Week 10 - Exercises-Solutions

Exercise solutions

Week 10

Investigation - interaction

Stata code and output

1) Compulsory reading: (Section 5.2.4. p 163-165)

This reading explains how to introduce a possible interaction the between *age* seen this time as a continuous variable and *arcus* (coded 0/1).

2) Reproduce the output

Perhaps it helps to write down the model first i.e.

$$log(p/(1-p)) = \beta_0 + \beta_1 arcus + \beta_2 age + \beta_3 age * arcus \tag{0.1}$$

where is the probability of CHD over the course of the study given the covariates. This model can be rewritten separately for patients without arcus

$$log(p/(1-p)) = \beta_0 + \beta_2 age \tag{0.2}$$

and patients with arcus

$$log(p/(1-p)) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) age \tag{0.3}$$

We clearly see that the slope of the association with age (i.e. the log-OR) is not the same in the two arcus groups (β_2 vs $\beta_2 + \beta_3$)

```
use wcgs.dta
logistic chd69 i.arcus##c.age, coef
## . use wcgs.. logistic chd69 i.arcus##c.age, coef
##
## Logistic regression                                     Number of obs =  3,152
##                                                         LR chi2(3)    =  53.33
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -858.93362                             Pseudo R2     = 0.0301
```

```
##
## ------------------------------------------------------------------------------
##        chd69 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##      1.arcus |   2.754185   1.140118     2.42   0.016     .5195952    4.988774
##          age |    .089647   .0148904     6.02   0.000     .0604623    .1188317
##              |
##  arcus#c.age |
##           1  |  -.0498298   .0233431    -2.13   0.033    -.0955814   -.0040782
##              |
##        _cons |  -6.788086   .7179977    -9.45   0.000    -8.195335   -5.380836
## ------------------------------------------------------------------------------
```

Note that you need to specify the type of variable you are using here. The default in Stata is *categorical* covariates. A code like *logistic chd69 arcus##age, coef* would return an ugly output will all different age values considered as categories (except the reference). The *c.age* option is absolutely necessary. You may forget the *i.* before *arcus* because it's coded 0/1 but in general it's safer to write the command as indicated in the textbook.

The analysis with age as a continuous variable confirms what we found with the dichotomised version of age at baseline; we have a significant interaction between *age* and *arcus*.

3) Association between *chd69* and *age* in patients without arcus? OR and 95% CI

The fitted model is $log(p/(1-p)) = -6.788 + 0.09age$ (up to rounding) with the association being described by $\widehat{\beta}_2 = 0.09$. To get the OR you can refit the model without the option coeff and get the OR for age (only) i.e. OR=1.09, 95% CI=(1.06 ; 1. 13). This means that for patients *without arcus* the odds of CHD is 9% bigger, 95% CI=(6% ; 13%) per additional year of age. If you wanted to describe the association for a 10-year age increment, you can 1) use the trick Vittinghof et al (2012) described, rescale age by dividing by 10 and repeat the procedure; 2) use lincom and type lincom 10*age, or. This gives you OR=2.45, OR=(1.83 ; 3.28) as indicated below

```
use wcgs.dta
logistic chd69 i.arcus##c.age
lincom age*10, or
## . use wcgs.. logistic chd69 i.arcus##c.age
##
## Logistic regression                                     Number of obs =  3,152
##                                                         LR chi2(3)    =  53.33
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -858.93362                             Pseudo R2     = 0.0301
##
```

```
## ------------------------------------------------------------------------------
##        chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##      1.arcus |   15.70823   17.90923     2.42   0.016     1.681347    146.7564
##          age |   1.093788    .016287     6.02   0.000     1.062328     1.12618
##              |
##  arcus#c.age |
##           1  |   .9513913   .0222084    -2.13   0.033     .9088444    .9959301
##              |
##        _cons |   .0011271   .0008093    -9.45   0.000     .0002759     .004604
## ------------------------------------------------------------------------------
## Note: _cons estimates baseline odds.
##
## . lincom age*10, or
##
##  ( 1)  10*[chd69]age = 0
##
## ------------------------------------------------------------------------------
##        chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          (1) |   2.450936   .3649546     6.02   0.000     1.830562    3.281553
## ------------------------------------------------------------------------------
```

4) Association between *chd69* and *age* in patients with arcus? OR and 95% CI.

