

Multiple Regression

A simple linear regression model is a summary of the relationship between a **dependent variable** (or **response variable**) Y and an **independent variable** (or **covariate variable**) X .

Y is assumed to be a random variable while, even if X is a random variable, we condition on it (assume it is fixed). Essentially, we are interested in knowing the behavior of Y given we know $X = x$.

A multiple regression model is a summary of the relationship between a **dependent variable** (or **response variable**) Y and multiple **independent variables** (or **covariates**) X .

Y is assumed to be a random variable while, even if the X_i are random variables, we again consider them to be fixed. We are interested in knowing the behavior of Y given we know

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n.$$

In simple linear regression, the population regression line was given by

$$E[y|x] = \mu_{y|x} = \alpha + \beta x$$

and the full linear regression model was

$$y = \alpha + \beta x + \epsilon,$$

where ϵ is a normally distributed random variable with mean 0 and variance σ^2 .

In multiple regression, the population regression line is given by

$$E[y|x_1, x_2, \dots, x_q] = \mu_{y|x_1, x_2, \dots, x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$$

and the full linear regression model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q + \epsilon,$$

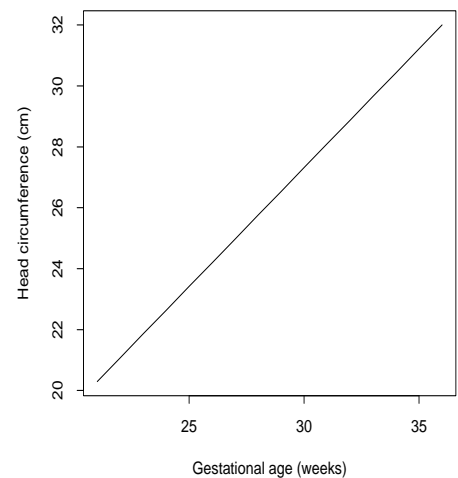
where again ϵ is a normally distributed random variable with mean 0 and variance σ^2 .

Multiple Linear Regression - Summary of Assumptions:

- Values of x_1, x_2, \dots, x_q are fixed.
- The outcomes of y are normally distributed (independent) random variables with mean $\mu_{y|x_1, x_2, \dots, x_q}$ and variance $\sigma^2_{y|x_1, x_2, \dots, x_q}$.
- $\sigma^2_{y|x_1, x_2, \dots, x_q}$ is the same for all sets of x_1, x_2, \dots, x_q values.
- The relationship between $\mu_{y|x_1, x_2, \dots, x_q}$ and x_1, x_2, \dots, x_q is described by

$$\mu_{y|x_1, x_2, \dots, x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$$

Consider an example of head circumference and gestational age for 100 infants. A regression line was fit to the data. The line is shown below.



We might want to know if head circumference also depends on birthweight. Of course, age and birthweight are related. We can let both age and birthweight be explanatory variables (covariates) in our model which predicts head circumference.

$$hc = \alpha + \beta_1 \text{age} + \beta_2 \text{bw} + \text{variation}$$

Written mathematically, this is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Using the method of least squares, we can get least square estimates ($\hat{\alpha} = 8.308$, $\hat{\beta}_1 = 0.4487$, and $\hat{\beta}_2 = 0.0047$):

$$\hat{y} = 8.3080 + 0.4487x_1 + 0.0047x_2$$

The value of R^2 with two explanatory variables is 0.7520. Recall that with just age in the model $R^2 = 0.6095$. We see that the new model (with both age and weight) better accounts for the variance in head circumference.

Something to be careful about...adding additional explanatory variables to the model will NEVER cause R^2 to decrease. Thus, we need to know if adding this variable was a good idea. In other words, does including weight really help predict (or explain some part of) head circumference.

Consider the estimated coefficient corresponding to gestational age (0.4487). The interpretation of this coefficient is that: given that a child's birth weight remains constant, each one-week increase in age corresponds to a 0.4487 unit (cm) increase in head circumference. More intuitively, given two infants with the same birth weight that differ in gestational age by one week (infant 1 is one week younger than infant 2), we would expect the head circumference of infant 2 to be 0.4487 cm larger.

Similarly, since the coefficient of birth weight is 0.0047, given two infants with the same gestational age that differ in birth weight by 1 unit (infant 1 is one gram heavier than infant 2), we would expect the head circumference of infant 2 to be 0.0047 cm larger.

We can test hypotheses about the coefficients, using methods similar to those in simple linear regression.

Recall that to test $H_0 : \beta = \beta_0$ against $H_A : \beta \neq \beta_0$, we consider

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$$

If the null is true, then t follows a t distribution with $n - 2$ degrees of freedom.

Similar formulas hold in the test $H_0 : \alpha = \alpha_0$ against $H_A : \alpha \neq \alpha_0$. Consider

$$t = \frac{\hat{\alpha} - \alpha_0}{se(\hat{\alpha})}$$

If the null is true, then t follows a t distribution with $n - 2$ degrees of freedom.

To test $H_0 : \beta_i = \beta_{i0}$ against $H_A : \beta_i \neq \beta_{i0}$, we still consider

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{\widehat{se}(\hat{\beta}_i)}$$

but in multiple regression, if the null is true, t as defined above follows a t distribution with $n - q - 1$ degrees of freedom, where q is the number of explanatory variables in the model.

For the head circumference example, it can be shown that

$$\widehat{se}(\hat{\beta}_1) = 0.0672 \text{ and } \widehat{se}(\hat{\beta}_2) = 0.00063.$$

Test $H_0 : \beta_1 = 0$ against the alternative of inequality; and test $\beta_2 = 0$ against the alternative of inequality.

$$\begin{aligned} t_1 &= \frac{\hat{\beta}_1 - \beta_{10}}{\widehat{se}(\hat{\beta}_1)} \\ &= \frac{0.4487 - 0}{0.0672} \\ &= 6.68 \end{aligned}$$

and

$$\begin{aligned} t_2 &= \frac{\hat{\beta}_2 - \beta_{20}}{\widehat{se}(\hat{\beta}_2)} \\ &= \frac{0.0047 - 0}{0.00063} \\ &= 7.47 \end{aligned}$$

Each of these values should be compared with values of a t -distribution with $100 - 2 - 1 = 97$ degrees of freedom. In each case, we'd reject the null.

Consider these questions:

Recall that R^2 went from 0.6095 to 0.7520 when we added an additional explanatory variable. Can we really compare these values, since adding additional explanatory variables always increases R^2 ?

If we did not reject the null $H_0 : \beta_i = 0$ for some β_i , should we leave that β_i in the model?

Consider a modification of the head circumference example. Suppose in addition to gestational age and birth weight, we also have information on whether or not the mother had toxemia during pregnancy. Forget about weight for the moment, and consider the effect of both age and toxemia on head circumference.

The model is

$$y = \alpha + \beta_1 x_1 + \beta_3 x_3 + \epsilon$$

where x_1 denotes age and x_3 denotes presence of toxemia (yes or no). Note that x_2 is skipped since x_2 was just used to represent birth weight.

The fitted model is $\hat{y} = 1.4956 + 0.8740x_1 - 1.4123x_3$

We could test hypotheses to confirm that the β_1 and β_3 are in fact significantly different from zero.

We could also think about separating the data into two groups according to toxemia status.

We could also consider models of interaction.

Model Selection

Until now, we have defined a particular model using some collection of chosen explanatory variables and then fit that model to data using the method of least squares. An important question is “What variables should be included in the model ?”

Idea: Define some measure of goodness of fit and add variables if they improve this measure, delete variables if they do not improve this measure.