AP Statistics – Chapter 12 Notes §12.1 Inference for Linear Regression

We can fit a least-squares line to any data relating two quantitative variables, but the results are useful only if the scatterplot shows a linear patter. How can we be sure of this? Could the linear relationship have happened just by chance? Also, can we estimate and test claims about the slope of the population (true) regression line that describes the relationship between two quantitative variables?

Inference for Linear Regression aims to answer these questions.

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*.	
L	_____________: The actual relationship between *x* and *y* is linear. For any fixed value of *x*, the mean response μ_y falls on the population (true) regression line $\mu_y = \alpha + \beta x$. The slope β and intercept α are usually unknown parameters.
	✓ 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.
I	____________________: Individual observations are independent of each other.
	✓Look at how the data were produced. Random sampling and random assignment help ensure the independence of individual observations. If sampling is done without replacement, remember to check that the population is at least 10 times as large as the sample (10% condition) [in time series data, consecutive observations are almost never independent]
N	_____________: For any fixed value of *x*, the response *y* varies according to a Normal distribution
	✓ Make a stemplot, histogram, or Normal probability plot of the residuals and check for clear skewness or other major departures from Normality.
E	______________________________: The standard deviation of *y* (call it σ) is the same for all values of *x*. The common standard deviation σ is usually an unknown parameter.
	✓Look at the scatter of the residuals above and below 'residual = 0' line in the residual plot. The amount of scatter should be roughly the same from the smallest to the largest *x*-value.
R	________________: The data come from a well-designed random sample or random experiment.
	✓See if the data were produced by random sampling or a randomized experiment.

**Don't overreact to minor violations of the conditions. Inference for regression is not very sensitive to lack of Normality, especially when we have many observations. Beware of influential observations!

See Example: The Helicopter Experiment (pg. 743)

Estimating the Parameters:

See Example: The Helicopter Experiment (pg.745)

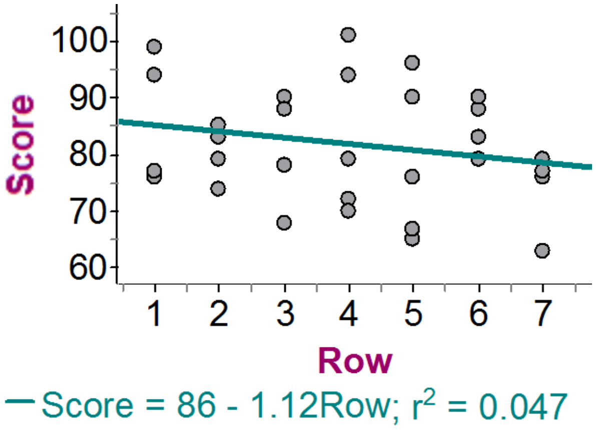
Constructing a Confidence Interval for the Slope: t Interval for the slope of a LSRL

Formula: Standard Error:

See Example: The Helicopter Experiment (pg.747)
Does Fidgeting Keep You Slim (pg. 749)

Example: Does seat location matter?
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 students were assigned (row 1 is closest to the front and row 7 is the farthest away). Here are the results, including a scatterplot and least-squares regression line:

