AP STAT

Ch 15- Inference for Linear Regression

hitting the main points

Conditions for Regression Inference:

- **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 $\mu_y = \alpha + \beta x$. The **slope β** and **intercept α** are usually unknown.
- **Independent**- individual observations are independent of each other
- **Normal**- for any fixed x, y varies according to a normal distribution
- **Equal Variance**- The standard deviation of y is the same for all values of x- the common standard deviation is usally unknown parameter
- **Random**- The data comes from a well designed randomized experiment

HOW TO ACTUALLY CHECK THE CONDITIONS:

- **Linear**- examine the scatterplot that the overall pattern is roughly linear. Check resituals centering at zero
- **Independent**- Look at how the data was produced. Random samplings and random assignements help endure independence
- **Normal**- make a stem plot, histogram, box plot or normal probability plot of residuals and check for skewness or otehr major departures from normal
- **Equal Variance**- Look at the scatter of the residuals above and below the residual = 0 line in the residual plot. The amount of scatter should be roughly the same.
- **Random**- random sampling/assignments

The Ho and Ha

First recall:

y= a + bx (least squares regression)

now because we are generalizing beyond the sample to the populations, the notation becomes: a = α and b = β $y = \alpha + \beta x$

y replaced w/ dep.
x replaced w/ indep. var

The Ho/Ha will be β focused (slope focused)

Ho: $\beta = 0$

Ha: $\beta \neq 0$ or < or >

Mini Tab vs Calculator output

Here is some data to put in L1 and L2

Infants who cry easily may be more eaily stimulated than others. This may be a sign of higher IQ. child development researchers explore the relationship between crying infants 4 to 10 days old and their later IQ scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured the intesity byt he number of peaks in the most active 20 seconds. They later recorded the childs IQ at age 3 using a known IQ test.The table contains the data from 38 infants. (Meaning does the crying intensity determine/ predict IQ- we will need a predition equation)

L1 Crying	L2 IQ	Crying	IQ	Crying	IQ	Crying	IQ
10	87	20	90	17	94	12	94
12	97	16	100	19	103	12	103
9	103	23	103	13	104	14	106
16	106	27	108	18	109	10	109
18	109	15	112	18	112	23	113
15	114	21	114	16	118	9	119
12	119	12	120	19	120	16	124
20	132	15	133	22	135	31	135
16	136	17	141	30	155	22	127
33	159	13	162				

1. Make a scatter plot of the data- draw sketch

r = .447
r² = 20.0%

$r = y - \hat{y} \rightarrow L2 - L3$

2. Find the regression equation (#8) be sure to define the variables in the equation.

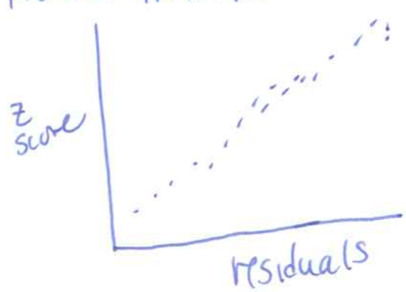
$y = 92.34 + 1.386x$

IQ = 92.34 + 1.386(crying intensity)

3. Check condition for doing a test for inference.

Residual Plot

Normal Prob Plot



Linearity - shows some but low r val
- no extreme outliers
- appears to be an association w/ cries & IQ

independent →

Normal → normal prob plot
Linear → normal

Residual Plot shows equal variance

Random

* Do the Data provide convincing evidence that theres a positive Linear relationship Between crying intensity in Infants and IQ

Ho + Ha are slope related

d) Write Ho and Ha for this.

Ho: $\beta = 0$ (diff = 0

Ha: $\beta > 0$ (difference >0)

e) go to TESTS>>LinRegTTest and record results Use $\alpha = .05$

linregTtest

t= 3.00

p = .002

df = 36

Reject Ho ; support Ha

f) Compare your output to the MINITAB outputs below- notice where you find all values for the test and equations

Predictor	Coeff	SECoeff	T	P
Constant	91.268	8.934	10.22	0.000
CryCount	1.4929	0.487	3.07	0.004
S=17.5	R-SQ=20.7%		R-SQ(adj)= 18.5%	

α intercept

β slope

* AP output

Split in half B/C thats for two sided test

g) Conclusion about infant crying and IQ?

h) Construct a confidence interval (using calc) for the LinReg at 90%. This will be generalizing the slope (beta)

FORMULA:

b± t*(SE_b)

df = n-2 *

1.4929 ± 1.687(.487)

SE St.Error

(.6713, 2.315)

