Week6-Exercises-Solutions

Exercise solutions

Week 6

There are a few issues with this dataset, namely the potential non-linearity that can be corrected with a log-transformation of fev1. However we will put this issue aside so that we can focus on collinearity.

Carrying out the regression, and calculating the variance inflation factors (VIF) gives some very high VIFs (e.g. armspan, ulna length, forearm length, and height), and some moderately high VIFs with age and weight. This makes a lot of sense, armspan, ulna length and forearm lengths are just different measures of "arm length". We would also expect children that are taller to have longer arms, weight more, and be older children. So lots of collinearity issues here.

The VIF is a measure of how much each covariate is associated with the other covariates in the model. To compute the VIF for the covariate age, for example, we could run a regression for age using all the other covariates:

$$\overline{age} = \alpha_0 + \alpha_1 wt + \alpha_2 armsp + \alpha_3 ulna + \alpha_4 farm$$

And the $VIF(age) = \frac{1}{1-R^2}$, where R^2 is computed from the model above.

We will first address the similar "arm length" measurements by running four models, one model for each armlength (with the other two removed).

These three models certainly elimate some of the collinearity, but there does still seem to be bad collinearity between height and the remaining armlength variable, and between age, weight (and potentially height). Let's next explore removing either height, or the armlenth variable to see how well this addresses the issue.

This seems to address the issue suitably. Although there is still collinearity between the remaining variables, all still remain strongly associated with the outcome (indicated by small p-values), and so despite their correlations, they are making independent contributions to explaining the variance in FEV1. You may also have noticed how removal of collinear variables has reduced the standard errors in the remaining variables considerably - a good thing.

Finally we must decide which analysis is most suitable for our research purpose. The final four models all are reasonably similar from a statistical perspective, with only a 2% difference observed in R-squared values (between 82% and 84% of variance explained across the models). Given this similarity, instead of choosing the model with the highest R-squared value, we may wish to consider which of the three measurements is most easily captured with low measurement error. Here we can rule out ulna length, as this is most likely the most difficult to measure accurately and non-invasively. Height, may be an obvious choice here - but it would depend on your sample. If your sample is likely to contain children who are unwell (lying in bed), or are in wheelchairs, height would be difficult to measure. Similarly, armspan may be difficult to capture for children with arm injuries. Given these factors, forearm length seems an appropriate choice for this model, and has only a marginally smaller R-squared value than height.

Stata code ::: {.cell}

```
import delimited "https://www.dropbox.com/s/2cs0z39v2ekni81/lungfun.csv?dl=1"

* Initial full model
reg fev1 age wt ht armsp ulna farm
vif

* Models with only 1 arm length measurement
reg fev1 age wt ht armsp
vif
reg fev1 age wt ht ulna
vif
reg fev1 age wt ht farm
vif

* Models with only 1 arm length measurement or with height (but not both)
reg fev1 age wt ht
vif
reg fev1 age wt armsp
vif
reg fev1 age wt ulna
vif
reg fev1 age wt farm
vif
```

:::

R code ::: {.cell}

```
library(car)
lungfun <- read.csv("https://www.dropbox.com/scl/fi/eee36d1h85s8gnejdqykx/lungfun.csv?rlkey=5
# Initial full model
lungfun.lm1 <- lm(fev1 ~ age + wt + ht + armsp + ulna + farm, data=lungfun)
summary(lungfun.lm1)
vif(lungfun.lm1)

# Models with only 1 arm length measurement
lungfun.lm3 <- lm(fev1 ~ age + wt + ht + armsp              , data=lungfun)
lungfun.lm2 <- lm(fev1 ~ age + wt + ht +         ulna       , data=lungfun)
lungfun.lm4 <- lm(fev1 ~ age + wt + ht +               farm , data=lungfun)
vif(lungfun.lm2)
vif(lungfun.lm3)
vif(lungfun.lm4)
summary(lungfun.lm2)
summary(lungfun.lm3)
summary(lungfun.lm4)

# Models with only 1 arm length measurement or with height (but not both)
lungfun.lm5 <- lm(fev1 ~ age + wt + ht, data=lungfun)
lungfun.lm6 <- lm(fev1 ~ age + wt + armsp, data=lungfun)
lungfun.lm7 <- lm(fev1 ~ age + wt + ulna, data=lungfun)
lungfun.lm8 <- lm(fev1 ~ age + wt + farm, data=lungfun)
vif(lungfun.lm5)
vif(lungfun.lm6)
vif(lungfun.lm7)
vif(lungfun.lm8)
summary(lungfun.lm5)
summary(lungfun.lm6)
summary(lungfun.lm7)
summary(lungfun.lm8)
```

:::