IDS 702

Logistic Regression - 2 (Estimation, interpretation, prediction)

Recall estimation procedure for MLR

Maximum likelihood estimation

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_i))$$

The **likelihood function** describes the joint probability of the observed outcome as a function of the input data and parameters of the chosen statistical model

Example data

Pumpkins!!!

- Want to characterize differences in two classes of pumpkin seeds
- Outcome: Ürgüp sivrisi or çerçevelik pumpkin seeds
- Predictors: area, perimeter, major axis length, minor axis length, convex area, diameter, eccentricity, solidity, extent, roundness, aspect ratio, compactness

KOKLU, M., SARIGIL, S., & OZBEK, O. (2021). The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.). Genetic Resources and Crop Evolution, 68(7), 2713-2726. Doi: https://doi.org/10.1007/s10722-021-01226-0

Interpreting coefficients

Back to odds ratios: start with binary predictor

Interpreting coefficients

Binary predictor

```
> summary(glm(class.fac~Area.bin,data=pumpkin,family=binomial))

Call:
glm(formula = class.fac ~ Area.bin, family = binomial, data = pumpkin)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.256  -1.035  -1.035   1.101   1.327

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34581    0.05742  -6.023 1.71e-09 ***
Area.bin1    0.52871    0.08077   6.546 5.91e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpreting coefficients

Categorical (>2 categories) predictor

```
> summary(glm(class.fac~Area.fac,data=pumpkin,family=binomial))

Call:
glm(formula = class.fac ~ Area.fac, family = binomial, data = pumpkin)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.396  -1.090  -1.013   1.267   1.351

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.39880    0.08160  -4.888 1.02e-06 ***
Area.fac2    0.19005    0.09946   1.911    0.056 .
Area.fac3    0.89855    0.11604   7.743 9.69e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpreting coefficients

Continuous predictor

```
> summary(glm(class.fac~Perimeter,data=pumpkin,family=binomial))

Call:
glm(formula = class.fac ~ Perimeter, family = binomial, data = pumpkin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1901  -1.0074  -0.6198   1.0380   2.1703

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.4819243  0.5218385  -18.17   <2e-16 ***
Perimeter    0.0083209  0.0004609   18.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-values

What is the null value?

```
> summary(glm(class.fac~Area.bin,data=pumpkin,family=binomial))

Call:
glm(formula = class.fac ~ Area.bin, family = binomial, data = pumpkin)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.256  -1.035  -1.035   1.101   1.327

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34581    0.05742  -6.023 1.71e-09 ***
Area.bin1    0.52871    0.08077   6.546 5.91e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confidence intervals

Know your scale!

```
> confint(glm(class.fac~Area.bin,data=pumpkin,family=binomial))
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) -0.4587311 -0.2336035
Area.bin1    0.3706665  0.6873130
> exp(confint(glm(class.fac~Area.bin,data=pumpkin,family=binomial)))
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.6320852 0.7916757
Area.bin1   1.4486999 1.9883657
```

Predictions

Predict probabilities or Y values?

```
> pumpkin.mod <- glm(class.fac~Perimeter,data=pumpkin,family=binomial)
> pumpkin$predprobs <- predict(pumpkin.mod,type="response")
> plot(pumpkin$predprobs)
```