MATH 4330 — Assignment 3

Instructor: Kevin McGregor **Due Date:** Monday Nov. 7th, 2022, 11:59 PM

Instructions: Hand in all questions to Crowdmark by the deadline. For written questions, you may upload a scan or photo of your work; however, I encourage you to use LaTeX (though it's not a requirement). Include a screenshot of your R code for each question in your Crowdmark submission. **Show your work, and make sure that all work you submit is your own**.

Question 1: Hourly carbon monoxide (CO) averages were recorded on summer weekdays at a measurement station in Los Angeles. The station was established by the Environmental Protection Agency as part of a larger study to assess the effectiveness of the catalytic converter. It was located about 25 feet from the San Diego Freeway, which in this particular area is located at 145 degrees north. It was located such that winds from 145 to 325 degress (which in the summer are the prevalent wind directions during the daylight hours) transport the CO emissions from the highway toward the measurement station. Aggregate measurements were recorded for each hour of the day 1 to 24 and the dataset is available in the file `CO2.txt`. Note: you can load this file in R using `read.table()` and setting the argument `header=TRUE`.

Hour	-	hour of the day, from midnight to midnight
CO	-	average summer weekday CO concentration (parts per million)
TD	-	average weekday traffic density (traffic count/traffic speed)
WS	-	average perpendicular wind-speed component (wind speed $\times$ cos(wind direction - 235 degrees))

(a) **[5 points]** Run a linear regression model to examine the effect of weekday traffic density and wind-speed component on CO concentration. Report the estimated slope parameters and their confidence intervals.

(b) **[3 points]** Examine residual plots in the model from part (a). Do you think any of the linear regression assumptions have been violated? Explain.

(c) **[5 points]** Run a weighted least-squares model using the same outcome and predictors from part (a). Have the estimates and confidence intervals changed much? Explain.

Question 2: Recall the chest pain dataset from Assignment 2. This dataset is from the Duke University Cardiovascular Disease Databank and consists of 2258 patients and 6 variables. The patients were referred to Duke University Medical Center for chest pain. The variables included in the dataset `acath2.csv` are the following:

- `sex`: sex of the patient (0=male, 1=female)

- `age`: age of the patient
- `cad.dur`: duration of symptoms of coronary artery disease
- `cholest`: cholesterol (in mg)
- `sigdz`: significant coronary disease by cardiac catheterization (defined as $\geq$ 75% diameter narrowing in at least one important coronary artery - 1=yes, 0=no)
- `tvdlm`: severe coronary disease (defined as three vessel or left main disease by cardiac catheterization - 1=yes, 0=no))

(a) [**4 points**] Run a logistic regression model to see the effect of cholesterol (continuous measure) on significant coronary disease (`sigdz`). Report the odds ratio and interpret. Calculate and interpret a 95% confidence interval for the odds ratio.

(b) [**3 points**] Calculate the predicted probability of significant coronary disease for an individual with cholesterol equal to 400.

(c) [**1 points**] Do you think the expression in (b) can be used to accurately predict significant coronary disease in the general population? Explain.

(d) [**4 points**] Run another logistic regression model to see the effect of cholesterol on `sigdz`, but this time adjust for age and sex. Report the odds ratio. From this new model fit do you think there is evidence that age and sex are confounders in the cholesterol/coronary disease relationship? Explain. (Hint: it has nothing to do with significance of the predictor variables).

(e) [**5 points**] Create an ROC curve for each of the models in part (a) and (d). Report the AUC for each one. Which model has better predictive accuracy? Explain.

