Extended Logistic Regression Example

Fundamental Techniques in Data Science

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Estimation

Interpretation

In Equations

Visualization

Example

Let's use logistic regression to predict the chances that Titanic passengers survived the sinking based on their age, sex, and ticket class.

```r
## Read the data:
titanic <- readRDS(here::here(dataDir, "titanic.rds"))

## Estimate the logistic regression model:
fit <- glm(survived ~ age + sex + class,
           data = titanic,
           family = "binomial")
```

Example

```
partSummary(fit, -1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.63492    0.37045   9.812  < 2e-16
age         -0.03427    0.00716  -4.787 1.69e-06
sexmale     -2.58872    0.18701 -13.843  < 2e-16
class2nd    -1.19911    0.26158  -4.584 4.56e-06
class3rd    -2.45544    0.25322  -9.697  < 2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.77  on 886  degrees of freedom
Residual deviance:  801.59  on 882  degrees of freedom
AIC: 811.59

Number of Fisher Scoring iterations: 5
```

Interpretation

Intercept

The expected log-odds of survival for a zero-year-old female passenger in first class are 3.63.

Age Effect

A passenger's age significantly predicts their probability of survival, after controlling for their gender and ticket class ($\beta = -0.03$, $z = -4.79$, $p < 0.001$).

- For each additional year of age, the expected log-odds of survival decrease by 0.03 units, after controlling for gender and ticket class.

Interpretation

Gender Effect

A passenger's gender significantly predicts their probability of survival, after controlling for their age and ticket class ($\beta = -2.59$, $z = -13.84$, $p < 0.001$).

- The expected log-odds of survival are 2.59 units lower for men/boys than for women/girls, after controlling for age and ticket class.

Interpretation

Class Effect

After controlling for age and gender, there is a significant difference in predicted survival probability between passengers in second class and passengers in first class ($\beta = -1.2$, $z = -4.58$, $p < 0.001$) and between passengers in third class and passengers in first class ($\beta = -2.46$, $z = -9.7$, $p < 0.001$).

- The expected log-odds of survival are 1.2 units lower for passengers in second class than for passengers in first class, after controlling for age and gender.
- The expected log-odds of survival are 2.46 units lower for passengers in third class than for passengers in first class, after controlling for age and gender.

Example

Compute odds ratios.

```
(or <- coef(fit) |> exp())

(Intercept)         age     sexmale    class2nd    class3rd 
 37.8988400   0.9663058   0.0751161   0.3014609   0.0858252 
```

Odds ratios smaller than 1.0 can be difficult to explain.

- We can ease interpretation by reciprocating the estimates.

```
1 / or

(Intercept)         age     sexmale    class2nd    class3rd 
 0.02638603  1.03486914 13.31272574  3.31717996 11.65158920 
```

Example

To convince ourselves that the above operation is sensible, we can compare the inverse odds ratios to the odds ratios we get from predicting the chances of dying.

```r
library(dplyr)
library(magrittr)

fit2 <- titanic |>
    mutate(died = relevel(survived, ref = "yes")) %$%
    glm(died ~ age + sex + class, family = "binomial")
```

Example

```
partSummary(fit2, -1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.63492    0.37045  -9.812  < 2e-16
age          0.03427    0.00716   4.787 1.69e-06
sexmale      2.58872    0.18701  13.843  < 2e-16
class2nd     1.19911    0.26158   4.584 4.56e-06
class3rd     2.45544    0.25322   9.697  < 2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.77  on 886  degrees of freedom
Residual deviance:  801.59  on 882  degrees of freedom
AIC: 811.59

Number of Fisher Scoring iterations: 5
```

Example

We get the same odds ratios that we derived through reciprocation.

```
coef(fit2) |> exp()

(Intercept)         age     sexmale    class2nd    class3rd
 0.02638603  1.03486914 13.31272574  3.31717996 11.65158920

1 / or

(Intercept)         age     sexmale    class2nd    class3rd
 0.02638603  1.03486914 13.31272574  3.31717996 11.65158920
```

Interpretation

Intercept

- The expected odds of survival for a zero-year-old female passenger in first class are 37.9.
- The expected odds of death for a zero-year-old female passenger in first class are 0.03.

Interpretation

AGE EFFECT

- For each additional year of age, the expected odds of survival change by a factor of 0.97 times, after controlling for gender and ticket class.
- For any two passengers with a one year age difference, the expected odds of the older passenger surviving are 0.97 times the expected odds of the younger passenger surviving, after controlling for their genders and ticket classes.
- For each additional year of age, the expected odds of death are 1.03 times higher, after controlling for gender and ticket class.

Interpretation

Gender Effect

- The expected odds of survival for men/boys are 0.08 times the expected odds of survival for women/girls, after controlling for age and ticket class.
- The expected odds of death are 13.31 times higher for men/boys than for women/girls, after controlling for age and ticket class.

