Probability, Likelihood, and Other Measures of Model Fit

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

Probabilistic generative model

- In any statistical analysis, we assume our data is drawn from some probability distribution.
- This is sometimes known as the *probabilistic generative model*, and in fact is exactly what we mean by the *statistical model*.
- This model is a model of the *statistical population*, which could also be described as the true generative model.
- In general in analyses, we aim to find a good, or good enough, model of the population.

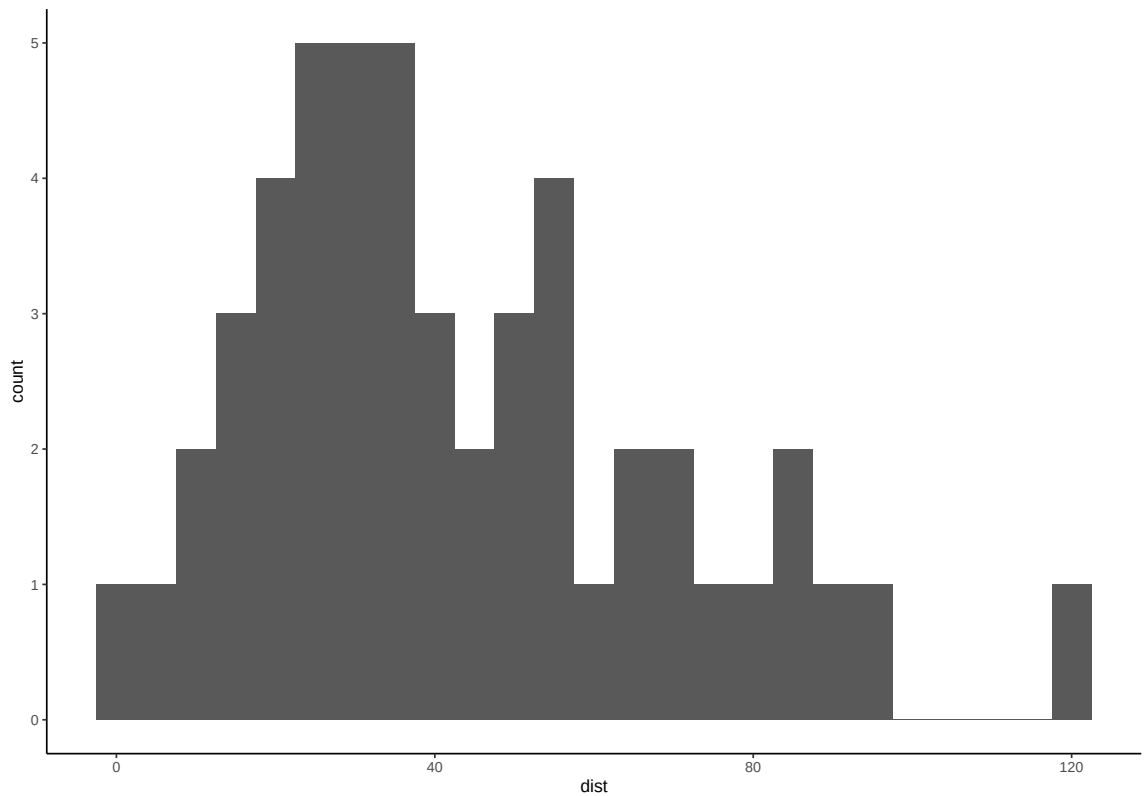
Model evaluation

- One general way we can evaluate a model is by asking if the data is compatible with model.
- One way to look at this is to calculate the probability of the data according to the model.
- If the probability of observing the data is relatively high in one model than in another, then the data is more compatible with the former than the latter model.
- We often refer to the probability of the data according to the model as the model's *likelihood*.

Example problem

- Let's assume we are analysing the `dist` variable from the `cars` data, which is depicted in the following scatterplot:

```
cars %>%
  ggplot(aes(x = dist)) + geom_histogram(binwidth = 5)
```



Probabilistic model

- One possible model of the `dist` variable is the following

$$y_i \sim N(\mu, \sigma^2) \quad \text{for } i \in 1...n,$$

  where y_i is the `dist` variable on observation i.

