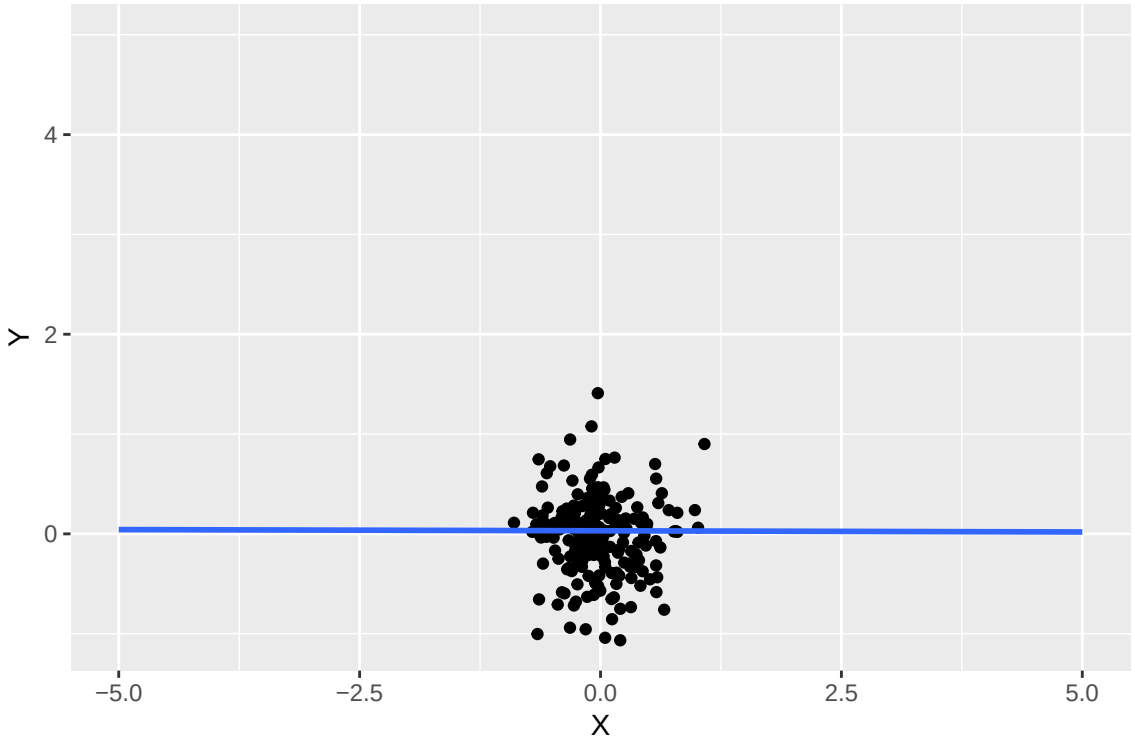
1. You fit a linear regression in R with `lm(Y~ X)`, involving two variables, `X` and `Y`, and 197 rows of data. The regression model is $\texttt{Y} = \beta_0 + \beta_1 \texttt{X} + \epsilon$. The scatter plot with the regression line looks like this:



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?

> **Solution:** The regression line is flat (horizontal) and sits at approximately $Y = 0$. Therefore we expect
>
> $$\hat{\beta}_1 \approx 0 \qquad \text{and} \qquad \hat{\beta}_0 \approx 0.$$

