Stats 111 HW4

Viraj Vijaywargiya

2023-03-02

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(epitools)
library(rmeta)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(nnet)
```

```
ifelse1 =function(test, x, y){ if (test) x else y}

glmCI <- function( model, transform=TRUE, robust=FALSE ){
    link <- model$family$link
    coef <- summary( model )$coef[,1]
    se <- ifelse1( robust, robust.se.glm(model)[,2], summary( model )$coef[,2] )
    zvalue <- coef / se
    pvalue <- 2*(1-pnorm(abs(zvalue)))

    if( transform & is.element(link, c("logit","log")) ){
        ci95.lo <- exp( coef - qnorm(.975) * se )
        ci95.hi <- exp( coef + qnorm(.975) * se )
        est <- exp( coef )
    }
    else{
```

```
        ci95.lo <- coef - qnorm(.975) * se
        ci95.hi <- coef + qnorm(.975) * se
        est <- coef
    }
    rslt <- round( cbind( est, ci95.lo, ci95.hi, zvalue, pvalue ), 4 )
    colnames( rslt ) <- ifelse1(    robust,
                    c("Est", "robust ci95.lo", "robust ci95.hi", "robust z value", "robust Pr(>|z|)"),
                    c("Est", "ci95.lo", "ci95.hi", "z value", "Pr(>|z|)") )
    colnames( rslt )[1] <- ifelse( transform & is.element(link, c("logit","log")), "exp( Est )", "Est" )
    rslt
    }

linContr.glm <- function( contr.names, contr.coef=rep(1,length(contr.names)), model, transform=TRUE ){
    beta.hat <- model$coef
    cov.beta <- vcov( model )

    contr.index <- match( contr.names, dimnames( cov.beta )[[1]] )
    beta.hat <- beta.hat[ contr.index ]
    cov.beta <- cov.beta[ contr.index,contr.index ]
    est <- contr.coef %*% beta.hat
    se.est <- sqrt( contr.coef %*% cov.beta %*% contr.coef )
    zStat <- est / se.est
    pVal <- 2*pnorm( abs(zStat), lower.tail=FALSE )
    ci95.lo <- est - qnorm(.975)*se.est
    ci95.hi <- est + qnorm(.975)*se.est

    link <- model$family$link
    if( transform & is.element(link, c("logit","log")) ){
        ci95.lo <- exp( ci95.lo )
        ci95.hi <- exp( ci95.hi )
        est <- exp( est )
        cat( "\nTest of H_0: exp( " )
        for( i in 1:(length( contr.names )-1) ){
            cat( contr.coef[i], "*", contr.names[i], " + ", sep="" )
            }
        cat( contr.coef[i+1], "*", contr.names[i+1], " ) = 1 :\n\n", sep="" )
        }
    else{
        cat( "\nTest of H_0: " )
        for( i in 1:(length( contr.names )-1) ){
            cat( contr.coef[i], "*", contr.names[i], " + ", sep="" )
            }
        cat( contr.coef[i+1], "*", contr.names[i+1], " = 0 :\n\n", sep="" )
        }
    rslt <- data.frame( est, se.est, zStat, pVal, ci95.lo, ci95.hi )
    colnames( rslt )[1] <- ifelse( transform && is.element(link, c("logit","log")), "exp( Est )", "Est"
    round( rslt, 8 )
}

lrtest <- function( fit1, fit2 ){
    cat( "\nAssumption: Model 1 nested within Model 2\n\n" )
    rslt <- anova( fit1, fit2 )
    rslt <- cbind( rslt, c("", round( pchisq( rslt[2,4], rslt[2,3], lower.tail=FALSE ), 4 ) ) )
```

```
    rslt[,2] <- round( rslt[,2], 3 )
    rslt[,4] <- round( rslt[,4], 3 )
    rslt[1,3:4] <- c( "", "" )
    names( rslt )[5] <- "pValue"
    rslt
}
```

1. 1a)

