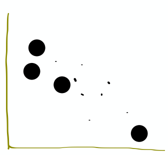
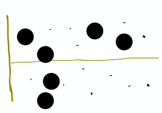
The Role of s and r^2 in Regression

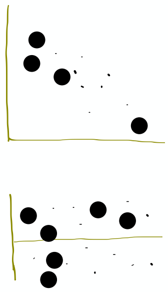
(pages 188-192)

We use Residual Plots
to determine if a LSRL
is appropriate.





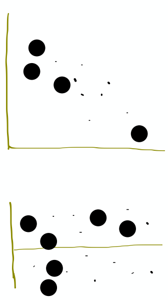
We use Residual Plots to determine if a LSRL is appropriate.



remember!

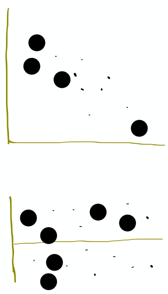
we can't use r to justify linearity
--- Need residual plots ---

We use Residual Plots to determine if a LSRL is appropriate.



If so, we can use S and r^2 to determine how good the predictions will be.
(How well does the line work)

We use Residual Plots to determine if a LSRL is appropriate.



If so, we can use

s and r^2

to determine how good the predictions will be.

(How well does the line work)

s_x

s_y

s Standard Deviations of the residuals

r^2 Coefficient of Determination

TARGET

Interpret the standard deviation of the residuals and r^2 and use these values to assess how well the least-squares regression line models the relationship between two variables.

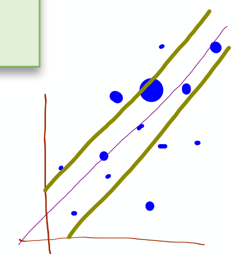
pick up →

AP Stats – Class Notes – Section 3.2 – Day 3
The Role of s and r^2 in Regression

The **standard deviation of the residuals** ***s***, measures the size of a typical residual. That is, ***s*** measures the typical distance between the actual *y* values and the predicted *y* values.

y

$\hat{y}$



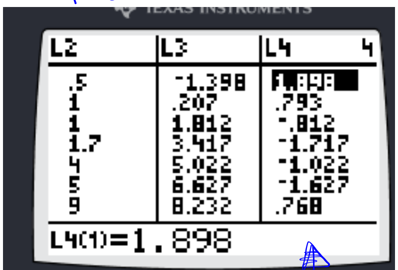
S is sometimes called the typical prediction error

S

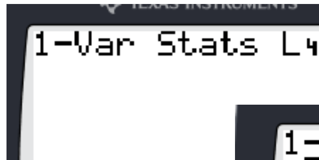
To assess how well the line fits all of the data, we need to consider the residuals for each observation, not just one. Using these residuals, we can estimate the "typical" prediction error when using the least-squares regression line.

Note: The sum of residuals will up to 0.

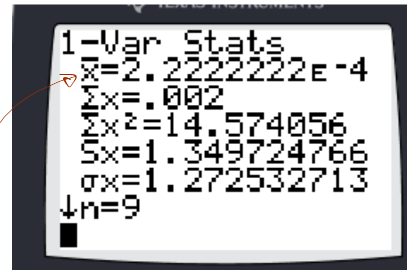
from the Candy Grab



Residuals



Sum of residuals



The **standard deviation of the residuals *s*** measures the size of a typical residual. That is, ***s*** measures the typical distance between the actual *y* values and the predicted *y* values.

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

Most likely you will be given this value.

We divide by n-2 rather than n-1. We used n-1 for s when we estimated the mean (used $\bar{x}$ for μ). Now we are estimating both slope and the y-intercept, so we use n-2. We subtract one more for each parameter we estimate.

Coefficient of Determination

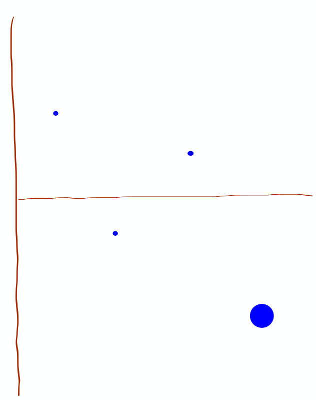
r^2 measures the fraction of the variability in the y variable that is accounted for by the LSRL using x.