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

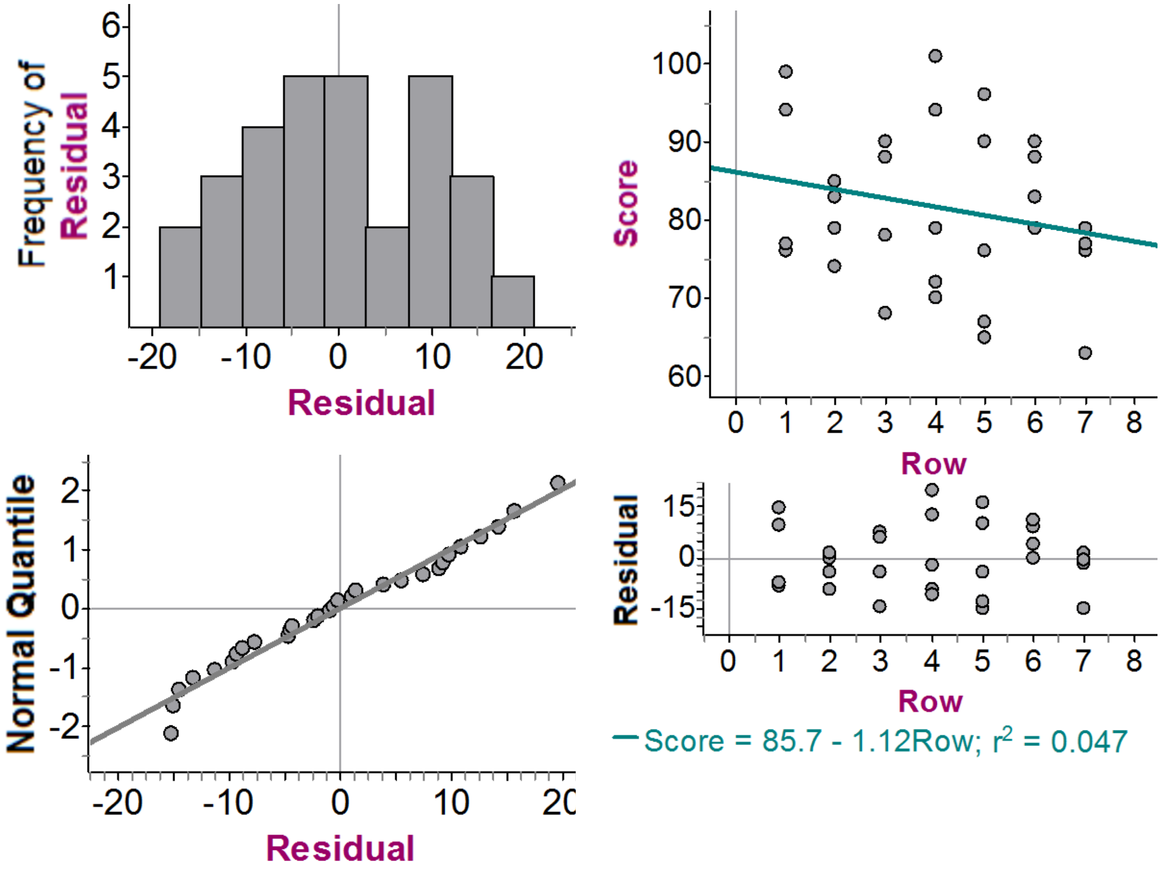


1. Interpret the slope of the least-squares regression line in this context.

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

3. Does the negative slope provide convincing evidence that sitting closer causes higher achievement or is it plausible that the association is due the chance variation in the random assignment? (How would you justify your answer to this question?)

4. We used Fathom to carry out a least-squares regression analysis for the "Does seat location matter?" Activity. A scatterplot, residual plot, histogram and Normal probability plot of the residuals are shown below.



Problem: Check whether the conditions for performing inference about the regression model are met.

Here is computer output for the least-squares regression analysis on the seating chart data.

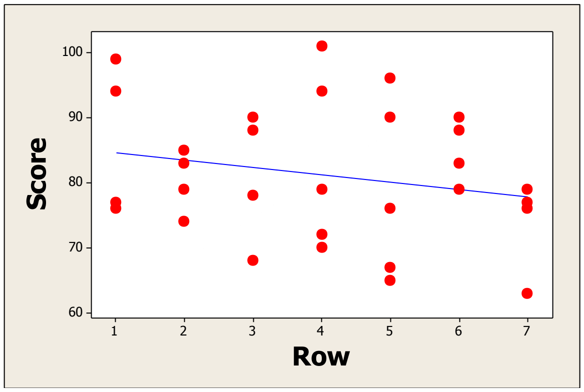
Regression Analysis: Score versus Row

```
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%
```

(a) State equation of the LSRL. Define any variable you use.

