Multiple Linear Regression

Dr. Kiah Wah Ong

Statistical Inference in Multiple Linear Regression

We would like to examine the following topics:

- Confidence intervals on regression coefficients.
- Confidence intervals on the mean response.
- Prediction intervals on a future observation.
- Hypothesis testing for β_j.

Statistical Inference in Multiple Linear Regression

Recall the multiple linear regression model is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

and when you have taken n data, we will write the model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \cdots, n$$

where ϵ_i are independent $\mathrm{N}(0, \sigma^2)$.

Statistical Inference in Multiple Linear Regression

Also, recall that

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

and based on

$$\hat{\sigma}^2 = \frac{SS_R}{n-k-1}$$

the estimated variance-covariance matrix is given by

$$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Confidence intervals on regression coefficients

Let C be the $(k+1) \times (k+1)$ matrix below

$$C = (\mathbf{X}^T \mathbf{X})^{-1} = (C_{ij})$$

then a $(1-\alpha) \times 100\%$ confidence interval for the regression coefficient $\beta_j, 0 \leq j \leq k$ is given by

$$\hat{\beta}_j \pm t_{\alpha/2, n-k-1} \sqrt{\hat{\sigma}^2 C_{jj}}$$

Confidence intervals on regression coefficients

Example

Using R and performing the multiple linear regression analysis on the data set called Experiment2, we obtain the 95% confidence intervals of the regression coefficients $\beta_0, \beta_1, \beta_2$ and β_3 as follows:

```
model3=lm(y~x1+x2+x3)
confint(model3,level=0.95)
```

```
> model3=lm(y~x1+x2+x3)
> confint(model3,level=0.95)
                2.5 %    97.5 %
(Intercept)  2.3409122 5.8873197
x1           1.7969460 2.1446658
x2           1.4688527 2.5535023
x3          -0.1036824 0.1554317
>
```

Confidence Interval for the Mean Response

Given the values of

$$x_{01}, x_{02}, \cdots, x_{0k}$$

we would like to construct a confidence interval for the mean response of y at the given level above. Namely,

$$E(y|x_1 = x_{01}, x_2 = x_{02}, \cdots, x_k = x_{0k}) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k}$$

Confidence Interval for the Mean Response

Let us write,

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

then the mean response is given by

$$\begin{aligned} E(y|\mathbf{x}_0) &= \mathbf{x}_0^T \boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k} \end{aligned}$$

Confidence Interval for the Mean Response

From

$$\begin{aligned} E(y|\mathbf{x}_0) &= \mathbf{x}_0^T\boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k} \end{aligned}$$

We see that the natural point estimator for $E(y|\mathbf{x}_0)$ is the following:

$$\hat{y}_0 = \mathbf{x}_0^T\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k}$$

and in fact

$$\hat{y}_0 \sim \mathrm{N}(\mathbf{x}_0^T\boldsymbol{\beta}, \sigma^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)$$

as explained in the next slide. This fact will help us construct confidence interval for the mean response $E(y|\mathbf{x}_0)$.

Confidence Interval for the Mean Response

Since $\hat{y}_0$ is a linear combination of the responses, it is normally distributed with

$$E(\hat{y}_0) = \mathbf{x}_0^T E(\hat{\boldsymbol{\beta}}) = \mathbf{x}_0^T \boldsymbol{\beta}$$

and

$$\begin{aligned} \mathrm{Var}(\hat{y}_0) &= \mathrm{Var}(\mathbf{x}_0^T \hat{\boldsymbol{\beta}}) \\ &= \mathbf{x}_0^T \mathrm{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}_0 \\ &= \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 \end{aligned}$$

Hence under the normality assumption on the model errors, a $(1 - \alpha) \times 100\%$ confidence interval on the mean response $E(y|\mathbf{x}_0)$ is

$$\hat{y}_0 \pm t_{\alpha/2, n-k-1} \sqrt{\hat{\sigma}^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}$$

