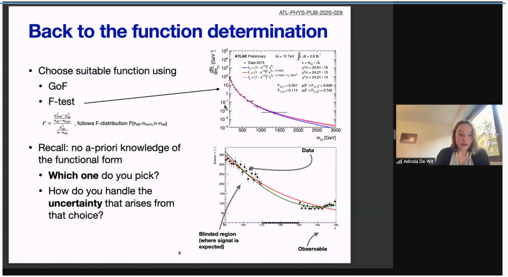
Methods of Applied Statistics I

STA2101H F LEC9101

Week 8

November 3 2021



Today

Start Recording

1. Upcoming events
2. Logistic Regression
3. Poisson Regression
4. In the News
5. Homework, Project (Hour 3)

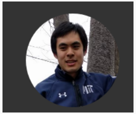
Upcoming

Reading Week!! Nov 8-13

- Thursday Nov 4 3.30

  Precise High-Dimensional Asymptotics for AdaBoost   Zoom Link

- Friday Nov 5 Toronto Data Workshop   Zoom link



Dear friends,

Toronto Data Workshop this Friday, 5 November, at noon (Toronto time) hosts Yun William Yu on the intersection of math and data science. Yun William Yu is an assistant professor in the math department at UofT whose research focuses on algorithmic methods for computational biology and medical informatics.

Link: https://utoronto.zoom.us/j/84277066292
Meeting ID: 842 7706 6292
Passcode: data_4_lyf

Recap

- last of linear models: factorial treatment structure, CR and RB designs, interaction plots, estimation of variance, comparison of group means
- regression with binomial response y: logistic transform, fitting by ML, interpretation of coefficients, Challenger data, linear predictor, variance-covariance matrix
- estimation of β; estimation of var(β), based on likelihood theory statistics secret sauce

Inference based on the likelihood function

Inference based on the likelihood function

- model: $y_i \sim f(y_i; \theta), i = 1, \dots, n$ independent
- joint density: $f(\underline{y}; \theta) = \prod_{i=1}^n f(y_i; \theta)$
- likelihood function $L(\theta; \underline{y}) = f(\underline{y}; \theta)$

- log-likelihood function $\ell(\theta; \underline{y}) = \log L(\theta; \underline{y}) = \sum_{i=1}^n \log f(y_i; \theta)$
- maximum likelihood estimate $\hat{\theta} = \arg\sup \ell(\theta; \underline{y})$; $\ell'(\hat{\theta}) = 0$
- Fisher information $j(\theta) = -\ell''(\theta)$

- two theorems:

$$(\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, I)$$

asymptotically normal

- likelihood ratio statistic

$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_p$$

p is dimension of θ

... Inference based on the likelihood function

- two theorems:

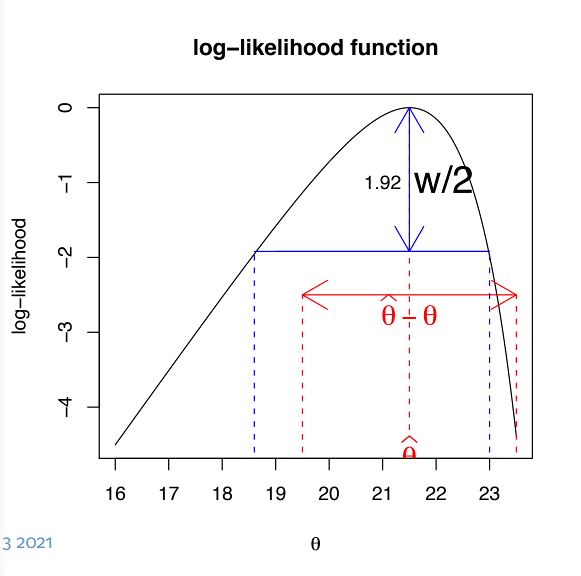
$$
\begin{aligned}
(\hat\theta - \theta) j^{1/2}(\hat\theta) &\xrightarrow{d} N(0, I) \\
w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} &\xrightarrow{d} \chi^2_p
\end{aligned}
$$

