Building Logistic Regression Models

AIC, Multicolliniearity, and Stepwise Methods

Demetris Athienitis

UF | UNIVERSITY of FLORIDA

Section 1

AIC

The AIC is an estimator of the relative quality of statistical models for a given set of data. Somewhat analogous to R^2-adjusted.

$$\mathrm{AIC} = 2(k+1) - 2\log(\hat{L})$$

Comprised of

- *"penalizing"* function $2(k+1)$ that penalizes for complicated models with a large k value, i.e. number of parameters
- maximum value of the likelihood function for the model, $\hat{L}$, so better fitting models have larger $\hat{L}$

Due to the "minus" sign in front og $\log(\hat{L})$, smaller values are desirable when comparing models.

AICc

AIC-correction (AICc) was developed that includes a correction for small sample sizes.

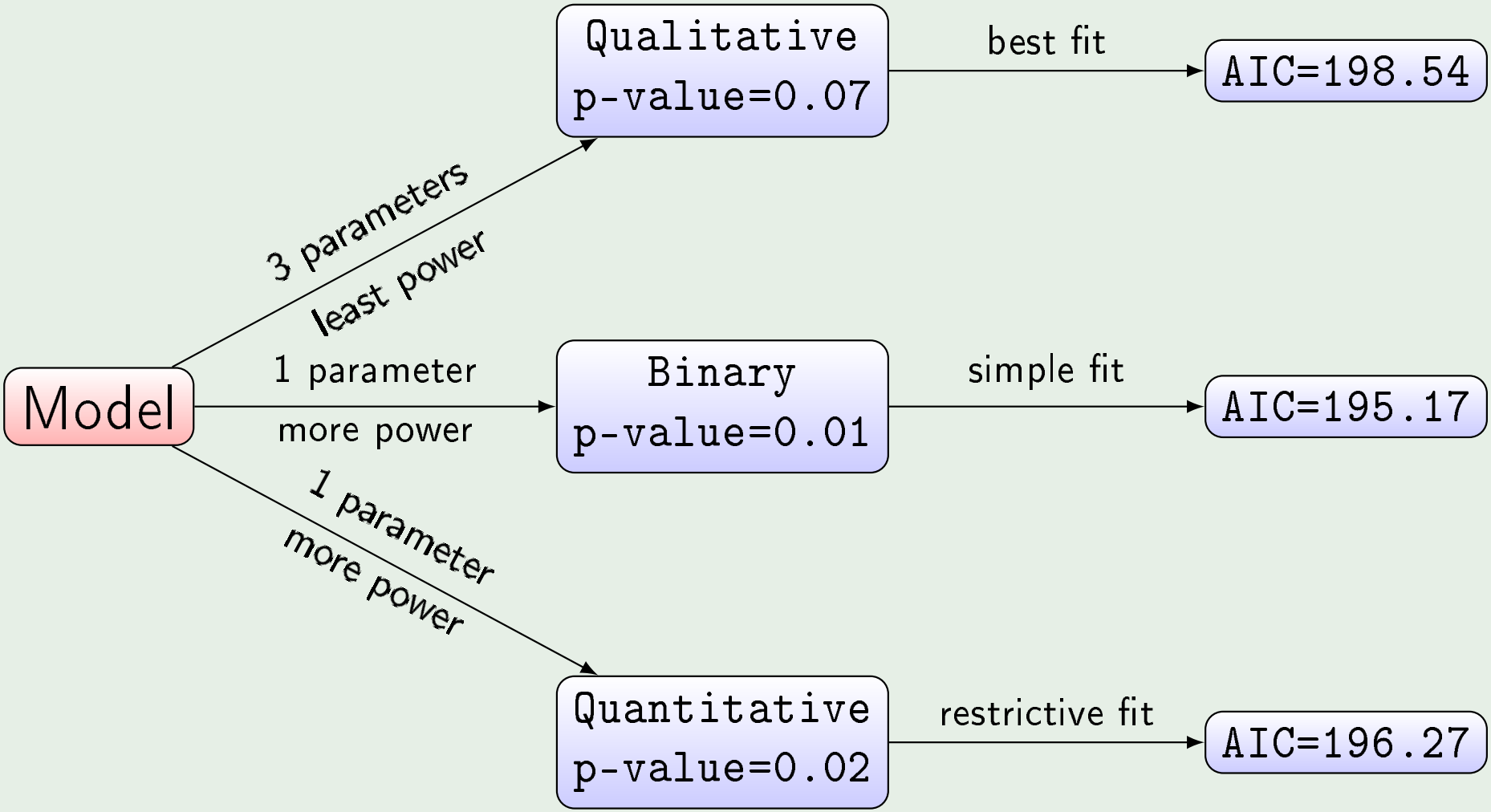
$$\text{AICc} = \text{AIC} + \frac{2(k+1)^2 + 2(k+1)}{n-k-2}$$

Remark

Other criteria exist such as the Bayesian Infromation Criterion (BIC) which simply have different penalizing functions.

Example (Horseshoe crab continued)

Results best illustrated on how "color" variable was used



Section 2

Multicollinearity

Multicollinearity is a phenomenon in which one predictor variable can be linearly predicted from the other predictors with a substantial degree of accuracy.
Effects:

- Coefficient estimates may change erratically in response to small changes in the model or the data
- Coefficient standard errors are inflated

Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.