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

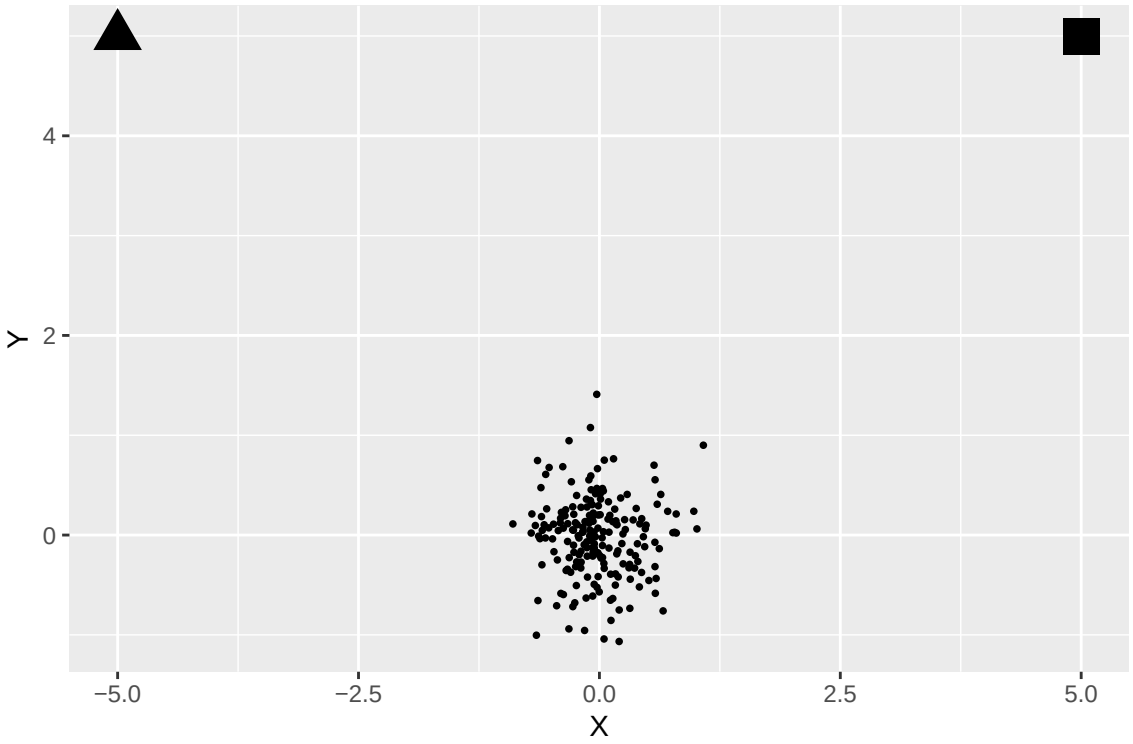
> **Solution:** Because the fitted line is essentially flat and the scatter shows almost no linear trend, $R^2 \approx 0$. The predictor X explains virtually none of the variation in Y.

(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?

> **Solution: No.** The regression line sits at $Y \approx 0$, so $\hat{\beta}_0 \approx 0$. The t-statistic $t = \hat{\beta}_0/\text{SE}(\hat{\beta}_0)$ will therefore be close to zero, yielding a large p-value well above 0.05. Equivalently, the null

hypothesis H_0: $\beta_0 = 0$ is entirely consistent with the graph: the mean response when $X = 0$ appears to be approximately zero.

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?

Solution: The triangle (▲) is located at approximately $(X, Y) \approx (-5, 5+)$, far to the upper left of the main cloud of points. This single high-leverage point pulls the fitted line so that it rises steeply as X decreases, inducing a negative slope. We therefore expect

$$\hat{\beta}_1 \approx -0.1 \text{ to } -0.2, \qquad \hat{\beta}_0 \approx 0 \text{ to } 0.5.$$

(The intercept remains close to its original value because the high-Y pull of the triangle is diluted by the 197 original points near $Y \approx 0$.)

(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?

Solution: Yes. The triangle lies at $X \approx -5$, far from the mean $\bar{X} \approx 0$ of the remaining data. The standard error of the slope is

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

Adding a point at $X = -5$ greatly increases $\sum (x_i - \bar{x})^2$, dramatically reducing $\mathrm{SE}(\hat{\beta}_1)$. Even though the estimated slope is only modestly negative, the much smaller standard error yields a large $|t|$-statistic, so the p-value will be well below 0.05.

