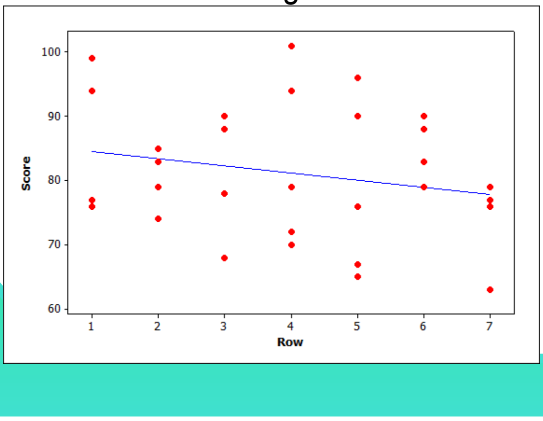
12.1 Inference for Linear Regression

vocab

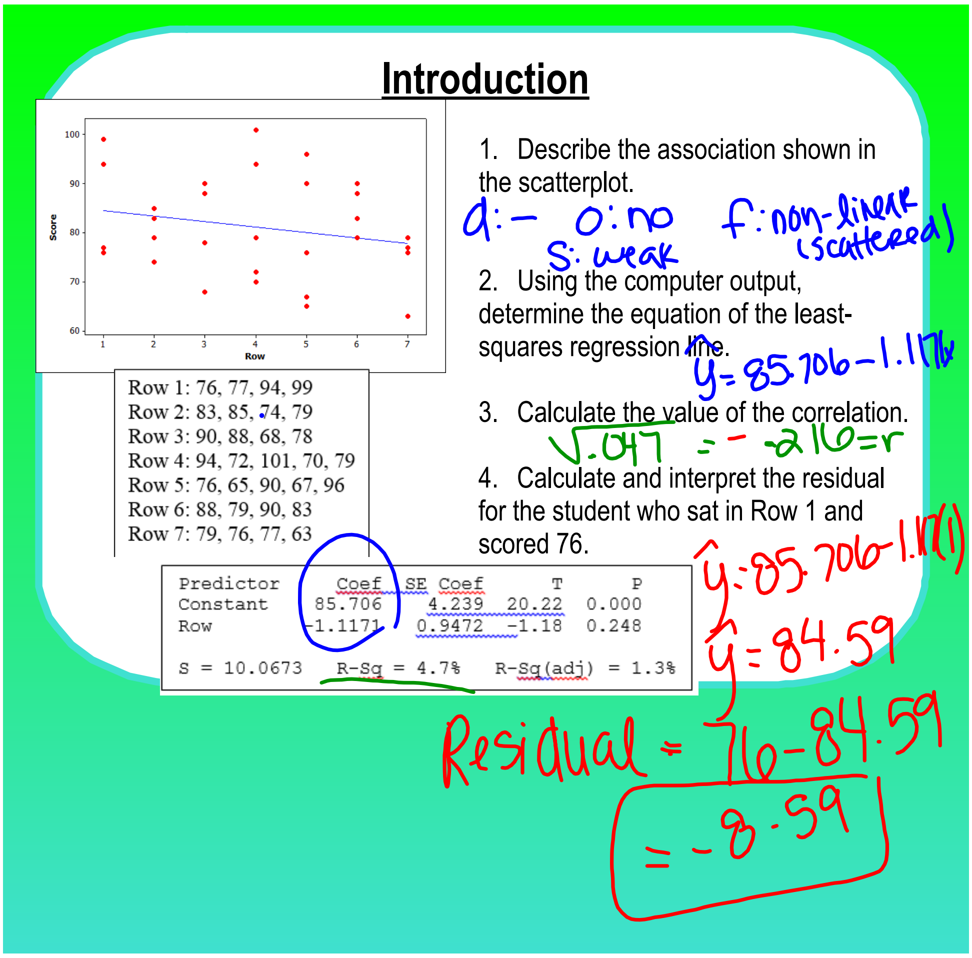
examples

Introduction

Many people believe that students learn better if they sit closer to the front of the classroom. Does sitting closer *cause* higher achievement, or do better students simply choose to sit in the front? To investigate, an AP Statistics teacher randomly assigned students to seat locations in his classroom for a particular chapter and recorded the test score for each student at the end of the chapter. The explanatory variable in this experiment is which row the student was assigned (Row 1 is closest to the front and Row 7 is the farthest away). Do these data provide *convincing* evidence that sitting closer causes students to get higher grades?



Row 1: 76, 77, 94, 99
Row 2: 83, 85, 74, 79
Row 3: 90, 88, 68, 78
Row 4: 94, 72, 101, 70, 79
Row 5: 76, 65, 90, 67, 96
Row 6: 88, 79, 90, 83
Row 7: 79, 76, 77, 63



Introduction

Row 1: 76, 77, 94, 99
Row 2: 83, 85, 74, 79
Row 3: 90, 88, 68, 78
Row 4: 94, 72, 101, 70, 79
Row 5: 76, 65, 90, 67, 96
Row 6: 88, 79, 90, 83
Row 7: 79, 76, 77, 63

```
Predictor      Coef    SE Coef      T      P
Constant     85.706      4.239  20.22  0.000
Row         -1.1171     0.9472  -1.18  0.248

S = 10.0673    R-Sq = 4.7%    R-Sq(adj) = 1.3%
```

