Schedule

Fr
~~Today~~ --- Day 2 of section 12.1

~~Tues~~ --- Day 3 of section 12.1
Mon

+ LCQ on AP Reviews (1 to 3)

TU
~~Wed~~ --- More Practice/Review 12.1

+ Start AP Exam Review ch 4

Wed
~~Thur~~ --- Quiz on Ch. 12 (section 12.1)

I suggest you have the 12.1 handout from last Thursday available

Big Idea: If the data come from a random sample or randomized experiment, the least-squares regression line is just an *estimate* of the population (true) least-squares regression line.

- Population line: $\mu_y = \beta_0 + \beta_1 x$
- Sample line: $\hat{y} = b_0 + b_1 x$

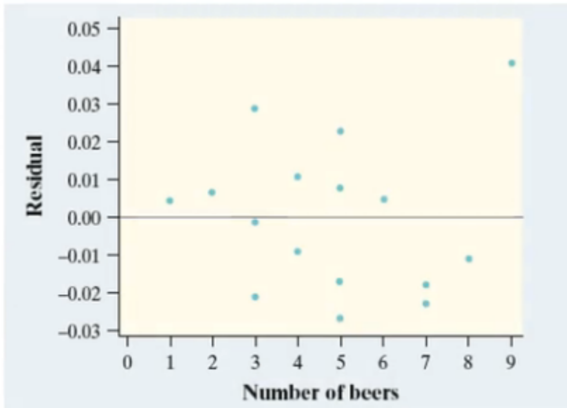
Big Idea: If the data come from a random sample or randomized experiment, the least-squares regression line is just an *estimate* of the population (true) least-squares regression line.

- Population line: $\mu_y = \beta_0 + \beta_1 x$
- Sample line: $\hat{y} = b_0 + b_1 x$

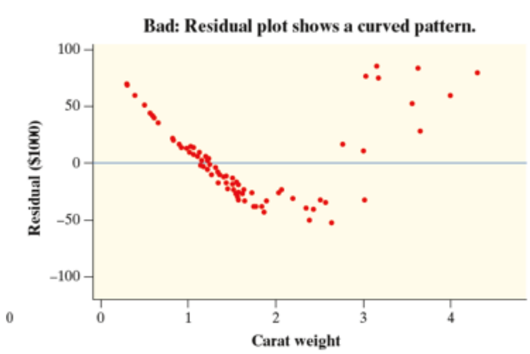
We checked for conditions for inference

Use the LINER acronym!

Linear: There is no leftover curved pattern in the residual plot, indicating that a linear model is appropriate.



Other Example - Linear

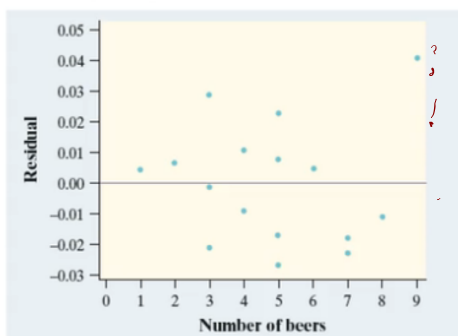


Residuals

BAD

Equal SD

Equal SD: The residual plot shows a similar amount of scatter about the residual = 0 line for each x = number of beers ✓



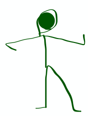
The variability of the residuals in the vertical direction should be **ROUGHLY** the same as you scan across each of the x-values. -**Look for major violations only.**

Big Idea: If the data come from a random sample or randomized experiment, the least-squares regression line is just an *estimate* of the population (true) least-squares regression line.

- Population line: $\mu_y = \beta_0 + \beta_1 x$
- Sample line: $\hat{y} = b_0 + b_1 x$

✓INTERPRET the values of β_0, β_1, σ, and SE_{b_1} in context, and DETERMINE these values from computer output.

helicoptor example from last class



Estimating the Parameters

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

Estimating the Parameters

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

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

4

8

AP® Exam Tip

We use the same notation as the AP® Statistics exam formula sheet for the equation of the sample regression line $\hat{y} = b_0 + b_1 x$.

However, your graphing calculator probably uses the notation $\hat{y} = a + bx$.

