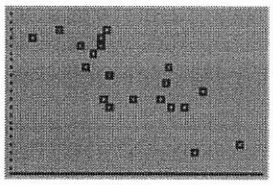
AP Statistics – Lesson Notes - Chapter 27: Inferences for Regression

Remember Linear Regression?

Data shows a possible relationship between the weight of a car and fuel efficiency:

Weight (lbs)	Fuel Eff. (mpg)
2,700	25
2,305	33
2,300	34
2,485	26
2,260	32
2,345	29
2,325	26
2,340	35
2,675	28
1,900	34
2,355	25
2,055	35
3,110	20
2,885	27
2,850	19
2,695	30
2,175	33
2,215	30
2,790	25
2,640	26



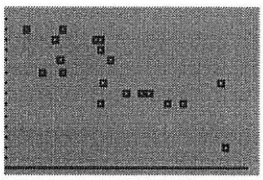
$\widehat{mpg} = 58.364 - .012(weight)$

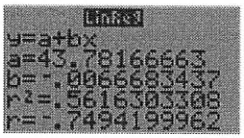
But these cars are just a sample of the population of all cars. If we had a different sample...

Weight (lbs)	Fuel Eff. (mpg)
2,485	28
2,885	27
2,695	28
2,680	25
2,655	27
3,065	22
2,970	25
2,690	24
2,910	26
2,975	25
2,920	21
2,575	24
2,935	23
2,920	27
3,640	17
3,295	21
3,380	21
3,145	22
3,200	22
3,610	23

...we would get a different LSRL equation:

$\widehat{mpg} = 43.782 - .007(weight)$



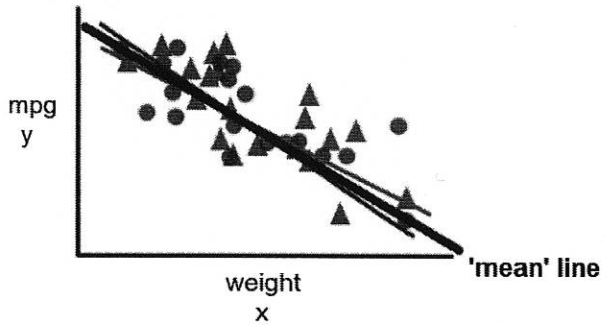


The population would have a 'mean' idealized regression line:

Each sample would produce its own LSRL with different ***statistics:*** $slope: a$

$intercept: b$

$\hat{y} = a + bx$

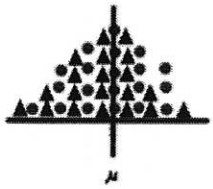


But a model for the entire population would be specified by ***parameters:*** $slope: \alpha$

$intercept: \beta$

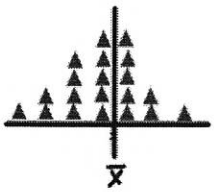
This model is called a 'mean' or idealized regression line: $\mu_y = \alpha + \beta x$

Inference for LSRL Regression



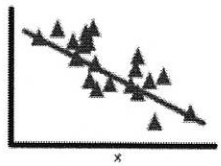
Out of a population (which has parameters)...

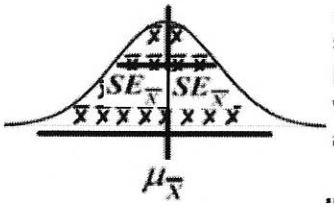
$\mu_y = \alpha + \beta x$



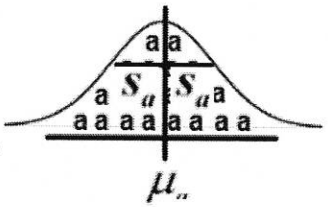
...we take a sample (which has statistics)...

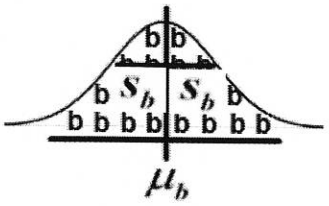
$\hat{y} = a + bx$





...if we took many samples, each would have its own statistics which vary according to a sampling distribution...





...we find a t-value (test statistics, number of standard devs away from expected) to determine p-value.

$$t = \frac{x - \mu_0}{SE_{\bar{x}}}$$

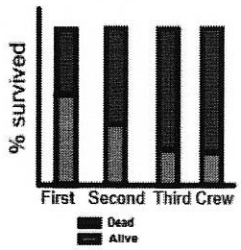
$$t = \frac{b - \beta_0}{s_b}$$

We use <u>slope</u> for LSRL inference

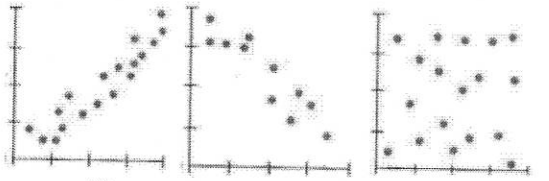
Why does LSRL Regression Inference Use Slope?