4330 Assignment 3

Ravish Kamath: 213893664

07 November, 2022

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Loading required package: carData
```

Question 1

Hourly carbon monoxide (CO) averages were recorded on summer week- days at a measurement station in Los Angles. The station was established by the Environmental Protection Agency as part of a larger study to assess the effectiveness of the catalytic converter. It was located about 25 feet from the San Diego Freeway, which in this particular area is located at 145 degrees north. It was located such that winds from 145 to 325 degrees (which in the summer are the prevalent wind directions during the daylight hours) transport the CO emissions from the highway toward the measurement station. Aggregate measurements were recorded for each hour of the day 1 to 24 and the data set is available in the file CO2.txt. Note: you can load this file in R using read.table() and setting the argument header = TRUE.

Hour - hour of the day, from midnight to midnight
CO - average summer weekday CO concentration (parts per million)
TD - average weekday traffic density (traffic count/traffic speed)
WS - average perpendicular wind-speed component (wind speed x cos(wind direction - 235 degrees))

(a) **[5 points]** Run a linear regression model to examine the effect of weekday traffic density and wind-speed component on CO concentration. Report the estimated slope parameters and their confidence intervals.
(b) **[3 points]** Examine residual plots in the model from part (a). Do you think any of the linear regression assumptions have been violated? Explain.
(c) **[5 points]** Run a weighted least-sqaures model using the same outcome and predictors from part(a). Have the estimates and confidence intervals changed much? Explain.

Solution

Part A

```
fit = lm(CO ~ Traffic + Wind, data = CO2df)
summary(fit)
```

```
##
## Call:
## lm(formula = CO ~ Traffic + Wind, data = CO2df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72858 -0.31710 -0.09629  0.22409  1.26554
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.274461   0.198137   6.432 2.25e-06 ***
## Traffic     0.018290   0.001343  13.616 6.85e-12 ***
## Wind        0.174747   0.056765   3.078   0.0057 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4987 on 21 degrees of freedom
## Multiple R-squared:  0.9495, Adjusted R-squared:  0.9447
## F-statistic: 197.5 on 2 and 21 DF,  p-value: 2.419e-14
```

