12.1 Inference for linear regression

We want to answer: Is there really a linear relationship between x & y in the population, or could the pattern we see in the scatterplot plausibly happen by chance?

Recall the LSRL has the form

$\hat{y} = a + bx$ where a is the y-intercept and b is the slope.

formula sheet: $\hat{y} = b_0 + b_1 x$

This is an estimate of the true population regression line

Confidence Intervals and significance tests for linear regression are about the slope of the population regression. As you would expect, if we took many samples of data from the population, the slopes would likely be different, but pretty close.

www.rossmanchance.com/applets

Sampling Distribution of a slope

Choose an SRS of n observations (x,y) from a population with size N. It has

predicted $y = \alpha + \beta x$

Let b be the slope of the sample LSRL. Then:

1. The mean of the sampling dist. of b is $\mu_b = \beta$
2. The standard deviation of the sampling dist. of b is

$$\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}} \text{ if } n \leq \frac{1}{10} N$$

3. The sampling distribution of b will be approx Normal if the values of of y follow a Normal distribution for each x.

Conditions for regression inference (LINER)

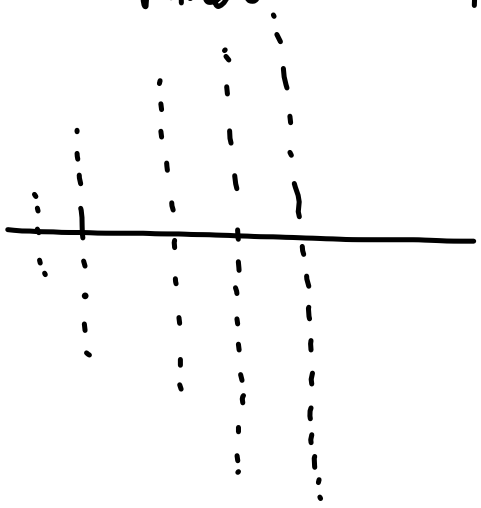
Linear: actual relationship between x and y is linear (scatterplot + residual plot)

Independent: observations are independent of each other and 10% condition is met.

Normal: for any fixed x value, y varies according to the Normal distribution (histogram of residuals)

Equal SD: standard deviation of y, σ, is the same for all values of x (use residual plot)

Random: Data comes from a random sample or randomized experiment.



t interval for slope

When conditions are met, a C% CI for the slope, β, of the population LSRL is:

$$b \pm t^* SE_b$$

where $SE_\beta = \frac{s}{s_x\sqrt{n-1}}$ where $s = \sqrt{\frac{\Sigma(\text{residuals})^2}{n-2}}$

and t^* is the critical value for the t dist. with $df = n-2$ having C% of its area between $-t^*$ and t^*

Check your understanding p. 752

State:

β = true slope of the LSRL comparing fat gain vs. NEA.

We will perform a 95% confidence interval for β.

Plan: t interval for slope

conditions:

Linear: scatterplot is fairly linear and a well scattered residual plot.

Independent: peoples fat gains are independent
No need to check 10% condition

Normal: Histogram is roughly symmetric, no outliers.

Equal standard deviation: residual plot shows that standards are approximately equal for each x.

And

Random: Random ✓

Do:

$$-.0034415 \pm -4.64(.0007414)$$

On my calc, use LinRegTInt,
XList is NEA change. in L_1
YList is Fat gain in L_2
using a confidence level of .95

$(-.005, -.0019)$

Conclude: We are 95% confident the interval of $(-.005, -.0019)$ captures the true slope of NEA change vs. Fat gain.

t-test for slope

suppose conditions are met. To test

$H_0: \beta = \beta_0$ against an alternative

compute $t = \frac{b - \beta_0}{SE_b}$. Find the p value by calculating the probability of a t statistic this large or larger in the direction of H_a.