Assignment 2 - SOLUTIONS

Instructor: Kevin McGregor

MATH 4330

Question 1:

(a)

After reading in the dataset, we code high vs. low colesterol based on the median using the `ifelse()` function.

```
# getting high/low cholesterol
med.chol <- median(dat$cholest)
chol <- ifelse(dat$cholest>med.chol, "high", "low")

# re-coding significant coronary disease by cardiac catheterization
# This is not necessary, but makes reading the output easier
sig.cd <- ifelse(dat$sigdz==1, "yes", "no")
```

Now, let's make a contingency table for high/low cholesterol vs. significant coronary disease. If we want to use the $\frac{ad}{bc}$ formula for the odds ratio, we need to make sure the table is oriented correctly. That is, by default, the ordering of the columns of the table are "no" "yes", but we need it to be "yes" "no". We can easily do this by exchanging the columns.

```
tab <- table(chol, sig.cd)
# Exchanging the columns of the table
tab <- cbind(tab[,2], tab[,1])
colnames(tab) <- c("yes", "no")
tab

##      yes  no
## high 816 313
## low  674 455
```

We can then calculate the odds ratio as:

```
# Odds ratio
OR <- tab[1,1]*tab[2,2]/(tab[1,2]*tab[2,1])
OR

## [1] 1.759938
```

Interpretation: The odds ratio is 1.7599, meaning that the odds of significant coronary disease is approximately 75.99% higher in the high-cholesterol group than in the low-cholesterol group.

(b)

For severe coronary disease we proceed in the same fashion as in part (a).

```
# re-coding severe coronary disease by cardiac catheterization
sev.cd <- ifelse(dat$tvdlm==1, "yes", "no")
tab2 <- table(chol, sev.cd)
tab2 <- cbind(tab2[,2], tab2[,1])
colnames(tab2) <- c("yes", "no")
OR <- tab2[1,1]*tab2[2,2]/(tab2[1,2]*tab2[2,1])
OR

## [1] 1.414939
```

Interpretation: The odds ratio is 1.414939, meaning that the odds of severe coronary disease is approximately 41.49% higher in the high-cholesterol group than in the low-cholesterol group.

(c)

This is a retrospective study, so we can't calculate the risk ratio directly. The only way to estimate the risk ratio is if the rare disease assumption holds. Coronary disease is actually not that rare, so it's unlikely that the rare disease assumption would hold.
However: since you were not given the population prevalence of coronary disease, you still get full marks if you estimated the risk ratio using the odds ratio as long as you specified the rare disease assumption.

Question 2:

(a)

Running the linear regression model:

```
lm.fit <- lm(cholest~age+sex, data=dat)
summary(lm.fit)

##
## Call:
## lm(formula = cholest ~ age + sex, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -208.28  -34.10   -4.92   28.47  339.19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 224.96743    5.83456  38.558  < 2e-16 ***
## age           0.03886    0.11307   0.344    0.731
## sex           9.78549    2.31139   4.234 2.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.43 on 2255 degrees of freedom
## Multiple R-squared:  0.008077,Adjusted R-squared:  0.007198
## F-statistic: 9.182 on 2 and 2255 DF,  p-value: 0.0001068
```

Based on the output, the **fitted** regression model is:

$$\texttt{cholest} = 224.9674 + 0.0388 \times \texttt{age} + 9.7855 \times \texttt{sex}$$

The parameter interpretations are:

- $\beta_0 = 224.9674$; since $\texttt{age} = 0$ falls well outside the range of the data (age range is 17-81), β_0 has no interpretation.
- $\beta_1 = 0.03886$; this means that if age increases by one year, then the mean cholesterol increases by 0.03886, assuming sex is held constant.
- $\beta_2 = 9.78549$; this means that the mean difference in cholesterol between females and males is 9.78549, assuming that age is held constant.

(b)

```
confint(lm.fit)

##                   2.5 %      97.5 %
## (Intercept) 213.5257660 236.4090892
## age          -0.1828807   0.2605961
## sex           5.2528151  14.3181654
```

The confidence intervals and interpretations are:

- `age`: $(-0.1828, 0.2606)$. Since the confidence interval overlaps with zero, we do not have evidence of association between age and cholesterol.
- `sex`: $(5.2528, 14.3181)$. Since the confidence interval does **not** overlap with zero, we have evidence of association between sex and cholesterol.

(c)

Our prediction for a 50-year-old female is:

```
predict(lm.fit, newdata = data.frame(age=50, sex=1))

##        1
## 236.6958
```

(d)

```r
predict(lm.fit, newdata = data.frame(age=10, sex=0))

##       1
## 225.356
```

We are much less confident of the prediction for a 10-year-old since this age falls well outside of the age range in the dataset (age range 17-81).