```
mcat = read.table("/Users/virajvijaywargiya/Downloads/MedGPA-2.txt", fill=TRUE, header=TRUE)

mcat$male = mcat$Sex
mod1 =glm(Acceptance~male, data = mcat, family="binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = Acceptance ~ male, family = "binomial", data = mcat)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.435  -1.084   0.940   0.940   1.274
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5878     0.3944   1.490    0.136
## maleM        -0.8109     0.5528  -1.467    0.142
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 73.594  on 53  degrees of freedom
## AIC: 77.594
##
## Number of Fisher Scoring iterations: 4
```

Null deviance: 75.791 on 54 degrees of freedom, Residual deviance: 73.594 on 53 degrees of freedom. The difference between the null deviance and the residual deviance is 2.197. A small difference between the null and residual deviances signifies that the model is a good fit for the data. In other words, most of the variability in the response variable is explained by the model. This suggests that the coefficient of sex in the model, which is -0.8109, is a good predictor of acceptance into med school. A negative coefficient for male indicates that being male is associated with a lower probability of being accepted into med school, compared to being female.

1b)

```
mod2 =glm(Acceptance~MCAT, data = mcat, family="binomial")
summary(mod2)
```

```
##
## Call:
## glm(formula = Acceptance ~ MCAT, family = "binomial", data = mcat)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7878  -1.0330   0.4256   0.9225   1.6601
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.71245    3.23645  -2.692  0.00710 **
## MCAT         0.24596    0.08938   2.752  0.00592 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 64.697  on 53  degrees of freedom
## AIC: 68.697
##
## Number of Fisher Scoring iterations: 4
```

Null deviance: 75.791 on 54 degrees of freedom, Residual deviance: 64.697 on 53 degrees of freedom. The null deviance is the same as in the previous model because the response variable and the sample size are the same. The null deviance is a measure of the total variability in the response variable when no explanatory variables are included in the model, and this variability is the same in both models.

However, the residual deviance is different because the model includes a different explanatory variable (MCAT score instead of sex). The residual deviance measures the amount of variability in the response variable that cannot be explained by the model after accounting for the explanatory variable(s), and this amount changes when a different variable is used in the model.

In this case, the residual deviance is 64.697 on 53 degrees of freedom, which is much smaller than the difference between the null and residual deviances in the previous model (2.197). This indicates that the model with MCAT scores as the explanatory variable is a much better fit for the data than the model with sex as the explanatory variable. The smaller difference between the null and residual deviances suggests that the MCAT score is a stronger predictor of acceptance into med school than sex. Additionally, the positive coefficient for MCAT score suggests that higher scores are associated with a higher probability of being accepted into med school.

1c) Can test the proposed model against the null model by looking at Null - Residual deviance (that is Null deviance minus Residual deviance) which has an approximate chi-squared distribution with degrees of freedom dfn - dfm = p (where p is the number of slope coefficients in the proposed model). In this case, Null - Residual deviance = 75.791 - 64.697.

The null hypothesis is that the proposed model (with MCAT scores as the explanatory variable) does not provide a better fit to the data than the null model (with no covariates). The alternative hypothesis is that the proposed model provides a better fit to the data than the null model.

1d)

```
mod3 =glm(Acceptance~male+MCAT+MCAT*male, data = mcat, family="binomial")
summary(mod3)
```

```
##
## Call:
## glm(formula = Acceptance ~ male + MCAT + MCAT * male, family = "binomial",
##     data = mcat)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.7857  -0.9770   0.3549   0.9417   2.0304
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.1804     4.3247  -1.429    0.153
## maleM        -7.2122     7.1083  -1.015    0.310
## MCAT          0.1887     0.1212   1.557    0.119
## maleM:MCAT    0.1697     0.1946   0.872    0.383
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 60.924  on 51  degrees of freedom
## AIC: 68.924
##
## Number of Fisher Scoring iterations: 5
```

Estimated model: log(u/1-u) = -6.1804 - 7.2122 maleM + 0.1887 MCAT + 0.1697 maleM*MCAT

From the model, a 1 unit increase in MCAT score is associated with an increase in the log odds of being accepted into med school by 0.1887 units. We can interpret the effect of a 1 unit increase in MCAT score on the odds of being accepted by exponentiating the coefficient for MCAT, exp(0.1887) = 1.2076. This means that a 1 unit increase in MCAT score is associated with a 20.76% increase in the odds of being accepted into med school, holding sex constant.

1e)

