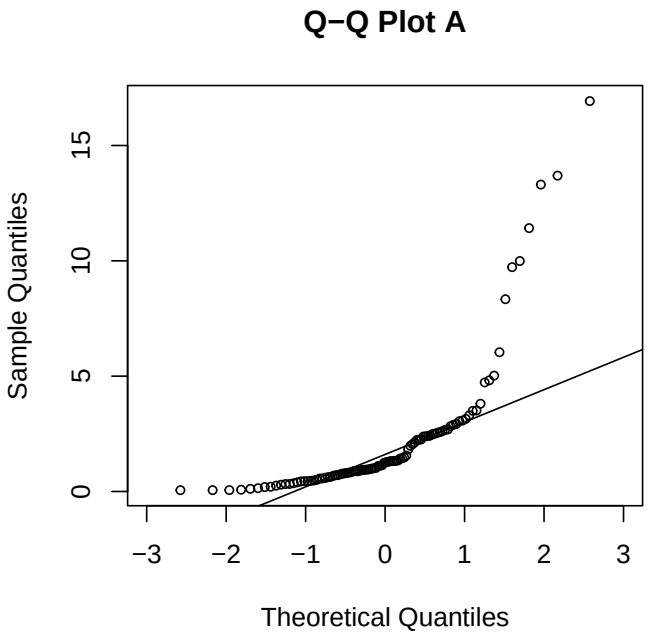
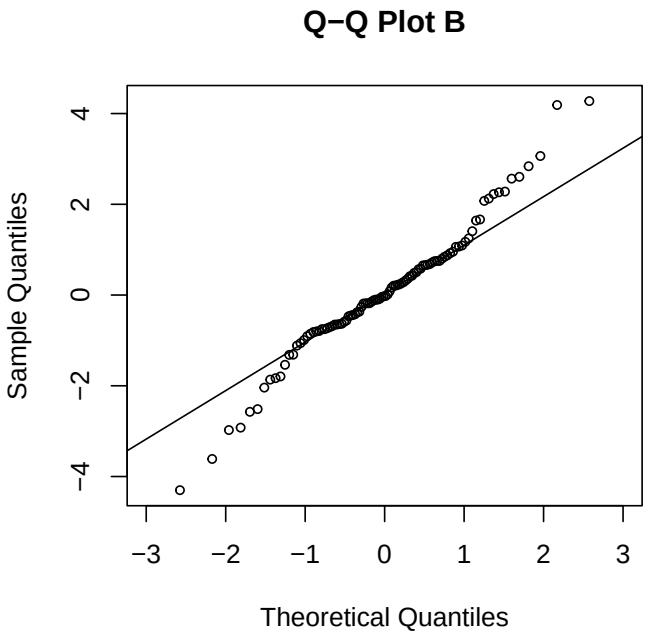
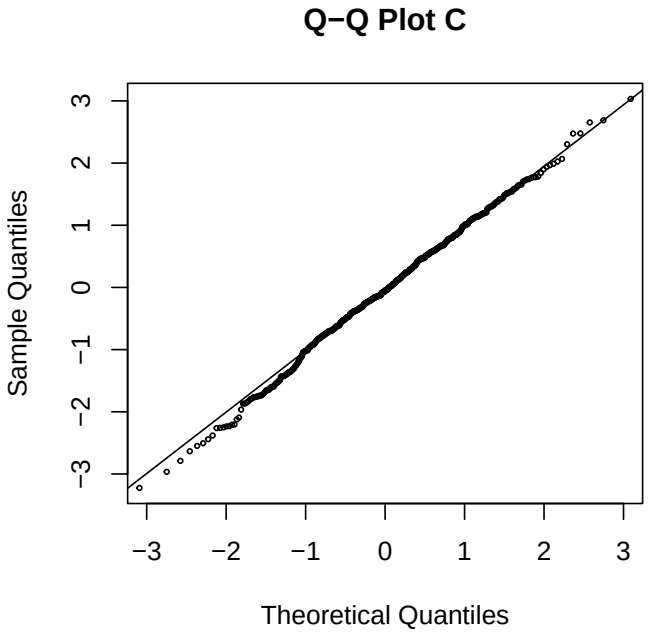
1. Each of the 3 Q-Q plots below is constructed by taking independent random samples from a probability distribution and applying the usual procedure to construct a Q-Q plot.







