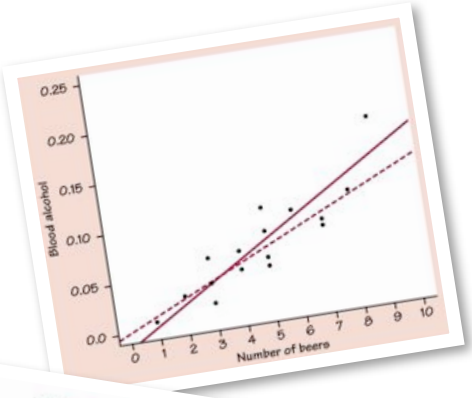
Chapter 14

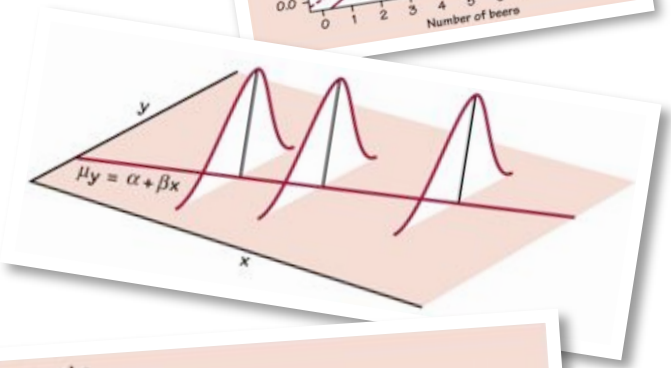
Inference for Regression

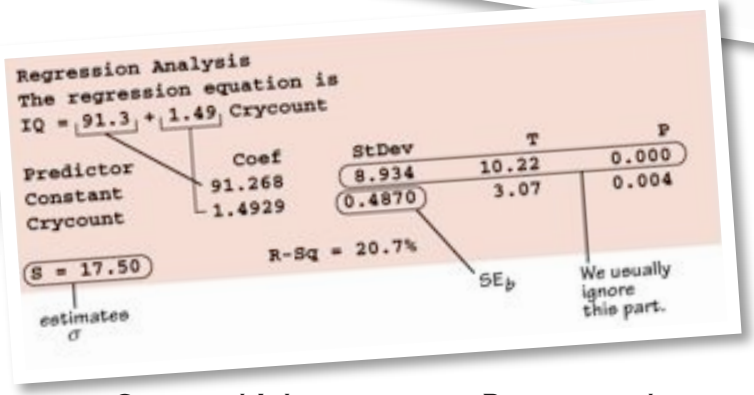
Our final topic of the year involves inference for the regression model. In Chapter 3 we learned how to find the Least Squares Regression Line for a set of bivariate data. In this chapter, we'll learn how to build a confidence interval for the true slope and how to perform a significance test to determine if there is evidence that a linear relationship exists between two variables.



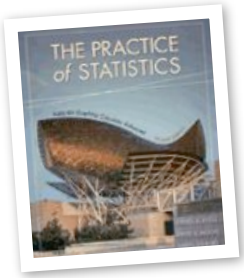
Inference for Regression

- 14.1 Inference about the Model
- 14.1 Testing the Relationship
- Significance Test Practice





AP Statistics Chapter 14: Chi-Square Procedures



"A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions."
~M.J. Moroney

Tentative Lesson Guide					
Date		**Stats**	**Lesson**	**Assignment**	**Done**
Mon	3/26	14.1	Inference for Regression	Rd 780-793 **Do 2-4, 6, 10, 11**	
Tues	3/27	14.1	Practice	14.1 Practice Page	
Wed	3/28	Rev	Inference Review	Inference Review Page	
Thu	3/29	**Rev**	**Review**	Rd 798-800 **Do 18, 19, 23, 24**	
Fri	3/30	**Ex**	**Exam Chapter 14**	***Organize Course Materials***	
Have a Safe, Enjoyable Spring Break!					
4/10 -	4/12	**Rev**	**Final Exam Review**		
Fri	4/13	**MC**	**Multiple Choice Final**	**40Q Multiple Choice Exam**	
Mon	4/16	**FRQ**	**Free Response Final**		
Tue	4/17	**FRQ**	**Free Response Final**	**Mr M's 31st Birthday!**	
4/18 - 5/7			**AP Statistics Exam Review**		
5/8			**AP Statistics Exam**		

