Inference for Regression

In this lesson (our last one!) we will be building confidence intervals and testing hypotheses about the slope of the regression line.

Regression model:

- Using y = ax + b , the slope and y intercept are statistics.
- The unknown parameters being estimated for a and b are β and α respectively.
- The "true" average linear model is therefore: $\mu_y = \beta x + \alpha$

Assumptions:

1. Linearity. The mean response (μ_y) has a linear relationship with x.
2. Normality (use histogram, boxplot)
3. Simple random sample
4. Independence. If SRS we can assume independence.

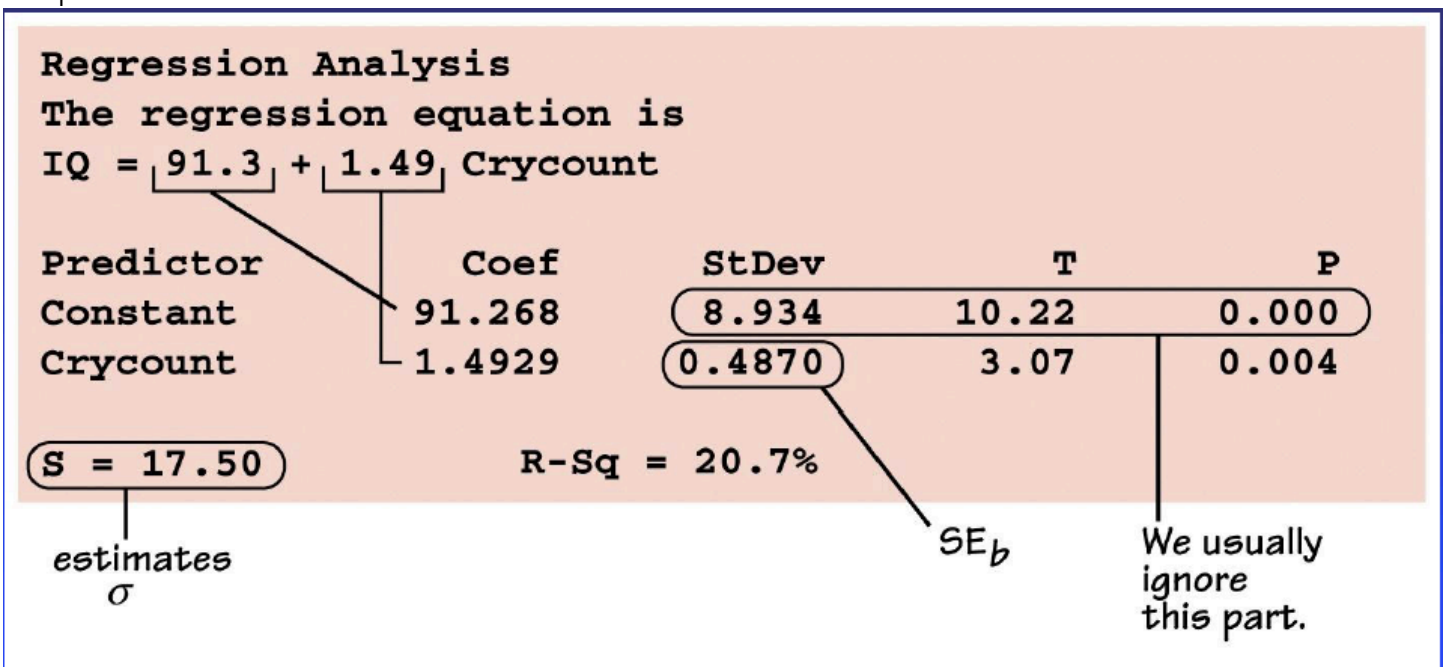
An example of data and output:

TABLE 14.1 Infants' crying and IQ scores

Crying	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	157
33	159	13	162				

Source: Samuel Karelitz et al., "Relation of crying activity in early infancy to speech and intellectual development at age three years," *Child Development*, 35 (1964), pp. 769–777.

Output:



Confidence Intervals

- Confidence intervals for the slope in regression gives a range of values where the true slope (β) might fall.
- The formula is:

$$b \pm t^* SE_b$$

- b = estimate of β
- t*= the upper (1-c)/2 critical value with n-2 df.

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

(Typically given in computer printout)

1. Write the LSRL for the data above.

2. Interpret the slope and y-intercept in the context of the problem.

3. Interpret the r (correlation coefficient) and r^2 (coefficient of determination) value.

4. Find a 90% confidence interval for the true slope of the LSRL.

Hypothesis Test:

Step 1: Same assumptions
Step 2:

- Define x and y of the population.
- Hypotheses:
- $H_o: \beta = 0$ (no linear relationship)
- The alternative hypothesis is always directional based on the wording of the problem:
- $H_a: \beta > 0$ (increasing slope)
- $H_a: \beta < 0$ (decreasing slope)
- $H_a: \beta \neq 0$

Why B = 0?

- Slope of zero would say that y doesn't tend to change linearly when x changes – there is no linear association between the two variables.
- If slope were zero, there wouldn't be much left of the regression equation.

- We will use a t-distribution with n – 2 df.

$$t = \frac{b}{SE_b}$$

- Use tcdf to find the p-value (if doing by hand).
- Conclusion is as usual.

Example:

- Conduct a two-tailed hypothesis test at an $\alpha = .01$ level to determine if the speed in ft/sec of top female runners predicts the number of steps they take per second. (Is there a linear relationship?)

Speed (ft/s)	15.86	16.88	17.50	18.62	19.97	21.06	22.11
Steps per Second	3.05	3.12	3.17	3.25	3.36	3.46	3.55

If we are asked for Standard Error without computer output:

- The calculator does not provide SE_b.
- However, a simple re-arrangement of the formula:

$$t = \frac{b}{SE_b}$$

- will give SE_b:

$$SE_b = \frac{b}{t}$$

- So for this test, SE_b = .0803/49.6582 = .00162

- This value estimates the variability in the sampling distribution of the estimated slope (how much we would expect sample slopes to vary from experiment to experiment).