STAT2203 STAT7203

Probability models and data analysis

Lecture 10.2

Multiple Regression

Multiple linear regression

A linear regression model that contains more than one explanatory variable is called a *multiple linear regression model*.

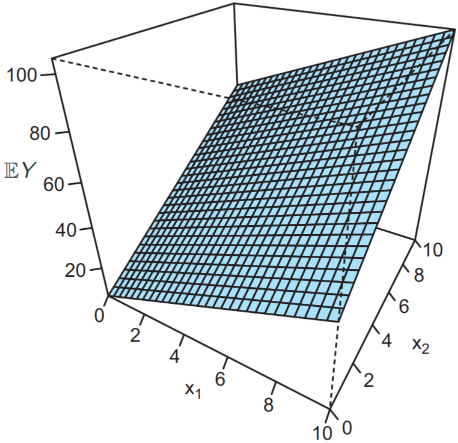
In a multiple linear regression model the response variable Y depends on $(p-1)$ explanatory variables, $x_1, \dots, x_{p-1}$, via the linear relationship

$$Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \ldots + \beta_{p-1}^\star x_{p-1} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

The regression coefficient β_i is the expected change in Y associated with a 1-unit increase in x_i while all other explanatory variables, i.e., $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, are held fixed.

Multiple linear regression



$$\mathbb{E}Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2$$

Models with Interactions

If an investigator has obtained observations on y, x_1, and x_2, one possible model is $Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \varepsilon$. However, other models can be constructed by forming new predictors that are mathematical functions of other predictors variables. For example, with $x_3 = x_1^2$ and $x_4 = x_1 x_2$, the model

$$Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \beta_3^\star x_3 + \beta_4^\star x_4 + \varepsilon,$$

is still a multiple linear regression model.

In general, it is not only permissible for some predictors to be mathematical functions of others but also often highly desirable in the sense that the resulting model may be much more successful in explaining variation in y than any model without such predictors.

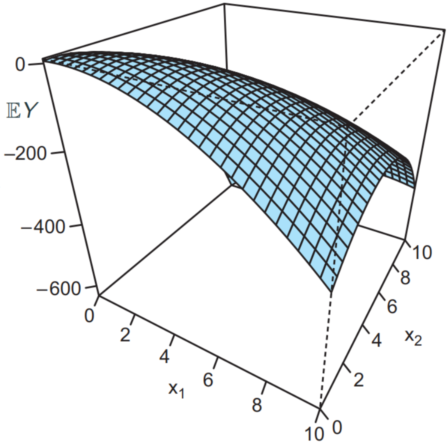
Models with Interactions

Example of such models include:

- First-order model without interaction:
  $Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \varepsilon$
- First-order model with interaction:
  $Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \beta_3^\star x_1 x_2 + \varepsilon$
- Second-order model without interaction:
  $Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \beta_3^\star x_1^2 + \beta_4^\star x_2^2 + \varepsilon$
- Second-order model with interaction:
  $Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \beta_3^\star x_1^2 + \beta_4^\star x_2^2 + \beta_5^\star x_1 x_2 + \varepsilon$

In general, the presence of interaction terms in the model implies that the expected change in Y depends not only on the variable being changed but also on the values of some of the fixed variables.

Estimation for Multiple Linear Regression



$$\mathbb{E}Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \beta_3^\star x_1^2 + \beta_4^\star x_2^2 + \beta_5^\star x_1 x_2$$

Multiple regression with categorical variables

Thus far we have explicitly considered only quantitative (numerical) predictor variables. Using simple numerical coding, qualitative (categorical) variables, can also be incorporated into a model.

A variable with just two possible categories can be represented as a "dummy" predictor variable, e.g., categories A and B:

$$x = \begin{cases} 1, & \text{if category A} \\ 0, & \text{otherwise} \end{cases}$$

Multiple regression with categorical variables

Suppose we want to model the travel time between locations in a certain city in terms of the distance between the locations as well as the types of vehicles, cars and trucks. One possible model is

$$Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \varepsilon,$$

where x_1 is the distance (quantitative) and x_2 is dummy/indicator variable indicating the type of car

$$x_2 = \begin{cases} 1, & \text{if vehicle is a truck} \\ 0, & \text{otherwise} \end{cases}$$

Multiple regression with categorical variables

The mean value of travel time depends on whether a vehicle is a car or a truck:

$$
\begin{aligned}
E(Y) &= \beta_0^\star + \beta_1^\star x_1, \quad \text{when a car is used,} \\
E(Y) &= \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star, \quad \text{when a truck is used.}
\end{aligned}
$$

The coefficient $\beta_2^\star$ is the difference in mean travel times between trucks and cars with distance held fixed; if $\beta_2^\star > 0$, on average it will take trucks longer to traverse any particular distance than it will for cars.

We can also consider interactions.

Multiple regression with categorical variables

A variable with more than two categories is represented with more dummy variables, e.g., categories A, B, and C

$$x_1 = \begin{cases} 1, & \text{if category A} \\ 0, & \text{otherwise} \end{cases}, \quad x_2 = \begin{cases} 1, & \text{if category B} \\ 0, & \text{otherwise} \end{cases}$$

More generally, incorporating a categorical variable with c possible categories into a multiple regression model requires the use of $c - 1$ indicator variables.

Multiple regression with categorical variables

Question: why not coding as below:

$$x = \begin{cases} 2, & \text{if category A} \\ 1, & \text{if category B} \\ 0, & \text{otherwise} \end{cases}$$

Answer: Because it imposes an ordering on the categories that may not be necessarily implied by the problem context.

