Past-paper walk-through

Analyzing Departures from Linearity ▪ 2.9

eXercise

Questions in this set

- 2021 Q6
- 2018 Q1
- 2016 Q6
- 2014 Q6

Reading the mark scheme

- **B1** a mark that stands on its own – the point is either made or not.
- **M1** a *method* mark: the right approach, even if the number is wrong.
- **A1** an *answer* mark, and it usually needs its M mark first.
- **C1** a working mark on the way to the answer.
- // separates answers the examiner accepts equally.
- **ecf** = error carried forward: a later part is marked on your own earlier answer.
- **owtte** = "or words to that effect" – the wording need not match.

Question 6

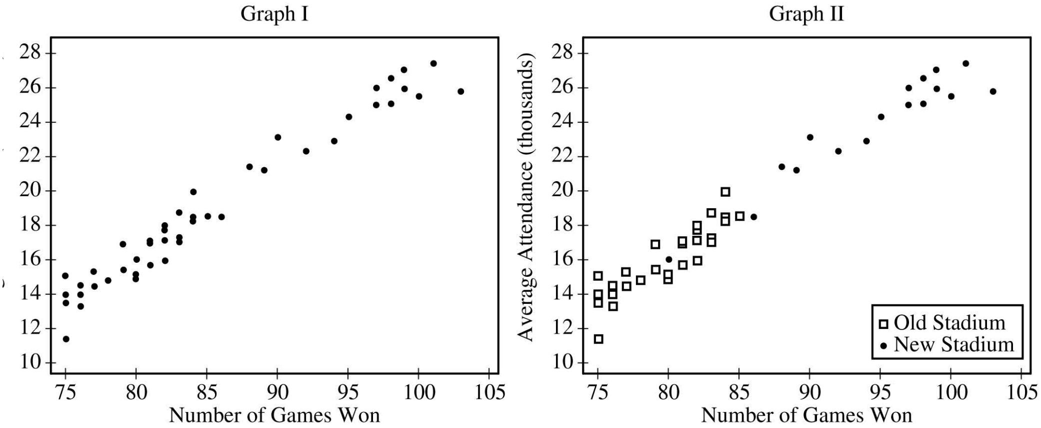
2021 ▪ Q6

Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.

[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]

(a) Compare the distributions of average attendance between the old and new stadiums.

Question 6

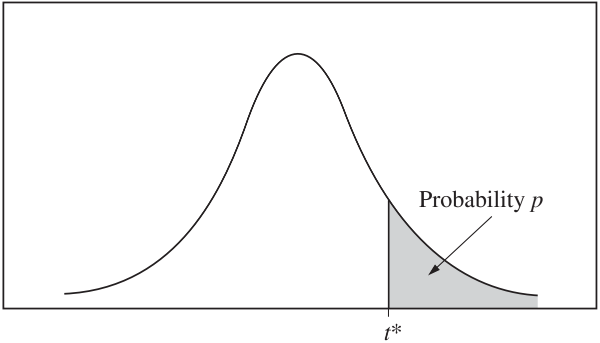


Graph I
Graph II
28
26
24
22
20
18
16
14
12
10
75
80
85
90
95
100
105
Number of Games Won
Average Attendance (thousands)
Old Stadium
New Stadium

Question 6

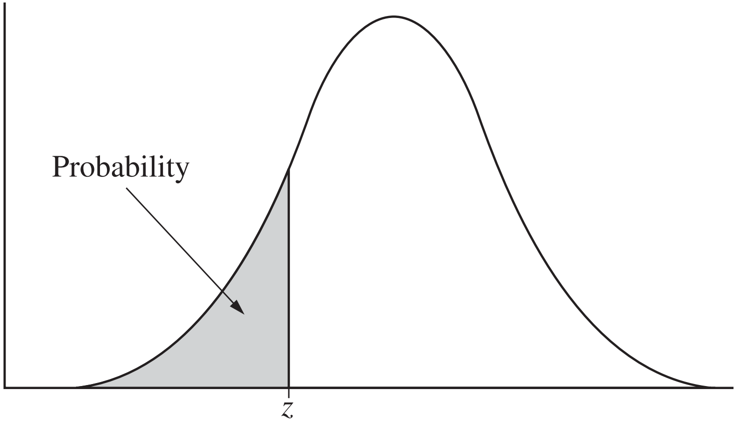
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.



