Statistics 112
Regression Cheatsheet
Section 1B - Ryan Rosario

I have found that the best way to practice regression is by brute force. That is, given nothing but a dataset and your mind, compute everything there is to compute about a regression model! So let's pretend that.

Part 1 - Simple Linear Regression

The simple linear regression model is, well, simple(r).

$$y = \beta_0 + \beta_1 x$$

where y is the response variable and is known if the data is given to us. **The above is the true, unobserved model.** Some oracle out there knows what the βs are, but humans are not as wise. The best we can do is *estimate* the values of β_0 and β_1 using the estimators b_0 and b_1!

The *fit* (or *estimated*) model is

$$\hat{y} = b_0 + b_1 x$$

where $\hat{y}$ is the *predicted* value of y based on our fitted model. Note that we can predict a y for any value of x **that is in the domain of** x. That is, if x goes from 0 to 100, then we can predict y for values of x from 0 to 100.

Suppose we have data that looks like the following

x	y
$\vdots$	$\vdots$

We need to fit a model, so we need to find b_0 and b_1. **For simple linear regression,**

$$b_1 = r\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b_1\bar{x}$$

where r is the correlation coefficient between x and y, s_x and s_y are the standard deviations of x and y respectively, and $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

Step 1: Find $\bar{x}, \bar{y}, s_x, s_y$

$$\bar{x} = \frac{\sum x_i}{n}, \qquad \bar{y} = \frac{\sum y_i}{n}, \qquad s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}, \qquad s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Step 2: Calculate r, the correlation coefficient.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y}}{(n-1)s_x s_y} = \frac{\sum z_x z_y}{n-1}$$

Step 3: Calculate the regression coefficient and intercept.

Given r,

$$b_1 = r\frac{s_y}{s_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Step 3: Write out the linear model using these estimates.

$$\hat{y} = b_0 + b_1 x$$

Because simple linear regression has only one regression coefficient, there is **no** F-test involved. Think about why this is true by looking ath te null/atlernate hypotheses as well as the null/alternate hypotheses for the t test. In this boring case where there is one predictor,

$$F = t^2$$

Step 5: Compute R^2 which is just, well, r raised to the power of 2!

Step 6: Determine whether or not the model and/or regression coefficient is statistically significant.

Using a confidence interval:

$$b_1 \pm t^* \text{SE}(\beta_1)$$

Since $H_0 : b_1 = 0$, we check to see if 0 is in the confidence interval. If it is, fail to reject H_0, x has no effect on y. Otherwise, reject H_0, x does have an effect on y.

Or using a hypothesis test.

$$t = \frac{b_1 - \beta_1}{\text{SE}(\beta_1)} = \frac{b_1 - 0}{\text{SE}(\beta_1)}$$

then find the critical t^* from the t-table using $n-2$ degrees of freedom.

If the t-values or the p-values are given USE THEM!

Step 7: Compute the standard error of the predicted values.

If you are asked about prediction error, compute s.

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-1}}$$

where $y_i - \hat{y}$ is the *residual* which is the error in predicting a particular value of y from a value of x.

Part 2 - Multiple Linear Regression

Simple linear regression was a special case of multiple regression which is unfortunate. The principle is still the same, but *how we account for variance is more tedious.*

The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where y is the response variable and is known if the data is given to us, but we only know y for values of y that are *in* the dataset! $x_1, x_2, \ldots, x_n$ is also known because it is given to us with the data. **The above is the true, unobserved model.** Some oracle out there knows what the βs are, but humans are not as wise. The best we can do is *estimate* the values of β_i using the estimators b_i!

The *fit* or *estimated* model is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Suppose we are given the following data. Let's compute a multiple regression entirely by hand...

x_1	x_2	$\ldots$	x_n	y
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

IMPORTANT NOTE: You will NOT be asked to compute the regression coefficients for a multiple linear regression. It requires linear algebra.

To compute by hand, you must be given either x_is or y_is. Using these x_is and a **model given to you**, you can then compute the predicted value of y, $\hat{y}$ if $\hat{y}$ is not given to you.

Step 1: Compute the residuals.

$$e_i = y_i - \hat{y}_i$$

Step 2: Compute the sum of squares.

The total variance in the response variable can be quantified using the sum of squares. The **total** sum of squares $SSTO$ quantifies *all* variance in the response variable. It is a sum of two specific sources of variance: the variance that is explained by the regression model (good) and variance that is not explained by the regression model (bad).

$$SSTO = \sum (y_i - \bar{y})^2$$

The sum of squares regression SS_{REG} is the amount of variance that is explained by the regression model. It can also be called sum-of-squares explained.

$$SS_{REG} = \sum (\hat{y}_i - \bar{y})^2$$

The sum of squares residual SSE, or the sum of squares error, is the amount of variance that is not explained by the regression model. That is, it is the amount of variance that is described by factors other than the predictors in the model.

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

note that $y_i - \hat{y}_i$ is the residual.

Step 3: Compute the mean squared error.

In Step 2 we computed the squared error. Now we want to essentially average it by dividing the "something." This something is the degrees of freedom. This yields that "average squared error."

$$MS_{REG} = \frac{SS_{REG}}{k}$$

$$MSE = \frac{SSE}{n-k-1}$$

where k is the number of predictors in the model and n is the sample size.

Step 4: Compute the F value to perform the F test.

F is sometimes called the signal-to-noise ratio. As in an audio system, noise is bad, it is erroneous sound. Signal is what our brains actually pay attention to; it is the part of the variance that our brains can explain! Thus, using this analogy,

$$F = \frac{\text{Signal}}{\text{Noise}} = \frac{MS_{REG}}{MSE} = \frac{\frac{SS_{REG}}{k}}{\frac{SSE}{n-k-1}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

Sometimes k is called the "degrees of freedom in the numerator" and $n-k-1$ is called the "degrees of freedom in the denominator." This is the terminology used by the F table.

Look up the degrees of freedom in the F table and find the critical F value F^*. If your calculated value of $F > F^*$, we reject the null hypothesis of the F test.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0$$
$$H_a : \beta_i \neq 0, \text{for some } i$$

If we reject H_0 this means that at least one of the β_is is not 0, that is, at least one of the explanatory variables helps to predict y. If we fail to reject H_0, all β_is are 0, and none of the predictors have any effect on the response variable. If we fail to reject, we stop. We cannot continue.

Step 5: Compute R^2.

Remember, R^2 is the proportion of the variance that is explained by the regression model. Recall that

$$SS_{REG}$$

is the amount of variance that can be explained by the regression model. Thus,

$$R^2 = \frac{SS_{REG}}{SS_T}$$

On the other hand,

$$1 - R^2 = 1 - \frac{SS_{REG}}{SS_T} = \frac{SS_T - SS_{REG}}{SS_T} = \frac{SS_{RES}}{SS_T}$$

which is the proportion of variance in y that is NOT explained by the regression model. That is, it is the proportion of the variance that is explained by factors other than the predictors in the model.

Step 6: Which regression coefficient(s) are significant? Which variables do help predict y?

Perform a t-test for every coefficient just as we did with simple linear regression. You need the standard error of β, $\text{SE}(\beta_i)$, to use the hypothesis test or confidence interval. *IF THE t-VALUE or p-VALUE IS GIVEN TO YOU, USE THEM!*

Step 7: Compute the Standard Error of the Estimate

The standard error of the estimate (if needed) gives an overall value of how accurate the prediction is.

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = s_y\sqrt{1-R^2}\sqrt{\frac{N}{N-2}}$$