1. Describe the association shown in the scatterplot.

d: – o: no f: non-linear (scattered)
s: weak

2. Using the computer output, determine the equation of the least-squares regression line.

$\hat{y} = 85.706 - 1.117x$

3. Calculate the value of the correlation.

$\sqrt{.047} = - -.216 = r$

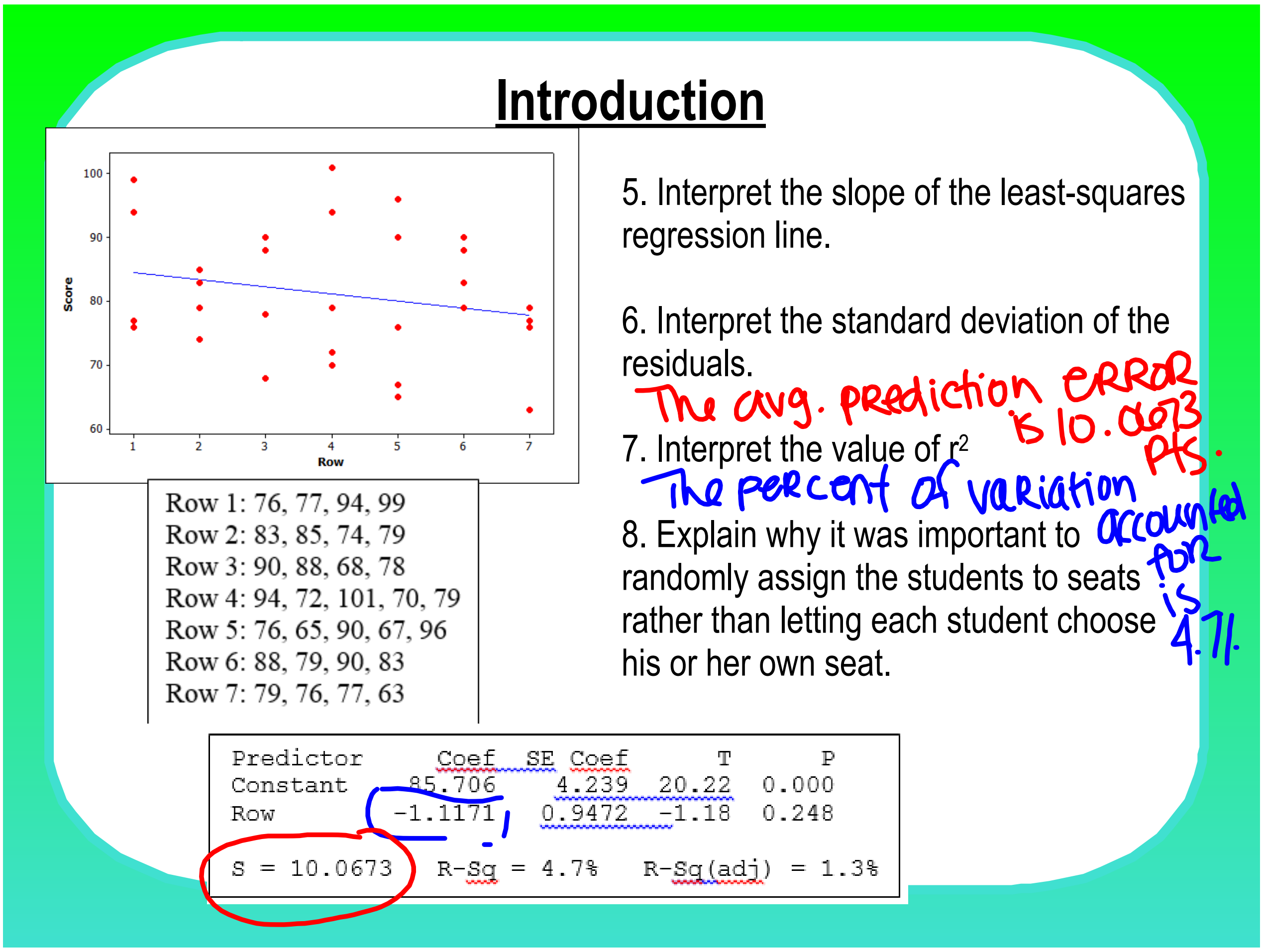
4. Calculate and interpret the residual for the student who sat in Row 1 and scored 76.

$\hat{y} = 85.706 - 1.117(1)$

$\hat{y} = 84.59$

Residual = 76 − 84.59

$\boxed{= -8.59}$



Introduction

Row 1: 76, 77, 94, 99
Row 2: 83, 85, 74, 79
Row 3: 90, 88, 68, 78
Row 4: 94, 72, 101, 70, 79
Row 5: 76, 65, 90, 67, 96
Row 6: 88, 79, 90, 83
Row 7: 79, 76, 77, 63

```
Predictor      Coef    SE Coef      T      P
Constant     85.706      4.239  20.22  0.000
Row         -1.1171     0.9472  -1.18  0.248

S = 10.0673    R-Sq = 4.7%    R-Sq(adj) = 1.3%
```

5. Interpret the slope of the least-squares regression line.

6. Interpret the standard deviation of the residuals.

The avg. prediction error is 10.0673 pts.

7. Interpret the value of r^2

The percent of variation accounted for is 4.7%.

8. Explain why it was important to randomly assign the students to seats rather than letting each student choose his or her own seat.