We've seen the idea of 'association' in three different parts of the course so far:

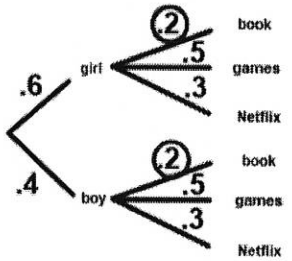
Unit 1



Unit 2

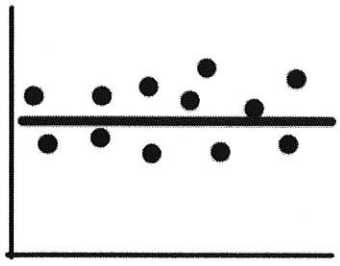


Unit 4



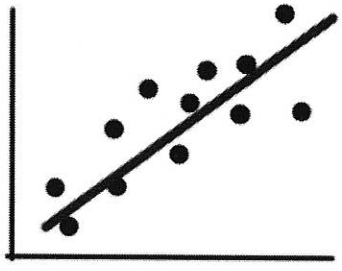
If $P(B) = P(B|A) = P(B|\bar{A})$, then A and B are independent.

All are different ways of picturing the same idea: if there is an association between two things (they are not independent) then varying one thing is connected with a change in the other thing.



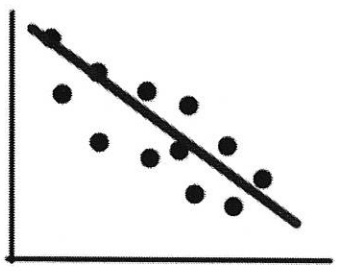
No association

$\mu_y = \alpha \quad \beta = 0$



Is an association (positive)

$\mu_y = \alpha + \beta x \quad \beta > 0$



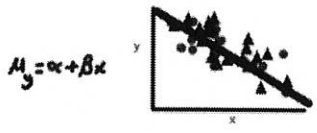
Is an association (negative)

$\mu_y = \alpha + \beta x \quad \beta < 0$

Hypotheses for LSRL Regression inference

For 1 sample means:

$$H_O : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$
$$(\mu > \mu_0)$$
$$(\mu < \mu_0)$$



For LSRL regression:

$$H_O : \beta = \beta_0$$
$$H_A : \beta \neq \beta_0$$
$$(\beta > \beta_0)$$
$$(\beta < \beta_0)$$

(almost always, $\beta_0 = 0$)

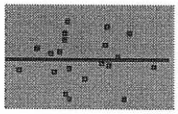
$$H_O : \beta = 0$$
$$H_A : \beta \neq 0$$
$$(\beta > 0)$$
$$(\beta < 0)$$

Conditions

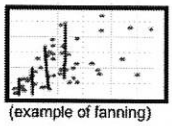
As always, there are conditions. Here, a bit trickier, because checking some of the conditions requires doing an analysis to be able to look at residuals, but we must first be convinced a linear regression is appropriate. We do the following, in order:

1) **Straight enough?** Make a ***scatterplot*** of the data, it must look approximately linear (if not, try re-expressing or stop).

2) **Independent?** Run a ***linear regression***, then ***plot residuals*** you must see no patterns.

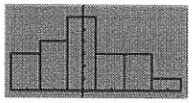


3) **Variance consistent for all x?** You should ***not see fanning*** in either the scatterplot or residuals plot.



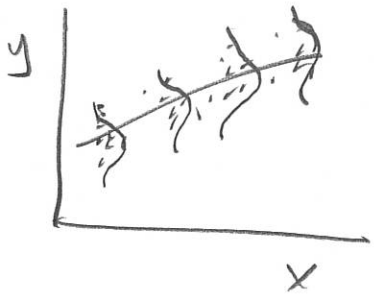
(example of fanning)

4) **Residuals Nearly Normal?** Load residuals into L3 so you can plot a histogram. Should be unimodal, symmetric, no outliers.



Inference with a calculator

Regression analysis is always done using technology, and we will also always use technology for doing inference with regression. There are underlying equations and the textbook does provide a good explanation of them and the basis for how the calculator is accomplishing the analysis (recommend reading the entire chapter).

