CH 8: Least Squares and Maximum Likelihood

Least Squares

Least squares is a common method for estimating regression coefficients in a linear model. Consider the model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where i indexes the i^{th} observation.

The residual sum of square are formally defined as:

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Hence, to minimize this value, we will take the derivatives with respect to β_0 and β_1

Then we also solve for β_1 plugging in the value of β_0

$$
\begin{aligned}
\frac{dRSS}{d\beta_1} &= -2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)x_i \\
(\text{set} = 0) \quad -2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)x_i &= 0 \\
\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)x_i &= 0 \\
\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\beta_0-\sum_{i=1}^{n}\beta_1 x_i^2 &= 0 \\
\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i(\bar{y}-\beta_1\bar{x})-\sum_{i=1}^{n}\beta_1 x_i^2 &= 0 \\
\sum_{i=1}^{n}x_i y_i-\bar{y}\sum_{i=1}^{n}x_i-\beta_1\bar{x}\sum_{i=1}^{n}x_i-\beta_1\sum_{i=1}^{n}x_i^2 &= 0 \\
\beta_1\left(\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i\right) &= \sum_{i=1}^{n}x_i y_i-\bar{y}\sum_{i=1}^{n}x_i \\
\beta_1\left(\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i\right) &= \sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}\frac{y_i}{n}n\bar{x} \\
\beta_1\left(\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i\right) &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i \\
\beta_1\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right) &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i \\
. &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i \\
. &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i \\
. &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i \\
\beta_1\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i \\
\beta_1 &= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}
\end{aligned}
$$

This calculation is actually considerably easier using matrix notation.

Let the model be specified as $\mathbf{y} = X\beta + \epsilon$, where

- $\mathbf{y}$ is a $n \times 1$ matrix (or vector), such that

- X is a $n \times 2$ matrix,

- β is 2×1 matrix,

- ϵ is a $n \times 1$ matrix (or vector), such that

To verify this matrix algebra, we can look at the first row of the matrix which results in

The residuals can be defined as

$$\mathbf{r} = \mathbf{y} - X\hat{\beta} = \begin{pmatrix} y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1 \\ y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2 \\ \vdots \\ y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n \end{pmatrix}$$

Then the sum of squares is written as:

$$\sum_{i=1}^{n} = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = \mathbf{r}^T \mathbf{r}$$

To minimize the sum of squares of the residuals, consider

$$\begin{aligned} \mathbf{r}^T \mathbf{r} =& (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \\ =& \mathbf{y}^T \mathbf{y} - \mathbf{y} T X\beta - \beta^T X^T \mathbf{y} + \beta^T X^T X\beta \\ =& \mathbf{y}^T \mathbf{y} - 2\beta^T X^T \mathbf{y} + \beta^T X^T X\beta \end{aligned}$$

Now take the derivative

$$\frac{d\,SSR}{d\,\beta^T} = -2X^t \mathbf{y} + 2X^T X\beta \tag{1}$$

and set = 0

$$0 - 2X^T \mathbf{y} + 2X^T X\beta \tag{2}$$

$$X^T \mathbf{y} = X^T X\beta \tag{3}$$

$$(X^T X)^{-1} X^T \mathbf{y} = \beta \tag{4}$$

σ is also estimated with the sum of the squared residuals

$$\hat{\sigma} = \sqrt{\frac{1}{n-k} \sum_{i=1}^{n} (y_i - X_i \hat{\beta})^2}$$

,

where k is the number of predictors in the model (including the intercept)

With least-squares estimation, there is no specification of a probability distribution. This is strictly a geometric procedure.

Recall a regression model can be written as

$$y = \beta_0 + \beta_1 x + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

or

Hence, the likelihood corresponds to how well the data point y corresponds to the normal density (or likelihood),

lm and stan_glm

With Bayesian inference, and the use of prior information, the posterior is a product of the the likelihood of the data as well as the prior.

The role prior distribution is often characterized as a penalty (to the likelihood) or regularization, where the prior can down weight some values of the parameter.

As we have seen, Bayesian inference enables simulated based approaches for summarizing parameter coefficients, contrasts, and predictions. Bayesian inference also allows uncertainty to be expressed using probability, as opposed to using confidence. However, the cost is specifying a prior, which can be subjective.

Given that maximum-likelihood methods are optimizaing the likelihood while Bayesian inference focus on the posterior (likelihood + prior), we'd expect differences in the results.

The default priors are weakly informative, so that posterior is not vastly different from the likelihood.

```
beer <- read_csv('http://math.montana.edu/ahoegh/Data/Brazil_cerveja.csv')
```

```
## Parsed with column specification:
## cols(
##   consumed = col_double(),
##   precip = col_double(),
##   max_tmp = col_double(),
##   weekend = col_double()
## )
```

```
beer %>% lm(consumed ~ max_tmp, data = .) %>% coef()
```

```
## (Intercept)     max_tmp
##   7.9749394   0.6548456
```

```
beer %>%
  stan_glm(consumed ~ max_tmp, data = ., refresh = 0, iter = 100000) %>% coef()
```

```
## (Intercept)     max_tmp
##    7.982674    0.654593
```

We can explicitly state that flat (uniform, uniformative) priors, so that the posterior and the likelihood are are the same

```
beer <- read_csv('http://math.montana.edu/ahoegh/Data/Brazil_cerveja.csv')
```

```
## Parsed with column specification:
## cols(
##   consumed = col_double(),
##   precip = col_double(),
##   max_tmp = col_double(),
##   weekend = col_double()
## )
```

```
beer %>% lm(consumed ~ max_tmp, data = .) %>% coef()
```

```
## (Intercept)     max_tmp
##   7.9749394   0.6548456
```

```
beer %>%
  stan_glm(consumed ~ max_tmp, data = .,
           refresh = 0, iter = 100000,
           prior_intercept = NULL,
           prior = NULL,
           prior_aux = NULL) %>% coef()
```

```
## (Intercept)     max_tmp
##   7.9775519   0.6546636
```

Uncertainty Intervals

In general, if you use classical methods and confidence intervals or Bayesian methods and credible intervals, it is important to be transparent about the model fitting process and interpretations.

I will be more lenient about this, but outside of my class, make sure you are clearly articulating the type of estimation and appropriate interpretation.