Note:
The purpose of this guide is to help you organize your studies for this chapter. The schedule and assignments may change slightly.

Keep your homework organized and refer to this when you turn in your assignments at the end of the chapter.

Class Website:
Be sure to log on to the class website for notes, worksheets, links to our text companion site, etc.

http://web.mac.com/statsmonkey

Don't forget to take your online quiz!. Be sure to enter my email address correctly!
http://bcs.whfreeman.com/yates2e

My email address is:

jmmolesky@isd194.k12.mn.us

Chapter 14 Objectives and Skills:

These are the expectations for this chapter. You should be able to answer these questions and perform these tasks accurately and thoroughly. Although this is not an exhaustive review sheet, it gives a good idea of the "big picture" skills that you should have after completing this chapter. The more thoroughly and accurately you can complete these tasks, the better your preparation.

Conditions for Inference

- ☐ Describe conditions necessary to perform inference about the model.
- ☐ Show inferential conditions are met for regression situations.

Confidence Interval for Slope

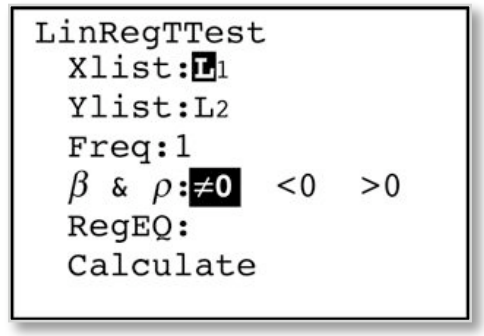
- ☐ Calculate and interpret a Level C confidence interval for the slope of the true regression line.
- ☐ Interpret the slope of the true regression line in the context of the situation.

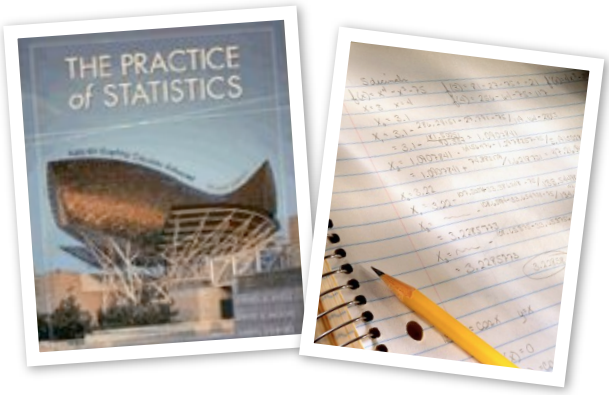
Test the Hypothesis of No Linear Relationship

- ☐ Perform a significance test on the Ho: slope = 0.
- ☐ Interpret computer output regarding a significance test on the slope of the true regression line.

Calculator Skills

- ☐ Enter bivariate data into the List Editor.
- ☐ Construct and interpret a Scatterplot.
- ☐ Calculate the LSRL.
- ☐ Interpret r and r^2.
- ☐ Construct and interpret a Residual Plot.
- ☐ Perform a LinRegTTest.
- ☐ Use LinRegTTest output to determine SEslope.





14.1: Inference about the Model

When a scatterplot shows a linear relationship between an explanatory x and a response y, we can use the LSRL fitted to the data to predict a y for a given x. However, the model fitted is just an estimated model of the *true* relationship between the variables. The slope and intercept are statistics that would take on different values if we sampled different data. We need to use inference to test the model's slope and intercept against their true parameter values.