Interpretation

Class Effect

- The expected odds of survival for passengers in second class are 0.3 times the expected odds of survival for passengers in first class, after controlling for gender and age.
- The expected odds of death are 3.32 times higher for passengers in second class than for passengers in first class, after controlling for gender and age.
- The expected odds of survival for passengers in third class are 0.09 times the expected odds of survival for passengers in first class, after controlling for gender and age.
- The expected odds of death are 11.65 times higher for passengers in third class than for passengers in first class, after controlling for gender and age.

Example in Equations

Here’s the symbolic representation of our logistic regression model:

$$\text{logit}(\pi_{died}) = \beta_0 + \beta_1 X_{age} + \beta_2 X_{male} + \beta_3 X_{2nd} + \beta_4 X_{3rd}$$

By fitting this model to the *titanic* data we get:

$$\text{logit}(\hat{\pi}_{died}) = -3.63 + 0.03X_{age} + 2.59X_{male} + 1.2X_{2nd} + 2.46X_{3rd}$$

Exponentiating the coefficients produces:

$$\frac{\hat{\pi}_{died}}{1 - \hat{\pi}_{died}} = \frac{\hat{\pi}_{died}}{\hat{\pi}_{survived}} = 0.03 \times 1.03^{X_{age}} \times 13.31^{X_{male}} \times 3.32^{X_{2nd}} \times 11.65^{X_{3rd}}$$

Exponentiating the Systematic Component

$$\text{logit}(\hat{\pi}_{died}) = -3.63 + 0.03X_{age} + 2.59X_{male} + 1.2X_{2nd} + 2.46X_{3rd}$$

$$e^{\text{logit}(\hat{\pi}_{died})} = e^{\left(-3.63 \, + \, 0.03X_{age} \, + \, 2.59X_{male} \, + \, 1.2X_{2nd} \, + \, 2.46X_{3rd}\right)}$$

$$\frac{\hat{\pi}_{died}}{\hat{\pi}_{survived}} = e^{-3.63} \times e^{0.03X_{age}} \times e^{2.59X_{male}} \times e^{1.2X_{2nd}} \times e^{2.46X_{3rd}}$$

$$= e^{-3.63} \times \left(e^{0.03}\right)^{X_{age}} \times \left(e^{2.59}\right)^{X_{male}} \times \left(e^{1.2}\right)^{X_{2nd}} \times \left(e^{2.46}\right)^{X_{3rd}}$$

$$= 0.03 \times 1.03^{X_{age}} \times 13.31^{X_{male}} \times 3.32^{X_{2nd}} \times 11.65^{X_{3rd}}$$

Visualization

We can visualize the model's predictions by generated predicted values for hypothetical passengers.

- We'll generate a set of hypothetical passengers whose attributes span the possible values in the real data.

```r
passengers <-
    expand.grid(
        age   = 1:100,
        sex   = c("male", "female"),
        class = c("1st", "2nd", "3rd")
    ) |>
    data.frame(stringsAsFactors = TRUE)
```

Visualization

View 10 random passengers.

```
slice_sample(passengers, n = 10)

   age    sex class
1   70 female   3rd
2   39   male   2nd
3   71   male   2nd
4   97 female   1st
5   28 female   1st
6    3 female   1st
7   72   male   2nd
8   21 female   1st
9   80   male   1st
10  26   male   3rd
```

Visualization

We can generate predictions on the scale of the linear predictor (i.e., log-odds), or we can generate predicted probabilities.

```r
passengers %<>%
    mutate(
        ## Predicted log odds of dying:
        etaHat = predict(fit2, newdata = ., type = "link"),

        ## Predicted probabilities of dying:
        piHat = predict(fit2, newdata = ., type = "response")
    )
```

We can then use the predicted probabilities of dying to classify:

```r
passengers %<>%
    mutate(dieHat = ifelse(piHat > 0.5, "dead", "alive") |> factor())
```

Visualization

View the predictions for 10 random passengers:

```r
slice_sample(passengers, n = 10)
```

```
   age    sex class     etaHat      piHat dieHat
1   94 female   1st -0.4130721 0.39817572  alive
2   32   male   2nd  1.2497144 0.77725041   dead
3   52 female   2nd -0.6535064 0.34219982  alive
4   36 female   1st -2.4010211 0.08309486  alive
5   98   male   2nd  3.5118632 0.97102344   dead
6    8 female   2nd -2.1616056 0.10325169  alive
7   54   male   3rd  3.2600916 0.96303405   dead
8   74 female   2nd  0.1005432 0.52511466   dead
9   35 female   2nd -1.2361811 0.22510142  alive
10  37   male   3rd  2.6774168 0.93568084   dead
```

