Stat Methods I - Homework 6

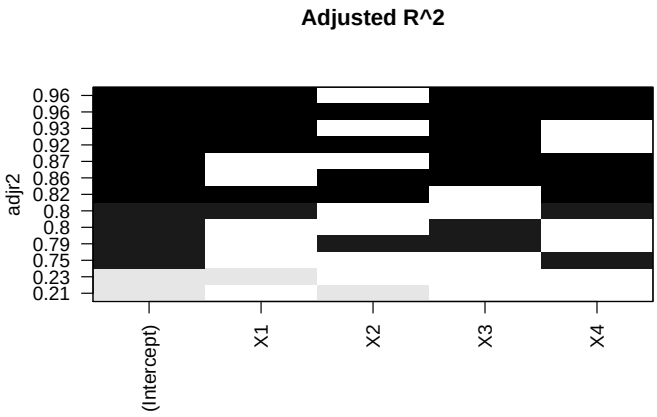
Justin Reising

November 27, 2018

9.11) Refer to Job Proficiency.

(a) Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression model according to the $R^2_{a,p}$ criterion.

```
# From the "leaps"" package
job.reg.subsets<-regsubsets(Score~X1+X2+X3+X4,data=job,nbest=4)
plot(job.reg.subsets, scale = "adjr2", main = "Adjusted R^2")
```



This plot was constructed using the best subsets algorithm in the "leaps" package in R. The way to read this plot in our model selection is there is a black block where the predictor variable being selected for the model with the corresponding adjusted R squared on the y-axis. So the 4 best models are as follows:

1. For $R^2_a = 0.96$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$
2. For $R^2_a = 0.96$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + +\beta_3 X_3 + \beta_4 X_4$
3. For $R^2_a = 0.93$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_3 X_3$
4. For $R^2_a = 0.92$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

(b) Since there is relatively little difference in $R^2_{a,p}$ for the four best subset models, what other criteria would you use to help in the selection of the best model? Discuss.

As a rule of thumb, we want to consider the model with the least number of predictors that maintain high adjusted R squared. As these are are quite similar in their adjusted R squared values, we may want to consider the Mallows' Cp criterion in addition to the adjusted R squared to investigate which model demonstrates the least amount of bias.

9.21) Refer to Job Proficiency. Problems 9.10 and 9.18. To assess internally the predictive ability of the regression model indentified in 9.18, compute the PRESS statistic and compare it to SSE. What does the comparison suggest about the validity of MSE as an indicator of the predictive ability of the fitted model?

$$\hat{Y} = -124.2 + 0.2963X_1 + 1.357X_3 + 0.5174X_4 \quad (9.18)$$

```
job.red.reg<- lm(Score~X1+X3+X4)
PRESS.statistic <- sum( (resid(job.red.reg)/(1-hatvalues(job.red.reg)))^2 )
print(paste("PRESS statistic= ", round(PRESS.statistic,2)))
```

```
## [1] "PRESS statistic=  471.45"
```

```
job.anova<-anova(job.red.reg)
SSE<- job.anova$`Sum Sq`[4]
PRESS.statistic/SSE
```

```
## [1] 1.353981
```

Note that the ratio of the $\frac{PRESS}{SSE}$ is approximately 1.3 which would indicate that the model is a good fit. Also note that the Adjusted R squared is approximately 0.95 which futher indicates that much of the variation is accounted for in the model.

10.2) A researcher stated: "One good thing about added-variable plots is that they are extremely useful for identifying model adequacy even when the predictor variables are not properly specified in the regression model." Comment.