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

Solution: Plot A. The sample quantiles curve sharply upward above the reference line on the right — the largest observed values are much larger than the corresponding normal quantiles would predict — indicating a long right tail. (Plot A was drawn from an exponential distribution, which is right-skewed with all mass on $(0, \infty)$.)

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

Solution: Plot B. The points fall *above* the reference line on the right and *below* it on the left — an S-shaped pattern symmetric about the center — indicating that both tails are heavier than normal. (Plot B was drawn from a t distribution, which is symmetric with heavier tails than the normal.)

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

Solution: Plot C. The points lie close to the reference line throughout, with only small random deviations, which is the expected pattern when the data come from a normal distribution.

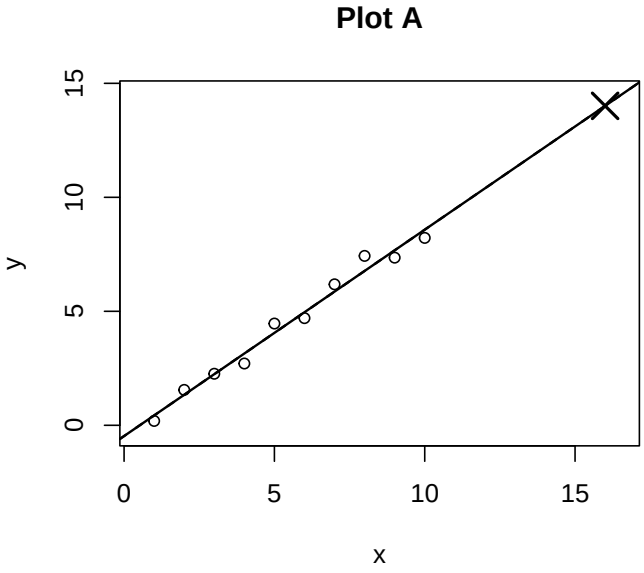
(d) (4 points) Which, if any, of these Q-Q plots is most likely to be constructed by taking samples from an exponential distribution?

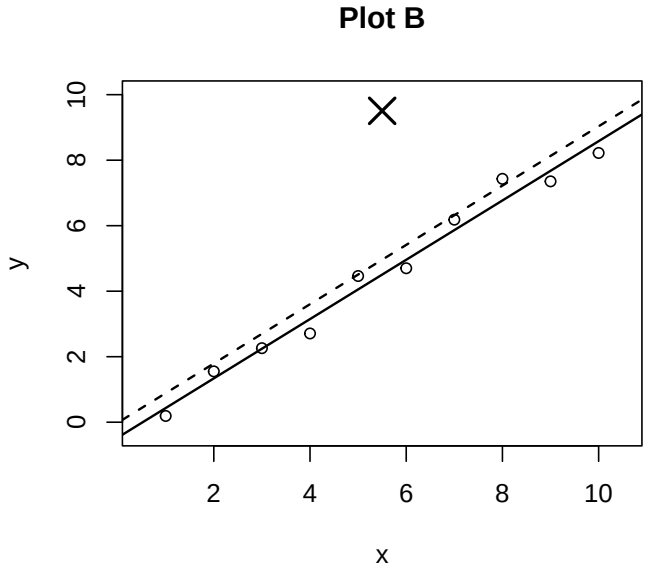
Solution: Plot A. The exponential distribution is strictly non-negative and right-skewed. Accordingly, Plot A has all sample quantiles ≥ 0, and the right side curves sharply upward away from the reference line.

