) At least one predictor is a perfect linear combination of the others.

14. (4 points) A researcher fits a regression and finds that all pairwise correlations between the predictors are below 0.4 in absolute value, yet several VIFs exceed 10. The most likely explanation is:

a) This is impossible; low pairwise correlations guarantee low VIFs.

b) There is a near-linear relationship involving *three or more* predictors simultaneously, even though no two predictors are strongly correlated in pairs.

c) The response variable is highly skewed and should be transformed.

d) The sample size is too small to estimate VIFs reliably.

15. (4 points) Which of the following actions is *least* likely to help when collinearity is detected among predictors?

a) Removing one of the correlated predictors from the model.

b) Combining the correlated predictors into a single composite variable (e.g. a principal component).

c) Collecting more data so that the predictors span a wider range.

d) Applying a log transformation to the response variable.

16. (4 points) Robust regression methods such as M-estimation (e.g. `rlm` in R) differ from ordinary least squares primarily in that they:

a) Minimize the sum of squared residuals, like OLS, but restrict the coefficients to be positive.

b) Downweight observations with large residuals so that outliers have less influence on the fitted line.

c) Require the errors to follow a t-distribution rather than a normal distribution.

d) Automatically remove outliers from the dataset before fitting.

17. (4 points) Least Trimmed Squares (LTS) regression achieves robustness by:

a) Replacing each residual by its absolute value before squaring.

b) Minimizing the sum of the smallest squared residuals (roughly half the data), while ignoring the observations with the largest residuals.

c) Adding a ridge penalty to the OLS objective function.

d) Fitting the regression only to observations whose leverage is below a threshold.