Assignment 3 - SOLUTIONS

Instructor: Kevin McGregor

MATH 4330

Question 1:

(a)

After reading in the dataset, we run linear regression:

```
co.fit <- lm(CO~Traffic+Wind, data=co.data)
summary(co.fit)

##
## Call:
## lm(formula = CO ~ Traffic + Wind, data = co.data)
##
## 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(co.fit)

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

For the estimated slope parameters we have:

- $\beta_1 = 0.0183$
- $\beta_2 = 0.1747$

The associated confidence intervals are:

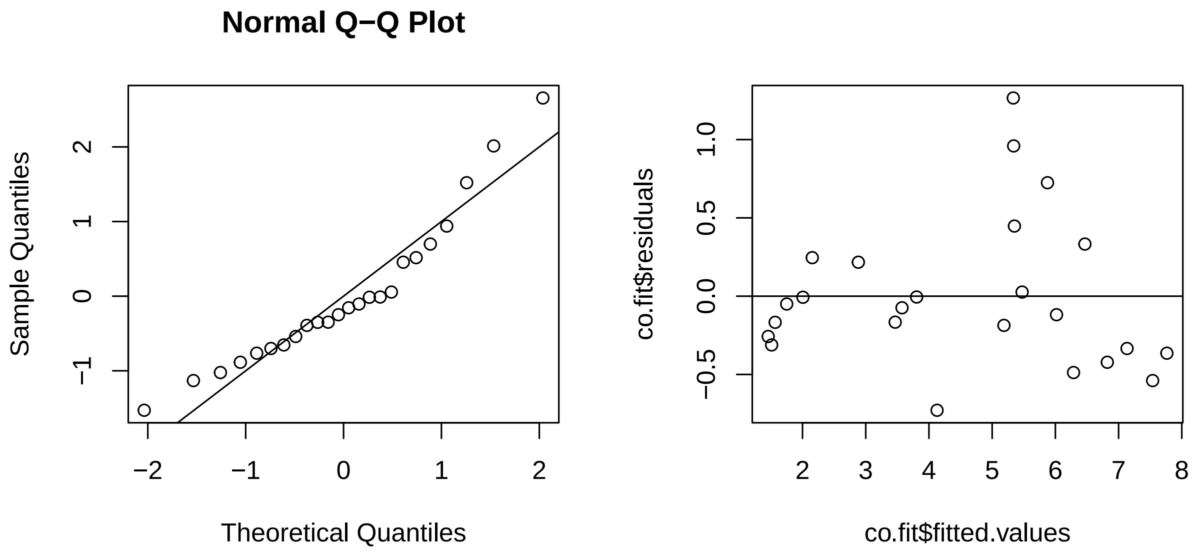
- $(0.0155, 0.0211)$
- $(0.0567, 0.2928)$

for β_1 and β_2, respectively.

(b)

Plotting the residuals, we have:

```
par(mfrow=c(1,2))
qqnorm(scale(co.fit$residuals))
abline(0,1)
plot(co.fit$fitted.values, co.fit$residuals)
abline(h=0)
```



Based on the residual plots, it is clear that the constant variance assumption has been violated; there is evidence of heteroskedasticity.

(c)

```
# WLS for Q1
res.fit <- lm(abs(co.fit$residuals) ~ co.fit$fitted.values)
# Weights
w <- 1/(res.fit$fitted.values^2)

# Running weighted least squares
wls.fit <- lm(CO~Traffic+Wind, data=co.data, weights=w)

summary(wls.fit)

##
## Call:
## lm(formula = CO ~ Traffic + Wind, data = co.data, 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
```

After running WLS, we can see that the slope estimates and confidence intervals have barely changed (though the CI for β_2 got a bit wider). Looks like the violation of the constant variance assumption did not matter too much in this case.

Question 2:

(a)

After loading in the data, we can run logistic regression as follows:

```
s.fit <- glm(sigdz~cholest, data=s.data, family = "binomial")
summary(s.fit)

##
## Call:
## glm(formula = sigdz ~ cholest, family = "binomial", data = s.data)
##
## 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
```

The odds ratio and associated CI are calculated as:

```
# OR
exp(s.fit$coefficients[2])

##  cholest
## 1.006246

# CI for OR
exp(confint(s.fit)[2,])

## Waiting for profiling to be done...

##    2.5 %   97.5 %
## 1.004389 1.008147
```

Interpretation: when cholesterol increases by one unit, the odds of significant coronary disease increases by a factor of 1.0062. The confidence interval for the OR is $(1.0043, 1.0081)$, meaning that we have evidence that cholesterol is positively associated with significant coronary disease (since the CI does not overlap with 1).

(b)

