DIAGNOSTICS & SUBSET SELECTION – EXTRA CONTENT

Dr. Aric LaBarr

Institute for Advanced Analytics

SUBSET SELECTION METHODS

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values…

$$AIC = -2\log(L) + 2p$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values…

$$AIC = -2\log(L) + 2p$$

Model better with lower AIC…

$$AIC_{p+1} < AIC_p$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values…

$$AIC = -2\log(L) + 2p$$

Model better with lower AIC…

$$AIC_{p+1} < AIC_p$$

$$-2\log(L_{p+1}) + 2(p+1) < -2\log(L_p) + 2(p)$$

$$2 < 2(\log(L_{p+1}) - \log(L_p))$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values…

$$AIC = -2\log(L) + 2p$$

Model better with lower AIC…

$$AIC_{p+1} < AIC_p$$

$$-2\log(L_{p+1}) + 2(p+1) < -2\log(L_p) + 2(p)$$

$$2 < 2(\log(L_{p+1}) - \log(L_p))$$

LRT!

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values…

$$AIC = -2\log(L) + 2p$$

Model better with lower AIC…

$$AIC_{p+1} < AIC_p$$

$$-2\log(L_{p+1}) + 2(p+1) < -2\log(L_p) + 2(p)$$

$$2 < \chi_1^2$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values…

$$AIC = -2\log(L) + 2p$$

Model better with lower AIC…

$$AIC_{p+1} < AIC_p$$

$$-2\log(L_{p+1}) + 2(p+1) < -2\log(L_p) + 2(p)$$

$$2 < \chi_1^2$$

Model better with variable below p-value…

$$1 - P(\chi_1^2 > 2) = 0.1573$$

P-value vs. BIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values…

$$BIC = -2\log(L) + p \times \log(n)$$

Model better with lower BIC…

$$BIC_{p+1} < BIC_p$$

$$-2\log(L_{p+1}) + \log(n)\,(p+1) < -2\log(L_p) + \log(n)\,(p)$$

$$\log(n) < \chi_1^2$$

Model better with variable below p-value…

$$1 - P(\chi_1^2 > \log(n)) = \cdots$$

P-value vs. BIC Selection

- For our birth weight data set, BIC selection is the same as the p-value selection with the following alpha:

$$1 - P\left(\chi_1^2 > \log(n)\right) = 1 - P\left(\chi_1^2 > \log(189)\right) = 0.022$$

- Lot of attention being given to p-values and how other selection techniques are better.
- Attention **should** be on significance level (α), **not** on p-value.
- **DON'T ALWAYS USE 0.05!**

GOODNESS-OF-FIT

Calibration

- **Calibration** measures how well predicted probabilities agree with actual frequency counts of outcomes.
- Helps detect bias!
  - Are predictions systematically too low or too high?

Calibration Curve

- Curve **above** 45° line indicates the model is predicting **lower** probabilities than actually observed.
- Curve **below** 45° line indicates the model is predicting **higher** probabilities than actually observed.

- Caveat:
  - Calibration depends on the observed proportion of events in the data, so models will likely have poor calibration on out-of-sample data.
  - Best used for goodness-of-fit in training, not on validation.