Sample Regression Line

Calculated off of a sample set of data (ex/ the height and weight of 20 random SDOHS students)

$\hat{y} = a + bx$

Population Regression Line

Calculated off of a population set of data (ex/ the height and weight of all SDOHS students)

$\hat{y} = \alpha + \beta x$

Sampling Distribution of b

all based on the possible values of b that could occur in random sampling if H_0 is true

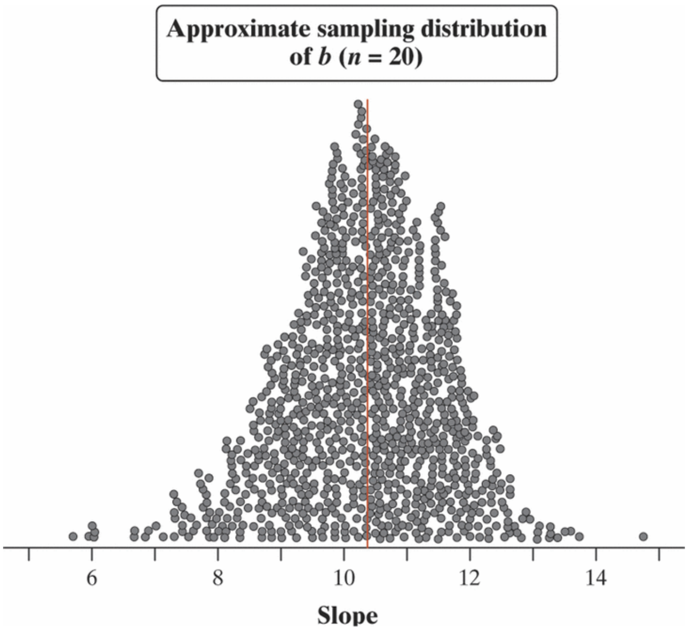
Shape: Symmetric, Unimodal, close to Normal

Center: The mean of all the b values in repeated samples (β)

Spread: The standard deviation of all the b values in repeated samples

$$SE_b = \frac{s}{s_x\sqrt{n-1}}$$

Sampling Distribution of b



Conditions for Regression Inference

Suppose we have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

LINEAR: The actual relationship between x and y is linear. For any fixed value of x, the mean response μ_y falls on the population regression line $\mu_y = \alpha+\beta x$

INDEPENDENT: Individual observations are independent of each other

Conditions for Regression Inference

NORMAL: For any fixed value of x, the response of y varies according to a Normal distribution.

EQUAL VARIANCE: The standard deviation of y (call it σ) is the same for all values of x.

RANDOM: The data come from a well-designed random sample or randomized experiment

use LINER to remember

12.1 Inference for Linear Regression (Day 2)

vocab

examples

Checking the Conditions

LINEAR: Examine the scatterplot to check that the overall pattern is roughly linear. Look for curved patterns in the residual plot. Check to see that the residuals center on the "residual = 0" line at each x-value in the residual plot.

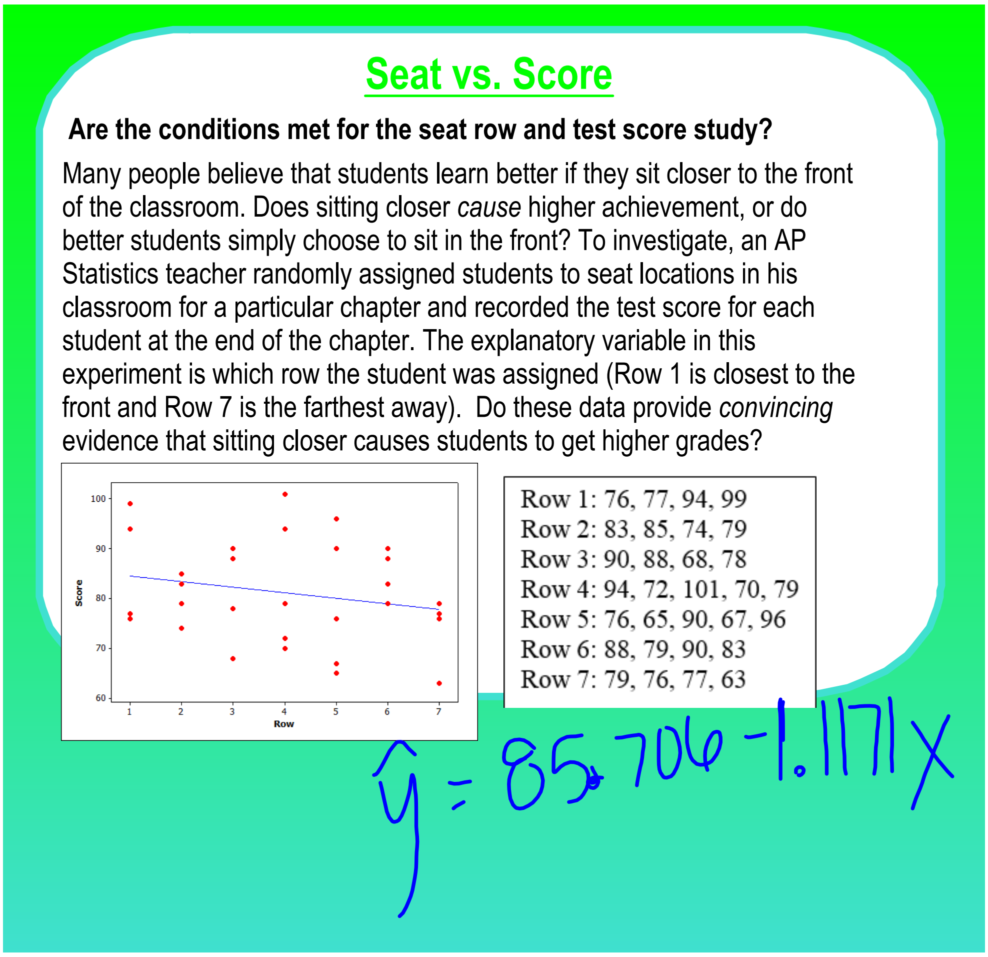
INDEPENDENT: 10% condition , OR each obs. ind. of one another

NORMAL: Make a stemplot, histogram, or Normal probability plot of the residuals and check for clear skewness or other major departures from Normality.

