Study Session Week of 11/6

Objectives:

- I will apply previous knowledge to solve miscellaneous problems involving concepts connected with linear regression
- I will review information about linear regression including R^2, correlation, residuals, slope and y-intercept.

Agenda:

- FR practice problems

You need a calculator AND formula sheets

1. The National Directory of Magazines tracks the number of magazines published in the United States each year. An analysis of data from 1988 to 2007 gives the following computer output. The dates were recorded as years since 1988. Thus, the year 1988 was recorded as year 0. A residual plot (not shown) showed no pattern.

Predictor	Coef	StDev	T	P
Constant	13549.9	2.731	7.79	0.000
Years	325.39	0.1950	10.0	0.000
S = 836.2	R-Sq = 84.8%	R-Sq (adj) = 80.6%		

$\hat{y} = a + bx$

$\hat{\text{magazines}} = 13549.9 + 325.39 \text{ years}$

(a) What is the value of the slope of the least squares regression line? Interpret the slope in the context of this situation.

(b) What is the value of the y-intercept of the least squares regression line? Interpret the y-intercept in the context of this situation.

(c) Predict the number of magazines published in the United States in 1999.

(d) What is the value of the correlation coefficient for number of magazines published in the US and years since 1988? Interpret this correlation.

1. **Solution**

Part (a):

The slope is 325.39 magazines per year. For each year since 1988, the number of magazines published in the US increases by about 325, on average.

Part (b):

The y-intercept is 13549.9 magazines. The predicted number of magazines published in the US in 1988 (year 0) is 13550 magazines.

Part (c):

1999 is year 11 (because 1999-1988 = 11).

$$\text{magazines} = 13549.9 + 325.39\ \text{year} = 13549.9 + (325.39)(11) = 17129$$

We predict that there were 17129 magazines published in the US in 1999.

Part (d):

Since the slope is positive, the correlation coefficient is the positive square root of 0.848: $r = +\sqrt{0.848} = 0.921$

Since the correlation coefficient is +0.921, there is a strong, positive linear relationship between the number of magazines published in the US and the year.

Scoring

All parts can be essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is correct if 1) the numerical value is correct, 2) correct units are given for the slope, 3) the interpretation is correct and in context, and 4) the interpretation distinguishes between the model and the data by using words like about, approximately, or on average. The slope value may be rounded in the interpretation.

Part (a) is partially correct if the student correctly does 2 or 3 of the items listed above.

Part (a) is incorrect if the student correctly does 0 or 1 of the items listed above.

Part (b) is correct if 1) the numerical value is correct, 2) correct units are given for the y-intercept, 3) the interpretation is correct and in context, and 4) the interpretation distinguishes between the model and the data by using words like about, approximately, or predicted. The y-intercept value may be rounded in the interpretation.

Part (b) is partially correct if the student correctly does 2 or 3 of the items listed above.

Part (b) is incorrect if the student correctly does 0 or 1 of the items listed above.

Part (c) is essentially correct if the student identifies 1999 as year 11 and uses that value in a correct regression equation to find the number of magazines in 1999.

Part (c) is partially correct if the student correctly identifies 1999 as year 11 but doesn't use that value in a correct regression equation

OR

Has a correct regression equation but uses the wrong year

Part (d) is essentially correct if the value for r is correct and the interpretation is correct and in context.

Part (d) is partially correct if only one of the value for r or the interpretation in context is correct.

Each essentially correct response is worth 1 point; each partially correct response is worth half a point.

