Linear regression cheatsheet

Gavin Band, WHG GMS Programme 2021

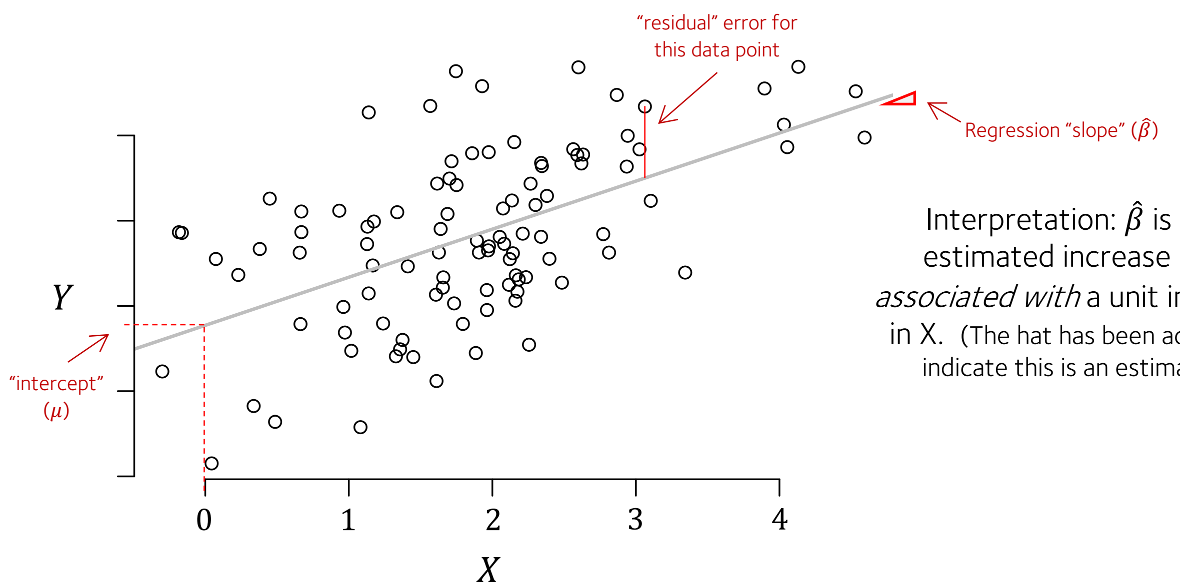
Linear regression models an outcome variable (Y) in terms of one or more predictor variables (X). The model asserts that Y is a linear combination of columns of X plus some noise. The noise is assumed to be Gaussian with some variance σ^2. The residual variance is assume to be the same for all data points).

$$Y = \mu + X_1\beta_1 + X_2\beta_2 + \cdots X_d\beta_d + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

Or using matrix notation:

$$Y = \mu + X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

Matrix multiplication of the d-dimensional *row* vector of predictors X and the d-dimensional *column vector* of parameters β



Interpretation: $\hat{\beta}$ is the estimated increase in Y *associated with* a unit increase in X. (The hat has been added to indicate this is an estimate.)

The likelihood function. The regression likelihood composes the above into a single formula – the likelihood of Y given X and the parameters. (It is simplest to write this if we instead imagine μ to be the first entry of β . This works out if we add a single 1 as the first entry of X:

For a single sample:

$$P(Y|X,\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\cdot\frac{(Y - X\beta)^2}{\sigma^2}}$$

Squared residual (distance) from regression line

The outcome values are assumed independent of each other (probabilities multiply). So for multiple samples the likelihood is:

For multiple samples: ($n = 1, \ldots, N$)

$$P(Y|X,\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\cdot\frac{\sum_n (Y_n - X_n\beta)^2}{\sigma^2}}$$

"sum of squared errors"

The exponent is negative. Maximising the likelihood is therefore the same as minimizing the sum of squared errors – it finds the 'best-fitting line'.

Basic linear regression (maximum likelihood) in R:

```
> fit = lm( Y ~ X, data = D )

> coefficients(fit)
 (Intercept)            X
0.0007606242 0.3135072376

> logLik(fit)
'log Lik.' -132.981 (df=3)

> residuals(fit)
         1          2          3          4
-0.6115976 -0.3239313  0.7034511 -0.2934937 . . .

> summary(fit)$coefficients
                Estimate Std. Error     t value     Pr(>|t|)
(Intercept) 0.0007606242 0.09412669 0.008080856 0.9935689071
X           0.3135072376 0.08512788 3.682780013 0.0003778035
```

This turns out to have an analytic solution:

$$\hat{\beta} = (X^tX)^{-1}X^tY$$

Maximum likelihood estimate (MLE)

$$\text{variance}(\hat{\beta}) = \sigma^2(X^tX)^{-1}$$

Variance of MLE

$$\text{se}(\hat{\beta}_j) = \sqrt{\sigma^2 {(X^tX)_{jj}}^{-1}}$$

Standard error of MLE

But what if you want to fit with prior information included? Use **`brms`** package:

```
> library( bmrs )

> fit = brm(
      Y ~ X,
      data = data,
      prior = set_prior( "normal(0,1)" )
)

> fit$fit
Inference for Stan model: ca2436c230608c2ca38ebc402110120d.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.
             mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
b_Intercept  0.25    0.00 0.05   0.16   0.22   0.25   0.28   0.34  3549    1
b_X         -0.05    0.00 0.04  -0.13  -0.08  -0.05  -0.02   0.03  4293    1
sigma        0.45    0.00 0.03   0.39   0.42   0.44   0.47   0.51  3729    1
lp__       -65.24    0.03 1.25 -68.44 -65.82 -64.91 -64.32 -63.81  1972    1


Samples were drawn using NUTS(diag_e) at Thu Nov 11 17:56:07 2021.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```