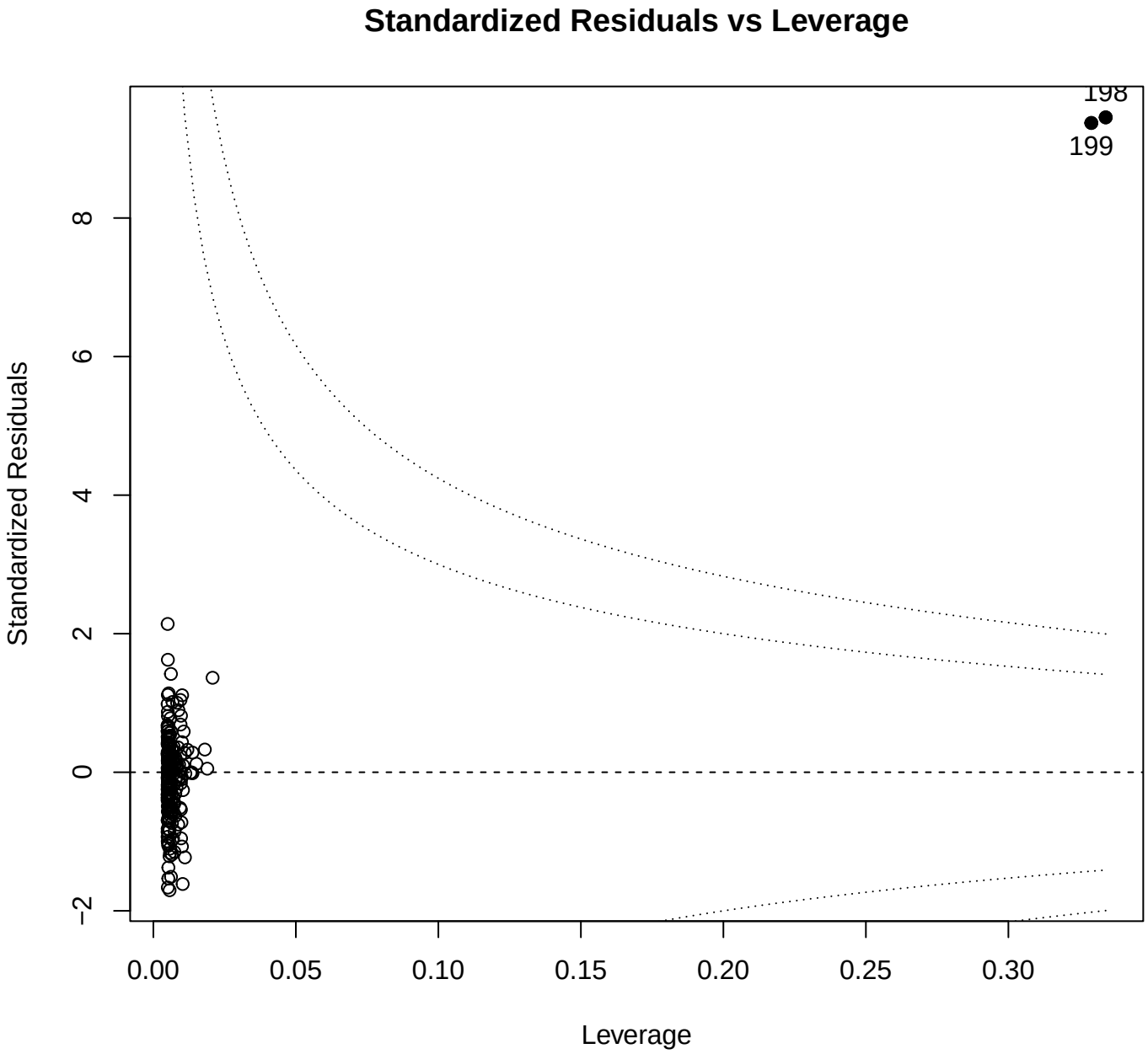
(f) (6 points) If we redo the regression *again* with *both* the square *and* triangle points added to the data set, what approximate values would you expect for the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ in the regression output, based on the graph above?

Solution: The square (■) is at approximately $(X, Y) \approx (+5,\ 5+)$, a mirror image of the triangle across $X = 0$. When both points are included their opposing effects on the slope approximately cancel: the triangle pulls the slope negative and the square pulls it positive by essentially the same amount. Their reinforcing upward pull on the intercept is diluted by the 197 original points. We therefore expect

$$\hat{\beta}_1 \approx 0, \qquad \hat{\beta}_0 \approx 0 \text{ to } 0.5,$$

very close to the original 197-point regression.

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.)

Solution: Both points 198 (square) and 199 (triangle) lie well outside the Cook's distance = 0.5 contour in the plot, with standardised residuals near 8 and leverage near 0.30. **Both the square and the triangle are influential.**

(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).)

Solution: No. As argued in (f), both the slope and intercept remain very close to their original values from the 197-point regression.

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

Solution: There is an apparent *conflict*: in (g) both added points are individually influential (large Cook's distances), yet in (h) the regression line has not meaningfully changed. How can two highly influential points leave the fit unchanged?

