AP Statistics
Unit 08 – Day 03 Notes
Regression Inference

Name Key 2018
Period____________

Sample Regression Line: the least squares regression line $\hat{y} = a + bx$ computed from the sample data.

Population (True) Regression Line: the LSRL $\mu_y = \alpha + \beta x$ based on the entire population of data.

Sampling distribution of slope b:

Choose an SRS of n observations (x,y) from a population of size N with least-squares regression line:

predicted y = $\alpha + \beta$x

Let b be the slope of the sample regression line. Then:

- The **mean** of the sampling distribution of b is $\mu_b = \beta$
- The **standard deviation** of the sampling distribution of b is

$$\boxed{\sigma_b = \frac{\sigma}{\sigma_x\sqrt{n}}}$$

as long as the 10% condition is satisfied: n ≤ (1/10)N

- The sampling distribution of b will be **approximately Normal** if the values of the response variable y follow a Normal distribution for each value of the explanatory variable x (the Normal condition)

Conditions for 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. We can check many of these conditions by looking at graphs of the residuals. Here's how to check each condition:

LINEAR: Examine the scatterplot to see if the overall pattern if roughly linear (no curve & no pattern in the residual plot).

INDEPENDENT: Look at how the data were produced (random samples or well-designed randomized experiments are the best since they of ensure independence of individual observations). If sampling without replacement, you must check the 10% condition. Also, avoid using time-series data (paired data). Example: measuring a person's height at age 4 and again at age 10. These data points are not independent because they come from the same individual and knowing something about the age 4 height may tell you something about age 10 height for that same individual.

NORMAL: Make a plot of the residuals. Check for skewness, outliers, or other major departures from Normality.

EQUAL SD: Look at the scatter of the residuals above and below the "residual = 0" line in the residual plot. The vertical spread of the residuals should be roughly the same form the smallest to the largest x-value.

RANDOM: See if the data came from a well-designed random sample or randomized experiment. If not, we can't make inferences about a larger population or about cause and effect.

STATE:

Confidence Interval:

We want to estimate the slope β of the population regression line relating ____________ to ______________ with C% confidence.

OR

Significance Test:

We want to perform a test of

$H_o: \beta = 0$
$H_A: \beta < 0$ (negative slope)
$H_A: \beta > 0$ (positive slope)
$H_A: \beta \neq 0$ (non-zero slope)

pick one based on the question

where β is the slope of the population regression line relating _________ to __________.

$\alpha =$_________

PLAN:

USE THE ACRONYMN **LINER**

- **Linear:** 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 $\mu_y = \alpha + \beta x$.
- **Independent:** Individual observations are independent of each other. When sampling without replacement, and also check the **10% condition**.
- **Normal:** For any fixed value of x, the response y varies according to a Normal distribution.
- **Equal Standard Deviation:** The standard deviation of y (call it σ) is the same for all values of x.
- **Random:** The data from from a well-designed random sample or randomized experiment.

Because the conditions are met, we will use **a t-interval for the slope of a regression line β** or we will use a **t-test for the slope of a regression line β.**

DO:

Confidence Interval (t-interval for slope):

Plug values into calculator & include: test name, df, t*, and the interval.

OR use the equation below:

$$\boxed{b \pm t^* \frac{s}{s_x\sqrt{n-1}}}$$

SE

use invT to find

with df = n – 2

with $\boxed{SE_b = \frac{s}{s_x\sqrt{n-1}}}$

If you are given a minitab output, you must use the equation!

Be sure to write down:

df =

Confidence interval (__________ , ___________)

OR

Significance Test (t-test for the slope):

Plug values into calculator & include: test name, df, test statistic, p-value, and graph.

OR use the equation below:

=0

$$\boxed{t = \frac{b - \beta_0}{\frac{s}{s_x\sqrt{n-1}}}}$$

SE

with df = n – 2

with $\boxed{SE_b = \frac{s}{s_x\sqrt{n-1}}}$

If you are given a minitab output, you must use the equation!

Be sure to write down:

df =

test statistic =

P-value =

CONCLUDE:

We are C% confident that the interval between ____________ and _____________ captures the slope of the population regression line relating __________ to __________ [in context].

OR

We reject or fail to reject the null because our P-value is __________ </> α. There is/isn't convincing evidence of a positive linear relationship between ___________ and _____________ [in context].

or negative, or non-zero

Technology:

Confidence Interval for slope of a line:

STAT → TESTS
G: LinRegTInt

Option G

Significance Test for slope of a line:

STAT→ TESTS
F: LinRegTTest

Option F

Example 1:

Do Beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, at random in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them. If so, more stumps should produce more beetle larvae. Minitab output for a regression analysis on these data is shown blow. Construct and interpret a 99% confidence interval for the slope of the population regression line. Assume that the conditions for performing inference are met.

Regression Analysis: Beetle larvae versus Stumps

Predictor	Coef	SE Coef	T	P
Constant	-1.286	2.853	-045	0.657
Stumps	11.894	1.136	10.47	0.000
S = 6.41939	R-Sq = 83.9%	R-Sq(adj) = 83.1%		

y-intercept (a) slope (b) SE (to use in equations)

State: We want to estimate the slope β of the population regression line relating number of beetle larvae clusters to number of cottonwood tree stumps cut by beavers with 99% confidence.

