Binary Regression and Non-linear Optimisation with R

Peter Tempfli

3/6/2019

1 a,

```
reg <- glm(lfp ~ age + k5 + k618 + wc, data=df, family=binomial("logit"))
summary(reg)
```

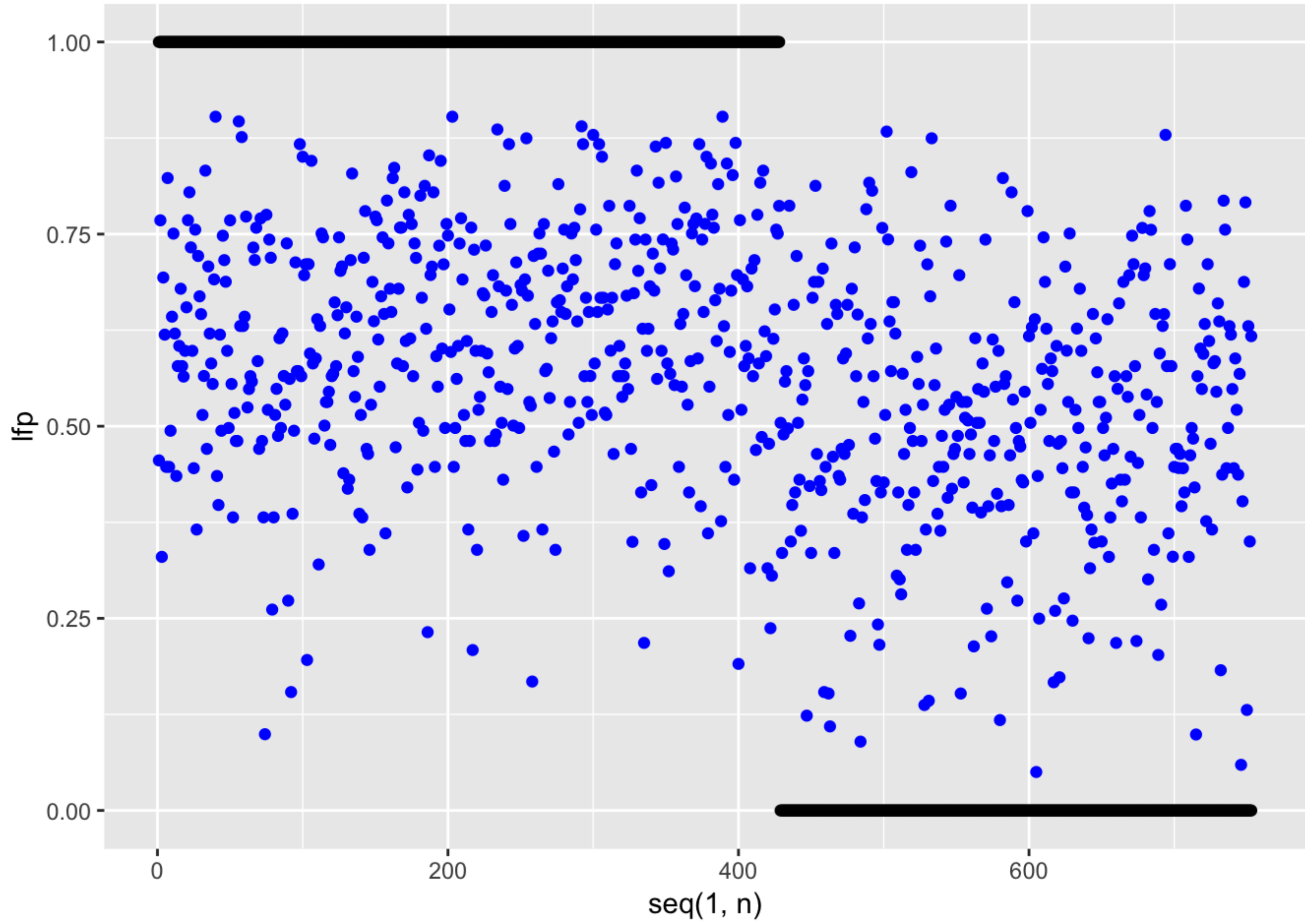
```
##
## Call:
## glm(formula = lfp ~ age + k5 + k618 + wc, family = binomial("logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0732  -1.1308   0.7134   1.0139   2.1498
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.44759    0.60443   5.704 1.17e-08 ***
## age         -0.06778    0.01236  -5.482 4.21e-08 ***
## k5          -1.45700    0.19234  -7.575 3.58e-14 ***
## k618        -0.10885    0.06620  -1.644      0.1
## wc           0.81433    0.18448   4.414 1.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  940.15  on 748  degrees of freedom
## AIC: 950.15
##
## Number of Fisher Scoring iterations: 4
```

```
## first prediction
pr1 <- predict(reg, df, type="response")
summary(pr1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05005 0.46390 0.57470 0.56839 0.69109 0.90279
```

```
n <- nrow(df)

ggplot(data=df, aes(y = lfp, x=seq(1, n))) + geom_point() + geom_point(aes(y=pr1), co
lor="blue")
```




```
#ggplot(data=data.frame(pr1), aes(x=pr1))+geom_histogram(binwidth = 0.02)
```