(b) Interpret the slope, *y* intercept (if possible), and standard deviation of the residuals.

Earlier, we used Minitab to analyze the results of an experiment designed to see if sitting closer to the front of a classroom causes higher achievement. We checked the conditions for inference earlier. Here is a scatterplot of the data.



Identify the standard error of the slope SE_b from the computer output. Interpret this value in context.

Calculate the 95% confidence interval for the true slope. Show your work.

Interpret the interval from part (b) in context.

(d) Based on your interval, is there convincing evidence that seat location affects scores?

CYU: Pg.750

Performing a Significance Test for the Slope: t test for the slope of the population regression line.

- State the Hypothesis: H_o: β = hypothesized value and H_a : β (<, >, ≠) hypothesized value (to determine whether there is actually a linear relationship between *x* and *y* in the population, let H_o: β = 0)
- Check Conditions (LINER)
- Compute the test statistic: Formula:
- Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with df = n – 2.
- Conclusion: (IN CONTEXT)
  If P-value < α significance level, we reject H_o. We have enough evidence to conclude H_a.
  If P-value > α significance level, we fail to reject $H_{o.}$ We do not have enough evidence to conclude H_a

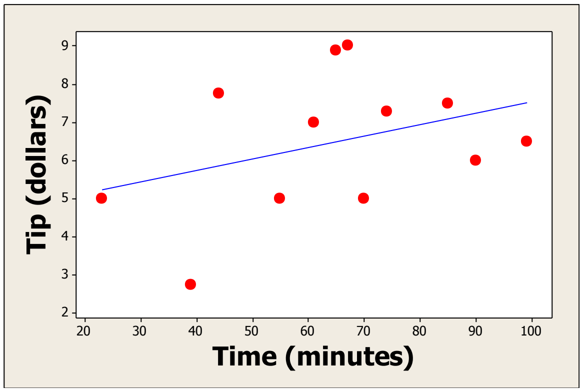
See Example: Crying and IQ. (pg.752)

Example: Tipping at a Buffet
Do customers who stay longer at buffets give larger tips? Charlotte, an AP statistics student who worked at an Asian buffet, decided to investigate this question for her second semester project. While she was doing her job as a hostess, she obtained a random sample of receipts, which included the length of time (in minutes) the party was in the restaurant and the amount of the tip (in dollars). Do these data provide convincing evidence that customers who stay longer give larger tips? Here is the data:

Time (minutes)	Tip (dollars)
23	5.00
39	2.75
44	7.75
55	5.00
61	7.00
65	8.88
67	9.01
70	5.00
74	7.29
85	7.50
90	6.00
99	6.50

Problem:

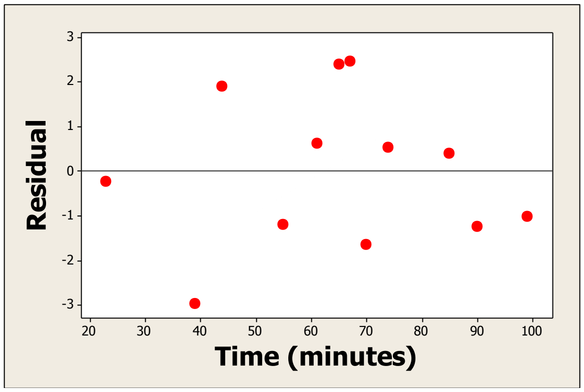
(a) Here is a scatterplot of the data with the least-squares regression line added. Describe what this graph tells you about the relationship between the two variables.