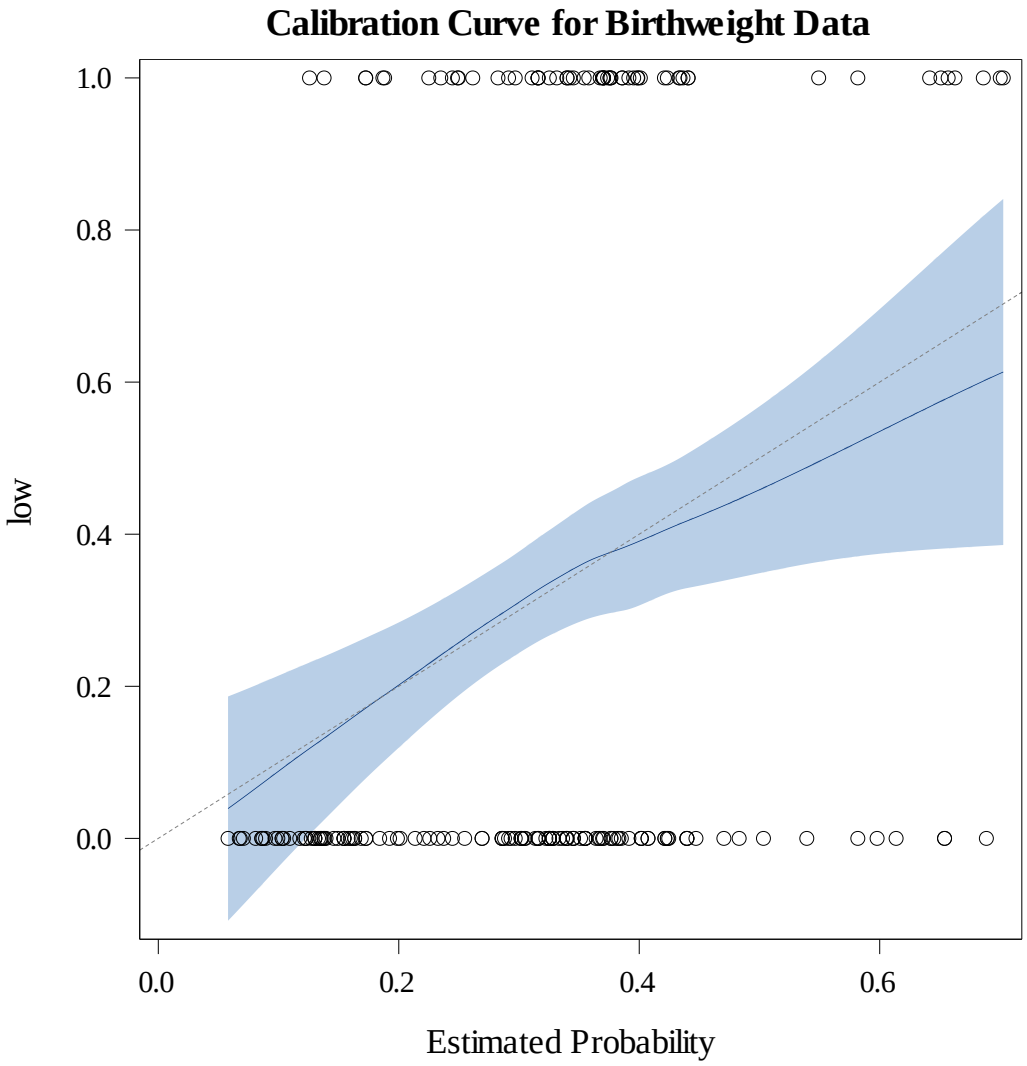
Calibration Curve – SAS

```
proc logistic data=logistic.lowbwt plots=effect;
   class race(ref='white') / param=ref;
   model low(event='1') = race lwt smoke;
   output out=cali predicted=PredProb;
run;

proc sort data=cali;
   by PredProb;
run;

proc sgplot data=cali noautolegend aspect=1;
   loess x=PredProb y=low / interpolation=cubic clm;
   lineparm x=0 y=0 slope=1 / lineattrs=(color=grey
                              pattern=dash);
   title 'Calibration Curve for Birthweight Data';
run;
```

Calibration Curve– SAS



Calibration Curve – R

```
cali.curve <- givitiCalibrationBelt(o = bwt$low,
                                    e = predict(logit.model,
                                        type = "response"),
                                    devel = "internal",
                                    maxDeg = 5)


plot(cali.curve, main = "Birth Weight Model Calibration Curve",
                 xlab = "Predicted Probability",
                 ylab = "Observed Low Birth Weight")
```

Calibration Curve – R