4 Complete Response

3 Substantial Response

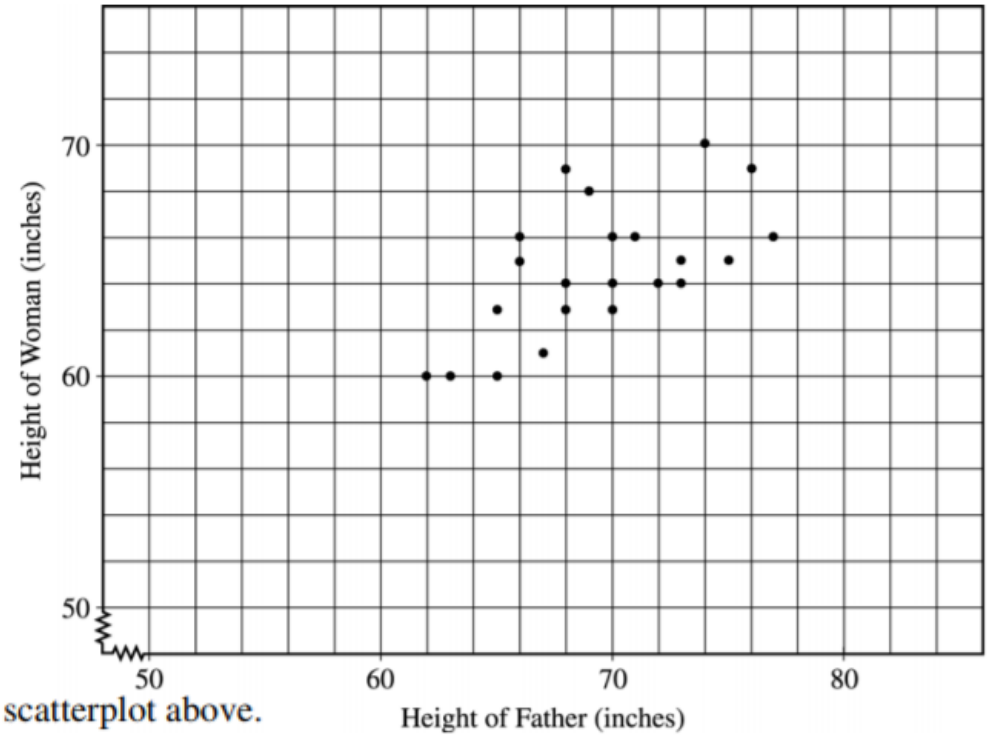
2 Developing Response

1 Minimal Response

If a response is between two scores (for example, 2.5 points), use a holistic approach to determine whether to score up or down depending on the strength of the response and communication.

2007B#4

4. Each of 25 adult women was asked to provide her own height (y), in inches, and the height (x), in inches, of her father. The scatterplot below displays the results. Only 22 of the 25 pairs are distinguishable because some of the (x,y) pairs were the same. The equation of the least squares regression line is $\hat{y} = 35.1 + 0.427x$.



(a) Draw the least squares regression line on the scatterplot above.

(b) One father's height was $x = 67$ inches and his daughter's height was $y = 61$ inches. Circle the point on the scatterplot above that represents this pair and draw the segment on the scatterplot that corresponds to the residual for it. Give a numerical value for the residual.

(c) Suppose the point $x = 84$, $y = 71$ is added to the data set. Would the slope of the least squares regression line increase, decrease, or remain about the same? Explain.

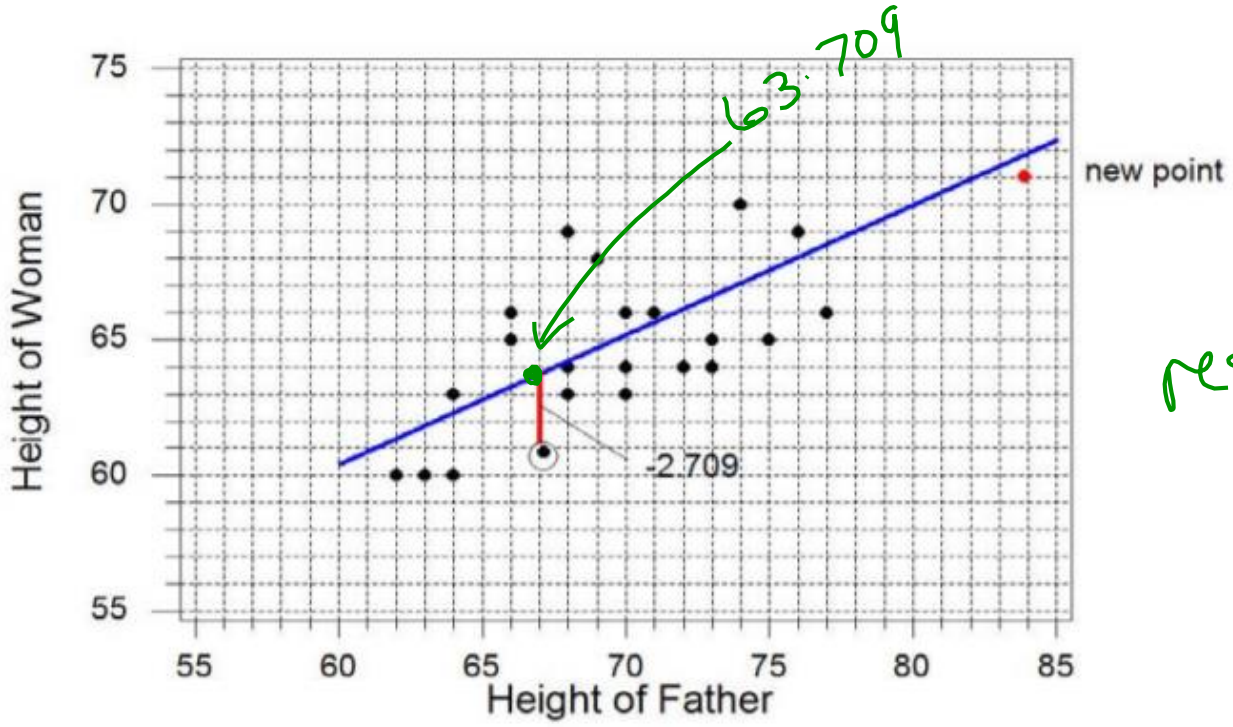
(Note: No calculations are necessary to answer this question.)