The fitted model is $log(p/(1-p)) = (-6.788+2.754)+(0.090-0.050)age = -4.034+0.04age$
Now the association of CHD with age is described by: $\widehat{\beta}_2 + \beta_3 = 0.04$ (on the log-odds scale).
To get an OR and its 95% we need to use *lincom* again as follows:

```
use wcgs.dta
logistic chd69 i.arcus##c.age, coef
lincom age + 1.arcus#c.age, or
## . use wcgs.. logistic chd69 i.arcus##c.age, coef
##
## Logistic regression                                     Number of obs =  3,152
##                                                         LR chi2(3)    =  53.33
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -858.93362                             Pseudo R2     = 0.0301
##
## ------------------------------------------------------------------------------
##        chd69 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
```

```
##      1.arcus |   2.754185   1.140118     2.42   0.016     .5195952    4.988774
##          age |    .089647   .0148904     6.02   0.000     .0604623    .1188317
##              |
##  arcus#c.age |
##           1  |  -.0498298   .0233431    -2.13   0.033    -.0955814   -.0040782
##              |
##        _cons |  -6.788086   .7179977    -9.45   0.000    -8.195335   -5.380836
## ------------------------------------------------------------------------------
##
## . lincom age + 1.arcus#c.age, or
##
##  ( 1)  [chd69]age + [chd69]1.arcus#c.age = 0
##
## ------------------------------------------------------------------------------
##        chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          (1) |    1.04062   .0187073     2.21   0.027     1.004593     1.07794
## ------------------------------------------------------------------------------
```

In patients with arcus, OR=1.04, 95% CI=(1.00 ; 1.08). This means that for those patients, the odds of CHD increases with age but at a slower rate, i.e the odds is 4% bigger, 95% CI=(0% ; 8%) per additional year of age. You can also notice on the plot given p. 164 that the probability of CHD occurrence is higher at a younger age. The two lines cross at a later age (around age 50), which means that older patients with arcus are at somewhat lower risk than patients without arcus. You can also get the OR for a 10-year increment by multiplying everything by 10 in the *lincom* command.

5) Can we interpret the coefficient of *arcus* alone? How can we get a more meaningful coefficient for *arcus*?

The coefficient for arcus (β_1) represents the effect of arcus for someone aged 0 (at birth), assuming we can extrapolate back to that age. It makes little sense. One way to overcome the problem is to centre age using a meaninful value e.g. the age sample mean= 46.275

```
use wcgs.dta
sum age
gen age_centred=age-46.3
logistic chd69 i.arcus##c.age_centred, coef
logistic chd69 i.arcus##c.age_centred
## . use wcgs.. sum age
##
##     Variable |        Obs        Mean    Std. dev.       Min        Max
```

```
## -------------+--------------------------------------------------------
##          age |      3,154    46.27869    5.524045         39         59
##
## . gen age_centred=age-46.3
##
## . logistic chd69 i.arcus##c.age_centred, coef
##
## Logistic regression                                     Number of obs =  3,152
##                                                         LR chi2(3)    =  53.33
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -858.93362                             Pseudo R2     = 0.0301
##
## ------------------------------------------------------------------------------
##               chd69 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
## --------------------+---------------------------------------------------------
##             1.arcus |   .4470638   .1448178     3.09   0.002     .1632261    .7309016
##         age_centred |    .089647   .0148904     6.02   0.000     .0604623    .1188317
##                     |
## arcus#c.age_centred |
##                   1 |  -.0498298   .0233431    -2.13   0.033    -.0955814   -.0040782
##                     |
##               _cons |   -2.63743    .087782   -30.05   0.000     -2.80948   -2.465381
## ------------------------------------------------------------------------------
##
## . logistic chd69 i.arcus##c.age_centred
##
## Logistic regression                                     Number of obs =  3,152
##                                                         LR chi2(3)    =  53.33
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -858.93362                             Pseudo R2     = 0.0301
##
## ------------------------------------------------------------------------------
##               chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## --------------------+---------------------------------------------------------
##             1.arcus |   1.563714   .2264537     3.09   0.002     1.177303    2.076952
##         age_centred |   1.093788    .016287     6.02   0.000     1.062328     1.12618
##                     |
## arcus#c.age_centred |
##                   1 |   .9513913   .0222084    -2.13   0.033     .9088444    .9959301
##                     |
##               _cons |   .0715449   .0062804   -30.05   0.000     .0602363    .0849765
## ------------------------------------------------------------------------------
```

```
## Note: _cons estimates baseline odds.
```

Now get get $\hat{\beta}_1 = 0.44$ and OR=1.56, 95% CI=(1.18 ; 2.08). The odds of CHD is 56% bigger, 95% CI=(18% ; 108%) for someone of average age with arcus (compared with someone of the same age without arcus). of course, the association with arcus depends on age by symmetry, as discussed in the examples provided in the textbook. These interpretation assumes that the model is correct (linearity, no confounding)

R code and output

1) Compulsory reading: (Section 5.2.4. p 163-165)

This reading explains how to introduce a possible interaction the between *age* seen this time as a continuous variable and *arcus* (coded 0/1).

