Homework 4

Gitanshu Munjal

November 4, 2014

1 Effects of collinearity on variances of betas and Yhats.

1.1

Make a MLR model with all explanatory variables and bd as the response variable. Report the code and the summary for the model. Determine is there is excessive collinearity. [10]

1.1.1 Code and Relevant Output

```
library(car)
library(HH)
library(leaps)
library(MASS)
library(DAAG)

birds<-read.table("C:\\Users\\gitanshu\\Desktop\\birds.csv",header=T,sep=",")
model<-lm(bd~grzinv+dist+height+peri+mammal+area+leg,birds)
summary(model)
```

```
Call:
lm(formula = bd ~ grzinv + dist + height + peri + mammal + area + leg, data = birds)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5284 -0.7912  0.2158  0.9800  3.0052

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.0481     0.1461 171.419   <2e-16 ***
grzinv        -1.7224     5.7661  -0.299    0.766
dist        -218.5275   499.9072  -0.437    0.663
height       196.6375   446.3263   0.441    0.661
peri         162.1747   368.9310   0.440    0.661
mammal        -8.2513    18.5205  -0.446    0.657
area          -1.0341     5.0022  -0.207    0.837
leg         -223.5453   507.8320  -0.440    0.661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 1.386 on 92 degrees of freedom
Multiple R-squared:  0.7428,    Adjusted R-squared:  0.7232
F-statistic: 37.96 on 7 and 92 DF,  p-value: < 2.2e-16
```

```
vif(model)
```

```
      grzinv          dist        height          peri        mammal          area           leg
    1617.970 12783783.739 12502844.195   6570401.610     15350.575      1105.641 14149460.854
```

Variance Inflation Factor (VIF) values over 5 are undesirable and over 10 indicate excessive collinearity. As demonstrated by the above results, **indeed there is excessive collinearity** in the current model as all VIF values are above 10.

1.2

Pretend that the first explanatory variable was not measured and create a second model without it. Report the same as in 3.1 [10]

1.2.1 Code and Relevant Output

```
modelminus<-lm(bd~dist+height+peri+mammal+area+leg,birds)
summary(modelminus)
```

```
Call:
lm(formula = bd ~ dist + height + peri + mammal + area + leg,
    data = birds)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4687 -0.8002  0.1927  0.9408  2.9999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.0549     0.1436 174.474  < 2e-16 ***
dist        -69.2525    13.0098  -5.323 7.04e-07 ***
height       63.3618    11.5795   5.472 3.75e-07 ***
peri         52.0097     9.5733   5.433 4.42e-07 ***
mammal       -2.7214     0.5311  -5.124 1.61e-06 ***
area          0.4590     0.1947   2.357   0.0205 *
leg         -71.9034    13.1469  -5.469 3.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 1.379 on 93 degrees of freedom
Multiple R-squared:  0.7426,    Adjusted R-squared:  0.726
F-statistic: 44.71 on 6 and 93 DF,  p-value: < 2.2e-16
```

```
vif(modelminus)
```

```
       dist      height        peri      mammal        area         leg
8743.655474 8498.801023 4467.834812   12.746025    1.691853 9576.817202
```

As demonstrated by the above results, **indeed there is excessive collinearity** in the current model as all VIF values except one (area) are above 10. However, the reduction in VIF values in comparison to the full model must be noted

1.3

Compare the estimated parameters in 3.1 and 3.2 and explain how and why they differ. Then, use the results as an example to briefly explain the effects of not including all variables in the model. [10]

Although both models found excessive collinearity, the estimates were much improved by removing the first explanatory variable as demonstrated by the **decrease in the std. error** for each of the estimates. Getting rid of that collinear variable's contribution to noise helped improve our ability to estimate by **reducing redundancy**. These results make the case for not including redundant variables and also for a stepwise removal of such redundancy, if found in the model, in order to **improve the signal to noise ratio**. Furthermore, removing such redundancy is advised as excessive collinearity coupled with extreme outliers could spell disaster for an analysis.