Would the correlation increase, decrease, or remain about the same? Explain.

(Note: No calculations are necessary to answer this question.)

Solution

Parts (a) and (b):



residual = real - predicted

When $x = 67$, $\hat{y} = 35.1 + 0.427(67) = 63.709$

and the residual $= y - \hat{y} = 61 - 63.709 = -2.709$.

Part (c):

See the new point indicated in the plot above. The slope would remain about the same since the new point is consistent with the linear pattern in the original plot (i.e., close to the line).

Each section is scored as either essentially correct (E), partially correct (P), or incorrect (I).

Section 1 (graphical parts of a and b) is essentially correct (E) if:

1. the regression line is drawn correctly on the scatterplot;
2. the point $(67, 61)$ is circled and the vertical segment corresponding to the residual is drawn on the scatterplot.

Section 1 is partially correct (P) if the response includes one of the above two elements.

Section 2 (numerical part of b) is essentially correct (E) if the residual is correctly comput
OR
the response states that the residual was approximated using the graph, a reasonable value
given, and the sign of the residual is correct.

Section 2 is partially correct (P) if the magnitude of the residual is correct but the sign is w

Section 3 (first part of (c)) is essentially correct (E) if it:

1. states that the slope will remain about the same (or change slightly);
2. provides an explanation based on the new point fitting the pattern in the origi

Section 3 is partially correct (P) if it states that the slope will be about the same, but the ex
incorrect.

4	**Complete Response**	
	All four sections essentially correct	
3	**Substantial Response**	
	Three sections essentially correct and no sections partially correct	
OR		
	Two sections essentially correct and two sections partially correct	
2	**Developing Response**	
	Two sections essentially correct and no sections partially correct	
OR		
	One section essentially correct and two sections partially correct	
OR		
	Four parts partially correct	
1	**Minimal Response**	
	One section essentially correct and no sections partially correct	
OR		
	No sections essentially correct and two sections partially correct	

NOTE: If the line is drawn incorrectly in part (a), and the answer to this part is consistent with the line drawn, section 3 is essentially correct (E).

Section 4 (second part of (c)) is essentially correct (E) if it:

1. states that the value of the correlation coefficient will increase;
2. provides an explanation based on the relative changes in s_x and s_y

   OR

   based on the fact that the new point fits the pattern AND is far out in the x direction,

   OR

   because the linear pattern is stronger.

Section 4 is partially correct (P) if it states that the value of the correlation coefficient will increase, but the explanation is missing or incorrect.

