Inference for Logistic Regression

EPI 204

Quantitative Epidemiology III

Statistical Models

Evans County, GA Dataset (1963)

- Data are in evans.dat (text, no header), evans.sas7bdat (SAS version 9 dataset), evans.sav (SPSS dataset), and evans.dta (Stata dataset) on the textbook website given in the syllabus.
- The data are from a cohort study in which 609 white males were followed for 7 years, with coronary heart disease as the outcome of interest.
- The variables are given on the next slide

Input for Evans.dat (tab delimited)

```
vars <- c("ID","CHD","CAT","AGE","CHL","SMK",
          "ECG","DBP","SBP","HPT","CH","CC")
evans <- read.table("evans.dat",header=F,
                     col.names=vars)
```

read.table, read.csv, etc. are all variants that can handle text file input with different defaults.

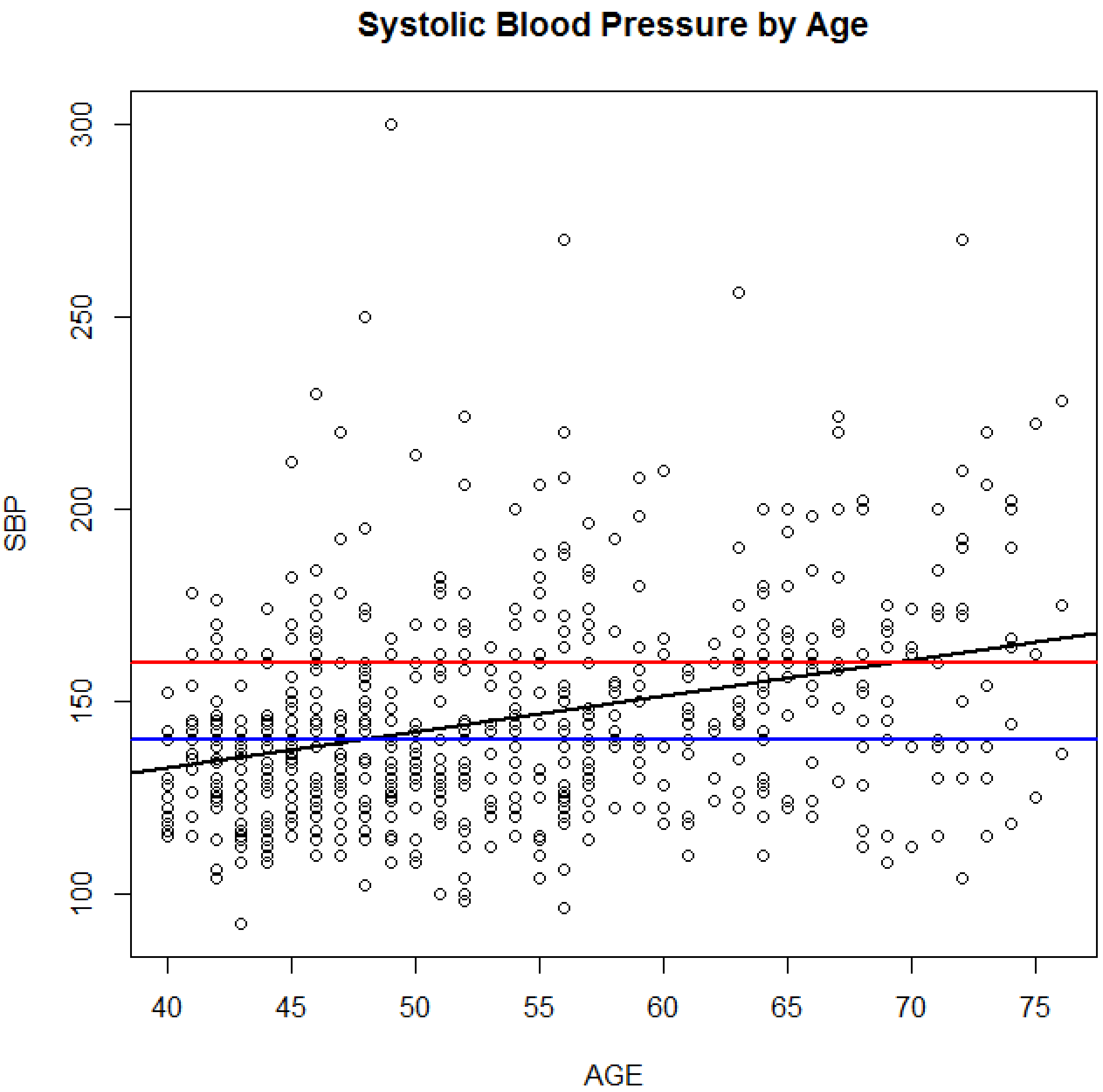
By default, reads strings as factors, unless stringsAsFactors=F. Often this option is a good idea (soon to be the default?).