Consider the following data based on problem 14.1. *Moleskius Primatium* is an extinct beast known to have inhabited the Iron Range region of Minnesota. Suppose 5 fossil specimens are found at a site on Little Sturgeon Lake. Examine the data to determine whether or not femur lengths could be good predictors of humerus lengths for this particular species.

Femur	Humerus	Predicted Humerus	Residual	(Residual)^2	(x Deviation)^2
x	y	y-hat	(y - y-hat)	(y - y-hat)^2	(x - x-bar)^2
38	41				
56	63				
59	70				
64	72				
74	84				
x-bar=		Totals:			

Enter the data in your calculator and make a scatterplot. Does there appear to be a relationship between the femur and humerus lengths? Find the LSRL...construct and interpret a residual plot!

Scatterplot	Residual Plot

LSRL: y-hat = a + bx _____ r=_____ r^2=_____

Interpret the LSRL, r, and r^2 values. (This is a good review of linear regression)

The LSRL we calculated is only an estimate of the true relationship between the variables. If we could measure ALL specimens, we would get another line...

$$\mu_y = \alpha + \beta x$$

We can estimate the unknown slope and intercept from our prediction equation using *a* and *b*. However, we also need to estimate the standard error about the line. This can then be used to create confidence intervals for β and test hypotheses about the linear relationship.

Estimate of α=_____ **Estimate of β =**_____

a and *b* are estimates of the true parameter values of the *true* regression line. If we were to add another observation to our scatterplot or sample different values, we'd probably see a change in *a* and *b*. Like other forms of inference, we can quantify this variability and use it to construct confidence intervals and perform significance tests.

Conditions for Inference about the Model

Before constructing a confidence interval or performing a significance test on the slope, we must check the following conditions:

- ☑ Observed ordered pairs (response values) are independent of each other.
- ☑ The true relationship is linear. (Is the scatterplot roughly linear?)
- ☑ The standard deviation of the response is constant. (Is the scatter about the LSRL consistent?)
- ☑ The response varies Normally about the true regression line. (Are the residuals approximately normally distributed?)

Standard Error:

Every x value has a y value associated with it. However, if we were to find another specimen with the same x value, there is a good chance it would have a different y value. Thankfully, we know these y values will vary according to a normal distribution. The standard deviation of this normal distribution determines how close observed points will fall to the true regression line. The smaller the standard deviation is, the closer they will fall to the line...the bigger it is, the more scattered they will be.

Residuals estimate how much the y values vary about the regression line. We can use the residuals to estimate the standard error...

Standard Error about the LSRL: $$s = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}}$$

Use this formula to calculate the standard error for our example.

Estimate of standard deviation: Standard Error s=_____

Confidence Intervals for the True Slope β:

The slope of the true regression line is the most important parameter in a regression problem. The slope is the rate of change of the response variable as the explanatory variable changes. Since *b* in the LSRL is only an estimate of β, we need to use a confidence interval to determine what it could be.

A level C confidence interval for β is given by

$$b \pm t^* SE_b$$

The standard error of the least-squares slope *b* is:

$$SE_b = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$$

and t^* is determined using (n-2) degrees of freedom.

You will rarely have to calculate the SE by hand. We'll learn how to use computer output or our calculator to provide that for us...

Find and interpret a 95% confidence interval for β based on our (humerus, femur) sample.

14.1: Testing the Relationship

The most common hypothesis about the slope is Ho: β=0. A regression line with slope=0 is horizontal. That is, y does not change when x changes, implying there is no linear relationship between x and y. Put another way, Ho says there is no linear relationship between x an y OR there is no correlation between x and y.

Significance Tests for the True Slope β:

We can use the Standard Error of the slope to perform a significance test to determine whether or not there is evidence to suggest the true slope is greater than, less than, or not equal to 0. We can calculate a t statistic with (n-2) degrees of freedom to determine if our *b* is significantly different than 0, suggesting a true slope other than 0 in the population. The structure and logic of the test is identical to those we have performed in Chapters 11-13.

