Chapter 10: Inference for Regression

Introduction

In this section will explain how to make statistical calculations when there is one quantitative response variable and one quantitative explanatory variable. Just as in chapter 2, we will use the regression line formula, $\hat{y}= b_0+ b_1x$. In this chapter, we investigate the extent to which a calculated regression line can be used to estimate the true regression line associated with the population. The population regression line can be denoted as $\beta_0+ \beta_1x$

10.1 Simple Linear Regression

Populations

Simple linear regression is used to examine the relationship between a response variable (y), and an explanatory variable (x). We expect that different values of x will corresond to different values of y. Suppose we want to record the change in blood pressure for two experimental groups: a treatment group and a placebo group. In this case, the explanatory variable would be treatment vs. placebo, and the response variable would be the blood pressure of study participants.

- The mean change in blood pressure may be different in the two populations. These averages are called μ_1 and μ_2.
- Individual changes in blood pressure vary within each population according to the normal distribution. This means that the majority of people in a group have approximately the same blood pressure, while a limited number of people have blood pressures that are extremely different from the rest of the group. It is assumed that the standard deviations for the populations are the same.

Subpopulations

In linear regression, the explanatory variable (x) can have a range of many different values. For example, you can give various amounts of calcium to different groups of participants. These values of x can be seen as sub-populations:

- Each value of x is associated with one subpopulation. Each subpopulation consists of all of the individuals in the population who have the same value of x.

The statistical model for simple linear regression assumes that the observed values of y for each value of x are normally distributed with a mean that is dependent on x. We use the symbol μ_y to denote the means of these subpopulations. The means may change as x changes in a pattern. With simple linear regression, we assume that the means will fall on a straight line when plotted against x. In short, the model of simple linear regression has two parts:

- There is a change in the mean of y when x changes. All averages are aligned. This is the regression line of the population is represented by $\mu_y= \beta_0+ \beta_1x$.

- Individual values of y (based on the same x) are normally distributed. These normal distributions all have the same standard deviation.

Residuals

The calculated regression is never perfect when it comes to the prediction of y-values on the basis of x-values. Therefore, the following rules apply:

- Data = fit + residual
- The "fit" consists of the sub-population averages that are found through $\mu_y = \beta_0 + \beta_1 x$.
- The "residual" represents the deviations of the data from the line that represents the population averages. We assume that these deviations are normally distributed and have standard deviation, σ. We use the Greek letter ε when we talk about the residual portion. The ε-values can be seen as 'noise', the portion of the data that cannot be explained with the regression line. These points show up in random clusters and should never form a line.

Model for Simple Linear Regression

The model for simple linear regression follows the following rules:
Given n number of observations of x and y, data points can be written as (x_1,y_1), (x_2,y_2),…, (x_n,y_n)
The observed response (y_i) is associated with explained and unexplained components: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. In this formula $\beta_0 + \beta_1 x_i$ the average response when x = x. The deviations (ε_i) are independent and normally distributed. They have a mean of 0 and standard deviation, σ. Thus, the parameters of the model are β_0, β_1 and σ.

Estimating Regression Parameters

As mentioned earlier, we want to use the regression line that we calculated, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. to estimate population parameters. This line can be found by using the following:

- $b_1 = r (s_y / s_x)$. In this formula, r is the correlation between x and y. The rest of the formula makes use of the standard deviations of x and y.
- b_0= : y-intercept - $b_1 \bar{x}$.
- The residual is: e_i = (observed y-value) - (predicted y-value). This is the same as: Yi - Yi. This is again the same as $y_i - b_0 - b_1 x_i$. The residuals (e_i) correspond to the residuals ε_i. The sum of e_i is always 0 and the εi of a population having a mean of 0.

This leaves the standard deviation (σ) as the last parameter to be calculated. This parameter measures the extent to which y-values of the population deviate from regression line. To calculate these parameters, we make use of residuals.

- First, we must calculate the variance of the regression line in the population (σ^2). We do this by using the variance of the sample: $s^2 = (\Sigma e^2 i) / n - 2$. This is the same as: $\Sigma (yi - yi)^2 / n - 2$.
- Then, we find the square root of the variance (s^2) to find σ.