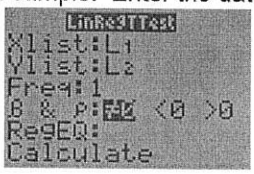


There a few things to know:

- Everything is based upon assuming that for each x value there is a variation in y values, and that these y values form a Normal distribution around the LSRL predicted y value for that x value (some value higher, some lower). This is only true if the sample is representative of the population.
- These 'normal' variations are assumed to all have the same variance for all x values (which is why it is important that the scatterplot not show 'fanning'.
- All of the same issues with outliers and leverage still apply.
- Larger sample sizes yield a more consistent regression slope.
- Samples which aren't 'bunched' together in x yield a more consistent regression slope.
- Higher correlation, $|r|$ closer to 1, yields a more consistent regression slope.
- The computations are based upon means, so t-distribution is used instead of Normal for p-value calculations.

A Linear Regression Slope t-Test

Let's run an analysis on the first mpg vs. weight data sample. Enter the data into L1, L2, then select STAT, TESTS, F:LinRegTTest with $\neq 0$

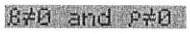


Format for inference answers:

1) State the type of test and H_o, H_A (and define symbols).
2) Check conditions.
3) Conduct test in calculator showing what was entered and report result.
4) State decision (significance level, p-value, reject or fail to reject H_o).
5) State conclusion in context of the problem.

How to interpret each of the output values:

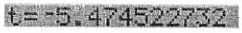
The form of the LSRL equation
a and b are the intercept and slope statistics for this particular sample. This should really be written: $\hat{y} = a + bx$



β is the slope parameter for the population LSRL.

ρ (rho) is the correlation parameter for the population (the parameter corresponding to the statistic r)

These are both zero because the null hypothesis is that there is no correlation which would produce a zero slope LSRL.



The value of the t-statistic (which is used to calculate p-value).
For inference for regression the t-statistic is...

$$t = \frac{b - \beta_0}{s_b}$$

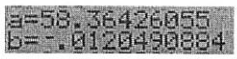
...which is how many standard deviations this particular sample's LSRL slope is away from the population LSRL slope (for the null hypothesis). (s_b is discussed below)

For this case (usually true) $\beta_0 = 0$ so: $t = \frac{b}{s_b}$

t= -5.47 is stating that this particular sample's LSRL slope is 5.47 standard deviations below the null hypothesis slope value of zero.

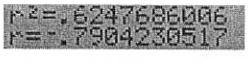
The p-value for the t-statistic, which is the probability that this sample's LSRL slope value would be this far away from the null hypothesis slope value due to natural sampling variation.

The degrees of freedom (based upon sample size) used to select the t-distribution used for p-value calculation.
df = n - 2
(2 because there are two sources of variation: a, and b)



The intercept and slope statistics for this sample's LSRL:

$$\hat{y} = 58.364 - .012x$$



r The correlation coefficient (strength/direction of association).
r^2 The coefficient of determination (% variation explained by LSRL).

The two different Standard Errors

s=2.968799193

The s reported by the calculator is the ***standard error of residuals.*** It represents a measure of the spread of the data points around the LSRL. It's equation is:

$$s = \sqrt{\frac{\sum\left(y_i - \hat{y}_i\right)^2}{n-2}} \quad \left(\textit{textbook labels this } s_e\right)$$

There is another standard error, the ***standard error of the slope*** which represents the amount of variation in the slopes of all sample LSRLs. It's equation is:

$$s_b = \frac{s}{\sqrt{\sum\left(x_i - \bar{x}_i\right)^2}} \quad \left(\textit{textbook labels this } SE\left(b_1\right)\right)$$

The calculator does not provide this value, but we can calculate it from info provided. Because $\beta_0 = 0$:

$$t = \frac{b}{s_b} \quad \text{so} \quad s_b = \frac{b}{t} \qquad \text{Here: } s_b = \frac{b}{t} = \frac{-.0120490884}{-5.474522732} = .0022$$

What we can conclude:

The estimated regression equation is: $\widehat{mpg} = 58.364 - .012\left(weight\right)$

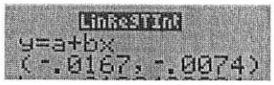
The p-value of $3.36*10^{-5}$ = .0000336 means that the association we see in the data in unlikely to have occurred by chance. We reject the null hypothesis that the slope is zero, and conclude that there is strong evidence that a relationship exists between weight and mpg.

The negative slope (b = -.012) and negative correlation (r = -.79) suggest that as weight increases, mpg decreases. *But to put values on the slope, it would be better for us to calculate a confidence interval for the slope.*

A Linear Regression Slope Confidence Interval

To run a confidence interval analysis with data in L1, L2, select STAT, TESTS, G:LinRegTInt with C-level: .95

Most of the output values are the same as for the t-test, but this test provides the confidence interval:



We can add to our conclusion:

We are 95% confident that the slope of the LSRL for the population is between -.0167 and -.0074. This means that for every additional pound a car weighs the mpg will decrease between .0074 and .0167 miles per gallon.