Quant III

Lab 4: MLE Application

Junlong Aaron Zhou

October 02, 2020

Outline

- Monte Carlo, Bootstrap
- Model Tests
- Predictive accuracy
- Seperation, Overdispersion

Monte Carlo, Bootstrap Revisited

- What we want: some quantities from a probalisitic distribution.
- Do we have the distribution?
- Yes, then we can do Monte Carlo. E.g. $\hat{\beta}^* \sim N(\hat{\beta}, \hat{\Sigma}^2_{\beta})$

Monte Carlo, Bootstrap Revisited

- What we want: some quantities from a probalisitic distribution.
- Do we have the distribution?
- Yes, then we can do Monte Carlo. E.g. $\hat{\beta}^* \sim N(\hat{\beta}, \hat{\Sigma}^2_{\beta})$
- No? Bootstap.
- Idea: any estimate we care about is a function of data. E.g. $\hat{\beta} = h(X, Y) = (X'X)^{-1}X'Y$
- If we keep sample $\{X, Y\}$ from population, we have the sampling distribution of $\hat{\beta}$.
- Nonparametric: sample the original data with replacement
- Parametric: sample from $f_{\hat{\beta}}(X)$

What quantity do we care?

- Suppose we care SAME:

$$SAME(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial x}\Pr(Y = 1|x, z_i)$$

- $S\hat{AM}E(X)$ is also a function of $\hat{\beta}$.
- Note we care about the $SAME$ for our given sample, the x we plug in is the x from original sample. However, we add variation from $\hat{\beta}$.

Which approach to choose?

- Do we have a parametric form? Can we sample directly?
- For regression: $\hat{\beta}$ and bootstrap has the same convergence rate.
- Recall $\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \sigma^2)$, so root-n rate.
- For empirical distribution, same as $I(x \leq t) \rightarrow F(t)$. It also has root-n rate.
- n refers to sample size.
- However, monte carlo simulation relies on the asymptotic distribution of $\hat{\beta}$, and problems emerge when Σ is misspecified.

AIC and BIC

- Let k be the number of parameters and n, the number of observations
- $AIC = -2ln\ L(\hat{\theta}; y) + 2k$ (Akaike information criterion)
- $BIC = -2ln\ L(\hat{\theta}; y) + kln\ n$ (Bayesian information criterion)
- Penalize complicated model

AIC and BIC

$\triangle$ BIC	Evidence against higher BIC
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
> 10	Very Strong

Prediction

- Consider a binary outcome $y_i \in \{0, 1\}$
- Our predicted value $\hat{y}_i \in \{0, 1\}$
- A general algorithm of prediction:

$$\hat{y}_i = \begin{cases} 0 & \hat{p}_i < \pi \\ 1 & \text{otherwise} \end{cases}$$

Prediction

Table 1: Confusion Matrix

	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Total
$y_i = 1$	a **TP**	b **FN**	$a + b$
$y_i = 0$	c **FP**	d **TN**	$c + d$
Total	$a + c$	$b + d$	N

- $\frac{c}{N}$ tells us type 1 error.
- $\frac{b}{N}$ tells us type 2 error.

Prediction, ctd

Table 2: Confusion Matrix

	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Total
$y_i = 1$	a **TP**	b **FN**	$a + b$
$y_i = 0$	c **FP**	d **TN**	$c + d$
Total	$a + c$	$b + d$	N

Prediction, ctd

Table 2: Confusion Matrix

	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Total
$y_i = 1$	a **TP**	b **FN**	$a + b$
$y_i = 0$	c **FP**	d **TN**	$c + d$
Total	$a + c$	$b + d$	N

- **Accuracy**: $\frac{\text{Number of correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$
- **Precision**: $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$

  Fraction of the cases predicted to be true, that were in fact true.

Prediction, ctd

Table 2: Confusion Matrix

	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Total
$y_i = 1$	a **TP**	b **FN**	$a + b$
$y_i = 0$	c **FP**	d **TN**	$c + d$
Total	$a + c$	$b + d$	N

- **Accuracy**: $\frac{\text{Number of correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$
- **Precision**: $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$

  Fraction of the cases predicted to be true, that were in fact true.
- **Recall**: $\frac{\text{number of TP}}{\text{number of TP} + \text{number of FN}} = \frac{a}{a+b}$

  Fraction of the cases that were in fact true, that method predicted were true.
- **F**:$2\frac{\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$ Harmonic mean of precision and recall.