Variable	Description
ID	Subject ID, one observation per subject
CHD	Coronary heart disease (1) or not (0)
CAT	High catecholamine level (1) or not (0)
AGE	Age in years
CHL	Cholesterol level
SMK	Ever smoked (1) or never smoked (0)
ECG	ECG abnormality (1) or not (0)
DBP	Diastolic blood pressure
SBP	Systolic blood pressure
HPT	= 1 if DBP ≥90 or SBP ≥ 160, otherwise = 0
CH	CAT*HPT
CC	CAT*CHL

```
> vars <- c("ID","CHD","CAT","AGE","CHL","SMK","ECG","DBP","SBP","HPT","CH","CC")
> evans <- read.table("evans.dat",header=F,col.names=vars)
> summary(evans)
       ID              CHD              CAT              AGE              CHL       
 Min.   :   21   Min.   :0.0000   Min.   :0.0000   Min.   :40.00   Min.   : 94.0  
 1st Qu.: 4242   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:46.00   1st Qu.:184.0  
 Median : 9751   Median :0.0000   Median :0.0000   Median :52.00   Median :209.0  
 Mean   : 9213   Mean   :0.1166   Mean   :0.2003   Mean   :53.71   Mean   :211.7  
 3rd Qu.:13941   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:60.00   3rd Qu.:234.0  
 Max.   :19161   Max.   :1.0000   Max.   :1.0000   Max.   :76.00   Max.   :357.0  
      SMK               ECG               DBP               SBP       
 Min.   :0.0000   Min.   :0.0000   Min.   : 60.00   Min.   : 92.0  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 80.00   1st Qu.:125.0  
 Median :1.0000   Median :0.0000   Median : 90.00   Median :140.0  
 Mean   :0.6355   Mean   :0.2726   Mean   : 91.18   Mean   :145.5  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:100.00   3rd Qu.:160.0  
 Max.   :1.0000   Max.   :1.0000   Max.   :170.00   Max.   :300.0  
 HPT                CH                 CC
 Min.   :0.0000   Min.   :0.0000   Min.   :  0.00
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  0.00
 Median :0.0000   Median :0.0000   Median :  0.00
 Mean   :0.4187   Mean   :0.1609   Mean   : 39.96
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:  0.00
 Max.   :1.0000   Max.   :1.0000   Max.   :331.00
```

(one of many possible exploratory plots)

```
> plot(SBP ~ AGE, data=evans)
> abline(coef(lm(SBP~AGE,data=evans)),lwd=2)
> abline(h=140,col="blue",lwd=2)
> abline(h=160,col="red",lwd=2)
> title("Systolic Blood Pressure by Age")
```


Systolic Blood Pressure by Age
SBP
100
150
200
250
300
AGE
40
45
50
55
60
65
70
75

```
> summary(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.680112   1.136363  -5.879 4.14e-09 ***
CAT          0.715810   0.340180   2.104  0.03536 *  
AGE          0.032770   0.015197   2.156  0.03105 *  
CHL          0.008608   0.003259   2.641  0.00827 ** 
SMK          0.802906   0.303001   2.650  0.00805 ** 
HPT          0.476272   0.289296   1.646  0.09970 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 401.95  on 603  degrees of freedom
AIC: 413.95

Number of Fisher Scoring iterations: 5
```

```
> drop1(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans),test="Chisq")
Single term deletions

Model:
CHD ~ CAT + AGE + CHL + SMK + HPT
       Df Deviance    AIC    LRT Pr(>Chi)
<none>      401.95 413.95
CAT     1   406.33 416.33 4.3805 0.036353 *
AGE     1   406.52 416.52 4.5682 0.032571 *
CHL     1   408.86 418.86 6.9088 0.008577 **
SMK     1   409.65 419.65 7.6990 0.005525 **
HPT     1   404.66 414.66 2.7097 0.099741 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HPT is not statistically significant, but omitting it causes a rise in the AIC, so some might keep it in the model.

Likelihood Ratio Test

- This is used to compare two statistical models that are *nested*, meaning that one (the *full model*) has all the terms of the other (the *reduced model*) plus one or more additional ones.
- For example, the full model might have CHD~CAT+AGE+CHL+SMK+HPT
- And the reduced model might have CHD~CAT+AGE+CHL+SMK (removing HPT) or CHD~CAT+CHL+SMK (removing AGE and HPT)

Likelihood Ratio Test

- If the full model has likelihood L_F and the reduced model has likelihood L_R, then statistical theory says that $-2\ln(L_R/L_F) = -2[\ln(L_R) - \ln(L_F)]$ has approximately a chi-squared distribution with df = the number of omitted variables (with categorical variables counting as one less than the number of categories).
- Since $D = -2[\ln(L) - \ln(L_0)]$, where L_0 is the likelihood of the maximal model, we can equally use
  $D_R - D_F = -2[\ln(L_R) - \ln(L_0)] + 2[\ln(L_F) - \ln(L_0)]$
  $= -2[\ln(L_R) - \ln(L_F)]$