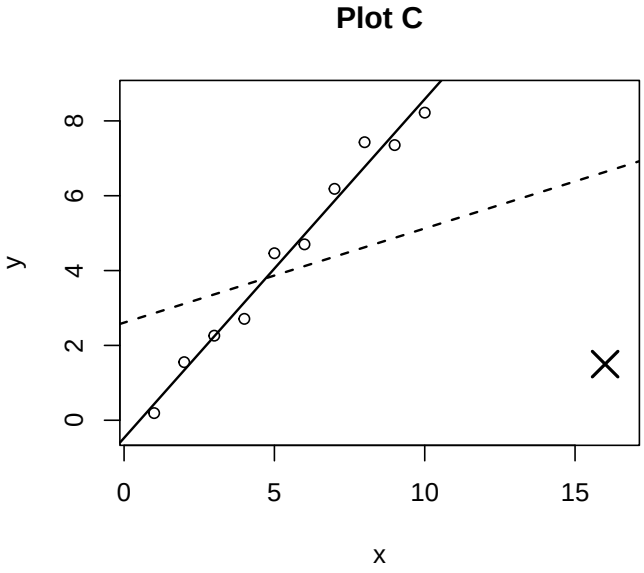
(e) (4 points) Which, if any, of these Q-Q plots shows a distribution with a short (or thin) left tail?

Solution: Plot A. The exponential distribution has an abrupt lower bound of 0, so there is essentially no left tail at all. In the Q-Q plot this appears as the leftmost points hugging just above zero while the reference line continues to predict increasingly negative values — the sample quantiles on the left fall *above* the reference line, indicating a short (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.)

Solution: Plots B and C.
In Plot B, the X is far above the solid regression line at its x-value, so it has a large residual. In Plot C, the X is far below the solid line at its x-value. In Plot A, the X falls essentially on the

extrapolated solid line, so it is not an outlier.

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

Solution: Plots A and C.
In both A and C the X has an extreme x-value, far to the right of the rest of the data. The leverage of a point depends only on how far its predictor value is from the center of the predictor values; the y-value is irrelevant. In Plot B the X is near the center of the x-range, so it has low leverage.

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

Solution: Plot C only.
Only in Plot C do the solid and dashed regression lines differ substantially — the dashed line has a noticeably different slope. In Plot A the X has high leverage but lies on the line, so the dashed line is nearly identical to the solid one. In Plot B the X is an outlier but has low leverage, so it shifts the intercept only slightly and does not meaningfully alter the slope.

(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.

Solution: A data point is influential if removing it from the dataset would substantially change the fitted line — that is, the slope or intercept would look quite different without that point. Put another way, an influential point is one that the regression analysis is heavily "dependent on": it is pulling the line noticeably toward itself.

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."

   Solution: On average, patients with low severity scores stayed 2 fewer days in hospital than patients with high severity scores.

   (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?

   Solution: For a low-severity patient ($S = 1$):

   $$\begin{aligned} \text{Private insurance } (I = 1): &\quad \hat{Y} = \hat{\beta}_0 + (-2)(1) + (-1.5)(1) \\ \text{Public/no insurance } (I = 0): &\quad \hat{Y} = \hat{\beta}_0 + (-2)(1) + (-1.5)(0) \end{aligned}$$

   The difference is $\hat{\beta}_2 = -1.5$ days. Model 2 predicts that low-severity patients with private insurance have, on average, a **1.5-day shorter** stay than low-severity patients without private 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?

   Solution: For a patient with private insurance ($I = 1$):

   $$\begin{aligned} \text{High severity } (S = 0): &\quad \hat{Y} = \hat{\beta}_0 + (-2)(0) + (-1.5)(1) + (1)(0)(1) = \hat{\beta}_0 - 1.5 \\ \text{Low severity } (S = 1): &\quad \hat{Y} = \hat{\beta}_0 + (-2)(1) + (-1.5)(1) + (1)(1)(1) = \hat{\beta}_0 - 2.5 \end{aligned}$$

   Change $= \hat{\beta}_1 + \hat{\beta}_3 = -2 + 1 = \mathbf{-1}$ day. Switching from high to low severity is associated with a 1-day *reduction* in length of stay.

