4 Multiple Linear Regression Model

4.1 Other explanatory variables

Whenever we fit a simple linear regression model there will be some amount of variation in the y_i that is not explained by the regression (that part of the R^2 less than 100%). Part of this remaining variation might be other explanatory variables. A multiple linear regression model is one that seeks to take into account more than one explanatory variable.

If we have 2 explanatory variables X_1 and X_2 and a response variable Y we can write the multiple linear regression model as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

For $i = 1, 2, \ldots, n$ observations of the form (x_{1i} , x_{2i} , y_i)

More generally we can have a multiple linear regression model with $p - 1$ explanatory variables $X_1, X_2, \ldots, X_{p-1}$ which we can write either as

$$E[y_i] = \mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1\,i}$$

$var(y_i) = \sigma^2$ for all $i = 1, \ldots, n$

$cov(y_i, y_j) = 0$ for all $i \neq j$

Or alternatively and equivalently as,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1\,i} + \varepsilon_i$$

$var(\varepsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$

$cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$

And we usually have the additional assumption of normality which can be written as either $y_i \sim N(\mu_i, \sigma^2)$ or as $\varepsilon_i \sim N(0, \sigma^2)$

We can also write the multiple linear regression model in matrix form. This is

$$\mathrm{Y} = \mathrm{X}\,\boldsymbol{\beta} + \varepsilon$$

where,

$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ the vector of responses

$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ the *design matrix*

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ the vector of parameters which are unknowns

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ the vector of random errors

4.2 Least Squares estimation in the multiple regression model

Algebraically we will find it easiest to work with the matrix form to derive the least squares estimates for **β** and then we will find that the results are the same as those found for the simple linear regression model in section 3 above.

Once again to find the least squares estimators we minimise the sum of squares of residuals that is

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1\,i}))^{\,2}$$

or

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i{}^2$$

$$S(\boldsymbol{\beta}) = \varepsilon^T \varepsilon$$

The least squares estimator $\widehat{\boldsymbol{\beta}}$ of the vector of unknown parameters $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y}$$

This is the same result as in section 3 above except that this time the identity matrix ***X*** has *p* columns for p – 1 explanatory variables whereas before it had 2 columns.

From the work we have already done on the simple linear regression model we also know that:

- $\widehat{\boldsymbol{\beta}}$ is an *unbiased estimator* for $\boldsymbol{\beta}$
- Var[$\widehat{\boldsymbol{\beta}}$] = $\sigma^2\ (\boldsymbol{X^T X})^{-1}$
- If Y = X **β** + ε with $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I})$ then $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X^T X}\right)^{-1})$

In finding the vector of fitted values $\widehat{\boldsymbol{Y}}$ we can use the *hat matrix* ***H*** where

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} = \boldsymbol{HY}$$

So

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-\mathbf{1}} \boldsymbol{X^T}$$

And recall from section 3 that $\boldsymbol{H}^T$ = ***H*** and ***HH*** = ***H***, the property of an idempotent matrix.

With the hat matrix we can now look at the residual vector ***e***

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{HY} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

Then

$$E[\boldsymbol{e}] = 0$$

Which we can show by:

$$E[\boldsymbol{e}] = (\boldsymbol{I} - \boldsymbol{H})E(\boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\ )E[\boldsymbol{Y}] = \ (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\ )\boldsymbol{X\beta} = \ \boldsymbol{X\beta} - \ \boldsymbol{X\beta}$$

Also

$$var(\boldsymbol{e}) = \ \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

Which we can show by:

$$var(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})var(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})^T = \ \sigma^2(\boldsymbol{I} - \boldsymbol{H})^{\,2} = \ \sigma^2(\boldsymbol{I} - 2\boldsymbol{H} - \ \boldsymbol{H}^2) = \ \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

The sum of all the elements in ***e*** is zero which is the same as the $\sum e_i = 0$ result we had before in section 2.

