Math 455 Class 2 January 23

Homeworks 1 and 2 are in Gradescope.

You should already have been added to Gradescope.

Course Web page:

people.math.binghamton.edu/dikran/455/

In regression output we see "p-value".

Review of Math 448: What is a p-value?

A p-value is defined in the context of a statistical test.

The p-value is the smallest level of significance for which the test statistic indicates the null hypothesis should be rejected.

(Wackerly Def. 10.2 p. 513) for a more exact version.

In order to make sense of this we need

to know two prior definitions.

"level of significance" and "statistical test"

A statistical test (Wackerly Ch. 10, p. 490)

Elements:
1. Null Hypothesis $H_0$
2. Alternative Hypothesis $H_a$
3. Test statistic.
4. Rejection Region.

Level of significance: (Wackerly Defn 10.1) p. 491

Type I error: rejecting $H_0$ when $H_0$ is true.

(Assuming $H_0$ & we can compute prob of type I error.) This prob. of type I error is denoted $\alpha$. The value of $\alpha$ is called the level of the test.

---

What does this mean in practice?

Imagine we are a drug company and we want the FDA to approve our drug that lowers blood pressure.

regression output is associated to the ~~Ho~~ alternative hypothesis $|\beta_1| \neq 0$, not $\beta_1 \leq 0$ as above.

Q: We think our drug works, otherwise we wouldn't spend $ billions on the clinical trial. Why do we assume it doesn't ($H_0$)?

A: When we analyze the data, we put on our skeptical statistician hats, and assume $H_0$. Then when we get a low p-value, we say to the FDA, "Look! We must reject the hypothesis $H_0$ that our drug doesn't work."

---

Regression software (R, lm) computes for us a "best" $\beta_0, \beta_1$ in

$$Y = \beta_0 + \beta_1 X + \text{error}.$$

Q: How are the "best" $\hat{\beta}_0, \hat{\beta}_1$ chosen?

A: "Least-Squares".

Skipping the calculus, the answer is:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ $\quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

$\hat{\beta}_0$ satisfies: $(\bar{X}, \bar{Y})$ is on the regression line: $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$.

Warning: You may see lots of different-looking formulas for $\hat{\beta}_1$, $\hat{\beta}_0$ (or the multivariate analogs.)

Example: $\hat{\beta}_1 = \rho \frac{\sigma_Y}{\sigma_X} = \frac{\rho \sigma_X \sigma_Y}{\sigma_X \sigma_X} = \frac{Cov(X,Y)}{Var(X)}$

Or: $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

This is the same because:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$= n \cdot \frac{1}{n} \sum_{i=1}^{n} x_i - n \cdot \bar{x}$$

$$= n \bar{x} - n \bar{x} = 0.$$

Thus $$\bar{Y} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) = \bar{Y} \cdot 0 = 0.$$

and the $-\bar{Y} \left( \sum_{i=1}^{n} x_i - \bar{x} \right)$ can be eliminated from the top of the fraction.