1 Inference for Quantitative Data: Slopes

1.1 Sampling Distributions and Confidence Intervals for Slope

Population Regression Line

- An "ideal" linear relationship can be described with a population regression line: $\mu_y = \alpha + \beta x$
  - Where μ_y represents the mean value of the response variable y for any given value of the explanatory variable x
  - α represents the population y-intercept and β represents the population slope
- An observed linear relationship can be described with a sample regression line: $\hat{y} = a + bx$
- If we took many LSRLs of the same size from the sample population, we can create a sampling distribution for our slope.

Sampling Distribution for the slope of a LSRL

- For a bivariate population with a given slope, β, a standard deviation of residuals σ, and a standard deviation of x-values σ_x.
- If you take all samples of size n and compute the slope of each of those samples, you get the sampling distribution:
  - Shape: The distribution of sample slopes is approximately normal.
  - Mean: $\mu_b = \beta$
  - Standard Deviation: $\sigma_b = \frac{\sigma}{\sigma_x\sqrt{n}}$, where σ is the standard deviation of residuals, σ_x is the standard deviation for the explanatory variable, and n is sample size.

Once we develop a sampling distribution for our slope, we can begin to ask and answer our inference questions:

- Is there a linear relationship between x and y in the population, or could the pattern we see happen just by chance?
- In the population, how much will the predicted value of y chance for each increase of 1 unit in x?

Conditions for Regression Inference

- Linear: x and y have a linear relationship. Check: make sure scatter plot can be described by a line
- Independent: If sampling without replacement, check the 10% condition.
- Normal Residuals: When x is fixed, y follows a normal distribution. Check: Make a histogram of the residuals and make sure it looks approximately normal. If the graph has outliers or strong skewness, n should be larger than 30.
- Equal SD: Standard deviation of residuals doesn't vary with x. Check: Make a residual plot and check for a random pattern
- Random: Random sampling (SRS) or random assignment (experiment)

Together, this makes up the acronym LINER, which can help you remember what conditions to check when creating a confidence interval or running a significance test.

A C% Confidence interval is created to estimate the slope β of the population (true) regression line.

$$b \pm t^* \left(\frac{s}{s_x\sqrt{n-1}} \right) \text{ with df } = n-2$$

- t^* has C% of the area between $-t^*$ and t^*

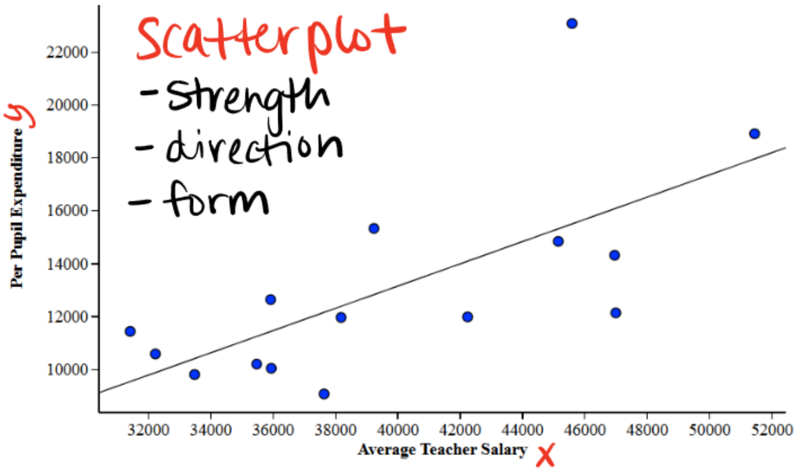
- b is the point estimate (slope from our sample data)
- $SE_b = \frac{s}{s_x\sqrt{n-1}}$ is the standard error of the slope
- s_x is the standard deviation of x-values
- s is the standard deviation of the residuals

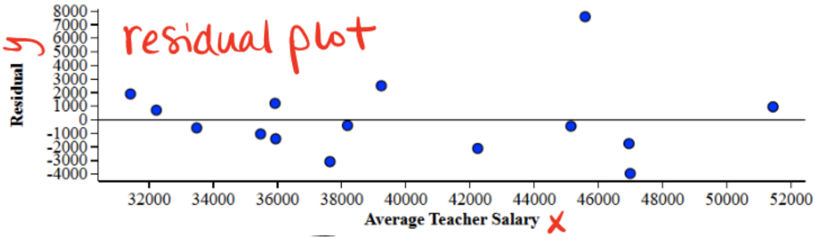
Interpretation: We are C% confident that the interval from ______& ______captures the true slope of the regression line between x-variable and y-variable.