2) Reproduce the output

Perhaps it helps to write down the model first i.e.

$$log(p/(1-p)) = \beta_0 + \beta_1 arcus + \beta_2 age + \beta_3 age * arcus \tag{0.4}$$

where is the probability of CHD over the course of the study given the covariates. This model can be rewritten separately for patients without arcus

$$log(p/(1-p)) = \beta_0 + \beta_2 age \tag{0.5}$$

and patients with arcus

$$log(p/(1-p)) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) age \tag{0.6}$$

We clearly see that the slope of the association with age (i.e. the log-OR) is not the same in the two arcus groups (β_2 vs $\beta_2 + \beta_3$)

```
wcgs <- read.csv("https://www.dropbox.com/s/uc29ddv337zcxk6/wcgs.csv?dl=1")
out<-glm(chd69 ~ arcus*age, family=binomial, data=wcgs)
summary(out)
##
## Call:
## glm(formula = chd69 ~ arcus * age, family = binomial, data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6350  -0.4579  -0.3832  -0.2950   2.5801
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.78809    0.71797  -9.455  < 2e-16 ***
## arcus        2.75418    1.14010   2.416   0.0157 *
## age          0.08965    0.01489   6.021 1.74e-09 ***
## arcus:age   -0.04983    0.02334  -2.135   0.0328 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1771.2  on 3151  degrees of freedom
## Residual deviance: 1717.9  on 3148  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 1725.9
##
## Number of Fisher Scoring iterations: 5
```

Note that you need to we assume that arcus is coded 0/1; otherwise you will have to define arcus as a factor or use *factor(arcus)* in the model. specify the type of variable you are using here. The default in Stata is *categorical* covariates. The analysis with age as a continuous variable confirms what we found with the dichotomised version of age at baseline; we have a significant interaction between *age* and *arcus*.

3) Association between *chd69* and *age* in patients without arcus? OR and 95% CI

The fitted model is $log(p/(1-p)) = -6.788 + 0.09age$ (up to rounding) with the association being described by $\hat{\beta}_2 = 0.09$. To get the OR you can simply take the exponential of the age coefficient $\hat{\beta}_1$ and do something similar for the 95% CI, yielding OR=1.09, 95% CI=(1.06 ; 1. 13). This means that for patients *without arcus* the odds of CHD is 9% bigger, 95% CI=(6% ; 13%) per additional year of age. If you wanted to describe the association for a 10-year age increment, you can 1) use the trick Vittinghof et al (2012) described, rescale age by dividing by 10 and repeat the procedure; 2) mutiply everything by 10 before exponentiting. This gives you OR=2.45, OR=(1.83 ; 3.28) as indicated below

```
out<-glm(chd69 ~ arcus*age, family=binomial, data=wcgs)
coef<-summary(out)$coef[,1]
SE<-summary(out)$coef[,2]
OR=exp(coef[3])
# 3rd element (3rd row of the table)
lower=exp(coef[3]-1.96*SE[3])
upper=exp(coef[3]+1.96*SE[3])
c(OR, lower, upper)
```

```
##      age      age      age
## 1.093788 1.062328 1.126180
# for a 10 year increment
OR=exp(10*coef[3])
lower=exp(10*(coef[3]-1.96*SE[3]))
upper=exp(10*(coef[3]+1.96*SE[3]))
c(OR, lower, upper)
##      age      age      age
## 2.450936 1.830571 3.281537
```

4) Association between *chd69* and *age* in patients with arcus? OR and 95% CI.