Estimation for Multiple Linear Regression

Suppose we have the data as

$$\begin{gathered}(x_{1,1}, x_{1,2}, \dots, x_{1,p-1}, y_1),\\ (x_{2,1}, x_{2,2}, \dots, x_{2,p-1}, y_2),\\ \vdots\\ (x_{n,1}, x_{n,2}, \dots, x_{n,p-1}, y_n).\end{gathered}$$

As in the case of simple regression, we look to find a hyperplane that best "fits" the data. For this, we can estimate $\beta_i^\star$'s in such a way the sum of squared errors (SSE) is minimized:

$$\min_{\beta_0,\beta_1,\dots,\beta_{p-1}} \sum_{i=1}^{n} (y_i - \overbrace{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1}}^{\hat{y}_i})^2.$$

Estimation for Multiple Linear Regression

Using a bit of advanced linear algebra/calculus, it can be shown that the estimates of $\beta_i^\star$'s are given by the solution to the following linear system, often known as the "normal equation"

$$\mathbf{X}^\intercal \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\intercal \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & \dots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The least squares estimate of σ^2 can be obtained by

$$\hat{\sigma}^2 = \text{MSE} = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{\text{SSE}}{n-p}.$$

Inference for on predictor variable

It can be shown that $\hat{\beta}_i \sim \mathcal{N}\left(\beta_i^\star, h_i(\sigma, \text{“data”})\right)$. The form of $h_i(\sigma, \text{“data”})$ is complicated[1], but nonetheless as before we get

$$\frac{\hat{\beta}_i - \beta_i^\star}{\text{se}(\hat{\beta}_i)} \sim t_{n-p}, \quad i = 0, 1, \ldots, p-1.$$

Software packages calculate $\text{se}(\hat{\beta}_i)$, which we use to construct CI for $\beta_i^\star$ or test hypotheses[2], e.g., $H_0 : \beta_i^\star = 0,$ against $H_1 : \beta_i^\star \neq 0.$

[1] $h_i(\sigma, \text{“data”})$ is $\sigma^2 \times i^{\text{th}}$diagonal element of $(\mathbf{X}^\intercal \mathbf{X})^{-1}$.

[2] $H_0 : \beta_i^\star = 0$ means that as long as other predictors are in the model, the predictor x_i provides no additional useful information about y. $H_1 : \beta_i^\star \neq 0$ means that the predictor x_i does provide useful information in addition to what is contained in the other predictors.

Hypothesis testing on group of predictors

In some situations, one may wish to know whether any of the predictors in a particular subset provide useful information about the response variable Y.

For example, a model to be used to predict students' test scores might include a group of variables such as family income, education levels, class size and spending per pupil. One interesting hypothesis is that the school characteristic predictors, i.e., class size and spending per pupil, are relevant in predicting students' performance.

Hypothesis testing on group of predictors

Consider the multiple linear regression model

$$Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \ldots + \beta_{p-1}^\star x_{p-1} + \varepsilon.$$

Suppose for some $1 \leq \ell \leq p-1$, we are aiming to test

$$\begin{cases} H_0: & \beta_\ell = \beta_{\ell+1} = \ldots = \beta_{p-1} = 0, \\ \\ H_1: & \text{at least of } \beta_\ell, \beta_{\ell+1}, \ldots, \text{ or } \beta_{p-1} \text{ is nonzero.} \end{cases}$$

Under the null hypothesis, our model reduces to

$$Y = \beta_0^\star + \beta_1^\star x_1 + \beta_2^\star x_2 + \ldots + \beta_{\ell-1}^\star x_{\ell-1} + \varepsilon.$$

Hypothesis testing on group of predictors

The test is carried out by fitting both the full and reduced models. Because the full model contains not only the predictors of the reduced model but also some extra predictors, it should fit the data at least as well as the reduced model, i.e., $\mathrm{SSE}_{\mathrm{full}} \leq \mathrm{SSE}_{\mathrm{reduced}}$.

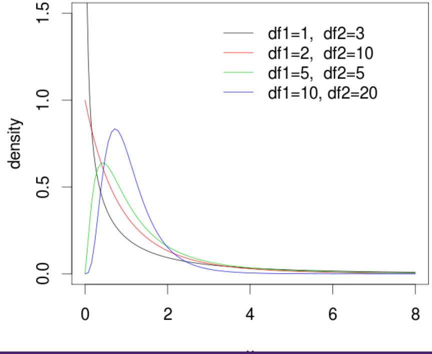
Intuitively, if $\mathrm{SSE}_{\mathrm{full}}$ is much smaller than $\mathrm{SSE}_{\mathrm{reduced}}$, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on relative reduction of SSE. Under H_0 and the homoscedastic normal noise model, we have

$$\frac{\left(\mathrm{SSE}_{\mathrm{reduced}} - \mathrm{SSE}_{\mathrm{full}}\right)/(p-\ell)}{\mathrm{SSE}_{\mathrm{full}}/(n-p)} \sim F_{p-\ell,n-p}.$$

The test is upper-tailed, that is the P-value is $\mathbb{P}(F_{p-\ell,n-p} \geq f)$, where f is the observed value of the **F-statistic**.

F distribution

This F-distribution is named after R.A. Fisher — one of the founders of modern statistics. Like Student's t distribution, this distribution is a family of distributions, this time depending on two parameters (called, as usual, degrees of freedom), $F_{\text{df1},\text{df2}}$.



Blood alcohol concentration and reaction times

A study explored the effects of increased blood alcohol concentration on reaction time. Participants were randomly selected and given 60 mL of vodka.

The blood alcohol concentration (BAC; g/dL) of participants was measured after 15 minutes and each participant then did a ruler test to measure their reaction distance (cm). The distances were then converted into reaction times (ms). The sex (male/female) of the participants was also recorded.