Methods II: Homework 1

Tim Vigers

29 January 2019

```
# Load libraries
library(car)
```

```
## Loading required package: carData
```

```
library(nortest)
library(MASS)
# Read in data
hyponat <- read.table("/Users/timvigers/Documents/School/UC Denver/Biostatistics/Biostatistical Methods
```

1. 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))
category <- lm(sodium ~ bmiC, data = hyponat)
summary(category)
```

```
##
## Call:
## lm(formula = sodium ~ bmiC, data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26.8314  -2.8314   0.1686   3.1686  15.1686 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  138.2973     0.7921 174.604  < 2e-16 ***
## bmiC(20,25]    2.5341     0.8476   2.990  0.00298 ** 
## bmiC(25,Inf]   1.5617     0.9617   1.624  0.10528    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.818 on 367 degrees of freedom
## Multiple R-squared:  0.02669,    Adjusted R-squared:  0.02139 
## F-statistic: 5.032 on 2 and 367 DF,  p-value: 0.006984
```

```
AIC(polymod,category)
```

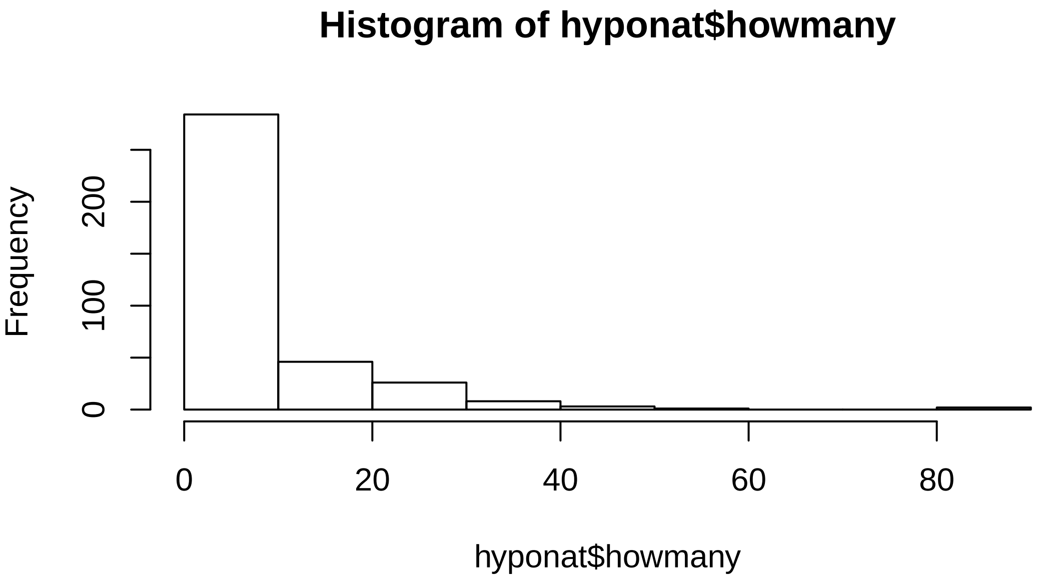
```
