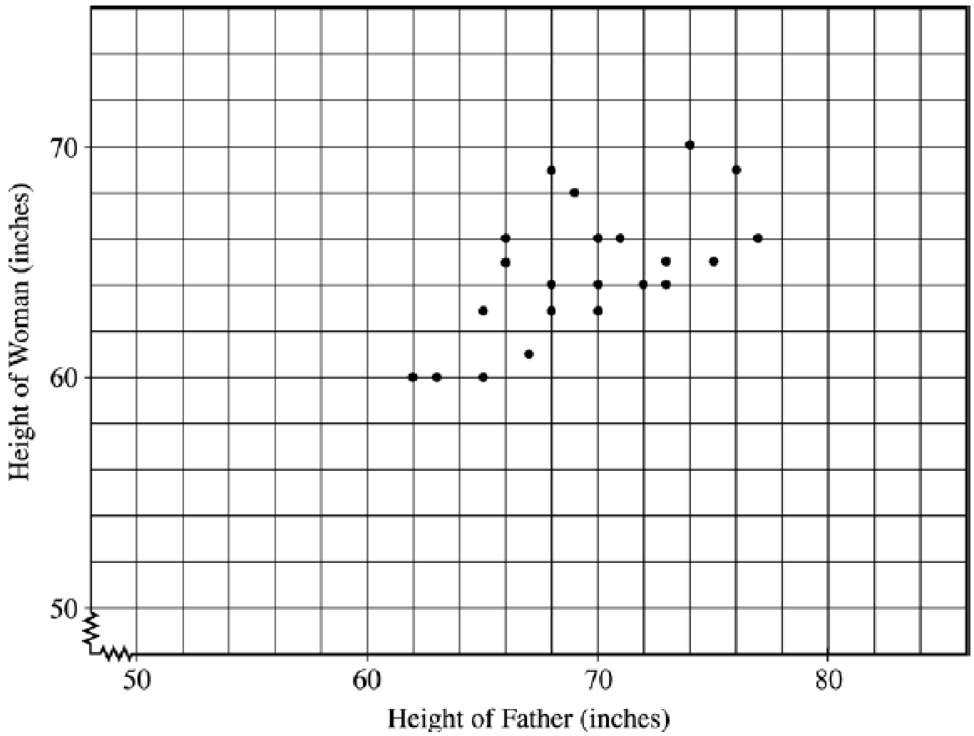
AP FRQ Review – Mr. Rich Name:________________________________ Per:_____ Seat:____

AP Stats Topic 2: Least Squares Regression

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.

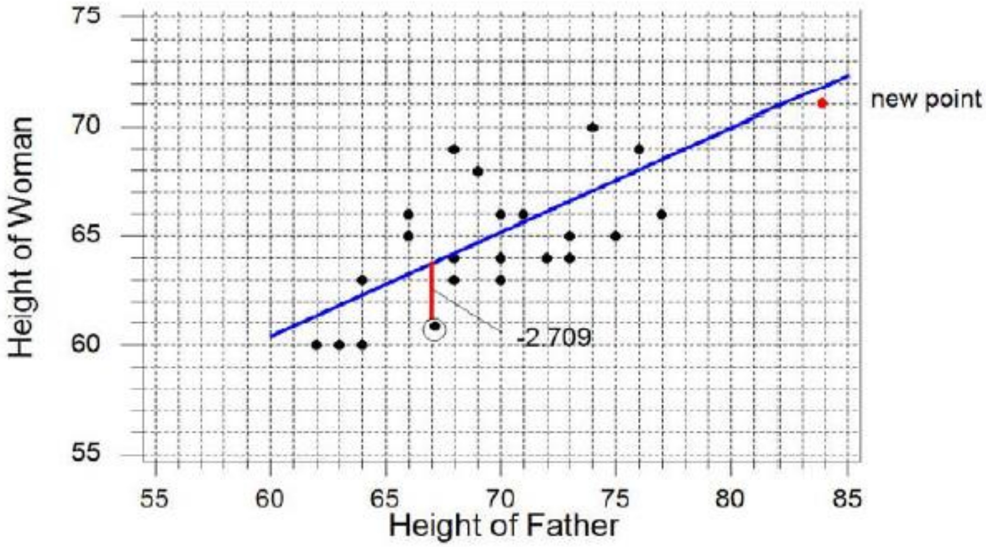
(c) Suppose the point $x = 84$, $y = 71$ is added to the data set. Would the slope of the least squares regression line increase, decrease, or remain about the same? Explain.