The resolution is that the two points are *jointly* influential but in *opposing directions* for the slope. Point 199 (triangle, $X \approx -5$, high Y) pulls the slope negative and the intercept upward; point 198 (square, $X \approx +5$, equally high Y) pulls the slope positive and the intercept upward by the same amount. Their effects on the slope cancel almost exactly. Their reinforcing upward pull on the intercept is small relative to the 197 original data points.

Cook's distance measures the influence of each point *in isolation*: if either point were added alone it *would* substantially alter the fit (as shown in parts (d)–(e)). But when both are present simultaneously, their net effect on the regression line is negligible. This illustrates a general caution: large Cook's distances do not guarantee a large net change in the model when multiple influential points are present and their individual effects are in opposing directions.

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.)

Solution:

- $\widehat{\alpha}_0 > 0$: the intercept is the predicted price of a brand-new laptop (0 years old, 0% wear), which should be positive.
- $\widehat{\alpha}_1 < 0$: holding age fixed, higher battery wear lowers price.
- $\widehat{\alpha}_2 < 0$: holding wear fixed, greater age lowers price.

(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.)

Solution: $\widehat{\beta}_1 < \widehat{\alpha}_1$.
(Both values are negative; the simple-regression coefficient on `wear` is *more negative* than the multiple-regression coefficient.)

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

Solution: This is an instance of **omitted-variable bias**. The omitted variable `age` is *positively* correlated with `wear` (older laptops have more battery wear) and has a *negative* effect on price ($\widehat{\alpha}_2 < 0$). The omitted-variable bias formula gives

$$\widehat{\beta}_1 \approx \widehat{\alpha}_1 + \widehat{\alpha}_2 \cdot \hat{\delta},$$

where $\hat{\delta} > 0$ is the coefficient from the auxiliary regression of `age` on `wear`. Since $\widehat{\alpha}_2 < 0$ and $\hat{\delta} > 0$, the correction term is negative, making $\widehat{\beta}_1 < \widehat{\alpha}_1$. Intuitively: in the simple regression, the coefficient on `wear` also picks up the negative effect of the omitted variable `age`, which correlates positively with `wear`, making the estimated effect of wear appear more strongly negative than it truly is after controlling for age.

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.)

Solution: With Age = 30, DrivingExperience = 10, HasAccidents = 1:

$$\hat{Y} = 270 + (-1.5)(30) + (-2.0)(10) + (60)(1) + (0.05)(30)(10) + (1.2)(30)(1).$$

(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.

Solution: Statement **(ii)** is correct. For fixed Age and DrivingExperience, the difference in predicted premium between a person with accidents ($X_3 = 1$) and without ($X_3 = 0$) is

$$\Delta\hat{Y} = 60(1) + 1.2 \cdot \text{Age} \cdot (1) - \big[60(0) + 1.2 \cdot \text{Age} \cdot (0)\big] = 60 + 1.2 \cdot \text{Age}.$$

Since $60 > 0$ and $1.2 > 0$, we have $\Delta\hat{Y} > 0$ for *every* non-negative Age: people with accidents always pay more, with no restriction on Age. Statement (iii) is incorrect because the condition "Age is high enough" is unnecessary.

(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.

Solution: False. Two reasons:

1. **Scale.** The coefficient 0.05 is applied to the product Age $\times$ DrivingExperience, which can be very large (e.g. $50 \times 30 = 1500$), making the actual contribution to the predicted premium potentially substantial.

2. **Statistical evidence.** The evidence for an interaction effect is measured by the t-statistic, $t = \hat{\beta}_4/\text{SE}(\hat{\beta}_4)$, not by the magnitude of $\hat{\beta}_4$ alone. A small coefficient can still be highly statistically significant if its standard error is equally small, and conversely a large coefficient can be non-significant if its standard error is large. Without the standard error we cannot draw any conclusion about evidence for the interaction.

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?

   Solution: One would intuitively expect both signs to be *positive*: higher manager salaries ought to attract better managers, and more formal training ought to improve staff performance, both improving customer satisfaction. Instead $\hat{\beta}_1 = -1.85$ and $\hat{\beta}_5 = -1.47$, implying (all else equal) that higher manager pay and more training are associated with *lower* satisfaction.

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

   Solution: In a multiple regression, each coefficient represents the estimated effect of that variable *while holding all the others constant*, which is not the same as its raw, unconditional association.

