Jiayuan Shi
MA684 HW9

1A. What null hypothesis is tested by the model chi-square? What are the degrees of freedom for this model chi-square? What is the p-value corresponding to this model chi-square?

Solution:

- The null hypothesis is none of the variables among age, sex, and political party in the model is significantly associated with the probability of voting.
- The degree of freedom for this model chi-square is 4.
- The p-value corresponding to this model chi-square is less than 0.01. We can reject the null hypothesis, and say that at least one of the variables in the model is significantly associated with the probability of a voting.

1B. The C-statistic (also called the AUC) for this model is 0.64. Give an interpretation for this C-statistic.

Solution: The C-statistic for this model is 0.64, which indicates a weaker predictive model. It is the probability that a registered voter who truly voted will have a higher predicted probability from the logistic regression equation than a registered voter who truly does not vote.

1C. What null hypothesis is being tested by the Hosmer-Lemeshow test? What can you conclude from the Hosmer-Lemeshow chi-square value and from the p-value from this test?

Solution: The null hypothesis is that the model is the model is appropriate (fits the data), and so accepting the null supports the goodness of the model.

The chi-square value is 9.33 with 8 df, and p-value is 0.3156, larger than 0.05, so we fail to reject the null hypothesis for the Hosmer-Lemeshow test. Hence, the model fits the data.

1D. From the full and reduced models, calculate the partial chi-square testing whether political party is associated with voting, controlling for age and sex. What are the degrees freedom for this chi-square? What is the p-value for this chi-square? What is the chance in the C-statistic due to adding political party to the model?

Solution: The partial chi-square testing whether political party is associated with voting, controlling for age and sex is 17.39.

Partial chi-square = model chi-square(full) – model chi-square(reduced) = 24.91-7.52 = 17.39

The degree freedom for this chi-square is 2. The p-value for this chi-square is less than 0.01.

The change in the C-statistic due to adding political party to the model is 0.64-0.58, which is 0.06. This indicates a better predictive model.

2A. What percent of women in this sample have high blood pressure?

```
> table(hypertension)
hypertension
  0   1
```

533 147

Solution:
Percent of women have high blood pressure = 147/(147+533) = 0.2162 = 21.62%.
About 21.62% of women in this sample have high blood pressure.

2B. Give a summary table presenting results from this logistic regression. The table should include columns for odds ratios, confidence intervals for the odds ratios, and p-values.

```
> log.out <- glm(hypertension ~ age + smoke + relevel(factor(race),ref='1'),
+ family=binomial(link=logit))
> summary(log.out)

Call:
glm(formula = hypertension ~ age + smoke + relevel(factor(race),
    ref = "1"), family = binomial(link = logit))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.1704  -0.7539  -0.5374  -0.4513   2.1851

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.578289   0.487966  -3.234  0.00122 **
age                             -0.005791   0.009726  -0.595  0.55155
smoke                            1.057311   0.193989   5.450 5.03e-08 ***
relevel(factor(race), ref = "1")2  0.689845   0.241574   2.856  0.00430 **
relevel(factor(race), ref = "1")3 -0.370916   0.344169  -1.078  0.28116
relevel(factor(race), ref = "1")4 -0.164278   0.347479  -0.473  0.63638
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 709.96  on 679  degrees of freedom
Residual deviance: 669.63  on 674  degrees of freedom
AIC: 681.63

> exp(coef(log.out))
                  (Intercept)                           age                         smoke
                    0.2063278                     0.9942257                     2.8786197
relevel(factor(race), ref = "1")2 relevel(factor(race), ref = "1")3 relevel(factor(race),
ref = "1")4
                    1.9934060                     0.6901021                     0.8485065 >
exp(confint(log.out))
Waiting for profiling to be done...
                                 2.5 %    97.5 %
(Intercept)                   0.07830377 0.5317231
age                          0.97540226 1.0133630
```

smoke 1.97321281 4.2256692
relevel(factor(race), ref = "1")2 1.23531107 3.1915894
relevel(factor(race), ref = "1")3 0.33722581 1.3135277
relevel(factor(race), ref = "1")4 0.41248113 1.6279665

Number of Fisher Scoring iterations: 4

Solution:

Variable	Odds Ratio	p-value	95% CI
Age	0.994	0.552	(0.975, 1.013)
Smoking	2.879	<0.001	(1.973, 4.226)
Race			
White	Ref.	---	---
Black	1.993	0.004	(1.235, 3.192)
Hispanic	0.690	0.281	(0.337, 1.314)
Asian	0.849	0.636	(0.412, 1.628)

2C. Is the overall model significant? How well does this model predict hypertension - report and interpret the C-statistic for the logistic regression model.

> lroc(log.out)$auc
[1] 0.6656775

Solution:
The overall model is significant. The model chi-square = 709.96 – 669.63 = 40.33, with 5 degrees of freedom, p-value < 0.01. Therefore, at least one of the variables in the model is significantly associated with the probability of hypertension in women.

The C-statistic for the logistic regression model predicting hypertension in women from age, race, and smoking is 0.67, which indicates a weaker predictive model.

2D. From this multiple logistic regression model, find the odds ratio and confidence interval comparing the odds of hypertension for two subjects who differ in age by 10 years.
Solution:
Odds Ratios:
OR(age, 10 years) = exp(-0.005791(10)) = 0.944 ; controlling for race and smoking, an subject who differs in age by 10 years has 0.944 times the odds of the other subject.