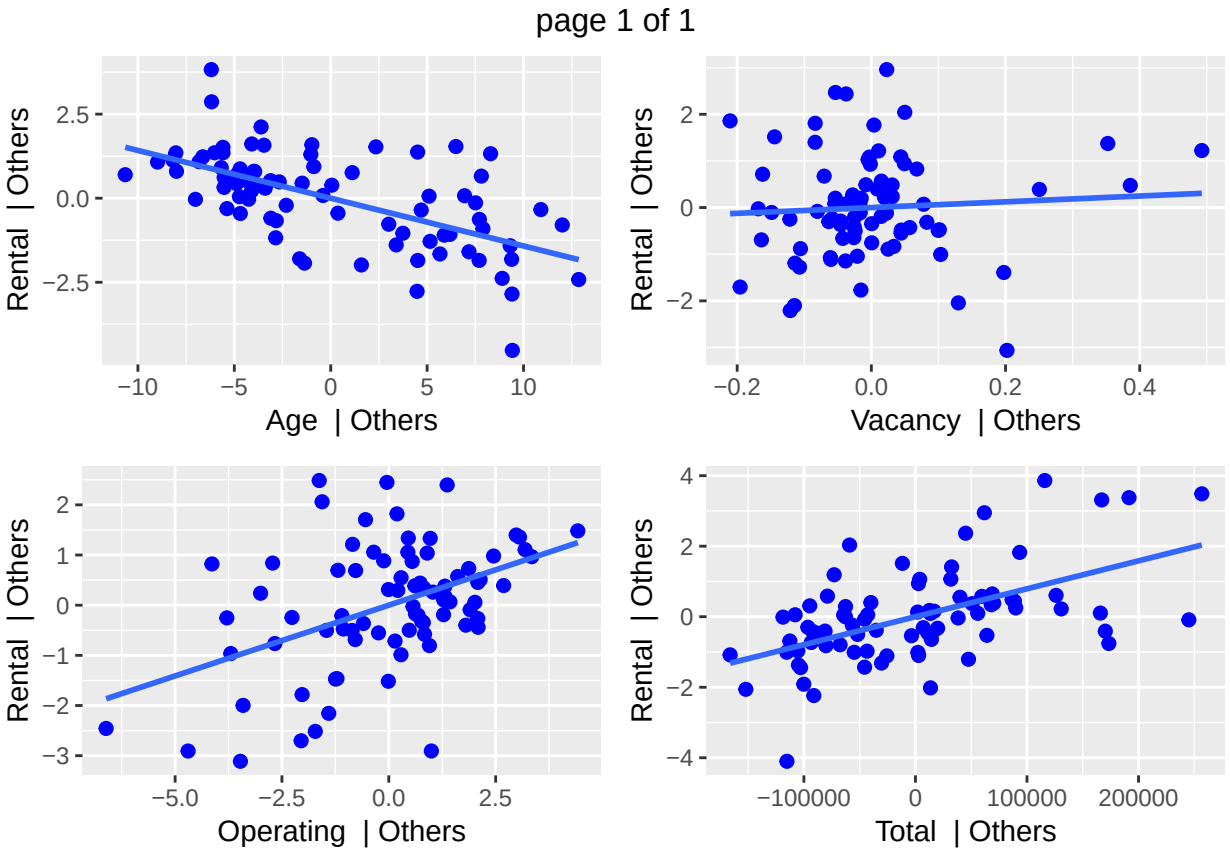
While the added-variable plots suggests the nature of the functional relation in which a predictor variable should be added to the regression model, it does not provide an analytical expression of the relation. Since added-variable plots for several variables are all concerned with marginal effects only, they may not be effective when the relations of predictor variables to the response variable are complex (Like Principal Components or Interaction Terms).

10.3) A student suggested: "If extremely influential outlying cases are detected in a data set, simply discard these cases from the data set." Comment.

One should not "simply" remove outliers without careful consideration. This is a very subjective process that depends on the domain and type of data. For example, if we were looking at all at-bats for Major League baseball players in a single season, we would expect to see significant outliers for batting average consisting primarily of Nation League pitchers that typically have less than 100 at bats in a season; and even less if they are an American League pitcher. If we remove them, then we restrict the scope of our study by say, considering players with a minimum of 200 plate appearances in a season. The criteria for removing outliers can heavily depend on the context of the data.

10.8) Refer to Commercial Properties.

(a) Prepare an added variable plot for each of the predictor variables.



(b) Do your plots in part (a) suggest that the regression relationships in the fitted regression function in problem 6.18c are inappropriate for any of the predictor variables? Explain.

It appears that Age and Total partial regression plots are indicating to be entered into the model in a linear way. Operating looks slightly flat in pattern but to have a bit scatter, however, Vacancy appears to be very flat in pattern which would indicate that Vacancy may not be needed (Which we described in 6.18).

10.12) Refer to Commercial Properties.

