3.2A

Least Squares

Regression

Aug 18-10:20 AM

A **regression line** summarizes the relationship between two variables, but only in settings where one of the variables helps explain or predict the other.

Regression line - a line that describes how a response variable y changes as an explanatory variable x changes. It is often used to make predictions. It is a model of the data.

If you find the equation of the regression line, it assumes the data is linearly associated. We need to check other things to make sure this assumption is valid.

Aug 18-10:21 AM

$\hat{y} = a + bx$

- $\hat{y}$ (read “y hat”) is the **predicted value** of the response variable *y* for a given value of the explanatory variable x.
- *b* is the **slope**, the amount by which *y* is predicted to change when *x* increases by one unit.
- *a* is the **y intercept**, the predicted value of *y* when *x* = 0.
- Sometimes it is better to use the variable names rather than x and y
- Ex. $\widehat{\text{Miles per Gallon}}$ = 1.9876 - 3.5674(Car weight in tons)

Aug 18-10:24 AM

AP Exam Common Errors:

1. When writing the equation of a regression line, students often forget the hat on the y. It is better to use the actual variable names.

2. Be sure to state that the slope is the predicted change. For example, a response that says, “The fat gain will go up 0.00344 kg for each added calorie” mistakenly implies that all data values will be on the regression line.

Aug 18-10:27 AM

Extrapolation

Use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line. Such predictions are often inaccurate.

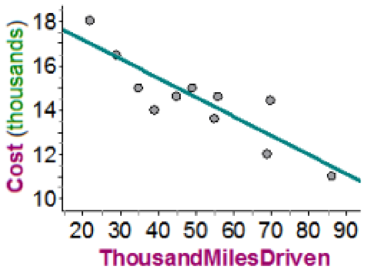
Don’t make predictions using values of x that are much larger or much smaller than those that actually appear in your data!

Aug 18-10:30 AM

Example

The following data show the number of miles driven and advertised price for 11 used Honda CR-Vs from the 2002 - 2006 model years.

Thousand Miles Driven	Cost (dollars)
22	17998
29	16450
35	14998
39	13998
45	14599
49	14988
55	13599
56	14599
69	11998
70	14450
86	10998



Aug 18-10:31 AM

a) Discuss DOFS in context of the problem.

Calculator - Find r value.

The graph shows a strong negative linear association between cost and miles driven (in thousands) for 2002-2006 Honda CRVs. That is, as the number of miles increase the cost decreases linearly. There does not appear to be any outliers.

Aug 18-10:32 AM

b) Find the regression equation. Identify the slope and y intercept of the regression line. Interpret each value in context.

The slope is -86.2.

The predicted price of a Honda CRV will decrease $86.20 for each additional thousand miles driven.

OR

For each additional thousand miles driven, the price will decrease, on average, by $86.20.

The y-intercept is 18773.

The predicted price of a 2002-2006 Honda CRV with 0 miles is 18773.

Oct 5-5:42 PM

c) Predict the price for a Honda CR-V with 50,000 miles.

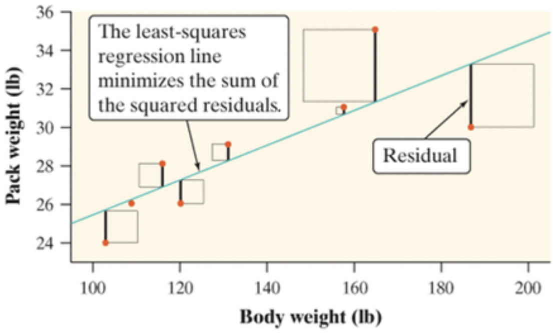
d) Predict the price for a Honda CR-V with 250,000 miles.

No line will pass exactly through all the points in a scatterplot. A good regression line makes the vertical distances of the points from the line as small as possible.

Aug 18-10:32 AM

Definition:

