Stats 111 HW3

Viraj Vijaywargiya

2023-02-15

```
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)
```

1. **1a)** I(Suburban) = 0, I(Rural) = 1, Age = 35. Putting this into the model,
   log(p/1-p) = B0 + B2 + B3(35). Therefore, the probability that a Rural person who is 35 years old will vote Republican is p = exp(B0 + B2 + B3(35)) / (1 + exp(B0 + B2 + B3(35))).
   Odds = p/(1-p). Therefore, from the expression for p we get odds = exp(B0 + B2 + B3(35)).

   1b) Odds ratio for Republican comparing Rural to Urban: odds(Rural) / odds(Urban) = exp(B2).
   Odds ratio for Republican comparing Rural to Suburban: odds(Rural) / odds(Suburban) = exp(B2 + B1).

   1c) The coefficient of Age, denoted as B3, represents the change in the log-odds of being Republican associated with a one-unit increase in Age, holding all other covariates constant.

   Since we have log(p/1-p) = B0 + B1 I(Suburban) + B2 I(Rural) + B3 Age, we can exponentiate both sides to obtain: p/1-p = exp(B0 + B1 I(Suburban) + B2 I(Rural) + B3 Age).

   This indicates that the odds of being Republican increase by a factor of exp(B3) for each one-unit increase in Age, holding all other covariates constant.

   To find the odds ratio for Republican comparing two people who differ in age by 20 years, we can use the equation for the odds that we derived earlier and plug in Age + 20 for one person and Age for the other person: odds(Age + 20) / odds(Age) = exp(B3(20)).

   This expression represents the ratio of the odds for someone who is 20 years older than the reference age (Age) to the odds for the reference age, holding all other covariates constant. Thus, the odds ratio for Republican comparing two people who differ in age by 20 years is exp(B3(20)).

   1d) The odds ratio for Republican comparing Rural to Urban for someone who is 35 years old can be obtained by comparing the log odds of the two groups: log(pRural/1-pRural) - log(pUrban/1-pUrban) = (B0 + B2 + B3(35) + B5(35)) - (B0 + B3*35) = B2 + B5*35.

So the odds ratio for Republican comparing Rural to Urban for someone who is 35 years old is exp(B2 + B5*35).

The odds ratio for Republican comparing Rural to Suburban for someone who is 35 years old can be obtained by comparing the log odds of the two groups: log(pRural/1-pRural) - log(pSuburban/1-pSuburban) = (B0 + B2 + B3(35) + B5(35)) - (B0 + B1 + B3(35) + B4(35)) = B2 - B1 + (B5 - B4)*35.

So the odds ratio for Republican comparing Rural to Suburban for someone who is 35 years old is exp(B2 - B1 + (B5 - B4)*35).

1e) The coefficient of Age (B3) in the model reflects the linear effect of age on the log odds of being a Republican, holding all other covariates constant. A positive B3 means that, as age increases, the log odds of being a Republican increase, and a negative B3 means that the log odds decrease as age increases.
To interpret the effect of age for each group separately, we can calculate the predicted log odds of being a Republican for a hypothetical individual at different ages, holding other covariates constant. For example, for urban residents, the predicted log odds of being a Republican for an individual at age A is: log(p_urban(A)/1-p_urban(A)) = B0 + B3*A.

For suburban residents, the predicted log odds of being a Republican for an individual at age A is: log(p_suburban(A)/1-p_suburban(A)) = B0 + B1 + B3(A) + B4(A).

For rural residents, the predicted log odds of being a Republican for an individual at age A is: log(p_rural(A)/1-p_rural(A)) = B0 + B2 + B3(A) + B5(A).

From these equations, we can see that the effect of age on the log odds of being a Republican differs across the three groups, as the coefficients of the interaction terms (B4 and B5) are not equal to zero. This suggests that the effect of age on the probability of being a Republican depends on the region of residence.

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","lot","

mod = glm(ac~pool, family=binomial(link="logit"), data=midwest)
summary(mod)
```

```
##
## Call:
## glm(formula = ac ~ pool, family = binomial(link = "logit"), data = midwest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6771   0.6281   0.6281   0.6281   0.6281
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.5231     0.1183   12.87   <2e-16 ***
## pool          2.0323     1.0211    1.99   0.0465 *
## ---
## 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: 465.87  on 520  degrees of freedom
```