```
> deviance(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans))
[1] 401.947
> deviance(glm(CHD~CAT+AGE+CHL+SMK,family=binomial,data=evans))
[1] 404.6566
> deviance(glm(CHD~CAT+CHL+SMK,family=binomial,data=evans))
[1] 409.3424
```

Test for omitting HPT from full model uses 404.6566 – 401.947 = 2.7096. Compare to chi-squared on 1df

```
> 1-pchisq(2.7096,1)
[1] 0.099746
```

(same as produced with drop1)

```
> deviance(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans))
[1] 401.947
> deviance(glm(CHD~CAT+AGE+CHL+SMK,family=binomial,data=evans))
[1] 404.6566
> deviance(glm(CHD~CAT+CHL+SMK,family=binomial,data=evans))
[1] 409.3424
```

Test for omitting both HPT and AGE from full model uses 409.3424 – 401.947 = 7.3954

Compared to a chi-squared on 2df

```
> 1-pchisq(7.3954,2)
[1] 0.02478046
```

```
> anova(glm(CHD~CAT+CHL+SMK,family=binomial,data=evans),
glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans),test="Chisq")
Analysis of Deviance Table

Model 1: CHD ~ CAT + CHL + SMK
Model 2: CHD ~ CAT + AGE + CHL + SMK + HPT
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       605     409.34
2       603     401.95  2   7.3955  0.02478 *
```

Interaction Terms

- We almost always observe a principle of hierarchy of models.
- If an interaction term such as CAT*HPT is in the model, then the main effects CAT and HPT are also in the model. If a three-way interaction such as CAT*CHL*HPT is in the model, then so are all three two way interactions CAT*CHL, CAT*HPT, and CHL*HPT as well as the three main effects.
- R will observe this in drop1() as long as the interactions are explicitly stated.

```
> summary(glm(CHD~CAT+CHL+SMK+HPT+CAT*CHL+CAT*HPT,binomial,data=evans))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.132296   0.913911  -2.333 0.019640 *
CAT         -12.719878   3.138573  -4.053 5.06e-05 ***
CHL          -0.005312   0.004166  -1.275 0.202248         #Don’t omit
SMK           0.698997   0.324996   2.151 0.031493 *
HPT           1.105883   0.328508   3.366 0.000762 ***
CAT:CHL       0.071175   0.014494   4.911 9.07e-07 ***
CAT:HPT      -2.221010   0.730937  -3.039 0.002377 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 352.92  on 602  degrees of freedom
AIC: 366.92

Number of Fisher Scoring iterations: 6
```

```
> drop1(glm(CHD~CAT+CHL+SMK+HPT+CAT*CHL+CAT*HPT,binomial,data=evans))
Single term deletions

Model:
CHD ~ CAT + CHL + SMK + HPT + CAT * CHL + CAT * HPT
        Df Deviance    AIC
<none>       352.92 366.92
SMK      1   357.93 369.93
CAT:CHL  1   399.88 411.88
CAT:HPT  1   362.06 374.06
```

Can't drop CAT, CHL, or HPT.

(because of hierarchy restrictions)

Interaction Terms

- Other than the hierarchical model for interactions, we can compare any two nested models.
- If we want to omit CHL, we also have to omit CAT*CHL, and we have a 2df comparison.
- CHL is quantitative, CAT*CHL is 0 whenever CAT = 0, and is equal to CHL when CAT = 1.
- Inclusion of the interaction means that the effect of high catecholamines is higher when cholesterol is high.
- It also means that the effect of cholesterol is greater when catecholamines are high.

Wald Tests

- Statistical theory provides an estimated variance-covariance matrix for the coefficients.

