

Chapter 5

Linear Regression

Linear regression tries to account for the variation in one variable by a linear relationship with another variable.

■ Knee surgery

The age and days required to recover were recorded for 15 patients who underwent knee surgery.

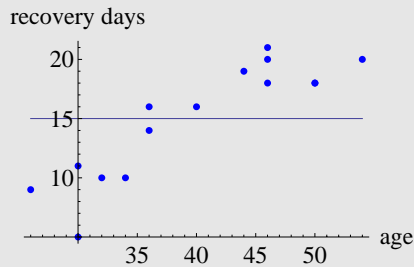
Age	26	30	30	32	34	36	36	40	44	46	46	46	50	50	54
Recovery	9	5	11	10	14	16	16	19	18	18	20	21	18	18	20

Your job is predict the recovery time for new patients.

○ First idea: Use the mean

You compute the mean recovery is 15 days and use this as your best guess.

Look at a scatter plot of the data with the a line at height 15, the mean recovery time.



Younger patients are below the line, older patients are above the line.

○ Better idea: The least squares line. Predicted values \hat{y} .

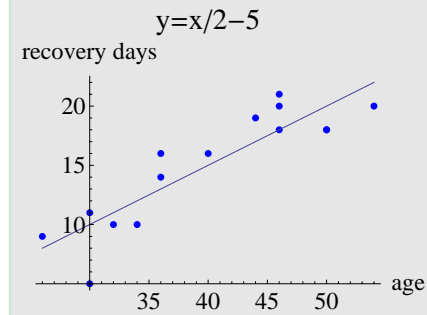
The plot also include the *least squares line*.

→ The least squares line is the line that best fits the data.

→ Each data set has a different least squares line.

→ You'll learn how to find the equation for the least squares line later. For now, they'll be given to you.

2 | ClassNotes145.nb



x	26	30	30	32	34	36	36	40	44	46	46	46	50	50	54
y	9	5	11	10	14	16	16	19	18	20	21	18	18	20	

Since the points roughly "line up" you say that age and recovery days

→ exhibit a linear relationship or

→ are linearly related

You can use the least squares line $y = x/2 - 5$ to predict a y value (recovery time) for any x value (age).

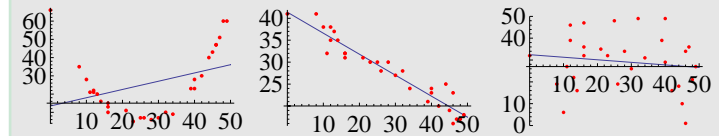
For the age $x = 30$ the least squares line gives a predicted value of $\hat{y} = 30/2 - 5 = 10$ recovery days.

→ predicted y values are denoted by \hat{y} (pronounced *y-hat*).

This is only a prediction! The two 30 year old patients you've already seen had recoveries of 5 and 11 days.

■ Assessing linear fit: Visual

Look at the scatterplots from three other bivariate numerical data sets.



The left plot shows a relationship between x and y , but not a linear relationship.

The middle plot shows a linear relationship between x and y since the points roughly "line up".

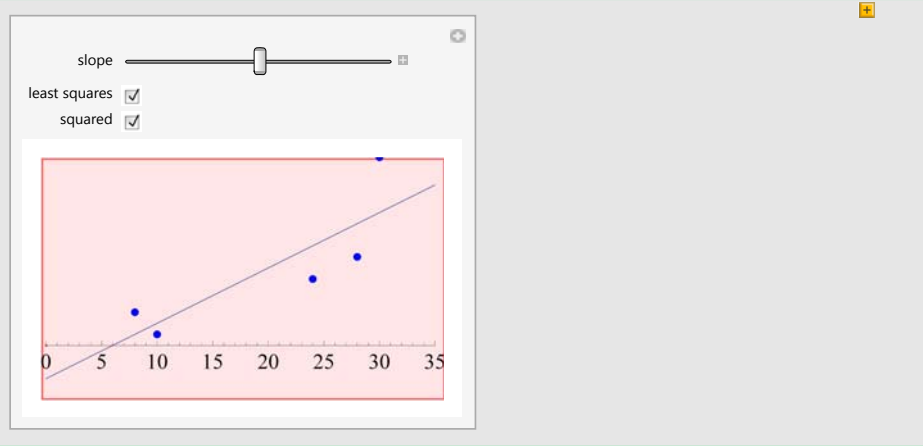
→ Since the line has negative slope (falls from left to right), x and y have a *negative linear relationship*.

The right plot doesn't show much of any relationship between x and y .

■ Assessing linear fit: Coefficient of determination

Here are four data points and the least squares line $y = -3 + \frac{1}{2}x$.

x	8	10	24	28	30
y	3	1	6	8	17



If you completely ignore the x coordinates and look at just the y values, then you measure their variation from their mean \bar{y} .

SSTot: measures the *total* variation in the y coordinates.

Note: $SSTot = S_{yy} = SS_y$ are three different notations in our text which all mean the same thing.

The vertical lines from the points to the line $y = \bar{y}$ show the deviations.

The area of the pink boxes show the squared deviations.

→ In table form, compute SSTot. Check your value for \bar{y} with your neighbor before computing the deviations.

