Chapter 1

Inference about the least-squares regression line

In chapter 2 we learned how use least-squares regression line to describe the linear relationship between two quantitative variables, based on the sampled data. Is this least-squares regression line still true for population data? We will answer this question in this chapter.

1.1 The basics

- **Notations**

  If the least-squares regression line is based on the sample data, the equation is

  $$\widehat{y} = a + bx$$

  If the least-squares regression line is based on the population, the corresponding parameters for statistics a and b are α and β.

  Recall from chapter 2 that s means the *standard deviation of the residuals*, which is interpreted as the "the typical error of the prediction". The corresponding parameter is σ.

  The least correlation r is a statistic to evaluate the strength of the linear relationship. The corresponding parameter is ρ.

- **The conditions for regression inference**

  - **Linear:** The actual relationship between x and y is linear. For any fixed value of x, the mean response my falls on the population (true) regression line $\mu_y = a + bx$.
  - **Independent:** Individual observations are independent of each other. When sampling without replacement, check the 10% condition.
  - **Normal:** For any fixed value of x, the response y varies according to a Normal distribution.
  - **Equal SD:** The standard deviation of y (call it σ) is the same for all values of x.
  - **Random:** The data come from a well-designed random sample or randomized experiment.

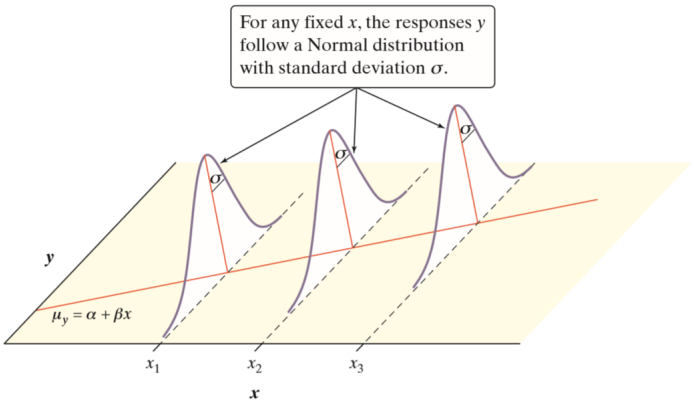


Figure 1.1: The normal and equal SD condition [1]

[1] This figure is adapted from *The practice of statistics, 5th edition, By Starnes etc.*

If all the above conditions are satisfied, we can do the **regression inference**. The sampling distribution of b is approximately normal.

$$b \sim N(\mu_b, \sigma_b), \qquad \mu_b = \beta,\ \sigma_b = \frac{\sigma}{\sigma_x\sqrt{n}}$$

Those conditions are rarely checked in free response problems. In most cases, the stem of the problem tells you that all the conditions are met.

1.2 Regression inference

Why the the confidence interval is about β instead of b?

- **Confidence interval about** β

The formula for the confidence interval are in the same form as before.

Statistic $\pm$ (**Critical value**) $\cdot$ (**Standard deviation of the statistic**)

From previous section we know that

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

In practice, we don't know σ, and replace it by s. The denominator $\sigma_x\sqrt{n}$ is replaced by $s_x\sqrt{n-1}$.

$$\mathbf{SE}_b = \frac{s}{s_x\sqrt{n-1}}$$

The statistic $t = \dfrac{b-\beta}{\mathbf{SE}_b}$ has a t distribution with degree of freedom $df = n-2$.

Thus the formula for confidence interval is

$$b \pm t^*\mathbf{SE}_b$$

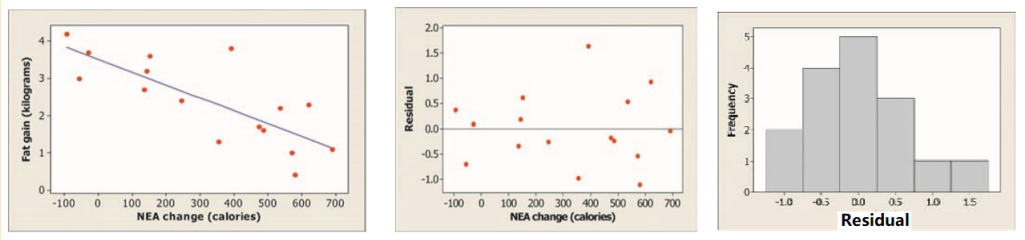
Most of the time, $\mathbf{SE}_b$ is given by the stem of the problem directly.

Keep slim

Does fidgeting keep you slim? Some people dont gain weight even when they overeat. Perhaps fidgeting and other nonexercise activity (NEA) explain whysome people may spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed a random sample of 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) as the response variable and change in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like—as the explanatory variable. Here are the data:

NEA change (cal):	94	57	29	135	143	151	245	355
Fat gain (kg):	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (cal):	392	473	486	535	571	580	620	690
Fat gain (kg):	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Minitab output from a least-squares regression analysis for these data is shown below.



Regression Analysis: Fat gain versus NEA change

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.03036	11.54	0.000
NEA change	−0.0034415	0.0007414	−4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq(adj) = 57.8%

(a) Give the standard error of the slop $\mathbf{SE}_b$ and interpret this value in the context.

(b) Suppose all the conditions are met. Construct and interpret a 95% confidence interval for the slope of the population (true) regression line.

(c) Interpret the result in (b).

(d) How would you check the *normal* and *equal SD* conditions?

- **Hypotheses test about β**

  The conditions to be checked is the same as the conditions checked in constructing confidence intervals for β.

  Suppose the null hypothesis is

  $$\mathbf{H}_0: \quad \beta = \beta_0$$

  The alternative hypothesis can be either of the following

  $$\mathbf{H}_a: \quad \beta \neq \beta_0, \qquad \mathbf{H}_a: \quad \beta > \beta_0, \qquad \mathbf{H}_a: \quad \beta < \beta_0$$

  The test statistic is

  $$t = \frac{b - \beta}{\mathbf{SE}_b}, \quad df = n - 2$$

  The way to calculate the P-value is the same as the tests we learned before. It depend on the $\mathbf{H}_a$

  Sometimes, you may be asked to perform a hypotheses test about population correlation ρ. The null hypothesis is $\mathbf{H}_0: \quad \rho = 0$ and the alternative hypothesis may be either of the following

  $$\mathbf{H}_a: \quad \rho \neq 0, \qquad \mathbf{H}_a: \quad \rho > 0, \qquad \mathbf{H}_a: \quad \rho < 0$$

  Since we know that β and ρ have the same sign. The above hypotheses test are the same as the following:

  The null hypothesis is $\mathbf{H}_0: \quad \beta = 0$. The alternative hypothesis may be either of the following:

  $$\mathbf{H}_a: \quad \beta \neq 0, \qquad \mathbf{H}_a: \quad \beta > 0, \qquad \mathbf{H}_a: \quad \beta < 0.$$

Crying and IQ

Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test 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 its intensity by the number of peaks in the most active 20 seconds. They later measured the childrens IQ at age three years using the Stanford-Binet IQ test. A random sample of 38 infants is studied and the results are shown below.

Regression Analysis: IQ versus Crycount

Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%

(a) What is the equation of the least-squares regression line for predicting IQ at age 3 from the number of crying peaks (crycount)? Interpret the slope and y intercept of the regression line in context.

(b) Interpret s

(c) Calculate the correlation r and interpret *the coefficient of determination.*

(d) Do the results provide convincing evidence of a positive linear relationship between crying counts and IQ in the population of infants? Suppose all the conditions for statistical inference are met.