Confidence Interval for the Mean Response

Example

Suppose we are given a point in $\mathbb{R}^3$, namely

$$x_{01} = 3.1, x_{02} = 4.1, x_{03} = 5.9$$

using R and performing the multiple linear regression analysis on the data set called Experiment2, we obtain the following 95% confidence interval of the mean response as:

```
predict(model3, data.frame(x1=3.1, x2=4.1, x3=5.9), interval="confidence", conf.level=0.95)
```

```
> predict(model3, data.frame(x1=3.1, x2=4.1, x3=5.9), interval="confidence", conf.level=0.95)
      fit      lwr      upr
1 18.6221 17.54363 19.70058
```

Prediction Interval for New Observations

Now, given the values of

$$x_{01}, x_{02}, \cdots, x_{0k}$$

we would like to construct a prediction interval on the future response given by

$$y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k} + \epsilon$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$.

Prediction Interval for New Observations

Using the same notation

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix},$$

$$\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k},$$

and under the normality assumption of the error ϵ, a $(1-\alpha)\%$ prediction interval for the future response y_0 is given by

$$\hat{y}_0 \pm t_{\alpha/2,n-k-1} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0)}$$

Prediction Interval for New Observations

The justification is similar to the one which we have seen in the simple linear regression model. Here is the simple outline:

Let us write $\mathbb{Y} = y_0 - \hat{y}_0$, then it can be shown that $\mathbb{Y}$ is normally distributed with mean

$$\mathbb{Y} = E(y_0) - E(\hat{y}_0) = \mathbf{x}_0^T \boldsymbol{\beta} - \mathbf{x}_0^T \boldsymbol{\beta} = 0$$

and variance

$$\mathrm{Var}(\mathbb{Y}) = \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 + \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$$

Prediction Interval for New Observations

It then follows that

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

and hence

$$\frac{y_0 - \hat{y}_0}{\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)}} \sim t_{n-k-1}$$

Accordingly, a $(1-\alpha)\%$ prediction interval on a future response y_0 at the level $\mathbf{x}_0$ is

$$\hat{y}_0 \pm t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)}$$

Prediction Interval for New Observations

Example

Suppose we are given a point in $\mathbb{R}^3$, namely

$$x_{01} = 3.1, x_{02} = 4.1, x_{03} = 5.9$$

using R and performing the multiple linear regression analysis on the data set called Experiment2, we obtain the following 95% prediction interval of

$$y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_3 x_{03} + \epsilon$$

as

```
predict(model3, data.frame(x1=3.1, x2=4.1, x3=5.9), interval="prediction", conf.level=0.95)
```

```
> predict(model3, data.frame(x1=3.1, x2=4.1, x3=5.9), interval="prediction", conf.level=0.95)
      fit      lwr      upr
1 18.6221 12.99196 24.25224
```

Hypothesis Testing on the Regression Coefficients

Recall that the least squares estimator $\hat{\beta}_j$ is normally distributed with mean β_j and variance $\sigma^2 C_{jj}$ where $j = 0, 1, \cdots, k$.

Hence for the test

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

the test statistics is

$$T = \frac{\hat{\beta}_j - 0}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-k-1}.$$

If we set the significance level at α, then H_0 is rejected if

$$|T| > t_{\alpha/2, n-k-1}$$

Hypothesis Testing on the Regression Coefficients

Alternative method using p-value:
We first predetermine a significance level at α.

We then compute the value of the test statistic

$$\frac{\hat{\beta}_j - 0}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

and call its value ν.

We then reject H_0 if the desired significance level α is at least as large as

$$\begin{aligned} p - \text{value} &= P(|T_{n-k-1}| > \nu) \\ &= 2P(T_{n-k-1} > \nu) \end{aligned}$$

Hypothesis Testing on the Regression Coefficients

Here is an example we have on R.

Example

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
   Min     1Q Median     3Q    Max
-6.505 -1.903 -0.402  2.079  7.253

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.11412    0.89331   4.605 1.26e-05 ***
x1           1.97081    0.08759  22.501  < 2e-16 ***
x2           2.01118    0.27321   7.361 6.23e-11 ***
x3           0.02587    0.06527   0.396    0.693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.784 on 96 degrees of freedom
Multiple R-squared:  0.8637,    Adjusted R-squared:  0.8595
F-statistic: 202.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

Hypothesis Testing on the Regression Coefficients

Why making such a test?