The fitted model is $log(p/(1-p)) = (-6.788+2.754)+(0.090-0.050)age = -4.034+0.04age$ Now the association of CHD with age is described by: $\hat{\beta}_2 + \beta_3 = 0.04$ (on the log-odds scale). To get an oR and its 95% we need to use the command *glht* of *lincomp* as follows:

```
library(multcomp)
## Loading required package: mvtnorm
## Loading required package: survival
## Loading required package: TH.data
## Loading required package: MASS
##
## Attaching package: 'MASS'
## The following object is masked from 'package:gtsummary':
##
##     select
## The following object is masked from 'package:dplyr':
##
##     select
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
##     geyser
lincom <- glht(out,linfct=c("age+arcus:age=0"))
lincom
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##                       Estimate
## age + arcus:age == 0  0.03982
```

```
out2<-summary(lincom)$test
OR<-exp(out2$coefficients)
lower<-exp(out2$coefficients -1.96*out2$sigma)
upper<-exp(out2$coefficients +1.96*out2$sigma)
cbind(OR,lower,upper)
##                        OR    lower   upper
## age + arcus:age 1.04062 1.004593 1.07794
# for a 10 year-increment
OR<-exp(10*out2$coefficients)
lower<-exp(10*(out2$coefficients -1.96*out2$sigma))
upper<-exp(10*(out2$coefficients +1.96*out2$sigma))
cbind(OR,lower,upper)
##                       OR    lower    upper
## age + arcus:age 1.4891 1.046887 2.118105
```

In patients with arcus, OR=1.04, 95% CI=(1.00 ; 1.08). This means that for those patients, the odds of CHD increases with age but at a slower rate, i.e the odds is 4% bigger, 95% CI=(0% ; 8%) per additional year of age. You can also notice on the plot given p. 164 that the probability of CHD occurrence is higher at a younger age. The two lines cross at a later age (around age 50), which means that older patients with arcus are at somewhat lower risk than patients without arcus. You can get the OR for a 10-year increment by multiplying everything by 10 before exponentiating. This gives you OR=1.49, OR=(1.05 ; 2.12).

5) Can we interpret the coefficient of *arcus* alone? How can we get a more meaningful coefficient for *arcus*?

The coefficient for arcus (β_1) represents the effect of arcus for someone aged 0 (at birth), assuming we can extrapolate back to that age. It makes little sense. One way to overcome the problem is to centre age using a meaninful value e.g. the age sample mean= 46.275

```
wcgs$age_centred<-wcgs$age-mean(wcgs$age,na.rm=TRUE)
out1<-glm(chd69 ~ arcus*age_centred, family=binomial, data=wcgs)
summary(out1)
##
## Call:
## glm(formula = chd69 ~ arcus * age_centred, family = binomial,
##     data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6350  -0.4579  -0.3832  -0.2950   2.5801
##
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.63934    0.08786 -30.039  < 2e-16 ***
## arcus              0.44813    0.14499   3.091   0.0020 **
## age_centred        0.08965    0.01489   6.021 1.74e-09 ***
## arcus:age_centred -0.04983    0.02334  -2.135   0.0328 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1771.2  on 3151  degrees of freedom
## Residual deviance: 1717.9  on 3148  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 1725.9
##
## Number of Fisher Scoring iterations: 5
coef<-summary(out1)$coef[,1]
SE<-summary(out1)$coef[,2]
OR=exp(coef[2])
lower=exp(coef[2]-1.96*SE[2])
upper=exp(coef[2]+1.96*SE[2])
c(OR,lower,upper)
##    arcus    arcus    arcus
## 1.565375 1.178146 2.079877
```

Now get get $\hat{\beta}_1 = 0.44$ and OR=1.56, 95% CI=(1.18 ; 2.08). The odds of CHD is 56% bigger, 95% CI=(18% ; 108%) for someone of average age with arcus (compared with someone of the same age without arcus). of course, the association with arcus depends on age by symmetry, as discussed in the examples provided in the textbook. These interpretation assumes that the model is correct (linearity, no confounding)

Investigation - predicted probability

The implicit assumption is that we are fitting the same model as in the notes, the response is *Chd69* and the covariates *age*, *bmi*, *chol*, *sbp*, *smoke*, *dibpat* considered previously. We will also delete the outlier in cholesterol (*chol=645*).

Stata code and output

1) calculate the predicted probability of CHD occurrence for a patient with the following characteristics: *age=50*, *BMI=27*, *chol=200*, *sbp=150*, *smoke=1*, *dibpat=0*. Give the 95% CI.