1.4

Start with all variables again and develop a good model using regsubsets() from the leaps package. Select the model with the minimum BIC. Report the model selected and its summary and analysis of variance. [10]

1.4.1 Code and Relevant Output

```
birds.best<-summary(best<-regsubsets(bd~., birds, nbest=1, nvmax=7, method=c("exhaustive")))
which.min(birds.best$bic)
birds.best$which[which.min(birds.best$bic),]
```

```
      X      grzinv        dist      height        peri      mammal        area         leg
  FALSE        TRUE        TRUE       FALSE       FALSE       FALSE        TRUE       FALSE
```

Based on the above stated results, **the best model with the minimum BIC includes 3 explanatory variables (grzinv, dist, area).**

```
bestmodel<-lm(bd~grzinv+dist+area,birds)
summary(bestmodel)
```

```
Call:
lm(formula = bd ~ grzinv + dist + area, data = birds)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4996 -0.9005  0.1404  0.9458  3.1988

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.0591     0.1435 174.676  < 2e-16 ***
grzinv        0.8147     0.1474   5.527 2.80e-07 ***
dist          1.5452     0.1454  10.630  < 2e-16 ***
area          1.1167     0.1518   7.356 6.39e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 1.381 on 96 degrees of freedom
Multiple R-squared:  0.7337,    Adjusted R-squared:  0.7254
F-statistic: 88.18 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
anova(bestmodel)
```

```
Analysis of Variance Table

Response: bd
          Df Sum Sq Mean Sq F value  Pr(>F)
grzinv     1    129   128.8    67.6 9.7e-13 ***
dist       1    272   272.3   142.9 < 2e-16 ***
area       1    103   103.1    54.1 6.4e-11 ***
Residuals 96    183     1.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1
```

1.5

Make predictions for bird abundance using the model from part a and the model from part d and two sets of predictor values: all predictors at their average value and all predictors equal average + 2 standard deviations. Obtain standard errors for all predictions. Compare predictions and s.e.'s and discuss the effects of collinearity and removal of redundant variables. How does the variance of Yhat change as function of predictor values and presence of collinearity? [10]

1.5.1 Code and Relevant Output

```
xvals<-as.data.frame(t(colMeans(birds[,c(3:9)])))
vars <- apply(birds[,c(3:9)],2,var)
xvals <- rbind(xvals,t(t(xvals)+2*sqrt(vars)))

predict(model, xvals, interval="confidence", level=0.90, se=T)
```

```
$fit
     fit    lwr   upr
1   25.2   24.9  25.4
2 -144.3 -808.7 520.1

$se.fit
      1       2
  0.139 399.844

$df
[1] 92

$residual.scale
[1] 1.39
```

```
predict(bestmodel,xvals[,c(1,2,6)],interval="confidence",level=0.9,se=T)
```

```
$fit
   fit  lwr  upr
1 25.2 24.9 25.4
2 31.9 31.1 32.6

$se.fit
    1     2
0.138 0.455

$df
[1] 96

$residual.scale
[1] 1.38
```

Comparing the above results makes it clear that there is not much difference in the ability to make predictions for the mean as the results using all predictors at their average values were nearly identical.

The real difference lies in making predictions about points far away from the mean as the results using all predictors at their average values summed + 2 standard deviations were found to be very different. The standard error for the fit of the reduced model was found to be higher than the full model which is as expected since we removed 4 explanatory variables. The fact that these variables were redundant is indicated in the observation that the standard error for the model fit is not impressively higher in their absence.

The confidence interval around the predicted values for bird abundance using the reduced model were found to be much narrower in comparison the confidence interval around the prediction for the full model. This observation makes the case for removing reducing excessive collinearity by removing redundant variables in order to improve our ability to estimate.