The sum of squares of residuals in matrix form is $\boldsymbol{e}^{\boldsymbol{T}}\boldsymbol{e}$ and

$$\boldsymbol{e}^{\boldsymbol{T}}\boldsymbol{e} = \ \boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y}$$

4.3 Analysis of Variance

The analysis of variance identity can be used for multiple linear regression and for regression in matrix form in the same way that it was for simple linear regression. That is,

Total sum of squares = Regression sum of squares + Residual sum of squares

$SS_T = SS_R + SS_E$

In matrix form the total sum of squares is

$$SS_T = \sum (Y_i - \bar{Y})^2 = \ \boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y} - n\bar{Y}^2$$

And the regression sum of squares is

$$SS_R = \sum (\hat{Y}_i - \bar{Y})^2 = \ \boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{H}\,\boldsymbol{Y} - \ n\bar{Y}^2$$

We have already seen that the residual sum of squares can be written as

$$SS_E = \sum (Y_i - \hat{Y}_i)^2 = \ \boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y}$$

It is possible to combine these to show the analysis of variance identity in matrix form and for multiple linear regression as we previously did with the simple linear regression model.

We can also produce an ANOVA table for a multiple linear regression with *n* observations and *p* – 1 explanatory variables and hence *p* parameters estimated (β_0 , $\beta_1 \ldots \beta_{p-1}$)

The ANOVA table is again in the format we have seen before

	d.f.	SS	MS	VR
Regression				
Residuals				
Total				

Where now the Regression row represents the multiple linear regression.

Now the degrees of freedom are:

- *n* – *p* for residuals (this is the general case of *n* – 2 when *p* = 2 in the simple linear regression model before)
- *p* – 1 for regression (this is the general case of 1 when *p* = 2 in the simple linear regression model before)
- *n* – 1 in total (as before)

We have already given the formulae for sums of squares. Mean squares are then those sums of squares divided by their respective degrees of freedom.

$$MS_R = \frac{SS_R}{p-1}$$

$$MS_E = \frac{SS_E}{n-p} = S^2$$

And once again $MS_E = S^2$ is an unbiased estimator for σ^2

Then the variance ratio or F statistic becomes

$$VR = \frac{MS_R}{MS_E} = \frac{\frac{SS_R}{p-1}}{S^2}$$

4.4 Overall test of significance of a multiple regression

We can use the Variance Ratio in the multiple regression ANOVA table to test whether the overall multiple regression has significance compared to a "null model" of a constant β_0 plus some random variation ε_i.

Our null hypothesis is

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

And the alternative hypothesis is that at least one of β_1, β_2, … β_{p-1} is not zero.

Our F-statistic, sometimes written *F** is the variance ratio in the ANOVA table

$$F^* = \frac{\frac{SS_R}{p-1}}{\frac{SS_E}{n-\ p}} = \frac{\frac{SS_R}{p-1}}{S^2}$$

Where the denominator is always an unbiased estimator of σ^2 but the numerator is only an unbiased estimator of σ^2 if the multiple regression assumptions (linear relationships, constant variance and normal distribution) are true.

Under H_0 we will have $F^* \approx 1$ so large values of F^* are required to reject H_0 (which is what we generally seek to do as we would like a model that has significance).

The F-test here compares F^* with the critical value of the Fisher's-F distribution on *p* – 1 and *n* – *p* degrees of freedom where we reject H_0 at 100(1 – α)% significance if $F^* > F_{n-p}^{p-1}(\alpha)$.

4.5 Inference about parameters in multiple regression models

We already have the distribution of the least squares estimators of the *p* model parameters

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2\ (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1})$$

So if we want the jth parameter estimator $\widehat{\beta_J}$ where j = 0, 1, …, *p* – 1, then

$\widehat{\beta_J} \sim N(\beta_j\ , \sigma^2\ c_{jj})$ where c_{jj} is the jth diagonal element of $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}$ where we count the diagonal elements 0, 1, …, p – 1 (i.e. the first diagonal element relates to β_0, the second one to β_1, and the last one to β_{p-1}.