Comments

- Every parameter except college attendece (`wc`) has negative effect (beta < 0)
- `k618` parameter has a very high P value, so it's effect is not sygnificant. Probably we should not use it in the model.

```
lfpVsPrediction = data.frame(lfp=ifelse(Mroz$lfp == 'yes', 1, 0), prediction=pr1)
lfpVsPrediction$pr = ifelse(lfpVsPrediction$prediction > 0.5, 1, 0)

sensitivity(as.factor(lfpVsPrediction$pr), as.factor(lfpVsPrediction$lfp))
```

```
## [1] 0.4953846
```

```
specificity(as.factor(lfpVsPrediction$pr), as.factor(lfpVsPrediction$lfp))
```

```
## [1] 0.7897196
```

True positives: 338 True negatives : 161 False positives : 164 False negatives : 90

Model sensitivity: 49% Model specificity: 78%

- The model can predict with good rate if a women won't participate (true negative). However, it can't predict confidentally if a women will participate (true positives).

1 b,

```
predict(reg,data.frame(age=30, wc=1, k5=1, k618=0), type="response")
```

```
##         1
## 0.6838588
```

1 c,

If Sue had another child, her probability to work is slightly lower. This is because `k618` beta is `-0.10885`

```
predict(reg,data.frame(age=30, wc=1, k5=1, k618=1), type="response")
```

```
##         1
## 0.6598694
```

1 d,

```
predict(reg,data.frame(age=25, wc=0, k5=1, k618=0), type="response")
```

```
##         1
## 0.5734959
```

1 e

College attendece beta is positive, so it's increasing the likelihood to work.

```
predict(reg,data.frame(age=25, wc=1, k5=1, k618=0), type="response")
```

```
##         1
## 0.7522139
```

1 f

According to this regression, higher family income implies lower likelihood to work (`inc` beta is negative, and P is very low, so significant).

```
reg2 <- glm(lfp ~ age + k5 + k618 + wc + hc + inc + lwg, data=df, family=binomial("lo
git"))
summary(reg2)
```

```
##
## Call:
## glm(formula = lfp ~ age + k5 + k618 + wc + hc + inc + lwg, family = binomial("logi
t"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1062  -1.0900   0.5978   0.9709   2.1893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
## age         -0.062871   0.012783  -4.918 8.73e-07 ***
## k5          -1.462913   0.197001  -7.426 1.12e-13 ***
## k618        -0.064571   0.068001  -0.950 0.342337
## wc           0.807274   0.229980   3.510 0.000448 ***
## hc           0.111734   0.206040   0.542 0.587618
## inc         -0.034446   0.008208  -4.196 2.71e-05 ***
## lwg          0.604693   0.150818   4.009 6.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  905.27  on 745  degrees of freedom
## AIC: 921.27
##
## Number of Fisher Scoring iterations: 4
```

2 a

We can reject the zero hypothesis.

```
hip0 <- glm(lfp ~ 1, family=binomial("logit"), data=df)
hip1 <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family=binomial("logit"), da
ta=df)
anova(hip0, hip1, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: lfp ~ 1
## Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       752    1029.75
## 2       745     905.27  7   124.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 b

As P-value is high for the second model compared to first, we can accept the zero hypothesis (so adding `k618` to the model doesn't give us any significant effect, which we already seen in the first exersice)

```
hip0 <- glm(lfp ~ k5, family=binomial("logit"), data=df)
hip1 <- glm(lfp ~ k5 + k618, family=binomial("logit"), data=df)
anova(hip0, hip1, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: lfp ~ k5
## Model 2: lfp ~ k5 + k618
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       751     994.75
## 2       750     994.53  1  0.22465   0.6355
```

2 c

We can reject the zero hypothesis – P is very low for the alternative hypothesis, so `lfp` DEPENDES on college attendence.

```
hip0 <- glm(lfp ~ 1, family=binomial("logit"), data=df)
hip1 <- glm(lfp ~ wc, family=binomial("logit"), data=df)
anova(hip0, hip1, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: lfp ~ 1
## Model 2: lfp ~ wc
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       752     1029.8
## 2       751     1014.7  1   15.076 0.0001033 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```