L
I
N
E
R

Conditions for inference for a slope

So we can estimate the population LSRL with the LSRL from a sample.

busy day today, tight for time...

Conditions for Inference for Regression

(For us, it means when doing inference for a slope)

L
I
N
E
R

Conditions for Inference for Regression

(For us, it means when doing inference for a slope)

Linear

Independent

Normal

Equal SD

Random

Conditions for Inference for Regression

(For us, it means when doing inference for a slope)

Linear - Scatter plots should show a roughly linear relationship and the residual plot should show random scatter with no curved pattern

Independent

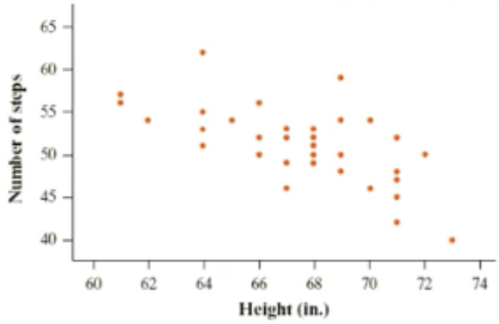
Normal

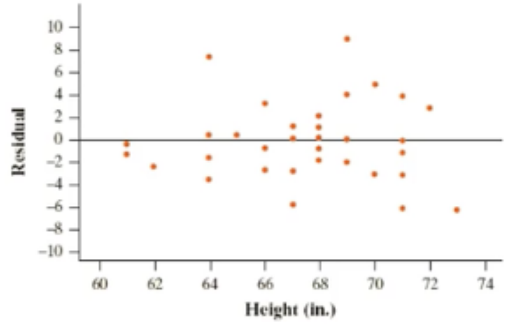
Equal SD

Random

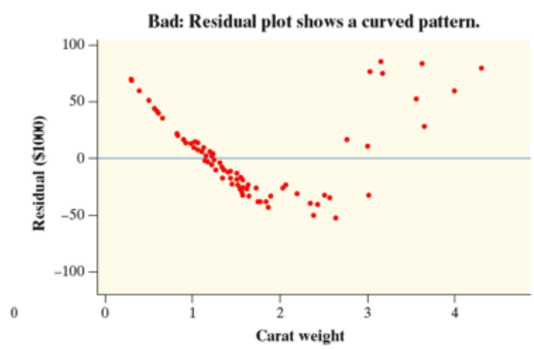
Linear

To check: Scatterplot should show a roughly linear relationship. Residual plot should show a random scatter with no curved pattern.





Other Example - Linear



Residuals

BAD

Conditions for Inference for Regression

(For us, it means when doing inference for a slope)

Linear - Scatter plots should show a roughly linear relationship and the residual plot should show random scatter with no curved pattern

Independent One response should not affect another response. If sampling is done w/out replacement, then check 10% condition.

Normal

Equal SD

Random

Conditions for Inference for Regression

(For us, it means when doing inference for a slope)

Linear - Scatter plots should show a roughly linear relationship and the residual plot should show random scatter with no curved pattern

Independent One response should not affect another response. If sampling is done w/out replacement, then check 10% condition.

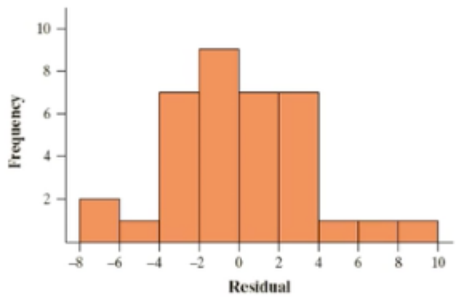
Normal Look at a graph (like a histogram) of the residuals but not a residual plot. Should see no strong skew or outliers. also satisfied if $n \geq 30$