```
linContr.glm <- function( contr.names, contr.coef=rep(1,length(contr.names)), model, transform=TRUE
    beta.hat <- model$coef
    cov.beta <- vcov( model )

    contr.index <- match( contr.names, dimnames( cov.beta )[[1]] )
    beta.hat <- beta.hat[ contr.index ]
    cov.beta <- cov.beta[ contr.index,contr.index ]
    est <- contr.coef %*% beta.hat
    se.est <- sqrt( contr.coef %*% cov.beta %*% contr.coef )
    zStat <- est / se.est
    pVal <- 2*pnorm( abs(zStat), lower.tail=FALSE )
    ci95.lo <- est - qnorm(.975)*se.est
    ci95.hi <- est + qnorm(.975)*se.est

    link <- model$family$link
    if( transform & is.element(link, c("logit","log")) ){
        ci95.lo <- exp( ci95.lo )
        ci95.hi <- exp( ci95.hi )
        est <- exp( est )
        cat( "\nTest of H_0: exp( " )
        for( i in 1:(length( contr.names )-1) ){
            cat( contr.coef[i], "*", contr.names[i], " + ", sep="" )
            }
        cat( contr.coef[i+1], "*", contr.names[i+1], " ) = 1 :\n\n", sep="" )
        }
    else{
        cat( "\nTest of H_0: " )
```

```
        for( i in 1:(length( contr.names )-1) ){
            cat( contr.coef[i], "*", contr.names[i], " + ", sep="" )
            }
        cat( contr.coef[i+1], "*", contr.names[i+1], " = 0 :\n\n", sep="" )
        }
    rslt <- data.frame( est, se.est, zStat, pVal, ci95.lo, ci95.hi )
    colnames( rslt )[1] <- ifelse( transform && is.element(link, c("logit","log")), "exp( Est )", '
    round( rslt, 8 )
}
linContr.glm(c("MCAT", "maleM:MCAT"), c(1,1), mod3)
```

```
##
## Test of H_0: exp( 1*MCAT + 1*maleM:MCAT ) = 1 :

##   exp( Est )    se.est    zStat       pVal  ci95.lo ci95.hi
## 1    1.43098 0.1522573 2.353646 0.01859033 1.061774 1.92857
```

1f)

2. 2a)

```
midwest = read.table("/Users/virajvijaywargiya/Downloads/MidwestSales.txt", fill=TRUE, header=FALSE
names(midwest)=c("id","price","sqft","bed","bath","ac","garage","pool","year","quality","style","lo