Checking the Conditions

EQUAL VARIANCE Look at the scatter of the residuals above and below the "residual = 0" line in the residual plot. The amount of scatter should be roughly the same from the smallest to the largest x-value.

RANDOM: Check that the data were collected randomly or that the experiment contains random assignment



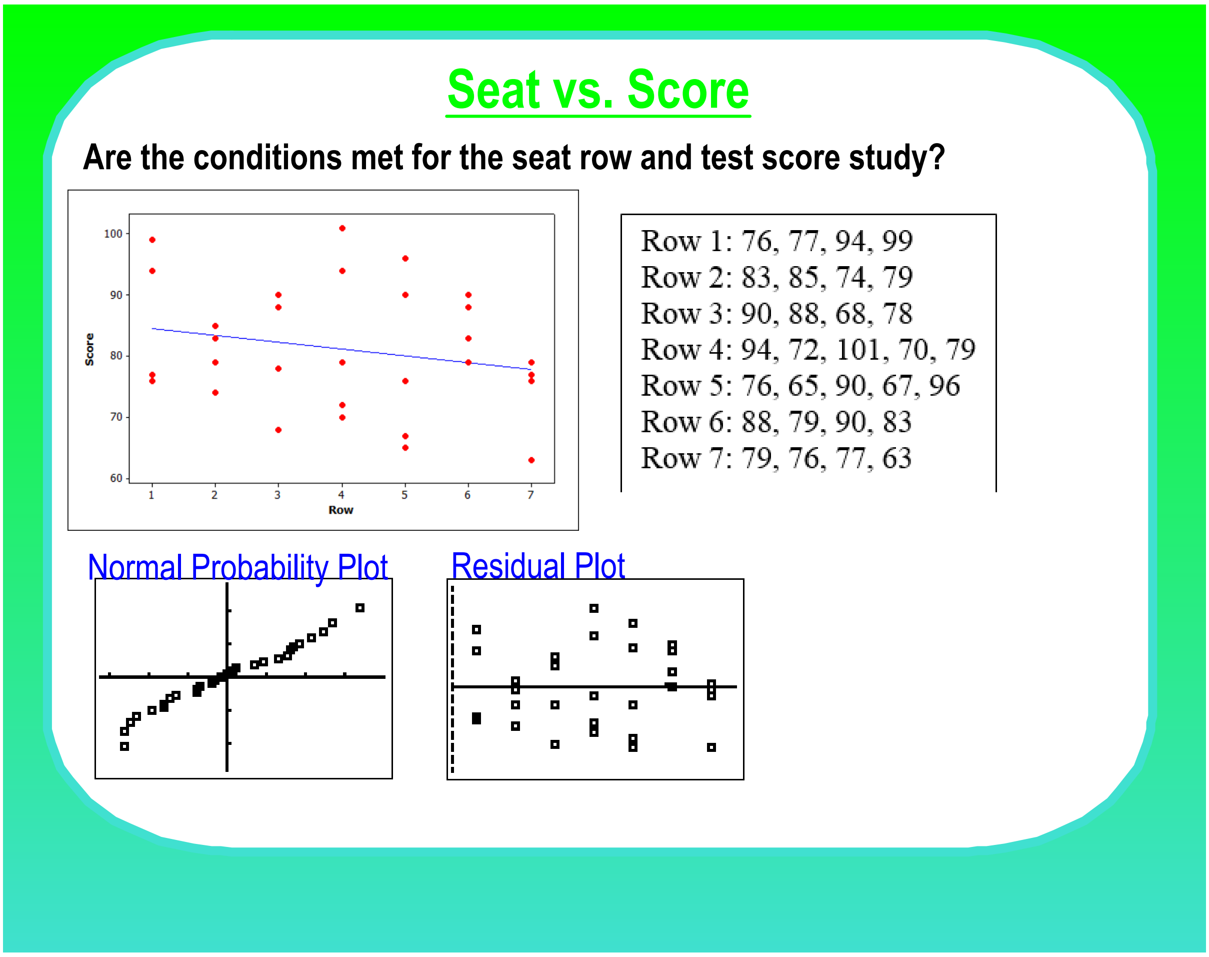
Seat vs. Score

Are the conditions met for the seat row and test score study?

Many people believe that students learn better if they sit closer to the front of the classroom. Does sitting closer *cause* higher achievement, or do better students simply choose to sit in the front? To investigate, an AP Statistics teacher randomly assigned students to seat locations in his classroom for a particular chapter and recorded the test score for each student at the end of the chapter. The explanatory variable in this experiment is which row the student was assigned (Row 1 is closest to the front and Row 7 is the farthest away). Do these data provide *convincing* evidence that sitting closer causes students to get higher grades?

