Stat 525 Homework 7

Kenny Flagg

October 31, 2016

1. *The analyses presented below are from a study about the relationship between waking with a sore throat after surgery as a function of duration (in minutes) and type of device. We have*

$$Y = \begin{cases} 1 & \text{sore throat} \\ 0 & \text{no sore throat} \end{cases} \qquad \text{type} = \begin{cases} 1 & \text{mask airway} \\ 0 & \text{tracheal tube} \end{cases}$$

The data will be provided to you in the file `sorethroat.csv`. *R code that may prove useful is also attached.*

(a) *Fit an additive model to the data and give me the summary results.*

```
sore.data <- read.csv('sorethroat.csv', header = TRUE)
sore.fit <- glm(y ~ type + duration, family = binomial, data = sore.data)
summary(sore.fit)


Call:
glm(formula = y ~ type + duration, family = binomial, data = sore.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3802  -0.5358   0.3047   0.7308   1.7821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.41734    1.09457  -1.295  0.19536
type        -1.65895    0.92285  -1.798  0.07224
duration     0.06868    0.02641   2.600  0.00931

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.180  on 34  degrees of freedom
Residual deviance: 30.138  on 32  degrees of freedom
AIC: 36.138

Number of Fisher Scoring iterations: 5
```

On the logit scale write down the two equations corresponding to mask and tube, respectively.

For the mask airway patients, the estimated log odds of waking with a sore throat are

$$\begin{aligned}\operatorname{logit}(\widehat{\pi}) &= -1.42 - 1.66 + 0.0687t \\ &= -3.08 + 0.0687t\end{aligned}$$

where t is the duration of surgery in minutes.

For the tracheal tube patients, the estimated log odds of waking with a sore throat are

$$\operatorname{logit}(\widehat{\pi}) = -1.42 + 0.0687t$$

where t is again the duration of surgery in minutes.

(b) *Estimate the ratio of the odds of a sore throat when using a mask airway to a sore throat when using the tracheal tube. Provide me with an approximate 95% confidence interval for the true odds ratio. Interpret the interval in terms of the problem.*

```
exp(coef(sore.fit)['type'])

     type
0.1903389

exp(confint(sore.fit)['type',])

     2.5 %     97.5 %
0.02608803 1.07717361
```

After adjusting for the duration of surgery, we are 95% confident that the true odds of a sore throat when using a mask airway ratio are between 0.0261 and 1.08 times the odds of a sore throat when using a tracheal tube.

(c) *Estimate the ratio of the odds of a sore throat for a surgery of 40 minutes to the odds for a surgery of 30 minutes. Provide me with an approximate 95% confidence interval for the true odds. Interpret the interval in terms of the problem.*

```
exp(coef(sore.fit)['duration'] * 10)

duration
1.987302

exp(confint(sore.fit)['duration',] * 10)

   2.5 %   97.5 %
1.290158 3.749682
```

After adjusting for the type of device, we are 95% confident that the true odds of a sore throat after a 40 minute surgery are between 1.29 and 3.75 times the odds of a sore throat after a 30 minute surgery.

(d) *Fit a model with an interaction between type and duration. Summarize the results and write down (on the logit scale) the two equations corresponding to mask and tube, respectively.*

```
sore.fit2 <- glm(y ~ type * duration, family = binomial, data = sore.data)
summary(sore.fit2)
```

```

Call:
glm(formula = y ~ type * duration, family = binomial, data = sore.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9707  -0.3779   0.3448   0.7292   1.9961

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.04979    1.46940   0.034   0.9730
type          -4.47224    2.46707  -1.813   0.0699
duration       0.02848    0.03429   0.831   0.4062
type:duration  0.07460    0.05777   1.291   0.1966

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.180  on 34  degrees of freedom
Residual deviance: 28.321  on 31  degrees of freedom
AIC: 36.321

Number of Fisher Scoring iterations: 6
```

For a patient with a mask airway with surgery duration t minutes, the estimated log odds of waking with a sore throat are

$$\begin{aligned}\operatorname{logit}(\widehat{\pi}) &= 0.0498 - 4.47 + (0.0285 + 0.0746)t \\ &= -4.42 + 0.103t.\end{aligned}$$

