HW4

Tianqi Wu

3/25/2020

```r
library(faraway)
attach(salmonella)
attach(gammaray)
attach(longley)
attach(prostate)
library(nlme)
library(lmtest)
```

Problem 1

Since p-value is 0.1341985 > 0.05, we fail to reject the null and conclude that there is no lack of fit.

```r
model.1 = lm(colonies~log(dose+1),data=salmonella)
summary(model.1)
```

```
##
## Call:
## lm(formula = colonies ~ log(dose + 1), data = salmonella)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.376  -6.882  -1.509   5.400  29.119
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     19.823      5.064   3.915  0.00123 **
## log(dose + 1)    2.396      1.128   2.125  0.04955 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.84 on 16 degrees of freedom
## Multiple R-squared:  0.2201, Adjusted R-squared:  0.1713
## F-statistic: 4.514 on 1 and 16 DF,  p-value: 0.04955
```

```r
model.1a=lm(colonies~factor(log(dose+1)),data=salmonella);
anova(model.1, model.1a)
```

```
## Analysis of Variance Table
##
## Model 1: colonies ~ log(dose + 1)
## Model 2: colonies ~ factor(log(dose + 1))
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
```

```
## 1     16 1881.1
## 2     12 1091.3  4    789.73 2.1709 0.1342
```

```
1-pf(2.1709,4,12)
```

```
## [1] 0.1341985
```

Problem 2

Adjusted R-squared from the WLS model is 0.968. We first apply log tranformation and use the errors to define the weights for WLS model.

```
# time series model
#model.2.ls = lm(flux~time, data=gammaray)
#dwtest(model.2.ls)
#model.2.gls = gls(flux~time,correlation=corARMA(p=1), weights = varFunc(~time), data=gammaray)
#summary(model.2.ls)

# WLS model
model.2 = lm(flux~time, data=log(gammaray), weight=1/abs(error))
summary(model.2)
```

```
##
## Call:
## lm(formula = flux ~ time, data = log(gammaray), weights = 1/abs(error))
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65432 -0.28032 -0.12943  0.07607  0.48068
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.27403    0.19195   58.73   <2e-16 ***
## time        -1.24804    0.02882  -43.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2511 on 61 degrees of freedom
## Multiple R-squared:  0.9685, Adjusted R-squared:  0.968
## F-statistic:  1876 on 1 and 61 DF,  p-value: < 2.2e-16
```

Problem 3

Condition numbers

Since the condition number is 110.54 $>$ 30, it is ill-conditioned and it indicates collinearity in data.

```
model.3 = lm(Employed ~ .,data=longley)
# condition number
x = model.matrix(model.3)[,-1]
x = x - matrix(apply(x,2, mean), nrow(x), ncol(x), byrow=TRUE)
x = x / matrix(apply(x, 2, sd), nrow(x), ncol(x), byrow=TRUE)
apply(x,2,mean)
```

```
## GNP.deflator           GNP    Unemployed  Armed.Forces    Population
## -4.774826e-16 -1.118897e-16 -5.724587e-17  5.767956e-17  8.949411e-16
##          Year
##  0.000000e+00
```

```r
apply(x,2,var)
```

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year
##            1            1            1            1            1            1
```

```r
e = eigen(t(x) %*% x)
sqrt(e$val[1]/e$val)
```

```
## [1]   1.000000   1.979048   4.757028  17.560372  42.470986 110.544153
```

Correlation between predictors

GNP.deflator, GNP, Population, Year and Employed are highly correlated to each other and it indicates collinearity in data.

```r
round(cor(longley), dig=2)
```

```
##              GNP.deflator  GNP Unemployed Armed.Forces Population Year Employed
## GNP.deflator         1.00 0.99       0.62         0.46       0.98 0.99     0.97
## GNP                  0.99 1.00       0.60         0.45       0.99 1.00     0.98
## Unemployed           0.62 0.60       1.00        -0.18       0.69 0.67     0.50
## Armed.Forces         0.46 0.45      -0.18         1.00       0.36 0.42     0.46
## Population           0.98 0.99       0.69         0.36       1.00 0.99     0.96
## Year                 0.99 1.00       0.67         0.42       0.99 1.00     0.97
## Employed             0.97 0.98       0.50         0.46       0.96 0.97     1.00
```

VIF

All the predictors have high VIF except Armed.Forces. For example, GNP has the highest VIF(1788.51) and it means that the se for the coef associated with GNP is 42.29078 times larger than it would have been without collinearity. High correlation, high VIF and high condition number indicate collinearity and we need to remove some variables.

```r
round(vif(x), dig=2)
```

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year
##       135.53      1788.51        33.62         3.59       399.15       758.98
```

```r
sqrt(1788.51)
```

```
## [1] 42.29078
```

Problem 4

Condition numbers

Since the condition number is $4.11 < 30$, there is no evidence of collinearity in data.

```
model.4 = lm(lpsa ~ .,data=prostate)
# condition number
x.4 = model.matrix(model.4)[,-1]
x.4 = x.4 - matrix(apply(x.4,2, mean), nrow(x.4), ncol(x.4), byrow=TRUE)
x.4 = x.4 / matrix(apply(x.4, 2, sd), nrow(x.4), ncol(x.4), byrow=TRUE)
apply(x.4,2,mean)
```

```
##        lcavol       lweight           age          lbph           svi
##  2.045230e-17 -4.350126e-16  4.115967e-16 -3.856742e-17  4.787256e-17
##           lcp       gleason         pgg45
##  1.319150e-17  4.583996e-17  1.443484e-17
```

```
apply(x.4,2,var)
```

```
##  lcavol lweight     age    lbph     svi     lcp gleason   pgg45
##       1       1       1       1       1       1       1       1
```

```
e.4 = eigen(t(x.4) %*% x.4)
sqrt(e.4$val[1]/e.4$val)
```

```
## [1] 1.000000 1.413945 1.839869 2.281503 2.613326 2.667878 3.552893 4.116210
```

Correlation between predictors

lcavol is relatively highly correlated with the response variable lpsa(0.73). pgg45 is relatively highly correlated with gleason(0.75) and they may be dependent on each other. We need to further examine them and decide whether remove one of them.

```
round(cor(prostate), dig=2)
```

```
##         lcavol lweight  age  lbph   svi   lcp gleason pgg45 lpsa
## lcavol    1.00    0.19 0.22  0.03  0.54  0.68    0.43  0.43 0.73
## lweight   0.19    1.00 0.31  0.43  0.11  0.10    0.00  0.05 0.35
## age       0.22    0.31 1.00  0.35  0.12  0.13    0.27  0.28 0.17
## lbph      0.03    0.43 0.35  1.00 -0.09 -0.01    0.08  0.08 0.18
## svi       0.54    0.11 0.12 -0.09  1.00  0.67    0.32  0.46 0.57
## lcp       0.68    0.10 0.13 -0.01  0.67  1.00    0.51  0.63 0.55
## gleason   0.43    0.00 0.27  0.08  0.32  0.51    1.00  0.75 0.37
## pgg45     0.43    0.05 0.28  0.08  0.46  0.63    0.75  1.00 0.42
## lpsa      0.73    0.35 0.17  0.18  0.57  0.55    0.37  0.42 1.00
```

VIF

Since all the predictors have very low VIF, there is no evidence of collinearity in data.

```
round(vif(x.4), dig=2)
```

```
##  lcavol lweight     age    lbph     svi     lcp gleason   pgg45
##    2.05    1.36    1.32    1.38    1.96    3.10    2.47    2.97
```