4 **Complete Response**

All four sections essentially correct

3 **Substantial Response**

Three sections essentially correct and no sections partially correct

OR

Two sections essentially correct and two sections partially correct

2 **Developing Response**

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and two sections partially correct

OR

Four parts partially correct

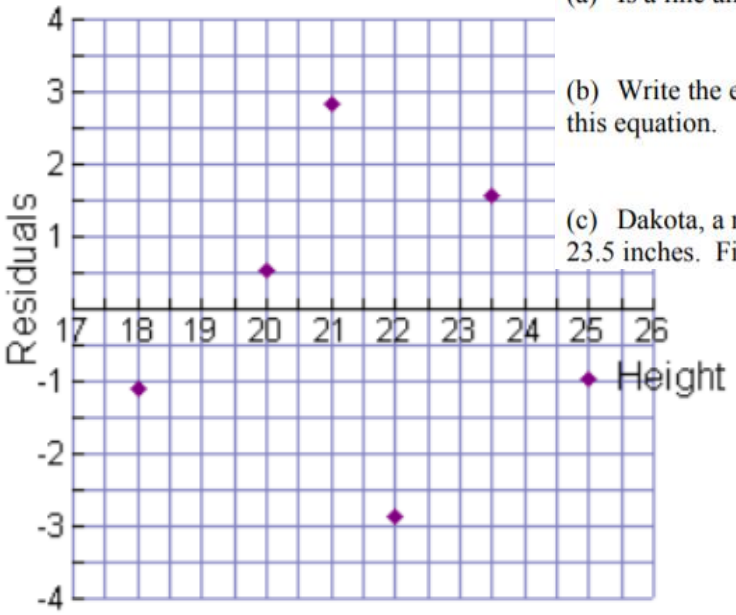
1 **Minimal Response**

One section essentially correct and no sections partially correct

OR

No sections essentially correct and two sections partially correct

2. The heights (in inches) and weights (in pounds) of six male Labrador Retrievers were measured. The height of a dog is measured at the shoulder. A linear regression analysis was done, and the residual plot and computer output are given below.



(a) Is a line an appropriate model to use for these data? What information tells you this?

(b) Write the equation of the least squares regression line. Identify any variables used in this equation.

(c) Dakota, a male Labrador, was one of the dogs measured for this study. His height is 23.5 inches. Find Dakota's predicted weight **and** Dakota's actual weight.

Predictor	Coef	StDev	T	P
Constant	−13.430	1.724	7.792	0.0000
Height	3.6956	0.4112	8.987	0.0004
S = 2.297	R-Sq = 95.3%	R-Sq (adj) = 90.6%		

2. **Solution**

Part (a):

Yes, a linear model is appropriate. The residual plot shows no pattern and a test for slope shows that there is a relationship ($H_0 : \beta_1 = 0$; $H_a : \beta_1 > 0$; where β_1 is the slope of the weight vs. height graph, $df = 4$, $t = 8.987$; p-value $= 0.0004$).

Part (b):

$\hat{y} = 3.6956x - 13.430$ where x = height in inches and $\hat{y}$ = predicted height in pounds

Part (c):

Dakota's predicted weight is $\hat{y} = (3.6956)(23.5) - 13.430 = 73.4$ pounds.

Dakota's residual (read from the graph) is approximately 1.55 to 1.6.

$$\text{residual} = y - \hat{y}$$

$$1.6 = y - 73.4$$

$$y = 73.4 + 1.6 = 75 \text{ pounds}$$

Dakota's actual weight is 75 pounds.

Scoring

Part (a) can be essentially correct (E) or incorrect (I). Parts (b) and (c) can be essentially correct (E), partially correct (P), or incorrect (I).

Part (a) can be essentially correct even if it fails to mention the linear regression t test as long as the residual graph is discussed.

Part (a) is incorrect if the only evidence for linearity given is the value of the correlation coefficient, $r = 0.976$.

Part (b) is essentially correct if the correct numerical values of both the slope and y-intercept are present in the equation **and** both variables are identified. Note: variable names (height and weight) may be used in the equation in place of x and y for full credit.

Part (b) is partially correct if the correct numerical values of both the slope and y-intercept are present in the equation, but the variables are not identified

OR

both variables are identified, but the numerical values of the slope and y-intercept are incorrect

OR

Only one numerical value is correct and only one variable is identified.

Part (c) is essentially correct if 1) the predicted weight is correct, 2) an appropriate residual (between 1.5 and 1.7) is read from the graph, and 3) the actual weight is computed correctly with work shown. The weights must be distinguished by labels (actual, predicted) or symbols (y for actual weight, $\hat{y}$ for predicted weight).

Part (c) is partially correct if two of the three tasks are completed correctly. The computation of the actual weight can be considered to be correct if an incorrect residual is substituted correctly into the residual formula.

Pact (c) is incorrect if zero or one of the tasks is completed correctly.

4 Complete Response

All parts essentially correct.