Question 6

Table entry for z is the probability lying below z.



Question 6

2021 ▪ Q6

Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.

[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]

(b) The following scatterplot shows average attendance versus year.

[Scatterplot titled with y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2, x-axis "Year" from 1970 to 2020 in increments of 10 (gridlines at 1970, 1980, 1990, 2000, 2010, 2020). Legend distinguishes open squares = Old

Question 6

Stadium and filled circles = New Stadium. Old Stadium square points span roughly 1970-2000, scattered between about 11 and 20 with no strong trend (values fluctuate between roughly 13-20 throughout). New Stadium circle points span roughly 2000-2018, showing a clear upward trend from about 16-21 in the early 2000s rising steadily to about 26-27.5 by the late 2010s.]
Compare the trends in average attendance over time between the old and new stadium.

Question 6

2021 ▪ Q6

Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.

[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]

(c) Consider the following scatterplots.

[Graph I: scatterplot titled "Graph I", y-axis "Average Attendance (thousands)" (label shown on Graph II) from 10 to 28 in increments of 2, x-axis "Number of Games Won" from 75 to 105 in increments of 5. All points plotted as plain dots (stadium not

Question 6

distinguished) showing a positive, fairly linear association: points cluster from about (75,11-15) at the low end rising to about (100-103,26-28) at the high end.]
[Graph II: scatterplot titled "Graph II", same axes as Graph I ("Average Attendance (thousands)" from 10 to 28, "Number of Games Won" from 75 to 105). Same data as Graph I but now marked with open squares = Old Stadium and filled circles = New Stadium (legend shown). Old Stadium (square) points cluster in the lower-left region, roughly games won 75-86 and attendance 11-20. New Stadium (circle) points cluster in the upper-right region, roughly games won 75-103 and attendance 16-27.5, appearing to lie along a steeper positive trend than the Old Stadium points.]
(c)(i) Graph I shows the average attendance versus number of games won for each year. Describe the relationship between the variables.
(c)(ii) Graph II shows the same information as Graph I, but also indicates the old and new stadiums. Does Graph II suggest that the rate at which attendance changes as

Question 6

number of games won increases is different in the new stadium compared to the old stadium? Explain your reasoning.

Question 6

2021 ▪ Q6

Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.

[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]

(d) Consider the three variables: number of games won, year, and stadium. Based on the graphs, explain how one of those variables could be a confounding variable in the relationship between average attendance and the other variables.

Solution 6

(a) Mark scheme

- The boxplots show the new stadium tended to have higher average per-game attendance (median $\approx 25,000) thantheoldstadium (median \approx 16,000). TheIQRs (spread/variability) aresimilarbetweenthetwostadiums, thoughtheoverallrange$ … $gameattendance. Creditrequires (1) acorrectcentercomparison (new > old, withstadiumsidentifiedandanexplicitcomparisonphrase), (2) acorrectvariabilitycomparison (s$ …

Solution 6

(b) Mark scheme

- In the new stadium (years 2000-2016), average per-game attendance shows an increasing (positive) linear trend, rising from about 16,000 in 2000 to about 27,000 in 2016. In the old stadium (years 1970-1999), there is no clear increasing or decreasing trend – attendance fluctuates around an average of about 16,000 attendees per game with no obvious pattern over time. Credit requires (1) describing the new-stadium trend as increasing/positive, (2) describing the old-stadium trend as flat/no association (or positive but much less steep than the new stadium), AND (3) context identifying both stadiums, the explanatory variable (time/year), and the response variable (average attendance) or its units.

Solution 6

(c)

Mark scheme

- This leaf is the shared stem introducing Graph I (a plain scatterplot of average attendance vs. number of games won, stadium not distinguished) and Graph II (the same data, but marked by stadium: old vs. new) that parts (c)(i) and (c)(ii) below are based on. It poses no separate question of its own, so there is no independent scorable answer here – the graded reasoning about these two graphs is in leaves 6ci and 6cii.

Solution 6

(c)(i) Mark scheme

- Graph I shows a strong, positive, linear (or nearly linear) relationship between average per-game attendance and number of games won across the team's history: attendance increases roughly linearly with wins, by about 500 additional attendees per game for each additional game won, and the scatter about this linear trend is relatively small and fairly consistent (homogeneous) across the range of wins. Credit (part of a combined 5-component part-(c) score with 6cii) for describing direction (positive), form (linear), and strength (strong/moderately strong) of the relationship, ideally also noting the roughly constant variation about the trend.