For a patient with a tracheal tube with surgery duration t minutes, the estimated log odds of waking with a sore throat are

$$\operatorname{logit}(\widehat{\pi}) = 0.0498 + 0.0285t.$$

(e) *Using the interaction model estimate the ratio of the odds of a sore throat for a surgery lasting 30 minutes when the mask is used to the odds of a sore throat surgery lasting 30 minutes when a tracheal tube is used. Provide me with an approximate 95% confidence interval. Interpret the interval in terms of the problem.*

```
coefs.pred <- rbind(0, 1, 0, 30)
logit.pi.hat <- coef(sore.fit2) %*% coefs.pred
logit.pi.hat
```

```
          [,1]
[1,] -2.234203
```

```
exp(logit.pi.hat)

          [,1]
[1,] 0.1070774

SE.e <- sqrt(t(coefs.pred) %*% vcov(sore.fit2) %*% coefs.pred)
SE.e

         [,1]
[1,] 1.128424

exp(logit.pi.hat + qnorm(c(0.025, 0.975)) * SE.e)

[1] 0.01172685 0.97771919
```

We are 95% confident that the true odds of a sore throat for a patient with a mask airway and a thirty minute surgery are between 0.0117 and 0.978 times the odds of a sore throat for a patient with a tracheal tube and a thirty minute surgery.

(f) *Do we need the interaction term? Carry out both a Wald test and a likelihood ratio test. Summarize the results and draw a conclusion in terms of the problem.*

Wald:

From the summary output in part (d), the interaction term has a Wald statistic of $z = 1.29$ with a p-value of 0.1966, which is not convincing evidence that the relationship between the odds of a sore throat and the type of device depends linearly on the surgery duration.

LRT:

```
anova(sore.fit, sore.fit2, test = 'Chisq')

Analysis of Deviance Table

Model 1: y ~ type + duration
Model 2: y ~ type * duration
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        32     30.138
2        31     28.321  1   1.8169   0.1777
```

The LRT statistic is $\chi^2_1 = 1.82$ with a p-value of 0.1777. Again, this is little to no evidence of an interaction between the device type and the surgery duration.

2. *Kyphosis is a disfiguring forward flexion of the spine following spinal surgery. Age in months of 18 subjects with kyphosis ($y = 1$) and 22 subjects without kyphosis ($y = 0$) are given below.*

```
y <- c(rep(1, 18), rep(0, 22))
age <- c(12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139,
         139, 157, 1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118,
         127, 131, 140, 151, 159, 177, 206)
```

   (a) *Fit two logistic regression models with age as the predictor. One model only incorporates age as a linear term while the second has age and age squared in it.*

```
fit <- glm(y ~ age, family = binomial)
fit.q <- glm(y ~ age + I(age^2), family = binomial)
```

Compare the two models. Does it appear that the relationship is quadratic? Justify your answer.

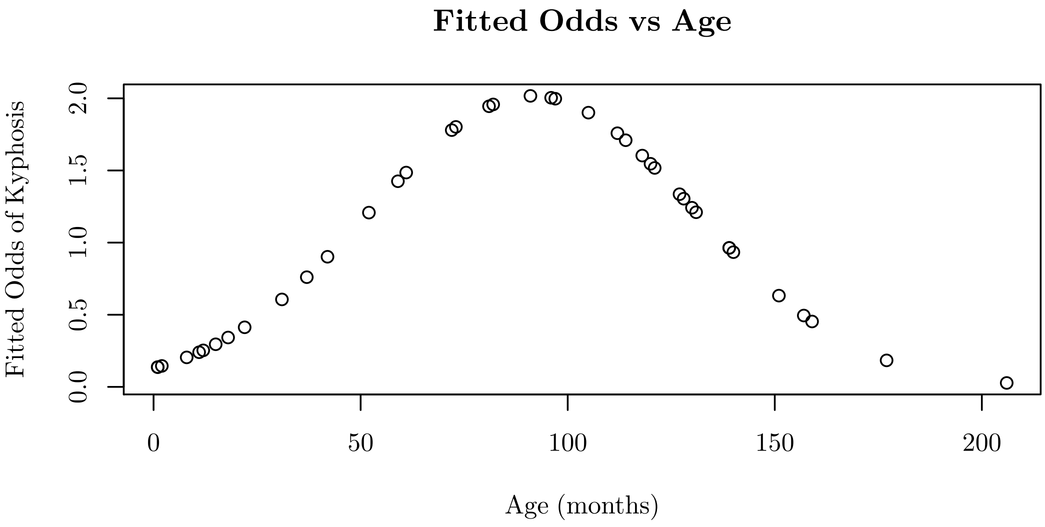
```
summary(fit)


Call:
glm(formula = y ~ age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3126  -1.0907  -0.9482   1.2170   1.4052

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.572693   0.602395  -0.951    0.342
age          0.004296   0.005849   0.734    0.463

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.051  on 39  degrees of freedom
Residual deviance: 54.504  on 38  degrees of freedom
AIC: 58.504

Number of Fisher Scoring iterations: 4

summary(fit.q)


Call:
glm(formula = y ~ age + I(age^2), family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.482  -1.009  -0.507   1.012   1.788

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.0462547  0.9943478  -2.058   0.0396
age          0.0600398  0.0267808   2.242   0.0250
```

```
I(age^2)    -0.0003279  0.0001564  -2.097   0.0360

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.051  on 39  degrees of freedom
Residual deviance: 48.228  on 37  degrees of freedom
AIC: 54.228

Number of Fisher Scoring iterations: 4

anova(fit, fit.q, test = 'Chisq')

Analysis of Deviance Table

Model 1: y ~ age
Model 2: y ~ age + I(age^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        38     54.504
2        37     48.228  1   6.2762  0.01224
```