Row 1: 76, 77, 94, 99
Row 2: 83, 85, 74, 79
Row 3: 90, 88, 68, 78
Row 4: 94, 72, 101, 70, 79
Row 5: 76, 65, 90, 67, 96
Row 6: 88, 79, 90, 83
Row 7: 79, 76, 77, 63

$\hat{y} = 85.706 - 1.1171x$



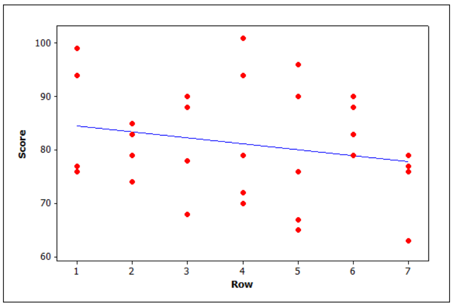
Seat vs. Score

Are the conditions met for the seat row and test score study?

Row 1: 76, 77, 94, 99
Row 2: 83, 85, 74, 79
Row 3: 90, 88, 68, 78
Row 4: 94, 72, 101, 70, 79
Row 5: 76, 65, 90, 67, 96
Row 6: 88, 79, 90, 83
Row 7: 79, 76, 77, 63

Seat vs. Score

Are the conditions met for the seat row and test score study?



Row 1: 76, 77, 94, 99
Row 2: 83, 85, 74, 79
Row 3: 90, 88, 68, 78
Row 4: 94, 72, 101, 70, 79
Row 5: 76, 65, 90, 67, 96
Row 6: 88, 79, 90, 83
Row 7: 79, 76, 77, 63

LINEAR- The scatterplot shows a somewhat linear formation and the residual plot shows no clear pattern, so a linear model is appropriate.

INDEPENDENT - each student's test score does not depend on the others (assuming no cheating)

NORMAL- The Normal Probability Plot is close to a straight line, so the distribution of the residuals is approximately Normal

EQUAL VARIATION-The residual plot shows an even distribution of residuals above and below the residual=0 line

RANDOM- Students were randomly assigned to their row

12.1 Inference for Linear Regression (Day 3)

vocab

examples

standard error of the slope

$$SE_b = \frac{s \leftarrow \text{st.dev. Resid}}{\text{St. dev. (x)} \rightarrow s_x\sqrt{n-1}}$$

interpret: how far we expect our estimate to be, on average, from the true slope value in repeated samples

z distribution

$$z = \frac{b - \beta}{\sigma_b}$$

t distribution

$$t = \frac{b - \beta}{SE_b}$$

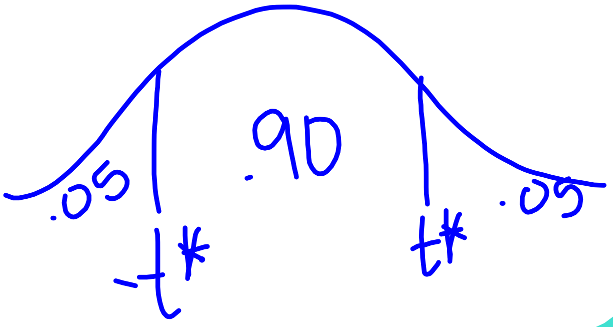
with df = n-2

t Interval for the Slope of a Least-Squares Regression Line

When the conditions for the regression inference are met, a level C confidence interval for the slope β of the population (true) regression line is :

$$b \pm t^* SE_b$$

where t* is the critical value for the t distribution with df = n-2 having area C between -t* and t*



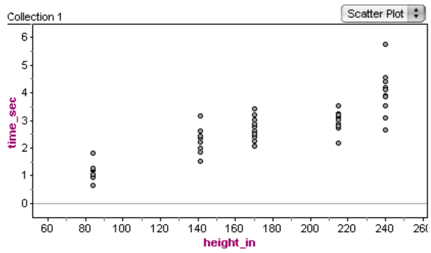
Helicopters

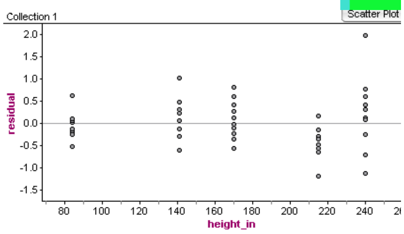
Construct and interpret a 95% confidence interval for the slope of the true regression line between drop height and flight time.

Collection 1

	time_sec
height_in	0.861262 0.0166569 -0.205145 0.00143354 0.741773

S1 = correlation( )
S2 = linRegrSlope(height_in, time_sec)
S3 = linRegrIntercept(height_in, time_sec)
S4 = linRegrSESlope(height_in, time_sec)
S5 = rSquared(height_in, time_sec)





STATE: We want to estimate β, the true slope for the regression line between drop height and drop time of paper helicopters at a 99% confidence level.

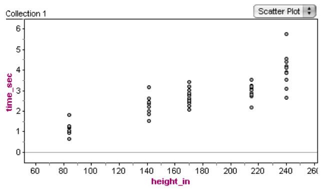
Helicopters

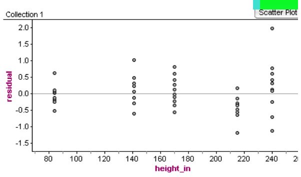
Construct and interpret a 95% confidence interval for the slope of the true regression line between drop height and flight time.