```
confint(fit)
```

```
##                  2.5 %     97.5 %
## (Intercept) 0.86241251 1.68650968
## Traffic     0.01549692 0.02108392
## Wind        0.05669662 0.29279680
```

$\beta_0 = 1.274461$ and C.I is **(0.8624, 1.6865)**.

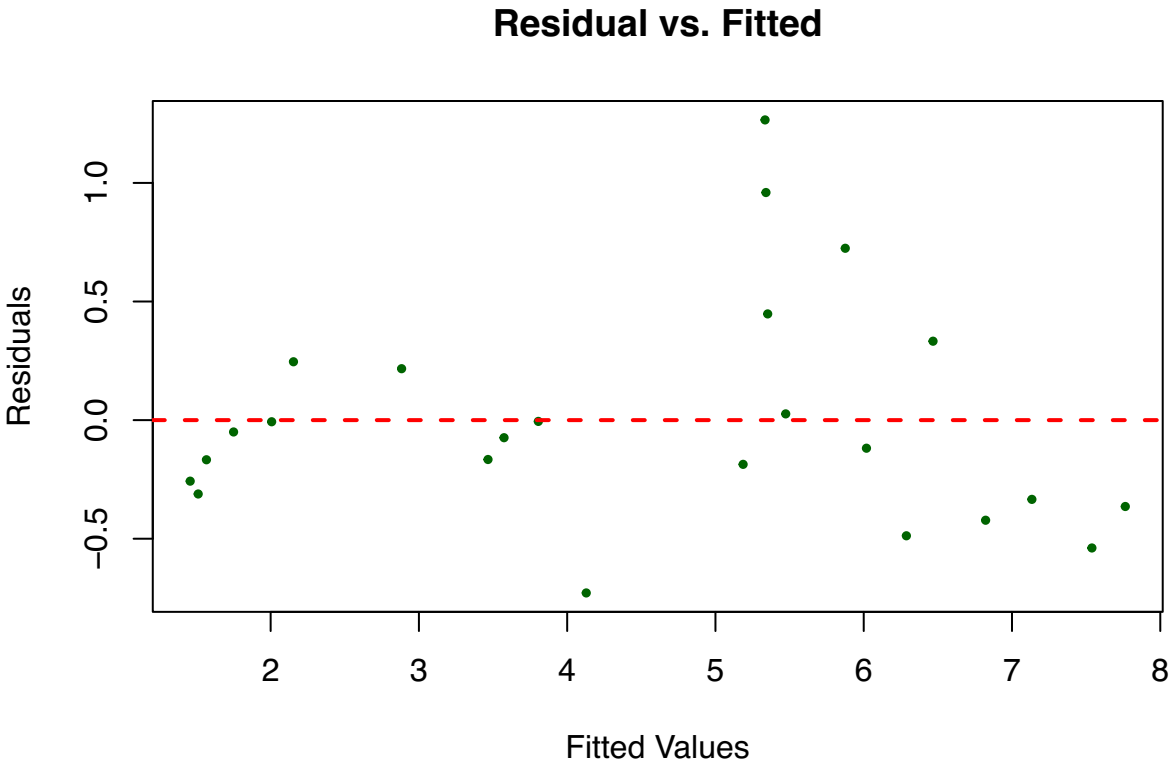
$\beta_1 = 0.018290$ and C.I. is **(0.0155, 0.0211)**

$\beta_2 = 0.174747$ and C.I. is **(0.0567, 0.2928)**

Part B

Residual vs. Fitted

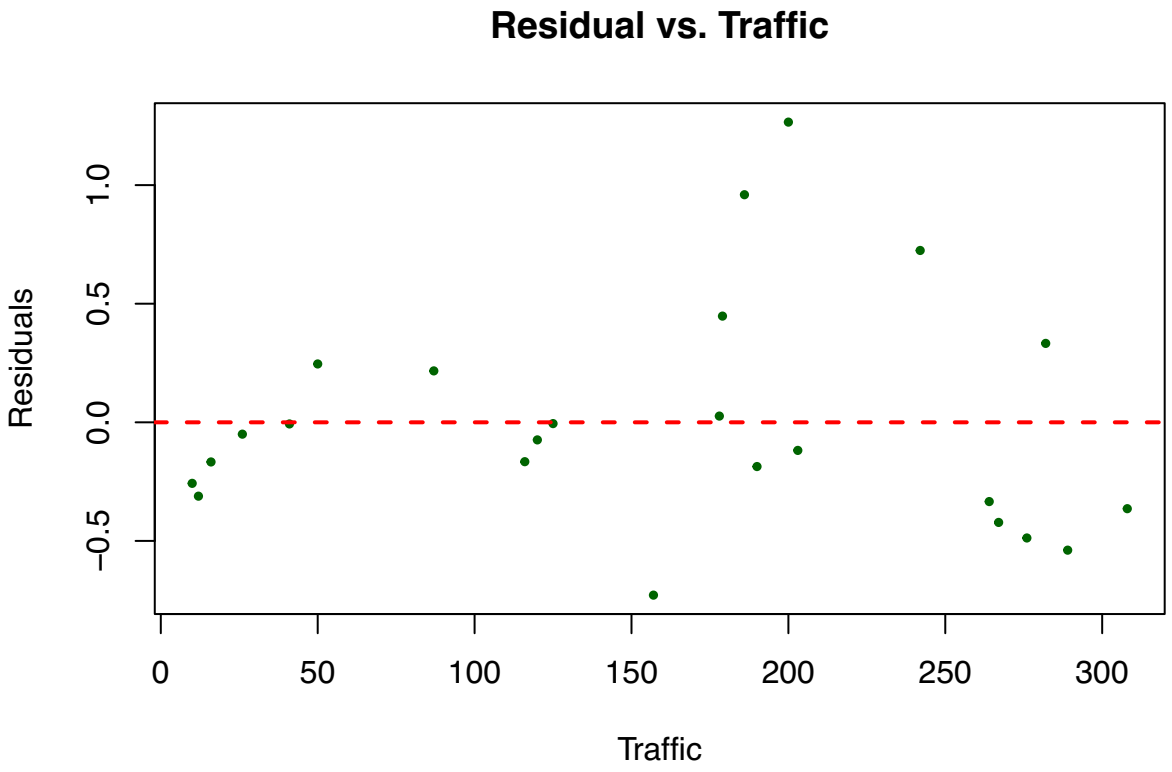