Plan: Because our conditions are met, we will construct a t-interval to estimate the slope β of the population regression line.

DO: $df = 23-2 = 21$ $\quad 11.894 \pm 2.831(1.136) = (8.678, 15.110)$

$t^* = 2.831$

Conclude: We are 99% confident that the interval from 8.678 to 15.110 (larvae clusters per stump) captures the slope β of the population regression line relating number of beetle larvae clusters to number of cottonwood tree stumps cut by beavers.

Example 2:

A researcher from the University of California, San Diego, collected data on average per capita wine consumption and heart disease death rate in a random sample of 19 countries for which data were available. The following table displays the data:

Alcohol from Wine (liters/year)	Heart Disease Death rate (per 100,000)	Alcohol from Wine (liters/year)	Heart Disease Death rate (per 100,000)
2.5	211	7.9	107
3.9	167	1.8	167
2.9	131	1.9	266
2.4	191	0.8	227
2.9	220	6.5	86
0.8	297	1.6	207
9.1	71	5.8	115
2.7	172	1.3	285
0.8	211	1.2	199
0.7	300		

Is there statistically significant evidence of a negative linear relationship between wine consumption and heart disease deaths in the population of countries? Carry out an appropriate significance test at the $\alpha = 0.05$ level.

State: $H_0: \beta = 0$ where β is the slope of the population regression line relating heart disease deathrate to wine consumption in the population of countries.
$H_A: \beta < 0$
$\alpha = 0.05$

Plan: Linear: the relationship between x and y is linear. There is no pattern in the residual plot.

Independent: country observations are independent (knowing one tells us nothing about another)
10% condition: 190 < all countries

Normal: a histogram of the residuals shows no skewness/outliers and is single-peaked

Equal SD: the residual plot shows a similar amount of scatter about the residual = 0 line, with a few points slightly closer to the line than others.

Random: this was a random sample of 19 countries.

Because our conditions are met, we will perform a t-test for the slope of β.

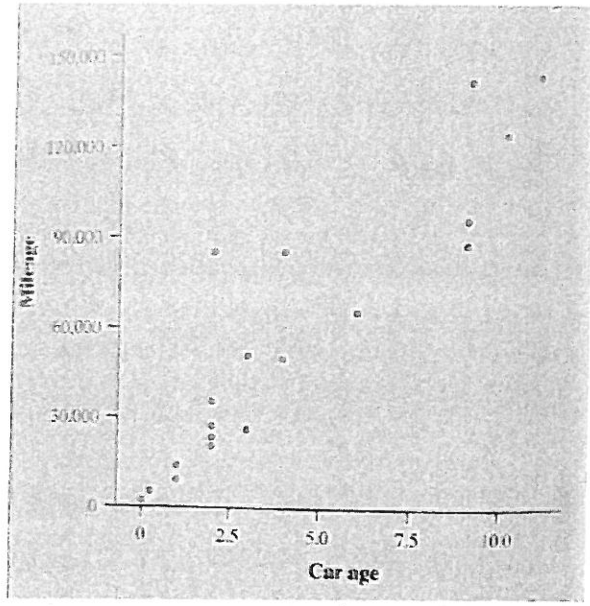
DO: df = 17
test statistic: −6.457
p-value: 0.000002957

Conclude: Because our p-value = 0.000002957 is less than our significance level $\alpha = 0.05$, we reject the null. There is convincing evidence that there is a negative relationship between heart disease and wine consumption in the population of countries

Example 3:

A random sample of AP Statistics teachers was asked to report the age (in years) and mileage of their primary vehicles. A scatterplot of the data is shown. A computer output from a least-squares regression analysis of these data is shown below (df=19). Assume that the conditions for regression inference are met.

Variable	Coef	SE Coef	T	P
Constant	7288.54	6591	1.11	0.2826
Car Age	11630.6	1249	9.31	<0.0001
S=19280	R-sq=82.0%		R-sq(Adj)=81.1%	



a) Verify that the 95% confidence interval for the slope of the population regression line is (9016.4, 14,244.8).

$11630.6 \pm 2.093(1249)$

$=(9016.4, 14244.8)$ ✓

b) A national automotive group claims that the typical driver puts 15,000 miles per year on his or her main vehicle. We want to test whether AP Statistics teachers are typical drivers. Explain why an appropriate pair of hypotheses for this test is $H_0: \beta = 15{,}000$ versus $H_A: \beta \neq 15{,}000$.

Because the automotive group claims that people drive 15000 miles per year. This says that for every 1 year, the mileage would increase by 15000 miles.

↑ just explaining slope

c) Compute the test statistic and P-value for the test in part (b). What conclusion would you draw at the $\alpha = 0.05$ significance level?

test statistic = 9.31

df = 19

p-value < 0.0001

Because our p-value < 0.0001 is less than our significance level $\alpha = 0.05$, we reject the null hypothesis. There is convincing evidence that the slope of the true regression line relating mileage to age in years differs from 15000.

d) Does the confidence interval in part (a) lead to the same conclusion as the test in part (c)? Explain.

15000 does not fit in the interval for part a which means it also shows evidence that the slope of the population regression line relating mileage to age of cars in years is not 15000.