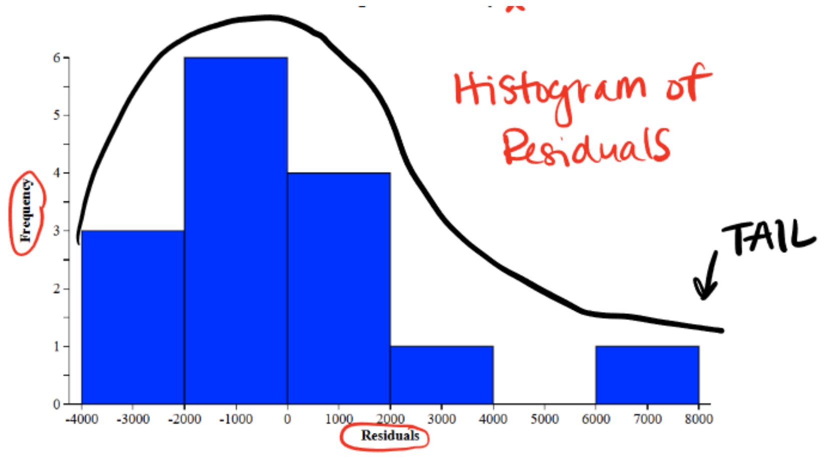
Example

The data below was obtained from 15 randomly selected large school districts from around the nation. It shows what their average teacher salary is and how much they spend per student (per pupil expenditure).

x Average Teacher Salary	y Per Pupil Expenditure
$31,418	$11,443
$32,226	$10,589
$33,483	$9,809
$35,474	$10,205
$35,923	$12,645
$35,943	$10,045
$37,636	$9,075
$38,181	$11,968
$39,236	$15,337
$42,240	$11,989
$45,147	$14,848
$45,589	$23,091
$46,954	$14,322
$46,992	$12,143
$51,443	$18,920







Data:

- $n = 15$
- $r = 0.684$
- $r^2 = 0.468$
- $s = 2865.8$
- $s_x = 6154.2$
- $s_y = 3786.1$
- $\hat{y} = -3679.1 + 0.4208x$

Construct and interpret a 95% confidence interval for the slope of the population regression line.

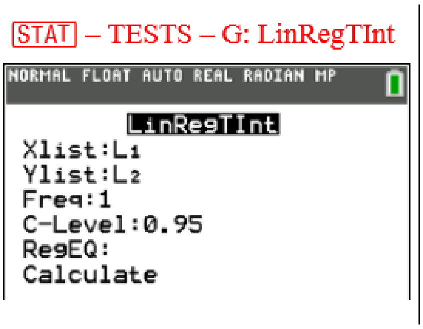
β = true population slope between average teacher salary and per pupil expenditure.

- Scatterplot shows a moderately positive linear relationship.
- $n = 15 \leq 0.10$(all school districts in the nation)
- A histograpm of the residuals appears skewed right. (This is the only condition not correctly met.)
- The residual plot shows random scatter around LSRL.
- Randomly selected 15 large school districts.

Linear Regression t-Interval for Slope

Using the formula given, we can use invT to get $t^* = 2.1604$ and the formula to get the confidence interval $(0.1519, 0.6897)$.

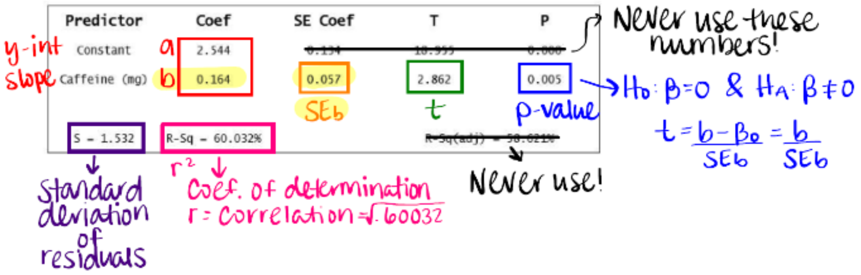
Calculator Steps:


STAT – TESTS – G: LinRegTInt

```
NORMAL FLOAT AUTO REAL RADIAN MP
         LinRegTInt
Xlist:L1
Ylist:L2
Freq:1
C-Level:0.95
RegEQ:
Calculate
```

We are 95% confident that the interval from 0.1519 to 0.6897 captures the true slope of the regression line between average teacher salary and per pupil expenditure. However, because our "Normal Residuals" condition was not met, we should be careful with this interpretation because it might not be correct.

Most AP Problems will not require you to do what we did in the previous example. Most inference questions come with a computer output, like what is pictured below.