Equal SD

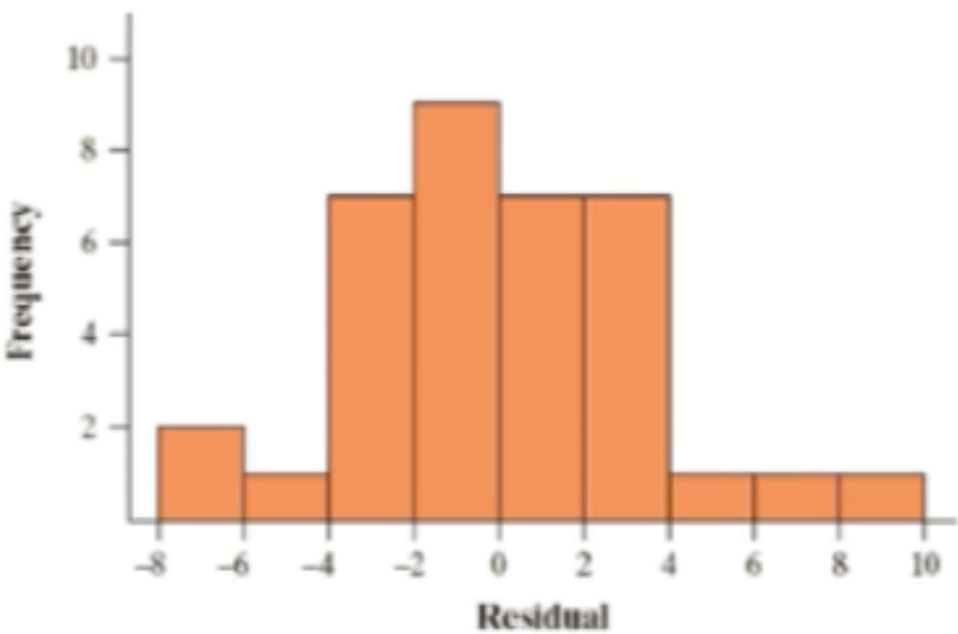
Random

Normality

To check: A graph of the residuals should not show strong skewness or outliers.



The histogram of the residuals does not show strong skewness or outliers.



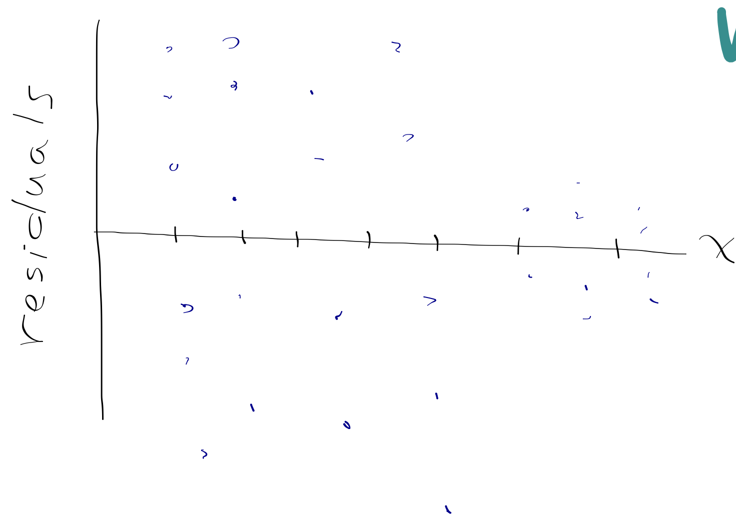
a dot plot could work in some cases

Conditions for Inference for Regression

(For us, it means when doing inference for a slope)

Linear -	Scatter plots should show a roughly linear relationship and the residual plot should show random scatter with no curved pattern
Independent	One response should not affect another response. If sampling is done w/out replacement, then check 10% condition.
Normal	Look at a graph (like a histogram) of the residuals but not a residual plot. should see no strong skew or outliers also satisfied if $n \geq 30$
Equal SD	The stand dev of y-values, σ_y, should not vary with x Residual Plot → look for approx. equal std. deviations for all x
Random	

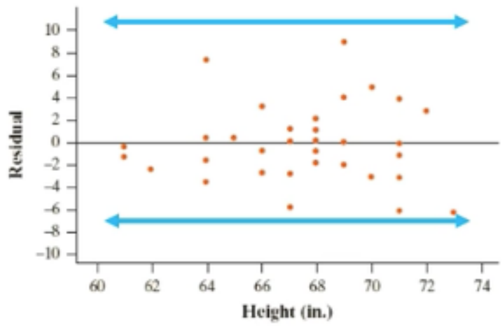
extreme example



would not meet

Equal SD

To check: The variability of the residuals should be roughly constant for all x values.



Looking for major violations only

The residual plot shows roughly equal amounts of scatter for all x values.