```
x.new <- list(cholest=400)
predict(s.fit, newdata=x.new, type="response")

##         1
## 0.8504561
```

The predicted probability for an individual with cholesterol equal to 400 is 0.8505.

(c)

No, the prediction would not apply to someone in the general population since this study was for people presenting with chest pain so this is likely not a representative sample of the general population.

(d)

Including age and sex as predictors is how we adjust for these two variables. The adjusted model is shown below:

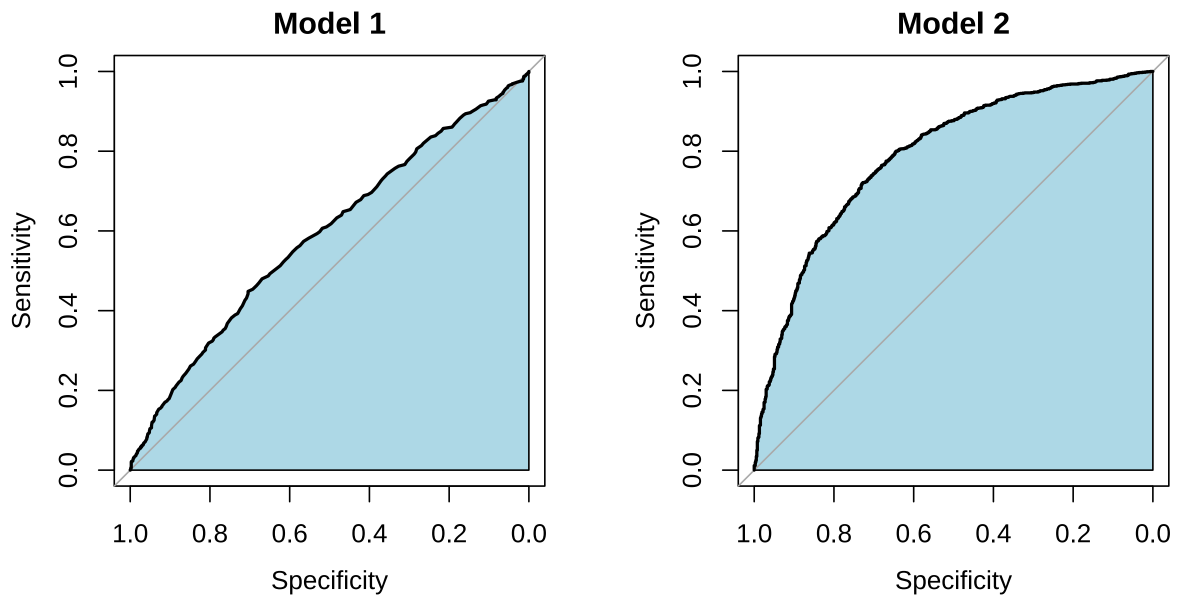
```
# Age and sex adjusted
s.fit.adj <- glm(sigdz~cholest+age+sex, data=s.data, family = "binomial")
summary(s.fit.adj)

##
## Call:
```

```
## glm(formula = sigdz ~ cholest + age + sex, family = "binomial",
##     data = s.data)
##
## 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
```

Recall that, if there is a confounding variable, this will cause the estimated relationship between the main predictor and outcome to be affected. So we need to check whether the beta for cholesterol changes after including age and sex in the model. The beta from the first logistic regression model was 0.006227 and from the second model was 0.009006. So $\frac{0.009006}{0.006227} = 1.4463$. Thus, the beta increased by 44% after including age and sex. This implies that either age or sex (or both) could be confounders in the cholesterol-coronary disease relationship.

(e)

```
library("pROC")

## Type 'citation("pROC")' for a citation.
##
## Attaching package:  'pROC'
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

# Predicted values for both models
pred.val.s <- predict(s.fit)
pred.val.s.adj <- predict(s.fit.adj)
# ROC curves and AUC
par(mfrow=c(1,2))
auc(s.data$sigdz, pred.val.s, plot=TRUE, auc.polygon=TRUE,
    auc.polygon.col="lightblue", asp=FALSE,
    main="Model 1")

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

## Area under the curve: 0.5887

auc(s.data$sigdz, pred.val.s.adj, plot=TRUE, auc.polygon=TRUE,
    auc.polygon.col="lightblue", asp=FALSE,
    main="Model 2")

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



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

The AUC for the first model is 0.5887, and for the second model it is 0.7887. Thus the second model has better predictive accuracy, since the AUC is greater for that model.