- In other words, we are modelling `dist` as normally distributed with a mean μ and standard deviation σ, but we do not know the values of the parameters μ and σ.

Model likelihood

- Assuming values for μ, σ, what the probability of the observed values of the `dist` variable, $y_1, y_2, y_3 \ldots y_n$?
- This is

$$P(y_1 \ldots y_n | \mu, \sigma).$$

- In this model, all y's are conditionally independent of one another, so the the joint probability is as follows:

$$P(y_1 \ldots y_n | \mu, \sigma) = \prod_{i=1}^{n} P(y_i | \mu, \sigma).$$

- We do not know the values of μ and σ, so we use their *maximum likelihood estimates*: $\hat{\mu}$ and $\hat{\sigma}_{\text{mle}}$:

$$P(y_1 \ldots y_n | \hat{\mu}, \hat{\sigma}_{\text{mle}}) = \prod_{i=1}^{n} P(y_i | \hat{\mu}, \hat{\sigma}_{\text{mle}}).$$

Model log likelihood

- The joint probability

$$P(y_1 \dots y_n | \mu, \sigma) = \prod_{i=1}^{n} P(y_i | \mu, \sigma).$$

will be a very small number (a product of probabilities), so we usually calculate its logarithm:

$$\log \left(\prod_{i=1}^{n} P(y_i | \mu, \sigma) \right) = \sum_{i=1}^{n} \log P(y_i | \mu, \sigma),$$

and with the maximum likelihood estimators for the unknowns, this is

$$\sum_{i=1}^{n} \log P(y_i | \hat{\mu}, \hat{\sigma}_{\text{mle}}).$$

Model log likelihood: calculations

```
y <- cars$dist
M0 <- lm(y ~ 1) # normal model
mu_hat <- coef(M0) # mle of mu
sigma_mle <- sqrt(mean(residuals(M0)^2)) # mle of sigma

dnorm(y, mean = mu_hat, sd = sigma_mle, log = TRUE) %>%
  sum()
#> [1] -232.9012
# same as
logLik(M0)
#> 'log Lik.' -232.9012 (df=2)
```

Regression models

- Often, for each observed value of the variable being modelled we have observed values of other variables (variously known as *covariates*, *predictor* variables, *independent* variables).
- For example, for each value of `dist`, we have the speed of the car `speed`.
- The first 10 observations of `cars` are:

```
head(cars, 10)
#>    speed dist
#> 1      4    2
#> 2      4   10
#> 3      7    4
#> 4      7   22
#> 5      8   16
#> 6      9   10
#> 7     10   18
#> 8     10   26
#> 9     10   34
#> 10    11   17
```

Regression probabilistic model

- Using the `speed` variable too, a potential model of the `dist` data is the following

$$y_i \sim N(\mu_i, \sigma^2) \quad \text{for } i \in 1...n,$$
$$\mu_i = \beta_0 + \beta_1 x_i,$$

  where y_i and x_i are the `dist` and `speed` variables on observation i.

- In other words, we are modelling `dist` as normally distributed around a mean that is a linear function of `speed`, and with a fixed variance σ^2. This is exactly a simple linear regression model.
- Here, we do not know the values of the parameters β_0, β_1, and σ^2.
- Note that this is a probabilistic model of the outcome variable only.

Regression model likelihood, log likelihood

- Assuming values for β_0, β_1, σ, what the probability of the observed values of the `dist` outcome variable, $y_1, y_2, y_3 \ldots y_n$ given the observed values of the `speed` predictor, $x_1, x_2, x_3 \ldots x_n$?
- This is

$$P(y_1 \ldots y_n | x_1 \ldots x_n, \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} P(y_i | x_i, \beta_0, \beta_1, \sigma).$$

- The log likelihood of the model is

$$\sum_{i=1}^{n} \log P(y_i | x_i, \beta_0, \beta_1, \sigma).$$

We do not know β_0, β_1, σ and so we use their maximum likelihood estimates, denoted $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}_{\text{mle}}$.