Just remember: The slope is always the coefficient of *x*, no matter what symbol is used.

$$b_0 \xrightarrow{\text{Estimates}} \beta_0$$

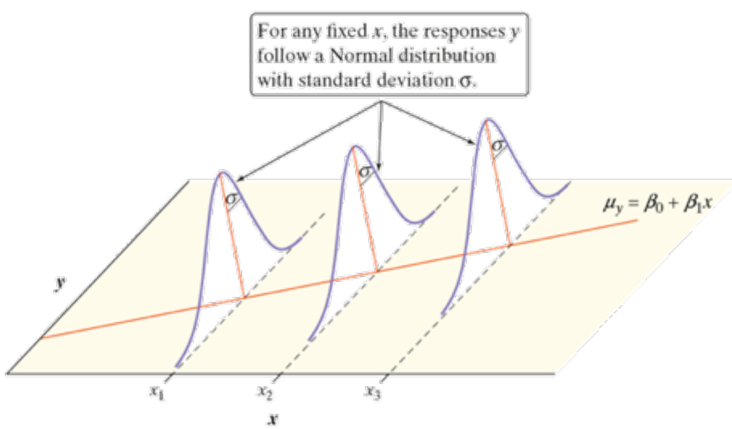
$$b_1 \xrightarrow{\text{Estimates}} \beta_1$$

$$s \xrightarrow{\text{Estimates}} \sigma$$

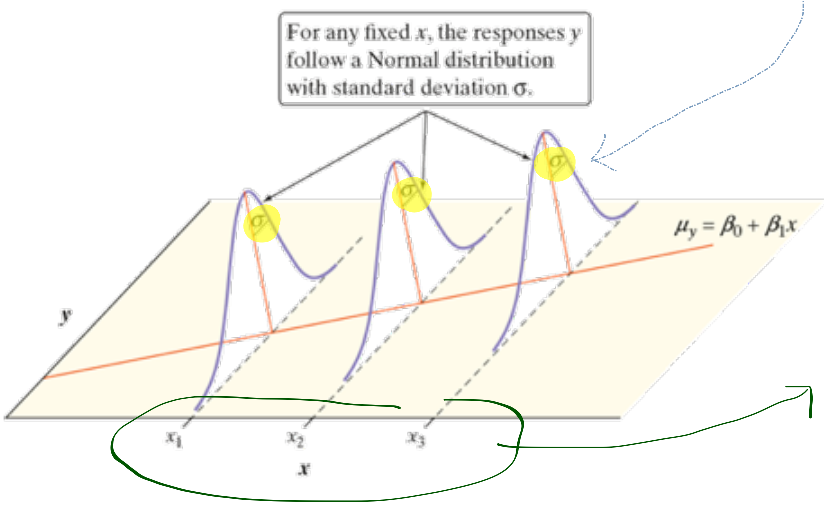
tricky

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**

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



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



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**

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

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**

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

for reasons beyond this course

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

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

We estimate the variability of the sampling distrib. of slope with the Std Error of the slope.

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

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

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$r = \frac{1}{n-1}\Sigma\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$b_1 = r\frac{s_y}{s_x}$$

$$s_{b_1} = \frac{\sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

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

$$s_{b_1} = \frac{\sqrt{\frac{\sum\left(y_i - \hat{y}_i\right)^2}{n-2}}}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}}$$

It will be very unlikely that you will have to use this formula.

(given computer output instead !!!!!)

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

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_1 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. Also, 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 cannot show strong skew or outliers. Make one using the applet and sketch it at right.

(4) **Equal SD:** The variability in the residuals in the vertical direction should be ROUGHLY the same as you scan across most of the x-values. The residual plot does not show a clear sideways Christmas tree patterns for example.

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

Dot Plot of Residuals

3. **There are five conditions to check.**

(1) **Linear:** The **scatterplot** needs to show a linear relationship. Also, 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** cannot show strong skew or outliers. Make one using the applet and sketch it at right.

(4) **Equal SD:** The variability in the residuals in the vertical direction should be ROUGHLY the same as you scan across most of the x-values. The residual plot does not show a clear sideways Christmas tree patterns for example.