Confidence intervals

Confidence intervals can be found, in general, by the formula:$\mu \pm t^* * SE_\mu$. Confidence intervals can also be found for β_0 and β_1:

- The confidence interval for the intercept β_0 is: $b_0 \pm t^* SEb_0$
- The confidence interval for the regression coefficient β_1 is: $b_1 \pm t^* SEb_1$.
- In these formulas, t* is the value t (n-2), with area C between t* and -t*.

Prediction intervals

Sometimes, we want to predict the value of y that lies far beyond the y-values in the data. In this case, we make use of a prediction interval. First, a sample of n observations needs to be drawn. Then, the 95% confidence interval (x*) should be calculated for a particular x-value.

- The prediction interval for a future observation of y from the subpopulation of x* is: $\hat{y} \pm t^* * SEb_y$. In this formula, t* is the value t (n-2), with area C between t* and -t*.

10.2 More Detail About Simple Linear Regression

Analysis of Variance (ANOVA) for Regression

By using an analysis of variance (ANOVA), we can find out the extent to which data can be explained by the part that fits to the regression line (fit), and the part that deviates from this line (residuals). The total variation in y is expressed by the deviations y_i – y-bar . If these abnormalities are all 0, all of the observations are equal and there would be no variation in y. There are two reasons why y_i would not equal y-hat:

- The values of y_i are associated with different values of x and are therefore different.
- Individual observations will differ from the average, since there is variation within the sub-population that is associated with a particular x-value.

The Model

As stated previously, we use a linear regression model using the formula: data = fit + residuals. If we look at this in terms of variance,

- SST = SSM + SSE, where SS stands for *sum of* squares, and T, M, and E stand for total, model, and error respectively.
- SST is calculated using the formula: $\Sigma (y_i\text{-y-bar})^2$.
- SSM is calculated using the formula: $\Sigma (y_i\text{-y-bar})^2$.
- SSE is calculated using the formula: $\Sigma(y_i - \hat{y}_i)^2$.

Degrees of Freedom and MS (mean square)

In addition, it is also possible to calculate the corresponding degrees of freedom for each source of variance. It is based on a similar formula:

- DFT = DFM + DFE, where DF = degrees of freedom. The degrees of freedom associated with the total, model, and error, can be calculated as follows:
- DFT = n-1.
- DFM = 1.
- DFE = n-2.

We find the MS for each source of variance by dividing the SS by the corresponding degrees of freedom (DF).

- If the MS is to be found for the total, it is done by calculating SST / DFT.
- The proportion of explained variance (r^2) can be found by calculating SSM / SST. The result shows us how much of the variance in y is explained by the model.

The ANOVA Test

The null hypothesis that the regression coefficient (β_0) of the population is 0, can be tested by using the *F-test*. This null hypothesis basically says that x is not linearly related to y. We conduct the F-test as follows:

- F = MSM / MSE

If the null hypothesis is true, this F-test has a distribution of one degree of freedom in the numerator and n-2 degrees of freedom in the denominator: F (1, n-2). These degrees of freedom are associated with MSM and MSE. Just as there exist a lot of t-tests, there are also many F-tests. If the regression coefficient is not 0 ($\beta_1 \neq 0$), then MSM is relatively large with respect to MSE. This means that large values of F provide evidence against the null hypothesis, in favour of the two-sided alternative.

The information that has been given so far is briefly summarized in the following ANOVA table:

Source	Degrees of Freedom	Sum of Squares	Mean Square (MS)	F
Model	1	$\Sigma(\hat{y}_i$-y-bar$)^2$	SSM/DFM	MSM/MSE
Error	n-2	$\Sigma(y_i - \hat{y}_i)^2$	SSE/DFE	
Totaal	n-1	$\Sigma(y_i$-y-bar$)^2$	SST/DFT	

Test for a Zero Population Correlation

We can also assess whether there is a correlation between two variables in the population. We use the Greek letter ρ in order to give the population correlation. If x and y are both normally distributed, then ρ = 0 means that x and y are not linearly correlated, or that x is independent of y. The alternative hypothesis can be one-sided or two-sided. To compute the *t-statistic*, we use the folowing formula:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where r is the sample correlation and n is the sample size.

The observed t-test is the same as the t test we would find when we tested the hypothesis $\beta_1 = 0$. This means that if there is no correlation in the population, that the regression coefficient is 0.