Prediction ctd

Table 3: Confusion Matrix

	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Total
$y_i = 1$	a **TP**	b **FN**	$a + b$
$y_i = 0$	c **FP**	d **TN**	$c + d$
Total	$a + c$	$b + d$	N

- **Hit Rate/True Positive Rate**: $\frac{\text{number of TP}}{\text{number of TP+number of FN}} = \frac{a}{a+b}$
- **True Negative Rate**: $\frac{\text{number of TN}}{\text{number of FP + number of TN}} = \frac{d}{c+d}$
- **False Positive Rate/False Alarm Rate**: $\frac{\text{number of FP}}{\text{number of FP + number of TN}} = \frac{c}{c+d} = 1 - \text{True Negative Rate}$

Receiver Operating Curve (ROC)

Table 4: Confusion Matrix

	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Total
$y_i = 1$	a **TP**	b **FN**	$a + b$
$y_i = 0$	c **FP**	d **TN**	$c + d$
Total	$a + c$	$b + d$	N

- Recall our algorithm:

$$\hat{y}_i = \begin{cases} 0 & \hat{p}_i < \pi \\ 1 & \text{otherwise} \end{cases}$$

Receiver Operating Curve (ROC)

Table 4: Confusion Matrix

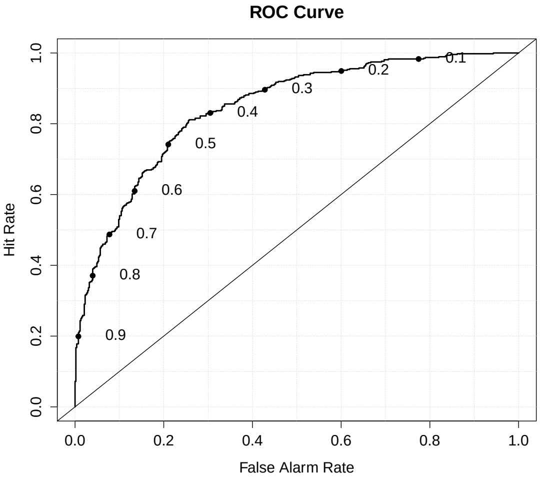
	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Total
$y_i = 1$	a **TP**	b **FN**	$a + b$
$y_i = 0$	c **FP**	d **TN**	$c + d$
Total	$a + c$	$b + d$	N

- Recall our algorithm:

$$\hat{y}_i = \begin{cases} 0 & \hat{p}_i < \pi \\ 1 & \text{otherwise} \end{cases}$$

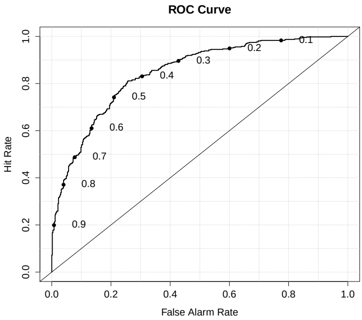
- When model is fixed (therefore $\hat{p}_i = E(y_i|x_i)$), we can only change π.
- We can change π to plot those statistics

Receiver Operating Curve (ROC)



- X-axis: False Alarm Rate
- Y-axis: Hit Rate

Receiver Operating Curve (ROC) ctd.



- Area Under Curve: varies between 0.5 (random draws) to 1 (perfect prediction).
- Larger AUC number means a better fit.

Separation

- Suppose y is binary and $y_i = 1(x_i < \tau)$.
- What would happen if we run `glm(y~x, family=binomial(link='logit')`?
- Coefficient of x and its variance approach infinity.

Penalized likelihood

- Add a penalty term to the likelihood:

$$L(\theta; y) - P(\theta),$$

  where P is a penalty function, typically increasing in $|\theta|$.
- Shrinks coefficients towards zero.
- Turns out very useful in prediction problems (relates to parsimony).
- Related: fixed-effects in binary dependent variable?

Over-dispersion

- It's a problem of model mis-specification
- When we use poisson, we impose the assumption that $E(Y) = Var(Y) = \lambda$
- It's might be violated.
- One approach to model the variance.

Over-dispersion ctd.

- Negative Binomial model is one approach
- $y \sim NB(n, p) = \binom{n+y-1}{y} p^y (1-p)^n$
- $E(y) = \frac{np}{1-p}$
- $Var(y) = \frac{np}{(1-p)^2}$

Over-dispersion ctd..

- Rewrite: $\mu = \frac{np}{1-p}$
- $Var(y) = \mu + \frac{\mu^2}{\beta}$
- where $\beta = np + (1 - p)$
- Degree of dispersion $\phi = 1 + \frac{1}{\beta}$
- Intuition: Poisson distribution is a binomial distribution when n approaches $+\infty$ and p is small
- Take away: Model specification is super important.