(5) **Random:** Either "SRS" or "Random Assignment"
assigned to rows

Dot Plot of Residuals

4. **Construct the interval:**

General Formula:

Specific Formula:

Work:

4. **Construct the interval:**

General Formula: Pt. Estim $\pm$ MOE Specific Formula: $b_1 \pm t^* \cdot SE_{b_1}$

$\uparrow$ df = n−2 = 30−2 = 28

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

$(-4.24, 1.21)$

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

5. **Conclude:**

We are 95% confident that the interval from (−4.24, 1.21) captures the true slope of the population between Row and Score

Confidence Intervals for Slope

Important ideas:

State Formulas Conditions

Confidence Intervals for Slope

Important ideas: State

Formulas

Conditions

Confidence Level

β_1

b_1

Confidence Intervals for Slope

Important ideas: State

Formulas

Conditions

Confidence Level

β_1 true slope of population LSRL

b_1 statistic — sample LSRL slope

point estim ± MOE

$b_1 \pm t^* \cdot SE_{b_1}$

$df = n-2$

1 sample t int for β_1

Confidence Intervals for Slope

Important ideas: State	Formulas	Conditions
Confidence Level	point estim $\pm$ MOE	**L** inear
β_1 true slope of population LSRL	$b_1 \pm t^* \cdot SE_{b_1}$	**I** ndependent
b_1 statistic — sample LSRL slope	$df = n-2$	**N** ormal
	1 sample t int for β_1	**E** qual SD
		R andom

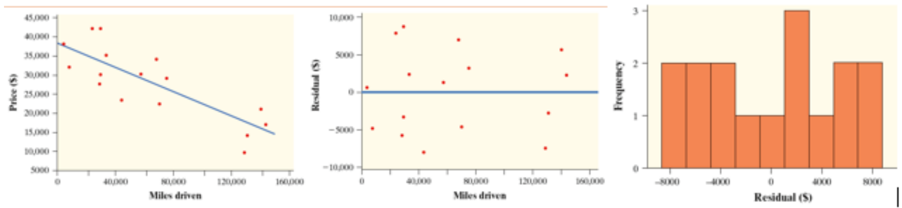
Now... a CI more formally.

Mileage vs Value

-we'll do together

-refer to this example when doing your HW

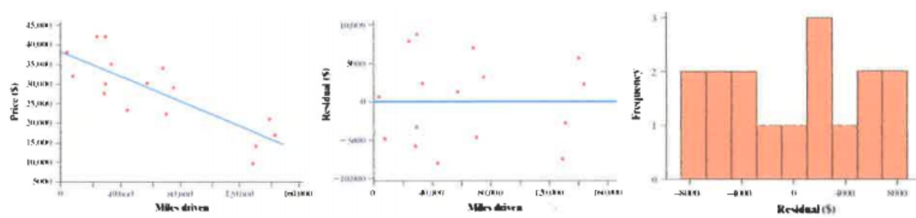
Mileage vs Value-Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 Super Crew 4 x 4 if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 Super Crew 4 x 4s was selected from among those listed for sale 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_1 = true slope of the population regression line relating y = price to x = miles driven for used FORD F-150 4x4's listed for sale on autotrader.com

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

$-.16292 \pm 1.761(.03096)$

$-.16292 \pm 0.05452$

$(-.21744, -.10840)$

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

LinRegTInt

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

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

You can still find b_1 SE_{b_1} by.....

If you use LinRegTTest (see page 785)

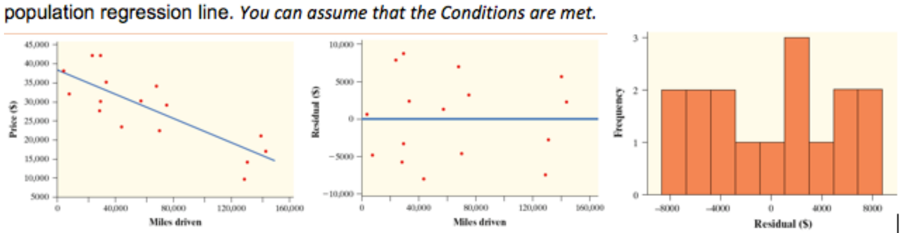
$b =$ slope $\quad SE_{b_1} = \dfrac{b}{t} \quad$ where $t = \dfrac{b-0}{SE_{b_1}}$

12.1....7, 9, 11

and study pp. 776-782

T-shirts

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^*