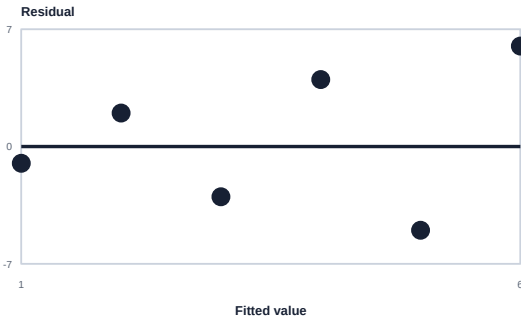


Regression Model Diagnostics

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

1. The residual plot coordinates are (fitted value, residual): (1, -1), (2, 2), (3, -3), (4, 4), (5, -5), (6, 6). Classify the spread as approximately constant or changing, and name the visible concern if any.

Reference: Judge vertical residual spread across fitted values. A widening fan suggests nonconstant variance; a stable band is consistent with constant spread.



2. The core data are (0, 1), (1, 3), (2, 5), (3, 7), (4, 9) and a candidate point is (8, 9). Compare the least-squares line with and without the candidate point. Is it influential for the fitted line?

Reference: A point is influential when including it substantially changes fitted coefficients. Compare both supplied-data fits; distance in x can create leverage.

Regression Model Diagnostics

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

3. A group selected for an unusually high first score has mean 96. On a comparable repeat with no intervention, its mean is 84; the population mean is 75. Explain the change using regression to the mean.

Reference: Extreme observed groups often include temporary variation. On repetition, their average tends to be less extreme even without a causal treatment effect.

4. Three model complexities have (training SSE, held-out SSE): simple=(19,17), medium=(13,8), complex=(3,16). Choose for future prediction and explain the complexity tradeoff.

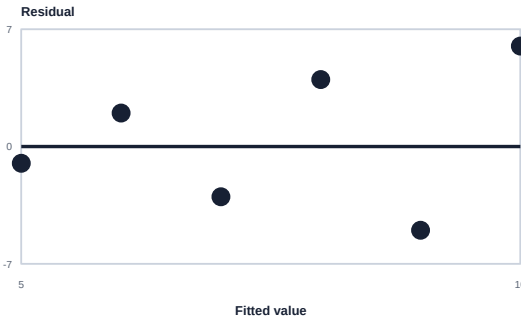
Reference: Training error often falls with complexity. Select predictive complexity using held-out error because it tests new rows.

Regression Model Diagnostics

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

5. The residual plot coordinates are (fitted value, residual): (5, -1), (6, 2), (7, -3), (8, 4), (9, -5), (10, 6). Classify the spread as approximately constant or changing, and name the visible concern if any.

Reference: Judge vertical residual spread across fitted values. A widening fan suggests nonconstant variance; a stable band is consistent with constant spread.



6. The core data are (0, 1), (1, 3), (2, 5), (3, 7), (4, 9) and a candidate point is (8, 4). Compare the least-squares line with and without the candidate point. Is it influential for the fitted line?

Reference: A point is influential when including it substantially changes fitted coefficients. Compare both supplied-data fits; distance in x can create leverage.

Regression Model Diagnostics

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

7. A group selected for an unusually high first score has mean 91. On a comparable repeat with no intervention, its mean is 83; the population mean is 75. Explain the change using regression to the mean.

Reference: Extreme observed groups often include temporary variation. On repetition, their average tends to be less extreme even without a causal treatment effect.

8. Three model complexities have (training SSE, held-out SSE): simple=(14,12), medium=(8,9), complex=(2,11). Choose for future prediction and explain the complexity tradeoff.

Reference: Training error often falls with complexity. Select predictive complexity using held-out error because it tests new rows.

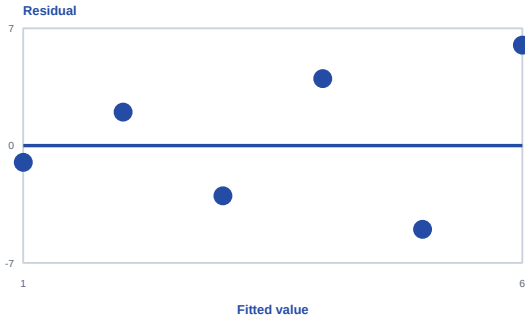
Regression Model Diagnostics

Worked Answer Key

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

1. The residual plot coordinates are (fitted value, residual): (1, -1), (2, 2), (3, -3), (4, 4), (5, -5), (6, 6). Classify the spread as approximately constant or changing, and name the visible concern if any.

Reference: Judge vertical residual spread across fitted values. A widening fan suggests nonconstant variance; a stable band is consistent with constant spread.



Answer: Changing spread; the widening fan suggests nonconstant variance

- 1. Compare absolute residual sizes at low and high fitted values: 1, 2, 3, 4, 5, 6.
- 2. The envelope widens as fitted values increase.