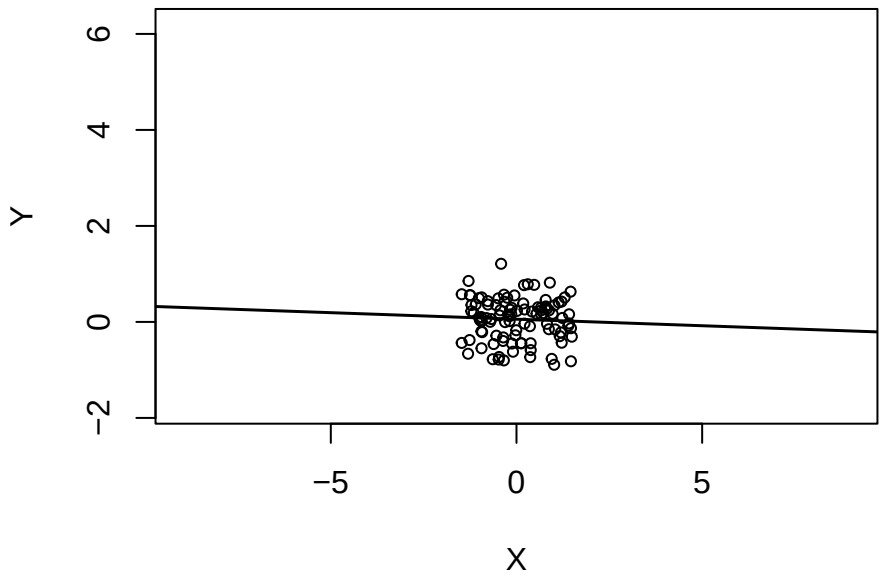
(d) (4 points) Suppose, for Model 3 we have the same coefficient estimates as in the previous part. What is the predicted change in length of stay associated with switching from high to low severity for a patient *without* private insurance?

Solution: For a patient without private insurance ($I = 0$):

$$\begin{aligned} \text{High severity } (S = 0): &\quad \hat{Y} = \hat{\beta}_0 \\ \text{Low severity } (S = 1): &\quad \hat{Y} = \hat{\beta}_0 + (-2)(1) = \hat{\beta}_0 - 2 \end{aligned}$$

Change $= \hat{\beta}_1 = \mathbf{-2}$ days. Switching from high to low severity is associated with a 2-day *reduction* in length of stay for patients without private insurance — a larger reduction than for patients with private insurance. The interaction term $\hat{\beta}_3 = 1 > 0$ reflects that the severity-related reduction in stay is partially offset for privately insured patients.

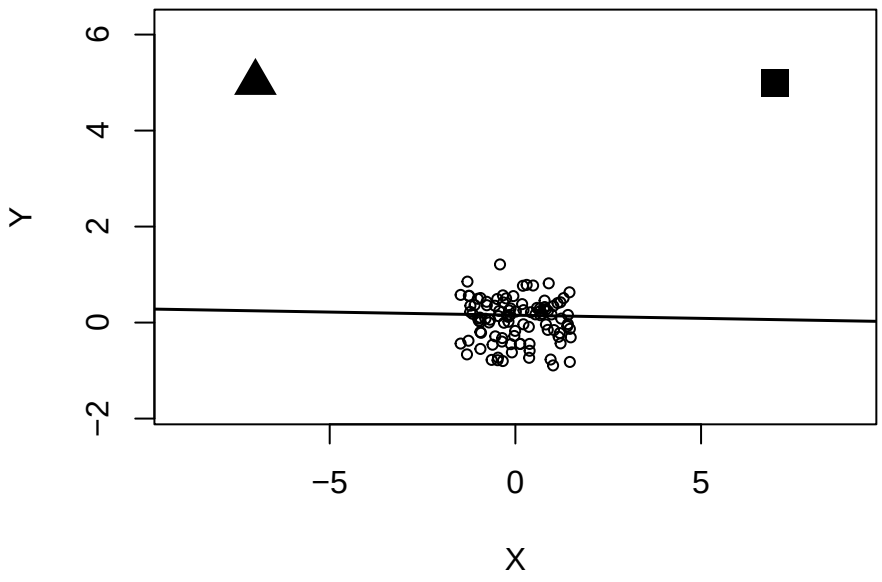
4. You fit a linear regression in R with `lm(Y~X)`, using 100 rows of data. The scatter plot with the regression line looks like this:



(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?