In this way we can make inference about β_j in the ways in which we did for β_1 in the simple linear regression model earlier. These are:

- Confidence intervals for β_j
- Tests of hypotheses with H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$

In line with the parameter confidence intervals we constructed in the simple linear model, our 100(1 – α)% confidence interval for β_j is

$$[a, b] = \widehat{\beta_J}\ \pm\ t_{n-p}(\alpha)\sqrt{S^2 c_{jj}}$$

The test statistic for H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$ is *T* where,

$$T = \frac{\widehat{\beta_J}}{\sqrt{S^2 c_{jj}}} \sim t_{n-p} \text{ under } H_0$$

We need to be very careful about the interpretation of these confidence intervals and tests of hypotheses. They only apply within the context of the whole *p* parameter model that is being fitted.

For example if we cannot reject $H_0: \beta_j = 0$ then:

- This does <u>not</u> mean that *X_j* has no explanatory power, rather that it has no <u>additional</u> explanatory power compared to the *p* – 1 parameter model that had all of the other betas apart from β_j
- Also this does not tell us about the model $y_i = \beta_0 + \beta_j x_{ji} + \varepsilon_i$ compared to the "null" model $y_i = \beta_0 + \varepsilon_i$, rather it tells us about the role of β_j within the whole *p* parameter model.

4.6 Confidence Intervals for μ in multiple regression

We might want to estimate the mean response, μ at a certain value of $\boldsymbol{x}$.

We already know that $\hat{\boldsymbol{\mu}} = \widehat{\boldsymbol{E}[\boldsymbol{Y}]} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$

Now say we want to estimate μ_0 at $x_0 = (1,\ x_{1,0}\ \dots\ x_{p-1,0})^T$ where

$\mu_0 = E[Y|X_1 = x_{1,0}\ \dots\ X_{p-1} = x_{p-1,0}]$

Our point estimate is

$$\hat{\mu}_0 = x_0^T\hat{\boldsymbol{\beta}}$$

With a multiple linear regression model that includes the assumption of a normal distribution we can develop a confidence interval for μ_0.

Now, $\hat{\mu}_0 = x_0^T\hat{\boldsymbol{\beta}}$ is a linear combination of the components of $\hat{\boldsymbol{\beta}}$ all of which are normally distributed therefore $\hat{\mu}_0$ must also be normal.

$E[\hat{\mu}_0] = E\left[x_0^T\hat{\boldsymbol{\beta}}\ \right] = x_0^T\boldsymbol{\beta} = \mu_0$ and

$var[\hat{\mu}_0] = var\left[x_0^T\hat{\boldsymbol{\beta}}\ \right] = x_0^T var(\hat{\boldsymbol{\beta}})x_0 = \sigma^2 x_0^T\ (\boldsymbol{X^T X})^{-1}x_0$

And putting all these together we have

$\hat{\mu}_0 \sim N(\mu_0, \sigma^2 x_0^T\ (\boldsymbol{X^T X})^{-1}x_0\ )$ from which it is straightforward to develop a 100(1 – α)% confidence interval for μ_0 which is

$$[a,b] = \hat{\mu}_0\ \pm\ t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{S^2 x_0^T\ (\boldsymbol{X^T X})^{-1}x_0}$$

4.7 Prediction Intervals in multiple regression

Now say we have a new set of $\boldsymbol{x}$ observations $x_0 = (1,\ x_{1,0}\ \dots\ x_{p-1,0})^T$ but we do not yet have the corresponding observation for the response *y₀*. When we predict *y₀* with a prediction interval we will need to take into account the random variation that comes with a new observation.