SSResid:

The least squares line for the data is $y = -3 + \frac{1}{2}x$.

Use this equation to fill in the predicted values \hat{y}_i column in the table.

For a given point (x_i, y_i) the *residual* is $y_i - \hat{y}_i$.

→ Called a "residual" since it measures the part of y_i not explained (left over) by the line.

In table form, compute SSResid the sum of the squared residuals for the least squares line.

→ Remember that SS in the SSResid stands for "sum of the squared"

→ Add to the plot on the right a graphical depiction of the residuals.

→ Add to the plot on the right a graphical depiction of the squared residuals.

The coefficient of determination r^2 :

A measure of the variation in y is given by $SSTot = \underline{\hspace{2cm}}$

A measure of the variation in y away from the least squares line is $SSResid = \underline{\hspace{2cm}}$

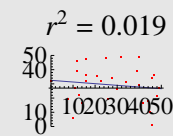
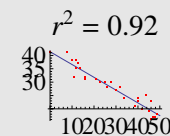
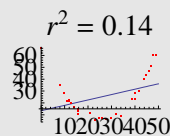
The amount of the variation in y accounted for by the least squares line is $SSTot - SSResid = \underline{\hspace{2cm}}$.

The fraction of the variation in y accounted for by the line is $\frac{SSTot - SSResid}{SSTot} = 1 - \frac{SSResid}{SSTot} = \underline{\hspace{2cm}}$.

residGrid[xy]

y	$y - \bar{y}$	$(y - \bar{y})^2$	x	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
3	-4	16	8	1	2	4
1	-6	36	10	2	-1	1
6	-1	1	24	9	-3	9
8	1	1	28	11	-3	9
17	10	100	30	12	5	25
Sum=35	Sum=0	Sum=154		Sum=0	Sum=48	

▫ Here are the data sets from above again.



Left set: Only 14% of the variation is explained by the least squares line, indicating a weak linear relationship.

Middle set, 92% of the variation in the y coordinates is explained by the least squares line: strong linear relationship.

Right set, less than 2% of the variation is explained by the least squares line.

The *coefficient of determination* for the least squares line is $r^2 = 1 - \frac{SSResid}{SSTot}$.

→ A value of r^2 near 1 means that almost all of the variation in y can be accounted for by the line.

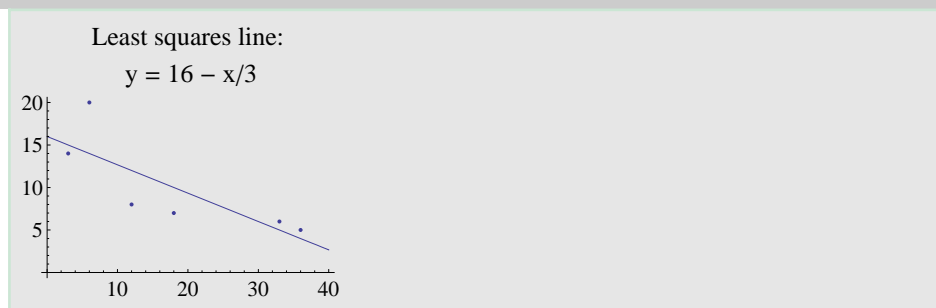
→ A value of r^2 near 0 means that almost none of the variation in y can be accounted for by the line.

■ **Assessing linear fit: correlation coefficient r .**

The correlation coefficient r :

- Always between -1 and 1 .
- A value near 1 indicates a strong positive linear relationship--line with positive slope (climbing).
- A value near -1 indicates a strong negative linear relationship--line with negative slope (falling).
- Assesses strength of linear relationship--see bottom of page 190.
- The slope of the least squares line and r have the same sign.
- r^2 equals the coefficient of determination

■ **Assessing linear fit: standard error**



Remember that the standard deviation $s = \sqrt{\frac{S_{yy}}{n-1}}$ measures a typical distance from y_i to the mean \bar{y} .

The *standard error about the least squares line*, denoted s_e measures a typical (vertical) distance from y_i to the least squares line.

▫ **By eye estimate of s_e**

From the plot above give an estimate of s_e .

To do this just estimate by eye the typical vertical distance from a data point to the least squares line.

▫ **Formula for s_e**

To compute s_e exactly use $s_e = \sqrt{\frac{SS_{Resid}}{n-2}}$.

Remember SS_{Resid} gives the sum of all the squared residuals.

You divide by $n-2$ to get an "average" squared residual.

Then take a square root to get an "average" residual.

▫ **Why $n-2$?**

In choosing a least squares line you see that a line has the freedom to go through any two points. So on average 2 of the n points will have zero residual. This leaves $n-2$ "degrees of freedom" for the calculation.

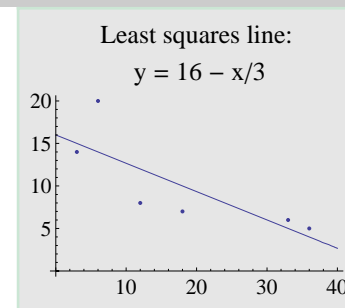
You don't need to memorize this--it's on your crib sheet.