Solution: The 100 points cluster near the origin with no discernible linear trend. We would expect $\hat{\beta}_1 \approx 0$, $\hat{\beta}_0 \approx 0$, and $R^2 \approx 0$.

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:



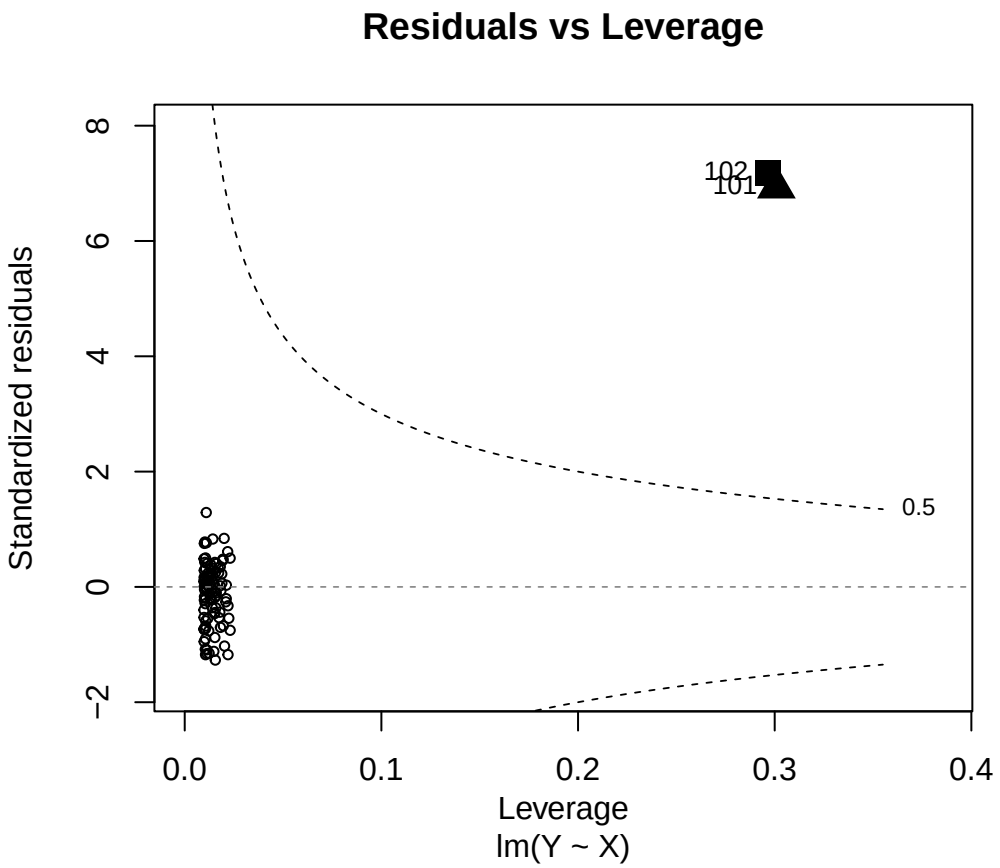
(b) (4 points) If we refit the regression using only the triangle point added to the original 100 rows (but *not* the square), what approximate values would you expect for $\hat{\beta}_1$ and $\hat{\beta}_0$?

Solution: The triangle (▲) is at a *large negative* x-value with a *large positive* y-value (top-left of the plot). This single extreme point, having high leverage, pulls the fitted line toward itself, creating a *negative* slope. We would expect $\hat{\beta}_1 < 0$ (noticeably negative) and $\hat{\beta}_0$ slightly above 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 $\hat{\beta}_0$?