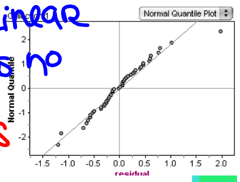
Collection 1

	time_sec
height_in	0.861262 0.0166569 -0.205145 0.00143354 0.741773

S1 = correlation( )
S2 = linRegrSlope(height_in, time_sec)
S3 = linRegrIntercept(height_in, time_sec)
S4 = linRegrSESlope(height_in, time_sec)
S5 = rSquared(height_in, time_sec)







PLAN- Linear: The scatter plot shows a linear pattern & the residual plot shows no clear pattern.

Independent - each helicopter's drop is independent of the others.

Normal - The Normal probability plot is fairly linear, so the residuals are approx. Normal

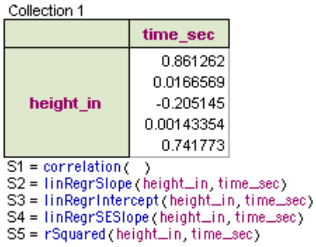
Equal Variance - There are about the same amount of dots above & below the residual=0 line.

Random - helicopters randomly assigned.

*we will use a t-interval to estimate β

Helicopters

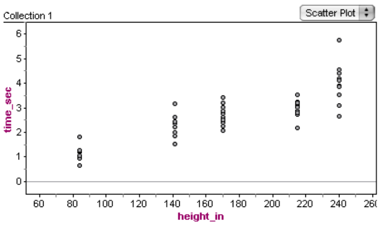
Construct and interpret a 95% confidence interval for the slope of the true regression line between drop height and flight time.

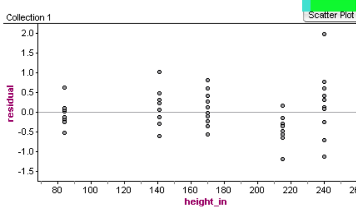


Collection 1

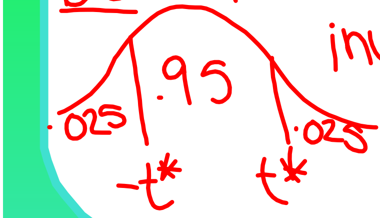
	time_sec
height_in	0.861262 0.0166569 -0.205145 0.00143354 0.741773

S1 = correlation( )
S2 = linRegrSlope(height_in, time_sec)
S3 = linRegrIntercept(height_in, time_sec)
S4 = linRegrSESlope(height_in, time_sec)
S5 = rSquared(height_in, time_sec)





DO: df=49-2=47



invT(.025, 47) = $\boxed{\pm 2.01 = \pm t^*}$

area, df

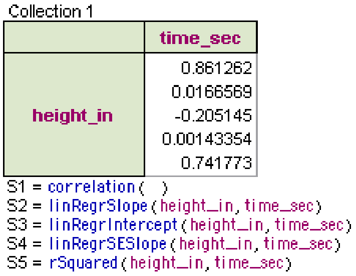
$b \pm t^*(SE_b)$

$.0166569 \pm 2.01(.00143354)$

$\boxed{(.0138, .0195)}$

Helicopters

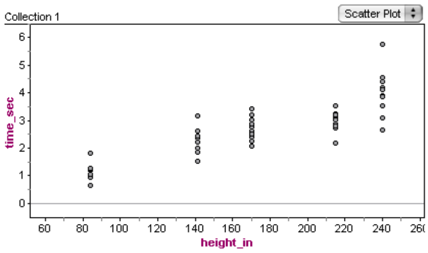
Construct and interpret a 95% confidence interval for the slope of the true regression line between drop height and flight time.

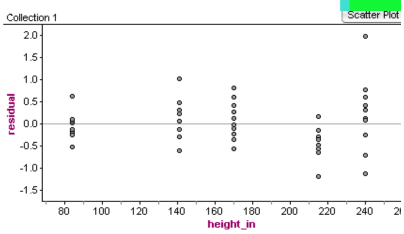


Collection 1

	time_sec
height_in	0.861262 0.0166569 -0.205145 0.00143354 0.741773

S1 = correlation( )
S2 = linRegrSlope(height_in, time_sec)
S3 = linRegrIntercept(height_in, time_sec)
S4 = linRegrSESlope(height_in, time_sec)
S5 = rSquared(height_in, time_sec)





CONCLUDE: We are 95% confident that the interval from .0138 to .0195 contains the true slope of the true Regression line relating height & time.

Fresh Flowers

For their second-semester project, two AP Statistics students decided to investigate the effect of sugar on the life of cut flowers. They went to the local grocery store and randomly selected 12 carnations. All the carnations seemed equally healthy when they were selected. When the students got home, they prepared 12 identical vases with exactly the same amount of water in each vase. They put one tablespoon of sugar in 3 vases, two tablespoons of sugar in 3 vases, and three tablespoons of sugar in 3 vases. In the remaining 3 vases, they put no sugar. After the vases were prepared and placed in the same location, the students randomly assigned one flower to each vase and observed how many hours each flower continued to look fresh. Here are the data and computer output. Construct and interpret a 99% confidence interval for the slope of the true regression line.