   - **Manager salary ($\hat{\beta}_1 < 0$):** Stores with higher manager salaries may tend to be in more challenging or competitive markets. Once we account for the neighborhood economic index and other store characteristics, those stores may face harder conditions that depress satisfaction in ways not fully captured by the predictors.
   - **Formal training ($\hat{\beta}_5 < 0$):** Stores that invest heavily in remedial training may do so precisely because they already have customer-service problems. Once we control for communication skills and experience—which reflect staff quality more directly—additional formal training (above and beyond those factors) may signal pre-existing difficulties rather than competence. Causation may run in the opposite direction: low satisfaction prompts more training.

   In short, the negative signs most likely reflect the presence of multicollinearity and/or confounding variables, not a genuine harmful effect of salary or training.

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

Solution: Variance Inflation Factor.

(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?

Solution: All four VIF values are well above the commonly used threshold of 10 (or even the more conservative threshold of 5). This indicates **severe multicollinearity** among X_1, X_2, X_3, and X_5: their standard errors are inflated by factors of roughly 9–12 relative to what they would be if the predictors were uncorrelated. As a result, individual coefficient estimates are highly unstable and their signs may be unreliable, which explains the counterintuitive results in (a).

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

Solution: Several remedial strategies are possible:

- **Drop redundant predictors.** Examine correlations and VIFs to identify which variables are most collinear, then remove one or more of them. For example, if formal training (X_5) is highly correlated with communication skills (X_4) or experience (X_2), one of these may be dropped.
- **Combine collinear predictors,** e.g. via a principal component or a composite index, reducing the effective number of predictors while retaining most of the information.
- **Use Ridge regression,** which adds an ℓ_2 penalty to the OLS criterion and produces more stable coefficient estimates under multicollinearity, at the cost of a small amount of bias.
- **Collect more data,** which improves the precision of estimates (though it does not eliminate the multicollinearity).

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.

Solution: The **cubic training RSS $\leq$ linear training RSS**.
The cubic model contains the linear model as a special case (set $\beta_2 = \beta_3 = 0$). Since OLS minimises RSS over all parameter values, it achieves an RSS at most as large as the linear model's

training RSS. In practice the cubic will fit the training data at least as well—and typically slightly better because it can adapt to random noise—regardless of whether the true relationship is linear.

(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.

Solution: We would expect the **linear test RSS $\leq$ cubic test RSS** (or roughly equal, but not appreciably worse).

When the true relationship is linear, the linear model is correctly specified: it has zero bias and low variance. The cubic model overfits the training noise by estimating two unnecessary parameters (β_2, β_3), increasing variance without reducing bias. On new test data this extra variance translates to larger prediction errors, so the linear model is expected to have the lower test RSS.

(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.

Solution: The **cubic training RSS $\leq$ linear training RSS**, for exactly the same reason as in (a): the cubic model has at least as many parameters, so OLS cannot achieve a higher RSS on the training data. This comparison is purely algebraic and does not depend on the true functional form.

(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.

Solution: There is not enough information to tell.

If the true relationship departs substantially from linearity, the cubic model's flexibility (lower bias) outweighs its higher variance, and we would expect a lower cubic test RSS. If the true relationship is only mildly non-linear, the cubic model may overfit and have a higher test RSS than the linear model. Without knowing how far the truth is from linear, we cannot determine which effect dominates.

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?

Solution: Each categorical variable with 8 levels requires $8 - 1 = 7$ dummy variables. With 8 such variables:

$$8 \times 7 = \mathbf{56} \text{ dummy variables.}$$

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

Solution: Including 56 dummy variables plus an intercept gives $p = 57$ parameters, which exceeds the number of observations $n = 50$. When $p > n$, the design matrix $\mathbf{X}$ does not have full column rank, so the normal equations $\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$ have no unique solution. Even if a solution is forced by dropping some dummies, having more free parameters than observations means the model can fit the training data perfectly (zero residuals) while having no predictive power for new 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 $\hat{\beta}_2$?

Solution: Both $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 > 0$. By assumption, higher nurse education and higher doctor experience are each associated with better recovery rates, so the partial regression coefficients should both be positive.

(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` ($\texttt{recovery} = \alpha_0 + \alpha_1 \texttt{nurseedu} + \epsilon$) 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 $\widehat{\beta}_1 > \widehat{\alpha}_1$.

