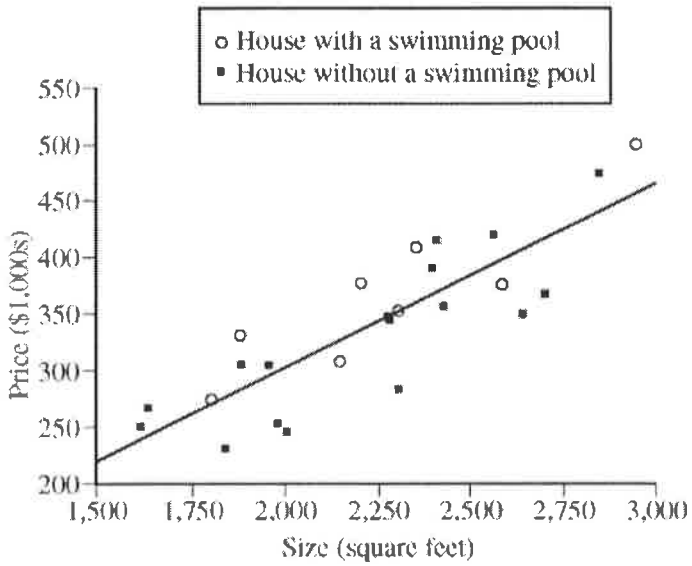
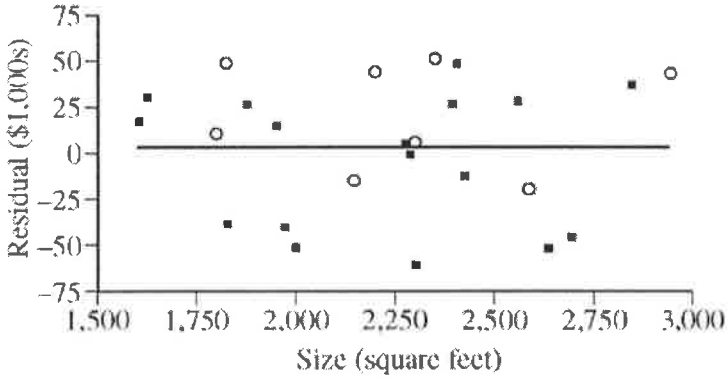
AP Statistics Exam Study Session Materials 4/17/14

2010 Exam B – Topics: Least Squares Regression Line (LSRL); Confidence Interval for Slope; Residual Plots

6. A real estate agent is interested in developing a model to estimate the prices of houses in a particular part of a large city. She takes a random sample of 25 recent sales and, for each house, records the price (in thousands of dollars), the size of the house (in square feet), and whether or not the house has a swimming pool. This information, along with regression output for a linear model using size to predict price, is shown below and on the next page.

Price ($1,000s)	Size (square feet)	Pool	Residual ($1,000s)
274	1,799	yes	6
330	1,875	yes	49
307	2,145	yes	–18
376	2,200	yes	42
352	2,300	yes	1
409	2,350	yes	50
375	2,589	yes	–23
498	2,943	yes	42
248	1,600	no	13
265	1,623	no	26
228	1,829	no	–45
303	1,875	no	22
303	1,950	no	10
251	1,975	no	–46
244	2,000	no	–57
347	2,274	no	1
345	2,279	no	–2
282	2,300	no	–69
389	2,392	no	23
413	2,410	no	44
353	2,428	no	–19
419	2,560	no	26
348	2,639	no	–58
365	2,701	no	–52
474	2,849	no	33