```
> coef(summary(hyp.glm))
               Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.37766146  0.3801845 -6.2539671 4.001553e-10
smokingYes  -0.06777489  0.2781242 -0.2436857 8.074742e-01
obesityYes   0.69530960  0.2850851  2.4389544 1.472983e-02
snoringYes   0.87193932  0.3975736  2.1931517 2.829645e-02

> round(vcov(hyp.glm),4)
            (Intercept) smokingYes obesityYes snoringYes
(Intercept)      0.1445    -0.0161    -0.0147    -0.1355
smokingYes      -0.0161     0.0774     0.0000    -0.0074
obesityYes      -0.0147     0.0000     0.0813    -0.0081
snoringYes      -0.1355    -0.0074    -0.0081     0.1581
```

Wald Tests

- Each coefficient has an estimated variance and therefore standard error.
- In general, the coefficients are correlated.
- Confidence intervals and tests for single coefficients are given in the summary(glm()) output or are easily derived.
- We need to do more work to find confidence intervals and tests for differences of coefficients as when a factor has more than two levels.
- The same trick allows us to get confidence intervals and tests for interaction terms.

```
> coef(summary(hyp.glm))
               Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.37766146  0.3801845 -6.2539671 4.001553e-10
smokingYes  -0.06777489  0.2781242 -0.2436857 8.074742e-01
obesityYes   0.69530960  0.2850851  2.4389544 1.472983e-02
snoringYes   0.87193932  0.3975736  2.1931517 2.829645e-02

CI for obesity log odds ratio 0.6953 ± (1.960)(0.2851)
0.6963 ± 0.5588
(0.1365, 1.2541)
CI for odds ratio use exp()
(1.15, 3.50)
> confint.default(glm(hyp.tbl~smoking+obesity+snoring,
                  binomial,hyp))
                  2.5 %     97.5 %
(Intercept) -3.12280942 -1.6325135
smokingYes  -0.61288823  0.4773385
obesityYes   0.13655304  1.2540662
snoringYes   0.09270929  1.6511693
```

```
> juul1.glm <-
glm(menarche~age+tanner,binomial,data=juul1)
> summary(juul1.glm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.7758     2.7630  -4.986 6.17e-07 ***
age           0.8603     0.2311   3.723 0.000197 ***
tanner2      -0.5211     1.4846  -0.351 0.725609
tanner3       0.8264     1.2377   0.668 0.504313
tanner4       2.5645     1.2172   2.107 0.035132 *
tanner5       5.1897     1.4140   3.670 0.000242 ***
```

The estimated log odds between tanner 5 and tanner 4 is 5.1897 – 2.5645 = 2.6252

How do we make a confidence interval or hypothesis test?

If X and Y are random variables then

$V(X-Y)=V(X)+V(Y)-2\text{Cov}(X,Y)$

more generally, if a and b are any numbers

$V(aX+bY)=a^2V(X)+b^2V(Y)+2ab\text{Cov}(X,Y)$

If B is a vector of coefficients $B=(\beta_0,\beta_1,\ldots,\beta_p)$

and V is the $(p+1)$ by $(p+1)$ covariance matrix of B,

and b is a vector of length $(p+1)$ of numbers, then

$$b^\top B=\sum_{i=0}^{p} b_i\beta_i \text{ has variance}$$

$$\mathrm{b}^\top Vb=\sum_{i=0}^{p}\sum_{j=0}^{p} b_i b_j v_{ij}$$

```
> c1 <- coef(juul1.glm)
> v1 <- vcov(juul1.glm)
> c1
(Intercept)         age     tanner2     tanner3     tanner4     tanner5 
-13.7758129   0.8603095  -0.5210667   0.8264390   2.5645049   5.1896586 
> v1
            (Intercept)         age     tanner2    tanner3    tanner4    tanner5
(Intercept)  7.63420854 -0.59398971 -0.08627975  0.2184034  0.4414426  0.6973030
age         -0.59398971  0.05338584 -0.08439333 -0.1117773 -0.1318233 -0.1548192
tanner2     -0.08627975 -0.08439333  2.20415172  1.2019693  1.2336584  1.2700108
tanner3      0.21840340 -0.11177725  1.20196929  1.5319140  1.3012763  1.3494243
tanner4      0.44144258 -0.13182328  1.23365842  1.3012763  1.4816528  1.4075578
tanner5      0.69730303 -0.15481918  1.27001076  1.3494243  1.4075578  1.9993267
> b1 <- c(0,0,0,0,-1,1) # this is the contrast = comparison of tanner 4 and 5
> t(b1) %*% c1
         [,1]
[1,] 2.625154
> t(b1)%*% v1 %*% b1
          [,1]
[1,] 0.6658638
```

2.625154 ± (1.960)√0.6658638

2.625 ± 1.599 or (1.025, 4.224) log odds ratio or (2.787, 68.33) odds ratio

Alternatives

- We can re-run the analysis with a different default level for the categorical variable so that the desired comparison is a single coefficient.