Solution: $\widehat{\beta}_1 > \widehat{\alpha}_1$.
The omitted-variable bias formula gives

$$\widehat{\alpha}_1 \approx \widehat{\beta}_1 + \widehat{\beta}_2 \cdot \hat{\delta},$$

where $\hat{\delta}$ is the slope from regressing `docexp` on `nurseedu`. Because `nurseedu` and `docexp` are *negatively* correlated, $\hat{\delta} < 0$. Since $\widehat{\beta}_2 > 0$ and $\hat{\delta} < 0$, the correction term is negative, so $\widehat{\alpha}_1 < \widehat{\beta}_1$. Intuitively: hospitals with more nurse education tend to have *less* experienced doctors (by the given correlation), which suppresses recovery. The simple regression cannot distinguish these two effects, so it attributes the negative influence of low doctor experience to nurse education, making $\widehat{\alpha}_1$ appear smaller than $\widehat{\beta}_1$.

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.

Solution: Each additional year of education is associated with approximately an **8% increase** in hourly wage, holding years of labor-market experience constant.
(More precisely, the multiplicative factor is $e^{0.08} \approx 1.083$, i.e. an 8.3% increase; for small coefficients the approximation $e^{\beta} - 1 \approx \beta$ is standard.)

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

Solution: Each additional year of labor-market experience is associated with approximately a **2% increase** in hourly wage, holding years of education constant.

(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.

Solution:

$$\widehat{\texttt{wage}} = \exp\big(3 + 0.08 \times 12 + 0.02 \times 10\big).$$

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**

> **Solution: Does not change.** R^2 is a correlation-based, dimensionless measure; scaling the response by a constant does not affect it.

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

> **Solution: Changes.** New value: $28.5660865/100 = \mathbf{0.286}$.

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

> **Solution: Changes.** New value: $0.1446422/100 = \mathbf{0.00145}$.

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

> **Solution: Does not change.** The t-statistic equals $\hat{\beta}/\mathrm{SE}(\hat{\beta})$; both numerator and denominator scale by the same factor of $1/100$, leaving the ratio (and hence the p-value) unchanged.

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

> **Solution: Changes.** New value: $3.803/100 = \mathbf{0.0380}$.

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**

Solution: Does not change. Rescaling a predictor is equivalent to multiplying a column of $\mathbf{X}$ by a constant, which does not change the column space or the fitted values.

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

Solution: Changes. New value: $(-0.0003369) \times 1000 = \mathbf{-0.337}$.

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

Solution: Changes. New value: $0.0009311 \times 1000 = \mathbf{0.931}$.

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

Solution: Does not change. The t-statistic equals $(-0.337)/0.931$, which is the same ratio as $(-0.0003369)/0.0009311$.

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

Solution: Does not change. The fitted values (and hence residuals) are identical to the original model; only the labelling of the `dpi` coefficient changes.

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		

Solution:

	GLS	**WLS**
A	No	No
B	Yes	Yes
C	Yes	No
D	No	No

Explanation.

- **A (Linearity).** Neither GLS nor WLS addresses non-linearity; both still assume a linear model.
- **B (Homoscedasticity).** WLS handles *heteroscedasticity* directly by weighting each observation inversely proportional to its variance. GLS handles the more general case $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$ with known $\mathbf{V}$, which includes heteroscedasticity as a special case.
- **C (Independence).** GLS allows $\mathbf{V}$ to be non-diagonal, accommodating correlated errors (e.g. time-series or spatial data). WLS assumes a diagonal covariance matrix (different variances but independent errors), so it does *not* relax independence.
- **D (Normality).** Neither method relaxes the normality assumption for purposes of inference (t- and F-tests); they modify the weighting of observations, not the distributional assumption on ϵ.

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?

Solution: No, it was not a mistake. In fact, including $\texttt{share}_A$ would have been a mistake.

By definition, $\texttt{share}_T + \texttt{share}_R + \texttt{share}_S + \texttt{share}_A \equiv 1$. If all four shares were included as predictors alongside an intercept, there would be an exact linear dependence among the columns of the design matrix (the four shares sum identically to the intercept column). This is the **dummy-variable trap**, a form of perfect multicollinearity: $\mathbf{X}^\top\mathbf{X}$ would be singular and OLS would have no unique solution.