mod = glm(ac~sqft+lot+pool, family="binomial", data=midwest)
summary(mod)
```

```
##
## Call:
## glm(formula = ac ~ sqft + lot + pool, family = "binomial", data = midwest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3262   0.1710   0.4158   0.6822   1.5197
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.309e+00  5.784e-01  -2.263  0.02361 *
## sqft         1.828e-03  3.097e-04   5.901 3.62e-09 ***
## lot         -3.431e-05  9.396e-06  -3.652  0.00026 ***
## pool         1.397e+00  1.037e+00   1.347  0.17792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 473.59  on 521  degrees of freedom
## Residual deviance: 407.26  on 518  degrees of freedom
## AIC: 415.26
##
## Number of Fisher Scoring iterations: 6
```

Estimated model: log(u(i)/1-u(i)) = -1.309 + 0.00183 sqft(i) - 0.0000343 lot(i) + 1.397 I(pool(i) = 1).

Null deviance: 473.59 on 521 degrees of freedom, Residual deviance: 407.26 on 518 degrees of freedom. The difference between the null deviance and the residual deviance is 66.33 on 3 degrees of freedom. This difference is large enough to suggest that the model with predictors is a better fit to the data than the null model.

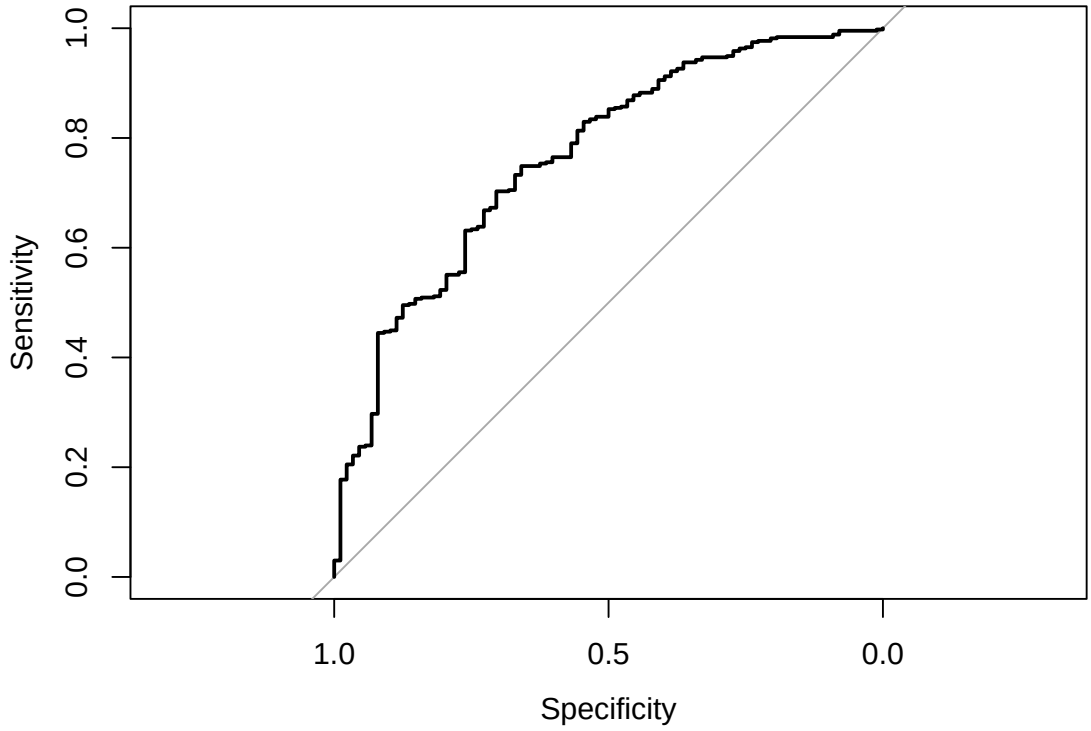
2b)

```
roc.curve = roc(midwest$ac~fitted(mod))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roc.curve)
```



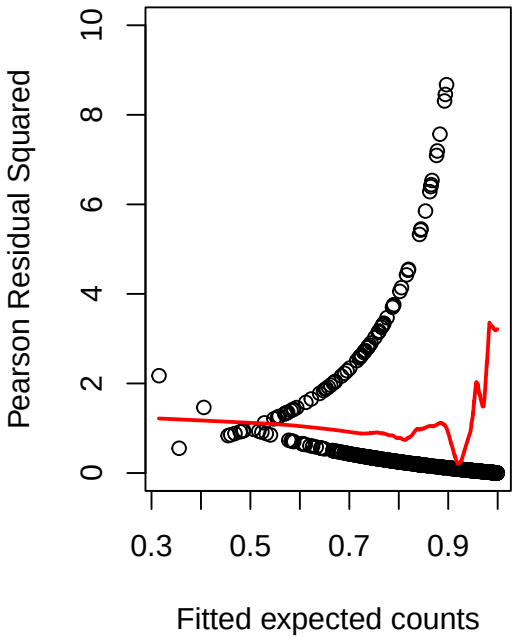

```
roc.curve
```

```
##
## Call:
## roc.formula(formula = midwest$ac ~ fitted(mod))
##
## Data: fitted(mod) in 88 controls (midwest$ac 0) < 434 cases (midwest$ac 1).
## Area under the curve: 0.7649
```

Area under the curve is 0.7649, which indicates that the model is moderately good at predicting whether the house will have air conditioning or not. This suggests that the model is able to distinguish between positive and negative cases better than random guessing, but there is still room for improvement.

2c)