Review: In computing s why do you divide by $n-1$ instead of n ?

Complete sentence(s) please.

▫ **Computing s_e by hand.**

x	3	6	12	18	33	36
y	14	20	8	7	6	5



Use a table to organize and present a hand computation of s_e , the standard error about the least squares line.

y	$y - \bar{y}$	$(y - \bar{y})^2$	x	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
14	4	16	3	15	-1	1
20	10	100	6	14	6	36
8	-2	4	12	12	-4	16
7	-3	9	18	10	-3	9
6	-4	16	33	5	1	1
5	-5	25	36	4	1	1
Sum=60	Sum=0	Sum=170		Sum=0	Sum=64	

▫ **Indication of a strong linear relationship**

Value of r^2 near 1.

Value of r near -1 or 1 .

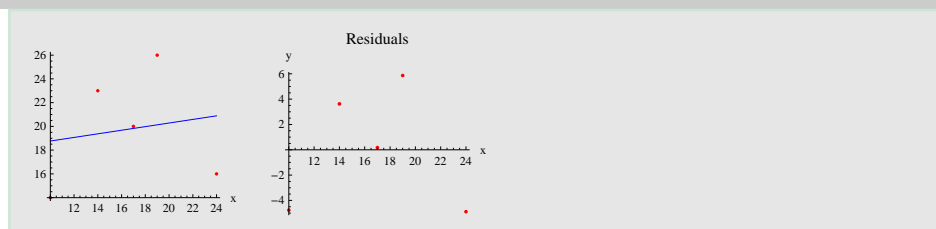
Value of s_e near 0.

■ **Assessing linear fit: residual plots**

A residual plot shows the pairs $(x_i, y_i - \hat{y}_i)$.

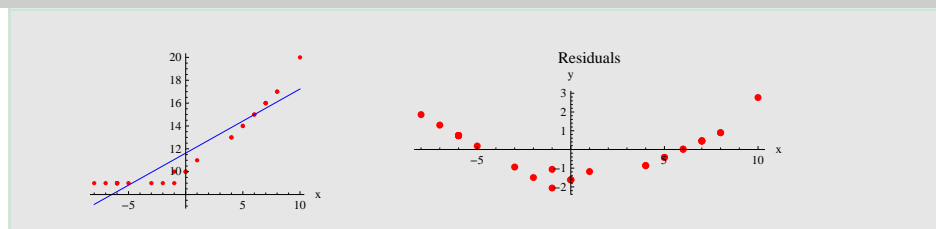
If the least squares line does a good job describing the relationship between x and y , then the residual plot should look random.

Here's a least squares line and residual plot for five points.



The residual plot looks pretty random, but there are so few points it might be hard to find any pattern.

Here's another example with 30 points.



The least squares line on the left seems to do a good job predicting y from the x .

But the residual plot on the right shows a strong curved shape.

→ Part of the relationship between x and y might not be linear.

○ A strong linear relationship results in a random scattering of points in the residual plot.

■ **Interpolation, extrapolation, interpretation**

$$y = x/2 - 5$$

Interpolation: Predicting a y value for an x value within the range of x values in the data.

Use the least squares line for the data to predict the recovery time for a 40 year old patient.

Since $x = 40$ is inside the range of x values from the data (26 to 54), you are interpolating for the value $x = 40$.

Extrapolation: Predicting a y value for an x value outside the range of x values in the data.

Use the least squares line for the data, to predict the recovery time for a 1 year old patient.

Since $x = 1$ is outside the range of x values from the data (26 to 54), you are extrapolating for the value $x = 1$.

Interpretation of slope: The least squares line $y = x/2 - 5$ has slope $1/2$ and y -intercept -5 . In the context of this problem, a slope of $1/2$ means that we expect that being one year older will on average result in $1/2$ of an additional recovery day.

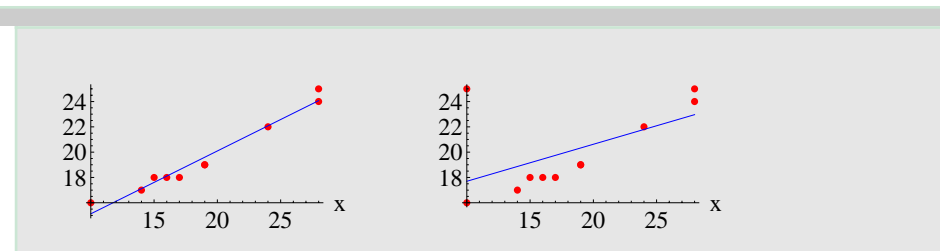
■ **The effect of outliers.**

The plot on the left shows a least squares line doing a great job of approximating the data.

The plot on the right has one additional point at $(10, 25)$.

This point is an outlier and you can see that it's pulled the least squares line way off the rest of the data.

→ The least squares line is susceptible to outliers.



■ **Word to the wise: p241.**

1. Correlation does not imply causation.
2. Any data set can be fitted with a least squares line. Indications of a good fit are:
3. r near zero DOES NOT imply no relationship
4. Be wary of extrapolation.
5. Be wary of the influence of outliers.

■ **HW**