General linear models

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

Simple linear regression

- Given a set of n bivariate data-points $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, a simple linear regression model assumes that for all $i \in 1 \ldots n$,

$$y_i = a + bx_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

- This is identical to saying that for all $i \in 1 \ldots n$,

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2), \\ \mu_i &= a + bx_i. \end{aligned}$$

- In other words, we are saying that are observed outcome variable values $y_1, y_2 \ldots y_n$ are samples from Normal distributions whose means are *linear functions* of the predictor variable's values $x_1, x_2 \ldots x_n$.

Multiple linear regression

- Given a set of observed values $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i$ is the array $[x_{1i}, x_{2i}, \ldots x_{Ki}]$, a *multiple* regression model assumes that for all $i \in 1 \ldots n$,

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma^2), \\
\mu_i &= b_0 + \sum_{k=1}^{K} b_k x_{ki}.
\end{aligned}
$$

- In other words, each observed outcome variable value y_i is a sample from a Normal distributions whose mean is a linear function of the values of the K predictor variables $x_{1i}, x_{2i}, \ldots x_{Ki}$.
- Note that a linear function is just a weighted sum.

Estimating the parameters

- Given a set of observed values, the aim of parameter estimation is to infer the possible values of $\mathbf{b} = [b_0, b_1 \dots b_k \dots b_K]$.
- The least-squares estimate of $\mathbf{b}$ is given by the set of parameters that minimize
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
where $\hat{y}_i = b_0 + \sum_{k=1}^{K} b_k x_{ki}$.
- This least-squares estimate is also the *maximum-likelihood* estimate.

Simple linear regression: Coefficient's of line of best-fit

- When we just have one predictor, the coefficients of the line of best fit is given by

$$\hat{b} = \frac{\mathrm{cov}(x, y)}{s_x^2} \quad \text{or} \quad r\,\frac{s_y}{s_x},$$

where r is Pearson's correlation coefficient, and

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Simple linear regression: Hypothesis testing $\hat{b}$

- If the true value of the slope is b then

$$\frac{\hat{b} - b}{S_{\hat{b}}} \sim t_{n-2}.$$

- Here, the standard error is

$$S_{\hat{b}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},$$

with $\hat{y} = \hat{a} + \hat{b}x_i$.

Confidence Intervals

- Confidence intervals on the true value of b are calculated as in the case of confidence intervals in a t-test.
- For example, in simple linear regression with n data-points, the relevant degrees of freedom are $n-2$. The 95% confidence interval is
$$\hat{b} \pm T(0.95, n-2) \times S_{\hat{b}},$$
where $\pm T(0.95, n-2)$ stands for the values between which lie 95% of the area in a t-distribution with $n-2$ degrees of freedom.

Hypothesis testing and confidence intervals in multiple regression

- In a multiple regression analysis with K predictors, hypothesis tests and confidence intervals on the individual predictors are similar to the case of simple linear regression.
- A null hypothesis test on the true value of the predictor b_k is based on the fact that if the true value of the predictor is 0 then

$$\frac{\hat{b}_k}{S_{\hat{b_k}}} \sim t_{n-K-1}.$$

- This is very similar to the case with one predictor variable. However, $S_{\hat{b_k}}$ is calculated differently.
- Likewise, the 95% confidence interval on b_k is

$$\hat{b}_k \pm T(0.95, n-K-1) \times S_{\hat{b}_k},$$

R^2: *The coefficient of determination*

- It can be shown that

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{RSS}},$$

  where TSS is *total* sum of squares, ESS is *explained* sum of squares, and RSS is *residual* sum of squares.

- The coefficient of determination R^2 is defined as

$$\begin{aligned} R^2 &= \frac{\text{ESS}}{\text{TSS}} = \text{Proportion of variation that is explained,} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \end{aligned}$$

Hypothesis testing R^2

- A test that $R^2 = 0$ is identical to a test that $b_1 = b_2 = \ldots = b_k = \ldots = b_K = 0$.
- If $R^2 = 0$ then

$$\frac{ESS/K}{RSS/(n-K-1)} \sim F_{(K,n-K-1)}.$$

- As before, if this statistic is greater than a critical threshold (i.e. the value of the F distribution below which lies e.g. 95% of the mass), then we may reject this (null) hypothesis that $R^2 = 0$.

Adjusted R^2

- By explaining proportion of variance explained, R^2 is used a *goodness of fit* measure.
- However, R^2 will always grow with K, the number of predictors.
- R^2 can be *adjusted* to counteract the artificial effect of increasing numbers of predictors as follows:

$$R^2_{\text{Adj}} = \underbrace{1 - \frac{\text{RSS}}{\text{TSS}}}_{R^2} \underbrace{\frac{n-1}{n-K-1}}_{\text{penalty}}$$

- While R^2_{Adj} is *not* identical to the proportion of variance explained, nor is $R^2_{\text{Adj}} = 0$ equivalent to $\beta_1 = \ldots = \beta_K = 0$.