```
par(mfrow=c(1,2))
presids = residuals(mod, type="pearson")
muhat = fitted(mod)
plot(muhat, presids^2, xlab="Fitted expected counts", ylab="Pearson Residual Squared", ylim=c(0,10)
sfit = supsmu(muhat, presids^2)
lines(sfit$x[order(sfit$x)] , sfit$y[order(sfit$x)], col="red", lwd=2)
```



The smoothed (red) line is roughly linear and around 1. This implies that our variance specification ,V(u), is appropriate. The smoothed red line is showing an approximation to the variance of the residuals given the fitted probability u.

2d)

```
summary(midwest[, c(3,8,12)])
```

```
##       sqft           pool             lot
##  Min.   : 980   Min.   :0.00000   Min.   : 4560
##  1st Qu.:1701   1st Qu.:0.00000   1st Qu.:17205
##  Median :2061   Median :0.00000   Median :22200
##  Mean   :2261   Mean   :0.06897   Mean   :24370
##  3rd Qu.:2636   3rd Qu.:0.00000   3rd Qu.:26787
##  Max.   :5032   Max.   :1.00000   Max.   :86830
```

```
midwest[which(hatvalues(mod) == max(hatvalues(mod))),]
```

```
##      id  price sqft bed bath ac garage pool year quality style   lot hwy
## 394 394 232900 1550   4    2  1      2    1 1962       3     2 14998   0
```

Observation 394 has the highest leverage. Compared to the average house that is 2261 square foot, has a lot size of 24370, and no pool, this observation has a smaller square footage and a smaller lot size but has a pool.
Specifically, the observation's square footage is about 68.5% of the average square footage, the lot size is about 61.5% of the average lot size, and it has a pool while the average house does not. These covariate values suggest that this observation is quite different from the average house in the dataset, and it may have a larger influence on the model's estimates than other observations.

2e)

```
linContr.glm( c("sqft" , "lot") , c(500,1500) , model=mod)
```

```
##
## Test of H_0: exp( 500*sqft + 1500*lot ) = 1 :

##   exp( Est )    se.est    zStat  pVal  ci95.lo  ci95.hi
## 1   2.368631 0.1515497 5.689963 1e-08 1.759941 3.187842
```

The 95% confidence interval for the odds ratio for ac comparing two houses that differ in
sqft by 500 and lot size by 1500 is (1.76, 3.19). This means that we are 95% confident that the true odds ratio lies between 1.76 and 3.19. We can say that if we compare two houses that differ in square footage by 500 and lot size by 1500, the odds of having air conditioning in the house with the larger square footage and lot size are between 1.76 and 3.19 times higher than the odds of having air conditioning in the house with the smaller square footage and lot size.
Since the confidence interval does not include the value 1, we can conclude that the difference in odds of having air conditioning between the two houses is statistically significant at the 0.05 level. This suggests that square footage and lot size are important predictors of air conditioning in houses, and houses with larger square footage and lot size are more likely to have air conditioning than houses with smaller square footage and lot size.

3. Interpreting the effects for each covariate,
   Age: The coefficient estimate for age is -1.320, which means that holding other variables constant, women who are 35 or younger are expected to have lower odds of using oral contraceptives compared to women over 35 years of age.

   Race: The coefficient estimate for race is 0.622, which means that holding other variables constant, white women are expected to have higher odds of using oral contraceptives compared to non-white women.

   Education: The coefficient estimate for education is 0.501, which means that holding other variables constant, women who have at least one year of college education are expected to have higher odds of using oral contraceptives compared to women who have less than one year of college education.

   Marital Status: The coefficient estimate for marital status is -0.460, which means that holding other variables constant, married women are expected to have lower odds of using oral contraceptives compared to unmarried women.

   Confidence Interval for Odds Ratio between Contraceptive Use and Education,
   Odds Ratio = exp(0.501) = 1.651, SE(Log Odds Ratio) = 0.077
   95% Confidence Interval = Odds Ratio +- (1.96 * SE(Log Odds Ratio)) = 1.651 +- (1.96 * 0.077) = (1.501, 1.813).
   Therefore, We are 95% confident that the true odds ratio between contraceptive use and education lies between 1.501 and 1.813. This means that women who have at least one year of college education are 1.501 to 1.813 times more likely to use oral contraceptives compared to women who have less than one year of college education, holding other variables constant. This result is statistically significant since the confidence interval does not include 1.

4.
```
nhanes = read.table( "/Users/virajvijaywargiya/Downloads/nhaneshw.txt", header=TRUE)
nhanes$agegrp = cut( nhanes$age, breaks=c(0,30,40,50,60,71), right=FALSE )
```

   1. 4a)

