Stat500(Section002): Homework #6

Due on Nov.10, 2021 at 3:00pm

Instructor:Naisyin Wang

Tiejin Chen **tiejin@umich.edu**

Problem 1

Part a

Using the following code to fit the model and compute the condition numbers.

```
library(faraway)
data(divusa)
lmod = lm(divorce~unemployed+femlab+marriage+birth+military,divusa)
x = model.matrix(lmod)[,-1]
e = eigen(t(x) %*% x)
sqrt(e$val[1]/e$val)
```

We get the result of condition number is :

```
[1]  1.000000  7.432684  8.532498 13.757290 25.150782
```

We can see that all the condition numbers are under 30. Hence it shows that there is no significant sign of collinearity.

Part b

Using the vif function to compute VIFs.

```
vif(x)
```

We get the result:

```
unemployed    femlab   marriage     birth   military
  2.252888  3.613276   2.864864  2.585485   1.249596
```

As we can see, all VIFs are quite small, and none of them are greater than 10. Hence there is no sign of collinearity. And we can get the summary.

```
Call:
lm(formula = divorce ~ unemployed + femlab + marriage + birth +
    military, data = divusa)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8611 -0.8916 -0.0496  0.8650  3.8300

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.48784    3.39378   0.733   0.4659
unemployed  -0.11125    0.05592  -1.989   0.0505 .
femlab       0.38365    0.03059  12.543  < 2e-16 ***
marriage     0.11867    0.02441   4.861 6.77e-06 ***
birth       -0.12996    0.01560  -8.333 4.03e-12 ***
military    -0.02673    0.01425  -1.876   0.0647 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.65 on 71 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9152
F-statistic: 165.1 on 5 and 71 DF,  p-value: < 2.2e-16
```

Problem 1 continued on next page. . .

We can see the military and unempolyed are two insignificant variables. However, their VIFs are even the smallest VIFs of all VIFs. Hence there is no evidence of that collinearity causes some predictors not to be significant.

Part c

We remove the military and unempolyed variables to get the new model.

```
lmod1 = lm(divorce~femlab+marriage+birth,data=divusa)
x1 = model.matrix(lmod1)[,-1]
e1 = eigen(t(x1) %*% x1)
sqrt(e1$val[1]/e1$val)
vif(x1)
```

We get the result:

```
[1]  1.000000  7.432012 13.659660
```

```
  femlab marriage    birth
1.893390 2.201891 2.008469
```

From the result, we know that the VIFs and condition number are all become smaller. Though there is no or little collinearity of previous model, removing insignificant predictors can still reduce collinearity.

Problem 2

Part a

Using the following code to fit the model and compute the condition numbers.

```
data(prostate)
lmod_pro = lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=prostate)
x_pro = model.matrix(lmod_pro)[,-1]
e_pro = eigen(t(x_pro) %*% x_pro)
sqrt(e_pro$val[1]/e_pro$val)
```

We get the result:

```
[1]   1.00000   2.78186  47.66094  52.22787  85.98499 103.73114
153.85414 243.30248
```

There is a wide range in the eigenvalues. And since we consider condition number which is greater than 30 as a large nubmer, there are 6 condition numbers are large.This means that problems are being caused by more than just one linear combination.

Part b

```
round(cor(prostate[1:8]),2)
```

Get the result:

Problem 2 continued on next page. . .

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
lcavol	1.00	0.19	0.22	0.03	0.54	0.68	0.43	0.43
lweight	0.19	1.00	0.31	0.43	0.11	0.10	0.00	0.05
age	0.22	0.31	1.00	0.35	0.12	0.13	0.27	0.28
lbph	0.03	0.43	0.35	1.00	-0.09	-0.01	0.08	0.08
svi	0.54	0.11	0.12	-0.09	1.00	0.67	0.32	0.46
lcp	0.68	0.10	0.13	-0.01	0.67	1.00	0.51	0.63
gleason	0.43	0.00	0.27	0.08	0.32	0.51	1.00	0.75
pgg45	0.43	0.05	0.28	0.08	0.46	0.63	0.75	1.00

We can find that $(gleason, pgg45), (lcp, lcavol), (lcp, svi), (lcp, pgg45)$ are some highly correlated pair predictors. More than one pair are highly correalted. All of them cause the collinearity and make many condition numbers large.

Part c

```
vif(x_pro)
```

Get the result:

```
  lcavol  lweight      age     lbph      svi      lcp  gleason    pgg45 
2.054115 1.363704 1.323599 1.375534 1.956881 3.097954 2.473411 2.974361 
```

We can find all VIFs are small.

Part d

From the condition numbers, we could say there is collinearity in the model.However, though there is a sign of collinearity, this collinearity does not effect the significant of predictors.