Conditions for Inference for Regression

(For us, it means when doing inference for a slope)

Linear - Scatter plots should show a roughly linear relationship and the residual plot should show random scatter with no curved pattern.

Independent One response should not affect another response. If sampling is done w/out replacement, then check 10% condition.

Normal Look at a graph (like a histogram) of the residuals but not residual plot. Should see no strong skew or outliers. also satisfied if $n > 30$

Equal SD The stand dev of y-values, σ_y, should not vary with x. Residual Plot → look for approx. equal std. deviations for all x.

Random Need random sample from poplation of interest or random assignments of treatments in an experiment.

Estimating the Parameters

When the conditions are met, we can do inference about the regression model $\cancel{\mu_y = \beta_0 + \beta_1 x}$. The first step is to estimate the unknown parameters.

$\mu_y = \alpha + \beta x$

Estimating the Parameters

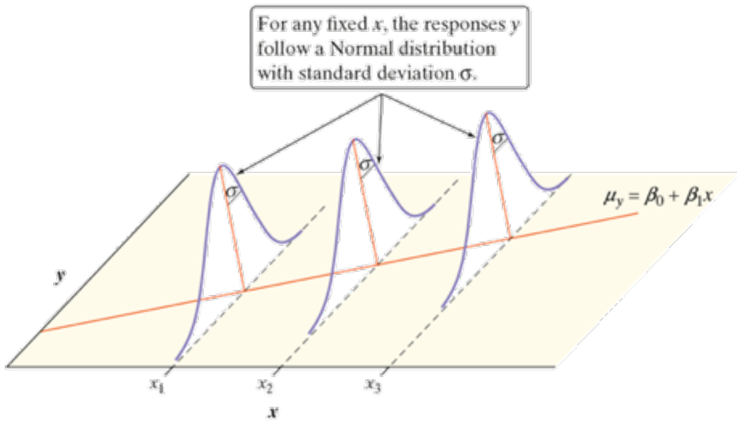
When the conditions are met, we can do inference about the regression model $\mu_y = \cancel{\beta_0 + \beta_1 x}$ $\alpha + \beta x$. The first step is to estimate the unknown parameters.

If we calculate the sample regression line $\cancel{\hat{y} = b_0 + b_1 x}$, the residuals estimate how much y varies about the population regression line.

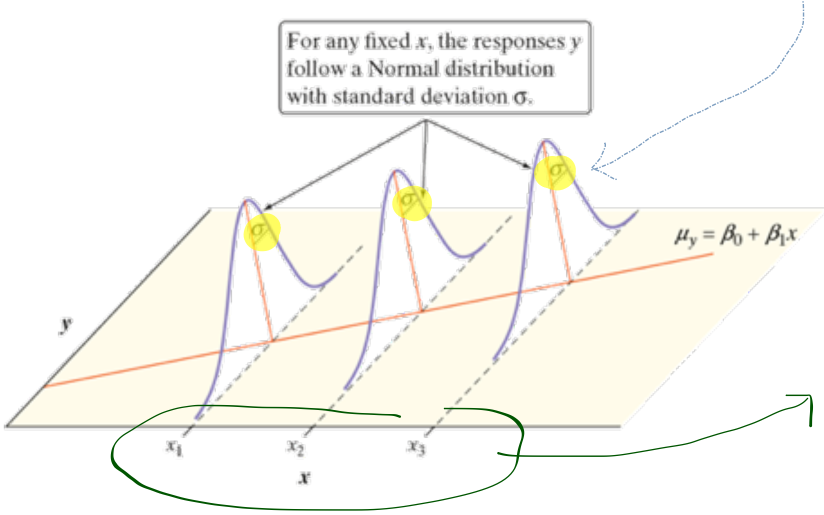
$\hat{y} = a + bx$

When the conditions are met, the sampling distribution of the slope b_1 is approximately Normal with mean $\mu_{b_1} = \beta_1$ and **standard deviation** of the slope

$$\sigma_{b_1} = \frac{\sigma}{\sigma_x \sqrt{n}}$$



$$\sigma_{b_1} = \frac{\sigma}{\sigma_x \sqrt{n}}$$



σ_x

When the conditions are met, the sampling distribution of the slope b_1 is approximately Normal with mean $\mu_{b_1} = \beta_1$ and **standard deviation**