(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.)

Parts (a) and (b):



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).

The correlation coefficient would increase. We know that $b = r\frac{s_y}{s_x}$. The added point will increase s_x more than it will increase s_y so $\frac{s_y}{s_x}$ will be less than 1. If the slope is to stay the same, r must increase.

OR

This point fits the pattern well and has an x value that is far from $\overline{x}$.

Windmills generate electricity by transferring energy from wind to a turbine. A study was conducted to examine the relationship between wind velocity in miles per hour (mph) and electricity production in amperes for one particular windmill. For the windmill, measurements were taken on twenty-five randomly selected days, and the computer output for the regression analysis for predicting electricity production based on wind velocity is given below. The regression model assumptions were checked and determined to be reasonable over the interval of wind speeds represented in the data, which were from 10 miles per hour to 40 miles per hour.

```
Predictor          Coef    SE Coef      T        P
Constant           0.137    0.126     1.09     0.289
Wind velocity      0.240    0.019    12.63     0.000


  S = 0.237         R-Sq = 0.873         R-Sq (adj) = 0.868
```

(a) Use the computer output above to determine the equation of the least squares regression line. Identify all variables used in the equation.

(b) How much more electricity would the windmill be expected to produce on a day when the wind velocity is 25 mph than on a day when the wind velocity is 15 mph? Show how you arrived at your answer.

(c) What proportion of the variation in electricity production is explained by its linear relationship with wind velocity?

(d) Is there statistically convincing evidence that electricity production by the windmill is related to wind velocity? Explain.

Part (a):

The equation of the least squares regression line is

$$\text{predicted electricity production} = 0.137 + 0.240 \times \text{wind velocity}.$$

Part (b):

The slope coefficient of 0.240 indicates that for each additional mph of wind speed, the expected electricity production increases by 0.240 amperes. Thus, the expected electricity production is $10 \times 0.240 = 2.40$ amperes higher on a day with 25 mph wind velocity as compared to a day with 15 mph wind velocity.

Part (c):

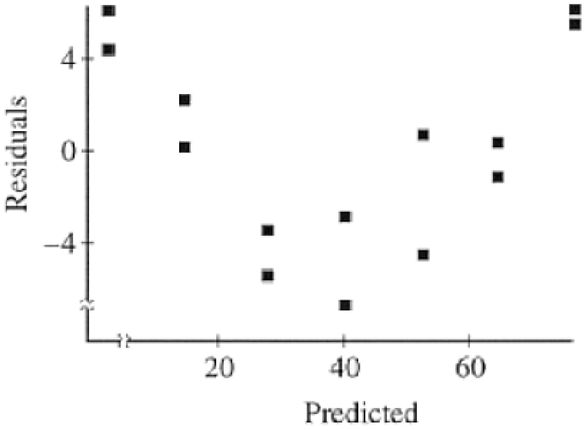
The proportion of variation in electricity production that is explained by the linear relationship with wind speed is R^2, which the regression output reports to be 0.873.

Part (d):

Yes, there is very strong statistical evidence that the population slope differs from zero, so electricity production is linearly related to wind speed. For testing the hypotheses $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$, where β represents the population slope, the output reveals that the test statistic is $t = 12.63$ and the *p*-value (to three decimal places) is 0.000. Because the *p*-value is so small (much less than both 0.05 and 0.01), the sample data provide very strong statistical evidence that electricity production is linearly related to wind speed.