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and two parts partially correct

1 Minimal Response

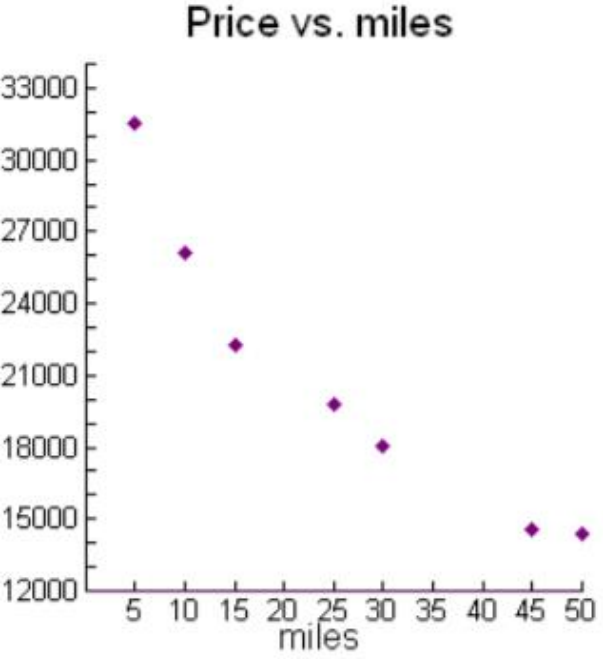
One part essentially correct and either zero or one part partially correct

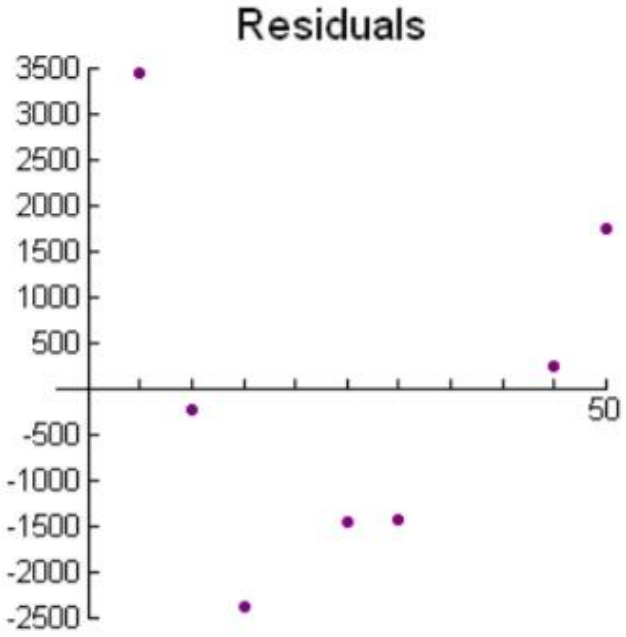
OR

No parts essentially correct and two parts partially correct

3. As more miles are driven in a car, the resale value of the car generally declines. This is called depreciation. For a certain make and model of car, information is gathered on the resale price in dollars and the number of miles driven (in thousands of miles). The scatterplot of price (y) versus miles (x), the residual plot, and the least squares regression line is shown for this data. In addition, the scatterplot, residual plot, and the accompanying best fit lines are shown for two other models using the common logarithm.

Model 1: $\hat{y} = 29784 - 343.58x$ $r = -0.9452$

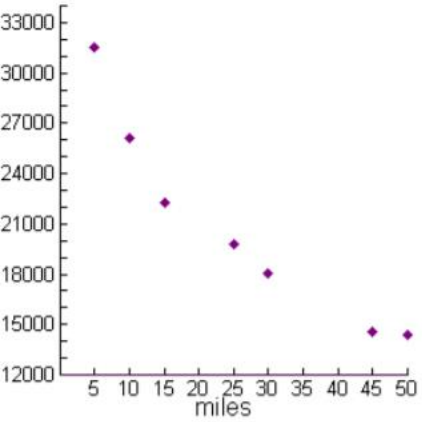


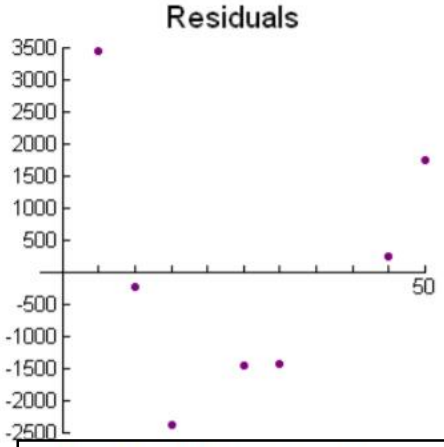


(a) Using Model 1, estimate a resale price for a car of this make and model which has been driven 35,000 miles.