Solution: The square (■) is at a *large positive* x-value with a *large positive* y-value (top-right). Its leverage-weighted pull on the slope is equal and opposite to that of the triangle: the triangle pushes the slope negative and the square pushes it positive. Together they largely cancel, so the slope remains approximately zero. Both points have high y-values, so they jointly pull the intercept slightly upward. We would expect $\hat{\beta}_1 \approx 0$ and $\hat{\beta}_0$ slightly positive (but close to 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?

Solution: Both added points are influential. In the Residuals vs. Leverage plot, Cook's distance contours are shown as dashed curves. Points 101 and 102 both appear far outside (well beyond) the Cook's distance = 0.5 contour, so both have Cook's distance $\gg 0.5$.

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

Solution: Yes, there is an *apparent* conflict, but it is resolved by understanding what Cook's distance actually measures.

Part (c) said that adding both points barely changes the slope or intercept from what they were with only the 100 main points. So the *combined* regression is not very different from the original.

Yet part (d) says both points are highly influential (Cook's $D \gg 0.5$). The resolution is that Cook's distance for a given point is defined as the change in the fit when *that single point* is removed from the current dataset, which *includes* both special points.

- If we remove only the *triangle* (leaving the square in), the square alone — at large positive x, large positive y — creates a strongly *positive* slope. This is very different from the flat fit with both points, so the triangle has enormous Cook's distance.
- Symmetrically, removing only the *square* (leaving the triangle in) produces a strongly *negative* slope, so the square also has enormous Cook's distance.

In other words: each point is highly influential relative to the model that includes the other, because their effects on the slope cancel each other out. Remove either one and the cancellation disappears, causing a dramatic change in the fit. Both points are genuinely influential — it is just that their influence happens to be in opposite directions, so they mask each other when present together.

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.

Solution: Using the bilinearity of covariance and the given independence assumptions:

$$\begin{aligned}
\text{Cov}(Y, X_1) &= \text{Cov}(2 + X_1 + 3X_2 + \epsilon,\ X_1) \\
&= \text{Cov}(2, X_1) + \text{Cov}(X_1, X_1) + 3\,\text{Cov}(X_2, X_1) + \text{Cov}(\epsilon, X_1) \\
&= 0 + \sigma_1^2 + 3\rho\sigma_1\sigma_2 + 0 \\
&= \sigma_1^2 + 3\rho\sigma_1\sigma_2.
\end{aligned}$$

(The first term is 0 because the covariance with a constant is 0; the last term is 0 by the independence of ϵ and X_1.)

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

Solution:

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X_1)}{\text{Var}(X_1)} = \frac{\sigma_1^2 + 3\rho\sigma_1\sigma_2}{\sigma_1^2} = 1 + 3\rho\,\frac{\sigma_2}{\sigma_1}.$$

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

Solution: With $\rho = -\frac{1}{2}$ and $\sigma_1 = \sigma_2 = 1$:

$$\hat{\beta}_1 = 1 + 3\left(-\tfrac{1}{2}\right) \cdot \frac{1}{1} = 1 - \tfrac{3}{2} = -\tfrac{1}{2}.$$

(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.

Solution: No. The true coefficient of X_1 is 1, but the estimated value is $-\frac{1}{2}$ — not only different, but of the opposite sign.

The reason is that X_1 and X_2 are negatively correlated: when X_1 is high, X_2 tends to be low. Since X_2 has a large positive effect on Y (coefficient 3), having a high X_1 typically predicts a low X_2, which in turn predicts a *lower* Y. When X_2 is left out of the model, the simple regression picks up this indirect, negative association between X_1 and Y and misattributes it to a direct (negative) effect of X_1. In reality, X_1 has a small *positive* direct effect on Y; the estimated negative sign is an artefact of the omitted X_2.

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

Solution: Omitted Variable Bias (also called confounding or confounding bias).

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.

Key principle: multiplying the response Y by a constant c multiplies every coefficient estimate and every standard error by c, and multiplies the residual standard error by c. The t-statistics, p-values, and R^2 are unchanged.