In a study of the application of a certain type of weed killer, 14 fields containing large numbers of weeds were treated. The weed killer was prepared at seven different strengths by adding 1, 1.5, 2, 2.5, 3, 3.5, or 4 teaspoons to a gallon of water. Two randomly selected fields were treated with each strength of weed killer. After a few days, the percentage of weeds killed on each filed was measured. The computer output obtained from fitting a least squares regression line to the data is shown below. A plot of the residuals is provided as well.

```
Dependent variable is: percent killed
R squared = 97.2%    R squared (adjusted) = 96.9%
s = 4.505 with 14 − 2 = 12 degrees of freedom
                  Sum of                        Mean
Source            Squares          df          Square      F-ratio
Regression        8330.16           1          8330.16       410
Residual           243.589         12           20.2990
Variable         Coefficient   s.e. of Coeff   t-ratio      Prob
Constant           -20.5893        3.242         -6.35     ≤ 0.0001
No. Teaspoons       24.3929        1.204         20.3      ≤ 0.0001
```



(a) What is the equation of the least squares regression line given by this analysis? Define any variables used in this equation.

(b) If someone uses this equation to predict the percentage of weeds killed when 2.6 teaspoons of weed killer are used, which of the following would you expect?

○ The prediction will be too large.

○ The prediction will be too small.

○ A prediction cannot be made based on the information given on the computer output.

Explain your reasoning.

4 Complete Response

(a) Correctly gives equation of the regression line as $\hat{y} = -20.5893 + 24.3929x$ (Could use y.) and defines both variables: $x = \#$ of teaspoons of weed killer; $\hat{y} = \%$ killed
OR
percent killed $= -20.5893 + 24.3929(\#$ of teaspoons of weed killer$)$

(b) Substitutes $x = 2.6$ into the regression equation to get a predicted value of 42.83224, and notes that the residuals around the predicted value of 42.8 (or the middle of the predicted values) are negative. Concludes that since the residual for this prediction is negative, the prediction is expected to be too large.
OR
Notes that $x = 2.6$ is about in the middle of the explanatory values and, hence, the predicted percent killed will be close to the middle of the predicted values. Concludes that since the residual for this prediction is negative, the prediction is expected to be too large.
OR
Notes that $x = 2.6$ is in the middle of the explanatory values and that the residuals as a function of the explanatory values must exhibit the same pattern of positive and negative residuals. Since the residuals in the middle of the explanatory values are negative, the predicted value is expected to be too large.

- Arithmetic errors in (b) that give reasonable predictions (i.e. predictions between 20 and 60) should not be penalized.

3 Substantial Response

Gives a correct answer for either parts (a) or (b) and a partially correct answer to the other part.

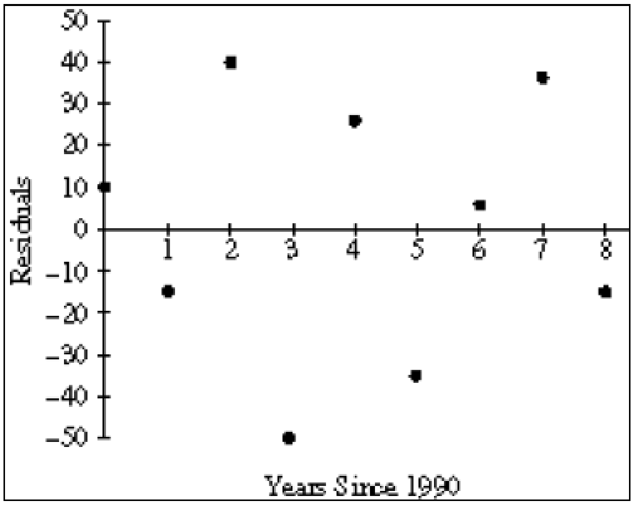
Partially correct answers include but are not limited to:

- (a) Gives the correct equation but fails to define both variables.
- (a) Switches the values for slope and y-intercept in the equation but defines both variables.
- (a) Defines both variables but gives only one correct coefficient in the correct place of the linear equation.
- (b) Correctly explains why the residual at $x = 2.6$ is negative, but incorrectly interprets this negative residual to mean that the predicted value will be too small.
- (b) States that the residual at $x = 2.6$ is negative and thus the predicted value will be too large but fails to specify where the predicted residual for $x = 2.6$ is relative to the other residuals on the residual plot.
- (b) Uses the correct model and gets an incorrect prediction that is not reasonable, but reasons correctly using this prediction. (Unreasonable predictions are below 20 or greater than 60.)
- (b) Gives incorrect residual but interprets it correctly.