```
plot(fit$fitted.values, fit$residuals, pch=19, col="darkgreen", cex=0.5, xlab = 'Fitted Values',
     ylab = 'Residuals', main = 'Residual vs. Fitted')
abline(h=0, lty=2, lwd=2, col="red")
```



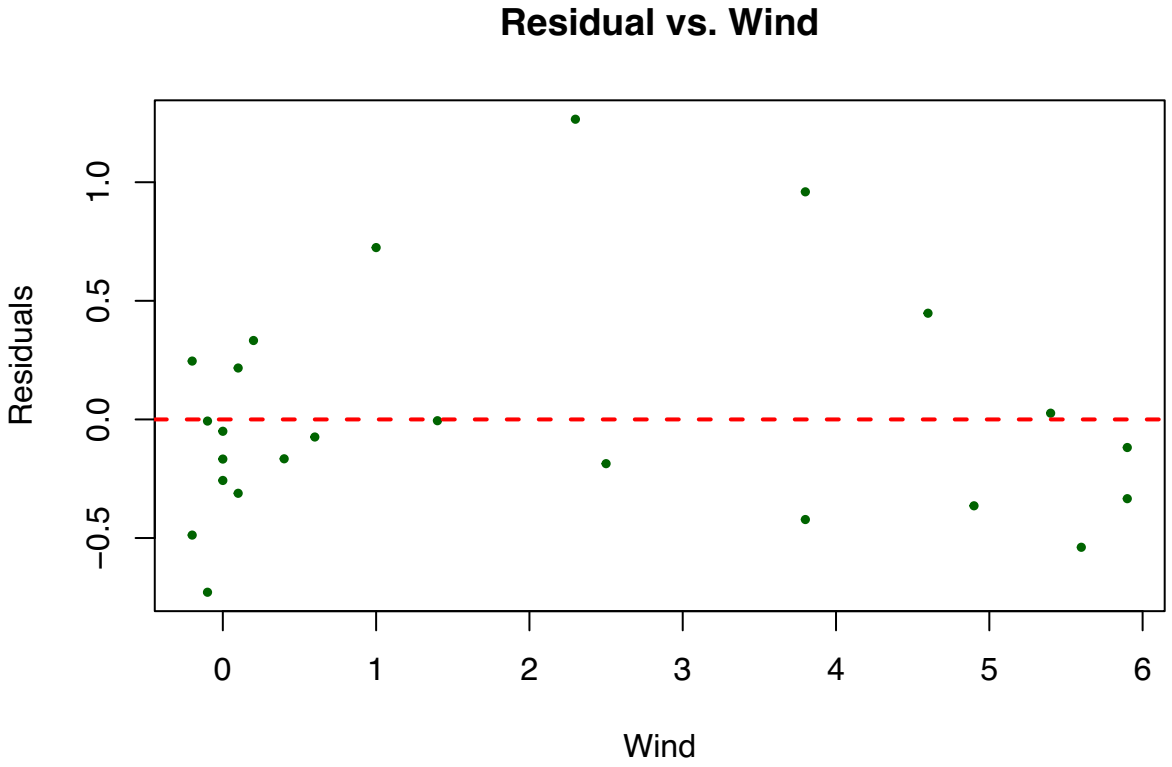
For the Residual vs. Fitted plot, I would say that there is no discernible change in the variation nor any pattern. Hence it has not violated the error variance is constant.

Residual vs. Predictors