Solution 6

(c)(ii) Mark scheme

- No – Graph II suggests the rate at which average attendance increases with games won is about the same for the two stadiums (or slightly larger for the old stadium): a line drawn through the old-stadium points would have roughly the same slope as (or a slightly steeper slope than) a line drawn through the new-stadium points, so there is no evidence the new stadium has a meaningfully different attendance-vs-wins rate than the old stadium. Credit (part of the same combined part-(c) score as 6ci) for correctly comparing the two implied slopes as about equal (or old $\geq new) withanexplanationreferencingfittingalinethrougheachstadium' spoints.$

Solution 6

(d) Mark scheme

- Number of games won could be a confounding variable for the effect of the new stadium on attendance. There is an association between attendance and stadium (part (a): new stadium had higher average attendance) and an association between attendance and games won (part (c): strong positive relationship); additionally, the team tended to win more games while playing in the new stadium than in the old stadium, so games won is also associated with stadium. Because these three variables are entangled, it is impossible to tell whether the higher attendance in the new stadium was actually caused by the stadium itself or simply by the team winning more games during that period – games won confounds the relationship between stadium and attendance. Credit requires (1)-(2) stating an

Solution 6

association between attendance and each of two explanatory variables (stadium, year, or wins), (3) stating an association between those same two explanatory variables, AND (4) explaining the confounding logic (the identified variable could be the true cause, or it's impossible to tell which variable is responsible).

Question 1

2018 ▪ Q1

The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output. [Scatterplot titled "Time (seconds)" vs. "Customers in Line"; y-axis (Time, seconds) from 0 to 1,250 in increments of 250; x-axis (Customers in Line) from 0 to 7. A least-squares regression line is drawn from about (0, 25) up through the point cloud to about (6.2, 1,150). Approximate data points (Customers in Line, Time): (0, 25), (1, 230), (1, 350), (2, 410), (2, 500), (2, 610), (3, 75), (3, 610), (4, 950), (5, 985), (6,

Question 1

1,150). The point at approximately (3, 75) lies well below the regression line and the rest of the data cloud.]

Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001

$S = 200.01$ $\qquad$ $R\text{-}Sq = 73.33\%$ $\qquad$ $R\text{-}Sq\ (adj) = 70.37\%$

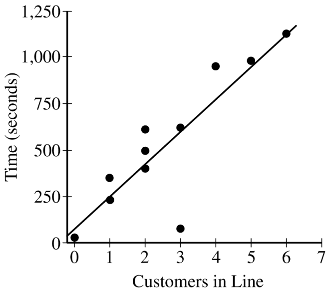
(a) Identify and interpret in context the estimate of the intercept for the least-squares regression line.

(b) Identify and interpret in context the coefficient of determination, r^2.

(c) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

Question 1

checkout. The data are shown in the following scatterplot along with the corresponding least-squ
line and computer output.



Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001

S = 200.01   R-Sq = 73.33%   R-Sq (adj) = 70.37%

Solution 1

(a) Mark scheme

- The estimate of the y-intercept is $\hat{\beta}_0 = 72.95$. In context: it is the predicted (estimated/average) time to finish checkout when there are zero other customers in line ahead of the selected customer. Full credit requires three things: (1) correctly identifying 72.95 as the intercept, (2) communicating the y-intercept concept in context that includes both time AND zero customers, and (3) indicating the value is a prediction/estimate/average (not a guaranteed value). A bare regression equation such as $\hat{y} = 72.95 + 174.40x$ does not by itself satisfy (1) unless the intercept is explicitly labeled, and cannot satisfy (3) at all. A response that calls 72.95 the slope earns no credit for (1) or (2).

Solution 1

(b) Mark scheme

- The coefficient of determination is $r^2 = 73.33\%$. Interpretation in context: about 73.33% of the variability in the time to finish checkout (including time waiting in line) is explained by (accounted for by) the linear relationship with the number of customers in line ahead of the selected customer. Full credit requires: (1) correctly identifying 73.33%, (2) a correct (possibly generic) interpretation of r^2 as the percent/proportion of variability in y explained by the linear relationship with x, and (3) that the interpretation includes context (time and number of customers) without reversing the roles of x and y. Common wrong interpretations that earn no credit for (2): variability in the *predicted* y explained, variability in 'the data' explained, or variability explained 'on average'.

