Methods II: Homework 1

Tim Vigers

27 January 2019

1. First consider transforming covariates and the outcome.

a. Is categorization necessary for BMI?

```
mod <- lm(sodium ~ bmi,data = hyponat)
summary(mod)
```

```
##
## Call:
## lm(formula = sodium ~ bmi, data = hyponat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.310  -2.382   0.535   3.271  15.668
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.54400    2.16471  64.463   <2e-16 ***
## bmi           0.03596    0.09326   0.386      0.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.876 on 368 degrees of freedom
## Multiple R-squared:  0.0004039,  Adjusted R-squared:  -0.002312
## F-statistic: 0.1487 on 1 and 368 DF,  p-value: 0.7
```

```
polymod <- lm(sodium ~ bmi + I(bmi^2),data = hyponat)
summary(polymod)
```

```
##
## Call:
## lm(formula = sodium ~ bmi + I(bmi^2), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4019  -2.8199   0.1535   3.0960  15.2932
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 85.94424   13.62912   6.306 8.24e-10 ***
## bmi          4.56748    1.14186   4.000 7.66e-05 ***
## I(bmi^2)    -0.09440    0.02371  -3.981 8.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.78 on 367 degrees of freedom
```

```
## Multiple R-squared:  0.04179,    Adjusted R-squared:  0.03657
## F-statistic: 8.003 on 2 and 367 DF,  p-value: 0.0003964
```

```
vif(polymod)
```

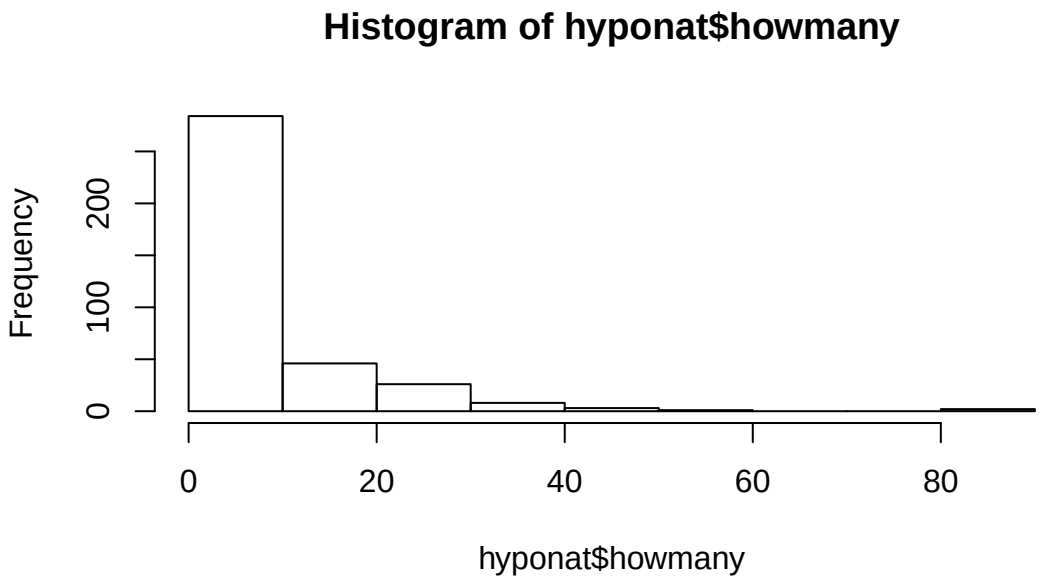
```
##      bmi I(bmi^2)
## 155.9472 155.9472
```

```
hyponat$bmiC <- cut(hyponat$bmi,c(0,20,25,Inf))
```

The quadratic BMI term is significant, and the VIF values for the polynomials are large. This just shows that there is indeed a quadratic relationship and that the polynomial terms are collinear (as we were told in the question). When this is the case, it's correct to make the variable categorical as long as doing so makes scientific sense. In the case of BMI, it does make sense to split people into categorical groups like underweight, normal, and overweight. This removes the collinearity concern, and the model is still easily interpretable.

b. Should the number of previous marathons run be dichotomized?

```
hist(hyponat$howmany)
```



The number of previous marathons is very skewed, which violates the assumption of normality. So, dichotomizing this variable at the median is a good idea.

c. Is there a quadratic relationship between weight change and sodium levels?

```
fit <- lm(sodium ~ poly(wtdiff,2), data = hyponat)
summary(fit)
```

```
##
## Call:
## lm(formula = sodium ~ poly(wtdiff, 2), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.0835  -2.4685   0.3256   2.6527  14.2696
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        140.373      0.224 626.688  < 2e-16 ***
## poly(wtdiff, 2)1  -42.313      4.309  -9.821  < 2e-16 ***
## poly(wtdiff, 2)2  -12.216      4.309  -2.835  0.00483 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.309 on 367 degrees of freedom
## Multiple R-squared:  0.2216, Adjusted R-squared:  0.2174
## F-statistic: 52.24 on 2 and 367 DF,  p-value: < 2.2e-16
```

There does appear to be a quadratic relationship between weight change and sodium levels ($p = 0.00483$).

d. Should fluid frequency be treated as a continuous variable or 2 indicator variables?

```
hyponat$fluidfr3 <- as.factor(hyponat$fluidfr3)
```

The levels of fluidfr3 are 1 = every one mile, 2 = every two miles, and 3 = every third mile or more. I don't see how this could be treated as a continuous variable, so I think it's best to keep it as a categorical variable (indicator functions). You could maybe use total water intake as a continuous variable if that information was available, but this data can't be treated as continuous.

e. The authors only used weight change and excluded the self-reported variables from the multivariable analysis. Is this an issue?