```
plot(CO2df$Traffic, fit$residuals, pch=19, col="darkgreen", cex=0.5, xlab = 'Traffic',
     ylab = 'Residuals', main = 'Residual vs. Traffic')
abline(h=0, lty=2, lwd=2, col="red")
```



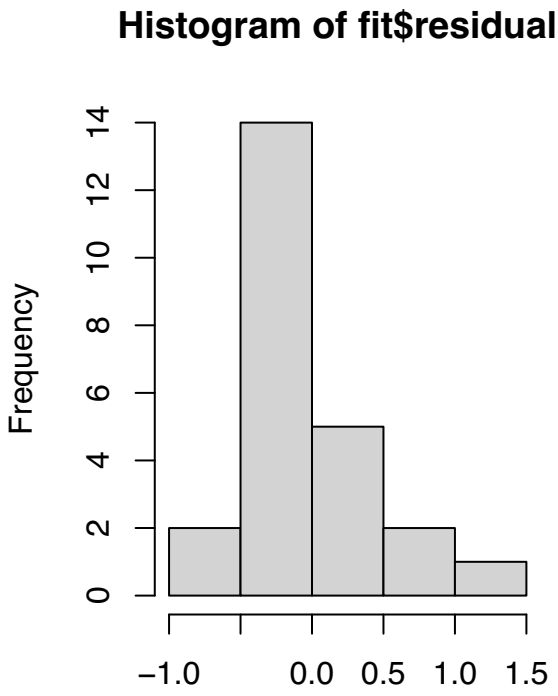
```
plot(CO2df$Wind, fit$residuals, pch=19, col="darkgreen", cex=0.5, xlab = 'Wind',
     ylab = 'Residuals', main = 'Residual vs. Wind')
abline(h=0, lty=2, lwd=2, col="red")
```

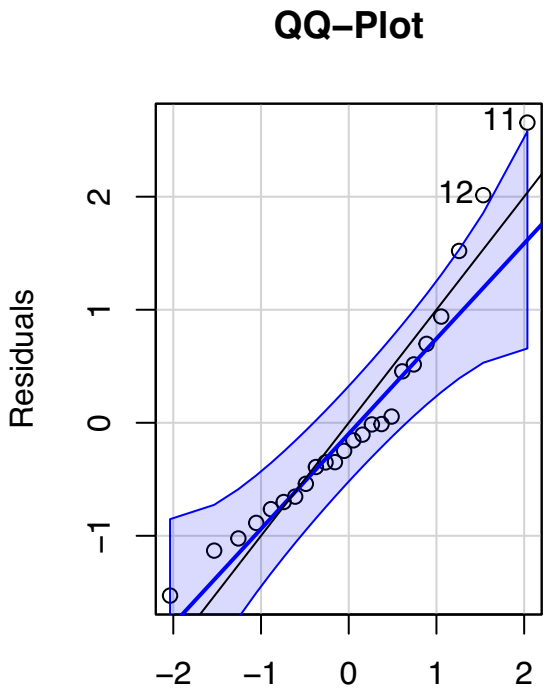


I would say for the first plot shows no clear pattern, hence there is no violation, however the second plot (Residual vs. Wind) does tend to show a parabola pattern. This may require squaring the wind variable.