##          df      AIC
## polymod   4 2212.750
## category  4 2218.535
```

Compare the categorical to the polynomial model (use AIC). Use mean centering to fix collinearity. Add some model diagnostics as well.

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.

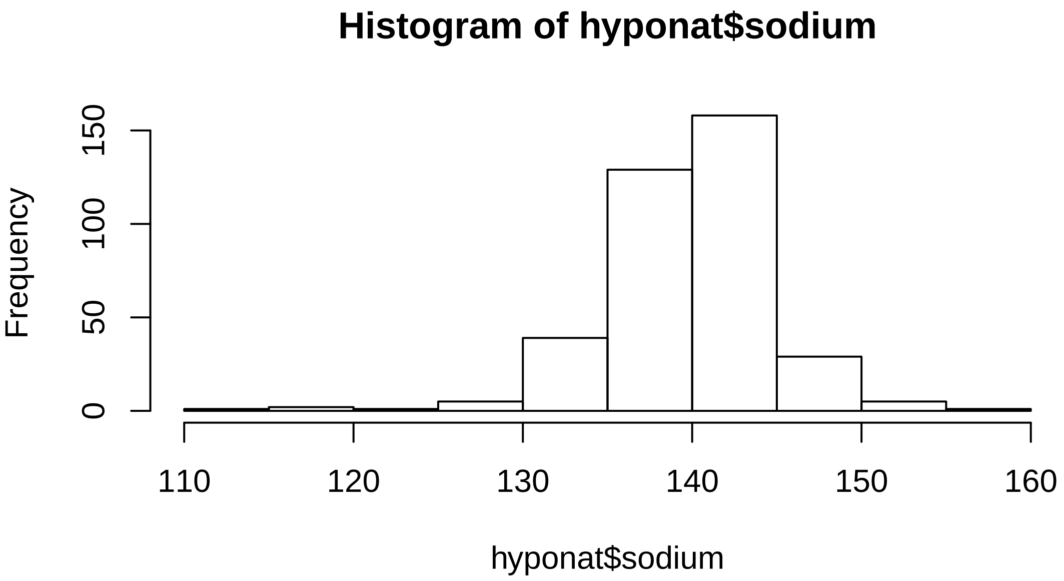
My main concern would be if one of those variables is really reporting different information, and by excluding them you're losing valuable data. Also, I worry a little bit about dropping 3 out of 4 variables, so it might be good to investigate the collinearity further and only drop 2.

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

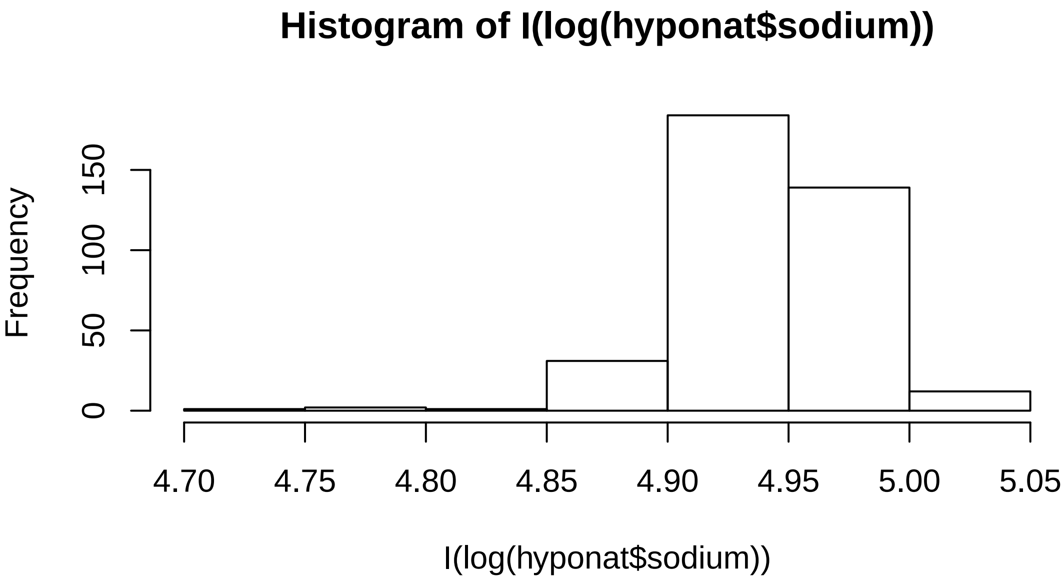
I'm more comfortable with this than the previous question, since you're only looking at two variables, and they pretty clearly tell you the same information. If you ran the whole marathon quickly, it seems safe to assume that your training pace was also fast. And since it's self reported (and possibly 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.

2. Run the single variable analyses.

