Day 8 - Fit Statistics

David Carlson

November 20, 2020

Fit Statistics

Pearson statistic

$$X^2 = \sum_{i=1}^{n} R^2_{Pearson} = \sum_{i=1}^{n} \left[\frac{Y - \mu}{\sqrt{Var[\mu]_i}} \right]^2$$

Statistic is distributed

$$\chi^2_{n-p}$$

. Useful for large sample sizes, and testing nested models.

```
xsq = sum(residuals(pop_logit, type = "pearson")^2)
pchisq(xsq, pop_logit$df.residual, lower.tail = F) #we are in the tails, but this is not really an appr
```

```
## [1] 0.03686815
```

Compare nested models: Compare to the null.

```
pop_logit_null = glm(intercon ~ aggdifxx + gdppc + polity2,
                 data=data,
                 family = binomial)
anova(pop_logit, pop_logit_null, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: intercon ~ country_pop + aggdifxx + gdppc + polity2
## Model 2: intercon ~ aggdifxx + gdppc + polity2
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      8211      10786
## 2      8212      11014 -1  -227.58 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Akaike Information Criterion (AIC)

Again, useful when $n >> p$. Also useful for comparing non-nested models.

$$\text{AIC} = -2\ell(\hat{\theta}|y) + 2p$$

```
pop_logit$aic
```

```
## [1] 10796.42
```

```
pop_logit_null$aic #larger, therefore 'worse'
```

```
## [1] 11022
```

Bayesian Information Criterion (BIC)

AIC favors more explanatory variables and greater fit, BIC favors more parsimonious models. BIC is based on Bayesian principles, but is an approximation appropriate for frequentist methods.

$$\text{BIC} = -2\ell(\hat{\theta}|y) = p\log(n)$$

```
BIC(pop_logit)
```

```
## [1] 10831.49
```

```
BIC(pop_logit_null)
```

```
## [1] 11050.06
```

Bootstrapping

We have not covered this yet. Even though it is not in the book or plan, it is an important concept to understand. Bootstrapping is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand. Bootstrapping offers advantages: - The bootstrap is quite general, although there are some cases in which it fails. - Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small. - It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically. - It is relatively simple to apply the bootstrap to complex data-collection plans (such as stratified and clustered samples).

```
iters = 1000
coefs = matrix(nrow = iters, ncol = 5)
dataComplete = na.omit(data[, c('intercon', 'country_pop', 'aggdifxx', 'gdppc', 'polity2')])
for(i in 1:iters){ #with such a simple model, a for loop is fine, but parallelize for more complex mode
  dataSub = dataComplete[sample(1:nrow(dataComplete), nrow(dataComplete), replace = T),]
  tempMod = glm(intercon ~ country_pop + aggdifxx + gdppc + polity2,
                data=dataSub,
                family = binomial)
  coefs[i, ] = tempMod$coefficients
}
ests = apply(coefs, 2, mean)
stdError = apply(coefs, 2, sd)
summary(pop_logit)
```

```
##
## Call:
## glm(formula = intercon ~ country_pop + aggdifxx + gdppc + polity2,
##     family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7248  -1.2371   0.9087   1.0745   1.3535
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.071172   0.058710  -1.212    0.225
## country_pop  0.451093   0.038043  11.857  < 2e-16 ***
```

```
## aggdifxx     0.037667   0.005560   6.775 1.25e-11 ***
## gdppc       -0.141813   0.035302  -4.017 5.89e-05 ***
## polity2      0.017915   0.003707   4.833 1.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11183  on 8215  degrees of freedom
## Residual deviance: 10786  on 8211  degrees of freedom
##   (8652 observations deleted due to missingness)
## AIC: 10796
##
## Number of Fisher Scoring iterations: 5
```

```
cbind(ests, stdError)
```

```
##             ests    stdError
## [1,] -0.07167112 0.059308711
## [2,]  0.45434747 0.042135995
## [3,]  0.03785714 0.005718328
## [4,] -0.14015516 0.034176916
## [5,]  0.01796151 0.003718941
```

```
cbind(ests, stdError, pop_logit$coefficients, summary(pop_logit)$coefficients[, 2])
```

```
##                    ests    stdError
## (Intercept) -0.07167112 0.059308711 -0.07117167 0.058709797
## country_pop  0.45434747 0.042135995  0.45109331 0.038042914
## aggdifxx     0.03785714 0.005718328  0.03766738 0.005560111
## gdppc       -0.14015516 0.034176916 -0.14181284 0.035301669
## polity2      0.01796151 0.003718941  0.01791514 0.003706969
```