$$\sigma_b \qquad \sigma_{b_1} = \frac{\sigma}{\sigma_x\sqrt{n}}$$

don't know σ for true regression line

We **ESTIMATE** the variability of the sampling distribution of b_1 with the ***standard error of the slope***

$$SE_{b_1} = \frac{s}{s_x\sqrt{n-1}}$$

so we estimate it with std. deviat. of the residuals

We also don't know the std. deviation of the population x-values, σ_x

When the conditions are met, the sampling distribution of the slope b_1 is approximately Normal with mean $\mu_{b_1} = \beta_1$ and **standard deviation**

$$\sigma_b \qquad \cancel{\sigma_{b_1}} = \frac{\sigma}{\sigma_x\sqrt{n}}$$

for reasons beyond this course

We **ESTIMATE** the variability of the sampling distribution of b_1 with the ***standard error of the slope***

$$S_b \qquad \cancel{SE_{b_1}} = \frac{s}{s_x\sqrt{n-1}}$$

We estimate the Variability of the sampling distrib. of slope with the Std Error of the Slope.

Look at the
very last part of
your formula sheet

Random Variable

For slope:

b

Parameters of Sampling Distribution		Standard Error* of Sample Statistic
$\mu_b = \beta$	$\sigma_b = \dfrac{\sigma}{\sigma_x\sqrt{n}}$,	$s_b = \dfrac{s}{s_x\sqrt{n-1}}$,
	where $\sigma_x = \sqrt{\dfrac{\sum(x_i-\bar{x})^2}{n}}$	where $s = \sqrt{\dfrac{\sum(y_i-\hat{y}_i)^2}{n-2}}$
		and $s_x = \sqrt{\dfrac{\sum(x_i-\bar{x})^2}{n-1}}$

You will have to interpret it.... like we did in the helicoptor example from the last class.

This standard error is interpreted as how far the sample slope typically varies from the population (true) slope if we repeat the data production process many times.

SE_{b_1} = 0.0002018; if we repeated the random assignment many times, the slope of the sample regression line would typically vary by about 0.0002018 from the slope of the true regression line for predicting flight time from drop height.

Aim today

✓CONSTRUCT and INTERPRET a confidence interval for the slope β_1 of the population (true) regression line.

Does Seat Location Matter

– Part II

Lesson 12.1: Day 2: Does seat location matter – Part 2?

Do students who sit in the front rows do better than students who sit farther away? Mrs. Gallas took a random sample of 30 students from her classes and found these results.

Row	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
Score	76	77	94	99	88	90	83	85	74	79	77	79	90	88	68	78	83	79

Row	4	4	4	4	4	4	5	5	5	5	5	5
Score	94	72	101	70	63	76	76	65	67	96	79	96

Line of best fit:________________

Slope: b = ________ SE_b = 1.33

$S_b = 1.33$

Row	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
Score	76	77	94	99	88	90	83	85	74	79	77	79	90	88	68	78	83	79

Row	4	4	4	4	4	4	5	5	5	5	5	5
Score	94	72	101	70	63	76	76	65	67	96	79	96

Line of best fit:________________

Slope: b = ________ SE_b = 1.33

1. If Mrs. Gallas were to take another random sample of 30 students, do you think the slope of the LSRL would be the same? Why?

2. **We are going to construct a 95% confidence interval for the slope of the population regression line. Identify the parameter and statistic.**

Parameter:________________________________ Statistic:____________

Row	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
Score	76	77	94	99	88	90	83	85	74	79	77	79	90	88	68	78	83	79

Row	4	4	4	4	4	4	5	5	5	5	5	5
Score	94	72	101	70	63	76	76	65	67	96	79	96

Line of best fit: $\hat{y} = 85.95 - 1.517x$

Slope: $b = -1.517$ $SE_b = 1.33$

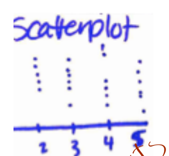
1. If Mrs. Gallas were to take another random sample of 30 students, do you think the slope of the LSRL would be the same? Why? No, every sample will lead to different results with a new LSRL and slope.