Linear Fit

Price = –28.144 + 0.165 Size

Summary of Fit

RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

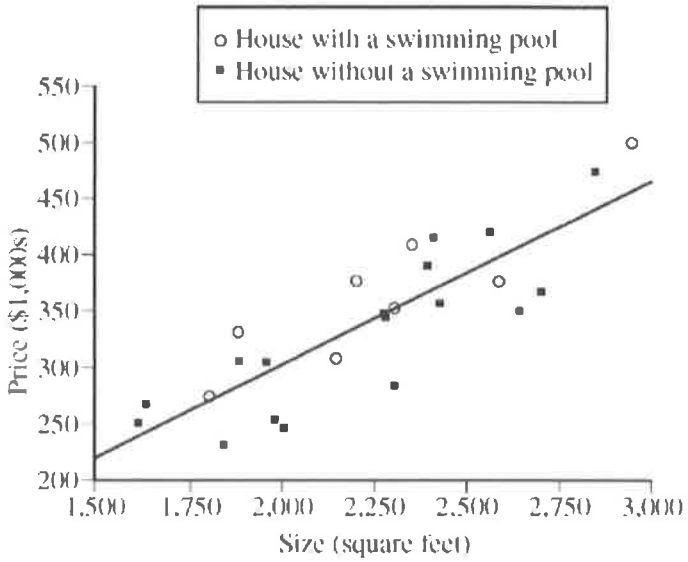
(a) Interpret the slope of the least squares regression line in the context of the study.

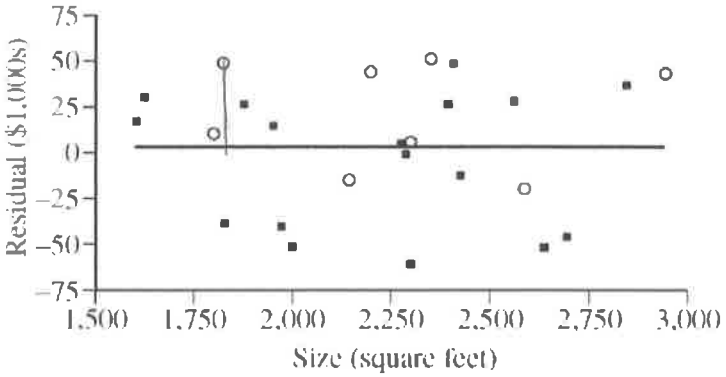
AP Statistics Exam Study Session Materials 4/17/14

2010 Exam B – Topics: Least Squares Regression Line (LSRL); Confidence Interval for Slope; Residual Plots

6. A real estate agent is interested in developing a model to estimate the prices of houses in a particular part of a large city. She takes a random sample of 25 recent sales and, for each house, records the price (in thousands of dollars), the size of the house (in square feet), and whether or not the house has a swimming pool. This information, along with regression output for a linear model using size to predict price, is shown below and on the next page.

Price ($1,000s)	Size (square feet)	Pool	Residual ($1,000s)
274	1,799	yes	6
330	1,875	yes	49
307	2,145	yes	–18
376	2,200	yes	42
352	2,300	yes	1
409	2,350	yes	50
375	2,589	yes	–23
498	2,943	yes	42
248	1,600	no	13
265	1,623	no	26
228	1,829	no	–45
303	1,875	no	22
303	1,950	no	10
251	1,975	no	–46
244	2,000	no	–57
347	2,274	no	1
345	2,279	no	–2
282	2,300	no	–69
389	2,392	no	23
413	2,410	no	44
353	2,428	no	–19
419	2,560	no	26
348	2,639	no	–58
365	2,701	no	–52
474	2,849	no	33





Linear Fit

Price = –28.144 + 0.165 Size

Summary of Fit

RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

(a) Interpret the slope of the least squares regression line in the context of the study.

For every additional square foot in a house's size, the predicted price increases by approximately $165.

(b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study.

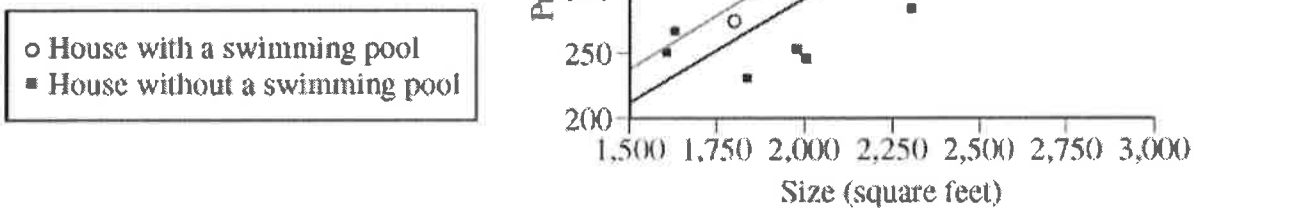
The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool.