Use the (humerus, femur) data to determine whether or not there is significant evidence to suggest a positive relationship between the humerus and femur lengths of *Moleskius Primatium*.

Hypotheses:

Conditions:

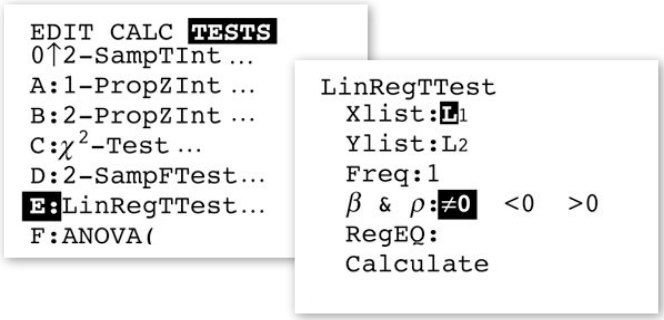
Sampling Distribution of *b* and Test Statistic:

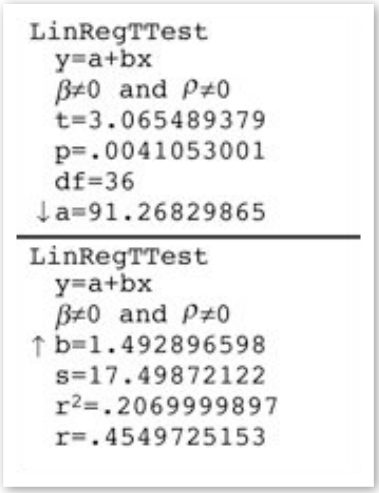
Conclusion:

Interpreting Calculator and Computer Output

The calculation of SEb can be tedious. When performing inference for regression, you should focus on the interpretation of statistical calculations rather than the calculations themselves. Be sure to re-familiarize yourself with your calculator's ability to construct Scatterplots, calculate LSRLs, and construct Residual Plots and Normal Quantile Plots. All are essential when performing inference on regression models.

To perform a significance test on regression, you must first enter your data into L1 and L2. Then, choose **STAT: TESTS: E:LinRegTTest...**





Choose the appropriate alternative hypothesis, leave RegEQ blank, and select Calculate.
The results are displayed over two screens. The first gives the t statistic and p-value for the observed slope. Scrolling down will display the statistic values, a and b, for the LSRL y-hat=a+bx that fits the observed data. Further, r and r^2 are presented, along with the standard error about the line, *s*.

How can you use *s* to calculate SE_b?

SE_b=_____

Minitab Regression Output

You should also familiarize yourself with common computer output for regression analyses. Consider the following Minitab output...identify the important values for inference regarding the relationship between beer consumption and blood alcohol content (from a study of 16 Ohio State students).

Predictor	Coef	StDev	T	P
Constant	-0.01270	0.01264	-1.00	0.332
Beers	0.017964	0.002402	7.48	0.000

S = 0.02044 R-Sq = 80.0%

LSRL=_____ SE_b=_____ r=_____ r^2=_____

14.1: Inference for the Model Practice

1. The article "The Risk of Teen Mothers Having Low Birth Weight Babies: Implications of Recent Medical Research for School Health Personnel" noted that adolescent females are much more likely to deliver low-birth weight babies than are adult females. Use the following data to support or refute this claim. Then, use a confidence interval to summarize the relationship between mother's age and birth weight.