2. **We are going to construct a 95% confidence interval for the slope of the population regression line. Identify the parameter and statistic.**

Parameter: β true slope of population LSRL Statistic: $b_1 = -1.517$

3. There are five conditions to check.

(1) **Linear:** The **scatterplot** needs to show a linear relationship AND the **residual plot** doesn't have a leftover curved pattern. Sketch each at right. ✓



(2) **Independent:** **Use 10% condition IF sampling without replacement** ✓

(3) **Normal:** A **dotplot of the residuals (or a histogram)** cannot show strong skew or outliers. Make one using the applet and sketch it at right. ✓ also meets $30 \geq 30$

Dot Plot of Residuals

(4) **Equal SD:** Look at Residual Plot - the variability in the residuals in the vertical direction should be ROUGHLY the same as you scan across most of the x-values. No sideways Christmas tree patterns, for example.

(5) **Random:** Either "SRS" or "Random Assignment"

4. **Construct the interval:**

General Formula:

Specific Formula:

Work:

4. **Construct the interval:**

General Formula: Pt. Estim ± MOE

Specific Formula: $b_1 \pm t^* \cdot SE_{b_1}$

$df = n-2 = 30-2 = 28$

Work: $-1.517 \pm 2.048 \times 1.33$ given

$(-4.24, 1.21)$

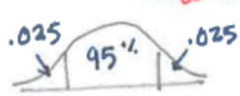


TABLE B or invT $\Big\}\ t^* = 2.048$

4. **Construct the interval:**

General Formula: Pt. Estim ± MOE

Specific Formula: $b \pm t^* \cdot SE_b$

$df = n-2 = 30-2 = 28$

Work: $-1.517 \pm 2.048 \times 1.33$ given

$(-4.24, 1.21)$

TABLE B or invT $t^* = 2.048$

5. **Conclude:**

We are 95% confident that the interval from (-4.24, 1.21) captures the true slope of the population regression line relating y = score and x = row

Confidence Intervals for Slope

Important ideas: State

Formulas

Confidence Intervals for Slope

Important ideas:

State

95% CI for β

Formulas

Confidence Intervals for Slope

Important ideas:

State

95% CI for β

β true slope of population LSRL

b statistic — sample LSRL slope

Formulas

point estim $\pm$ MOE

$b_1 \pm t^* \cdot S_b$

df = n − 2

where

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

std dev of resid

std dev of x-values

Confidence Intervals for Slope

Important ideas:

State

95% CI for β

β true slope of population LSRL

b statistic– sample LSRL slope

Formulas

point estim ± MOE

$b_1 \pm t^* \cdot S_b$

$df = n-2$

where $S_b = \frac{s}{S_x\sqrt{n-1}}$

s std dev of resid

S_x std dev of x-values

Formal name → 1 sample t int for β_1 which you would use in "PLAN"

for some TI-84 LinRegTInt

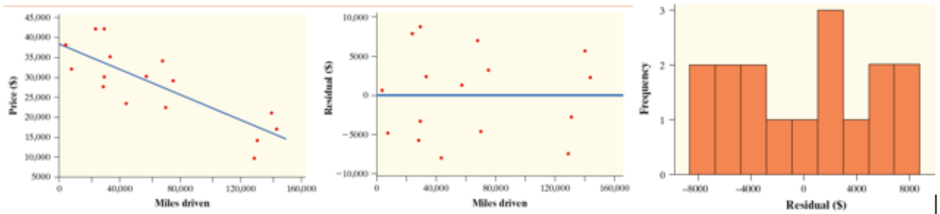
Now... a CI more formally.

Mileage vs Value

-we'll do together

-refer to this example when doing your HW

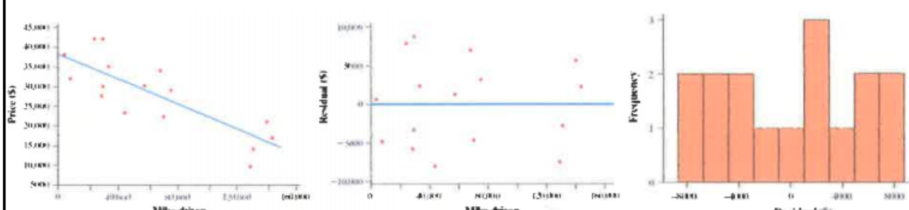
Mileage vs Value-Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 Super Crew 4 x 4 if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 Super Crew 4 x 4s was selected from among those listed for sale on autotrader.com. The number of miles driven and price (in dollars) were recorded for each of the trucks. Here is some computer output from a least-squares regression analysis of these data. Construct and interpret a 90% confidence interval for the slope of the population regression line. *You can assume that the Conditions are met.*