By omitting $\texttt{share}_A$, the model remains identifiable. The coefficients β_T, β_R, β_S are then interpreted as the effect on `perf` of shifting one unit of budget share *into* teaching, research, or student services, respectively, *from* the residual category $\texttt{share}_A$, which serves as the reference baseline.

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.

Solution:

A. True. Forward stepwise adds one predictor at each step and never removes any, so the k-variable model's predictors are by construction a subset of the $(k+1)$-variable model's predictors.

B. True. Backward stepwise removes one predictor at each step and never adds any, so the $(k+1)$-variable model contains all predictors of the k-variable model plus one more; hence the k-variable model's predictors are a subset.

C. False. Forward and backward stepwise are independent procedures that can select entirely different subsets of predictors. There is no guarantee that the k-variable backward model's predictors appear in the $(k+1)$-variable forward model.

D. False. By the same argument as C, the k-variable forward model's predictors need not be a subset of the $(k+1)$-variable backward model's predictors.

E. False. Best subset selection optimises independently over all $\binom{p}{k}$ possible k-variable models for each k. The best k-variable subset and the best $(k+1)$-variable subset are found by separate optimisations and there is no nesting requirement, so the former need not be a subset of the latter.

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.

> **Solution: Answer: D.** AIC $= -2 \log \hat{L} + 2p$. It measures fit (via the log-likelihood) penalised by the number of parameters p; smaller values are preferred. It does not address outliers, variable importance, or multicollinearity directly.

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.

> **Solution: Answer: D.** Adding a predictor always weakly increases R^2, but the AIC penalty for the extra parameter may outweigh the improvement in fit. AIC can therefore prefer the simpler model with lower R^2, provided that the simpler model also has fewer parameters (which is precisely the situation in which the penalty saves more than the fit loses).

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.

Solution: Answer: B. In WLS each observation receives weight $w_i = 1/\sigma_i^2$, so observations with larger variance (less reliable) are down-weighted, reducing their influence on the coefficient estimates.

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.

Solution: Answer: C. Ridge adds an ℓ_2 penalty $\lambda \sum_j \beta_j^2$ that shrinks correlated coefficients toward each other, stabilising estimates under multicollinearity. Option A describes Lasso (which uses an ℓ_1 penalty to achieve exact zeros). Option B also describes Lasso. Option D is incorrect: Ridge shrinks all coefficients but never sets any exactly to zero.

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.

Solution: Answer: B. Regularisation (Ridge/Lasso) deliberately introduces a small amount of bias by shrinking coefficients toward zero. The gain is a substantial reduction in variance, which often lowers the overall prediction error (MSE = bias^2+ variance + irreducible noise). Models with more predictors have *lower* bias and *higher* variance (opposite of A). Low R^2 indicates a poor fit but does not imply high bias per se.

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.

Solution: Answer: A. The standard Box-Cox family involves power transformations $Y^{(\lambda)}$ (including $\log Y$ for $\lambda = 0$), all of which require $Y > 0$. The Yeo–Johnson transformation extends Box-Cox to non-positive values.

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.

Solution: Answer: B. Larger λ imposes a stronger penalty, shrinking the coefficients more aggressively toward zero. Setting $\lambda = 0$ recovers OLS.

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.

Solution: Answer: A. The ℓ_1 penalty in Lasso encourages sparsity (exact zeros) via the geometry of the ℓ_1 ball, whereas the ℓ_2 penalty in Ridge shrinks all coefficients toward zero but essentially never achieves exact zeros.

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.

Solution: Answer: A. AIC penalty per parameter $= 2$; BIC penalty $= \log n$. For any $n \geq 8$ we have $\log n > 2$, so BIC penalises complexity more heavily and tends to favour smaller models, especially as n grows.

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.

Solution: Answer: B. The penalty shrinks coefficient estimates, reducing their variance and the model's tendency to overfit. While this introduces a small bias, the net effect on MSE (bias^2+ variance) is often beneficial, especially when predictors are correlated or p is large relative to n.

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

Solution: Answer: C. Bradford Hill's (1965) criteria include: Strength, Consistency, Specificity, Temporality, Biological Gradient (Dose-Response), Plausibility, Coherence, Experiment, and Analogy. "Linearity" does not appear among them.

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.

Solution: Answer: B. Robust regression (e.g. IRLS with Huber or bisquare weights) iteratively assigns lower weights to observations with large residuals, limiting the influence of outliers on the coefficient estimates without entirely discarding those observations.