Multiple regression

Table of contents

1 When to use it?

Multiple regression, and multiple correlation, are tool used to examine the combined relations between multiple predictors and a dependent variable.

When correlating one quantitative dependent variable (Yi) with multiple (more than one) quantitative variable ($Xi_1, Xi_2, ...Xi_k$). it is specially useful when several independent variables work better at predicting a dependent variable compared to a single independent variable. For example, one can estimate percent body fat using the skinfold measurement technique. Although there are several techniques to estimate body site, researchers have used multiple regression techniques to identify the best predictors. See excerpt below:

> The practical equation including age, race, height, weight and waist circumference had high predictive ability for lean body mass (men: R 2=0 · 91, standard error of estimate (SEE)=2 · 6 kg; women: R 2=0 · 85, SEE=2 · 4 kg) and fat mass (men: R 2=0 · 90, SEE=2 · 6 kg; women: R 2=0 · 93, SEE=2 · 4 kg). Waist circumference was a strong predictor in men only. Addition of other circumference and skinfold measures slightly improved the prediction model. @lee2017

2 Variables

2.1 Independent variables

One or more quantitative (interval or ratio) variables.

2.2 Dependent variable

One quantitative of interval or ratio level variable.

3 Stating the Hypotheses

Null hypothesis

F test for the complete regression model

H_0 : the variance explained by all the independent variables together (the complete model) is 0 in the population

$H_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$

t test for the individual regression coefficient β_k

$H_0 : \beta_k = 0$

Alternative hypothesis

F test for the complete regression model

H_a : not all population regression coefficients are 0

$H_a : \beta_1 \neq \beta_2 \neq ... \neq \beta_K \neq 0$

t test for the individual regression coefficient β_k

$H_a : \beta_k \neq 0$ (two sided)

$H_a : \beta_k > 0$ (right sided)

$H_0 : \beta_k < 0$ (left sided)

The multiple linear regression best fit line has an equation of the following form:

$$\hat{y} = a + \beta(x)$$

where, a is the intercept (the value of y when x is $= 0$), β represents the slope of the line and x is the explanatory variable (the predictor).

---

4 Assumptions

Adapted from Navarro and Foxcroft (2019) @navarro2022

1. **Normality**. It assumes that the residuals are normally distributed. It's actually okay if the predictors X and the outcome Y are non-normal, so long as the residuals are normal.
2. Linearity. Assumes that the relationship between X and Y is linear (both simple regression or a multiple regression).
3. **Homogeneity of variance**. The regression model assumes that each residual is comes from a normal distribution with mean 0, and with a standard deviation that is the same for every single residual. In practice, it's impossible to test the assumption that every residual is identically distributed. Instead, what we care about is that the standard deviation of the residual is the same for all values of $\hat{Y}$, and all values of every predictor Xi in the model.
4. **Independence**. The residuals are independent of one another.

Also, pay attention to the following:

1. **Uncorrelated predictors**. In a multiple regression model, you don't want your predictors to be too strongly correlated with each other. Predictors that are too strongly correlated with each other (referred to as "collinearity") can cause problems when evaluating the model.

2. **No “bad” outliers**. There is an implicit assumption that your regression model isn’t being too strongly influenced by one or two anomalous data points. This raises questions about the adequacy of the model and the trustworthiness of the data in some cases.

5 Test statistic

F test for the complete regression model. Refer to the One-Way ANOVA test.

t test for individual β_k

$$t = \frac{b_k}{SE_{b_k}}$$

For one independent variable:

$$SE_{b_1} = \frac{\sqrt{\sum(y_j - \hat{y}_j)^2/(N-2)}}{\sqrt{\sum(x_j - \bar{x})^2}} = \frac{s}{\sqrt{\sum(x_j - \bar{x})^2}}$$

with s the sample standard deviation of the residuals, x_j the score of subject j on the independent variable x , and $\bar{x}$ the mean of x.

6 Sampling distributions

Refer to the sampling distributions of the F test (One-Way ANOVA) and the t test (Independent-Samples t test).

7 Significance

Refer to the steps used for the F test (One-Way ANOVA) and the t test (Independent-Samples t test).

8 Confidence Intervals

Refer to the StatKat website[1] for a detailed explanation. I will show below how to calculate it using `jamovi`.

9 Effect size

For linear regression we calculate R^2 as the effect size. This is the amount of variance in the dependent variable y that is explained by the sample regression equation (the independent variable(s)). The reader is referred to StatKat[2] for more detailed information.

10 Example

Data: parenthood data found under lsj-data library in jamovi. Notice that you may need to install the lsj-data module @datalabcc2018 in jamovi.

Below is the formula for a straight line when there are more than one independent (predictor) variable:

$$y = a + b_1 Xi1 + b_2 Xi2 + b_j Xi$$

Where, a is the intercept, b is the slope, and x is a given value. The intercept is where the line touches the y axis, which in the graph in the left is between 80 and 90. This is the expected value of Yi when Xi is equal to 0. The slop is the tilt of the best fit line.

To run the linear regression, click on `Regression - Linear Regression` analysis in jamovi, using the `parenthood` data set.

Then specify `dani.grump` as the **`Dependent Variable`**, `dani.sleep` , and `baby.sleep` as the variables entered in the `Covariates` box. This gives the results shown above.

Predictor	Estimate
Intercept	125.966
`dani.sleep`	-8.950
`baby.sleep`	0.011

With these values, we can create the linear regression equation:

[1] Confidence interval for linear regression - https://bit.ly/3rVJOUp

[2] Regression effect size on StatKat: https://bit.ly/3rVJOUp

$$\hat{Y} = 125.96 + (-8.94)X$$

10.1 Interpretation:

The slope: if one increases Xi by 1 unit, then one is decreasing Yi by 8.94. In other words, for each additional hour of sleep, Dani reduce her grumpiness level (points), which in turn will improve her mood.

The intercept: recall that the a is the predicted value of Yi when Xi is equal to 0. Thus, if Dani gets zero hours of sleep ($Xi = 0$), then her grumpiness will reach about ($Yi = 125.96$), which is lot since the scale goes up to 100.

10.2 Assumption Checks

Normality: QQ-plots + Shapiro-Wilk test

Outlier

Interpretation: Values greater than 1 is often considered large and indicates the presence of outlier.

11 Web Resources

I have created a list of additional resources on this topic that can be accessed by scanning the following QR code:

add qr code here

References