```
## AIC: 469.87
##
## Number of Fisher Scoring iterations: 6
```

Estimated model: P(Yi = 1) = Bˆ(0) + Bˆ(1) Xi(1) = 1.5231 + 2.0323pool(i).
The coefficient on pool in the logistic regression model is 2.0323. This means that when all other variables in the model are held constant, a house with a pool is 2.0323 times more likely to have air conditioning than a house without a pool. In other words, the odds of a house having air conditioning are exp(2.0323) times higher when the house has a pool compared to when it does not have a pool. For example, if the odds of a house with no pool having air conditioning are 1 in 10, then the odds of a house with a pool having air conditioning are exp(2.0323) times higher, or approximately 7 times higher (since exp(2.0323) is approximately 7).

2b) Null hypothesis H0: B1 = 0 (there is no significant effect of having a pool on the probability of a house having air conditioning). Alternate hypothesis Ha: B1 != 0 (there is a significant effect of having a pool on the probability of a house having air conditioning).
Test statistic = 1.99 and p-value = 0.0465. Since the p-value (0.0465) is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is evidence to suggest that having a pool has a significant effect on the probability of a house having air conditioning.

2c)

```
mod2 = glm(ac ~ pool + sqft + pool*sqft, family=binomial(link="logit"), data=midwest)
summary(mod2)
```

```
##
## Call:
## glm(formula = ac ~ pool + sqft + pool * sqft, family = binomial(link = "logit"),
##     data = midwest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1701   0.2188   0.4398   0.7038   1.1105
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.8124228  0.5611205  -3.230  0.00124 **
## pool         5.8968034  3.5734828   1.650  0.09891 .
## sqft         0.0016459  0.0002913   5.651 1.59e-08 ***
## pool:sqft   -0.0018386  0.0012385  -1.485  0.13767
## ---
## 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: 418.49  on 518  degrees of freedom
## AIC: 426.49
##
## Number of Fisher Scoring iterations: 6
```

Estimated model: P(Yi = 1) = Bˆ(0) + Bˆ(1) Xi(1) + Bˆ(2) Xi(2) + Bˆ(3) Xi(1)Xi(2) = -1.8124228 + 5.8968034 pool(i) + 0.0016459 sqft(i) - 0.0018386 pool(i) sqft(i).

2d) If the house has no pool (i.e., pool = 0), then the effect of a 1 sqft increase in square footage on the odds of having air conditioning is given by the coefficient of sqft, which is 0.0016459. This means that a 1 sqft increase in square footage is associated with an increase of exp(0.0016459) = 1.001648 in the odds of having air conditioning, when the house has no pool.

If the house has a pool (i.e., pool = 1), then the effect of a 1 sqft increase in square footage on the odds of having air conditioning is given by the sum of the coefficients of sqft and the interaction term pool*sqft, which is 0.0016459 - 0.0018386 = -0.0001927. This means that a 1 sqft increase in square footage is associated with a decrease of exp(-0.0001927) = 0.9998074 in the odds of having air conditioning, when the house has a pool.

For a house with no pool, a 500 sqft increase in square footage is associated with an increase of exp(0.0016459 * 500) = 1.825932 in the odds of having air conditioning. For a house with a pool, a 500 sqft increase in square footage is associated with a decrease of exp(-0.0001927 * 500) = 0.904242 in the odds of having air conditioning.

3. **3a)**

```
wcgs = read.csv("/Users/virajvijaywargiya/Downloads/wcgs.csv", header=TRUE)

m9 = glm(chd~as.factor(smoke), family=binomial(link="logit"), data=wcgs)
summary(m9)
```

```
##
## Call:
## glm(formula = chd ~ as.factor(smoke), family = binomial(link = "logit"),
##     data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5355  -0.4265  -0.3497  -0.3497   2.3769
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.7636     0.1042 -26.535  < 2e-16 ***
## as.factor(smoke)1   0.4122     0.1628   2.533   0.0113 *
## as.factor(smoke)2   0.8035     0.1835   4.379 1.19e-05 ***
## as.factor(smoke)3   0.8938     0.2011   4.445 8.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1781.2  on 3153  degrees of freedom
## Residual deviance: 1751.7  on 3150  degrees of freedom
## AIC: 1759.7
##
## Number of Fisher Scoring iterations: 5
```

Estimated model: log(odds of CHD) = -2.7636 + 0.4122 * I(smoke(i) = 1) + 0.8035 * I(smoke(i) = 2) + 0.8938 * I(smoke(i) = 3)

3b) The estimated effect of going from the non-smoker group 0 to group 2 (21-30 cigs/day) is an increase in log-odds of CHD by 0.8035 units, holding other variables constant. This corresponds to an

estimated increase in the odds of CHD by exp(0.8035) = 2.23, or about 123%, for a person who goes from not smoking to smoking 21-30 cigarettes per day.

log(odds of CHD) for group 3 - log(odds of CHD) for group 1 = 0.8938 - 0.4122 = 0.4816. Therefore, the estimated effect of going from group 1 (1-20 cigs/day) to group 3 (31+ cigs/day) is an increase in log-odds of CHD by 0.4816 units, holding other variables constant. This corresponds to an estimated increase in the odds of CHD by exp(0.4816) = 1.62, or about 62%, for a person who goes from smoking 1-20 cigarettes per day to smoking 31 or more cigarettes per day.

3c)