To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.

Linear Fit (Pool = yes)
Price = −11.602 + 0.166 size

Linear Fit (Pool = no)
Price = −27.382 + 0.160 size



(d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is $(-0.099, 0.110)$. Based on this interval, is there a significant difference in the two slopes? Explain your answer.

(b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study.

A residual of 49 means the actual price of the house is \$49,000 above the predicted value from the model.

The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool.

Average residual value for houses with pool $= \frac{149}{8} = 18.6$

$= \$18,600$

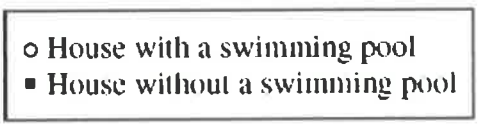
Average residual value for houses without pool $= \frac{-150}{17} = -8.8 = -\$8,800$

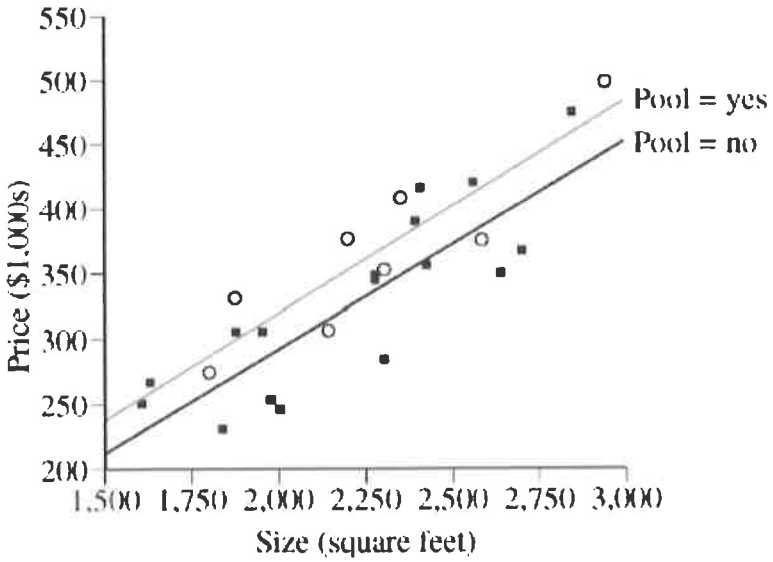
Based on the residuals, a house with a pool is on average $18,600 - (-8800) = \$27,400$ higher in value than a house without a pool that is the same size.

To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.

Linear Fit (Pool = yes)
Price = −11.602 + 0.166 size

Linear Fit (Pool = no)
Price = −27.382 + 0.160 size





(d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is $(-0.099, 0.110)$. Based on this interval, is there a significant difference in the two slopes? Explain your answer.

No, the interval does not indicate a significant difference in the two slopes because zero is on the interval.

(e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c) ?

β_1 = true slope of the LSRL for all homes in this part of the city

Linear Regression T-interval

- Linear Model is appropriate
  ↳ scatterplot looks linear
- Random sample
- House prices are independent (esp. with random sample)
- 

  Residuals appear somewhat symmetric, no outliers → approx. normal
- Spread of residuals is fairly constant for all x values

$invT(.025, 23) = 2.069$ (n-2)

$.165 \pm 2.069(.0213) =$

$(.121, .209)$

I am 95% confident that the true slope of the LSRL used to estimate house prices in this part of the city is between .121 and .209.

(e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c) ?

Since there is not a significant difference in the two slopes I will used the difference in the y-intercepts to find the vertical distance between the two regression lines.

$-11.602 - (-27.382) = 15.78 = \$15,780$

(pool) (no pool)

So a pool will increase the predicted value of a house by approximately \$15,780. Quite different from estimate in part c.