The likelihood ratio test has a statistic of $\chi^2_1 = 6.28$ with a p-value of 0.01224, providing strong evidence that there is a quadratic assiciation between the log odds of kyphosis and age.

(b) *Interpretation of quadratic effects can be problematic. A graphical assessment is often helpful. Plot the estimated odds of kyphosis versus age. Provide me with a copy of the plot.*

```
# Note: fitted(fit.q) gives probabilities, not odds or log odds.
# predict(fit.q, type = 'link') gives the log odds.
plot(age, exp(predict(fit.q, type = 'link')), main = 'Fitted Odds vs Age',
     xlab = 'Age (months)', ylab = 'Fitted Odds of Kyphosis')
```



(c) *Summarize the relationship.*

The estimated relationship between kyphosis and age has clear curvature, with the odds of kyphosis being near 0.1 for newborn infants, increasing to about 2 for children aged 100 months, and then decreasing with age, leveling off near zero around 200 months.

(d) *STAT 525 Only: At what age is the estimated odds of kyphosis the greatest (I want a general answer given as a function of the estimated regression coefficients)? Give an approximation for the variance of this estimated age. Now use the general formulas you just derived to give an estimate of the age and the variance of that estimate for the kyphosis problem.*

First, note that the logarithm is a one-to-one function, so the odds and log odds are maximized by the same value.

$$\begin{aligned}\frac{d}{dt}\mathrm{logit}(\widehat{\pi}(t)) &= \frac{d}{dt}\left(\widehat{\beta}_0+\widehat{\beta}_1 t+\widehat{\beta}_2 t^2\right)\\ &= \widehat{\beta}_1+2\widehat{\beta}_2 t.\end{aligned}$$

Setting this equal to zero, we find that the estimated log odds (and therefore the estimated odds as well) are maximized by an age of $t_{\max} = -\dfrac{\widehat{\beta}_1}{2\widehat{\beta}_2}$.

Example 5.5.27 in Casella and Berger gives a cute delta-method result for the variance of a ratio, so

$$\mathrm{Var}\left(\frac{\widehat{\beta}_1}{\widehat{\beta}_2}\right) \approx \left(\frac{\beta_1}{\beta_2}\right)^2\left(\frac{\mathrm{Var}\left(\widehat{\beta}_1\right)}{\beta_1^2}+\frac{\mathrm{Var}\left(\widehat{\beta}_2\right)}{\beta_2^2}-2\frac{\mathrm{Cov}\left(\widehat{\beta}_1,\widehat{\beta}_2\right)}{\beta_1\beta_2}\right)$$

using the fact that $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are MLEs and therefore asymptotically unbiased. Then a very crude estimate is