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

> **Solution: Does not change.** R^2 is a ratio of variances that is scale-invariant.

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

> **Solution: Changes.** New intercept $= 52.3416 \times 100 =$ **5234.16** (i.e. 5230 to 3 s.f.).

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

> **Solution: Changes.** New SE $= 8.1247 \times 100 =$ **812.47** (i.e. 812 to 3 s.f.).

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

> **Solution: Does not change.** The t-statistic (and hence the p-value) equals the coefficient divided by its standard error; both are multiplied by the same factor $c = 100$, so the ratio — and the p-value — are unchanged. The p-value remains 0.258.

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

Solution: Changes. New RSE $= 12.43 \times 100 =$ **1243** (i.e. 1240 to 3 s.f.).

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.

Key principle: replacing predictor X by $10X$ divides the corresponding coefficient (and its standard error) by 10. The t-statistic, p-value, R^2, RSE, and all other coefficients are unchanged.

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

Solution: Does not change. Rescaling a predictor does not affect R^2.

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

Solution: Does not change. The intercept is the fitted value when all predictors are zero; rescaling `procedures` does not change the fitted values, so the intercept is unaffected.

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

Solution: Changes. Because the predictor has been multiplied by 10, its coefficient is divided by 10 to keep the fitted values unchanged. New coefficient $= 1.8534/10 =$ **0.185** (to 3 s.f.).

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

Solution: Does not change. The t-statistic is (new coefficient)/(new SE) $= (1.8534/10)/(0.2419/10) = 1.8534/0.2419$, the same as before. Hence the p-value is unchanged.

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

Solution: Does not change. The RSE depends on the residuals, which are unchanged when a predictor is rescaled (because the fitted values are unchanged).

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?

Solution: The Box-Cox procedure selects a transformation of the **response** Y.

(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?

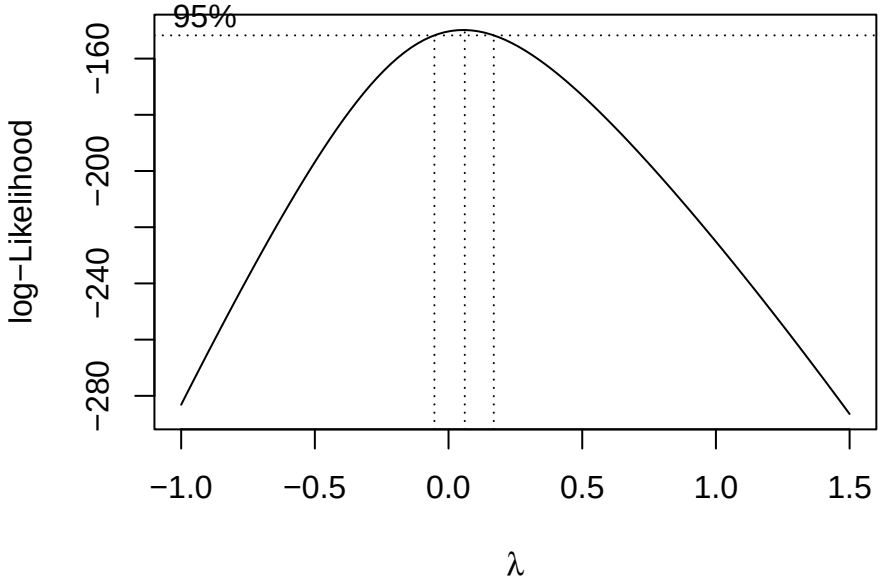
Solution: The Box-Cox procedure considers the family of power transformations

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & \lambda \neq 0, \\ \log Y & \lambda = 0. \end{cases}$$

The parameter λ is estimated by maximum likelihood. Different values correspond to familiar transformations: $\lambda = 1$ means no transformation, $\lambda = 1/2$ gives a square-root transformation, $\lambda = 0$ gives a log transformation, and $\lambda = -1$ gives a reciprocal transformation. The procedure selects the λ that makes the transformed data most consistent with a normal linear model (constant variance, normally distributed errors).