Solution 1

(c) Mark scheme

- The outlier is the point with $x = 3$ and y close to 0 (a y-value strictly between 0 and 250). It is an outlier because the combination of its x and y values does not fit the pattern of the rest of the data: specifically its y-value (checkout time) is much lower than would be expected/predicted for x=3 customers in line, given the remaining data (equivalently, it has an unusually large residual relative to the other residuals). Full credit requires (1) correctly identifying the outlier (circled on the graph, or valid coordinates given: x=3 and 0<y<250) and (2) describing an unusual feature relative to the REST of the data sufficient to justify calling it an outlier. A statement with no comparison to the remaining points (e.g. just 'it has a large residual' or 'it's nowhere near the line') does not satisfy (2), nor does an

Solution 1

appeal to the outlier's influence on the regression coefficients or correlation without explicit numerical calculation.

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Question 6

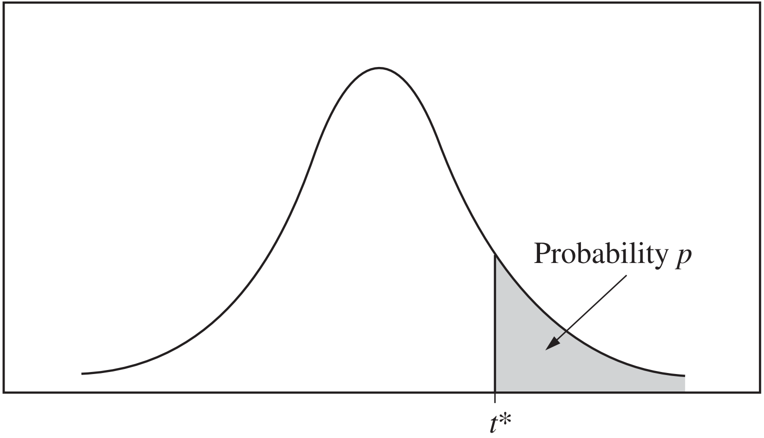
Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

(a) Does the scatterplot support the newspaper's report about number of semesters and starting salary? Justify your answer.

Question 6

Table entry for p and C is the point $t*$ with probability p lying above it and probability C lying between $-t*$ and $t*$.



Question 6

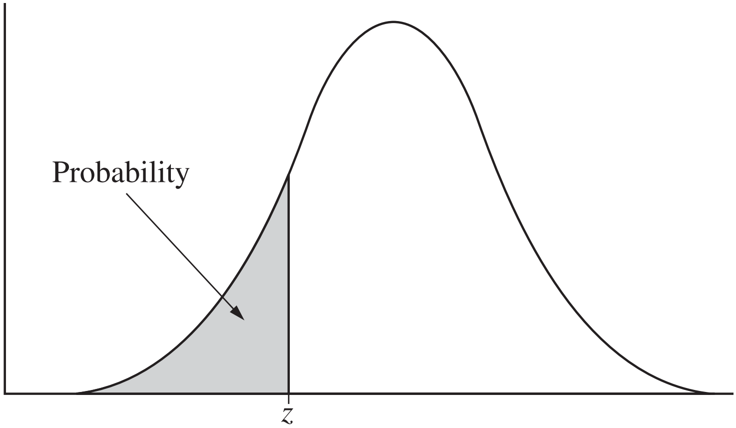
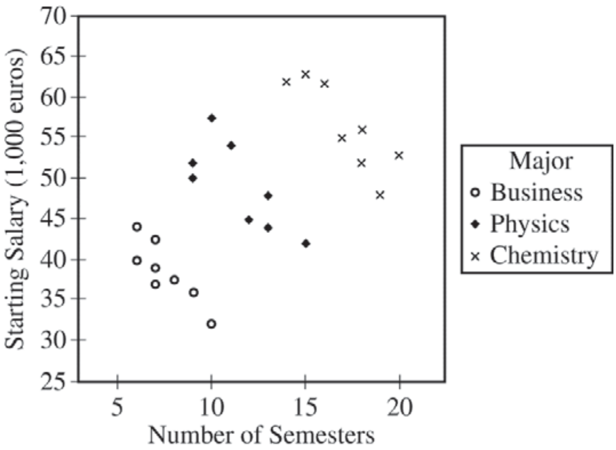


Table entry for z is the probability lying below z.