2 Developing Response

Gives a correct answer to one of (a) or (b) but not a partially correct answer to the other
OR
gives a partially correct response to both (a) and (b).

Lydia and Bob were searching the Internet to find information on air travel in the United States. They found data on the number of commercial aircraft flying in the United States during the years 1990-1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



```
Predictor        Coef      Stdev   t-ratio       p
Constant      2939.93      20.55    143.09   0.000
Years         233.517      4.316     54.11   0.000

s = 33.43
```

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

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

c. What is the value of the intercept of the least squares regression line?
Interpret the intercept in the context of this situation.

d. What is the predicted number of commercial aircraft flying in 1992 ?

e. What was the actual number of commercial aircraft flying in 1992 ?

Solution:

a. Yes. Test for slope indicates that the linear model is useful (H_o: BETA is equal to 0, H_a: BETA is not equal to 0, t = 54.11, p-value = .000) and the residual plot shows no pattern, indicating a linear model is appropriate.

b. Slope = 233.517 aircraft/year
On average, the number of commercial aircraft flying in the U.S. increased by approximately 233.517 each year. (OK if rounded to 234 in interpretation)

c. Intercept = 2939.93 aircraft
Predicted number of commercial aircraft that were flying in 1990 (since x = 0 corresponds to year 1990) was 2939.93. (OK if rounded to 2940 in interpretation)

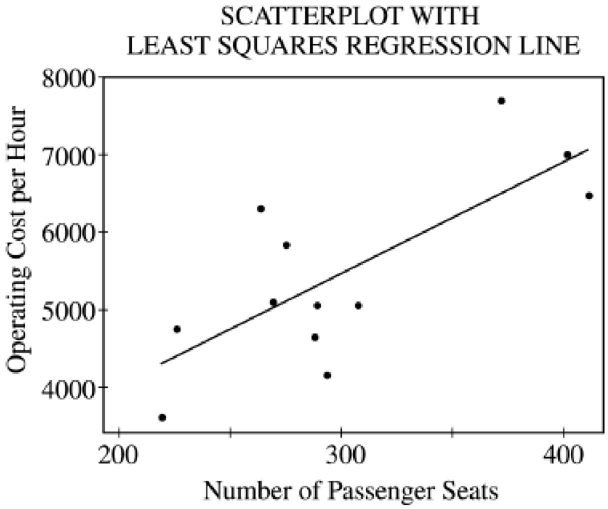
d. For 1992, x = 2, so predicted number of commercial aircraft flying is
2939.93 + 233.517(2) = 3406.964 aircraft

e. From the residual plot, the residual for 1992 is +40, so actual - predicted = 40 and
Actual = 3406.964 + 40 = 3446.964 aircraft
Since actual number flying must be an integer, actual must have been 3447.

Notes:

- Part (a) can be considered essentially correct even if it fails to mention the t test, as long as it discusses the residual plot.
- Parts (b) and (c) should draw the distinction between the model and the data. They can be considered essentially correct if the student incorporates the idea of estimation using words such as on average, predicted, approximately, about, etc.
- Parts (b) and (c) can be considered partially correct if the student (1) incorrectly identifies the values for the slope and intercept but gives an essentially correct interpretation OR (2) correctly identifies the values for the slope and intercept but gives an incomplete interpretation or an interpretation not in context for one or both.
- Parts (d) and (e) can be considered essentially correct if incorrect numbers from previous parts are correctly substituted.
- Part (e) can be considered essentially correct even if it fails to round to an integer.

Commercial airlines need to know the operating cost per hour of flight for each plane in their fleet. In a study of the relationship between operating cost per hour and number of passenger seats, investigators computed the regression of operating cost per hour on the number of passenger seats. The 12 sample aircraft used in the study included planes with as few as 216 passenger seats and planes with as many as 410 passenger seats. Operating cost per hour ranged between $3,600 and $7,800. Some computer output from a regression analysis of these data is shown below.