Here we compute the linear predictor, its 95% CI and transform it to the probability scale using the reciprocal of logit. This is done automatically using the *pr* option in *adjust* or *margins*

```
use wcgs.dta

drop if missing(chd69) | missing(bmi) | missing(age) | missing(sbp) | missing(smoke) | mis
drop if chol ==645
** n=3141 observations
logistic chd69 age chol sbp bmi smoke dibpat, coef
adjust age=50 bmi=27 chol=200 sbp=150 smoke=1 dibpat=0, ci pr
## . use wcgs..
## . drop if missing(chd69) | missing(bmi) | missing(age) | missing(sbp) | missing(smoke)
## (12 observations deleted)
##
## . drop if chol ==645
## (1 observation deleted)
##
## . ** n=3141 observations
## . logistic chd69 age chol sbp bmi smoke dibpat, coef
##
## Logistic regression                                     Number of obs =  3,141
##                                                         LR chi2(6)    = 184.34
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -794.92603                             Pseudo R2     = 0.1039
##
## ------------------------------------------------------------------------------
##        chd69 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          age |   .0604453    .011969     5.05   0.000     .0369866    .0839041
##         chol |   .0106408   .0015267     6.97   0.000     .0076485    .0136332
##          sbp |   .0180675   .0041204     4.38   0.000     .0099917    .0261433
##          bmi |   .0549478   .0265311     2.07   0.038     .0029478    .1069478
##        smoke |   .6038582   .1410863     4.28   0.000     .3273341    .8803823
##       dibpat |   .6965686   .1443722     4.82   0.000     .4136043     .979533
##        _cons |  -12.27086   .9821107   -12.49   0.000    -14.19577   -10.34596
## ------------------------------------------------------------------------------
##
## . adjust age=50 bmi=27 chol=200 sbp=150 smoke=1 dibpat=0, ci pr
##
## ------------------------------------------------------------------------------
##      Dependent variable: chd69     Equation: chd69     Command: logistic
## Covariates set to value: age = 50, bmi = 27, chol = 200, sbp = 150, smoke = 1, dibpat =
```

```
## ------------------------------------------------------------------------------
##
## ---------------------------------------------
##       All |         pr          lb          ub
## ----------+----------------------------------
##           |    .089248    [.064873    .121591]
## ---------------------------------------------
##      Key:  pr         =  Probability
##            [lb , ub]  =  [95% Confidence Interval]
```

The predicted CHD probability for that patient's profile is 8.9%, 95% CI=(6.5% ; 12.2%)

2) Represent the probability of an event as a function of age for a particular patient profile, e.g. use *BMI=27*, *chol=200*, *sbp=150*, *smoke=1*, *dibpat=0* and let *age* free to vary.

The plot can be produced using the command *marginplot* after running the appropriate *margins* command

```
use wcgs.dta

drop if missing(chd69) | missing(bmi) | missing(age) | missing(sbp) | missing(smoke) | mis
drop if chol ==645
** n=3141 observations
logistic chd69 age chol sbp bmi smoke dibpat, coef
margins,  at(age=(20(5)60) bmi=27 chol=200 sbp=150 smoke=1 dibpat=0)
marginsplot,  name(temp1)
## . use wcgs..
## . drop if missing(chd69) | missing(bmi) | missing(age) | missing(sbp) | missing(smoke)
## (12 observations deleted)
##
## . drop if chol ==645
## (1 observation deleted)
##
## . ** n=3141 observations
## . logistic chd69 age chol sbp bmi smoke dibpat, coef
##
## Logistic regression                                     Number of obs =  3,141
##                                                         LR chi2(6)    = 184.34
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -794.92603                             Pseudo R2     = 0.1039
##
## ------------------------------------------------------------------------------
```

```
##        chd69 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          age |   .0604453    .011969     5.05   0.000     .0369866    .0839041
##         chol |   .0106408   .0015267     6.97   0.000     .0076485    .0136332
##          sbp |   .0180675   .0041204     4.38   0.000     .0099917    .0261433
##          bmi |   .0549478   .0265311     2.07   0.038     .0029478    .1069478
##        smoke |   .6038582   .1410863     4.28   0.000     .3273341    .8803823
##       dibpat |   .6965686   .1443722     4.82   0.000     .4136043     .979533
##        _cons |  -12.27086   .9821107   -12.49   0.000    -14.19577   -10.34596
## ------------------------------------------------------------------------------
##
## . margins,  at(age=(20(5)60) bmi=27 chol=200 sbp=150 smoke=1 dibpat=0)
##
## Adjusted predictions                                  Number of obs = 3,141
## Model VCE: OIM
##
## Expression: Pr(chd69), predict()
## 1._at: age    =  20
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 2._at: age    =  25
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 3._at: age    =  30
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 4._at: age    =  35
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
```

```
## 5._at: age    =  40
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 6._at: age    =  45
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 7._at: age    =  50
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 8._at: age    =  55
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 9._at: age    =  60
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
##
## ------------------------------------------------------------------------------
##              |            Delta-method
##              |     Margin   std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          _at |
##           1  |   .0157318   .0058583     2.69   0.007     .0042498    .0272138
##           2  |   .0211656   .0067627     3.13   0.002      .007911    .0344201
##           3  |    .028422   .0076689     3.71   0.000     .0133912    .0434527
##           4  |   .0380693   .0085776     4.44   0.000     .0212576    .0548811
##           5  |   .0508201   .0096192     5.28   0.000     .0319669    .0696733
```

```
##           6  |   .0675417   .0112361     6.01   0.000     .0455193    .0895641
##           7  |   .0892479   .0143243     6.23   0.000     .0611728     .117323
##           8  |   .1170543   .0199997     5.85   0.000     .0778555     .156253
##           9  |   .1520774   .0291205     5.22   0.000     .0950022    .2091526
## ------------------------------------------------------------------------------
##
## . marginsplot,  name(temp1)
##
## Variables that uniquely identify margins: age
```