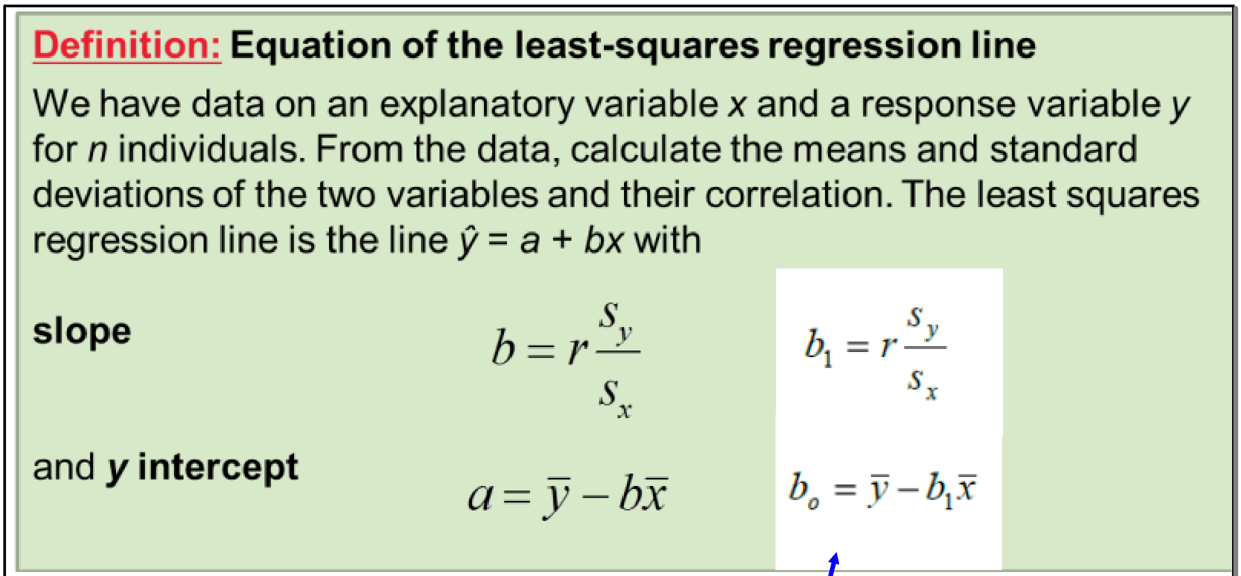
The **least-squares regression line** of *y* on *x* is the line that makes the sum of the squared residuals as small as possible.



$$\boxed{\hat{y} = b_o + b_1 x}$$

This is how it appears on AP Formula Sheet!!

Aug 18-10:38 AM



Definition: Equation of the least-squares regression line

We have data on an explanatory variable *x* and a response variable *y* for *n* individuals. From the data, calculate the means and standard deviations of the two variables and their correlation. The least squares regression line is the line $\hat{y} = a + bx$ with

slope $b = r\frac{s_y}{s_x}$ $b_1 = r\frac{s_y}{s_x}$

and ***y* intercept** $a = \bar{y} - b\bar{x}$ $b_o = \bar{y} - b_1\bar{x}$

These formulas are on the AP formula sheet!!

You Try:

The equation of the least-squares regression line is $\hat{y} = 106.5 - 0.782x$, where $\hat{y}$ = July temperature in F° and x = latitude.

1. Interpret the slope of the least-squares line in the context of the problem.

The predicted average July temperature will decrease 0.782 degrees F for each increase of one degree latitude.

2. Predict the average July temperature for a city at a latitude of 42 degrees. Show your work.

3.2B

Least Squares

Regression

Oct 5-5:59 PM

Residual - the difference between an observed value of the response variable and the value predicted by the regression line.

residual = actual y - predicted y (AP)

$$= \quad y \quad - \quad \hat{y}$$

Oct 5-5:49 PM

Example

McDonalds Beef Sandwiches

Carbs (g)	31	33	34	37	40	40	45	37	38
Fat (g)	9	12	23	19	26	42	29	24	28

Oct 5-5:55 PM

(a) Calculate the equation of the least-squares regression line using technology. Make sure to define variables! Sketch the scatterplot with the graph of the least-squares regression line.

Oct 5-5:56 PM

(b) Interpret the slope and *y*-intercept in context.

(c) Calculate and interpret the residual for the Big Mac, with 45g of carbs and 29g of fat.

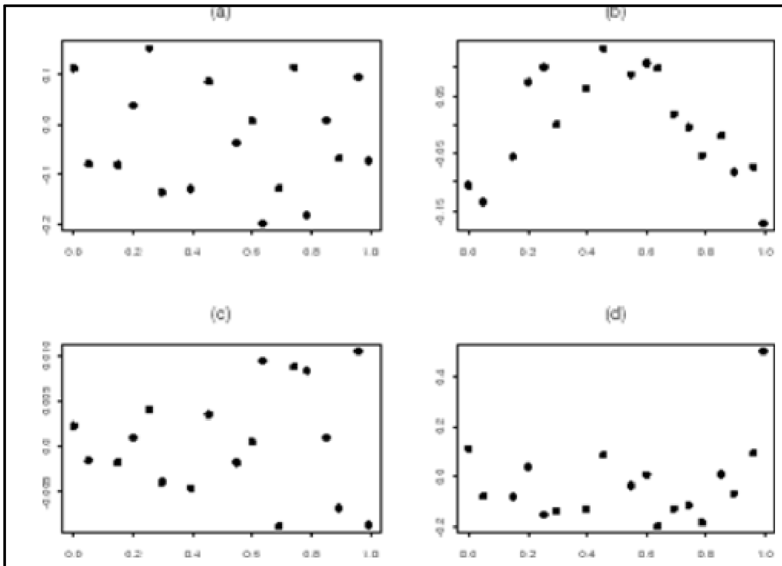
Residual Plot

A scatterplot of the residuals against the explanatory variable.