Mother's Age	15	17	18	15	16	19	17	16	18	19
Birth Weight (g)	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

2. The article "Root Dentine Transparency: Age Determination of Human Teeth Using Computerized Densitometric Analysis" described a study in which the objective was to predict age from percentage of a tooth's root with transparent dentine. Use the following MINITAB output to decide whether a LSRL is useful in predicting age:

```
Predictor       Coef      Stdev     t-ratio   p
Constant        32.08     13.32     2.41      0.043
Percent         0.5549    0.3101    1.79      0.111
s = 14.30       R-sq=28.6%     R-sq(adj)=19.7%     n=10
```

What is the LSRL equation? Would the study's data provide significant evidence of a linear relationship between percent and age? Construct and interpret a 95% CI for the slope of the relationship.

3. The article "Effect of Temperature on the pH of Skim Milk" reported on a study involving the temperature (°C) and pH of milk under experimental conditions. Do the following data strongly suggest a negative linear relationship between temp and pH? If so, estimate the slope.

Temperature	4	4	24	24	25	38	40
pH	6.85	6.79	6.63	6.65	6.72	6.62	6.52
Temperature	45	50	55	60	67	70	78
pH	6.5	6.48	6.42	6.38	6.34	6.32	6.34

4. The following MINITAB output summarizes data on x=treadmill run time to exhaustion and y= 20km ski time for a sample of 11 biathletes. Use this output to describe the relationship between run time and 20km ski time. Be complete…estimate the true relationship.

```
Predictor       Coef      Stdev     t-ratio   p
Constant        88.796    5.750     15.44     0.000
tread           -2.3335   0.5911    -3.95     0.003
s = 2.188       R-sq=63.4%     R-sq(adj)=59.3%     n=11
```

Inference Review Practice

The following problems cover inferential topics from Chapters 11, 12, 13, and 14. Use the methods learned this semester to answer the following questions regarding means, proportions, categorical relationships, and regression models.

1. Fruit flies show a daily cycle of rest and activity. Researchers, wondering if fruit flies respond differently when resting, used a sensor to determine if the flies moved in response to vibration under each state (walking and resting). Of the 64 walking flies, 54 responded to vibration. Of the 32 flies who were resting, 4 responded to vibration. Is there significant evidence to suggest the flies respond differently to vibration in the different states of rest and walking?

2. UNC studied student performance in a course required by chemical engineering majors. The question of interest was whether or not there was a relationship between time spent in extracurricular activities and academic performance in the course. Is there significant evidence to suggest a relationship between hours spend in extra curricular activities and grade?

	Extracurricular Hours per Week		
Grade	**<2**	**2 to 12**	**>12**
C or Better	11	68	3
D or F	9	23	5

3. The 3-point line was installed in college basketball in 1986. Since then, the number of 3-pointers attempted per game has increased. However, data suggests the percent made has actually *decreased*. Use the following minitab output to construct and interpret a 95% confidence interval for the slope of the true regression line that predicts % made from number attempted. {Note: the scatterplot suggests a linear relationship and the residuals are approximately normally distributed.}

```
Predictor    Coef         Stdev       t-ratio     p
Constant     42.8477      5.750       77.40       0.000
Taken        -0.47620     0.5911      -13.70      0.000
s = 0.4224          R-sq= 91.7%       R-sq(adj)=91.2%    n=19
```

4. Do female mice have more endurance than male mice? The following data are from an experiment that measured how long mice could spend on a physical task before exhaustion. Is there evidence to suggest females have significantly higher endurance?

Group	**n**	**Mean**	**Std Dev**
Female	162	1.4	26.09
Male	135	6.7	6.69

5. In a restaurant worker survey, 68 of 100 randomly selected employees indicated work stress had a negative impact on their personal lives. Estimate the true proportion of restaurant employees who feel this way with 98% confidence.

6. In a recent experiment to determine whether or not pleasant odors have an effect on performance, 21 subjects were timed as they completed a series of mazes wearing an unscented mask, then re-timed while wearing a scented mask. The differences (unscented-scented) were calculated and were found to have an average of 0.96 sec with a standard deviation of 12.55 sec. Construct a 95% Confidence Interval for the true mean difference and use it to make a determination about pleasant odors and performance.