Stats 111 Homework 3

Owen Lin

1. a. The probability that a Rural person that is 35 years old will vote Republican is $\frac{e^{\beta_0+\beta_2+35*\beta_3}}{(1+e^{\beta_0+\beta_2+35*\beta_3})}$. The odds is $e^{\beta_0+\beta_2+35*\beta_3}$.
   b. Odds ratio for Republican comparing Rural to Urban is $\frac{e^{\beta_0+\beta_2}}{e^{\beta_0}} = e^{\beta_2}$
   c. Holding all other covariates constant, the odds ratio for Republican comparing two people who differ in age by 20 years leads to $e^{20*\beta_3}$ times higher estimated odds of being Republican.
   d. Odds ratio for Republican comparing Rural to Urban for someone who is 35 years old is $\frac{e^{\beta_0+\beta_2+35*\beta_5}}{e^{\beta_0}} = e^{\beta_2+35*\beta_5}$
   The odds ratio for Republican comparing Rural to Suburban for someone who is 35 years old is $\frac{e^{\beta_2+35*\beta_5}}{e^{\beta_1+35*\beta_4}}$. For a urban person, one unit increase in Age leads to e^{β_3} times higher estimated odds of being Republican.
   For a suburban person, one unit increase in Age leads to $e^{\beta_3+\beta_4}$ times higher estimated odds of being Republican.
   For a rural person, one unit increase in Age leads to $e^{\beta_3+\beta_5}$ times higher estimated odds of being Republican.

2. a. $\ln \frac{p}{1-p} = 1.52 + 2.03 * Pool$
Having a Pool leads to $e^{2.03}$ times higher estimated odds of the house has air-conditioning.

```
MidwestSales = read.table("D:\\Coding\\Stats111\\Data\\MidwestSales.txt", fill=TRUE, header=FALSE)
names(MidwestSales)=c("id","price","sqft","bed","bath","ac","garage","pool","year","quality","style","lot","hwy")
lr_2a = glm(ac~pool, family=binomial(link="logit"), data=MidwestSales)
summary(lr_2a)
```

```
##
## Call:
## glm(formula = ac ~ pool, family = binomial(link = "logit"), data = MidwestSales)
##
## 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
```

b. $H_0 : \beta_1 = 0$
   $H_a : \beta_1 \neq 0$
   test statistic: 1.99 ~ N(0,1)
   p-value: 0.0465
   conclusion: We reject null on a 0.05 significant level and conclude that $\beta_1 \neq 0$

c. $\ln \frac{p}{1-p} = -1.81 + 5.90 * Pool + 0.0016 * sqft - 0.0018 * Pool * sqft$

```
lr_2c = glm(ac~pool+sqft+pool*sqft, family=binomial(link="logit"), data=MidwestSales)
summary(lr_2c)
```

```
##
## Call:
## glm(formula = ac ~ pool + sqft + pool * sqft, family = binomial(link = "logit"),
##     data = MidwestSales)
##
## 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
```

d. For a house with pool, 1 sqft increase leads to $e^{0.0016-0.0018}$ times the estimated odds of having ac.
For a house with pool, 500 sqft increase leads to $e^{500*(0.0016-0.0018)}$ times the estimated odds of having ac.
For a house without pool, 1 sqft increase leads to $e^{0.0016}$ times the estimated odds of having ac.
For a house without pool, 500 sqft increase leads to $e^{500*0.0016}$ times the estimated odds of having ac.

3. a. $\ln \frac{p}{1-p} = -2.76 + 0.41 * Smoke1 + 0.80 * Smoke2 + 0.89 * Smoke3$

```
wcgs = read.csv("D:\\Coding\\Stats111\\Data\\wcgs.csv", fill=TRUE, header = T)
lr_3a = glm(chd~as.factor(smoke), family=binomial(link="logit"), data=wcgs)
summary(lr_3a)
```

```
##
## 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
```

b. Comparing group 2 (21-30 cigs/day) to non-smoker 0 group, the odds ratio for obtaining CHD is estimated to be $e^{0.8035}$.
Comparing group 3 (31+ cigs/day) to group 1 (1-20 cigs/day), the odds ratio for obtaining CHD is estimated to be $e^{0.8935-0.4122}$.

c. $\ln \frac{p}{1-p} = -3.00 + 0.30 * Smoke1 + 1.06 * Smoke2 + 0.72 * Smoke3 + 0.83 * bp + 0.35 * Smoke1 * bp - 0.91 * Smoke2 * bp + 0.36 * S$

```
lr_3c = glm(chd~as.factor(smoke)+bp+bp*as.factor(smoke), family=binomial(link="logit"), data=wcgs)
summary(lr_3c)
```

```
##
## 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
```

d. The predicted probability of chd for some with high blood pressure (bp=1) who has smoke=3 (31+ cigs/day) is $\frac{e^{-3+0.72+0.83+0.36}}{(1+e^{-3+0.72+0.83+0.36})}$.

   The estimated probability for someone with normal blood pressure (bp=0) who has smoke=3 is $\frac{e^{-3+0.72}}{(1+e^{-3+0.72})}$

e. For high blood pressure (bp = 1), comparing smoke = 3 group to non-smoker 0 group, the odds ratio for obtaining CHD is estimated to be $e^{0.72+0.36}$.

   For low blood pressure (bp = 0), comparing smoke = 3 group to non-smoker 0 group, the odds ratio for obtaining CHD is estimated to be $e^{0.72}$.

f. If the model was fit with smoke being a quantitative variable with values 0,1,2,3, then the model would assume for each unit increase in smoke level, the effect is the same on the probabilities of CHD.

4. The purpose of link function is to transform the mean response such that it makes more sense than a simple linear model. For example, identity link function is the response itself: g(y) = y. It is not often used with the binomial/Bernoulli parameter p because the probability is between 0 and 1, but the idenity link function propose no restriction on the estimated outcome.