$$\widehat{\mathrm{Var}}\left(t_{\max}\right) \approx \frac{1}{4}\widehat{\mathrm{Var}}\left(\frac{\widehat{\beta}_1}{\widehat{\beta}_2}\right) \approx \left(\frac{\widehat{\beta}_1}{2\widehat{\beta}_2}\right)^2\left(\frac{\mathrm{Var}\left(\widehat{\beta}_1\right)}{\widehat{\beta}_1^2}+\frac{\mathrm{Var}\left(\widehat{\beta}_2\right)}{\widehat{\beta}_2^2}-2\frac{\mathrm{Cov}\left(\widehat{\beta}_1,\widehat{\beta}_2\right)}{\widehat{\beta}_1\widehat{\beta}_2}\right).$$

```
beta1.hat <- coef(fit.q)['age']
beta2.hat <- coef(fit.q)['I(age^2)']
t.max <- -beta1.hat / (2 * beta2.hat)
t.max


     age
91.53975


var.hat.beta1 <- vcov(fit.q)['age', 'age']
var.hat.beta2 <- vcov(fit.q)['I(age^2)', 'I(age^2)']
cov.hat <- vcov(fit.q)['age', 'I(age^2)']
var.hat.t <- (beta1.hat / (2 * beta2.hat))^2 *
  (var.hat.beta1 / (beta1.hat^2) +
     var.hat.beta2 / (beta2.hat^2) -
     2 * cov.hat / (beta1.hat * beta2.hat))
var.hat.t
```

```
     age
134.9694
```

We estimate that the odds of kyphosis are greatest at the age of 91.5 months. The variance of this estimate is approximately 134.97.

3. *Duchenne Muscular Dystrophy (DMD) is a genetic disease transmitted from mothers to their children. Male offspring generally do not live very long but females may be silent carriers - they do not get sick but are capable of transmitting DMD to their offspring. Blood levels of 2 enzymes (creatine kinase (CK) and hemopexin (H)) were evaluated as possible screening tools. The data are available in an R package* `Sleuth3`*.*

   (a) *Fit a model with the explanatory variables* `log(ck)` *and* `h`*. Summarize the results.*

```
library(Sleuth3)
MD.dat <- ex2012
names(MD.dat) <- c('group', 'ck', 'h')
MD.dat$group <- ifelse(MD.dat$group == 'Control', 0, 1)
fit.1 <- glm(group ~ log(ck) + h, data = MD.dat, family = binomial)
summary(fit.1)
```

```

Call:
glm(formula = group ~ log(ck) + h, family = binomial, data = MD.dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.89706  -0.38782  -0.16696   0.09903   2.60371

Coefficients:
             Estimate Std. Error z value   Pr(>|z|)
(Intercept) -28.91340    5.80017  -4.985 0.00000062
log(ck)       4.02043    0.82910   4.849 0.00000124
h             0.13652    0.03654   3.736   0.000187

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.840  on 119  degrees of freedom
Residual deviance:  61.992  on 117  degrees of freedom
AIC: 67.992

Number of Fisher Scoring iterations: 7
```

(b) *Create a classification table using a cutpoint of $c = 38/120$. Compare the accuracy, sensitivity, specificity, $PV+$, and $PV-$ values (summarize your results in a table). Interpret these measures in terms of the problem. Be careful with $PV+$ and $PV-$ as these are case-control data and those values are not applicable for a general population.*

```
pi.hat <- predict(fit.1, type = 'response')
y.hat <- ifelse(pi.hat >= 38/120, 1, 0)
table(Group = MD.dat$group, Predicted = y.hat)
```

```
     Predicted
Group  0  1
    0 73  9
    1  4 34
```

Quantity	Value
Accuracy	$(73 + 34)/120 = 0.892$
Sensitivity	$34/(34 + 4) = 0.895$
Specificity	$73/(73 + 9) = 0.890$
$PV+$	$34/(34 + 9) = 0.791$
$PV-$	$73/(73 + 4) = 0.948$

Overall, this rule correctly classifies 89.2% of the individuals in the sample as having DMD or not. Of the individuals with DMD, 89.5% are correctly identifed, and 89.0% of the individuals without DMD are correctly classified. Of the individuals predicted to have DMD, 79.1% actually have DMD, and of the individuals predicted to not have DMD, 94.8% did not actually have DMD.

