

Stat 218 - Day 35 Inference for Regression

Recall that we have been studying relationships between two *quantitative* variables. We have determined how to fit a **least squares regression line** to model the relationship between the two variables: $\hat{y} = b_0 + b_1x$.

Today we will suppose that our regression analysis is based on a random sample from a population and that we want to draw inferences about the linear model for the population:

$Y = \beta_0 + \beta_1X$. We will primarily be interested in drawing inferences (conducting hypothesis tests and constructing confidence intervals) about the population slope coefficient β_1 .

- Typical null hypothesis is that $\beta_1 = 0$ (no linear relationship)
- Compare to t -distribution with $(n-2)$ degrees of freedom

$$SE(b_1) = \frac{s_{Y|X}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- The standard error of the sample slope coefficient is:
 - Minitab calculates standard error, test statistic, two-sided P -value

(a) Do b_0 and b_1 represent statistics or parameters? Explain.

(b) Do β_0 and β_1 represent statistics or parameters? Explain.

Example: House prices (cont.)

Reconsider the sample data on house prices in Bakersfield (`RealEstate.mtw`).

a) Produce regression output for predicting price from size (`Stat > Regression > Regression`). Report the sample slope coefficient and its standard error, along with the appropriate symbols.

b) State the null and alternative hypotheses, in symbols and in words, for testing whether there is a positive slope coefficient in the population.

c) Determine (by hand) the test statistic and P-value.

d) Summarize the conclusion that you would draw from this test.

e) Produce a 95% confidence interval for the population slope coefficient.

f) Interpret this interval.

If we let ρ denote the correlation coefficient between two variables in a population, then we can conduct a test of $H_0: \rho=0$ based on the sample correlation coefficient r , using the test statistic

$t_s = r \sqrt{\frac{n-2}{1-r^2}}$. The degrees of freedom for determining the P-value from the t -distribution is again $(n-2)$.

g) Confirm that this produces the same test statistic value that appears in the Minitab output above.

Example: Nenana ice break competition (cont.)

Recall that you analyzed data on the arrival time of spring in Nenana, Alaska for every year since 1917 (`NenanaIceBreak.mtw`).

(a) Use Minitab to calculate the correlation coefficient between arrival date of spring (in `c7`) and year (in `c1`).

(b) Treat this as a sample correlation coefficient, and calculate the test statistic for testing whether it differs significantly from zero.

(c) Use the t -table to determine the P -value as accurately as possible. Interpret this P -value, and summarize your conclusion.

Example: Draft lottery (cont.)

Recall that early in the course, we studied the 1970 draft lottery, in which each of the 366 birthdays of the year was assigned a draft number between 1 and 366.

(a) Use Minitab to calculate the correlation coefficient between draft number and sequential birth date. Also report the corresponding P -value.

(b) How often would a fair, random lottery process produce a correlation coefficient as large (in absolute value) as the one observed in 1970? Explain how you know.

(c) Summarize the conclusion that you would draw from this analysis. Also describe the reasoning process behind your conclusion.