```
m10 = glm(chd~as.factor(smoke)+bp+bp*as.factor(smoke), family=binomial(link="logit"), data=wcgs)
summary(m10)
```

```
##
## Call:
## glm(formula = chd ~ as.factor(smoke) + bp + bp * as.factor(smoke),
##     family = binomial(link = "logit"), data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7585  -0.4637  -0.3602  -0.3114   2.4701
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.0023     0.1312 -22.889  < 2e-16 ***
## as.factor(smoke)1      0.2992     0.2095   1.428 0.153154
## as.factor(smoke)2      1.0599     0.2140   4.953  7.3e-07 ***
## as.factor(smoke)3      0.7199     0.2689   2.677 0.007425 **
## bp                     0.8263     0.2175   3.799 0.000145 ***
## as.factor(smoke)1:bp   0.3507     0.3385   1.036 0.300133
## as.factor(smoke)2:bp  -0.9121     0.4348  -2.098 0.035922 *
## as.factor(smoke)3:bp   0.3575     0.4154   0.861 0.389501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1781.2  on 3153  degrees of freedom
## Residual deviance: 1708.3  on 3146  degrees of freedom
## AIC: 1724.3
##
## Number of Fisher Scoring iterations: 5
```

Estimated model: log(odds of CHD) = -3.0023 + 0.2992*I(smoke=1) + 1.0599*I(smoke=2) + 0.7199*I(smoke=3) + 0.8263*bp + 0.3507*I(smoke=1)bp - 0.9121*I(smoke=2)bp + 0.3575*I(smoke=3)bp.

3d) Predicted/estimated probability of chd for some with high blood pressure (bp=1) who has smoke=3 (31+ cigs/day): exp(-3.0023+0.7199+0.8263+0.3575) / (1 + exp(-3.0023+0.7199+0.8263+0.3575)) = 0.250.

Estimated probability for someone with normal blood pressure (bp = 0) who has smoke = 3: exp(-3.0023+0.7199) / (1 + exp(-3.0023+0.7199)) = 0.0926.

3e) For individuals with bp=0, the estimated log-odds of CHD for those who smoke 31+ cigs/day is 0.7199. Thus, compared to non-smokers (smoke=0), the estimated log-odds of CHD for individuals who smoke 31+ cigs/day and have normal blood pressure (bp=0) is increased by 0.7199.

For individuals with bp=1, the estimated log-odds of CHD for those who smoke 31+ cigs/day is (0.7199+0.3575) = 1.0774. Thus, compared to non-smokers (smoke=0), the estimated log-odds of CHD for individuals who smoke 31+ cigs/day and have high blood pressure (bp=1) is increased by 1.0774.

Therefore, we can conclude that the estimated effect of going from nonsmoker (smoke=0) to 31+ cigs/day (smoke=3) on the log-odds of CHD depends on the level of blood pressure. For individuals with normal blood pressure (bp=0), the estimated effect is an increase in log-odds of CHD by 0.7199, while for individuals with high blood pressure (bp=1), the estimated effect is an increase in log-odds of CHD by 1.0774.

3f) By fitting the model with smoke as a quantitative variable with values 0, 1, 2, 3, the assumption being made is that the effect of smoking on CHD risk is linear, which implies that the log odds of CHD increase linearly with the number of cigarettes smoked. This assumption also restricts the effect of smoke on CHD to be the same across all levels of smoke, meaning that the effect of increasing smoking by one unit (e.g., from 0 to 1, or from 1 to 2) has the same effect on the log odds of CHD, regardless of the starting level of smoking.

4. In a Generalized Linear Model (GLM), the link function is used to relate the expected value of the response variable to the linear predictor. The purpose of the link function is to transform the output of a linear equation into the range of the response variable. This is necessary because the response variable may have a non-linear relationship with the predictor variables, which cannot be captured by a linear model.

   The identity link function is a link function that simply relates the expected value of the response variable to the linear predictor without any transformation. It is often used with continuous variables or count data in Poisson regression. However, it is not often used with the binomial/Bernoulli parameter, p, because the range of p is restricted between 0 and 1, while the range of the linear predictor is unbounded. Using the identity link function in this case may result in predicted values that fall outside the valid range of the response variable, which can lead to invalid inference and poor model performance. Instead, the logit or probit link functions are commonly used in binomial/Bernoulli regression, as they transform the linear predictor into the range of probabilities (i.e., between 0 and 1).