2019_sem_1_past_exam

Michael Jones

12/02/2022

R

Question 1)

a)

i)

$$logit(p_i) = \beta_0 + \beta_1 \times duration_i + \beta_2 \times ageB_i + \beta_3 \times ageC_i + \beta_4 \times own.houseYes_i$$

$$Y_i \sim Binomial(n_i, p_i)$$

Where Y_i is the number of customers who defaulted on their loan from the ith group. n_i is the number of customers in the ith group And p_i refers to the probability of a customer defaulting on their loan

$ageB$ is a dummy variable. It takes value 1 for groups between the ages 30 and 50, 0 otherwise.

$ageC$ is a dummy variable. It takes value 1 for groups older than 50, 0 otherwise

$duration$ refers to the duration of the loan in months

$own.houseYes$ is a Dummy variable, that takes the value 1 if the group owns their house, 0 otherwise.

i)

$Y_i \sim Binomial(n_i, p_i)$ Our response variable is assumed to have a binomial distribution.

b)

A one month increase in duration of the loan will lead to a between 0.026 and 0.050 increase in the log odds that a custom will default.

```
100*(exp(c(0.026, 0.050))-1)
```

```
## [1] 2.634095 5.127110
```

c)

The deviance does NOT suggest the model does not fit the data. We have done a hypothesis test with the H0 the model is appropriate. Given a p-value of 0.29, we do NOT reject the Null. The model is appropriate.

d)

The GAM plot of duration is approximately linear (we can fit a straight line between the dotted line). We do not need to add polynomial terms to the models.

e)

Model A is more appropriate. My preference is not very strong, as they models have approximately identical AICc. I prefer Model A because we should avoid over-complicating the model if it doesn't substantially increase predictive ability. Furthermore the extra variable (duration^2) has a high p-vaoue.

f)

i)

The deviance of model E is 0

ii)

Null deviance is 95.764

iii)

Saturated Model (E) Log Likelihood: $\ell_s = -42.182$

Deviance Model B: $D_B = 25.261$

$D_B = 2(\ell_s - \ell_B)$

$25.261 = 2(-42.182 - \ell_B)$

$12.6305 = -42.182 - \ell_B$

$54.8125 = -\ell_B$

$\ell_B = -54.8125$

```
12.6305+42.182
```

```
## [1] 54.8125
```

The Maximized log likelihood of model B is -92.704

g)

H0: The submodel (null Model) is appropriate. All our explanatory variables loan, age house, ownership are unrelated to probabiity a customer defaults their loan.

h)

H1: The submodel (null model) is not appropriate. At least one of our explanatory terms are necessary.

i and j)

```
diffInDeviance <- 95.765-28.366
df <- 29-25
1-pchisq(diffInDeviance, df)
```

```
## [1] 8.026912e-14
```

We reject the null hypothesis, the submodel is not appropriate and at least one of our explanatory terms (duration, age, own.house) are necessary variables.

Question 2)

a:d) Skip didn't study this

Question 3)

a)

When plotting an explanatory variable whose relationship with log(y) decays over time

b)

lm(log(y) ~ log(x), data = data)

c)

```
beta_0 = -0.453
beta_1 = 0.0163
xfit=2
yfit=3

ypred = exp(beta_0  + beta_1*log(xfit))

ypred
```

```
## [1] 0.6429414
```

The fitted value is 0.64

d)

$log(y) = \beta_0 + \beta_1 \times log(x)$

If x is tripled: $log(y) = \beta_0 + \beta_1 \times log(3) \times log(x)$

$y = exp(\beta_0) \times exp(\beta_1) \times 3x$

Therefore y is multiplied by 3 times.

mean of y increases 3 times.

Question 4)

a)

An explanatory variable is factor that attempts to describe our repsonse variables

b)

A predictive model is a model whose primary purpose is to determine future observations.

c)

Measurement errors have little effect on the predictive power of a model, because they are biased.

Question 5)

a)

$log(y) = \beta_0 + \beta_1 \times x$

$\log(70) - \beta_0 = \beta_1 \times x$ $\frac{log(70) - \beta_0}{\beta_1} = x$

```
beta_0 = 4.270399
beta_1 = -0.000478

x = (log(70) - beta_0)/beta_1
x
```

```
## [1] 45.82376
```

b)

```
n = nrow(eurof)
n.sims = 1000
xActual = 45.81

for (i in n.sims) {
```

```
  samp <- sample(1:nrow(eurof), replace = T)
  boot.df <- eurof[samp,]

  fit <- lm(pulse ~age, family = 'poisson', data = boot.df)

  est.b0 = coef(fit)[1]
  est.b1 = coef(fit)[2]
  xsim[i] = (log(70) - est.b0)/est.b1


}

c(2*xActual - quantile(xsim, probs = 0.975),
  2*xActual - quantile(xsim, probs = 0.025))
```

Question 6)

$$logit(p_i) = \beta_0 + \beta_1 \times x2_i + f_1(x3_i) + \beta_2 \times x4_i + f_2(x4_i^2)$$

p_i refers to the probability that y01 takes the value 1. $beta_i$ reflect linear relationships f_i are the smoothing functions of GAM

Question 7)

Test set is data that we do NOT use for model training. We use test set to gain an honest estimate of the mean squared prediction error (MSPE)

Question 8)

a)

```
X2 = 40
X3 = 6.5

beta_1 = -6
beta_2 = 0.03
beta_3 = 0.5

logitY = beta_1 + beta_2*X2 + beta_3*X3

y = plogis(logitY)
y
```

```
## [1] 0.1750863
```

There is a 17.5% chance they get an A

b)

$logit(0.5) = beta_1 + beta_2 \times X2 + \beta_3 \times 6.5$

$0 - beta_1 - beta_3 \times 6.5 = \beta_2 \times X2$

$\frac{-beta_1 - beta_3 \times 6.5}{\beta_2} = X2$

```
X2 = (-beta_1 - beta_3*6.5)/beta_2
X2
```

```
## [1] 91.66667
```

They would have to study for approximately 91.7 hours.

Question 9) REGSUBSETS (Didnt Study this)

Question 10)

a)

We say that M is a modifier of the effect of X on Y when the average causal effect of X on Y varies across levels of M. We handle them by allowing interactions in statistical models.

b)

Design Experiments and observational studies

c)

```
fit <- glm(y ~ A + B, family = 'poisson', data = data)
```

d)

```
fit <- glm(B ~ A, family = 'poisson', data = data)
```

e)

```
fit <- glm(y ~ A, family = 'poisson', data = data)
```