Our point estimate for *y₀* is $\hat{\mu}_0$ which is the same as $\hat{y}_0$

With our Normal distribution assumption for the *y*

$$\hat{y}_0 \sim N(\mu_0, \sigma^2 x_0^T \ (\boldsymbol{X^T X})^{-1} x_0 \ )$$

$$\hat{y}_0 - \mu_0 \ \sim N(0, \sigma^2 x_0^T \ (\boldsymbol{X^T X})^{-1} x_0 \ )$$

$$\hat{y}_0 - (\mu_0 + \ \varepsilon_0) \ \sim N(0, \sigma^2 x_0^T \ (\boldsymbol{X^T X})^{-1} x_0 + \ \sigma^2)$$

So

$$\hat{y}_0 - y_0 \sim N(0, \sigma^2 (1 + x_0^T \ (\boldsymbol{X^T X})^{-1} x_0))$$

Standardising gives us

$$\frac{\hat{y}_0 - y_0}{\sqrt{\sigma^2 (1 + x_0^T \ (\boldsymbol{X^T X})^{-1} x_0)}} \sim N(0,1)$$

And replacing the unknown σ^2 with our estimate S^2 gives

$$\frac{\hat{y}_0 - y_0}{\sqrt{S^2 (1 + x_0^T \ (\boldsymbol{X^T X})^{-1} x_0)}} \sim t_{n-p}$$

Which allows us to develop the 100(1 – α)% prediction interval for y_0 which is

$$\hat{y}_0 \ \pm \ t_{n-p} \left(\frac{\alpha}{2}\right) \sqrt{S^2 (1 + x_0^T \ (\boldsymbol{X^T X})^{-1} x_0)}$$

5 Model building

In building a multiple regression model we have two objectives which seem to be in conflict:

- having a model that describes the data as well as possible
- having a model that is as simple as possible (the principle of parsimony)

Selecting a model – or a subset of the potential explanatory variables – that gives a suitable balance between these objectives can be more art than science. There is no one correct answer. The interaction between the explanatory variables makes this even more complex because a combination of say three explanatory variables may explain more, or demonstrate better modelling properties (normal distribution, constant variance) than any of the three explanatory variables when used in a simple linear regression.

So in this section we will look at a number of approaches to deciding which explanatory variables to keep in a multiple linear regression model.

5.1 Using the F test to delete variables

Let us say we have a multiple linear regression model with *p – 1* explanatory variables and *p* parameters. With an ANOVA table we can carry out a test of the overall model and see that not all of the β parameters are zero and hence the multiple linear regression model has some significance and some explanatory power. But perhaps we could delete some of the explanatory variables to leave a simpler model that still contains explanatory power.

We do this with a *Subset test*. We are looking to see whether the *p* parameter model could be reduced to a *q* parameter model (*q* < *p*).

We are looking to see whether we can keep $x_1, \dots, x_{q-1}$ but remove $x_q, \dots, x_{p-1}$. *Note that in practice we will not necessarily be keeping variables in number order. For example in a six variable, 7 parameter model where we look to remove 2 variables it is not necessarily the case that x_5 and x_6 are the variables to be deleted first, but rather the two that contribute least to model significance. We will cover how to identify which variables to consider for deletion later.*

More specifically we are interested in whether these variables under consideration for deletion significantly increase the sum of squares due to regression or significantly reduce the sum of squares due to residuals compared with the simpler model that does not include them. This is sometimes referred to as the "extra sum of squares principle". The idea is that we seek models that maximise the proportion of sums of squares that are due to regression and minimise the proportion due to residuals.

We seek the *extra sum of squares* due to $x_q, \dots, x_{p-1}$ given that $x_1, \dots, x_{q-1}$ are already in the model. This can be written $SS(x_q, \dots, x_{p-1} \mid x_1, \dots, x_{q-1})$

Extra SS = {Regression SS under the full model} – {Regression SS under the reduced model}

and

Extra SS = {Residual SS under the reduced model} – {Residual SS under the full model}

Let $\boldsymbol{\beta}_1^T = (\beta_0, \dots, \beta_{q-1})$ and $\boldsymbol{\beta}_2^T = (\beta_q, \dots, \beta_{p-1})$

so that $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$

that is we have split the parameter vector $\boldsymbol{\beta}$ into one vector for the reduced model with *q* parameters and another vector with the additional *p – q* parameters we are considering for deletion.

similarly we can split up the ***X*** matrix into ***X₁*** and ***X₂*** where ***X₁*** contains a columns of 1's and then *q – 1* columns with n observations for explanatory variables $x_1, \dots, x_{q-1}$ and ***X₂*** contains *p – q* columns with *n* observations for explanatory variables $x_q, \dots, x_{p-1}$.