- two approximations

$$
\begin{aligned}
\hat\theta_k &\dot\sim N(\{\theta_k, j^{-1}(\hat\theta)_{kk}\} \\
w(\theta) &\dot\sim \chi^2_p
\end{aligned}
$$

- compare two models using change in likelihood ratio statistic — nested models

... Inference based on the likelihood function



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

maximum likelihood estimate $\partial \ell(\beta; y)/\partial \beta = 0$

$$\hat{\beta}_0 = 5.08498, \quad \hat{\beta}_1 = -0.11560 \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}}$$

$$\text{var}(\hat{\beta}) \doteq j^{-1}(\hat{\beta})$$

```
> vcov(logitmodcorrect)
            (Intercept)  temperature
(Intercept)   9.3175983 -0.142564339
temperature  -0.1425643  0.002211221
```

Nested models

- Comparing two models:
- likelihood ratio test
$$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$
compares the maximized log-likelihood function under model A and model B
- example
model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$
model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$

- when model B is nested in model A, LRT is approximately χ^2_ν distributed, under model B
- $\nu = dim(A) - dim(B)$

... nested models

```
> logitmodcorrect <- glm(cbind(r,m-r) ~ temperature, family = binomial, data = shuttle2)
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)
> summary(logitmodcorrect2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.520195   3.486784   0.723   0.4698
temperature -0.098297   0.044890  -2.190   0.0285 *
pressure     0.008484   0.007677   1.105   0.2691
---
    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 16.546  on 20  degrees of freedom
AIC: 36.106
Number of Fisher Scoring iterations: 5
```

... nested models

```
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)

> anova(logitmodcorrect,logitmodcorrect2)
Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ temperature
Model 2: cbind(r, m - r) ~ temperature + pressure
  Resid. Df Resid. Dev Df Deviance
1        21     18.086
2        20     16.546  1   1.5407
```

...nested models

- Model A: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$
- Under Model B, the change in deviance is (approximately) an observation from a χ^2_1
- $\Pr(\chi^2_1 \geq 1.5407) = 0.22$: this is a p-value for testing $H_0 : \beta_2 = 0$

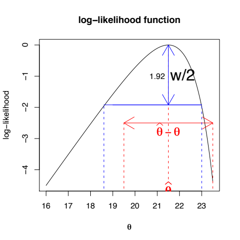
ELM-1 p.30

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

- based on profile log-likelihood
- `confint(logitmodcorrect):`
  `( -0.212`~~`2262`~~`, -0.0244`~~`701`~~` )`



$\ell_p(\beta_1)$, details to follow

ELM-1 p. 31

Binary data

- each response is Binary: $y_i = 0, 1$ instead of $0, 1, \dots, m_i$
- explanatory variables x_i^T as usual
- same model

$$\mathrm{pr}(y_i = 1 \mid x_i) = p_i(\beta) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

- example: SM 10.18
- example HW6: "The math group, the single dependent variable of this work, was coded as a dichotomous variable (1: math group vs. 0: nonmath group)."
- "To classify the math vs. nonmath groups, we also executed a binary logistic regression."

- example `wcgs` data, ELM-2, Ch.2

Binary data


```
> data(wcgs, package="faraway")
> head(wcgs); help(wcgs) #latter not shown
     age height weight sdp dbp chol behave cigs
2001  49     73    150 110  76  225     A2   25
2002  42     70    160 154  84  177     A2   20
2003  42     69    160 110  78  181     B3    0
2004  41     68    152 124  78  132     B4   20
2005  59     70    150 144  86  255     B3   20
2006  44     72    204 150  90  182     B4    0
     dibep chd  typechd timechd   arcus
2001     B  no     none    1664  absent
2002     B  no     none    3071 present
2003     A  no     none    3071  absent
2004     A  no     none    3064  absent
2005     A yes infdeath    1885 present
```