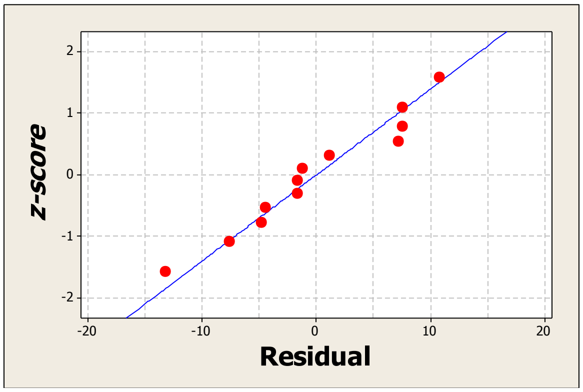


Minitab output from a linear regression analysis on these data is shown below.

Regression Analysis: Tip (dollars) versus Time (minutes)

```
Predictor         Coef  SE Coef    T     P

Constant         4.535    1.657  2.74  0.021

Time (minutes)  0.03013  0.02448  1.23  0.247

S = 1.77931   R-Sq = 13.2%   R-Sq(adj) = 4.5%
```





(b) What is the equation of the least-squares regression line for predicting the amount of the tip from the length of the stay? Define any variables you use.

(c) Interpret the slope and *y* intercept of the least-squares regression line in context.

(d) Carry out an appropriate test to answer Charlotte’s question.

CYU: Pg. 755

AP Statistics – Chapter 12 Notes §12.2 Transforming to Achieve Linearity

When two-variable data show a curved relationship, we could perform simple ‘transformations’ of the data that can straighten a nonlinear pattern. Once the data have been transformed to achieve linearity, we can use least-squares regression to generate a useful model for making predictions. If the conditions for regression inference are met, we can estimate or test a claim about the slope of the population (true) regression line using the transformed data.

See Example: Health and Wealth (pg.766)

Applying a function such as a logarithm or square root to a quantitative variable is called _______________________ the data. This process amounts to changing the scale of measurement that was used when the data was collected.

TRANSFORMING WITH POWERS AND ROOTS:

The power model:

Examples: -distance that an object dropped falls related to time

-time it takes a pendulum to complete on back-and-forth swing

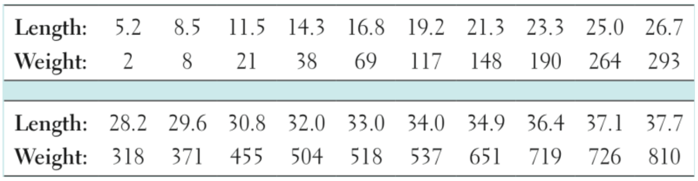
-intensity of a light bulb related to distance from the bulb.

Strategies for transforming the data to achieve linearity:

1. Raise the values of the explanatory variable *x* to the *p* power and plot the points _____________.

2. Take the *p*th root of the values of the response variable *y* and plot the points _____________.

Example: Go Fish



Length:	5.2	8.5	11.5	14.3	16.8	19.2	21.3	23.3	25.0	26.7
Weight:	2	8	21	38	69	117	148	190	264	293
Length:	28.2	29.6	30.8	32.0	33.0	34.0	34.9	36.4	37.1	37.7
Weight:	318	371	455	504	518	537	651	719	726	810

a)Input the data in your calculator. (Decide whether length or weight should be the explanatory variable – use L1 for x)

b)Make a scatterplot of weight vs length.

What do you see?

c) Transform the power model: weight = a $(\text{length})^3$. Then make a scatterplot of the transformation.

d) Make Residual plots of the transformations and describe what you see.

TRANSFORMING WITH LOGARITHMS:

The exponential model: The logarithmic model:

Examples: -Populations of living things tend to grow exponentially
-Money also displays exponential growth when interest is compounded each time period.

**If a variable grows exponentially, its logarithm grows linearly.

See Examples: Money, Money, Money (pg.772)
Moore's Law and Computer Chips (pg.773)

CYU: Pg.776

When we apply the logarithm transformation to the response variable *y* in an exponential model, we produce a linear relationship. To achieve linearity from a power model, we apply logarithm transformation to both variables:

1) A power model has the form _____________, when *a* and *p* are constants.

2) Take the log of both sides of this equation.

3) Look carefully: the power *p* in the power model becomes the __________ of the straight line that links log *y* to log *x*.