```
lapply( split( nhanes, nhanes$male), summary)
```

```
## $`0`
##       age              wt              male        htn              rtid       
##  Min.   :20.00   Min.   : 35.90   Min.   :0   Min.   :0.0000   Min.   :   2.0  
##  1st Qu.:30.00   1st Qu.: 61.90   1st Qu.:0   1st Qu.:0.0000   1st Qu.: 870.5  
##  Median :42.00   Median : 72.14   Median :0   Median :0.0000   Median :1764.0  
##  Mean   :43.11   Mean   : 75.71   Mean   :0   Mean   :0.1751   Mean   :1758.4  
##  3rd Qu.:56.00   3rd Qu.: 85.05   3rd Qu.:0   3rd Qu.:0.0000   3rd Qu.:2645.5  
##  Max.   :70.00   Max.   :191.10   Max.   :0   Max.   :1.0000   Max.   :3528.0  
##      agegrp    
##  [0,30) :449  
##  [30,40):401  
##  [40,50):357  
##  [50,60):278  
##  [60,71):394  
##               
##               
## $`1`
##       age              wt              male        htn              rtid       
##  Min.   :20.00   Min.   : 42.70   Min.   :1   Min.   :0.0000   Min.   :   1.0  
##  1st Qu.:32.00   1st Qu.: 72.10   1st Qu.:1   1st Qu.:0.0000   1st Qu.: 897.2  
##  Median :44.00   Median : 82.00   Median :1   Median :0.0000   Median :1769.0  
##  Mean   :45.12   Mean   : 84.84   Mean   :1   Mean   :0.2067   Mean   :1772.5  
##  3rd Qu.:60.00   3rd Qu.: 94.50   3rd Qu.:1   3rd Qu.:0.0000   3rd Qu.:2648.8  
##  Max.   :70.00   Max.   :193.30   Max.   :1   Max.   :1.0000   Max.   :3529.0  
##      agegrp    
##  [0,30) :319  
##  [30,40):332  
##  [40,50):330  
##  [50,60):254  
##  [60,71):415  
##
```

   4b)

```
fit1.full = glm( htn ~ factor(agegrp) + wt + male, family=binomial, data=nhanes )
glmCI( fit1.full )
```

```
##                       exp( Est ) ci95.lo ci95.hi  z value Pr(>|z|)
## (Intercept)               0.0127  0.0076  0.0212 -16.7476   0.0000
## factor(agegrp)[30,40)     2.2562  1.4529  3.5036   3.6234   0.0003
## factor(agegrp)[40,50)     4.8381  3.2085  7.2955   7.5230   0.0000
## factor(agegrp)[50,60)     7.5369  4.9961 11.3700   9.6283   0.0000
## factor(agegrp)[60,71)    14.8658 10.0885 21.9052  13.6461   0.0000
## wt                        1.0154  1.0108  1.0200   6.6467   0.0000
## male                      0.9694  0.8062  1.1655  -0.3309   0.7407
```

```
fit1.red = glm( htn ~ wt + male, family=binomial, data=nhanes )
lrtest( fit1.red, fit1.full )
```