OR – Use size = 2250

$\widehat{price}_{(pool)} = -11.602 + .166(2250)$

$= 361.898 = \$361,898$

$\widehat{price}_{(no\ pool)} = -27.382 + .16(2250)$

$= 332.618 = \$332,618$

$361,898 - 332,618 =$

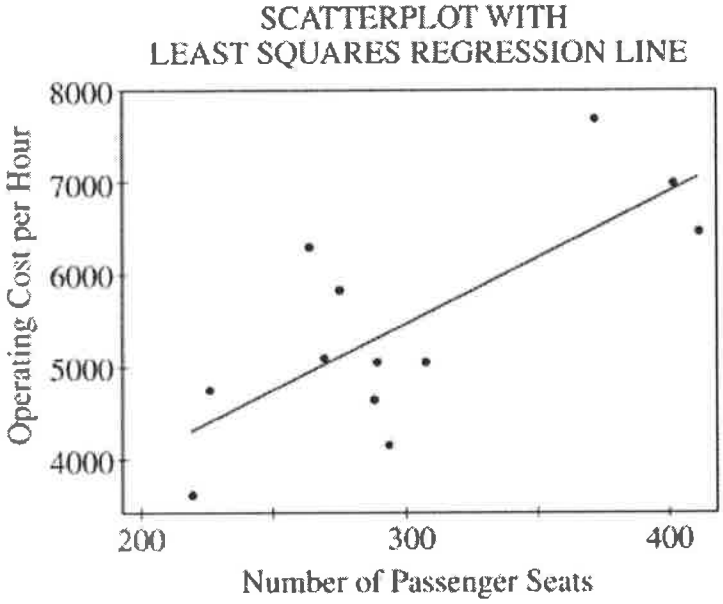
\$29,280

So a pool will increase predicted price of a house by approximately \$29,280.

this is very similar to the estimate from the residuals in part c.

2002 Exam – Topics: LSRL; Correlation and Linearity; Outliers and Influential Points

4. 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.



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%	

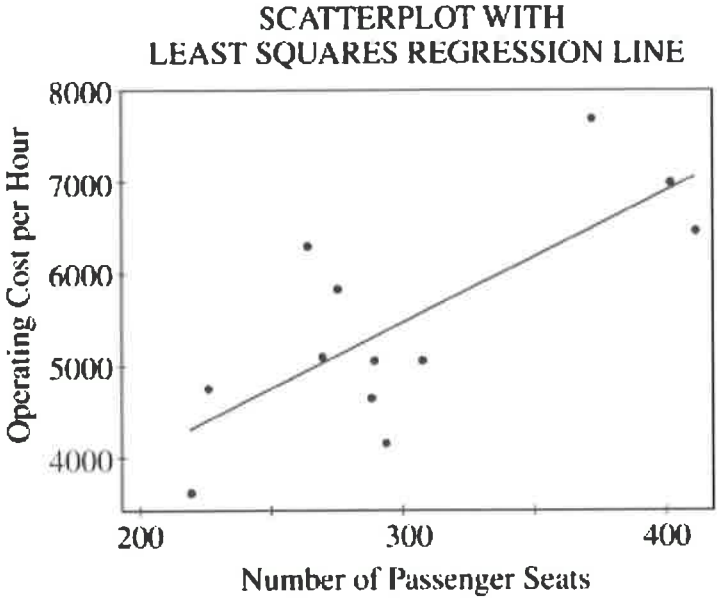
(a) What is the equation of the least squares regression line that describes the relationship between operating cost per hour and number of passenger seats in the plane? Define any variables used in this equation.

(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?

2002 Exam – Topics: LSRL; Correlation and Linearity; Outliers and Influential Points

4. **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.**



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.**

Operating Cost per hour = 1136 + 14.673 (seats)

(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.**

r = .755 There is a moderate, positive linear relationship between operating cost per hour and number of passenger seats.

(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?**

No. The LSRL is being influenced by the values around 400 seats as well as those near 225 seats. Values from 250-350 seats show a negative association instead of positive.