Question 6



Starting Salary (1,000 euros)
70
65
60
55
50
45
40
35
30
25
5
10
15
20
Number of Semesters
Major
Business
Physics
Chemistry

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(b) Identify the slope of the least-squares regression line, and interpret the slope in context.

An independent researcher received the data from the newspaper and conducted a new analysis by separating the data into three groups based on the major of each person. A revised scatterplot identifying the major of each person is shown below.

[Revised scatterplot, same axes as before (y-axis "Starting Salary (1,000 euros)" approximately 25 to 65, x-axis "Number of Semesters" approximately 5 to 20), with points marked by three different symbols per the legend "Major: Business (square), Physics (circle), Chemistry (diamond)". Business-major points cluster at lower semesters with lower-to-moderate salaries; Physics- and Chemistry-major points spread across higher semesters with a range of salaries, generally higher for more semesters.]

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(c) Based on the people in the sample, describe the association between starting salary and number of semesters for the business majors.

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.

[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]

The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(d) Based on the people in the sample, compare the median starting salaries for the three majors.

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(e) Based on the analysis conducted by the independent researcher, how should the newspaper report be modified to give a better description of the relationship between the number of semesters and the starting salary for the people in the sample?

Solution 6

(a)

Mark scheme

- Yes – the scatterplot supports the newspaper's report, because it shows a positive association between number of semesters needed to complete an academic program and starting salary (in general, needing more semesters is associated with a higher starting salary). Full credit requires both addressing the positive association AND using it to justify that the scatterplot supports the report.

Solution 6

(b)

Mark scheme

- The slope of the least-squares regression line is 1.1594 (from $\hat{y} = 34.018 + 1.1594x$). Interpretation: for each additional semester needed to complete an academic program, the predicted (model-estimated / average) starting salary increases by about €1,159.40 (equivalently 1.1594 thousand euros). Full credit requires the correct numerical slope value AND an in-context interpretation using nondeterministic language (e.g. "predicted"/"estimated"/"on average" – not a deterministic claim that salary WILL increase).

Solution 6

(c)

Mark scheme

- For the business majors alone, there is a strong, negative, linear association between number of semesters and starting salary: business majors who need a greater number of semesters to complete their academic program tend to have LOWER starting salaries. Full credit requires stating the association is negative, describing it as strong and/or linear, and referring to both variables (semesters, salary) in context.

Solution 6

(d)

Mark scheme

- Comparing the three majors by median starting salary: business majors have the lowest median starting salary (around €38,000), followed by physics majors (around €48,000), with chemistry majors having the highest median starting salary (around €55,000). Full credit requires a correct comparison of all three majors, with reasonable median-salary values (or explicit reference to "median starting salary").

Solution 6

(e) Mark scheme

- The newspaper's report should be modified to account for major: overall, majors that take longer to complete DO tend to have higher starting salaries (chemistry highest, physics next, business lowest – the original positive association holds across majors). However, WITHIN each individual major, students who take a greater number of semesters tend to have LOWER starting salaries (a negative association) – the opposite of the overall pattern. Full credit requires noting the negative within-major association for each major AND still acknowledging the overall positive association across majors (a Simpson's-paradox-type reversal).

Question 6

2014 ▪ Q6

Jamal is researching the characteristics of a car that might be useful in predicting the fuel consumption rate (FCR); that is, the number of gallons of gasoline that the car requires to travel 100 miles under conditions of typical city driving. The length of a car is one explanatory variable that can be used to predict FCR. Length of a car is measured in inches, and FCR is measured in gallons per 100 miles.
Jamal examined the scatterplot and determined that a linear model would be a reasonable way to express the relationship between FCR and length. A computer output from a linear regression is shown below.

Question 6

[Scatterplot titled "Graph I" plotting FCR (y-axis, 0 to 8) versus Length (inches) (x-axis, 150 to 210). Points are scattered showing a positive association between length and FCR, roughly increasing from about 3-4 at length 150-160 up to about 6-7 at length 200-210, with considerable scatter. One point is labeled A, located at approximately (175, 5.88).]