3) Contrast with a plot of the CHD probability vs age for *smoke=0*, the other characteristics remaining the same. Draw the 2 plots side-by-side.

The plot can be produced using the command *marginplot* after running the appropriate *margins* command (twice) amd combining the plots

```
use wcgs.dta
drop if missing(chd69) | missing(bmi) | missing(age) | missing(sbp) | missing(smoke) | mis
drop if chol ==645
** n=3141 observations
logistic chd69 age chol sbp bmi smoke dibpat, coef
margins,  at(age=(20(5)60) bmi=27 chol=200 sbp=150 smoke=1 dibpat=0)
marginsplot,  name(temp2)

margins,  at(age=(20(5)60) bmi=27 chol=200 sbp=150 smoke=0 dibpat=0)
marginsplot,  name(temp3)

graph combine temp2 temp3
## . use wcgs.. drop if missing(chd69) | missing(bmi) | missing(age) | missing(sbp) | miss
## (12 observations deleted)
##
## . drop if chol ==645
## (1 observation deleted)
##
## . ** n=3141 observations
## . logistic chd69 age chol sbp bmi smoke dibpat, coef
##
## Logistic regression                                     Number of obs =  3,141
##                                                         LR chi2(6)    = 184.34
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -794.92603                             Pseudo R2     = 0.1039
##
```

```
## ------------------------------------------------------------------------------
##        chd69 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          age |   .0604453    .011969     5.05   0.000     .0369866    .0839041
##         chol |   .0106408   .0015267     6.97   0.000     .0076485    .0136332
##          sbp |   .0180675   .0041204     4.38   0.000     .0099917    .0261433
##          bmi |   .0549478   .0265311     2.07   0.038     .0029478    .1069478
##        smoke |   .6038582   .1410863     4.28   0.000     .3273341    .8803823
##       dibpat |   .6965686   .1443722     4.82   0.000     .4136043     .979533
##        _cons |  -12.27086   .9821107   -12.49   0.000    -14.19577   -10.34596
## ------------------------------------------------------------------------------
##
## . margins,  at(age=(20(5)60) bmi=27 chol=200 sbp=150 smoke=1 dibpat=0)
##
## Adjusted predictions                                      Number of obs = 3,141
## Model VCE: OIM
##
## Expression: Pr(chd69), predict()
## 1._at: age    =  20
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 2._at: age    =  25
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 3._at: age    =  30
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 4._at: age    =  35
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
```

```
##        dibpat =   0
## 5._at: age    =  40
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 6._at: age    =  45
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 7._at: age    =  50
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 8._at: age    =  55
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
## 9._at: age    =  60
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   1
##        dibpat =   0
##
## ------------------------------------------------------------------------------
##              |            Delta-method
##              |     Margin   std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          _at |
##           1  |   .0157318   .0058583     2.69   0.007     .0042498    .0272138
##           2  |   .0211656   .0067627     3.13   0.002      .007911    .0344201
##           3  |    .028422   .0076689     3.71   0.000     .0133912    .0434527
##           4  |   .0380693   .0085776     4.44   0.000     .0212576    .0548811
```

```
##            5  |   .0508201   .0096192     5.28   0.000     .0319669    .0696733
##            6  |   .0675417   .0112361     6.01   0.000     .0455193    .0895641
##            7  |   .0892479   .0143243     6.23   0.000     .0611728     .117323
##            8  |   .1170543   .0199997     5.85   0.000     .0778555     .156253
##            9  |   .1520774   .0291205     5.22   0.000     .0950022    .2091526
## ------------------------------------------------------------------------------
##
## . marginsplot,  name(temp2)
##
## Variables that uniquely identify margins: age
##
## .
## . margins,  at(age=(20(5)60) bmi=27 chol=200 sbp=150 smoke=0 dibpat=0)
##
## Adjusted predictions                                  Number of obs = 3,141
## Model VCE: OIM
##
## Expression: Pr(chd69), predict()
## 1._at: age    =  20
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
## 2._at: age    =  25
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
## 3._at: age    =  30
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
## 4._at: age    =  35
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
```

```
##        dibpat =   0
## 5._at: age    =  40
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
## 6._at: age    =  45
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
## 7._at: age    =  50
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
## 8._at: age    =  55
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
## 9._at: age    =  60
##        chol   = 200
##        sbp    = 150
##        bmi    =  27
##        smoke  =   0
##        dibpat =   0
##
## ------------------------------------------------------------------------------
##              |            Delta-method
##              |     Margin   std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          _at |
##           1  |   .0086623   .0033133     2.61   0.009     .0021683    .0151563
##           2  |   .0116833   .0038495     3.04   0.002     .0041385    .0192281
##           3  |    .015741   .0043978     3.58   0.000     .0071215    .0243606
##           4  |   .0211779   .0049563     4.27   0.000     .0114637    .0308922
```

```
##           5  |   .0284384   .0055894     5.09   0.000      .0174834    .0393934
##           6  |   .0380911   .0065384     5.83   0.000      .0252761    .0509062
##           7  |   .0508488   .0083492     6.09   0.000      .0344847    .0672129
##           8  |   .0675792   .0118174     5.72   0.000      .0444175    .0907409
##           9  |   .0892963   .0177387     5.03   0.000       .054529    .1240636
## ------------------------------------------------------------------------------
##
## . marginsplot,  name(temp3)
##
## Variables that uniquely identify margins: age
##
## .
## . graph combine temp2 temp3
```