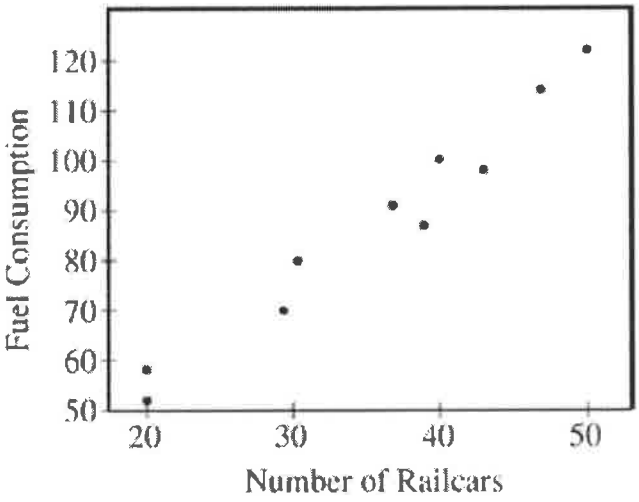
2005 Exam – Topics: LSRL; Residual Plots; Estimating Population Parameters; Generalizing results

3. **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.**

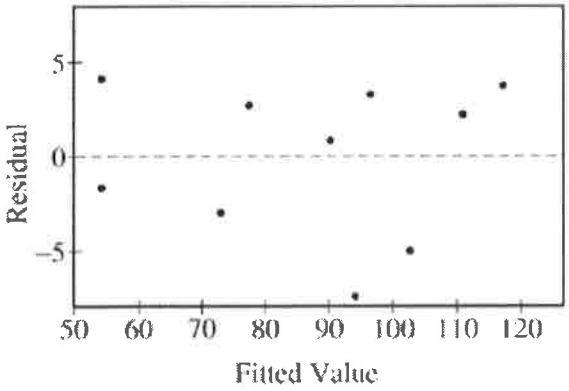
A random sample of 10 trains on this route has yielded the data in the table below.

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



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.

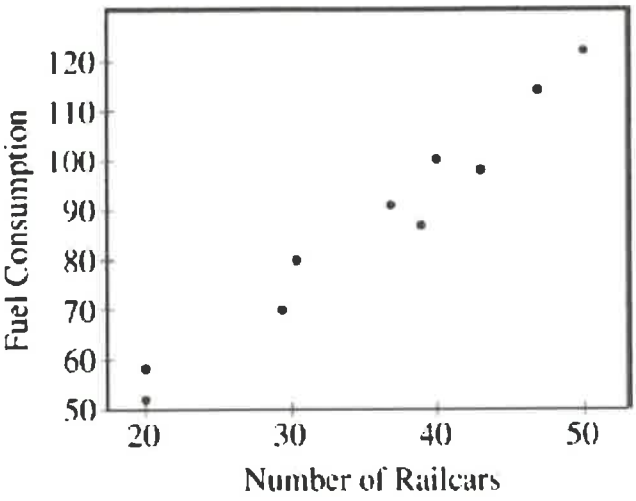
2005 Exam – Topics: LSRL; Residual Plots; Estimating Population Parameters; Generalizing results

3. **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.**

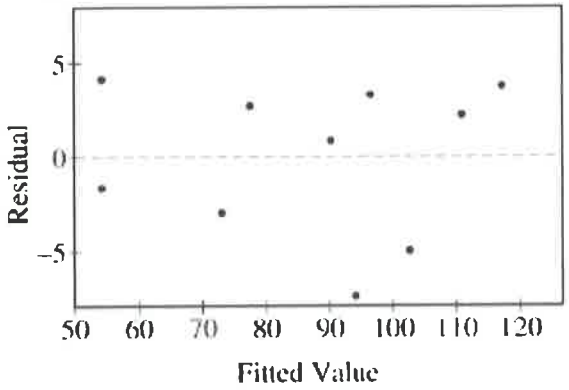
A random sample of 10 trains on this route has yielded the data in the table below.