```
##
## Assumption: Model 1 nested within Model 2
```

```
##   Resid. Df Resid. Dev Df Deviance pValue
## 1      3526   3377.187
## 2      3522   2989.850  4  387.337      0
```

The model: log(P[hypertension=1]) = 0.0127 + 2.2562(agegrp[30,40)) + 4.8381(agegrp[40,50)) + 7.5369(agegrp[50,60)) + 14.8658(agegrp[60,71)) + 1.0154(wt) + 0.9694(male).

A typical B coefficient for one of the age group dummy variables represents the change in the log odds of having hypertension when compared to the reference category (ages 20-29). For example, the coefficient of 2.2562 for age group [30,40) means that the log odds of having hypertension for individuals in the age group [30,40) is 2.2562 higher than that of individuals in the reference category, adjusting for sex and weight.

The null hypothesis is that there is no global effect of age on hypertension, i.e., all age group coefficients are equal to 0. The alternative hypothesis is that at least one age group coefficient is not equal to 0. We can use a likelihood ratio test to test this hypothesis, which compares the logistic regression model with age group as a predictor to the null model without age group as a predictor.

From the output, all age group coefficients are significantly different from 0 with very small P-values, indicating that age is a significant predictor of hypertension after adjusting for sex and weight. The coefficients of the age groups increase with age, suggesting a strong positive association between age and hypertension. The weight coefficient is also significant, indicating that higher weight is associated with higher odds of having hypertension. However, the male coefficient is not significant, indicating that sex is not a significant predictor of hypertension after adjusting for age and weight.

4c)

```
fit2 = glm( htn ~ factor(agegrp) + wt + male + factor(agegrp)*male, family=binomial, data=nhane
glmCI(fit2)
```

```
##                             exp( Est ) ci95.lo ci95.hi  z value Pr(>|z|)
## (Intercept)                     0.0069  0.0034  0.0143 -13.4667   0.0000
## factor(agegrp)[30,40)           2.1729  1.0029  4.7079   1.9673   0.0491
## factor(agegrp)[40,50)           8.2170  4.1307 16.3458   6.0022   0.0000
## factor(agegrp)[50,60)          14.7492  7.4409 29.2353   7.7093   0.0000
## factor(agegrp)[60,71)          32.8006 16.9616 63.4303  10.3733   0.0000
## wt                              1.0158  1.0112  1.0204   6.7781   0.0000
## male                            2.7329  1.2657  5.9010   2.5599   0.0105
## factor(agegrp)[30,40):male      1.0208  0.3971  2.6244   0.0428   0.9659
## factor(agegrp)[40,50):male      0.3863  0.1628  0.9166  -2.1576   0.0310
## factor(agegrp)[50,60):male      0.2893  0.1217  0.6876  -2.8080   0.0050
## factor(agegrp)[60,71):male      0.2315  0.1019  0.5258  -3.4953   0.0005
```

```
glmCI(fit2, transform=FALSE)
```

```
##                                Est ci95.lo ci95.hi  z value Pr(>|z|)
## (Intercept)                -4.9745 -5.6985 -4.2505 -13.4667   0.0000
## factor(agegrp)[30,40)       0.7761  0.0029  1.5492   1.9673   0.0491
## factor(agegrp)[40,50)       2.1062  1.4184  2.7940   6.0022   0.0000
## factor(agegrp)[50,60)       2.6912  2.0070  3.3754   7.7093   0.0000
```

```
## factor(agegrp)[60,71)       3.4904  2.8310  4.1499  10.3733   0.0000
## wt                          0.0157  0.0112  0.0202   6.7781   0.0000
## male                        1.0054  0.2356  1.7751   2.5599   0.0105
## factor(agegrp)[30,40):male  0.0206 -0.9236  0.9648   0.0428   0.9659
## factor(agegrp)[40,50):male -0.9513 -1.8154 -0.0871  -2.1576   0.0310
## factor(agegrp)[50,60):male -1.2402 -2.1058 -0.3746  -2.8080   0.0050
## factor(agegrp)[60,71):male -1.4633 -2.2839 -0.6428  -3.4953   0.0005
```