Then the full model is

$$\mathrm{Y} = \mathrm{X}\,\boldsymbol{\beta} + \varepsilon$$

$$\mathrm{Y} = \boldsymbol{X_1}\,\boldsymbol{\beta_1} + \boldsymbol{X_2}\,\boldsymbol{\beta_2} + \varepsilon$$

and the reduced model is

$$\mathrm{Y} = \boldsymbol{X_1}\,\boldsymbol{\beta_1} + \varepsilon$$

We can calculate the SS_R and SS_E for both the full and the reduced models. We will call them:

SS_R^{Full} and SS_E^{Full}

SS_R^{Red} and SS_E^{Red}

these use the same formulae that we developed in the previous section for sums of squares under multiple linear regression models but with the appropriate vector $\boldsymbol{\beta}$ and matrix ***X*** for the full / reduced model.

Then extra sum of squares is

$SS_{extra} = SS_R^{Full} - SS_R^{Red} = SS_E^{Red} - SS_E^{Full} = \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{Y} - \widehat{\boldsymbol{\beta_1}}^{\boldsymbol{T}} \boldsymbol{X_1}^{\boldsymbol{T}} \boldsymbol{Y}$ in matrix form.

Once we have calculated the extra sum of squares we need to test whether that amount is significant or not. We do this with a test of hypotheses.

$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$

H_1: at least one of these parameters is not zero.

If we reject H_0 then there is evidence that at least some of the additional variables $x_q, \dots, x_{p-1}$ are significant and should be included in the model.

If we cannot reject H$_0$ then we should delete the variables $x_q, \dots, x_{p-1}$.

Under H$_0$ the test statistic *F** follows a Fisher F distribution

$$F^* = \frac{\left(\frac{SS_{extra}}{p-q}\right)}{s^2}$$

here s^2 is found from MS$_E$ in the full model

and under H$_0$ $F^* \sim F_{n-p}^{p-q}$

so we reject H$_0$ at α significance level if $F^* > F_{n-p}^{p-q}(\alpha)$

We may set out the calculation for this test in a particular form of ANOVA table.

Source	d.f.	SS	MS	VR = F*
$x_1, \dots, x_{q-1}$	q – 1	$SS(x_1, \dots, x_{q-1})$		
$x_q, \dots, x_{p-1} \| x_1, \dots, x_{q-1}$	p – q	SS_{extra}	$\frac{SS_{extra}}{p-q}$	$\frac{\left(\frac{SS_{extra}}{p-q}\right)}{s^2}$
Overall Regression	p – 1	SS_R		
Residual	n – p	SS_E	s^2	
Total	n – 1	SS_T		

There are two special cases where the F test can be replaced by a t test:

- where p – q = 1 so only one explanatory variable is being considered for deletion
- where there is a natural ordering of the explanatory variables X$_1$, X$_2$, X$_3$, … so that we naturally consider deleting them one at a time sequentially according to that order.

For deleting one explanatory variable (in this case we will consider deleting *X$_{p-1}$* but our one variable for deletion does not need to be the one with the highest subscript) our test statistic is

$$t = \frac{\hat{\beta}_{p-1}}{se(\hat{\beta}_{p-1})}$$

where $se(\hat{\beta}_{p-1})$ is the estimated standard error of the relevant beta parameter. The `summary()` function for a `lm()` linear model in R will include this standard error estimate in its output for each coefficient.

Under H_0: $\beta_{p-1} = 0$, this t statistic $t \sim t_{n-p}$ and we complete a two-sided test of *t* at our chosen level of significance. It can be shown that under the null hypothesis *F** = *t*^{2}.

Where there is a natural ordering of the *X$_i$* variables, we can perform a sequence of t tests to consider deletion of these variables in reverse order.