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

If we failed to reject H_0, then the regressor x_j can be considered as insignificant and hence can be deleted from the model, while preserving the other regressors.

Hypothesis Testing on the Regression Coefficients

From our example, we see that

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
   Min     1Q Median     3Q    Max
-6.505 -1.903 -0.402  2.079  7.253

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.11412    0.89331   4.605 1.26e-05 ***
x1           1.97081    0.08759  22.501  < 2e-16 ***
x2           2.01118    0.27321   7.361 6.23e-11 ***
x3           0.02587    0.06527   0.396    0.693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.784 on 96 degrees of freedom
Multiple R-squared:  0.8637,    Adjusted R-squared:  0.8595
F-statistic: 202.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

at any reasonable significant level α, x_3 is insignificant and can be deleted from the model.

Test for Significance of Regression

The test is performed to determine if there is a linear relationship between the response y and any of the regressor variables $x_1, x_2, \cdots, x_k$.

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 = \beta_j \neq 0 \quad \text{for at least one} \quad j$$

Test for Significance of Regression

In our example,

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
   Min     1Q Median     3Q    Max
-6.505 -1.903 -0.402  2.079  7.253

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.11412    0.89331   4.605 1.26e-05 ***
x1           1.97081    0.08759  22.501  < 2e-16 ***
x2           2.01118    0.27321   7.361 6.23e-11 ***
x3           0.02587    0.06527   0.396    0.693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.784 on 96 degrees of freedom
Multiple R-squared:  0.8637,    Adjusted R-squared:  0.8595
F-statistic: 202.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

the p-value associated with the F test is 2.2×10^{-16}, hence at any reasonable significant level α, H_0 is rejected. This means that at least one of the regressor is significantly related to the outcome variable.

R and Adjusted R-Square

After fitting a linear model, we are interested not only in knowing whether a linear relationship exists, but also in measuring the quality of the fit of the model to the data.

Let us define the following:

- Total sum of squared deviations in y from its mean $\overline{y}$

$$\mathrm{SST} = \sum (y_i - \overline{y})^2$$

- Sum of squares due to regression

$$\mathrm{SSR} = \sum (\hat{y}_i - \overline{y})^2$$

- Sum of squared residuals (errors)

$$\mathrm{SSE} = \sum (y_i - \hat{y}_i)^2$$

R and Adjusted R-Square

We can think of

$$\hat{y}_i - \overline{y}$$

as an explained deviation from the mean while

$$y_i - \hat{y}_i$$

as an unexplained deviation, hence the ratio

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{\mathrm{SSR}}{\mathrm{SST}}$$

can be interpreted as the ratio "variability explained by the model" over " total variability of the data".

R-square is called the goodness-of-fit index or coefficient of determination.

R and Adjusted R-Square

In our example

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
   Min     1Q Median     3Q    Max
-6.505 -1.903 -0.402  2.079  7.253

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.11412    0.89331   4.605 1.26e-05 ***
x1           1.97081    0.08759  22.501  < 2e-16 ***
x2           2.01118    0.27321   7.361 6.23e-11 ***
x3           0.02587    0.06527   0.396    0.693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.784 on 96 degrees of freedom
Multiple R-squared:  0.8637,    Adjusted R-squared:  0.8595
F-statistic: 202.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

the value of R^2 for the Experiment2 data is 0.8637, showing that about 86% of the total variation in y can be accounted for by the three regressors x_1, x_2 and x_3 in the model.

R and Adjusted R-Square

The Adjusted R-square, R_a^2 is also used for judging the goodness of fit and is defined as

$$R_a^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

where n is the number of data and k is the number of regressors.

R and Adjusted R-Square

Reasons for the Adjusted R^2:

One shortfall with R^2 is that it increases when we add independent regressors to the model. This is misleading as some of the added regressors might be useless with minimal importance.

Adjusted R^2 overcome this issue by adding a penalty if we add independent regressors that does not improved the model.

If useless regressors are added to the model, Adjusted R^2 will decrease.

If useful regressors are added to the model, Adjusted R^2 will increase.