SCATTERPLOT WITH
LEAST SQUARES REGRESSION LINE

```
Predictor        Coef        StDev          T         P
Constant         1136        1226           0.93      0.376
Seats            14.673      4.027          3.64      0.005

S = 845.3       R-Sq = 57.0%               R-Sq (adj) = 52.7%
```

(a) What is the equation of the least squares regression line that describes the relationship between operating cost per hour and number of passenger seats in the plane? Define any variables used in this equation.

(b) What is the value of the correlation coefficient for operating cost per hour and number of passenger seats in the plane? Interpret this correlation.

(c) Suppose that you want to describe the relationship between operating cost per hour and number of passenger seats in the plane for planes only in the range of 250 to 350 seats. Does the line shown in the scatterplot still provide the best description of the relationship for data in this range? Why or why not?

Part (a):

Predicted cost = 1136 + 14.673 (number of passenger seats)

OR

$\hat{y} = 1136 + 14.673x$ where y = operating cost per hour
and x = number of passenger seats

Part (b):

- The value of the correlation coefficient

  $r = +\sqrt{0.570} = 0.755$ (r is positive because the scatterplot shows a positive association)

- The interpretation of correlation

  There is a moderate (or strong) positive linear relationship between operating costs per hour and number of passenger seats.

  OR

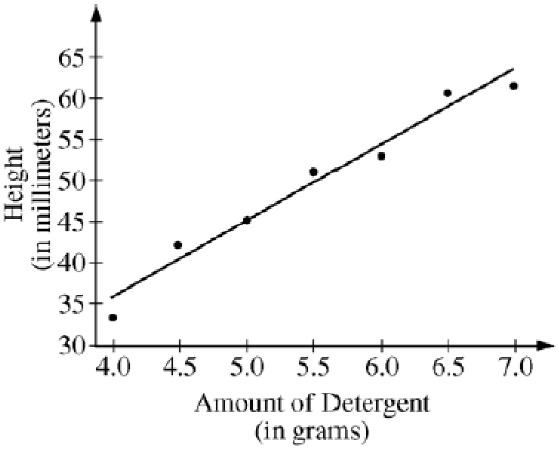
  Fifty-seven percent of the variability in operating cost per hour can be explained by a linear relationship between cost and number of passenger seats AND the relationship is positive.

Part (c):

No. The equation of the least-squares regression line is influenced by the three points in the upper right-hand corner and the two points in the lower left-hand corner of the scatterplot. The seven remaining points (with number of seats in the 250 to 350 range) would have a negative correlation. Hence, the slope of the recalculated least-squares regression line is negative.

A manufacturer of dish detergent believes the height of soapsuds in the dishpan depends on the amount of detergent used. A study of the suds' heights for a new dish detergent was conducted. Seven pans of water were prepared. All pans were of the same size and type and contained the same amount of water. The temperature of the water was the same for each pan. An amount of dish detergent was assigned at random to each pan, and that amount of detergent was added to the pan. Then the water in the dishpan was agitated for a set amount of time, and the height of the resulting suds was measured.

A plot of the data and the computer output from fitting a least squares regression line to the data are shown below.



```
Predictor      Coef    SE Coef        T        P
Constant     -2.679      4.222    -0.63    0.554
Amount       9.5000     0.7553    12.58    0.000

S = 1.99821   R-Sq = 96.9%     R-Sq(adj) = 96.3%
```

(a) Write the equation of the fitted regression line. Define any variables used in this equation.

(b) Note that $s = 1.99821$ in the computer output. Interpret this value in the context of this study.

(c) Identify and interpret the standard error of the slope.

Part (a):

The regression line is $\hat{y} = -2.679 + 9.5x$, where $\hat{y}$ represents the estimated (or predicted) mean height of the soapsuds and x represents the amount of detergent added to the pan.