```
relevel {stats}   Reorder Levels of Factor

The levels of a factor are re-ordered so that the
level specified by ref is first and the others are
moved down.

relevel(x, ref, ...)

x     an unordered factor.
ref   the reference level, typically a string.
```

Contrasts

- A contrast is a weighted combination of factor levels in which the weights add up to 1.
- We can change the coding of a factor with five levels from an intercept and four 0/1 variables to a set of contrasts.
- This can be complex and hard to implement.
- There are other R packages that can handle these separately such as `multcomp`.

Interaction Terms

- An interaction means that the effect of one variable depends on the level of another.
- To get a numerical measure of effect such as odds ratio, we need to specify the level of the other one.
- We can set the other variable at the modal level, the median level, or the mean level, or at a variety of levels
- We will illustrate this with the evans data.

```
> summary(glm(CHD~CAT+CHL+HPT+CAT*HPT,binomial,evans))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.575185   0.758746  -6.030 1.64e-09 ***
CAT          2.190774   0.498842   4.392 1.12e-05 ***
CHL          0.008495   0.003256   2.609  0.00909 **
HPT          0.912656   0.324423   2.813  0.00491 **
CAT:HPT     -1.681469   0.600829  -2.799  0.00513 **
```

Effect of CAT when HPT = 0 is 2.190774 with inferences given on the line.

Effect of CAT when HPT = 1 is 2.191 - 1.681 = 0.5093

We can analyze the latter using b1 = c(0,1,0,0,1)

```
> b1 <- c(0,1,0,0,1)
> t(b1) %*% c1
          [,1]
[1,] 0.5093054
> v <- vcov(evans1.glm)
> t(b1) %*% v %*% b1
          [,1]
[1,] 0.1245581
> sqrt(t(b1) %*% v %*% b1)
          [,1]
[1,] 0.3529279
```

$\eta = -4.575 \ + \ 2.191 \, \text{CAT} \ + \ 0.008495 \, \text{CHL} \ + \ 0.9127 \, \text{HPT} \ - \ 1.6815 \, \text{CAT} \times \text{HPT}$

When HPT = 0, this is

$\eta = -4.575 \ + \ 2.191 \, \text{CAT} \ + \ 0.008495 \, \text{CHL}$

and the effect of CAT is 2.191

Inference can be read off the coefficient table

When HPT = 1, this is

$\eta = -4.575 \ + \ 2.191 \, \text{CAT} \ + \ 0.008495 \, \text{CHL} \ + \ 0.9127 \ - \ 1.6815 \, \text{CAT}$

and the effect of CAT is $2.191 \ - \ 1.6815$

Inference is more complex

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.575185   0.758746  -6.030 1.64e-09 ***
CAT          2.190774   0.498842   4.392 1.12e-05 ***
CHL          0.008495   0.003256   2.609  0.00909 **
HPT          0.912656   0.324423   2.813  0.00491 **
CAT:HPT     -1.681469   0.600829  -2.799  0.00513 **
```

Effect of CAT when HPT = 0 is 2.190774 with inferences given on the line.

Effect of CAT when HPT = 1 is 2.191 - 1.681 = 0.5093

```
> b1 <- c(0,1,0,0,1)
> t(b1) %*% c1
[1,] 0.5093054
> v <- vcov(evans1.glm)
> sqrt(t(b1) %*% v %*% b1)
[1,] 0.3529279
```

Log odds ratio = 0.5093054, odds ratio = 1.664

Log odds ratio CI is 0.5093 ± (1.960)(0.3529) 0.5093 ± 0.6917

(-.1824, 1.2010) and odds ratio CI is (0.83, 3.32)

Catecholamine level has a significant effect on non-hypertensives, but a smaller and non-significant effect on hypertensives. 41.9% of subjects are hypertensive, so we could use b1 <- c(0,1,0,0,0.419) for average effect.

Homework: Due 4/15/21

- For the Byssinosis data, there are three levels of workspace. Test the hypothesis that each differs from the other (two are in the summary table and the third requires the covariance matrix). Find 95% confidence intervals for the three odds ratios.
- Do the same for the three levels of employment time.
- Fit a model omitting race and sex but including the smoking by workspace interaction. Find the estimated effect of smoking for each of the three workspaces separately, test the hypothesis of no effect, and find 95% confidence intervals for the odds ratios.