The CHD probability increases by age and is higher for smokers.

R code and output

1) calculate the predicted probability of CHD occurence for a patient with the following characteristics: *age=50*, *BMI=27*, *chol=200*, *sbp=150*, *smoke=1*, *dibpat=0*. Give the 95% CI.

Here we compute the linear predictor, its 95% CI and transform it to the probability scale using the reciprocal of logit (called expit).

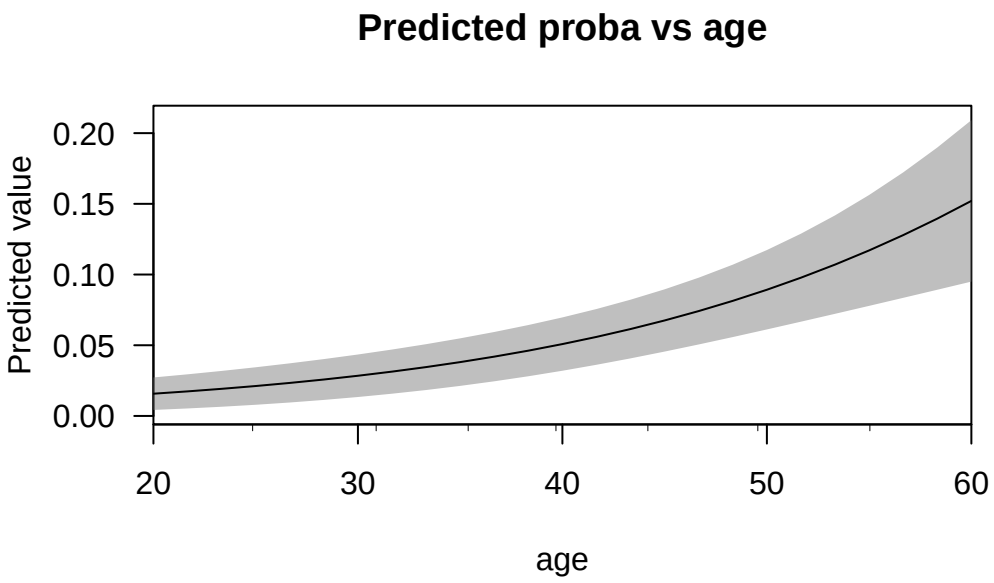
```
myvars <- c("id","chd69", "age", "bmi", "chol", "sbp", "smoke", "dibpat")
wcgs1 <- wcgs[myvars]
wcgs1 <- wcgs1[wcgs1$chol <645,]
wcgs1cc <- na.omit(wcgs1) # 3141 x 11
model1 <- glm(chd69 ~ age + chol + sbp + bmi + smoke + dibpat, family=binomial, data=wcgs1
new <- data.frame(age = 50, bmi=27, chol =200, sbp=150, smoke=1, dibpat=0)
out <- predict(model1, new, type="link",se.fit=TRUE)
mean<-out$fit
SE<-out$se.fit
CI=c(mean-1.96*SE,mean+1.96*SE)
f.expit<-function(u){exp(u)/(1+exp(u))}
f.expit(c(mean,CI))
##          1          1          1
## 0.08924790 0.06487243 0.12159140
```