```
lrtest( fit1.full, fit2 )
```

```
##
## Assumption: Model 1 nested within Model 2
```

```
##   Resid. Df Resid. Dev Df Deviance pValue
## 1      3522   2989.850
## 2      3518   2956.953  4   32.897      0
```

To interpret a typical B coefficient for one of the age-by-sex interaction terms, we can use the output from "glmCI(fit2, transform=FALSE)". For example, the B coefficient for the interaction term "factor(agegrp)[40,50):male" is -0.9513. This means that the odds of hypertension for males in the age group 40-50 are expected to be 0.9513 times the odds of hypertension for females in the same age group. In other words, males in this age group are expected to have a lower odds of hypertension than females in the same age group, after adjusting for weight and the other variables in the model. This coefficient is statistically significant (p-value = 0.0310), indicating that there is evidence for a difference in the effect of age on hypertension between males and females in this age group.

Null hypothesis: The effect of age on hypertension is the same for males and females (i.e., all interaction terms are equal to 0). Alternative hypothesis: The effect of age on hypertension is different for males and females (i.e., at least one interaction term is not equal to 0).

The output from the lrtest gives a p-value of 0, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the effect of age on hypertension is different for males and females.

4d)

```
linContr.glm( contr.names=c("(Intercept)", "factor(agegrp)[60,71)", "wt", "male", "factor(agegr
```

```
##
## Test of H_0: exp( 1*(Intercept) + 1*factor(agegrp)[60,71) + 85.543*wt + 1*male + 1*factor(ag
```

```
##   exp( Est )    se.est     zStat  pVal   ci95.lo   ci95.hi
## 1   0.549176 0.1033585 -5.798616 1e-08 0.4484691 0.6724972
```

1.
```
linContr.glm( contr.names=c("(Intercept)", "factor(agegrp)[60,71)", "wt", "male", "factor(agegr
```

```
##
## Test of H_0: exp( 1*(Intercept) + 1*factor(agegrp)[60,71) + 85.543*wt + 1*male + 1*factor(ag
```

```
##   exp( Est )    se.est     zStat  pVal   ci95.lo   ci95.hi
## 1   0.549176 0.1033585 -5.798616 1e-08 0.4484691 0.6724972
```

2.
```
linContr.glm( contr.names=c("(Intercept)", "factor(agegrp)[60,71)", "wt", "male", "factor(agegr
```

```
##
## Test of H_0: 1*(Intercept) + 1*factor(agegrp)[60,71) + 85.543*wt + 1*male + 1*factor(agegrp)
```

```
##          Est    se.est     zStat  pVal    ci95.lo    ci95.hi
## 1 -0.5993364 0.1033585 -5.798616 1e-08 -0.8019154 -0.3967574
```

3.
```
exp(-0.599)/(1+exp(-0.599))
```

```
## [1] 0.3545725
```

```
exp(-0.802)/(1+exp(-0.802))
```

```
## [1] 0.3095979
```

```
exp(-0.397)/(1+exp(-0.397))
```

```
## [1] 0.4020333
```

```
linContr.glm( contr.names=c( "factor(agegrp)[40,50)", "factor(agegrp)[60,71)", "factor(agegrp)
```

```
##
## Test of H_0: exp( -1*factor(agegrp)[40,50) + 1*factor(agegrp)[60,71) + -1*factor(agegrp)[40
```

```
##   exp( Est )    se.est    zStat    pVal  ci95.lo  ci95.hi
## 1    2.39212 0.1750172 4.983398 6.2e-07 1.697494 3.370993
```

4.
```
linContr.glm( contr.names=c( "factor(agegrp)[40,50)", "factor(agegrp)[60,71)"), contr.coef=c(-1
```

```
##
## Test of H_0: exp( -1*factor(agegrp)[40,50) + 1*factor(agegrp)[60,71) ) = 1 :
```

```
##   exp( Est )   se.est    zStat pVal  ci95.lo  ci95.hi
## 1     3.9918 0.177032 7.819164    0 2.821492 5.647532
```

4e) Based on the logistic regression analysis, age and BMI were found to be significant predictors of hypertension. The odds of hypertension increased with increasing age and BMI. Additionally, males had higher odds of hypertension compared to females. The odds ratio comparing a 64 year old male to a 42 year old male was also significant, indicating that age was a strong predictor of hypertension. The odds ratio comparing a 64 year old female to a 42 year old female was not significant, indicating that age may not be as strong a predictor of hypertension for females. Overall, age and BMI are important factors to consider when assessing an individual's risk for hypertension.