Regression Analysis: Price ($) versus Miles driven

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13 R-Sq = 66.4% R-Sq(adj) = 64.0%



Regression Analysis: Price ($) versus Miles driven

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13 R-Sq = 66.4% R-Sq(adj) = 64.0%

If doing "PLAN" the test would be t-interval for the slope

Regression Analysis: Price ($) versus Miles driven

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13 R-Sq = 66.4% R-Sq(adj) = 64.0%

State

90% CI for β

$\beta \rightarrow$ true slope of the population regression li relating y = price and x = miles driven for used Ford 4x4's

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13 R-Sq = 66.4% R-Sq(adj) = 64.0%

Not the t-value

Do:

$df = 16-2 = 14$, $t^* = 1.761$ ← from TABLE B

$-.16292 \pm 1.761(.0396)$

$-.16292 \pm 0.05452$

$(-.21744, -.10840)$

Conclude

Do:

$df = 16 - 2 = 14$, $t^* = 1.761$ ← from TABLE B

$-.16292 \pm 1.761(.0396)$

$-.16292 \pm 0.05452$

$(-.21744, -.10840)$

Conclude

We are 90% confident that the interval from $-.21744$ to $-.10840$ captures the slope of the population regression line relating y = price to x = miles driven for FORD F-150's listed on auto trader.com

Conclude

We are 90% confident that the interval from $-.21744$ to $-.10840$ captures the slope of the population regression line relating y = price to x = miles driven for FORD F-150's listed on auto trader.com

Note: the CI only contains negative values as plausible values for the slope. Because the interval does not contain 0, we have convincing evidence that that there is a linear relationship.

page 781
Technology Corner
on how to do t intervals for the slope

if given
raw data

LinRegTInt

the calculator gives
(−.02173, −.1084)
using df = 14

```
NORMAL FLOAT AUTO REAL RADIAN CL

LinRegTInt
Xlist:L1
Ylist:L2
Freq:1
C-Level:.9
RegEQ:
Calculate
```

```
NORMAL FLOAT AUTO REAL RADIAN CL

LinRegTInt
y=a+bx
(-.2173,-.1084)
b=-.1628114837
df=14
s=5737.55499
a=38254.8639
r²=.664549225
r=-.8151988868
```

AP® Exam Tip

When you see a list of data values on an exam question, wait a moment before typing the data into your calculator. Read the question through first. Often, information is provided that makes it unnecessary for you to enter the data at all. This can save you valuable time on the AP® Statistics exam.

Older TI-84's (TI-83's) may not have this option

You can still find b and S_b by .....

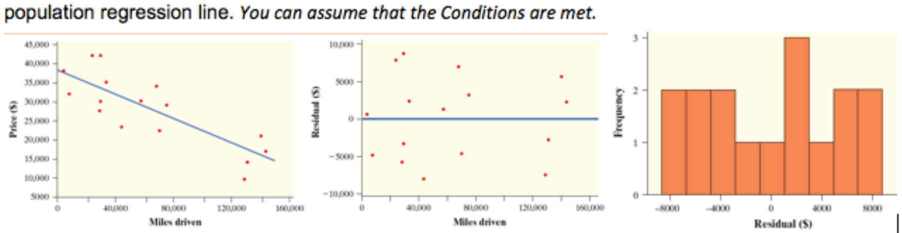
If you use LinRegTTest (see page 785)

$b =$ slope $\quad S_b = \frac{b}{t} \quad$ where $t = \frac{b - 0}{S_b}$

12.1....3, 5, 9, 11

and study pp. 776-782

Mileage vs Value–Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 Super Crew 4 x 4 if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 Super Crew 4 x 4s was selected from among those listed for sale on autotrader.com. The number of miles driven and price (in dollars) were recorded for each of the trucks. Here is some computer output from a least-squares regression analysis of these data. Construct and interpret a 90% confidence interval for the slope of the population regression line. *You can assume that the Conditions are met.*



Regression Analysis: Price ($) versus Miles driven

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13 R-Sq = 66.4% R-Sq(adj) = 64.0%

not t^*