Sugar (tbs.)	Freshness (hours)
0	168
0	180
0	192
1	192
1	204
1	204
2	204
2	210
2	210
3	222
3	228
3	234

```
Predictor      Coef       SE Coef      T      P
Constant       181.200    3.635        49.84  0.000
Sugar (tbs)    15.200     1.943        7.82   0.000

S = 7.52596   R-Sq = 86.0%  R-Sq(adj) = 84.5%
```

Fresh Flowers

Construct and interpret a 99% confidence interval for the slope of the true regression line.

```
Predictor      Coef       SE Coef      T      P
Constant       181.200    3.635        49.84  0.000
Sugar (tbs)    15.200     1.943        7.82   0.000

S = 7.52596   R-Sq = 86.0%  R-Sq(adj) = 84.5%
```

Sugar (tbs.)	Freshness (hours)
0	168
0	180
0	192
1	192
1	204
1	204
2	204
2	210
2	210
3	222
3	228
3	234

STATE: We want to estimate the true slope β of the true regression line relating sugar and freshness of flowers at a 99% confidence level

Fresh Flowers

Construct and interpret a 99% confidence interval for the slope of the true regression line.

```
Predictor     Coef       SE Coef      T      P
Constant      181.200    3.635        49.84  0.000
Sugar (tbs)   15.200     1.943        7.82   0.000

S = 7.52596   R-Sq = 86.0%  R-Sq(adj) = 84.5%
```

Sugar (tbs.)	Freshness (hours)
0	168
0	180
0	192
1	192
1	204
1	204
2	204
2	210
2	210
3	222
3	228
3	234

PLAN: Linear - The scatter plot shows a linear pattern and the residual plot is scattered with no clear pattern

Independent - each flower's freshness is independent of the others

Normal - The Normal Probability Plot shows forms close to a line, so the residuals are approx. Normal

Equal Variance - There are about the same amount of dots above and below the residual = 0 line on the residual plot.

Random - The flowers were randomly assigned

*We will use a t interval to estimate β

Fresh Flowers

Construct and interpret a 99% confidence interval for the slope of the true regression line.

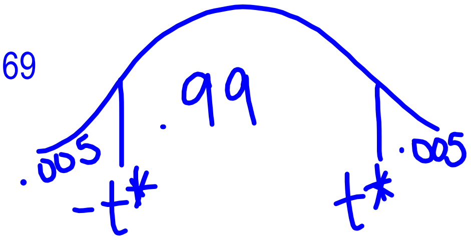
```
Predictor     Coef         SE Coef        T      P
Constant      181.200      3.635          49.84  0.000
Sugar (tbs)   15.200 b     1.943 SEb      7.82   0.000

S = 7.52596   R-Sq = 86.0%  R-Sq(adj) = 84.5%
```

Sugar (tbs.)	Freshness (hours)
0	168
0	180
0	192
1	192
1	204
1	204
2	204
2	210
2	210
3	222
3	228
3	234

DO: df = 12 - 2 = 10

± t = invT(.005,10) = ±3.169

area of below



15.2 ± 3.169 (1.943) = (9.0415, 21.359)

$b \pm t^* (SE_b)$

Fresh Flowers

Construct and interpret a 99% confidence interval for the slope of the true regression line.

```
Predictor     Coef        SE Coef      T      P
Constant      181.200  a  3.635        49.84  0.000
Sugar (tbs)   15.200   b  1.943  SE b  7.82   0.000

S = 7.52596   R-Sq = 86.0%  R-Sq(adj) = 84.5%
```

Sugar (tbs.)	Freshness (hours)
0	168
0	180
0	192
1	192
1	204
1	204
2	204
2	210
2	210
3	222
3	228
3	234

CONCLUDE: We are 99% confident that the interval from 9.0415 to 21.359 contains the true slope relating sugar (tbs) and freshness (hours) of flowers

12.1 Inference for Linear Regression (Day 4)

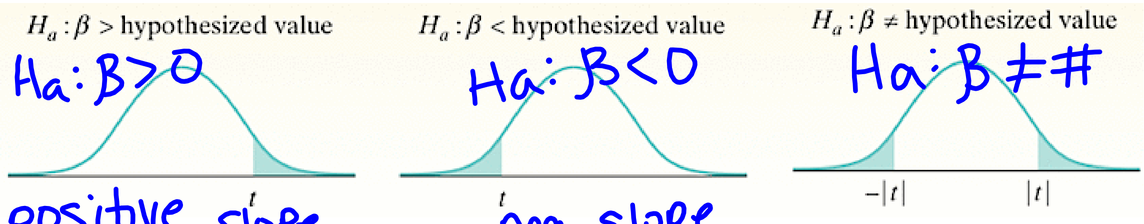
vocab

examples

t Test for Slope

H$_0$: β = hypothesized value $\boxed{H_0: \beta = 0}$

H$_a$ options:



$H_a: \beta > 0$ $H_a: \beta < 0$ $H_a: \beta \neq \#$

positive slope neg. slope

*define β

DO: $t = \dfrac{b - \beta_0}{SE_b}$, tcdf p-value

Crying & IQ

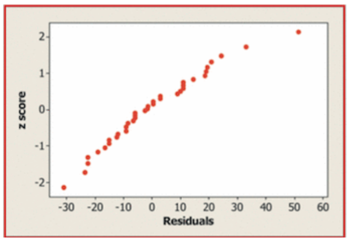
Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds. They later measured the children's IQ at age three years using the Stanford-Binet IQ test. The table below contains data from a random sample of 38 infants

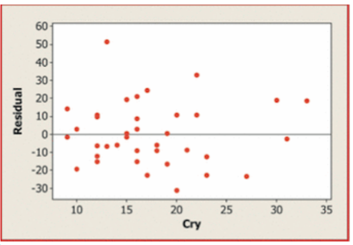
Crying	IQ	Crying	IQ	Crying	IQ	Crying	IQ
10	87	20	90	17	94	12	94
12	97	16	100	19	103	12	103
9	103	23	103	13	104	14	106
16	106	27	108	18	109	10	109
18	109	15	112	18	112	23	113
15	114	21	114	16	118	9	119
12	119	12	120	19	120	16	124
20	132	15	133	22	135	31	135
16	136	17	141	30	155	22	157
33	159	13	162				

Crying & IQ

Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants? Carry out an appropriate test to help answer this question.

