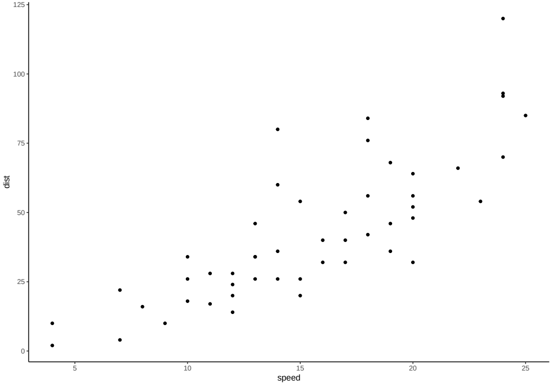
Probability, Likelihood, and Other Measures of Model Fit

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

Example problem

- Let's assume we have the `cars` data, which is depicted in the following scatterplot:



Example problem

- 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
```

Probabilistic model

- A potential model of the `cars` data is the following

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

  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.
- 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.

Conditional probability of any observation

- Given our model specification, we can ask what is the probability of any given value of `dist`, assuming a given value of `speed`, for any given values of the parameters β_0, β_1, σ^2.
- For example, we can ask, what is the probability that `dist` $= 50$ if `speed` $= 15$ if β_0, β_1, and σ have the values -20, 4, 15, respectively, i.e.

$$P(y = 50 | x = 15, \beta_0 = -20, \beta_1 = 4, \sigma = 15),$$

- We can do this for *any* values of `dist`, `speed`, and β_0, β_1, and σ.

Conditional probability of any observation

- If $x = 15$, and $\beta_0 = -20$, $\beta_1 = 4$, $\sigma = 15$, then the value of y has been assumed to be drawn from a normal distribution with mean

$$\begin{aligned}\mu &= \beta_0 + \beta_1 x,\\ \mu &= -20 + 4 \times 15,,\\ \mu &= 40\end{aligned}$$

  and a standard deviation of $\sigma = 15$.

- And so the probability that $y = 50$ when $x = 15$, and $\beta_0 = -20$, $\beta_1 = 4$, $\sigma = 15$, is the probability of a value of 50 in a normally distributed random variable whose mean is 40 and whose standard deviation is 15.

Conditional probability of any observation

- The probability (density) that a normal random variable, with mean of 40 and standard deviation of 15, takes the value of 50 can be obtained from this probability density function for normal distributions:

$$P(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|y-\mu|^2}{2\sigma^2}\right)$$

- Using R, this is

```
y <- 50; mu <- 40; sigma <- 15
1/sqrt(2*pi*sigma^2) * exp(-0.5 * (y-mu)^2/sigma^2)
#> [1] 0.02129653
```

or just

```
dnorm(y, mean = mu, sd = sigma)
#> [1] 0.02129653
```

Conditional probability of any observation

- We can use the R function `prob_obs_lm` (from `utils.R`) for the probability of an observation in any (simple) linear regression:

```
# e.g.
prob_obs_lm(y = 50,
            x = 15,
            beta_0 = -20, beta_1 = 4, sigma = 15)
#> [1] 0.02129653
```

Conditional probability of all observed data

- 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
$$P(y_1 \ldots y_n | x_1 \ldots x_n, \beta_0, \beta_1, \sigma).$$
- In this model, all y's are conditionally independent of one another, given that values of $x_1, x_2, x_3 \ldots x_n$, so the the joint probability is as follows:
$$P(y_1 \ldots y_n | x_1 \ldots x_n, \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} P(y_i | x_i, \beta_0, \beta_1, \sigma).$$

Conditional log probability of all observed data

- The joint probability

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

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

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

Log conditional probability of all observed data

- For example, the log probability of all the `dist` values given the `speed` values, and assuming certain values for β_0, β_1, σ can be calculated as follows:

```
y <- cars$dist; x <- cars$speed
beta_0 = -20; beta_1 = 4; sigma = 15
prob_obs_lm(y, x, beta_0, beta_1, sigma, log = TRUE) %>%
  sum()
#> [1] -206.805
```

Log conditional probability of all observed data

- The log conditional probability can

$$\sum_{i=1}^{n} \log P(y_i | x_i, \beta_0, \beta_1, \sigma) = \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{|y_i - \mu_i|^2}{2\sigma^2} \right) \right),$$

$$= -\frac{n}{2} \log (2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} |y_i - \mu_i|^2,$$

where $\mu_i = \beta_0 + \beta_1 x_i$.

- This is calculated by `log_prob_obs_lm` (in `utils.R`):