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



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.**

Yes. The scatterplot shows a strong linear relationship between fuel consumption and number of railcars, and there is no obvious pattern in the residuals.

(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.**

2.15(25) = $53.75 for each additional railcar added (slope)

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

96.7% of the variation in fuel consumption can be explained by the variation in the number of railcars.

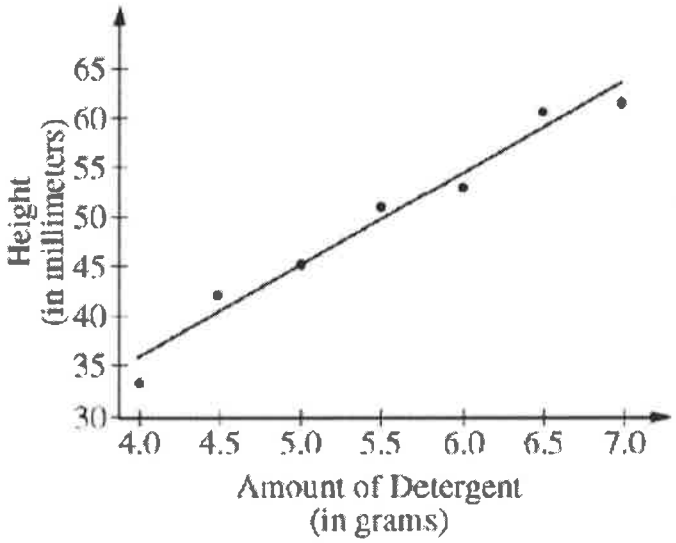
(d) **Would it be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train had 65 railcars? Explain.**

No. 65 railcars is outside the data set far enough that it would not be reasonable to predict its fuel consumption – called extrapolation.

2006 Exam – Topics: LSRL; Slope of a Regression Line; Sampling Distributions

2. 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.

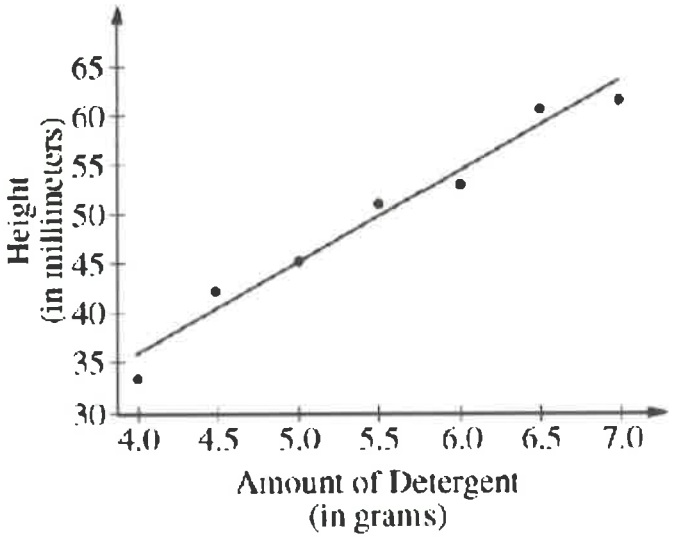
(b) Note that $s = 1.99821$ in the computer output. Interpret this value in the context of this study.

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

2006 Exam – Topics: LSRL; Slope of a Regression Line; Sampling Distributions

2. **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.



standard error

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.**

$\widehat{\text{Height}} = -2.679 + 9.5(\text{Amount of Detergent})$

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

S is the standard deviation of the residuals. This value describes the variation in the height of soapsuds for a given amount of detergent.
OR Height of soapsuds are typically within 1.99 mm of their predicted values.

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

The standard error of the slope is .755 mm/gram of detergent. This value describes how much we would expect the sample slope to vary between samples.

2007 Exam – Topics: LSRL; Test for the Slope of a LSRL

6. A study was designed to explore subjects' ability to judge the distance between two objects placed in a dimly lit room. The researcher suspected that the subjects would generally overestimate the distance between the objects in the room and that this overestimation would increase the farther apart the objects were.