Crying	IQ	Crying	IQ	Crying	IQ	Crying	IQ
10	87	20	90	17	94	12	94
12	97	16	100	19	103	12	103
9	103	23	103	13	104	14	106
16	106	27	108	18	109	10	109
18	109	15	112	18	112	23	113
15	114	21	114	16	118	9	119
12	119	12	120	19	120	16	124
20	132	15	133	22	135	31	135
16	136	17	141	30	155	22	157
33	159	13	162				





Regression Analysis: IQ versus Crycount

```
Predictor      Coef     SE Coef       T       P
Constant     91.268     8.934      10.22   0.000
Crycount  b   1.4929    0.4870  SEb  3.07   0.004
S = 17.50  R-Sq = 20.7%  R-Sq(adj) = 18.5%
```

Crying & IQ

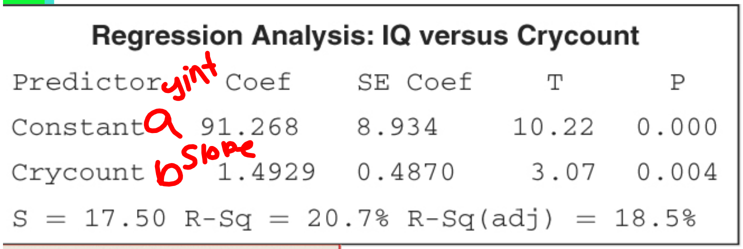
Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants? Carry out an appropriate test to help answer this question.

State: H_0: $\beta = 0$ H_a: $\beta > 0$ $\alpha = 0.05$

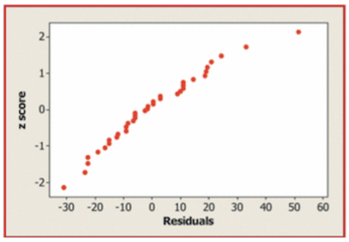
β is the true slope for the population regression line relating crying and IQ for all infants

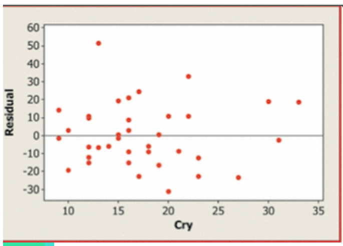
Crying & IQ

Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants? Carry out an appropriate test to help answer this question.



```
Regression Analysis: IQ versus Crycount

Predictor yint  Coef     SE Coef       T       P
Constant a    91.268     8.934     10.22   0.000
Crycount b slope 1.4929  0.4870     3.07   0.004

S = 17.50  R-Sq = 20.7%  R-Sq(adj) = 18.5%
```





PLAN: Linear - The residual plot shows no clear pattern.

Independent - assuming at least 380 infants.

Normal- The Normal probability plot shows that the residuals are approx. Normal since it is fairly linear.

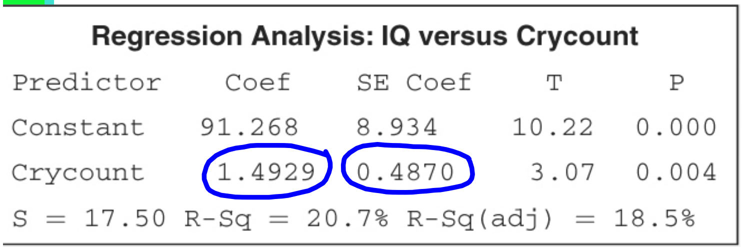
Equal Variance- There are about the same amount of dots above and below the residual=0 line on the residual plot.

Random - "random sample of infants"

*we will use a t test for β

Crying & IQ

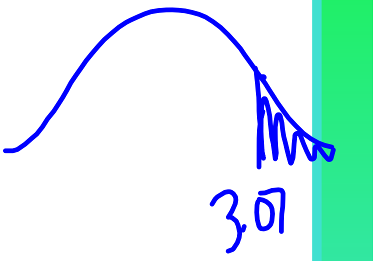
Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants? Carry out an appropriate test to help answer this question.



```
Regression Analysis: IQ versus Crycount

Predictor      Coef     SE Coef       T       P
Constant     91.268     8.934     10.22   0.000
Crycount     1.4929     0.4870     3.07   0.004

S = 17.50  R-Sq = 20.7%  R-Sq(adj) = 18.5%
```

DO:

$t = \frac{1.4929 - 0}{.4870} = 3.07$ df = 38-2 = 36



p-value: tcdf(3.07, 1E99, 36) = .0020
lower, upper, df

Crying & IQ

Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants? Carry out an appropriate test to help answer this question.

CONCLUDE: .0020 < .05 --> We reject H_0 --> We can conclude H_a: $\beta>0$. There is convincing evidence that the true slope for the regression line relating the population of infants crying and IQ is positive.