MIE237

Neil Montgomery

2016-03-01

Recap

Model:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \cdots \beta_k x_{ki} + \varepsilon_i \\
\boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\varepsilon}
\end{aligned}
$$

with the ε_i i.i.d. $N(0, \sigma^2)$.

Solution and details. The $(\boldsymbol{X}'\boldsymbol{X})^{\mathbf{-1}}$ is key to the whole operation.

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

`trees` example

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

The "overall" hypothesis test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \text{Any } \beta_i \neq 0 \end{aligned}$$

Similar to before. Last time we had $SST = SSR + SSE$ with $n-1$, k, and $n-(k+1)$ degrees of freedom. From this we can define $MSR = SSR/k$ and we use:

$$\frac{MSR}{MSE} \sim F_{k,n-(k+1)}$$

(Note: there is no $T^2 = F$ relationship to be had, unlike in the simple regression case.)

`trees` example again

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

`trees` ANOVA table

```
trees %>%
  lm(Volume ~ Girth + Height, data = .) %>%
  anova
```

```
## Analysis of Variance Table
##
## Response: Volume
##           Df Sum Sq Mean Sq  F value  Pr(>F)
## Girth      1 7581.8  7581.8 503.1503 < 2e-16 ***
## Height     1  102.4   102.4   6.7943 0.01449 *
## Residuals 28  421.9    15.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In R the regression SS line is split up into all the 1 degree of freedom components (i.e. one for each input variable). They could be added up to get SSR.

R^2

Same as before:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Same old meaning, now with even more potential for abuse!

(Note: square root of R^2 is now nothing in particular in multiple regression)

Mean response (with confidence) - I

The concept is the same as before. Let's say we have some new vector $\boldsymbol{x_0}$ defined as:

$$\boldsymbol{x_0}' = [1 \quad x_{10} \quad x_{20} \quad \cdots \quad x_{k0}]$$

We need an estimate and a standard error for the estimate.

The estimated mean response is simply $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \cdots \hat{\beta}_k x_{k0}$, or $\hat{y}_0 = \boldsymbol{x_0}'\hat{\boldsymbol{\beta}}$.

The variance is:

$$\mathrm{Var}\left(\boldsymbol{x_0}'\hat{\boldsymbol{\beta}}\right) = \boldsymbol{x_0}'\,\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{x_0} = \sigma^2 \boldsymbol{x_0}'\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}$$

Mean response (with confidence) - II

As usual the 95% confidence interval then becomes:

$$\boldsymbol{x_0}'\hat{\boldsymbol{\beta}} \pm t_{n-(k+1),0.025}\sqrt{MSE}\sqrt{\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}}$$

This is definitely work for the computer only!

The prediction interval for a new response at

$$\boldsymbol{x_0}$$

is similarly:

$$\boldsymbol{x_0}'\hat{\boldsymbol{\beta}} \pm t_{n-(k+1),0.025}\sqrt{MSE}\sqrt{1+\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}}$$

`trees` example

Let's find the intervals at a `Girth` of 11 inches and a `Height` of 72 feet.

```
x_0 <- data_frame(Girth = 11, Height = 72)
predict(trees_lm, newdata=x_0, interval = "c")
```

```
##       fit     lwr      upr
## 1 18.2282 16.4088 20.04759
```

```
predict(trees_lm, newdata=x_0, interval = "p")
```

```
##       fit      lwr      upr
## 1 18.2282 10.07113 26.38526
```