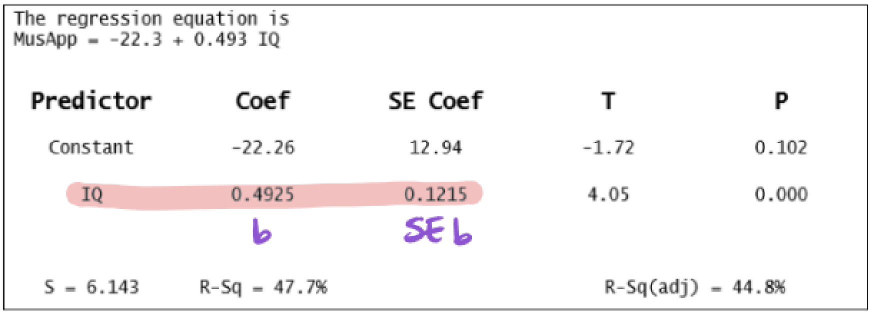
```
Predictor        Coef     SE Coef      T          P
Constant         2.544    ~~0.134~~    ~~18.955~~ ~~0.000~~
Caffeine (mg)    0.164    0.057        2.862      0.005

S = 1.532   R-Sq = 60.032%              ~~R-Sq(adj) = 58.021%~~
```

y-int
slope
a
b
SEb
t
p-value
Never use these numbers!
$\rightarrow H_0: \beta = 0$ & $H_A: \beta \neq 0$
$t = \frac{b - \beta_0}{SE_b} = \frac{b}{SE_b}$
Standard deviation of residuals
r^2
Coef. of determination
r = Correlation = $\sqrt{.60032}$
Never use!

Example

A study attempted to establish a linear relationship between IQ score and musical aptitude. The following table is a partial printout of the regression analysis based on a sample of 20 individuals.



```
The regression equation is
MusApp = -22.3 + 0.493 IQ

Predictor      Coef      SE Coef      T        P
Constant      -22.26     12.94       -1.72     0.102
IQ             0.4925    0.1215       4.05     0.000

S = 6.143     R-Sq = 47.7%               R-Sq(adj) = 44.8%
```

Construct and interpret a 99% confidence interval for the slope of the regression line. Does it suggest a linear relationship? Assume all assumptions and conditions for inference have been met.

β = true population slope between IQ score and musical aptitude

All assumptions and conditions met

Linear Regression t-Interval for Slope

$t^* = \text{invT(area} = .995, \text{ df} = 18) = 2.8784$

$0.4925 \pm 2.8784(0.1215) = (0.1428, 0.8422)$

We are 99% confident that the interval from 0.1428 to 0.8422 captures the true slope of the regression line between IQ score and musical aptitude.

1.2 Significance Test for a Slope

- The significance test for a slope is called a Linear Regression t-Test for Slope.
- It can help us answer three different questions with the hypotheses

Is the relationship between the explanatory and response variable negative?

- $H_0 : \beta = 0$
- $H_A : \beta < 0$

Is there a relationship between the explanatory and response variable?

- $H_0 : \beta = 0$
- $H_A : \beta \neq 0$

Is the relationship between the explanatory and response variable positive?

- $H_0 : \beta = 0$
- $H_A : \beta > 0$

Conditions for Regression Inference: Same for the confidence interval

The test statistics is $t = \frac{\text{statistic-parameter}}{\text{standard error}}$, or $\frac{b}{SE_b}$, where $SE_b = \frac{s}{s_x\sqrt{n-1}}$.

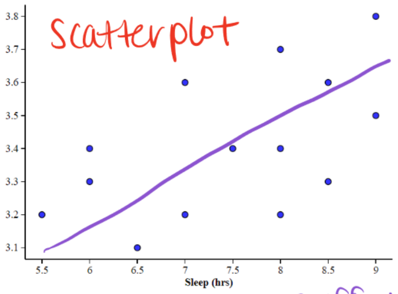
df = n-2, p- value is calculated using your calculator, in the direction of the alternative hypothesis. p-value = tcdf(lower, upper, df)

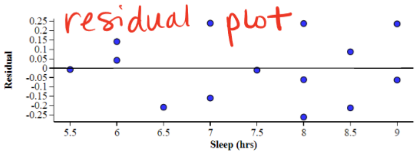
- In your conclusion, you would state the results of your significance test (reject or fail to reject) and then interpret the findings in context
- Note: Having a low p-value and finding evidence of the alternative hypothesis of some linear association does not mean that the association is strong

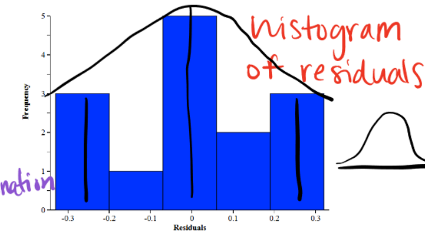
Example

A school counselor is concerned that the number of hours of sleep his students get each night is affecting their GPA in a negative way. He selects a random sample of 14 seniors in his district and asks them how many hours of sleep they get on a typical school night. He then uses school records to determine the most recent grade-point average (GPA) for each student. His data are given below.