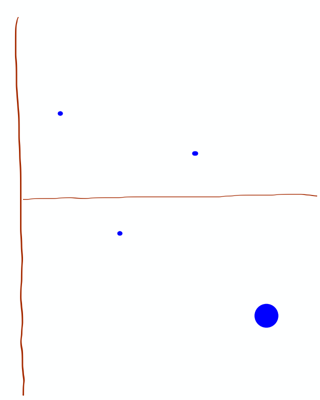
$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}$$

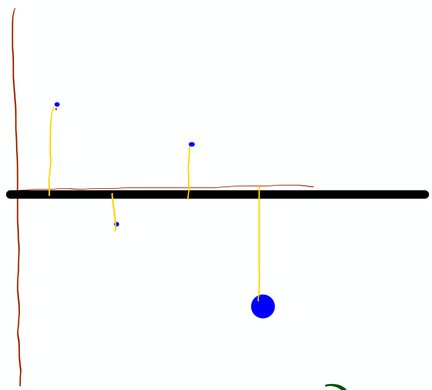
The **coefficient of determination** $\boldsymbol{r^2}$ measures the percent reduction in the sum of squared residuals when using the least-squares regression line to make predictions, rather than the mean value of *y*.

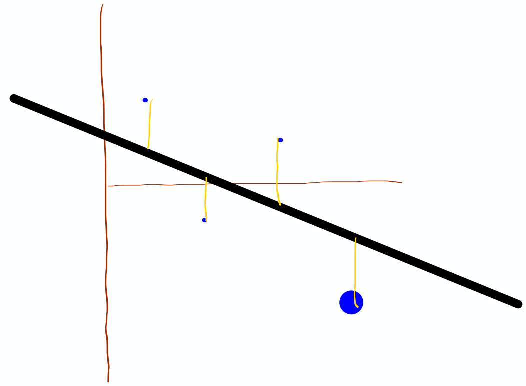
In other words, r^2 measures the percent of the variability in the response variable that is accounted for by the least-squares regression line.

r^2 tells us how much better the LSRL does at predicting values of *y* than simply guessing the mean *y* for each value in the dataset.



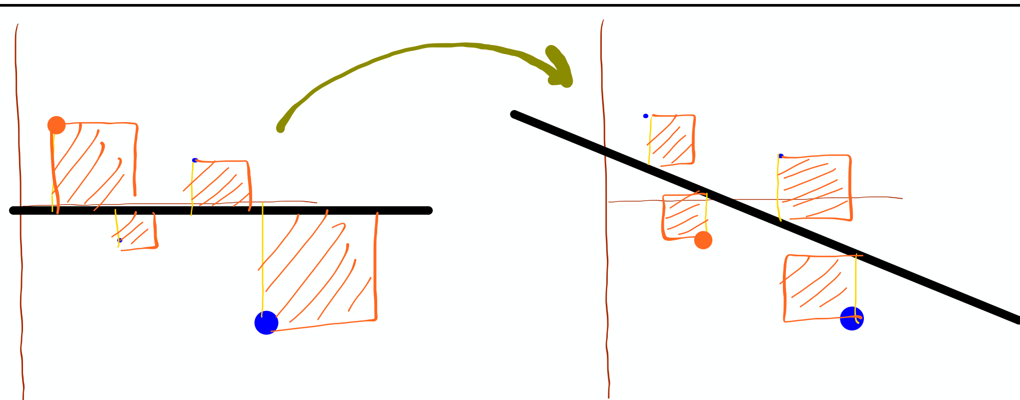






Sum of $(\text{resid})^2$
with mean of
y-values

Sum of $(\text{resid})^2$
with LSRL



Sum of $(\text{resid})^2$
with mean of
y-values

Sum of $(\text{resid})^2$
with LSRL

s and r^2

Big Ideas:

Standard Deviation of Residuals (s)

Interpretation:

Coefficient of Determination r^2

Interpretation:

s and r^2

Big Ideas:

Standard Deviation of Residuals (s)

Interpretation:

"The actual y-context is typically about S away from the #predicted by the LSRL".

Coefficient of Determination r^2

Interpretation:

"About r^2 % of the variability in y-context is accounted for by the LSRL when x = x-context

Ninth-grade students at the Webb Schools go on a backpacking trip each fall. Students are divided into hiking groups of size 8 by selecting names from a hat. Before leaving, students and their backpacks are weighed. The data here are from one hiking group.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

Analyze the data using stapplet.com.

1. Find the LSRL of the data. Write it below.

$\hat{y} = 16.265 + 0.091x$

$\widehat{\text{Backpack weight}} = 16.265 + 0.091(\text{Body Weight})$

2. Find and interpret s. $n=8$ $S = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} = \sqrt{\frac{30.904}{8-2}} = 2.27$

The actual backpack weight is typically about 2.27 lbs away from the weight predicted by the LSRL with x = Body weight

3. Find and interpret the value of r^2.

$r^2 = .63$

About 63% of the variability

2. Find and interpret s.

$S = 2.27$

The actual backpack weight is typically about 2.27 lb. away from the weight predicted by the LSRL with x = the body weight.