Confidence Interval:
CI for the one-year odds ratio = (0.975, 1.013)
log(0.975) = -0.0253
log(1.013) = 0.0129
OR(age, 10 years) = exp(-0.0253 (10)) = 0.776
OR(age, 10 years) = exp(0.0129 (10)) = 1.138
CI for the ten-year odds ratio is (0.776, 1.138); controlling for race and smoking, confidence interval comparing the odds of hypertension for two subjects who differ in age by 10 years is (0.776, 1.138). The confidence interval contains 1, which indicates

no significance of the odds of hypertension for two subjects who differ in age by 10 years.

2E. Based on this model, describe the association between race and hypertension, including a statement about the overall significance of race in the model (based on a multiple-partial test for the 3 dummy variables for race) and a comparison across races based on the odds ratios relating to race.

```
> log.reduced <- glm(hypertension ~ age + smoke,
+ family=binomial(link=logit))
> summary(log.reduced)

Call:
glm(formula = hypertension ~ age + smoke, family = binomial(link = logit))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.9359  -0.8366  -0.5680  -0.5364   2.0120

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.356461   0.470966  -2.880  0.00397 **
age         -0.008217   0.009581  -0.858  0.39111
smoke        1.004348   0.190711   5.266 1.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 709.96  on 679  degrees of freedom
Residual deviance: 680.88  on 677  degrees of freedom
AIC: 686.88

Number of Fisher Scoring iterations: 4
```

Solution:
Full model: chi-square = 709.96-669.63 = 40.33
The model chi-square for this Full model is 40.33 with 5 df;
Reduced model: chi-square = 709.96-680.88 = 29.08
The model chi-square from the reduced model (see above) is 29.08 with 2 df.

The partial chi-square, testing the significance of adding the 3 dummy variables for race is,
Chi-square (3 df) = 40.33– 29.08 = 11.25, which gives a p-value 0.01. We can reject the null hypothesis of non-significantly association. Hence, race is significantly associated with hypertension in women.
From the odds ratios, we see that Blacks have 1.993 times the odds of hypertension than Whites, and the difference is significant; Hispanics and Asians have lower odds of hypertension than Whites, although the difference is not significant given the p-value (>0.05) and CIs (contains 1) for the odds ratios.

2F. Based on this model, describe the association between smoking and hypertension.

Solution: Smoking is significantly associated with hypertension in women, because in the logistic regression model predicting hypertension in women from age, race, and smoking, its p-values is less than 0.001. Also, 95% CI for OR of smoking is (1.973, 4.226), does not contain null value of 1.0, indicating hypertension in women significantly differs with smoking.

Controlling for age and race, smoking has 2.879 times the odds of hypertension than non-smokers.

3A. Based on the test for this interaction term, does the effect of smoking differ for older vs. younger women? Explain.

```
> log.int <- glm(hypertension ~ age + factor(race) + smoke + age*smoke,
+ family = binomial(link=logit))
> summary(log.int)

Call:
glm(formula = hypertension ~ age + factor(race) + smoke + age *
    smoke, family = binomial(link = logit))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.2251  -0.7856  -0.5376  -0.4154   2.3070

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.54740    0.68507  -0.799  0.42426
age           -0.02796    0.01452  -1.926  0.05415 .
factor(race)2  0.68207    0.24218   2.816  0.00486 **
factor(race)3 -0.39184    0.34639  -1.131  0.25796
factor(race)4 -0.17709    0.34813  -0.509  0.61096
smoke         -0.86467    0.93459  -0.925  0.35487
age:smoke      0.04111    0.01965   2.092  0.03647 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 709.96  on 679  degrees of freedom
Residual deviance: 665.21  on 673  degrees of freedom
AIC: 679.21

Number of Fisher Scoring iterations: 4
```

Solution: The p-value for the interaction variable (p=0.036) indicates significant interaction, so the effect of smoking significantly differs for older vs. younger women.

3B. Regardless of the significance of the interaction, use this model to find the odds ratio comparing the odds of hypertension for:

Solution:

In the interaction model, the effect of smoking is shared between two terms:
-0.8647 (smoking) + 0.0411 (smoking * age)

The odds ratio of hypertension for a 30 year old smoker vs. a 30 year old non-smoker is:
OR = exp(-0.8647 + 0.0411 * 30) = 1.4453
A 30 year old smoker has about 1.4 times the odds of hypertension than a 30 year old non-smoker in women.

The odds ratio of hypertension for a 50 year old smoker vs. a 50 year old non-smoker is:
OR = exp(-0.8647 + 0.0411 * 50) = 3.2881
A 50 year old smoker has about 3.3 times the odds of hypertension than a 50 year old non-smoker in women.

3C. I ran a Hosmer-Lemeshow test on this interaction model, placing subjects into 10 categories based on their predicted risk of hypertension. The Hosmer-Lemeshow chi-square statistic was 7.30 with 8 degrees of freedom. Find the p-value for this test statistic, and give an interpretation of the results of this test.

```
> 1-pchisq(7.30,8)
[1] 0.5046378
```

Solution: The p-value for this test statistic is 0.5046, larger than 0.05. We fail to reject the null hypothesis, which is that the model is appropriate (fits the data), and so accepting the null supports the goodness of the model. Hence, it shows non-significant evidence of a problem with the model (non-significant lack of fit).