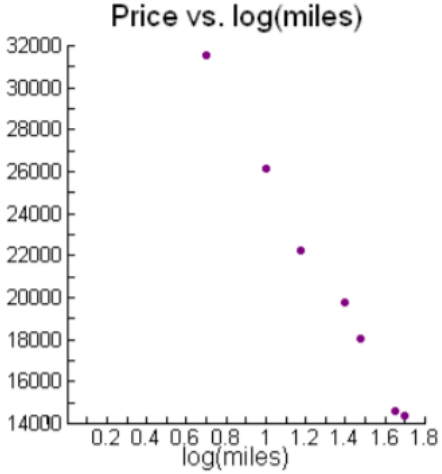
3. As more miles are driven in a car, the resale value of the car generally declines. This is called depreciation. For a certain make and model of car, information is gathered on the resale price in dollars and the number of miles driven (in thousands of miles). The scatterplot of price (y) versus miles (x), the residual plot, and the least squares regression line is shown for this data. In addition, the scatterplot, residual plot, and the accompanying best fit lines are shown for two other models using the common logarithm.

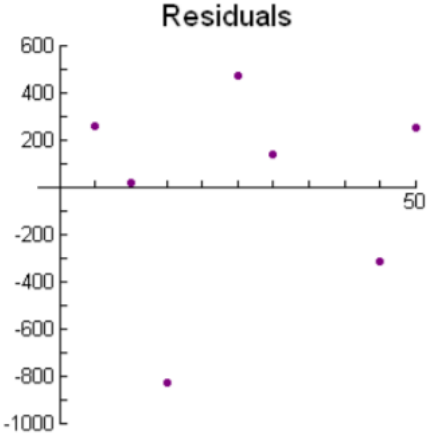
Model 1: $\hat{y} = 29784 - 343.58x$ $\quad r = -0.9452$





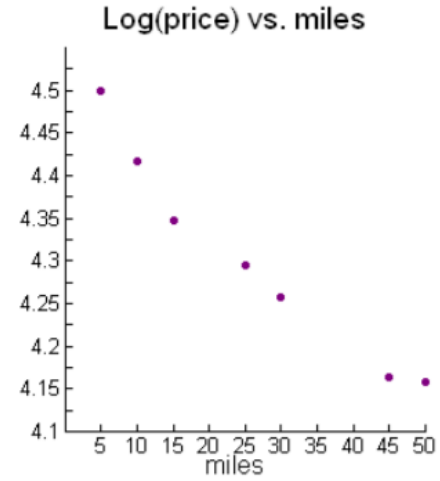
Model 3: $\hat{y} = 43254 - 17153 \log x$ $\quad r = -0.9975$

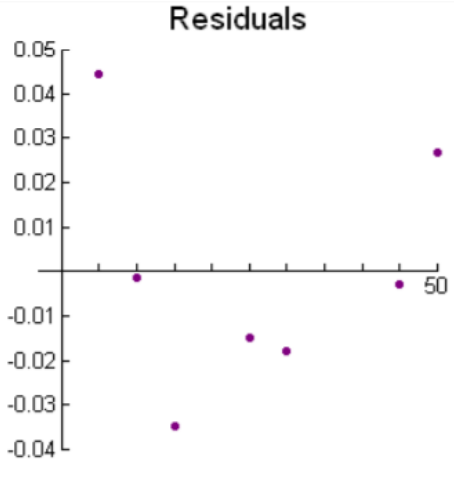




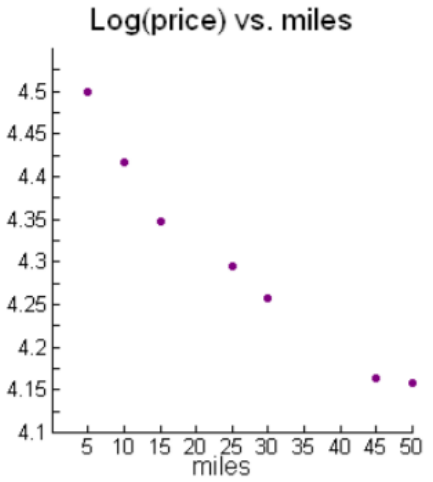
(b) Model 1 is not the most appropriate to use to compute an estimated resale price. Explain why it is not appropriate, and determine whether Model 2 or Model 3 is better.