Regression log likelihood: calculation

```
x <- cars$speed # y is dist, defined above
M1 <- lm(y ~ x) # normal linear model
# for each observation y_i, there is a mu_i
mu_hat <- predict(M1)
sigma_mle <- sqrt(mean(residuals(M1)^2)) # mle of sigma

dnorm(y, mean = mu_hat, sd = sigma_mle, log = TRUE) %>%
  sum()
#> [1] -206.5784
# same as
logLik(M1)
#> 'log Lik.' -206.5784 (df=3)
```

Likelihood ratios

- We have two model of the `dist` variable.
- The normal model has a log likelihood of -233, and the regression model has a log likelihood of -207. Let's denote these log likelihoods as $\log \mathcal{L}_0$ and $\log \mathcal{L}_1$, respectively.
- The log of the ratio of likelihoods is as follows:

$$\begin{aligned} \log\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) &= \log \mathcal{L}_1 - \log \mathcal{L}_0, \\ &= -207 - -233, \\ &= 26. \end{aligned}$$

- In other words,

$$\frac{\mathcal{L}_1}{\mathcal{L}_0} = e^{26} \approx 10^{11}$$

Residual sum of squares (RSS)

- The sum of squared residuals in normal linear models when using the maximum likelihood estimators is

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} |y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)|^2, \\ &= \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 \end{aligned}$$

- Using R, the residuals are obtained by

```
head(residuals(M1))
#>         1         2         3         4         5         6
#>  3.849460 11.849460 -5.947766 12.052234  2.119825 -7.812584
```

or

```
head(y - mu_hat)
#>         1         2         3         4         5         6
#>  3.849460 11.849460 -5.947766 12.052234  2.119825 -7.812584
```

Residual sum of squares and log likelihood

- The RSS is a measure of the model's lack of fit.
- The model's log likelihood and its RSS are related as follows:

$$\log \mathcal{L} = -\frac{n}{2}\left(\log(2\pi) - \log(n) + \log(\text{RSS}) + 1\right)$$

```
rss <- sum(residuals(M1)^2)
n <- length(y)

-(n/2) * (log(2*pi) - log(n) + log(rss) + 1)
#> [1] -206.5784
logLik(M1)
#> 'log Lik.' -206.5784 (df=3)
```

- In two normal linear models of the same data, the differences in likelihood or determined only by differences in RSS.

Root mean square error

- The larger the sample size, the larger the RSS.
- An alternative to RSS as a measure of model fit is the square root of the mean of the squared residuals, known as the *root mean square error* (RMSE):

$$\text{RMSE} = \sqrt{\frac{\text{RSS}}{n}},$$

- This is $\hat{\sigma}_{\text{mle}}$.

Mean absolute error

- Related to RMSE is the mean absolute error (MAE), which the mean of the absolute values of the residuals.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

- In R

```
mean(abs(residuals(M1)))
#> [1] 11.58012
```

Deviance

- Deviance is used as a measure of model fit in generalized linear models.
- Strictly speaking, the deviance of model M_0 is

$$2\left(\log \mathcal{L}_s - \log \mathcal{L}_0\right),$$

  where $\log \mathcal{L}_0$ is the log likelihood (at its maximum) of model M_0, and $\log \mathcal{L}_s$ is a *saturated* model, i.e. one with as many parameters as there are data points.
- When comparing two models, M_0 and M_1, the saturated model is the same, and so the difference of the deviances of M_0 and M_1 is

$$(-2\log \mathcal{L}_0) - (-2\log \mathcal{L}_1),$$
$$\mathcal{D}_0 - \mathcal{D}_1,$$

  and so the deviance of M_0 is usually defined simply as

$$-2\log \mathcal{L}_0.$$

Differences of deviances

- Differences of deviances are equivalent to log likelihood ratios:

$$\begin{aligned}\mathcal{D}_0 - \mathcal{D}_1 &= -2\log\mathcal{L}_0 - -2\log\mathcal{L}_1, \\ &= -2\left(\log\mathcal{L}_0 - \log\mathcal{L}_1\right), \\ &= -2\log\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right), \\ &= 2\log\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right).\end{aligned}$$