```
#Normality Plot
par(mfrow = c(1,2))
hist(fit$residuals, xlab = 'Residuals')
qqPlot(scale(fit$residuals), main ='QQ-Plot', ylab = 'Residuals'); abline(0,1)
```

```
## [1] 11 12
```





Based of the histogram and qqplot, it does seem violate the assumption that it comes from a normal distribution.

Part C

```
res.fit <- lm(abs(fit$residuals) ~ fit$fitted.values)
w <- 1/(res.fit$fitted.values^2)
wls.fit <- lm(CO ~ Traffic + Wind, data=CO2df, weights=w)
summary(wls.fit)
```

```
##
## Call:
## lm(formula = CO ~ Traffic + Wind, data = CO2df, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1957 -0.9313 -0.2387  0.6485  3.0661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.153799   0.111087  10.386 9.93e-10 ***
## Traffic     0.019011   0.001232  15.429 6.24e-13 ***
## Wind        0.184738   0.062544   2.954  0.00758 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.309 on 21 degrees of freedom
## Multiple R-squared:  0.9672, Adjusted R-squared:  0.9641
## F-statistic: 309.6 on 2 and 21 DF,  p-value: 2.609e-16
```

```
confint(wls.fit)
```

```
##                  2.5 %     97.5 %
## (Intercept) 0.92278197 1.38481646
## Traffic     0.01644902 0.02157387
## Wind        0.05467042 0.31480494
```

No the estimates and confidence intervals have not changed much, which means that the residuals do follow a constant variance.

Question 2

Recall the chest pain data set from Assignment 2. This data set is from the Duke University Cardiovascular Disease Databank and consists of 2258 patients and 6 variables. The patients were referred to Duke University Medical Center for chest pain. The variables included in the data set acath2.csv are the following:

- sex: sex of the patient (0 = male, 1 = female)
- age: age of the patient
- cad.dur: duration of symptoms of coronary artery disease
- cholest: cholesterol (in mg)
- sigdz: significant coronary disease by cardiac catheterization (defined as $\geq 75\%$ | diameter narrowing in at least one important coronary artery - 1 = yes, 0 = no)
- tvdlm: severe coronary disease (defined as three vessel or left main disease by cardiac | catheterization - 1 = yes, 0 = no)

(a) **[4 points]** Run a logistic regression model to see the effect of cholesterol (continuous measure) on significant coronary disease (sigdz). Report the odds ratio and interpret. Calculate and interpret a 95% confidence interval for the odds ratio.

(b) **[3 points]** Calculate the predicted probability of significant coronary disease for an individual with cholesterol equal to 400.

(c) **[1 point]** Do you think the expression in (b) can be used to accurately predict significant coronary disease in the general population? Explain.

(d) **[4 points]** Run another logistic regression model to see the effect of cholesterol on sigdz, but this time adjust for age and sex. Report the odds ratio. From this new model fit do you think there is evidence that age and sex are confounders in the cholesterol/coronary disease relationship? Explain. (Hint: it has nothing to do with significance of the predictor variables).

(e) **[5 points]** Create an ROC curve for each of the models in part (a) and (d). Report the AUC for each one. Which model has better predictive accuracy? Explain.

Solution

Part A

```
fit = glm(sigdz ~ cholest, data=df, family=binomial)
summary(fit)
```

```
##
## Call:
## glm(formula = sigdz ~ cholest, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2140  -1.3669   0.8247   0.9406   1.4279
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.7525280  0.2186516  -3.442 0.000578 ***
## cholest      0.0062268  0.0009525   6.538 6.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2895.3  on 2257  degrees of freedom
## Residual deviance: 2849.7  on 2256  degrees of freedom
```

```
## AIC: 2853.7
##
## Number of Fisher Scoring iterations: 4
```

```
beta1_OR = exp(fit$coefficients[2])
```

Since our odds ratio is **1.006246**, the odds of having significant coronary disease by cardiac catheterization **increases by a factor of 1.006246** as your cholesterol increases by 1 unit.

```
CI = confint(fit, level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
new_CI = c(exp(CI[2,1]), exp(CI[2,2]))
new_CI
```

```
## [1] 1.004389 1.008147
```

Since the confidence interval **does not overlap 1**, then we can say **there is an association** between significant coronary disease and cholesterol.

Part B

```
n.dat <- data.frame(cholest = 400)
predict(fit, newdata = n.dat, type="response")
```

```
##         1
## 0.8504561
```

Part C

I would say that the expression in (b) cannot be used to accurately predict significant coronary disease for a general population because the data set used, were for patients that already had some prior chest pain. These are two different populations of interest, hence, it is not a good predictive model.

Part D

```
fit2 = glm(sigdz ~ cholest + age + sex, data = df, family = binomial)
summary(fit2)
```

```
##
## Call:
## glm(formula = sigdz ~ cholest + age + sex, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4867  -0.8699   0.5259   0.7691   2.4029
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.177898   0.380091 -10.992   <2e-16 ***
## cholest      0.009006   0.001076   8.367   <2e-16 ***
## age          0.069987   0.005832  12.001   <2e-16 ***
## sex         -2.094385   0.113306 -18.484   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2895.3  on 2257  degrees of freedom
## Residual deviance: 2347.4  on 2254  degrees of freedom
## AIC: 2355.4
##
## Number of Fisher Scoring iterations: 4
```

```
betas = as.vector(fit2$coefficients)
exp(betas)
```

```
## [1] 0.0153307 1.0090463 1.0724942 0.1231459
```

β_1 OR would be 1.0090463.

β_2 OR would be 1.0724942.

β_3 OR would be 0.1231459.

To identify whether Age and Sex are confounders for significant coronary diseases and cholesterol level, we should see a difference in our estimated coefficient for cholesterol. When comparing the two models, fit and fit2, we can see that the β_1 coefficient for cholesterol has changed, though not by much, shows that Age and Sex are a potential confounder to the model.

Part E