```
vars <- colnames(hyponat)[-c(which(colnames(hyponat)=="sodium"))]
univar <- lapply(vars, function(x){
  summary(lm(as.formula(paste0("sodium ~ ",x)), data = hyponat))
})
univar
```

```
## [[1]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1339  -2.1339  -0.0206   2.9794  14.9794
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 141.0206     0.3075 458.603  < 2e-16 ***
## female       -1.8867     0.5249  -3.595 0.000369 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.793 on 368 degrees of freedom
## Multiple R-squared:  0.03392,    Adjusted R-squared:  0.0313
## F-statistic: 12.92 on 1 and 368 DF,  p-value: 0.000369
##
##
## [[2]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.5557  -2.4258   0.4443   3.3860  15.5432
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 140.58045    0.31964 439.812   <2e-16 ***
## howmany      -0.02474    0.02327  -1.063    0.288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.869 on 368 degrees of freedom
## Multiple R-squared:  0.003062,   Adjusted R-squared:  0.000353
## F-statistic:  1.13 on 1 and 368 DF,  p-value: 0.2884
##
##
## [[3]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.2152  -2.3795   0.4178   3.1970  15.6206
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.64576    1.07859 129.471   <2e-16 ***
## age           0.01898    0.02736   0.694    0.488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.874 on 368 degrees of freedom
## Multiple R-squared:  0.001306,   Adjusted R-squared:  -0.001408
## F-statistic: 0.4811 on 1 and 368 DF,  p-value: 0.4883
##
##
## [[4]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.7711  -2.0888   0.2289   2.9112  16.2289
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 141.0888     0.3717 379.563  < 2e-16 ***
## lwobup01     -1.3176     0.5043  -2.613  0.00935 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.832 on 368 degrees of freedom
## Multiple R-squared:  0.01821,    Adjusted R-squared:  0.01554
## F-statistic: 6.826 on 1 and 368 DF,  p-value: 0.009353
##
##
## [[5]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.2390  -2.2390   0.2551   3.1316  15.2551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 140.7449     0.4921 285.997   <2e-16 ***
## wateld01     -0.5059     0.5740  -0.881    0.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.872 on 368 degrees of freedom
## Multiple R-squared:  0.002107,   Adjusted R-squared:  -0.0006048
## F-statistic: 0.777 on 1 and 368 DF,  p-value: 0.3786
##
##
## [[6]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4699  -2.4699   0.5301   3.4571  16.2381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 140.4699     0.2602 539.873   <2e-16 ***
## urinat3p     -1.7080     1.0922  -1.564    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.861 on 368 degrees of freedom
## Multiple R-squared:  0.006602,   Adjusted R-squared:  0.003903
## F-statistic: 2.446 on 1 and 368 DF,  p-value: 0.1187
##
##
## [[7]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.7480  -2.7368   0.2632   3.2520  15.2632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.7368     0.3316 421.356  < 2e-16 ***
## fluidfr32     1.0112     0.5394   1.875 0.061645 .
## fluidfr33     3.1455     0.8866   3.548 0.000439 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.794 on 367 degrees of freedom
## Multiple R-squared:  0.03617,    Adjusted R-squared:  0.03091
## F-statistic: 6.885 on 2 and 367 DF,  p-value: 0.00116
##
##
## [[8]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.1017  -2.5612   0.3042   2.5762  14.5866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.4968     0.2434 573.117   <2e-16 ***
## wtdiff       -1.4092     0.1449  -9.728   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.35 on 368 degrees of freedom
## Multiple R-squared:  0.2046, Adjusted R-squared:  0.2024
## F-statistic: 94.64 on 1 and 368 DF,  p-value: < 2.2e-16
##
##
## [[9]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.7264  -2.5159   0.1344   2.7186  16.6333
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.802143   1.420510 101.937  < 2e-16 ***
## runtime      -0.019501   0.006157  -3.168  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.812 on 368 degrees of freedom
## Multiple R-squared:  0.02654,    Adjusted R-squared:  0.0239
## F-statistic: 10.03 on 1 and 368 DF,  p-value: 0.001666
##
##
## [[10]]
##
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4851  -2.4851   0.1256   2.8313  16.0718
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 146.712879   1.968890  74.516  < 2e-16 ***
## trainpse     -0.012975   0.003997  -3.246  0.00128 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.809 on 368 degrees of freedom
## Multiple R-squared:  0.02784,    Adjusted R-squared:  0.0252 
## F-statistic: 10.54 on 1 and 368 DF,  p-value: 0.001276
## 
## 
## [[11]]
## 
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), 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
## 
## 
## [[12]]
## 
## Call:
## lm(formula = as.formula(paste0("sodium ~ ", x)), data = hyponat)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26.8314  -2.8314   0.1686   3.1686  15.1686 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  138.2973     0.7921 174.604  < 2e-16 ***
## bmiC(20,25]    2.5341     0.8476   2.990  0.00298 ** 
## bmiC(25,Inf]   1.5617     0.9617   1.624  0.10528    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.818 on 367 degrees of freedom
## Multiple R-squared:  0.02669,    Adjusted R-squared:  0.02139 
## F-statistic: 5.032 on 2 and 367 DF,  p-value: 0.006984
```

a. Which variables are associated with sodium levels at the 0.05 level of significance?

*Add bmiC to these

Based on the above output, the variables significantly associated with sodium are: female, lwobup01, fluidfr3, wtdiff, runtime, trainpse. So sex, whether you use anti-inflammatory medications, fluid intake frequency, weight change during the marathon, how quickly you run the marathon, and your training pace are all associated with sodium levels.

b. How do these univariate analyses compare to the original paper where sodium levels were dichotomous?

The paper concluded that "considerable weight gain while running, a long racing time, and bodymass index extremes were associated with hyponatremia, whereas female sex, composition of fluids ingested, and use of nonsteroidal anti-inflammatory drugs were not." So our analyses agree that weight change and running time are associated with sodium. However, we did not find that BMI was associated (when treated as a continuous or a categorical variable), and did find that sex and use of NSAIDs were.

3. Multivariable analyses with stepwise regression based on AIC