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

Solution: All values of the response Y must be **strictly positive** ($Y > 0$). This is necessary because the power transformations Y^λ are undefined (or complex) for $Y \leq 0$ when λ is not a positive integer, and $\log Y$ is undefined for $Y \leq 0$.



(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?

Solution: We would apply the **log transformation**, i.e. replace Y by $\log Y$.

The reason is twofold. First, the maximum log-likelihood occurs at $\lambda \approx 0$, and since $\lambda = 0$ in the Box-Cox family corresponds exactly to the log transformation, the data strongly suggest that the log scale is most appropriate. Second, $\lambda = 0$ lies well within the 95% confidence interval, meaning we cannot reject $\lambda = 0$ in favour of any other value; we therefore prefer the log transformation over the exact maximum-likelihood value because it is simpler and far more interpretable.

Multiple Choice: answers.

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?

Solution: b) An F-test (partial F-test / ANOVA). The F-test directly compares the RSS of the two nested models and accounts for the difference in the number of parameters. Adjusted R^2 and condition numbers are not formal hypothesis tests. Individual t-tests only assess each predictor separately and do not constitute a simultaneous test of the group.

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

Solution: c) Multicollinearity among the predictors. The VIF of predictor X_j equals $1/(1-R_j^2)$, where R_j^2 is the R^2 from regressing X_j on all other predictors. A large VIF indicates that X_j is nearly linearly predictable from the others, i.e. high collinearity.

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

Solution: c) The point lies far from the center of the predictor space, but the fitted model passes close to it. Leverage depends only on x; a small residual means the point is well-fitted. A point with high leverage but small residual therefore conforms to the trend of the data and is not influential (small Cook's distance), so option (a) is wrong.

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

Solution: b) The remaining coefficient estimates to be biased if the omitted variable is correlated with any included predictor. This is the classical omitted variable bias result, as illustrated in Question 5. When the omitted variable is uncorrelated with the predictors, the remaining estimates are unbiased but the standard errors are inflated; when it is correlated, the estimates are biased.

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

Solution: b) Individual t-tests may fail to flag predictors as significant even though the predictors are collectively highly significant. Collinearity inflates the standard errors of individual coefficient estimates, making t-statistics small. However, the overall F-test (which tests all predictors jointly) is not similarly inflated, so it can remain significant. The fitted values and RSE are not strongly affected.

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?

Solution: b) There is moderate-to-strong collinearity; the parameter estimates may be unstable. A condition number around 30 is a common warning threshold; a value of 450 is very high and strongly indicative of collinearity. A condition number below 500 is *not* automatically acceptable — option (a) is false. A large condition number indicates near-singular design matrix, not heteroscedasticity (c) and not exact collinearity (d).

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:

Solution: b) There is a near-linear relationship involving *three or more* predictors simultaneously, even though no two predictors are strongly correlated in pairs. Pairwise correlations detect only two-variable collinearity. VIFs detect multicollinearity of any order. For example, if $X_3 \approx 2X_1 - X_2$, the VIFs of all three will be large even if the pairwise correlations are modest.

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

Solution: d) Applying a log transformation to the response variable. A log transformation of Y addresses non-normality or heteroscedasticity of the errors; it has no effect on the relationships *among* the predictors and therefore does not reduce collinearity. Options (a), (b), and (c) all directly reduce the collinearity among the X variables.

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

Solution: b) Downweight observations with large residuals so that outliers have less influence on the fitted line. M-estimation replaces the OLS sum of squared residuals with a sum of a less rapidly growing function of the residuals, effectively applying smaller weights to observations with large residuals. It does not restrict coefficients (a), require a t-distribution (c), or delete observations (d).

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

Solution: b) Minimizing the sum of the smallest squared residuals (roughly half the data), while ignoring the observations with the largest residuals. LTS discards the observations that fit worst (those with the largest squared residuals) and minimizes the sum of the remaining squared residuals. This makes it highly resistant to outliers. It is not related to absolute values (a), ridge penalties (c), or leverage (d).