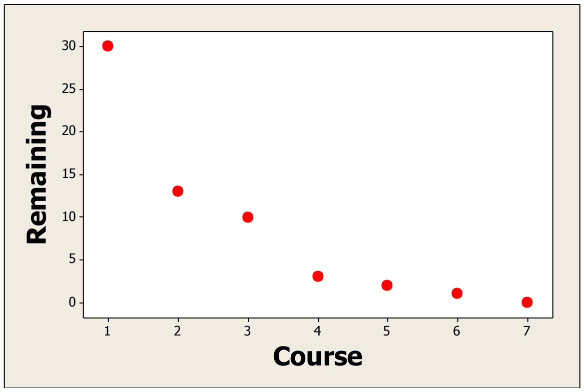
**If a power model describes the relationship between two variables, a scatterplot of the logarithms of both variables should produce a linear pattern. Then we can fit a least-squares regression line to the transformed data and use the linear model to make predictions.

See Example: What's a Planet, Anyway? (pg.778)

Example: More M&M's

A student opened a bag of M&M's, dumped them out, and ate all the ones with the M on top. When he finished, he put the remaining 30 M&M's back in the bag and repeated the same process over and over until all the M&M's were gone. Here is a table and scatterplot showing the number of M&M's remaining at the end of each "course".

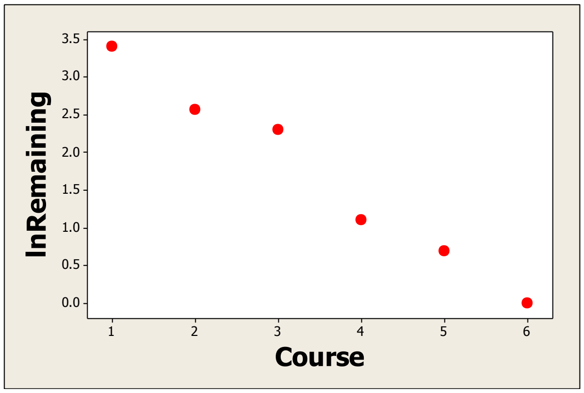
Course	M&M's remaining
1	30
2	13
3	10
4	3
5	2
6	1
7	0



Since the number of M&M's should be cut in half after each course, an exponential model should describe the relationship between the variables.

Problem:

(a) A scatterplot of the natural log of the number of M&M's remaining versus course number is shown below. The last observation in the table is not included since ln(0) is undefined. Explain why it would be reasonable to use an exponential model to describe the relationship between the number of M&M's remaining and the course number.



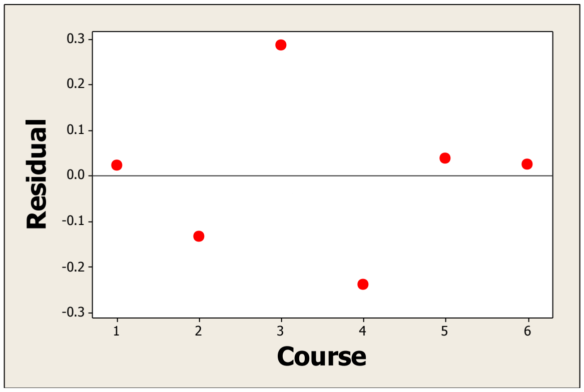
(b) Minitab output from a linear regression analysis on the transformed data is shown below. Give the equation of the least-squares regression line defining any variables you use.

Regression Analysis: LnRemaining versus Course

```
Predictor     Coef  SE Coef      T     P

Constant    4.0593   0.1852   21.92  0.000

Course    -0.68073  0.04755  -14.32  0.000

S = 0.198897   R-Sq = 98.1%   R-Sq(adj) = 97.6%
```

(c) Use your model from part (b) to predict the original number of M&M's in the bag.

(d) A residual plot of the linear regression in part (b) is shown below. Discuss what this graph tells you about the appropriateness of the model.


Residual
Course