- Help us assess how well a regression line fits the data
- Should not show an obvious pattern
  - Data points should be randomly scattered and there should be a balance
- Residuals should be relatively small in size
- Mean of the residuals is 0

Rectangular shape!!!

Aug 18-10:44 AM



Oct 5-5:51 PM

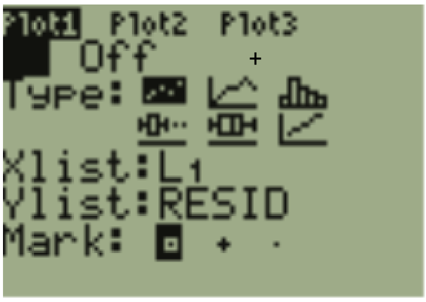
AP Exam Common Error:

When asked if a linear model is appropriate, do not use the correlation to justify linearity. Associations can be clearly nonlinear and still have correlation close to 1. Only a residual plot can adequately address whether a line is an appropriate model for the data.

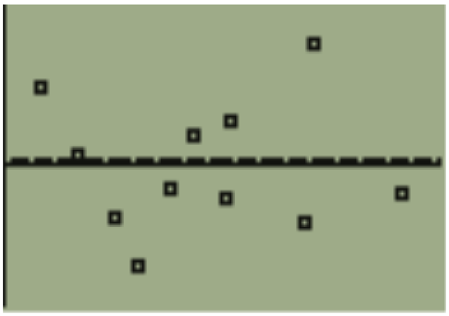
Example

Refer to the Honda CRV data to calculate and interpret a residual plot.

(Instructions p. 178 in text)



```
Plot1 Plot2 Plot3
 Off
Type:
Xlist:L1
Ylist:RESID
Mark: □ + ·
```



The residual plot shows no obvious pattern. The points are randomly scattered, balanced, and appear to take on a rectangular shape. Therefore, the linear model is appropriate.

Standard deviation of the residuals(s):

- Gives the approximate size of a “typical” or “average” prediction error (residual)

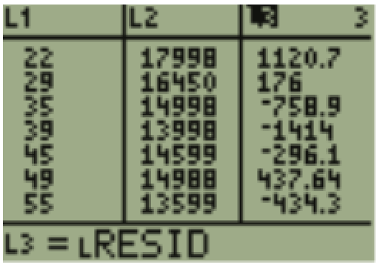
$$s = \sqrt{\frac{\Sigma residuals^2}{n-2}} = \sqrt{\frac{\Sigma(y_i - \hat{y})^2}{n-2}}$$

- Same as standard deviation formula except for denominator

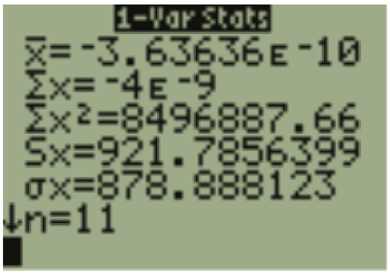
- The reason the denominator is n - 2 is because of the fact that you can only estimate the amount of variability from the LSRL when you have 3 or more points

Example

Calculate and interpret the standard deviation of the residuals for the Honda CR-V data.



```
L1      L2      L3      3
22      17998   1120.7
29      16450   176
35      14998   -758.9
39      13998   -1414
45      14599   -296.1
49      14988   437.64
55      13599   -434.3
L3 = LRESID
```



```
1-Var Stats
x̄=-3.63636E-10
Σx=-4E-9
Σx²=8496887.66
Sx=921.7856399
σx=878.888123
↓n=11
```

s = 922

The predicted cost differs from the actual cost by an average of $922.

Coefficient of Determination: r^2 in regression

- r^2 and s both measure how well the least-squares regression line models the data (how much scatter there is from the least-squares regression line)
- s is measured in the units of the response variable, r^2 is on a standard scale

It is the correlation coefficient squared!!

"_____% of the variation in (y variable) is accounted for by (x variable)."

OR

"_____% of the variation in (y variable) is accounted for by the regression model relating it to (x variable)."