```
log_prob_obs_lm(y, x, beta_0, beta_1, sigma)
#> [1] -206.805
```

The likelihood function

- The following is a function over the space of values of $y_1 \ldots y_n$:

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

- In other words, $x_1 \ldots x_n$ and β_0, β_1, and σ are fixed (like the parameters of a function) and $y_1 \ldots y_n$ are free variables and so $P(y_1 \ldots y_n | x_1 \ldots x_n, \beta_0, \beta_1, \sigma)$ is a function over the $y_1 \ldots y_n$ space.
- If, however, we treat $y_1 \ldots y_n$ and $x_1 \ldots x_n$ as fixed, and treat β_0, β_1, and σ as free variables, then

$$\mathcal{L}(\beta_0, \beta_1, \sigma | \vec{y}, \vec{x}) = \prod_{i=1}^{n} P(y_i | x_i, \beta_0, \beta_1, \sigma)$$

defines a function over the three dimensional β_0, β_1, σ space.
- The function is known as the *likelihood function*.

The log likelihood function

- The log likelihood function is just the log of the likelihood function.
- In the present example, it is

$$\log \mathcal{L}(\beta_0, \beta_1, \sigma | \vec{y}, \vec{x}) =$$

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|^2,$$

where $y_1 \ldots y_n$ and $x_1 \ldots x_n$ are assumed to be fixed.

Maximum likelihood estimation

- The values of β_0, β_1, σ that maximize $\mathcal{L}(\beta_0, \beta_1, \sigma | \vec{y}, \vec{x})$ are the maximum likelihood estimators of the (random variables) β_0, β_1, σ.
- The values of β_0, β_1, σ that maximize $\log \mathcal{L}(\beta_0, \beta_1, \sigma | \vec{y}, \vec{x})$ are those that maximize $\mathcal{L}(\beta_0, \beta_1, \sigma | \vec{y}, \vec{x})$.
- We usually denote estimators by $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}$.
- By definition, the maximum likelihood estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}$ are the values of β_0, β_1, σ that maximize the probability of the observed data.

Maximum likelihood estimation

- In a simple linear regression, the maximum likelihood estimators for the linear coefficients are those that minimize the following

$$\sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|^2,$$

- In general, for linear regression, the maximum likelihood estimators always minimize the sum of squared residuals.
- In R, for the `cars` data, the maximum likelihood estimators for β_0 and β_1 are obtained as follows:

```
M1 <- lm(dist ~ speed, data = cars)
coef(M1)
#> (Intercept)       speed
#>  -17.579095    3.932409
```

Maximum likelihood estimation

- The maximum likelihood estimator for σ^2 is the following:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)|^2,$$

- Using R, we can obtain this as follows:

```
mean(residuals(M1)^2)
#> [1] 227.0704
```

and so the maximum likelihood estimator for σ is

```
mean(residuals(M1)^2) %>% sqrt()
#> [1] 15.06886
```

The model's log likelihood

- When we speak of the log likelihood of a model, we mean the maximum value of the model's log likelihood function.
- In other words, it is the value of the log likelihood function at its maximum likelihood estimators' values.
- In yet other words, **it is the (log) probability of the observed data given the maximum likelihood estimates of its parameters**.

The model's log likelihood

```r
beta_0_mle <- coef(M1)['(Intercept)']
beta_1_mle <- coef(M1)['speed']
sigma_mle <- mean(residuals(M1)^2) %>% sqrt()

log_prob_obs_lm(y, x,
                beta_0 = beta_0_mle,
                beta_1 = beta_1_mle,
                sigma = sigma_mle) %>% sum()
#> [1] -206.5784
# same as ...
logLik(M1)
#> 'log Lik.' -206.5784 (df=3)
```

Residual sum of squares

- The sum of squared residuals in a simple linear model is

$$\mathrm{RSS} = \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|^2.$$

- The RSS when using the maximum likelihood estimators is

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

Residual sum of squares and log likelihood

- The residual sum of squares (RSS) when using the maximum likelihood estimators is

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

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

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

Residual sum of squares and log likelihood

```r
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)
```

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):

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

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

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

$$y_i \sim \text{Bernoulli}(\theta_i), \quad \text{for } i \in 1 \ldots n$$
$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 x_i$$

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

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

```r
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
```