Polychotomous Predictor Variables

- A variable such as $x \in \{\text{english}, \text{Scottish}, \text{Welsh}\}$ can not be reasonably recoded as $x \in \{0, 1, 2\}$.
- In this situation, we recode as follows:

	x_1	x_2
English	0	0
Scottish	0	1
Welsh	1	0

- The two variables used to code the single categorical are sometimes called *dummy* variables.
- In general, for a variable with L possible values (i.e. levels), we need $L - 1$ dummy variables.

Polychotomous Predictor Variables

- In a model with one categorical variable with three possible values, the regression equation is

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i.$$

where x_{1i} and x_{2i} collectively code the categorical variable value.

- Using the case of {English, Scottish, Welsh}, coded as above, we have

$$\begin{aligned} y_i &\sim N(b_0, \sigma^2) \quad \text{if group = English,} \\ y_i &\sim N(b_0 + b_2, \sigma^2) \quad \text{if group = Scottish,} \\ y_i &\sim N(b_0 + b_1, \sigma^2) \quad \text{if group = Welsh.} \end{aligned}$$

- If $\mu_1 \triangleq b_0$, $\mu_2 \triangleq b_0 + b_2$, $\mu_3 = b_0 + b_1$, then this is identical to a one-way Anova model with three groups.

Polychotomous Predictor Variables

- Continuing with the previous example, the coefficients b_0, b_1 and b_2 have the interpretation:

b_0	"Mean of English group",
$b_0 + b_1$	"Mean of Welsh group",
$b_0 + b_2$	"Mean of Scottish group",

and so

b_2	"difference of means of Welsh and English",
b_3	"difference of means of Scottish and English".

Mixing categorical predictor and continous predictors

- Let's say we have an outcome variable $y_1, y_2 \ldots y_n$ and one continous variable $x_{11}, x_{12}, \ldots x_{1n}$ and one categorical variable $g_1, g_2, \ldots g_n$, where e.g. each $g_i \in \{\text{english}, \text{scottish}, \text{welsh}\}$, we can then recode each g_i with x_{2i} and x_{3_i} as above, and then perform a multiple linear regression:

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma^2), \\
\mu_i &= b_0 + b_1 x_{1i} + \underbrace{b_2 x_{2i} + b_3 x_{3i}}_{\text{categorical variable}}
\end{aligned}
$$

- In this case we have

$$
\begin{aligned}
y_i = b_0 + b_1 x_{1i} + \epsilon_i \quad &\text{(english)}, \\
y_i = b_0 + b_1 x_{1i} + b_3 + \epsilon_i \quad &\text{(scottish)}, \\
y_i = b_0 + b_1 x_{1i} + b_2 + \epsilon_i \quad &\text{(welsh)}.
\end{aligned}
$$

Mixing categorical predictor and continous predictors

- If we rewrite

$$\begin{aligned} y_i &= b_0 + b_1 x_{1i} + \epsilon_i \quad \text{(english)}, \\ y_i &= b_0 + b_1 x_{1i} + b_3 + \epsilon_i \quad \text{(scottish)}, \\ y_i &= b_0 + b_1 x_{1i} + b_2 + \epsilon_i \quad \text{(welsh)}. \end{aligned}$$

  as

$$\begin{aligned} y_i &= b_0 + b_1 x_{1i} + \epsilon_i \quad \text{(english)}, \\ y_i &= (b_0 + b_3) + b_1 x_{1i} + \epsilon_i \quad \text{(scottish)}, \\ y_i &= (b_0 + b_2) + b_1 x_{1i} + \epsilon_i \quad \text{(welsh)}, \end{aligned}$$

  we notice that these are *varying intercept* linear regression model.

Varying intercept linear models

Given data

```
## # A tibble: 10 x 3
##          x country   score
##      <dbl> <chr>     <dbl>
##  1  0.940 english    6.35
##  2 -1.20  scottish   5.81
##  3 -0.602 welsh      8.99
##  4 -0.217 welsh      8.84
##  5 -0.538 scottish   5.62
##  6 -1.86  english    3.98
##  7 -0.988 english    4.86
##  8 -1.94  english    4.24
##  9  0.972 welsh     10.2
## 10  0.442 welsh     10.7
```

we can do

```
M <- lm(score ~ x + country, data=Df)
```

to perform a varying-intercept model.

Varying intercept linear models