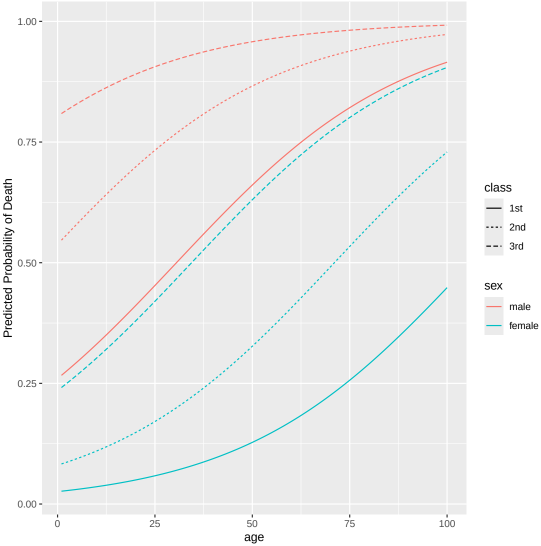
Visualization

We can visualize the predicted probabilities of dying.

```r
library(ggplot2)

ggplot(passengers, aes(age, piHat, color = sex)) +
  geom_line(aes(linetype = class)) +
  ylab("Predicted Probability of Death")
```

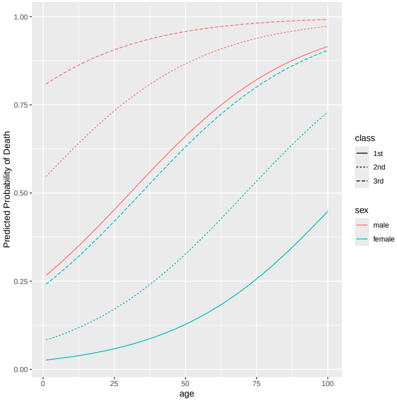
Visualization



Visualization

Our model reflects very strong sociocultural dynamics:

- For any given age, males are less likely to survive than females.
- For young males, the model predicts large differences in survival rates between classes.
  - These differences diminish for older males.
- For females, class strongly predicts survival rates, regardless of age.

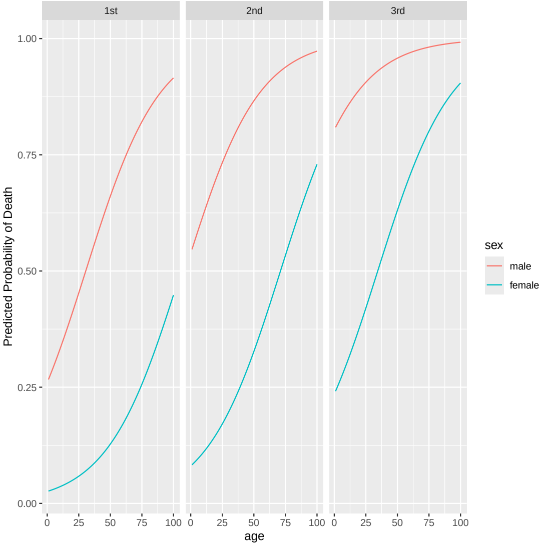


Visualization

Alternatively, we could facet on the class factor.

```r
ggplot(passengers, aes(age, piHat, color = sex)) +
  geom_line() +
  facet_wrap(vars(class)) +
  ylab("Predicted Probability of Death")
```

Visualization

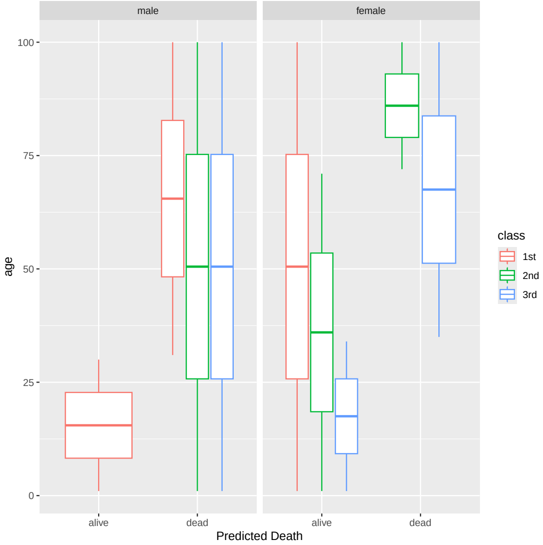


Visualization

We can also visualize the classifications the model would make.

```r
ggplot(passengers, aes(dieHat, age, color = class)) +
  geom_boxplot() +
  facet_wrap(vars(sex)) +
  xlab("Predicted Death")
```

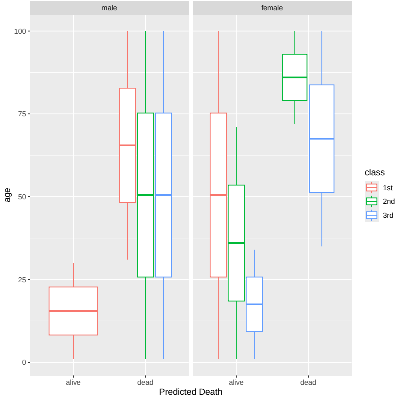
Visualization



Visualization

Our model has some very strong opinions:

- For males, the model only predicts survival for 1st class.
- For females, the model never predicts death for 1st class.


male
female
age
alive
dead
alive
dead
Predicted Death
class
1st
2nd
3rd