Part (b):

The value $s = 1.99821$ mm is the standard deviation of the residuals. This statistic measures a typical amount of variability in the vertical distances from the observed height of the soapsuds to the regression line.

OR

The value $s = 1.99821$ mm is a measure of variation in the height of soapsuds for a given amount of detergent.

Part (c):

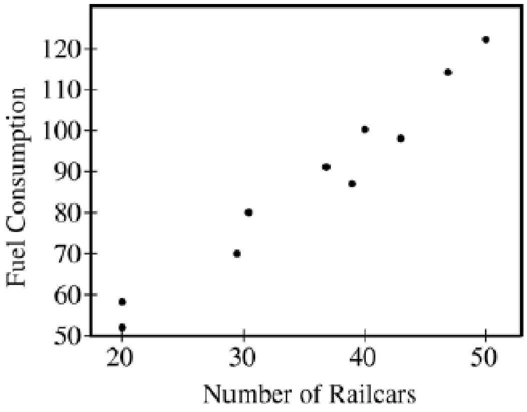
The standard error of the estimated slope parameter is 0.7553 mm per gram. Thus, the standard deviation of the estimated slope for predicting the height of soapsuds by using an amount of detergent is estimated to be 0.7553 mm per gram. This value estimates the variability in the sampling distribution of the estimated slope (i.e., how much we would expect sample slopes to vary from experiment to experiment).

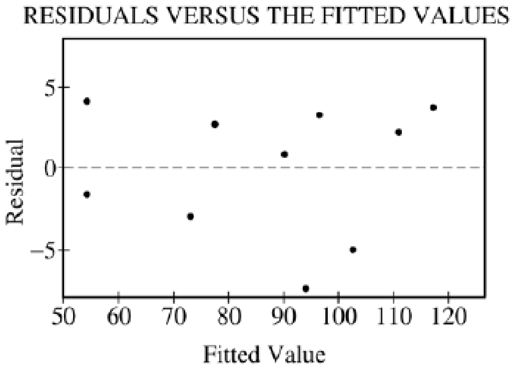
The Great Plains Railroad is interested in studying how fuel consumption is related to the number of railcars for its trains on a certain route between Oklahoma City and Omaha.

A random sample of 10 trains on this route has yielded the data in the table below.

Number of Railcars	Fuel Consumption (units/mile)
20	58
20	52
37	91
31	80
47	114
43	98
39	87
50	122
40	100
29	70

A scatterplot, a residual plot, and the output from the regression analysis for these data are shown below.




RESIDUALS VERSUS THE FITTED VALUES

```
The regression equation is
Fuel Consumption = 10.7 + 2.15 Railcars

Predictor    Coef       StDev        T        P
Constant     10.677     5.157        2.07     0.072
Railcar      2.1495     0.1396      15.40     0.000

S = 4.361  R-Sq = 96.7%   R-Sq(adj) = 96.3%
```

(a) Is a linear model appropriate for modeling these data? Clearly explain your reasoning.

(b) Suppose the fuel consumption cost is \$25 per unit. Give a point estimate (single value) for the change in the average cost of fuel per mile for each additional railcar attached to a train. Show your work.

(c) Interpret the value of r^2 in the context of this problem.

(d) Would it be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train had 65 railcars? Explain.

Part (a):

Yes, the linear model is appropriate for these data. The scatterplot shows a strong, positive, linear association between the number of railcars and fuel consumption, and the residual plot shows a reasonably random scatter of points above and below zero.

Part (b):

According to the regression output, fuel consumption will increase by 2.15 units for each additional railcar. Since the fuel consumption cost is $25 per unit, the average cost of fuel per mile will increase by approximately $\$25 \times 2.15 = \$53.75$ for each railcar that is added to the train.

Part (c):

The regression output indicates that $r^2 = 96.7\%$ or 0.967. Thus, 96.7% of the variation in the fuel consumption values is explained by using the linear regression model with number of railcars as the explanatory variable.

Part (d):

No, the data set does not contain any information about fuel consumption for any trains with more than 50 cars. Using the regression model to predict the fuel consumption for a train with 65 railcars, known as extrapolation, is not reasonable.