2 Regression diagnostics and validation.

2.1

Do a multiple linear regression of SalePr on YrHgt, FtFrBody, PrctFFB, Frame, BkFat, SaleHt and SaleWt.

2.1.1 Code and Relevant Output

```
bulls<-read.table("C:\\Users\\gitanshu\\Desktop\\Bulls.txt",header=T,sep=",")
bully<- lm(SalePr ~ YrHgt + FtFrBody + PrctFFB + Frame + BkFat + SaleHt + SaleWt, bulls)
summary(bully)
```

```
Call:
lm(formula = SalePr ~ YrHgt + FtFrBody + PrctFFB + Frame + BkFat +
    SaleHt + SaleWt, data = bulls)

Residuals:
   Min     1Q Median     3Q    Max
  -927   -280    -44    220   1613

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3156.894   4105.719   -0.77   0.4446
YrHgt          25.931    111.403    0.23   0.8166
FtFrBody       -2.201      1.071   -2.06   0.0437 *
PrctFFB       -30.098     26.717   -1.13   0.2639
Frame         360.317    172.903    2.08   0.0409 *
BkFat        2571.550    790.788    3.25   0.0018 **
SaleHt         84.397     64.091    1.32   0.1923
SaleWt          0.363      0.608    0.60   0.5523
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 461 on 68 degrees of freedom
Multiple R-squared:  0.504,     Adjusted R-squared:  0.453
F-statistic: 9.86 on 7 and 68 DF,  p-value: 2.01e-08
```

2.2

Test the asumption of normality. [5]

2.2.1 Code and Relevant Output

```
shapiro.test(residuals(bully))
qqplot(bully)
```

```
        Shapiro-Wilk normality test

data:  residuals(bully)
W = 0.925, p-value = 0.0002499
```

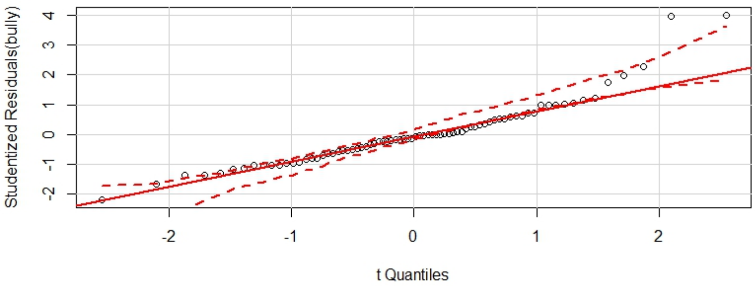


Figure 1: Quantile plot of residuals to test for normality

The significant p-value for the above produced Shapiro-Wilk test indicates that assumption of normality has been violated i.e. the **residuals are NOT normally distributed.** One possible cause for this violation could be extreme outliers (we see two such points lying outside the confidence interval). The other possibility could be lack of homogeneity of variances which might need a transformation to remedy.

2.3

Test the homogeneity of variance. In addition to other potential checks, use Breed as grouping variable. [10]

2.3.1 Code and Relevant Output

```
residualPlots(bully)
```

```
           Test stat Pr(>|t|)
YrHgt          0.677    0.501
FtFrBody      -0.715    0.477
PrctFFB       -1.565    0.122
Frame          0.528    0.599
BkFat         -1.197    0.235
SaleHt         0.146    0.884
SaleWt        -0.256    0.799
Tukey test     4.441    0.000
```

