Methods II: Homework 1

Tim Vigers

27 January 2019

1. First consider transforming covariates and the outcome.

a. Is categorization necessary for BMI?

```r
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
```

```r
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
```

```r
vif(polymod)
```

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

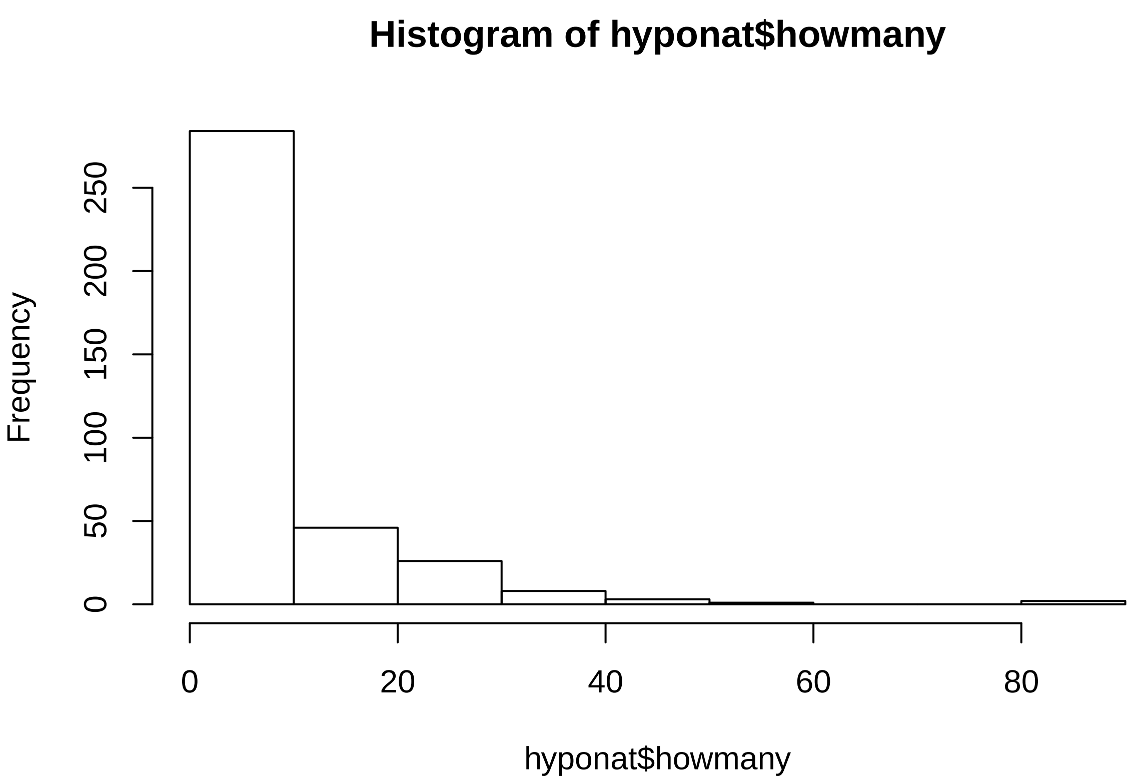
```r
hyponat$bmiC <- cut(hyponat$bmi,c(0,20,25,Inf))
```

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.

WHY CATEGORICAL?

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

```r
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?

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

```
##
## Call:
## lm(formula = sodium ~ wtdiff + I(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) 139.84844    0.27113 515.800  < 2e-16 ***
## wtdiff       -1.61468    0.16075 -10.044  < 2e-16 ***
## I(wtdiff^2)  -0.16980    0.05988  -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
```

```r
anova(mod,polymod)
```

```
## Analysis of Variance Table
##
## Model 1: sodium ~ wtdiff
## Model 2: sodium ~ wtdiff + I(wtdiff^2)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    368 6962.1
## 2    367 6812.9  1    149.24 8.0395 0.00483 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the partial F test is significant at the 0.05 level, there does appear to be a quadratic relationship between weight change and sodium levels.