LINEAR REGRESSION WITH A SINGLE-PREDICTOR

ETHAN LEVIEN

Contents

1. Linear regression 1
2. Least squares for the regression model 1
3. Correlation and standardization 2
4. p-values 2

1. Linear regression

The subject of this course is mostly linear regression models. In the simplest case where we have a single independent variable, x, a regression model for the relationship between Y and x is

$$y \sim \text{Normal}(\alpha + \beta x, \sigma) \tag{1}$$

Written another way, our model for Y is

$$y = \alpha + \beta x + \xi, \quad \xi \sim \text{Normal}(0, \sigma). \tag{2}$$

We are interested in the associated statistical inference problem, but first, let's assume a and b are known and think about some of the predictions we can make.

Exercise 1. *Suppose that a regression model for the relationship between time and is*

$$y \sim \text{Normal}(1.4x + 5, 1.) \tag{3}$$

What is the chance that the incumbent receives more that 50% of the vote if the economic growth was 4%?

2. Least squares for the regression model

Now suppose we have some data D consisting of x and y pairs:

$$D = \{(x_1, y_1), \dots, (x_n, y_n)\} \tag{4}$$

One can show that an of β and α is

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \tag{5}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{6}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y respectively.

Exercise 2. *????*

We call this least squares estimator because it comes minimizing the squares of the residuals,

$$r_i = \hat{\beta} x_i + \hat{\alpha} - y_i. \tag{7}$$

By minimizing r_i we ensure that the probability we see the data is as high as possible:

$$p(y|x) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-r_i^2/(2\sigma^2)} \tag{8}$$

What about $\hat{\sigma}$? Recall that the estimator of a the standard deviation of a Normal distribution is

$$\sqrt{\frac{1}{n-1}\sum y_i^2} \tag{9}$$

If we want to estimate the standard error, we might expect to replace y_i with r_i (can you see why?); however, this does not account for the fact that we don't know $\hat{\alpha}_i$. This additional degree of freedom causes the Equation (9) to over estimate the variance. Correcting for this yields

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum y_i^2} \tag{10}$$

You can see the formal derivation of all this in [].

Exercise 3. *Are the residuals the same things as ξ?*

Date: April 2022.

3. Correlation and standardization

We start with an exercise in rewriting the estimator of β: Let $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ be the estimators of the variance in x and y. The sample correlation of two samples x and y is given by

$$r_{x,y}^2 = \frac{1}{n-2}\sum(x_i - \bar{x})(y_i - \bar{y}) \tag{11}$$

Note that $\hat{\beta} = r_{x,y}/\hat{\sigma}_x^2$. Sometimes it is useful to *standardize* the variables before performing a regression, meaning we transform them into standard Normal random variables.

Exercise 4. *Show that for a sample of a random variable* $x_1, x_2, \ldots, x_n$

$$z_i = \frac{x_i - \bar{x}}{\hat{\sigma}_x^2} \tag{12}$$

If we perform a regression on the standardized variables, then

$$\hat{\beta} \tag{13}$$

The quantity ρ is called the *correlation coefficient*. We will take

Exercise 5. *Is the correlation coefficient symmetric; that is, i*

4. p-values

We've already discussed p-values and mentioned some potential reasons to avoid using them. However, they play a central role and statistics and we must therefore understand them in the context of regression.