Chapter 11: Multiple Regression

11.1: Multiple Regression

In many cases, variable y is influenced by a number of explanatory variables. For example, suppose you want to predict scores on a math test. In that case, you can look at different variables: IQ, motivation and attitude.

The simple linear regression model assumes that the average response variable y depends on x. The corresponding formula is: $\mu_y = \beta_0 + \beta_1 x$. However, if we want to account for more variables, we can use this formula:

- $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.

This is called the *population regression* formula.

Multiple Linear Regression Model

We combine the regression line for the population and assumptions about the variance in order to create a multiple linear regression model. The subpopulation means describe the "fit" portion of the model. The residuals cover the variance, which cannot be explained on the basis of the model. We also use the symbol ε here when we talk about how far an individual observation is different from the mean of the subpopulation. These deviations are normally distributed with a mean of 0 and an unknown standard deviation that does not depend on the values of x. These are assumptions that we can verify by looking at the residuals.

- The statistical model of multiple linear regression is: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$.
- The mean response, μ_y, is a linear function of all explanatory variables: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.
- The deviations (ε_i) are Normally distributed with a mean of 0 and a standard deviation σ. Briefly summarised, N(0,σ). The parameters of the model are thus $\beta_0 + \beta_1$, $\beta_2, \ldots, \beta_p$ and σ.

Estimating Parameters with Multiple Regression

- As with simple linear regression we create in the estimation of parameters (β) use of sample values (b).
- $b_0, b_1, b_2, \ldots, b_p$ are used to estimate $\beta_0, \beta_1, \beta_2, \ldots \beta_p$
- For the i^{th} observation, the predicted response $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$. The i^{th} residual (e_i) is the difference between the observed response and the predicted response: $y_i - \hat{y}_i$. This is the same as: $y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip}$.
- The sum of squared residuals can be found by using: $\Sigma(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip})^2$.
- The parameter σ^2 is estimated using s^2. We find s^2 by using the following: $\Sigma e^2_i / n-p-1$. In this formula, n = sample size and p = the number of explanatory variables
- The standard deviation (σ) can be found by finding the root of s^2.

Confidence Intervals for Multiple Regression

We can calculate confidence intervals and perform significance tests for the regression coefficients of all the variables (β_j).

- The confidence interval for β_j is $b_j \pm t^* SE_{bj}$, where SE_{bj} is the standard error of b_j and t

* is the value of t (n - p - 1)

- To test the hypothesis $b_j = 0$, we calculate a t-test: $t = b_j / SE_{bj}$.
- The alternative hypothesis can be either one-sided or double-sided.

ANOVA-Table for Multiple Regression

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

Significance Testing for Regression Coefficients in Multiple Regression

- In case of multiple regression, we can test the null hypothesis which states that regression $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. The null hypothesis actually says so that none of the x-variables are explanatory variables of the y-variable.
- The alternative hypothesis is that at least one of the regression coefficients (βj) is not 0. This hypothesis is actually saying that at least one of the x-variables is a predictor of the y-variable.
- The F-value is found as follows: MSM / MSE. If the null hypothesis is true, then F has the F (p, np-1) distribution.
- Finally, we can calculate how much variance in y is explained by all the explanatory variables together: R^2 = SSM / SST.