Designing and interpreting linear models

Wenda Zhou

June 12, 2017

Why linear models?

Linear models form a flexible class of models that is widely applicable and simple to use.

Why linear models?

Linear models form a flexible class of models that is widely applicable and simple to use.
Important to understand and interpret estimation and testing in linear models.

Fitting linear models in R

We can fit linear models and generalised linear models using the `lm` or `glm` functions.
We can inspect the result of the fit using the `summary` function.

Interpreting a `lm` fit

```
Call:
lm(formula = y ~ x1 + x2, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66986 -0.95467 -0.04739  0.71725  2.43705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1357     0.3412  -0.398    0.693
x1           -0.2692     0.1889  -1.425    0.162
x2            1.3092     0.1712   7.647 3.96e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.082 on 37 degrees of freedom
Multiple R-squared:  0.9181,    Adjusted R-squared:  0.9137
F-statistic: 207.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

The output includes four sections: call, residuals, coefficients and some statistics.

Interpreting a `lm` fit

These sections refer to the following:

Call	a description of the model fitted
Residuals	some statistics about the residuals
Coefficients	the coefficients fitted
Statistics	some statistics about the general fit

Interpreting a `lm` fit

Residuals

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.66986 -0.95467 -0.04739  0.71725  2.43705
```

This section gives the quartiles of the residuals, and can be useful to see if there is a large skew in the residuals.
It is usually better to plot the residuals.

Interpreting a `lm` fit

Coefficients

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1357     0.3412  -0.398    0.693
x1           -0.2692     0.1889  -1.425    0.162
x2            1.3092     0.1712   7.647 3.96e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This section includes information about each coefficient.
The *estimate* and *std. error* column gives information about the estimate and its standard error.
The `Pr(>|t|)` columns gives the p-value, and the last column displays a graphical summary of that information.

Interpreting a `lm` fit

Statistics

```
Residual standard error: 1.082 on 37 degrees of freedom
Multiple R-squared:  0.9181,    Adjusted R-squared:  0.9137
F-statistic: 207.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

This section includes general statistics about the model (for example, the R^2 statistic).

Interpreting coefficients

The linear model is given by:

$$\mathbb{E}\, y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \tag{1}$$

The coefficient β_1 may be interpreted as the unit change in the average response for each unit change in x_1 with all other variables fixed.

Confidence interval for coefficients

Can be shown that $\hat{\beta}_1$ is approximately normal.
Use confidence interval for normal observation:

$$[\hat{\beta}_1 - \mathrm{SE} \times z_{1-\alpha/2}, \hat{\beta}_1 + \mathrm{SE} \times z_{1-\alpha/2}] \tag{2}$$

where $z_{1-\alpha/2}$ verifies:

$$\mathrm{P}(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2 \tag{3}$$

and can be computed in R by

$$z_{1-\alpha/2} = \mathrm{qnorm}(\alpha/2, \mathrm{lower.tail} = \mathrm{TRUE}) \tag{4}$$

Significance for coefficients

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1357     0.3412  -0.398    0.693
x1           -0.2692     0.1889  -1.425    0.162
x2            1.3092     0.1712   7.647 3.96e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test conditional significance of a single variable: does the variable of interest explain more of the outcome than what the other variables already explain?

Significance for coefficients

Conditional significance can be delicate. A variable can go from significant to non-significant easily.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.22052    0.54051  -0.408    0.686
x1           1.10423    0.09272  11.910 2.14e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1357     0.3412  -0.398    0.693
x1           -0.2692     0.1889  -1.425    0.162
x2            1.3092     0.1712   7.647 3.96e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Significance for coefficients

Significant is not the same as important.

Significance

Significance indicates that we are statistically certain that there exists an effect.

Importance

Importance indicates that the size of the effect implies that it should be taken into account.

Interpreting categorical variables

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1235     0.2979  10.486 1.24e-12 ***
catB          2.3946     0.4146   5.775 1.27e-06 ***
catC          5.9761     0.4864  12.286 1.26e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Categorical variable with three levels A, B and C.

Interpreting categorical variables

Coefficients for a categorical variable can no longer be interpreted as the unit change for each unit increase.

Interpreting categorical variables

Coefficients for a categorical variable can no longer be interpreted as the unit change for each unit increase.
Instead, represent difference between levels.

Interpreting categorical variables

Coefficients for a categorical variable can no longer be interpreted as the unit change for each unit increase.
Instead, represent difference between levels.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1235     0.2979  10.486 1.24e-12 ***
catB          2.3946     0.4146   5.775 1.27e-06 ***
catC          5.9761     0.4864  12.286 1.26e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpreting categorical variables

Ordered

For ordered variables we are often interested in understanding whether the response increases or decreases with the factor. We will often use a polynomial contrast.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.8480     2.9624  -1.974 0.056086 .
cat.L        -4.7559     2.0878  -2.278 0.028768 *
cat.Q        -0.1344     0.2989  -0.450 0.655719
x             1.1947     0.2937   4.067 0.000247 ***
```

Interpreting categorical variables

Ordered

Polynomial contrasts decompose the factor into a linear, quadratic, cubic, etc. effects.
There are as many effects as the number of levels minus 1.