The predicted CHD probability for that patient's profile is 8.9%, 95% CI=(6.5% ; 12.2%)

2) Represent the probability of an event as a function of age for a particular patient profile, e.g. use *BMI=27*, *chol=200*, *sbp=150*, *smoke=1*, *dibpat=0* and let *age* free to vary.

The plot can be produced using the command *cplot* available in the *margins* library

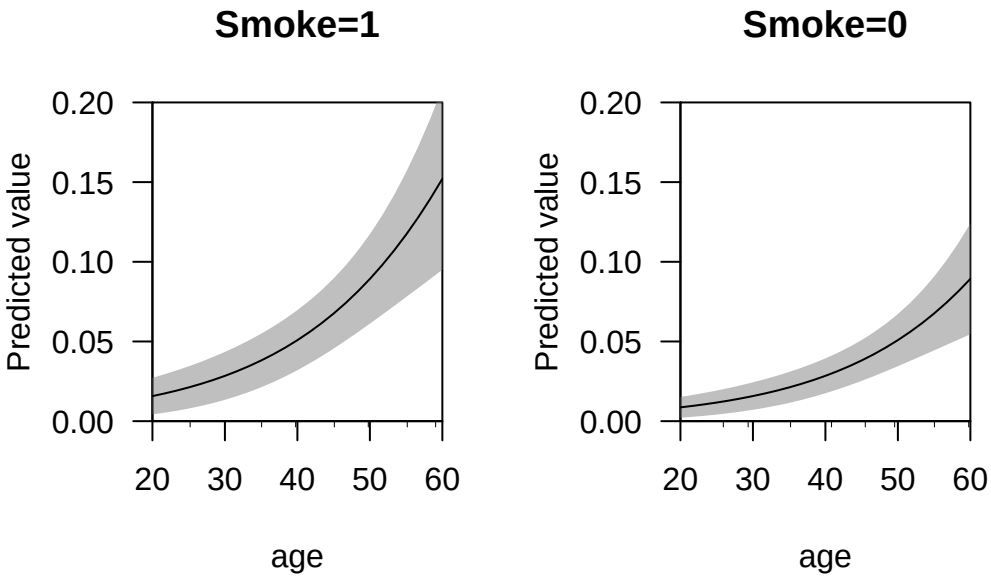
```
require(margins)
## Loading required package: margins
new <- data.frame(age=seq(20,60,5),bmi=27, chol =200, sbp=150, smoke=1, dibpat=0)
cplot(model1, what = "prediction", data=new,main = "Predicted proba vs age")
```



3) Contrast with a plot of the CHD probability vs age for *smoke=0*, the other characteristics remaining the same. Draw the 2 plots side-by-side.

Again *cplot* can be used to produce these plots. The CHD probability increases by age and is higher for smokers.

```
par(mfrow=c(1,2))
new <- data.frame(age=seq(20,60,5),bmi=27, chol =200, sbp=150, smoke=1, dibpat=0)
cplot(model1, what = "prediction", data=new,main = "Smoke=1", ylim=c(0,0.20))
new <- data.frame(age=seq(20,60,5),bmi=27, chol =200, sbp=150, smoke=0, dibpat=0)
cplot(model1, what = "prediction", data=new,main = "Smoke=0",ylim=c(0,0.20))
```


Smoke=1
Smoke=0
Predicted value
0.00
0.05
0.10
0.15
0.20
20
30
40
50
60
age