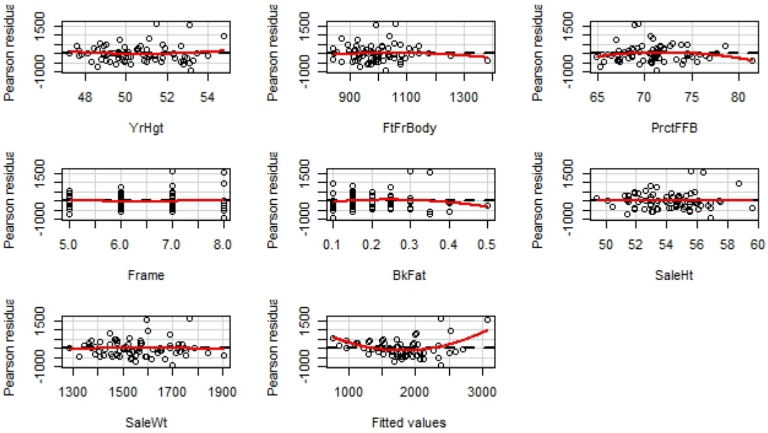


Figure 2: Residual plots

The above resuls from the residual plots indicate that indeed all variances are homogenuous as the p-values for the supporting test,”Pr(>|t|)”, were found to be not significant at the 95 percent signficance level. However, the plot for Residual Vs Fitted values was found to be non-linear indicating that the assumption regarding additivity might be violated (again, possibly remedied by removal of redundancy and/or transformation).

```
leveneTest(residuals(bully)~as.factor(Breed),bulls)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2    0.51    0.6
      73
```

The levene’s test for homogeneity of variances using Breed as a grouping variable was found to be not significant at the 95 percent significant level indicating that the variances were homogenuous.

2.4

Inspect the added variable or ”leverage” plots to test for lack of linearity. [5]

2.4.1 Code and Relevant Output

```
avPlots(bully)
```

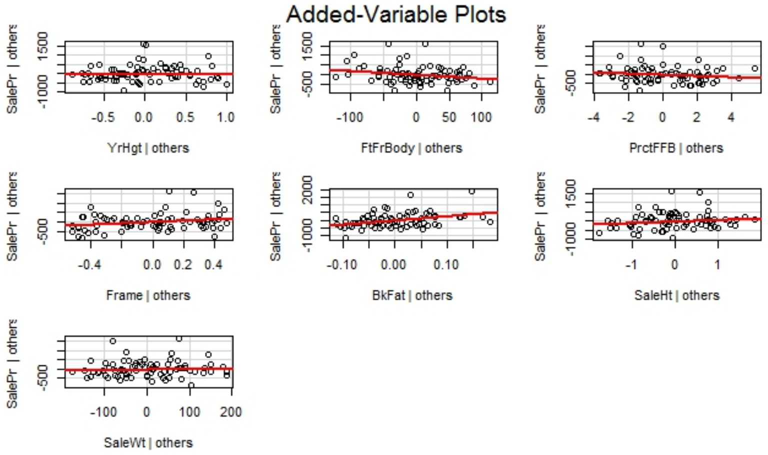


Figure 3: Added Variable plots

The added variable plots indicated **no departure from linearity.**

2.5

Determine if there is excessive collinearity. [10]

2.5.1 Code and Relevant Output

```
vif(bully)
```

```
   YrHgt  FtFrBody   PrctFFB     Frame     BkFat    SaleHt    SaleWt
   13.13      3.48      2.69      9.06      1.77      5.83      2.20
```

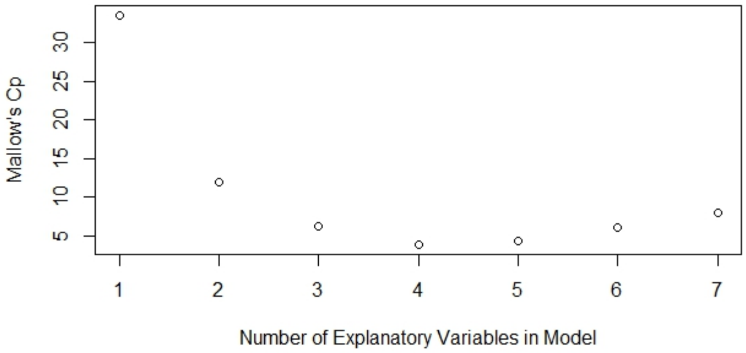
As mentioned earlier, VIF values above 5 are undesirable and above 10 are excessive. Using that definition, indeed **there is evidence for excessive collinearity.**