(c) *Install the Epi package if you have not already done so and use the ROC function to create a ROC curve using fitted results. Show the plot in your write-up.*

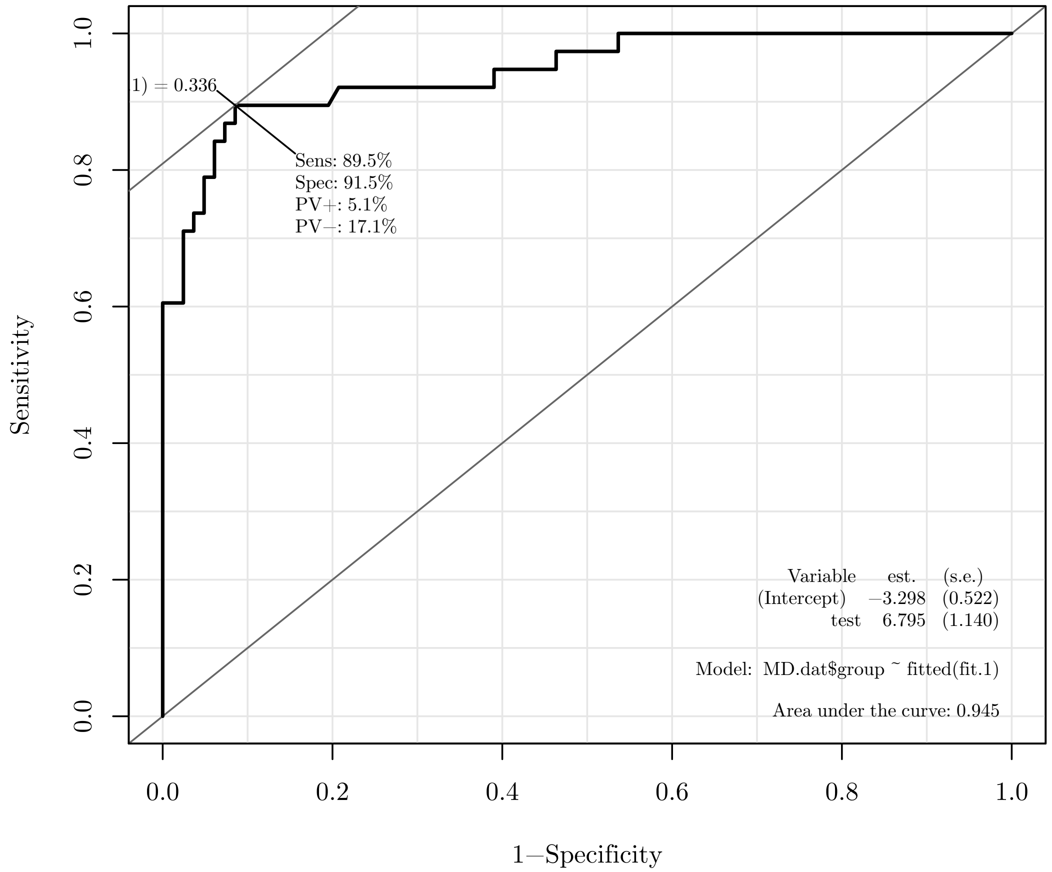
Compare the results commenting on

i. *the optimum cutpoint*

ii. *Sensitivity, Specificity, $PV+$, and $PV-$. Recall that the $PV+$ and $PV-$ values in the plot are incorrect but you can use them to get the correct values.*

iii. *AUC and the implication it has in terms of using these two enzymes as a screening tool.*

```
library(Epi)
# additive model
ROC(fitted(fit.1), MD.dat$group, plot = 'ROC')
```



Quantity	Part (b)	Optimal
Sensitivity	$34/(34+4) = 0.895$	0.895
Specificity	$73/(73+9) = 0.890$	0.915
$PV+$	$34/(34+9) = 0.791$	0.949
$PV-$	$73/(73+4) = 0.948$	0.829

The optimum cutpoint is 0.336, very close to the cutpoint of $38/120 = 0.317$ used in part (b). The sensitivity and specificity are roughly the same, but the optimal cutpoint rule has a much larger $PV+$ and a much smaller $PV-$, suggesting that it finds the "easiest" cases of DMD but missed quite a few of them. But, with an AUC of 0.945, this classifcation rule would still be considered to have "outstanding predictive ability."