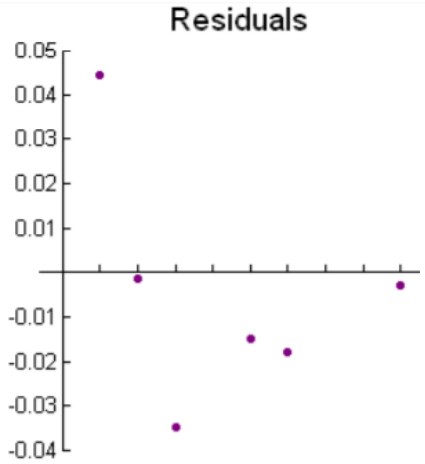
Model 2: $\log \hat{y} = 4.4901 - .0071910x$ $\quad r = -0.9765$



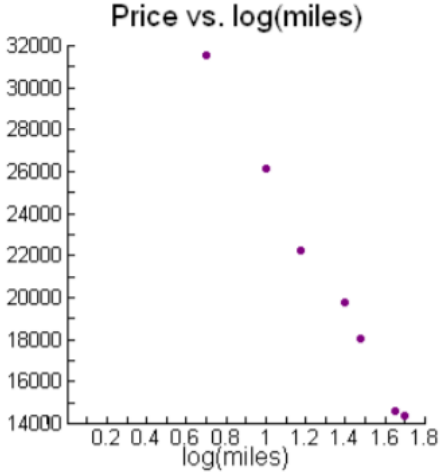


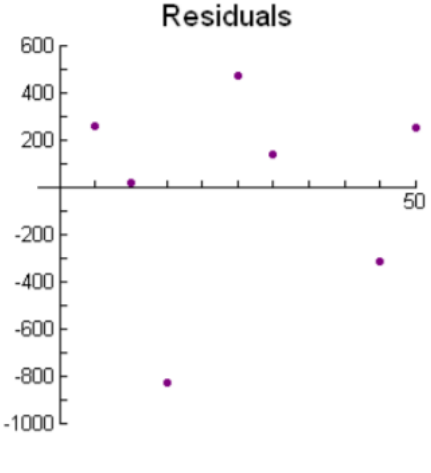
Model 2: $\log \hat{y} = 4.4901 - .0071910x$ $r = -0.9765$





Model 3: $\hat{y} = 43254 - 17153 \log x$ $r = -0.9975$





(c) Use the model you chose in part (b) to estimate a resale price for a car of this make and model that has been driven 35,000 miles.

3. **Solution**

Part (a):

Using Model 1 gives an estimated resale price of $17,758.70:

$$\hat{y} = 29784 - (343.58)(35) = \$17758.70$$

Part (b):

Model 1 is not appropriate because the scatterplot shows a slight curve and the residual plot shows a pattern. Model 3 is best because the scatterplot looks straightest and the residual plot has no pattern. (Model 2 suffers from the same problems as model 1: a curved scatterplot and a residual plot with a pattern.)

Part (c):

Using Model 3 gives an estimated resale price of $16,768.60:

$$\hat{y} = 43254 - (17153)\log 35 = \$16768.60$$

Scoring

Parts (a) and (c) can be essentially correct (E) or incorrect (I). Part (b) can be essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is essentially correct if the correct answer is given and work is shown.
Part (a) is incorrect if the answer is wrong OR if the answer is correct but no work is shown.

Part (b) is essentially correct if the student discusses the pattern in the residual plot as a shortcoming AND chooses Model 3 because its residual plot is scattered.
Part (b) is partially correct if the student correctly does only one of the above.

Part (c) is essentially correct if the student makes a correct prediction based on the model chosen in part (b) and work is shown. Note: If Model 2 is chosen, the predicted resale price is \$17,314.70:

$$\log \hat{y} = 4.4901 - (.0071910)(35) = 4.23841$$

$$\hat{y} = 10^{4.238415} = \$17314.70$$

4 Complete Response

All parts essentially correct.

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

1 Minimal Response

One part essentially correct and either zero or one part partially correct