Nov-2

Nancy

2/11/2021

Binary data

```
data(wcgs, package = "faraway")
head(wcgs) #not run: str(wcgs); plot(wcgs); help(wcgs)
```

```
##      age height weight sdp dbp chol behave cigs dibep chd  typechd timechd
## 2001  49     73    150 110  76  225     A2   25     B  no     none    1664
## 2002  42     70    160 154  84  177     A2   20     B  no     none    3071
## 2003  42     69    160 110  78  181     B3    0     A  no     none    3071
## 2004  41     68    152 124  78  132     B4   20     A  no     none    3064
## 2005  59     70    150 144  86  255     B3   20     A yes infdeath    1885
## 2006  44     72    204 150  90  182     B4    0     A  no     none    3102
##        arcus
## 2001  absent
## 2002 present
## 2003  absent
## 2004  absent
## 2005 present
```

… Binary responses

- where's the epsilon? There isn't one
- what's the model? It has two parts
- Regression.

$$\mathbb{E}(y_i) = p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- Probability distribution.

$$y_i \sim Bernoulli(p_i)$$

- What are these parts in linear regression?
- Regression

$$\mathbb{E}(y_i) = \mu_i = x_i^{\mathrm{T}}\beta$$

- Probability distribution

$$y_i \sim Normal(\mu_i, \sigma^2)$$

Binomial responses

- if you add a lot of Bernoulli's together, all with the same p_i, you get
- how could they have the same p_i in our model?
- $p_i = function(x_i^{\mathrm{T}}\beta)$
- different observations with the same p_i are called covariate classes
- Example 10.18 in SM – Table 10.8 has 23 rows of binomials

  sample sizes vary from 1 to 6
- `data(nodal)` in `library(SMPracticals)` has 53 rows of binary observations
- `R` expects `cbind(r, m-r)` in `glm` with binomial dat
- a, but if all observations are binary you can get away with `r` only
- see `?family` (check `Details`)
- you can also specify proportions y_i/n_i, but then you need to use `weights`

Table 10.8 Data on nodal involvement (Brown, 1980).

m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
1	1	1	1	0	1	1
1	1	1	0	1	1	1
1	1	1	0	0	1	1
1	0	1	0	1	0	0
1	1	0	1	1	1	0
1	0	0	1	1	0	0
1	1	0	1	0	1	0
1	1	0	0	1	0	1

Can we predict nodal involvement from other measurements?

Nov-2

Nancy

2/11/2021

Binary data

```
data(wcgs, package = "faraway")
head(wcgs) #not run: str(wcgs); plot(wcgs); help(wcgs)
```

→ .Rmd

```
##      age height weight sdp dbp chol behave cigs dibep chd  typechd timechd
## 2001  49     73    150 110  76  225     A2   25     B  no     none    1664
## 2002  42     70    160 154  84  177     A2   20     B  no     none    3071
## 2003  42     69    160 110  78  181     B3    0     A  no     none    3071
## 2004  41     68    152 124  78  132     B4   20     A  no     none    3064
## 2005  59     70    150 144  86  255     B3   20     A yes infdeath    1885
## 2006  44     72    204 150  90  182     B4    0     A  no     none    3102
##        arcus
## 2001  absent
## 2002 present
## 2003  absent
## 2004  absent
## 2005 present
## 2006  absent
```

Special to the binomial

- likelihood ratio test for logistic model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
- this model is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- residual deviance compares fitted model to saturated model

- under the fitted model, approximately distributed as χ^2_{n-q}
  if each n_i "large"

ELM-1 p.29

```
> summary(Ex1018.glm)

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40.710  on 22  degrees of freedom
 Residual deviance: 18.069  on 17  degrees of freedom
 AIC: 41.69
```

... example 10.18 variable selection

```
> step(ex1018binom)

Coefficients:
(Intercept)        stage         xray         acid
     -3.052        1.645        1.912        1.638

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:^^I    40.71
Residual Deviance: 19.64 ^^IAIC: 39.26
```

– we can drop `age` and `grade` without affecting quality of the fit

– in other words the model can be simplified by setting two regression coefficients to zero

– several mistakes in text on pp. 491,2;

– deviances in Table 10.9 are incorrect as well `http://statwww.epfl.ch/davison/SM/` has corrected version

… example 10.18: variable selection

- `step` implements stepwise regression
- evaluates each fit using AIC = $-2\ell(\hat{\beta}; y) + 2p$
- penalizes models with larger number of parameters