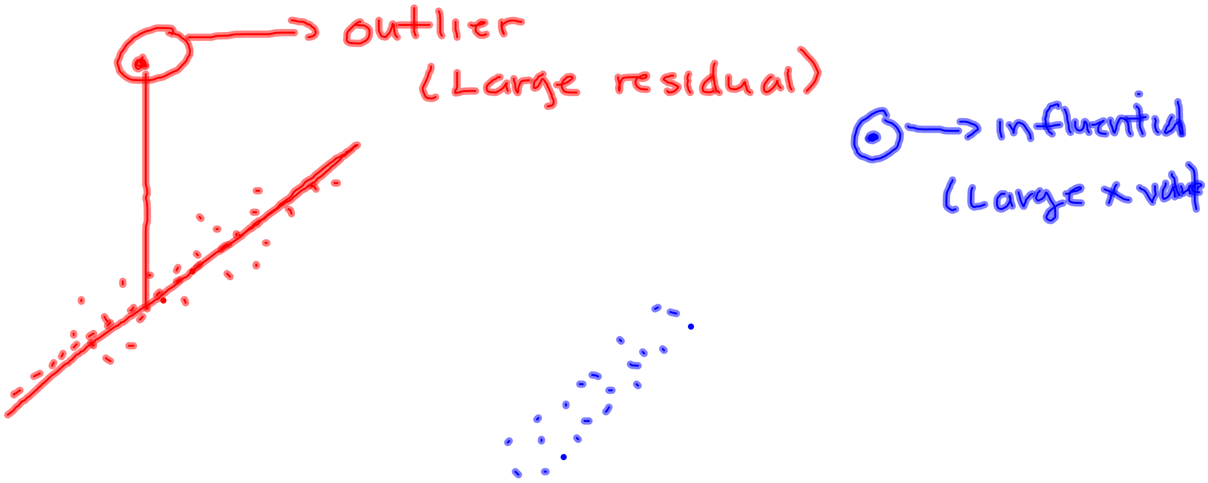
Oct 6-9:13 PM

Example

For the Honda data, Calculate r^2 and discuss what it means in context to the problem.

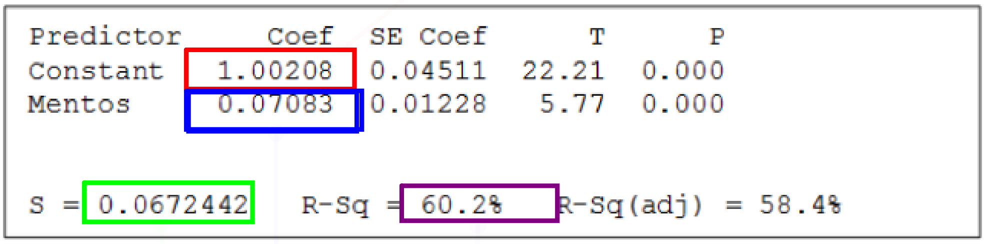
76.4% of the variation in cost is accounted for by mileage (in thousands).

Oct 6-9:15 PM



Oct 8-10:30 AM

Mentos Example



```
Predictor      Coef   SE Coef      T       P
Constant    1.00208  0.04511   22.21   0.000
Mentos      0.07083  0.01228    5.77   0.000


S = 0.0672442   R-Sq = 60.2%   R-Sq(adj) = 58.4%
```

y-intercept

slope

r^2

standard deviations
of residuals

Oct 7-5:48 PM

(a) What is the equation of the least-squares regression line? Define any variables you use.

$$\widehat{\text{Amount Expelled}} = 1.00208 + 0.07083(\text{\# of Mentos})$$

Oct 7-5:50 PM

(b) Interpret the slope of the least-squares regression line.

The predicted amount of coke expelled increases by 0.07083 cups for each additional mento added.

Oct 7-5:49 PM

(c) What is the correlation?

$r = \sqrt{0.602} = 0.776$. The data shows a strong positive linear association.

(d) Is a linear model appropriate for this data? Explain.

Yes. The residuals are relatively small and show no apparent pattern.

(e) Would you be willing to use the linear model to predict the amount of Diet Coke expelled when 10 mentos are used? Explain.

No. This would require extrapolation well outside the range of the data. We can not be certain the linear regression model would apply for 10 mentos.

(f) Calculate and interpret the residual for bottle of diet coke that had 2 mentos and lost 1.25 cups.

$\widehat{\text{Amount Expelled}} = 1.00208 + 0.07083(2) = 1.14$ cups

Residual = Actual - Predicted = 1.25 - 1.14 = 0.11 cups

The predicted amount of Diet Coke expelled is 0.11 cups less than the actual amount of Diet Coke expelled.

(g) Interpret the values of r^2 and *s*.

$r^2 = 0.602$

60.2% of the variation in the amount of Diet Coke expelled can be explained by the linear model relating it to the number of mentos.

$s = 0.067$

The predicted amount of Diet Coke expelled differs from the actual amount expelled by an average of 0.067 cups.

(h) If the amount expelled was measured in ounces instead of cups, how would the values of r^2 and s be affected? Explain.

r^2 would not change because it is unitless. *s* would change because of the linear transformation.

Oct 8-11:37 AM