```
pred.val1 = predict(fit)
pred.val2 = predict(fit2)
par(mfrow = c(1,2))

auc(df$sigdz, pred.val1, plot=TRUE, auc.polygon=TRUE,
    auc.polygon.col="lightblue", asp=FALSE, main = 'Model with Cholesterol')
```

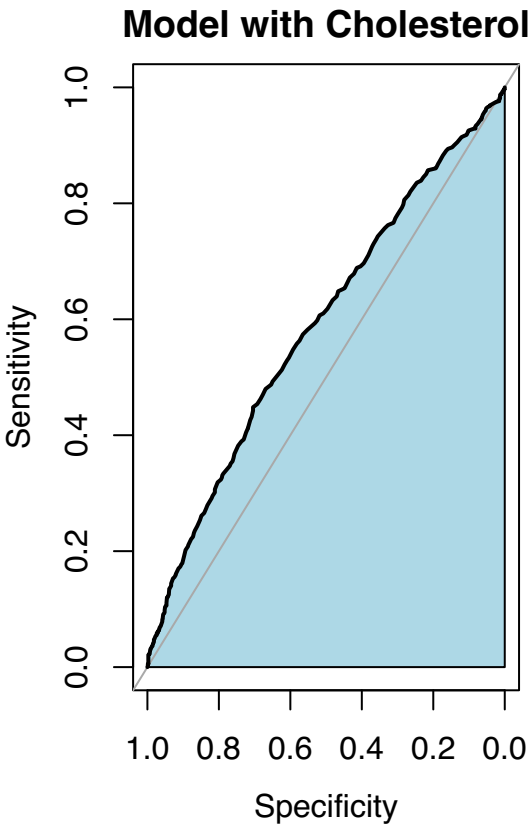
```
## Setting levels: control = 0, case = 1
```

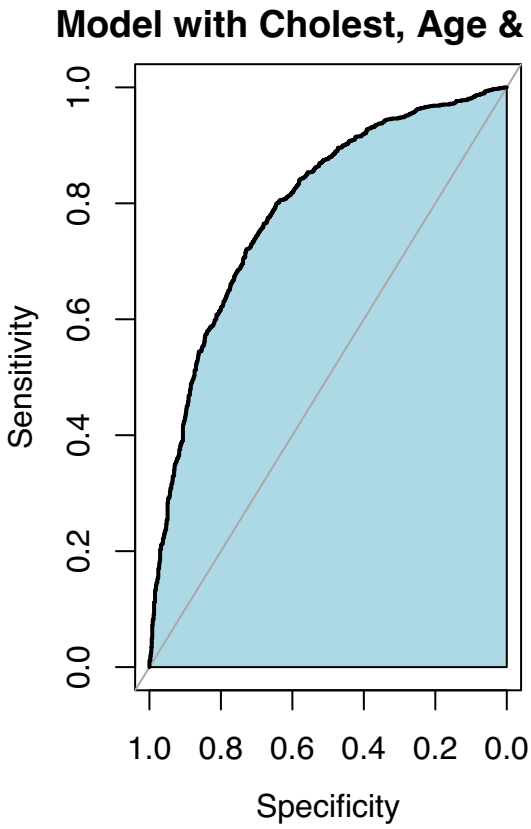
```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.5887
```

```
auc(df$sigdz, pred.val2, plot=TRUE, auc.polygon=TRUE,
    auc.polygon.col="lightblue", asp=FALSE, main = 'Model with Cholest, Age & Sex')
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```






```
## Area under the curve: 0.7887
```

I would say that the model with the added predictors, age and sex, would be the better model. If we see based of the ROC curves, the 2nd model to be closer to the top left corner of the plot. This signifies that we are getting higher value for sensitivity as well as specificity. Furthermore, the AUC for the second model is 0.7887 which is better than the previous model with only 0.5887.