```r
sigvars <- c("female","lwobup01","fluidfr3","wtdiff","runtime","trainpse","bmiC")
sigvars <- paste(sigvars,collapse = " + ")
formula <- as.formula(paste0("sodium ~ ",sigvars))
stepwise <- stepAIC(lm(formula, data = hyponat),direction = "both",trace = 0)
stepwise
```

```
##
## Call:
## lm(formula = sodium ~ lwobup01 + wtdiff + bmiC, data = hyponat)
##
## Coefficients:
##  (Intercept)      lwobup01        wtdiff   bmiC(20,25]  bmiC(25,Inf]
##     138.7320       -0.7481       -1.3379        1.4928        0.8858
```

a. What predictors are included in the final model?

Using both forward and backward stepwise regression based in AIC, change in weight and anti-inflammatory usage are both associated with sodium level, and so is BMI category.

b.

There are a couple of problems with this method. First, it's possible that some variables are significant in a multiple regression model, but are not significant when tested on their own. Second, it doesn't take polynomial associations (e.g. BMI) or collinearity into account. Lastly, this way of approaching things doesn't really think about the scientific question. All of the variables in this data set make sense to test, but many data sets include lots of variables that don't make sense with the question at hand, so just using this approach without thinking can end up including nonsensical variables in the model.

4. Partial F test with all covariates with a p-value less than 0.1

```r
sigvars <- c("female","lwobup01","fluidfr3","wtdiff","runtime","trainpse","bmiC")
sigvars <- paste(sigvars,collapse = " + ")
formula <- as.formula(paste0("sodium ~ ",sigvars))
anova(lm(formula, data = hyponat))
```

```
## Analysis of Variance Table
##
## Response: sodium
##            Df Sum Sq Mean Sq F value    Pr(>F)
## female      1  296.9  296.91 15.8502 8.295e-05 ***
## lwobup01    1  123.1  123.09  6.5711  0.010771 *
## fluidfr3    2  223.1  111.55  5.9548  0.002856 **
## wtdiff      1 1275.1 1275.09 68.0698 3.009e-15 ***
## runtime     1   12.3   12.32  0.6575  0.417983
## trainpse    1    5.7    5.71  0.3051  0.581053
## bmiC        2   72.8   36.38  1.9419  0.144925
## Residuals 360 6743.6   18.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. What predictors are included in the final model?

The predictors in this model are the same as above. Changing the alpha level to 0.1 instead of 0.05 did not alter which variables were considered significant (i.e. there were no p values > 0.05 and < 0.1).

b. What are the results of the F test?

Based on this ANOVA table, we would keep sex, NSAID usage, fluid frequency, and weight change in the model.

5. Why do you think that there are more significant covariates in the final model for a binary outcome than there are for a continuous outcome?