X	Sleep (hours)	9	8.5	9	7	7.5	6	7	8	5.5	6	8.5	6.5	8	8
y	GPA	3.8	3.3	3.5	3.6	3.4	3.3	3.2	3.2	3.2	3.4	3.6	3.1	3.4	3.7







coeff. of determination

Equation	n	r^2	s	s_x
$\hat{y} = 2.65 + 0.102x$	14	0.313	0.18	1.15

$\hat{y} = a + bx$

Do these data provide convincing evidence, at the 0.05 significance level, of a positive linear relationship between the hours of sleep students typically get and their academic performance?

β = true pop. slope between hours of sleep students typically get and their GPA.

$H_0 : \beta = 0,\ H_A : \beta > 0$

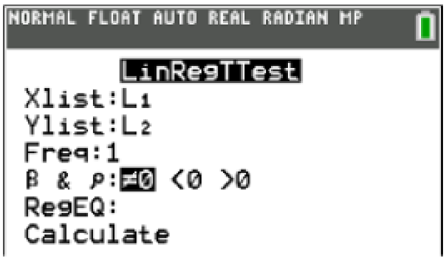
- Scatterplot shows a weak positive linear relationship.
- $n = 14 \leq 0.10$(all HS Seniors)
- Histogram of residuals doesn't appear normal but no strong skew or outliers
- Residual plot shows random scatter
- Random sample of 14 HS seniors

Linear Regression t-Test for Slope

$t = \frac{0.102}{\frac{0.18}{1.15\sqrt{13}}} = 2.3496$, tcdf with this gives $p = 0.0184$.

You can also run this on the calculator

$\boxed{\text{STAT}}$ – TESTS – F: LinRegTTest



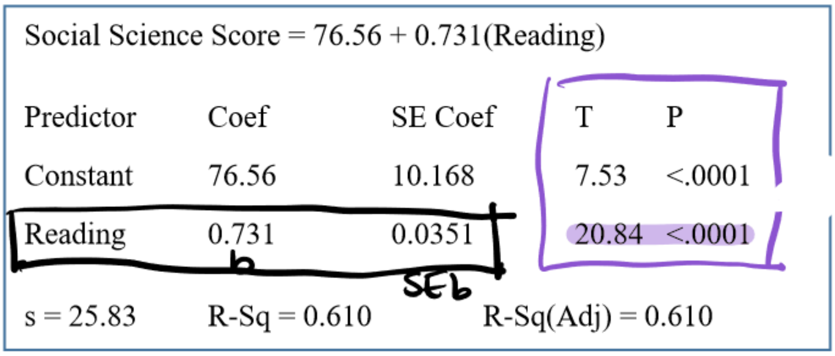
```
NORMAL FLOAT AUTO REAL RADIAN MP

        LinRegTTest
 Xlist:L1
 Ylist:L2
 Freq:1
 β & ρ:≠0 <0 >0
 RegEQ:
 Calculate
```

Since the p-value of 0.0184 is less than $\alpha = 0.05$, we reject the null. There is convincing evidence of a positive linear relationship between hours of sleep per night and GPA for HS seniors.

Example

The computer output given shows a regression analysis of an honors social science course (score in points) versus a reading comprehension score (in points) for 25 sophomores at your school.



```
Social Science Score = 76.56 + 0.731(Reading)

Predictor    Coef     SE Coef    T      P
Constant     76.56    10.168     7.53   <.0001
Reading      0.731    0.0351     20.84  <.0001

s = 25.83    R-Sq = 0.610    R-Sq(Adj) = 0.610
```

Carry out a hypothesis test for these data to determine if there is a linear relationship and interpret your results in the context of the problem. Assume all assumptions and conditions for inference have been met.

$\beta =$ true pop. slope between social science score and reading score.

$H_0 : \beta = 0$, $H_A : \beta \neq 0$

Linear Regression t-Test for Slope

$t = \frac{0.731}{0.0351} = 20.8262$, tcdf with this gives $p \approx 0 \times 2 \approx 0$.

Since the p-value of approx. 0 is less than $\alpha = 0.05$, we reject the null. There is convincing evidence of a linear relationship between social science score and reading score for your school's sophomores.