- Clearly, $\frac{\mathcal{L}_1}{\mathcal{L}_0}$ the factor by which the likelihood of model M_1 is greater than that of model M_0.
- Therefore, the difference of the deviance of models M_0 and M_1 ($D_0 - D_1$), gives the (two times) the logarithm of the factor by the likelihood of model M_1 is greater than that of model M_0.
- The larger $D_0 - D_1$, the greater the likelihood of M_1 compared to M_0.

Logistic regression example

```
cars_df <- mutate(cars, z = dist > median(dist))
M2 <- glm(z ~ speed,
          data = cars_df,
          family = binomial(link = 'logit')
)

logLik(M2)
#> 'log Lik.' -17.73468 (df=2)
deviance(M2)
#> [1] 35.46936
logLik(M2) * -2
#> 'log Lik.' 35.46936 (df=2)
```

Conditional probability in logistic regression

- The model in a logistic regression (with one predictor) is

$$y_i \sim \text{Bernoulli}(\theta_i), \quad \text{for } i \in 1 \dots n$$

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 x_i$$

- The conditional probability of $y_1, y_2 \dots y_n$ given $x_1, x_2 \dots x_n$ is

$$\prod_{i=1}^{n} \theta_i^{y_i} (1-\theta_i)^{1-y_i},$$

  where each θ_i is

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 x_i$$

Conditional probability in logistic regression

- The logarithm of the conditional probability of $y_1, y_2 \dots y_n$ is

$$\log\left(\prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}\right) = \sum_{i=1}^{n} \log\left(\theta_i^{y_i}(1-\theta_i)^{1-y_i}\right),$$

$$= \sum_{i=1}^{n} \left(y_i \log\theta_i + (1-y_i)\log(1-\theta_i)\right),$$

$$= \sum_{i=1}^{n} y_i \log\theta_i + \sum_{i=1}^{n} (1-y_i)\log(1-\theta_i)$$

Conditional probability in logistic regression

```
theta <- predict(M2, type = 'response')
sum(log(theta[cars_df$z])) + sum(log(1-theta[!cars_df$z]))
#> [1] -17.73468
```

```
z <- pull(cars_df, z)
sum(z * log(theta) + (1-z) * log(1 - theta))
#> [1] -17.73468
```

Deviance residuals

- Deviance residuals are values such that their sum of squares is equal to the model's deviance.
- We know that the sum, for $i \in 1 \dots n$, of the following is the log likelihood:

$$y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i),$$

and so the sum of the following, for $i \in 1 \dots n$, is the deviance:

$$-2 \left(y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i) \right).$$

- So the sum of the *squares* of the following, for $i \in 1 \dots n$, is the deviance:

$$\sqrt{-2 \left(y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i) \right)}.$$

- All of these values will necessarily be positive.
- It is conventional for deviance residuals to be negative when $y_i = 0$ and positive when $y_i = 1$.

Deviance residuals

```
d <- sqrt( -2 * (z * log(theta) + (1-z) * log(1 - theta)))
sum(d^2)
#> [1] 35.46936
```

```
d[c(1, 25, 35, 50)]
#>          1         25         35         50
#> 0.05724272 1.00995907 0.71599367 0.11291237
```

```
residuals(M2)[c(1, 25, 35, 50)]
#>           1          25          35          50
#> -0.05724272 -1.00995907  0.71599367  0.11291237
z[c(1, 25, 35, 50)]
#> [1] FALSE FALSE  TRUE  TRUE
(ifelse(z, 1, -1) * d)[c(1, 25, 35, 50)]
#>           1          25          35          50
#> -0.05724272 -1.00995907  0.71599367  0.11291237
```