I think this approach sort of makes sense. Weight difference is probably the best measure of fluid loss/intake (assuming they're consuming a negligible amount of solid food), and the other three variables are reporting similar information. When this is the case, dropping the self-reported variables makes sense as they're most likely the least accurate.

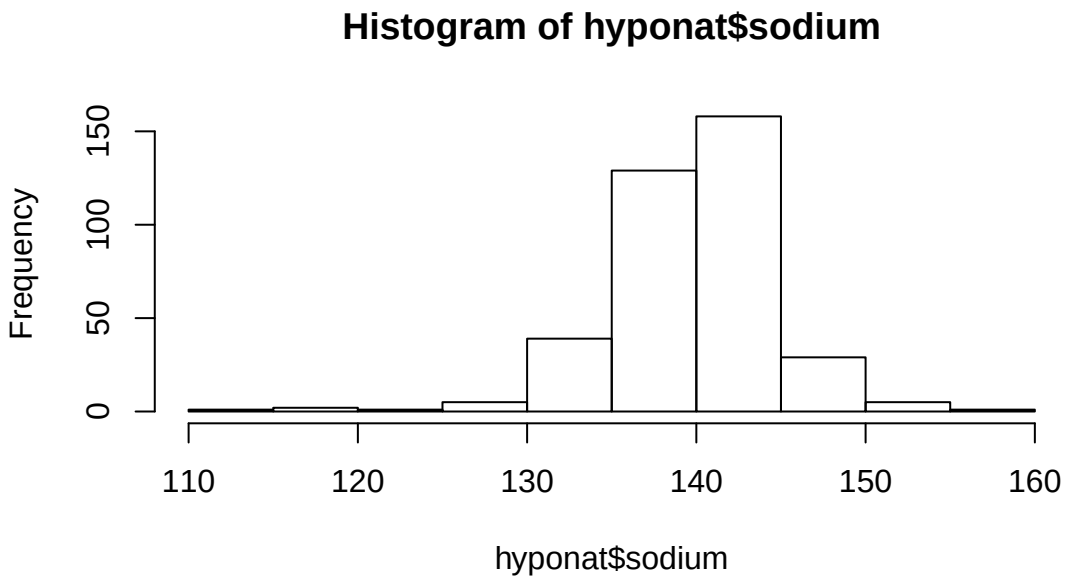
My main concern would be if one of those variables is really reporting different information, and by excluding them you're losing valuable data. Also, I worry a little bit about dropping 3 out of 4 variables, so it might be good to investigate the collinearity further and only drop 2.

f. Only running time was used in the multivariable model and not training pace since it is self-reported. Is this an issue?

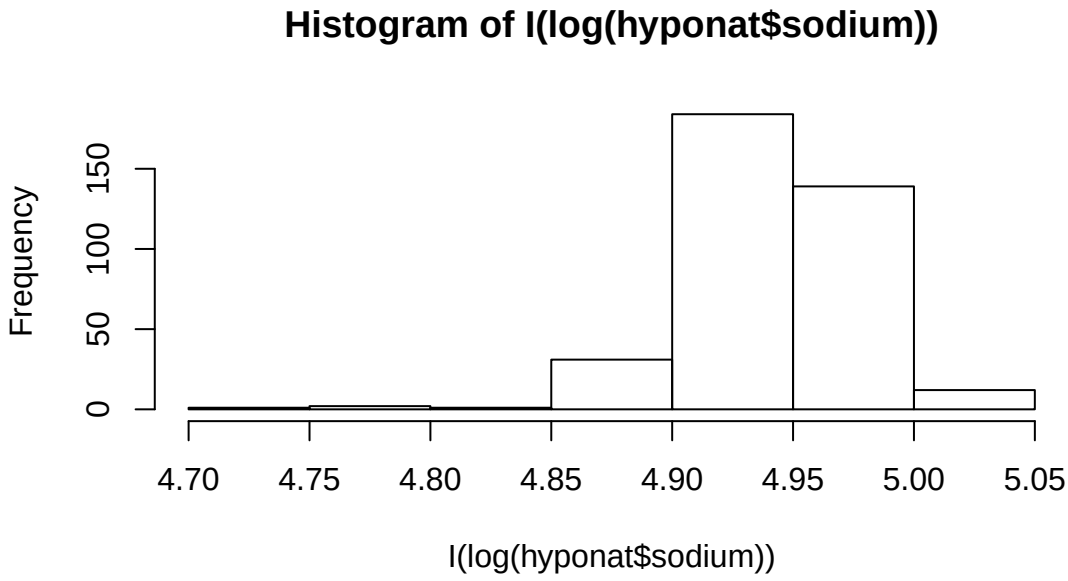
I'm more comfortable with this than the previous question, since you're only looking at two variables, and they pretty clearly tell you the same information. If you ran the whole marathon quickly, it seems safe to assume that your training pace was also fast. And since it's self reported (and possibly hard to measure accurately yourself), you have to worry about inaccuracy or people intentionally overestimating how quickly they run.

g. Should the outcome sodium levels be log transformed?

```
hist(hyponat$sodium)
```




```
hist(I(log(hyponat$sodium)))
```




```
lillie.test(hyponat$sodium)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  hyponat$sodium
## D = 0.10252, p-value = 5.076e-10
```

```
lillie.test(I(log(hyponat$sodium)))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  I(log(hyponat$sodium))
## D = 0.11078, p-value = 8.407e-12
```

Log transforming the outcome clearly doesn't change much (the log transformed outcome is still not normally distributed, and the histograms look very similar). So, I would keep the outcome as it is, especially since it's generally best to avoid transforming the outcome if possible.