2.6

Use the leaps package to select a good reduced model as necessary. [10]

2.6.1 Code and Relevant Output

```
bestbulls<-summary(best<-regsubsets(SalePr ~ .-Breed, bulls, nbest=1,
nvmax=7, method=c("exhaustive")))

plot(bestbulls$cp~1+as.numeric(rownames(bestbulls$which)),
xlab="Number of Explanatory Variables in Model",ylab="Mallow's Cp")
```



```
bestbulls$which[which.min(bestbulls$cp),]
```

```
YrHgt    FtFrBody     PrctFFB       Frame       BkFat      SaleHt      SaleWt
FALSE        TRUE       FALSE        TRUE        TRUE        TRUE       FALSE
```

The best model was thus picked by minimizing Mallow's Cp. This model included four explanatory variables (FtFrBody, Frame, BkFat, and SaleHt).

```
betterbully<-update(bully, ~ .- YrHgt -PrctFFB - SaleWt)
summary(betterbully)
```

```
Call:
lm(formula = SalePr ~ FtFrBody + Frame + BkFat + SaleHt, data = bulls)

Residuals:
   Min      1Q  Median      3Q     Max
-841.4 -329.9   -40.1   193.2  1687.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4511.119   2047.267   -2.20   0.0308 *
FtFrBody       -2.745      0.802   -3.42   0.0010 **
Frame         375.466     95.807    3.92   0.0002 ***
BkFat        3157.195    616.125    5.12  2.5e-06 ***
SaleHt        110.763     49.709    2.23   0.0290 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 457 on 71 degrees of freedom
Multiple R-squared:  0.49,      Adjusted R-squared:  0.462
F-statistic: 17.1 on 4 and 71 DF,  p-value: 7.49e-10
```

2.7

Perform a k-fold validation. Report your choice of k, the average MSE and compare with the original MSE. [10]

2.7.1

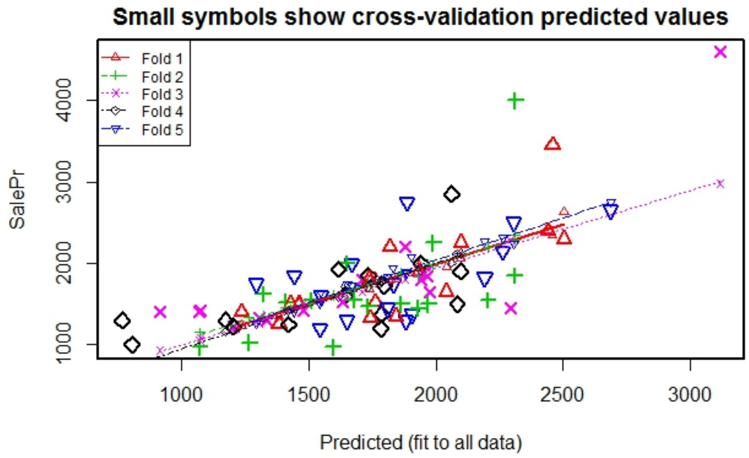
For the analysis here, a K value of 5 was chosen so as to allow sufficient observations for each K group while not compromising on number of groups.