   The two objects were placed at random locations in the room before a subject estimated the distance (in feet) between those two objects. After each subject estimated the distance, the locations of the objects were rerandomized before the next subject viewed the room.

   After data were collected for 40 subjects, two linear models were fit in an attempt to describe the relationship between the subjects' perceived distances (y) and the actual distance, in feet, between the two objects.

   $$\text{Model 1: } \hat{y} = 0.238 + 1.080 \times (\text{actual distance})$$

   The standard errors of the estimated coefficients for Model 1 are 0.260 and 0.118, respectively.

   $$\text{Model 2: } \hat{y} = 1.102 \times (\text{actual distance})$$

   The standard error of the estimated coefficient for Model 2 is 0.393.

(a) Provide an interpretation in context for the estimated slope in Model 1.

(b) Explain why the researcher might prefer Model 2 to Model 1 in this context.

(c) Using Model 2, test the researcher's hypothesis that in dim light participants overestimate the distance, with the overestimate increasing as the actual distance increases. (Assume appropriate conditions for inference are met.)

2007 Exam – Topics: LSRL; Test for the Slope of a LSRL

6. **A study was designed to explore subjects' ability to judge the distance between two objects placed in a dimly lit room. The researcher suspected that the subjects would generally overestimate the distance between the objects in the room and that this overestimation would increase the farther apart the objects were.**

The two objects were placed at random locations in the room before a subject estimated the distance (in feet) between those two objects. After each subject estimated the distance, the locations of the objects were rerandomized before the next subject viewed the room.

After data were collected for 40 subjects, two linear models were fit in an attempt to describe the relationship between the subjects' perceived distances (y) and the actual distance, in feet, between the two objects.

Model 1: $\hat{y} = 0.238 + 1.080 \times$ (actual distance)

The standard errors of the estimated coefficients for Model 1 are 0.260 and 0.118, respectively.

Model 2: $\hat{y} = 1.102 \times$ (actual distance)

The standard error of the estimated coefficient for Model 2 is 0.393.

(a) Provide an interpretation in context for the estimated slope in Model 1.

For every additional foot in distance between the two objects, there is an increase of approximately 1.08 feet in the perceived distance apart.

(b) Explain why the researcher might prefer Model 2 to Model 1 in this context.

The y-intercept in model 1 doesn't make sense since we would expect a person to distinguish if two objects are side by side. A y-intercept of zero, as in model 2 makes more sense.

(c) Using Model 2, test the researcher's hypothesis that in dim light participants overestimate the distance, with the overestimate increasing as the actual distance increases. (Assume appropriate conditions for inference are met.)

β_1 = true slope between the perceived distances and actual distances

$H_0: \beta_1 = 1$ (will estimate same dist... $y=x$)

$H_a: \beta_1 > 1$ (will overestimate dist)

t-test for slope

$t = \frac{1.102 - 1}{.393} = .2595$

$df = 40 - 2 = 38$

p-value $= P(t > .2595) =$ tcdf(.2595, 100, 38) $= .3983$

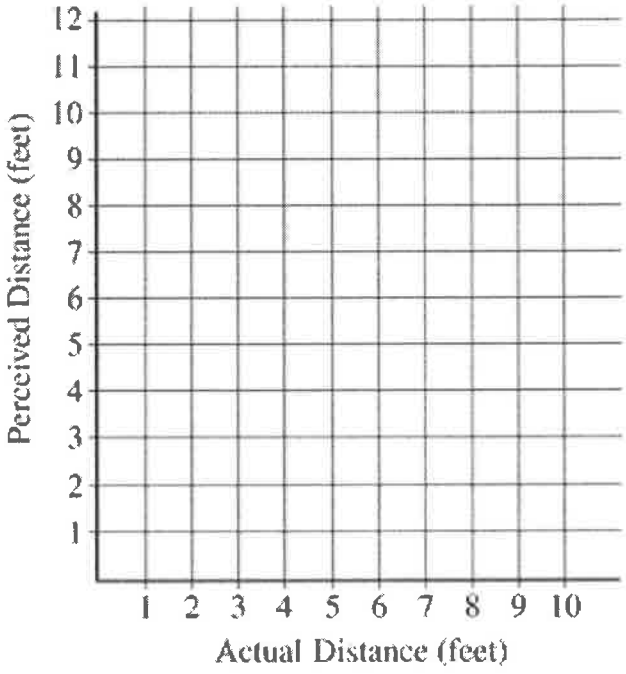
Since the p-value of .398 > α of .05, I fail to reject H_0. There is insufficient evidence to suggest the subjects overestimate the distances between objects as the actual distance increases.