- we can also compare fits by comparing deviances
-

```
> update(ex1018binom, . ~ . - aged - stage)

Call:  glm(formula = cbind(r, m - r) ~ grade + xray + acid, family = binomial,
    data = nodal2)

Coefficients:
(Intercept)        grade         xray         acid
     -2.734        1.420        1.750        1.797

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:^^I    40.71
Residual Deviance: 21.28 ^^IAIC: 40.9

> deviance(ex1018binom)
[1] 18.06869
> pchisq(21.28-18.07,df=2,lower=F)
[1] 0.2008896
```

AIC

- as terms are added to the model, deviance always decreases
- because log-likelihood function always increases
- similar to residual sum of squares

- Akaike Information Criterion penalizes models with more parameters
- $$AIC = 2\{-\ell(\hat{\beta}; y) + p\}$$

SM (4.57)

- comparison of two model fits by difference in AIC

Deviance residuals

`glm.diag; library(SMPracticals)`

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q    Median      3Q      Max
-1.4989  -0.7726  -0.1265  0.7997  1.4351
```

Deviance: $2\sum_{i=1}^{n}[y_i \log\{y_i/n_i p_i(\hat\beta)\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i p_i(\hat\beta))\}]$

approximately χ^2_{n-q}

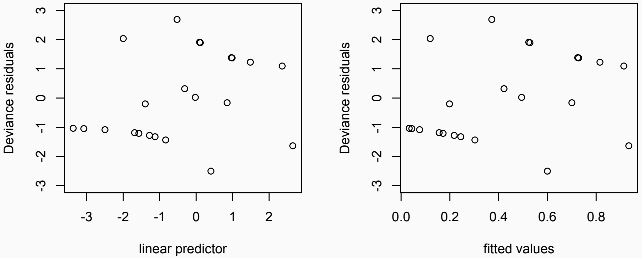
$$r_{Di} = \pm\surd(2[y_i \log\{y_i/n_i\hat p_i\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i\hat p_i)\}])$$

... example 10.18: residuals

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```



Generalized linear models

`glm` has several options for `family`

```
binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

Each of these is a member of the class of generalized linear models

Generalized: distribution of response is not assumed to be normal

Linear: some transformation of $E(y_i)$ is of the form $x_i^T\beta$

link function

Poisson regression

- the Poisson distribution is a useful starting point for data that counts events
- $$f(y_i \mid x_i) = \frac{1}{y!}\mu_i^{y_i} e^{-\mu_i}, y_i = 0, 1, \ldots,$$
- $$f(y_i \mid x_i) = \exp\{y_i \log \mu_i - \mu_i - \log(y_i!)\}$$
- canonical parameter
$$\theta_i = \log(\mu_i)$$
- linear model:
$$\log(\mu_i) = x_i^T \beta$$
- equivalently
$$E(y_i) = \mu_i = \exp(x_i^T \beta)$$

$\longrightarrow$ .Rmd part 3

HW 6

- coding 1 for ”lack math”, 0 otherwise; p.6 + data
- t-test with 84 (and 83) df; Fig 2

Welch’s t-test

- how many predictors in logistic regression? p.2,3
- conclusions p. 4

HW4

HW Question Week 4

STA2101F 2021

Due October 14 2021 11.59 pm

Homework to be submitted through Quercus

Part 1: Data set for project Okay to submit October 21

Please submit details about the data you will use for your project. Ideally the data will have a single response or outcome variable of interest, and several potential explanatory variables. You should submit with this homework:

(1) the data source: both bibliographic and a web link
(2) the number of observations and the number of potential explanatory variables
(3) a description of the response variable
(4) a description of the potential explanatory variables
(5) the scientific question(s) of interest

When you submit the final project, it will consist of the parts listed in Slide 3 on October 6.

Part 2: Question(s) for marking

There has been a lot of talk this week about rapid testing in the schools. On one hand there seems no harm in using rapid antigen tests on a regular basis, but on the other hand if a lot of the tests give incorrect results, especially flagging as covid-related too often, then children will unnecessarily miss school. This seems to be the main concern from the public health officials who are cautioning a slower approach.

Tests for Covid19 (or any screening for that matter), are assessed by their false positive and