Variance Inflation Factor

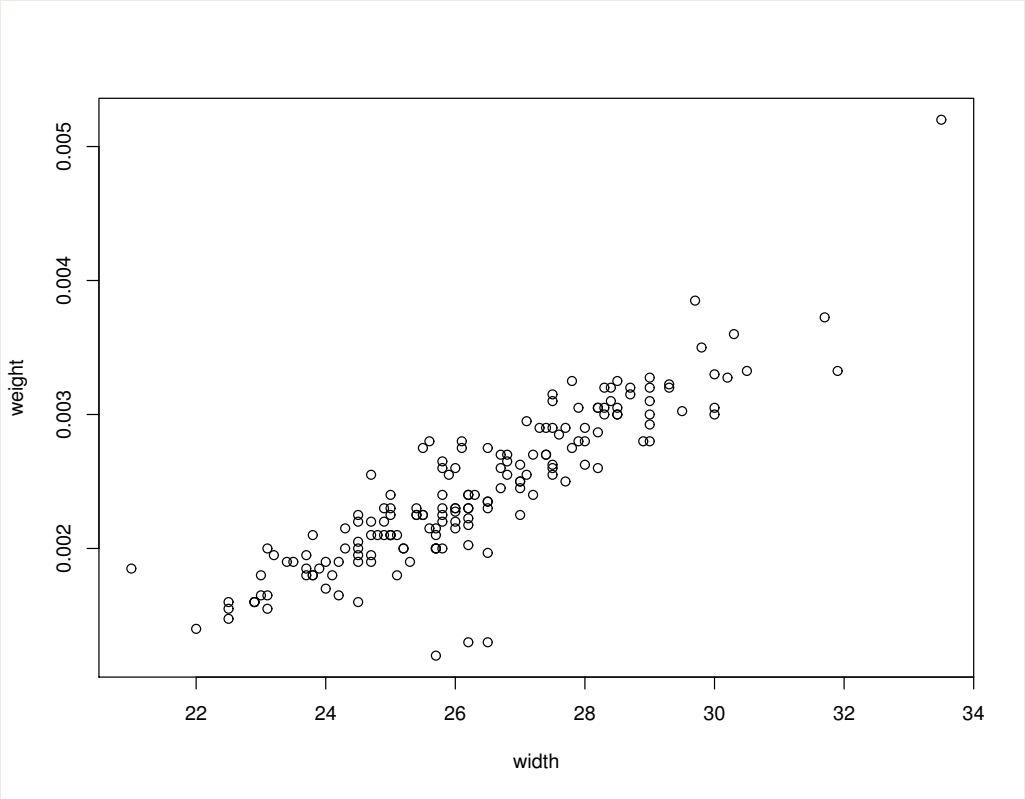
A useful tool is the *Variance Inflation Factor* (VIF). The square root of the variance inflation factor tells you how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model.

If the variance inflation factor of a predictor variable were 5.27 ($\sqrt{5.27} = 2.3$) this means that the standard error for the coefficient of that predictor variable is 2.3 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables.

Example (Horseshoe crab continued)

Consider the weight and width of a crab that are likely to be correlated. Could use either, but will see that width is slightly better.

```
> cor(weight,width)
[1] 0.8868715
```



Example (continued)

Only weight

```
> fit.we=glm(y ~ weight, family=binomial(link=logit))
> summary(fit.we)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
AIC: 199.74
```

Only width

```
> fit.wi=glm(y ~ width, family=binomial(link=logit))
> summary(fit.wi)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
```

Section 3

Stepwise Selection Algorithms

- **Backward** - Start with a full model and *remove* 1 factor/predictor at a time, based on a criterion, until stopping is reached.
- **Forward** - Start with a reduced simple model and *add* 1 factor/predictor at a time, based on a criterion, until stopping is reached.
- **Both** - Start with any model (of varying complexity) and at each step add or remove a variable.

Common criteria include (but not limited to)

- AIC
- L.R.T. p-values

Example (Horseshoe crab continued)

Consider model with 2 way interactions as the fullest and model with no predictors as most reduced.

```
> stepAIC(fit.w,scope=list(upper=~width*color+width*spine+
  color*spine, lower=~1),direction="both")


Start:  AIC=198.45
y ~ width

         Df Deviance    AIC
+ color   3   187.46 197.46
<none>        194.45 198.45
+ spine   2   194.43 202.43
- width   1   225.76 227.76


Step:  AIC=197.46
y ~ width + color
```

Number of predictors

Remark

There is a study that suggests ≥ 10 outcomes of each type per model predictor (where dummy variables for qualitative predictors are considered individual predictors).

Example (Horseshoe crab)

In this example there were 173 crabs, 111 had a male satellite while 62 did not. Hence, choosing the smaller count of the two

$$\frac{62}{10} \approx 6 \text{ predictors}$$

That is why we do not fit a model with all 3-way interactions. Fit and you will fins some SE are huge and some estimates are NA.

We learned

- AIC
- Multicollinearity and VIF
- AIC stepwise model selection criterion