2. The core data are (0, 1), (1, 3), (2, 5), (3, 7), (4, 9) and a candidate point is (8, 9). Compare the least-squares line with and without the candidate point. Is it influential for the fitted line?

Reference: A point is influential when including it substantially changes fitted coefficients. Compare both supplied-data fits; distance in x can create leverage.

Answer: Without: $\hat{y} = 1 + 2x$. With: $\hat{y} = 2.667 + 1x$. The point is influential.

- 1. The five core points give intercept 1 and slope 2.
- 2. After adding (8,9), intercept=2.667 and slope=1; the slope change is 1.

Regression Model Diagnostics

Worked Answer Key

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

3.

A group selected for an unusually high first score has mean 96. On a comparable repeat with no intervention, its mean is 84; the population mean is 75. Explain the change using regression to the mean.

Reference: Extreme observed groups often include temporary variation. On repetition, their average tends to be less extreme even without a causal treatment effect.

Answer: The repeat is closer to the population mean; this pattern alone does not establish a causal effect

1.

Distance from population mean changes from 21 to 9.

2.

Selection on an extreme first measurement plus ordinary variation can produce this movement; no intervention was given, so do not invent a treatment cause.

4.

Three model complexities have (training SSE, held-out SSE): simple=(19,17), medium=(13,8), complex=(3,16). Choose for future prediction and explain the complexity tradeoff.

Reference: Training error often falls with complexity. Select predictive complexity using held-out error because it tests new rows.

Answer: Choose the medium model; its held-out SSE 8 is smallest

1.

Training SSE decreases 19 to 13 to 3.

2.

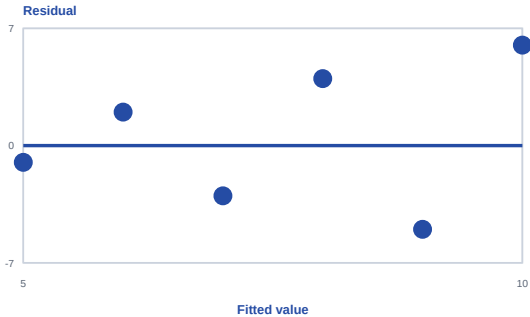
Held-out SSE is 17, 8, 16; it improves and then worsens, evidence of overfitting by the complex model.

Regression Model Diagnostics
Worked Answer Key

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

5. The residual plot coordinates are (fitted value, residual): (5, -1), (6, 2), (7, -3), (8, 4), (9, -5), (10, 6). Classify the spread as approximately constant or changing, and name the visible concern if any.

Reference: Judge vertical residual spread across fitted values. A widening fan suggests nonconstant variance; a stable band is consistent with constant spread.



Answer: Changing spread; the widening fan suggests nonconstant variance

- 1. Compare absolute residual sizes at low and high fitted values: 1, 2, 3, 4, 5, 6.
- 2. The envelope widens as fitted values increase.

6. The core data are (0, 1), (1, 3), (2, 5), (3, 7), (4, 9) and a candidate point is (8, 4). Compare the least-squares line with and without the candidate point. Is it influential for the fitted line?

Reference: A point is influential when including it substantially changes fitted coefficients. Compare both supplied-data fits; distance in x can create leverage.

Answer: Without: $\hat{y} = 1 + 2x$. With: $\hat{y} = 3.708 + 0.375x$. The point is influential.

- 1. The five core points give intercept 1 and slope 2.
- 2. After adding (8,4), intercept=3.708 and slope=0.375; the slope change is 1.625.

Regression Model Diagnostics

Worked Answer Key

Use finite residual plots and error tables. Describe association and prediction evidence without inventing causal effects.

7.

A group selected for an unusually high first score has mean 91. On a comparable repeat with no intervention, its mean is 83; the population mean is 75. Explain the change using regression to the mean.

Reference: Extreme observed groups often include temporary variation. On repetition, their average tends to be less extreme even without a causal treatment effect.

Answer: The repeat is closer to the population mean; this pattern alone does not establish a causal effect

1.

Distance from population mean changes from 16 to 8.

2.

Selection on an extreme first measurement plus ordinary variation can produce this movement; no intervention was given, so do not invent a treatment cause.

8.

Three model complexities have (training SSE, held-out SSE): simple=(14,12), medium=(8,9), complex=(2,11). Choose for future prediction and explain the complexity tradeoff.

Reference: Training error often falls with complexity. Select predictive complexity using held-out error because it tests new rows.

Answer: Choose the medium model; its held-out SSE 9 is smallest

1.

Training SSE decreases 14 to 8 to 2.

2.

Held-out SSE is 12, 9, 11; it improves and then worsens, evidence of overfitting by the complex model.