The researchers also wanted to explore whether the performance on this task differed between subjects who wear contact lenses and subjects who do not wear contact lenses. A new variable was created to indicate whether or not a subject wears contact lenses. The data for this variable were coded numerically (1 = contact wearer, 0 = noncontact wearer), and this new variable, named "contact," was included in the following model.

$$\text{Model 3: } \hat{y} = 1.05 \times (\text{actual distance}) + 0.12 \times (\text{contact}) \times (\text{actual distance})$$

The standard errors of the estimated coefficients for Model 3 are 0.357 and 0.032, respectively.

(d) Using Model 3, sketch the estimated regression model for contact wearers and the estimated regression model for noncontact wearers on the grid below.



(e) In the context of this study, provide an interpretation of the estimated coefficients for Model 3.

The researchers also wanted to explore whether the performance on this task differed between subjects who wear contact lenses and subjects who do not wear contact lenses. A new variable was created to indicate whether or not a subject wears contact lenses. The data for this variable were coded numerically (1 = contact wearer, 0 = noncontact wearer), and this new variable, named "contact," was included in the following model.

Model 3: $\hat{y} = 1.05 \times (\text{actual distance}) + 0.12 \times (\text{contact}) \times (\text{actual distance})$

The standard errors of the estimated coefficients for Model 3 are 0.357 and 0.032, respectively.

1.05 .12

(d) Using Model 3, sketch the estimated regression model for contact wearers and the estimated regression model for noncontact wearers on the grid below.

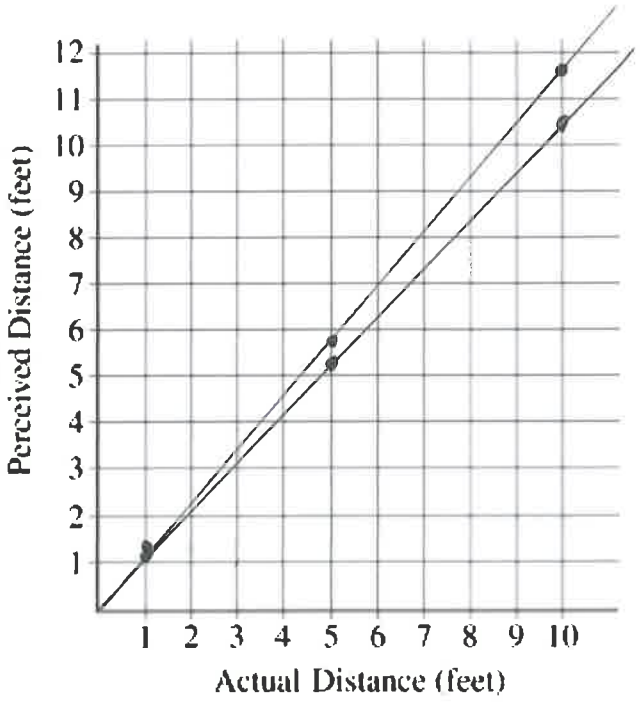
No contacts

$\hat{y} = 1.05(\text{dist})$

Contacts

$\hat{y} = 1.05(\text{dist}) + .12(\text{dist})$

$\hat{y} = 1.17(\text{dist})$



(e) In the context of this study, provide an interpretation of the estimated coefficients for Model 3.

For every additional foot that the two object are apart, the perceived distance apart increases by approximately 1.05 feet. For those who wear contacts, there is an additional increase of approximately .12 feet in the perceived distance for every foot the objects are apart.