CSCI 3022

intro to data science with probability & statistics

November 12, 2018

Statistical regression
&
Inference in Regression

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- **HW6** posted tonight!. Giddyup!

Last time on CSCI3022: SLR

- Given data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ fit a simple linear regression of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

- Compute estimates of the intercept and slope parameters by minimizing:

$$SSE = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

- The least-squares estimates of the parameters are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Residuals

- The **fitted** or **predicted** values ____________________ are obtained by substituting $x_1,\ldots x_n$ into the equation of the estimated regression line.

- The **residuals** are the differences between the observed and fitted *y* values:

Residuals

- Why are the residuals estimates of the error?

Maximum likelihood estimates

- Rather than minimizing the sum of the squared errors to find the parameters of the model, we can *maximize the likelihood of the data* by changing the parameters.
- You already know **maximum likelihood estimates** but we never called them that before.
- Imagine that we flip a biased coin and get 5 heads and 1 tails. What is the maximum likelihood estimate of the coin's bias, p?

Maximum likelihood estimates

- Three steps:
    1. Assume the parameter *p* is fixed (for now).
    2. What is the probability that we observe 5H and 1T, given p? Note: this probability is called *the likelihood*. If we take a log, this is now called the *log likelihood*.
    3. Take the derivative of step 2 with respect to p and set equal to zero. In other words, maximize the likelihood of getting 5H and 1T by finding the optimal p.

Maximum likelihood estimates

MLE (generally)

- **Maximum Likelihood Estimation** asks: what are the *parameters* that best explain the data that we see?
- **In practice**, this means that we usually go through three steps:

1. Write down the probability of getting the data, given the probability distribution and the parameter(s) of interest. (This is the likelihood.)
2. Take a log to get the *log-likelihood*.
3. Take a derivate with respect to the parameter, set equal to zero, and solve to find the MLE value of the parameter. (Don’t forget to put a hat on it 🎩)

MLE for simple linear regression

1. P(data | params)
2. Take a log.
3. Derivative = 0

The punchline:

- Maximum Likelihood and Least-Squares are solving **the same problem**
- Important: this means that when we are solving the least-squares problem, what are we *always, implicitly assuming about the errors?*
-
-
-

For the rest of today:

- **How can we**:
  - Estimate the variance in the population of estimates?
  - Quantify the goodness-of-fit in our simple linear regression model?
  - Perform inference on the regression parameters?

Estimating the variance

- The parameter σ^2 determines the spread of the data about the true regression line. [We experimented with this in the notebooks!]

Estimating the variance

- The divisor (n-2) in the estimate of σ^2 is the number of *degrees of freedom* (abbreviated df) associated with the estimate of SSE.

- This is because to obtain $\hat{\sigma}^2$, the two parameters $\hat{\alpha}$ and $\hat{\beta}$ must first be estimated, which results in a loss of 2 degrees of freedom.

The coefficient of determination

- The coefficient of determination, R^2 quantifies how well the model explains the data.

- R^2 is a value between 0 and 1.

The coefficient of determination

The **sum of squared errors** (SSE)

can be interpreted as a measure of how much variation in *y* is left unexplained by the model: how much variation *cannot* be attributed to a linear relationship?

The **regression sum of squares** is given by

A quantitative measure of the total amount of variation in observed *y* values is given by the so-called **total sum of squares**

The coefficient of determination

- The sum of squared deviations about the least-squares line is smaller than the sum of squared deviations about any other line, i.e. SSE < SST unless the horizontal line itself is the least-squares line
- The ratio SSE/SST is the proportion of total variation in the data that cannot be explained by the simple linear regression model, and the coefficient of determination is

The coefficient of determination

The coefficient of determination

- Note: R^2 is the proportion of total variation in the data that is explained by the model.
- But: R^2 does *not* tell you that you necessarily have the correct model!

Inference about parameters

- The parameters in simple linear regression have distributions! We demonstrated this in the in-class notebook last time.
- From these distributions, we can conduct hypothesis tests (e.g.:   ), compute confidence intervals, etc.
- **Distributions**:

Inferences about the parameters

- **Confidence intervals**:

- **Tests**: