Multiple Linear Regression Models

CHRIS SUTTON, MARCH 2024

Modelling more complex relationships between variables

Simple Linear Regression → Multiple Linear Regression

The multiple linear regression model

model with $p - 1$ explanatory variables $X_1, X_2, \ldots, X_{p-1}$

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

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

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

$$\varepsilon_i \sim N(0, \sigma^2)$$

The multiple linear regression model

model with $p - 1$ explanatory variables $X_1, X_2, \ldots, X_{p-1}$

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

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

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

$$y_i \sim N(\mu_i, \sigma^2)$$

The multiple linear regression model

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

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p-1,1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1,n} & \dots & x_{p-1n} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Multiple linear regression in R

Response variable observations in vector `y`

If we have four explanatory variables with their observations in vectors `x1 x2 x3 x4`

To construct the multiple linear regression in an R object called `mlrm` (for example) and then display the results

```
mlrm <- lm(y ~ x1 + x2 + x3 + x4)
summary(mlrm)
```

To calculate the fitted values and store them as `yhat` and the standardised residuals and store them as `d`

```
yhat <- fitted(mlrm)
d <- rstandard(mlrm)
```

Residuals in multiple linear regression

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

With

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

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

The sum of the elements in $\boldsymbol{e}$ is zero as before

The sum of squares of residuals in matrix form is $\boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y}$

Analysis of Variance

The analysis of variance identity still holds in multiple regression

That is

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

$$SS_T = SS_R + SS_E$$

ANOVA table

We can produce the ANOVA table for a regression with n observations, $p - 1$ explanatory variables and p parameters in the same format as before

	d.f.	SS	MS	VR
Regression				
Residuals				
Total				

Now the Regression row represents the multiple linear regression

Degrees of freedom

Residuals	$n - p$	This is the general case of $n – 2$ when we had $p = 2$ in the simple linear regression model
Regression	$p – 1$	This is the general case of $2 – 1 = 1$ when we had $p = 2$ in the simple linear regression model
Total	$n – 1$	As before

Mean Squares

Mean Squares are again the Sums of Squares divided by degrees of freedom

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

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

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

Variance Ratio

The VR or F Statistic becomes

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

Once again we can use the VR as a test statistic in an overall test of significance of the multiple regression model

Setting out hypotheses tests

1. State the null and alternative hypotheses
2. Give a formula for the test statistic
3. Calculate the test statistic
4. State the assumption the statistic follows under H0
5. Calculate or show the degrees of freedom
6. State the significance level of the test
7. Compare the critical value of the distribution with the test statistic [or calculate p]
8. State the conclusion of the test

Overall test of significance

Does the multiple regression model have statistical significance?

Over a "null model" of constant β_0 plus some random variation

Our null hypothesis is

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

Our alternative hypothesis is that at least one of $\beta_1, \beta_2, \ldots \beta_{p-1}$ is not zero

We would like to reject H_0

We can use the Variance Ratio from the multiple regression ANOVA

Test F statistic

Our F statistic sometimes written $F*$ is the variance ratio from ANOVA

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

The denominator is always an unbiased estimator of σ^2

The numerator is only an unbiased estimator of σ^2 if the regression assumptions hold true

The F test

Under H_0 we will have $F^* \approx 1$

So large values of F^* are required to reject H_0

We compare F^* with the critical value of the F distribution

On $p - 1$ and $n - p$ degrees of freedom

We reject H_0 at $100(1 - \alpha)\%$ significance

if $F^* > F_{n-p}^{p-1}(\alpha)$

Inference about parameters in multiple regression models

We already have the distribution of the least squares parameters

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

Now consider the j[th] parameter estimator $\hat{\beta}_j$ where j = 0, 1, …, p – 1

$\hat{\beta}_j \sim N(\beta_j , \sigma^2 c_{jj})$

Where:

- c_{jj} is the j[th] diagonal element of $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$
- Counting the p diagonal elements 0, 1, …, p – 1
- So the first diagonal element relates to β_0, the next to β_1 , …, the last to β_{p-1}

Inference about β_j

With this normal distribution for β_j we can make inference about β_j in the same way that we did for β_1 in the simple linear regression model

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

Inference about β_j

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 $H_0\colon \beta_j = 0$ versus $H_1\colon \beta_j \neq 0$ is *T* where

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

Need to be careful here

Care needed with the interpretation of these tests of hypothesis

They only apply within the context of the whole *p* parameter model

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 $y_i = \beta_0 + \beta_j x_{ji} + \varepsilon_i$
  - rather it tells us about the role of β_j within the whole *p* parameter model

Estimating the mean response

To estimate the mean response, μ at a certain value of $\boldsymbol{X}$

$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{E}[\boldsymbol{Y}]} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

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

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

And our point estimate is

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

Confidence intervals for μ_0

Using the normal distribution assumption

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

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

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

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

Confidence intervals for μ_0

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

It is straightforward to develop a 100(1 – α)% confidence interval for μ_0

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

Prediction intervals

If we have a new set of ***x*** observations $x_0 = (1, \; x_{1,0} \; \ldots \; x_{p-1,0})^T$

But we do not yet have the corresponding response observation y_0

We can construct a 100(1 – α)% prediction interval

- Which takes account of the random variation that comes with a new observation

Our point estimate for y_0 is $\hat{\mu}_0$ = $\hat{y}_0$

And we have our normal distribution assumption for the y_i's

Developing the prediction interval

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

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

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

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

Developing the prediction interval

Standardising this normal distribution gives

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

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

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

Prediction interval

Our 100(1 – α)% prediction interval for y_0 is

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

Confidence intervals in R

The R commands remain the same as those used in simple linear regression.

If a multiple regression model called `model` has been run using `lm()` and existing observations are saved in a data frame called `x_obs` then

To construct a 99% confidence interval and store it in `conf` use

```
conf <- predict(model, newdata = data.frame(x=x_obs), interval = 'confidence', level = 0.99)
```

If no `level =` is specified, R will default to 95%

Prediction intervals in R

The R commands to create prediction intervals are very similar

If new observations are saved in a data frame called `x_new` then

To construct a 90% prediction interval and store it in `pred` use

```
pred <- predict(model, newdata = data.frame(x=x_new),
interval = 'prediction', level = 0.90)
```

You can display the first 6 rows of `pred` with the command `head(pred)`