3. Find and interpret the value of r^2.

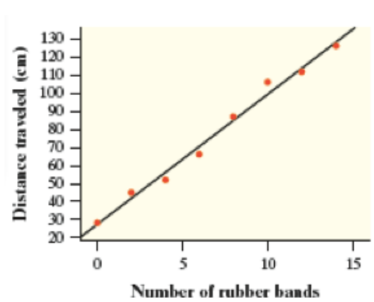
$r^2 = 0.632$

About 63.2% of the variability in backpack weight is accounted for by the LSRL with x = body weight.

The last Barbie bungee jump *Interpreting s and r²*

Mrs. Gallas's class performed the "Barbie Bungee" activity. They connected rubber bands one at a time in a chain to Barbie's feet and then measured the distance that Barbie travels on her (last) bungee jump. The distance is measured from the edge of the jumping platform to the lowest point that Barbie's head reaches.

Here is the scatterplot of data from one of the groups with the regression line $\hat{y} = 27.42 + 7.21x$. For this model, technology gives $s = 4.11$ and $r^2 = 0.989$.



(a) Interpret the value of s.

(a) Interpret the value of *s*.

The distance travelled by Barbie is typically about 4.11 cm away from the distance predicted by the LSRL with x = # rubber bands.

(b) Interpret the value of r^2.

About 98% of the variability in dist. travelled by Barbie is accounted for by the LSRL with x = # rubber bands.

Interpreting Computer Regression Output
(pages 192-194)

You are not expected to able to use the software but you are expected to interpret the output.

From the output, be sure you can find the:

slope b_1
y-intercept b_0
S
r^2

$\hat{y} = b_0 + b_1 x$

Minitab

Slope y intercept

Predictor	Coef	SE Coef	T	P
Constant	$\boxed{38257}$	2446	15.64	0.000
Miles Driven	$\boxed{-0.16292}$	0.03096	-5.26	0.000

r^2

$\boxed{S = 5740.13}$ $\boxed{R\text{-}Sq = 66.4\%}$ R-Sq(adj) = 64.0%

Standard deviation of the residuals

JMP

Summary of Fit

RSquare	$\boxed{0.664248}$	r^2
RSquare Adj	0.640266	
Root Mean Square Error	$\boxed{5740.131}$	Standard deviation of the residuals
Mean of Response	27833.69	
Observations (or Sum Wgts)	16	

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	$\boxed{38257.135}$	2445.813	15.64	<.0001
Miles Driven	$\boxed{-0.162919}$	0.030956	-5.26	0.0001

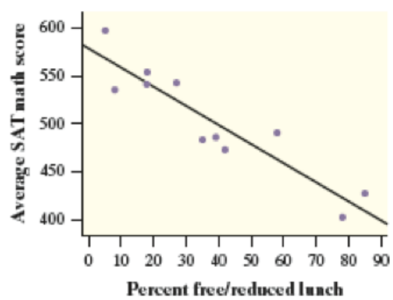
y intercept Slope

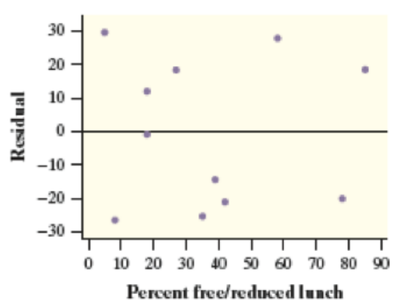
Can we predict a school's average SAT math score? *Interpreting regression output*

A random sample of 11 high schools was selected from all the high schools in Michigan. The percent of students who are eligible for free/reduced lunch and the average SAT math score of each high school in the sample were recorded.

Students with household income below a certain threshold are eligible for free/reduced lunch.

Here are a scatterplot with the least-squares regression line added, a residual plot, and some computer output:





Predictor	Coef	SE Coef	T	P
Constant	577.9	12.5	46.16	0.000
Foot length	−1.993	0.276	−7.22	0.000
S = 23.3168	R-Sq = 85.29%	R-Sq(adj) = 83.66%		

(a) Is a line an appropriate model to use for these data? Explain how you know the answer.

Because the scatter plot shows a linear association and the residual plot shows no leftover pattern, the line is appropriate

shows a random scatter

(b) Find the correlation.

$r = \pm\sqrt{0.8529} = \pm 0.924$

because the relationship is negative

$r = -0.924$

(c) What is the equation of the least-squares regression line that describes the relationship between percent free/reduced lunch and average SAT math score? Define any variables that you use.

$\hat{y} = 577.9 - 1.993x$ where

$\hat{y}$ is the predicted average SAT math score.

and x is percent free/reduced lunch.

(d) By about how much do the actual average SAT math scores typically vary from the values predicted by the least-squares regression line with *x* = percent free/reduced lunch?

$S = 23.3168$ so the actual average SAT math scores typically vary by about 23.3168 from the values predicted by the regression line using x = percent/free lunch.

See your
LCQ

Assignment:

3.2 55, 57, 59, 67

pp. 188-192