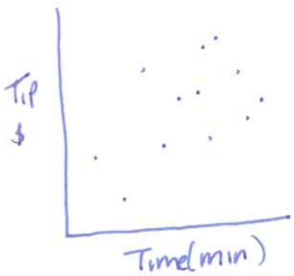
90% confident the true pop slope is Btwn .6713 and 2.315

IQ increases roughly ~~.6713~~ between .6713 and 2.3149 points per each additional peak of crying

Another Example:

Do customers who stay longer at a buffer give larger tips? Charlotte, an AP stat student who worked at an Asian Buffet, decided to investigate this question for a second semester project. While doing her job as the hostess, she ontained a random sample of receipts which included the lenght of time (in min) the party was in the restaurant and the amount of the tip (in dollars). Do the data provide convincing evidence that customers who stay longer give larger tips.

Scatter Plot

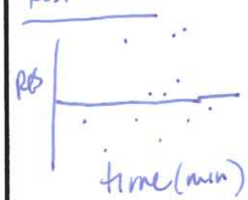


$r = .363$

Time(min)	Tip (dollars)
23	5
39	2.75
44	7.75
55	5
61	7
65	8.88
67	9.01
70	5
74	7.29
85	7.5
90	6
99	6.5

A) Produce a scatterplot and check conditions.

Resid Plot



Normal PP

- Linear → some though weak assoc btween length of stay and tip amt.
- indep. → yes (10% rule)
- normal → npp appears strong Linear.
- Equal Variance → Resid plot appears even space
- Random → states random

B) What is the equation for the least squares regression line prediciting the amount of tip from length of stay. Define variables!

$\hat{y} = 4.535 + 0.030x$ OR predicted tip $= 4.535 + .030$ time $\quad y$ tip $\quad x$ time $\quad$ tip / min

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

tip will increase .030 as time increases by one min

D) Carry out an appropriate test to answer Charlottes question. Use a 0.05 significance level.

$H_0: \beta = 0$ use $\alpha = .05$

$H_a: \beta > 0$

$\beta = 0$ → no impact - no ~~real~~ relationship
$\beta > 0$ → there is a pos. relationship

LinTTest → $t = 1.23$, $p = .1235$

Since $p = .1235 > \alpha (.05)$ we Fail to Reject the H_0 We do not have convincing evidence to say time will increase tip

E) Write a 95% confidence interval (and interpret) for the slope

Using MINITAB output below:

FORMULA:

b± t*(SE_b)

df = n − 2

Regression Analysis

Tips Vs Time

Predictor	Coef	SE Coef	T	P
Constant	4.535	1.657	2.74	0.021
Time	0.03013	0.02448	1.23	0.247
S=1.77931	R-Sq= 13.2%	R-sq(adj)= 4.5%		

Split in half b/c thats for a 2 sided test

$b \pm T(SE_b)$

n = 12
df = 10

$.03013 \pm \underbrace{2.23(.02448)}$

$.03013 \pm .0546$

$(-.02447, .08473)$

With 95% confidence the mean tip increase over time will Be between -.02447 and .08473.

CHECK for UNDERSTANDING

Is Wine good for your heart

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 was available. The data is displayed below: (alcohol is liters per year)

L1 Alcohol x	L2 HD Death Rate y	L1 Alcohol	L2 HD Death rate
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		

A) 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 an alpha= 0.05.

I will conduct a LinRegTTest to determine if there is a negative linear relationship between wine consumption + heart disease deaths.

Conditions ($\hat{y} = 260.56 - 22.969x$)

- Linearity: r=.71. The scatterplot shows some neg. linear relationship given data
- indep → data was taken independently (each country drinking indep.)
- random → says random sample
- normal → normal prob plot shows somewhat normal (linear)
- Equal variance → Equal scatter above and below the zero line

LinReg T Test

$T = -6.45$

$p = 2.96 \times 10^{-6} \approx 0$

$df = 17$

$H_0: \beta = 0$

$H_a: \beta < 0$

with small pvalue we can Reject Ho + support the alt. that there is a neg linear relationship between wine consumption + heart Des. deaths (more wine, less death)

B) Calculate and interpret a 95% confidence interval for the slope β of the population regression line.

$b \pm t^* SE_b$ $b = -22.97$ $s = 37.87$

$-22.97 \pm 2.11(3.57)$

$SE_b = \dfrac{s}{s_x\sqrt{n-1}}$ $\dfrac{37.87}{(2.5)\sqrt{18}}$

-22.97 ± 5.422

$SE_b = 3.57$ 2var stats S_x

$(-28.39, -17.548)$

With 95% confidence the true slope for population regression line will be captured Between -28.39 and -17.548