**Biostatistics 140.654: Applied Generalized Linear Models**
**Fourth Term, 2015**

**Midterm**

**Instructions**: This **is a closed book quiz. Do not consult with any other person or any materials not on the attached pages in answering the questions below. Choose the best answer(s) for each question and circle its letter(s) clearly. Good luck.**

**Name**
**(print)**_____

**By signing my name, I agree to abide by the Johns Hopkins University School of Public Health Academic Code:**

**Signature**_____

Below find a logistic regression using the NMES expenditure and disease data with outcome variable *bigexp*=1 if expenditures exceed $1,000 per year; 0 otherwise and predictor variables: presence of a major smoking caused disease (*mscd*) and age (*agem65* = age – 65 in years; *age_sp65* = age-65 when age > 65 and 0 otherwise).

```
glm(formula = bigexp ~ mscd + agem65 + age_sp65, family = binomial,
    data = data1)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.579619   0.037966 -15.267   <2e-16 ***
mscd         1.591648   0.068527  23.226   <2e-16 ***
agem65       0.027661   0.002799   9.882   <2e-16 ***
age_sp65    -0.006414   0.005996  -1.070    0.285
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15263  on 11586  degrees of freedom
Residual deviance: 14165  on 11583  degrees of freedom
AIC: 14173

> vcov(lr.new1)
              (Intercept)          mscd        agem65      age_sp65
(Intercept)  1.441394e-03 -5.872843e-04  8.154566e-05 -1.833519e-04
mscd        -5.872843e-04  4.695994e-03 -2.221828e-05  9.835147e-06
agem65       8.154566e-05 -2.221828e-05  7.835390e-06 -1.372802e-05
age_sp65    -1.833519e-04  9.835147e-06 -1.372802e-05  3.595175e-05
```

(1). The probability of a big expenditure for a 65 year-old with mscd is estimated in this model to be (select single best answer)

(a). 0.36

(b). 0.64

(c). 0.74

(d). 0.98

(e). 0.999

(2). A 95% confidence interval for the probability of a big expenditure for a 65 year old with mscd is (show your work)

3). The ratio of odds of a big expenditure for a 75 year-old without major smoking-caused disease (mscd=0) as compared to a 65 year old without mscd is (select single best answer)
(a). 0.94
(b). 1.022
(c). 1.029
(d). 1.24
(e). 1.32

The model above is revised by dropping the two age variables. The results become

```
glm(formula = bigexp ~ mscd, family = binomial, data = data1)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.74267    0.02109  -35.22   <2e-16 ***
mscd         1.80551    0.06706   26.92   <2e-16 ***
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15263  on 11586  degrees of freedom
Residual deviance: 14416  on 11585  degrees of freedom
AIC: 14420
```

4). Using the two models above, the likelihood ratio test statistic and distribution for the null hypothesis that age does not improve the prediction of whether a person has a large expenditure is (select single best answer)
(a). 847, Chi-square with 1 df
(b). 1098, Chi-square with 3 df
(c). 15264, Chi-square with 1 df
(d). 251, Chi-square with 2 df
(e). can not say from the available information

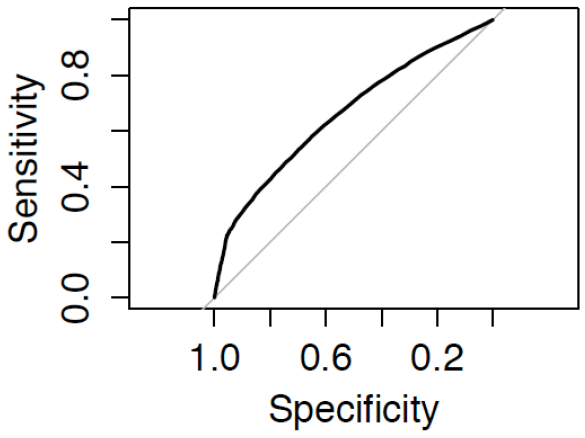
5). The coefficient for mscd in the two models above (select single best answer)
(a). are not the same indicating that age modifies the effect of mscd on the risk of big expenditure
(b). are not the same indicating that age confounds the effect of mscd on big expenditure
(c). are similar indicating that the there is some quantitative but not qualitative effect modification
(d). are similar indicating that the there is some quantitative but not qualitative confounding
(e). is statistically significant indicating that the model is correct

A third model was fit with an interaction between the major smoking caused disease indicator and the age variables. The results are provided below

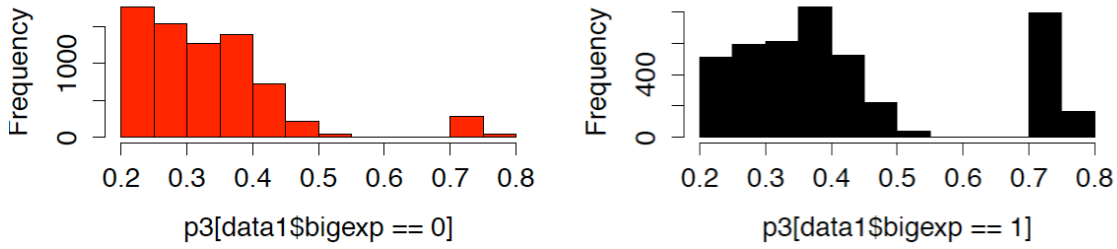
```
glm(formula = bigexp ~ mscd * (agem65 + age_sp65), family = binomial,
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7035  -0.9085  -0.7406   1.2911   1.7534

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.586684   0.039337 -14.914   <2e-16 ***
mscd           1.601041   0.116778  13.710   <2e-16 ***
agem65         0.028340   0.002901   9.771   <2e-16 ***
age_sp65      -0.003468   0.006377  -0.544    0.587
mscd:agem65   -0.032299   0.013406  -2.409    0.016 *
mscd:age_sp65  0.013275   0.021255   0.625    0.532
---

Null deviance: 15263  on 11586  degrees of freedom
Residual deviance: 14149  on 11581  degrees of freedom
AIC: 14161
```

6). In the space below, draw a figure of the log odds of being a big expenditure person as a function of age and mscd. Label the axes. Indicate on your graph each of the regression coefficients in the output above.

The third model (with interaction) was used to predict whether a person is likely to be a big spender based upon their age and mscd status. The ROC curve for this model and a plot of the predicted values for those without (bigexp=0) and with (bigexp=1) expenditures above $1,000 are shown below.




**Data: p3 in 7307 controls, 4280 cases**
**Area under the curve: 0.6625; 95% CI: 0.6521-0.6729**




7). At c=0.5, the sensitivity of the prediction based upon this model is roughly (select single best answer)
(a). 5%
(b). 25%
(c). 65%
(d). 95%
(e). can not tell from the data shown

8). At c=0.5, the specificity of the prediction method is roughly (select single best answer)
(a). 5%
(b). 25%
(c). 65%
(d). 95%
(e). can not tell from the data shown

9). The chance that a big spender (>$1,000) drawn at random will have a predicted value from model 3 that is less than that for a small spender (<$1,000) is approximately (select single best answer)
(a). 5%
(b). 25%
(c). 65%
(d). 95%
(e). can not tell from the data shown

10). In a sentence or two, explain how bootstrapping can be used to obtain a cross-validated estimate and confidence interval for the area under the ROC curve?