Linear Fit FCR = -1.595789 + 0.0372614 * Length

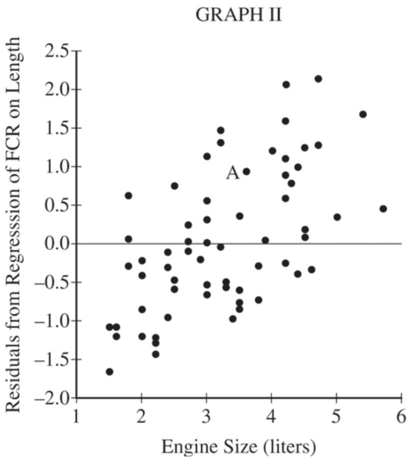
Summary of Fit

Statistic	**Value**
RSquare	0.250401
Root Mean Square Error	0.902382
Observations	66

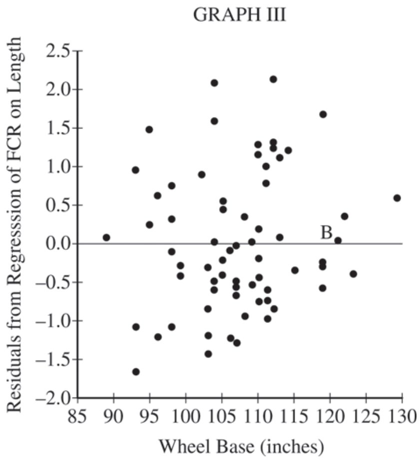
Question 6

(a) The point the graph labeled A represents one car of length 175 inches and an FCR of 5.88. Calculate and interpret the residual for the car relative to the least squares regression line.

Question 6



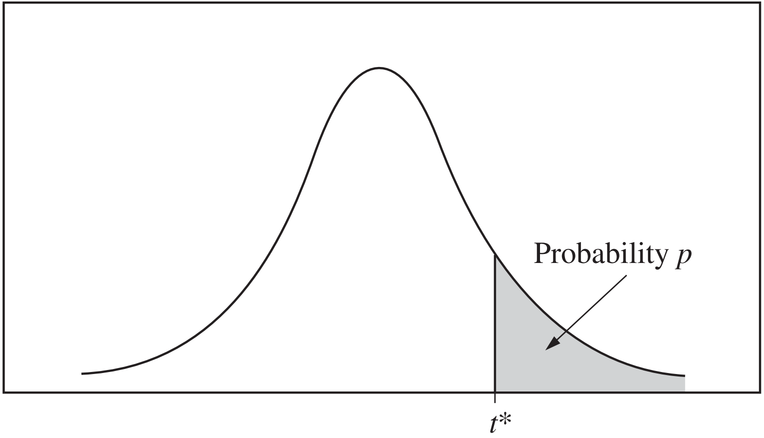
GRAPH II
Residuals from Regresssion of FCR on Length
2.5
2.0
1.5
1.0
0.5
0.0
–0.5
–1.0
–1.5
–2.0
1
2
3
4
5
6
A
Engine Size (liters)



GRAPH III
Residuals from Regression of FCR on Length
2.5
2.0
1.5
1.0
0.5
0.0
–0.5
–1.0
–1.5
–2.0
85
90
95
100
105
110
115
120
125
130
B
Wheel Base (inches)

Question 6

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.



Question 6

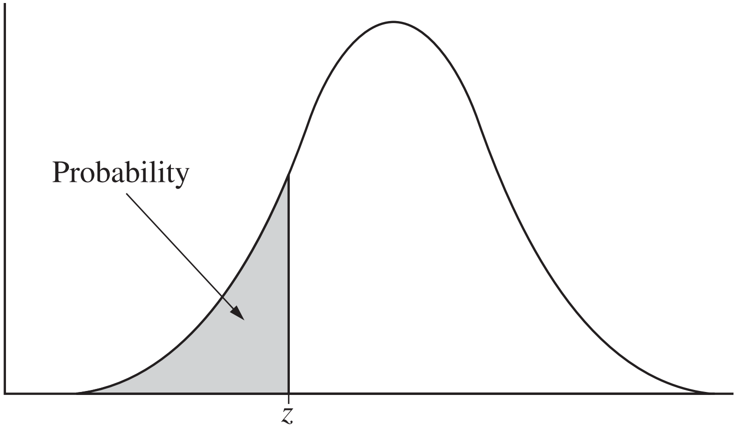


Table entry for z is the probability lying below z.