2.7.2 Code and Relevant Output

```
cv.lm(df=bulls, betterbully,m=5, seed=floor(1000*runif(1)))
```

```
Analysis of Variance Table

Response: SalePr
          Df   Sum Sq Mean Sq F value  Pr(>F)
FtFrBody   1   302808  302808    1.45   0.233
Frame      1  7981585 7981585   38.19 3.6e-08 ***
BkFat      1  4950958 4950958   23.69 6.6e-06 ***
SaleHt     1  1037571 1037571    4.97   0.029 *
Residuals 71 14837102  208973
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1



fold 1
Observations in test set: 15
                7   10   16     17     22     23   30     37   45   55   58   59   65   68   69
Predicted    2097 1763 2503 1933.1 1460.8 2442.3 1235 1428.7 1740 1738 2458 2042 1817 1383 1842
cvpred       2068 1833 2631 1983.9 1521.4 2414.9 1239 1424.8 1800 1687 2344 1959 1788 1357 1811
SalePr       2250 1525 2300 1900.0 1500.0 2400.0 1400 1500.0 1325 1800 3450 1650 2200 1250 1350
CV residual   182 -308 -331  -83.9  -21.4  -14.9  161   75.2 -475  113 1106 -309  412 -107 -461

Sum of squares = 2227846    Mean square = 148523    n = 15

fold 2
Observations in test set: 16
                2    3    8   14     25   33   40   44   47   49   51     62   64   66   67
76
Predicted    1990 1321 2313 2204 1408.8 1648 1865 1595 1266 1072 1929 1508.9 1733 2309 1679 1971
cvpred       1973 1350 2293 2199 1452.1 1656 1888 1610 1333 1151 1837 1501.8 1752 2353 1693 1993
SalePr       2250 1625 4000 1550 1525.0 2000 1500  975 1025  975 1450 1550.0 1475 1850 1550 1500
CV residual   277  275 1707 -649   72.9  344 -388 -635 -308 -176 -387   48.2 -277 -503 -143 -493

Sum of squares = 5036765    Mean square = 314798    n = 16

fold 3
Observations in test set: 15
                1    4      6   18   19   24   26   32     34     41   42   54     56     61
74
Predicted    1880 3119 1204.5 1072 1976 1074 1941  915 1334.9 1304.0 1710 2291 1632.4 1967.2 1480.0
cvpred       1808 2983 1181.8 1053 1933 1094 1840  930 1329.9 1262.4 1646 2296 1612.9 1936.6 1515.6
SalePr       2200 4600 1225.0 1400 1650 1425 1800 1400 1300.0 1325.0 1800 1450 1525.0 1850.0 1425.0
CV residual   392 1617   43.2  347 -283  331  -40  470  -29.9   62.6  154 -846  -87.9  -86.6
-90.6

Sum of squares = 4069311    Mean square = 271287    n = 15

fold 4
Observations in test set: 15
               12   20   21   28   31   36     38   46   48   52     53   57   60     70   75
Predicted    2062 2085 1787 1643 1174  767 1204.4 1733  805 1784 1940.5 1617 2096 1797.1 1421
cvpred       2009 2117 1787 1720 1164  717 1182.4 1711  677 1774 1928.2 1596 2125 1795.9 1419
SalePr       2850 1500 1375 1600 1300 1300 1225.0 1850 1000 1200 2000.0 1925 1900 1725.0 1250
CV residual   841 -617 -412 -120  136  583   42.6  139  323 -574   71.8  329 -225  -70.9 -169
```

```
Sum of squares = 2284133    Mean square = 152276    n = 15

fold 5
Observations in test set: 15
                5      9    11    13    15    27    29    35    39    43    50    63    71    72    73
Predicted    2264 1547.7 1441 2688 1670 2307 1883 1651 1884 1904 1832 2192 1294 1810 1542
cvpred       2310 1640.5 1402 2758 1725 2241 1874 1737 1890 2082 1945 2269 1268 1841 1502
SalePr       2150 1600.0 1850 2650 2000 2500 1300 1300 2750 1375 1750 1825 1750 1450 1200
CV residual -160  -40.5  448 -108  275  259 -574 -437  860 -707 -195 -444  482 -391 -302

Sum of squares = 2854132    Mean square = 190275    n = 15

Overall (Sum over all 15 folds)
      ms
216739
```



The average MSE calculated from the above validation (216739) was found to be 1.03 times the original MSE (208973) indicating that our reduced model is indeed good. Within reasonable limitations, our reduced model is able to make roughly accurate predictions.