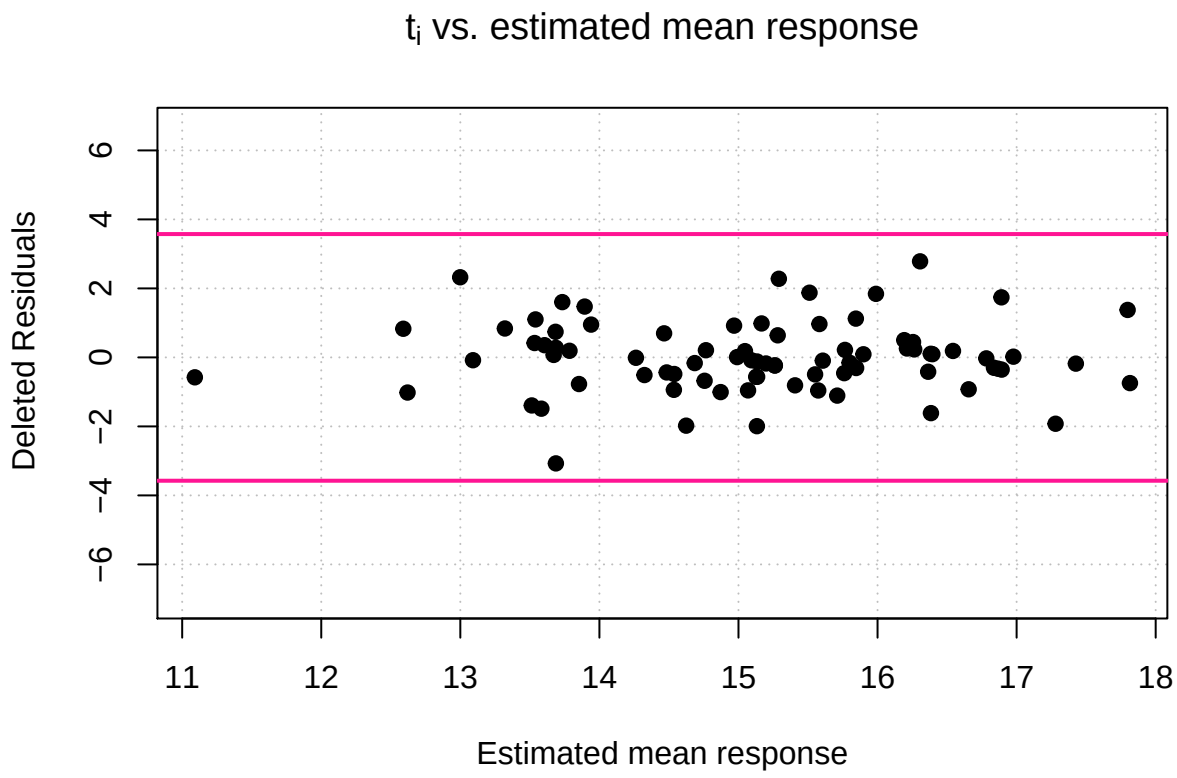
(a) Obtain the studentized deleted (externally studentized) residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.01$. State the decision rule and conclusion.

```
outlierTest(comm.reg.full, cutoff = .01)
```

```
## No Studentized residuals with Bonferonni p < 0.01
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 6 -3.072105          0.0029614      0.23988
```

From a preloaded R package, we can quickly apply the Bonferroni Outlier Test and see that no externally studentized residuals outliers. The plot below can further illustrate this as none of the the residual points fall outside of the critical value bands.

```
## [1] 2.642983
```



(b) Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

```
X<-cbind(rep(1,length(Age)),Age,Operating,Vacancy,Total)
H<- X%*%solve(t(X)%*%X)%*%t(X)
#diag(H)

#Similarly with built in function
Comm.Inf.Measures<-influence.measures(comm.reg.full)
#Comm.Inf.Measures$is.inf[,9] -> diag(H)
which(Comm.Inf.Measures$is.inf[,9] == "TRUE")
```

```
##  3  8 61
##  3  8 61
```

(d) Cases 61, 8, 3, and 53 appear to be outlying X observations, and cases 6 and 62 appear to be outlying Y observations. Obtain the $DIFFITS$, $DFBETAS$, and Cook's Distance values for each case to assess its influence. What do you conclude?

```
potential.outliers<- Comm.Inf.Measures$infmat[c(3,8,61,53,6,62), -c(7,9)]
potential.outliers
```

```
##         dfb.1_      dfb.Age     dfb.Oprt    dfb.Vcnc     dfb.Totl
## 3  -0.23178572 -0.155328321  0.236413643  0.10078041 -0.011493947
## 8  -0.01421023 -0.007197892  0.003014250  0.09551927  0.012599064
## 61 -0.05541528  0.024248530 -0.007608429  0.54571270  0.003819789
## 53 -0.01962803 -0.023983533 -0.024340442  0.41796384  0.048967863
## 6   0.19511546 -0.564851536 -0.176722251 -0.61719137  0.448172929
## 62  0.27581469 -0.333496175 -0.259470370  0.06272880  0.405078144
##         dffit      cook.d
## 3  -0.2842805 0.016306196
## 8   0.1164136 0.002744609
## 61  0.6387208 0.081662174
## 53  0.5252264 0.054981894
## 6  -0.8735488 0.137366521
## 62  0.6903319 0.087535890
```

```
#DFFITS
which(abs(potential.outliers[,6]) > 2*sqrt(p/n))
```

```
## 61 53  6 62
##  3  4  5  6
```

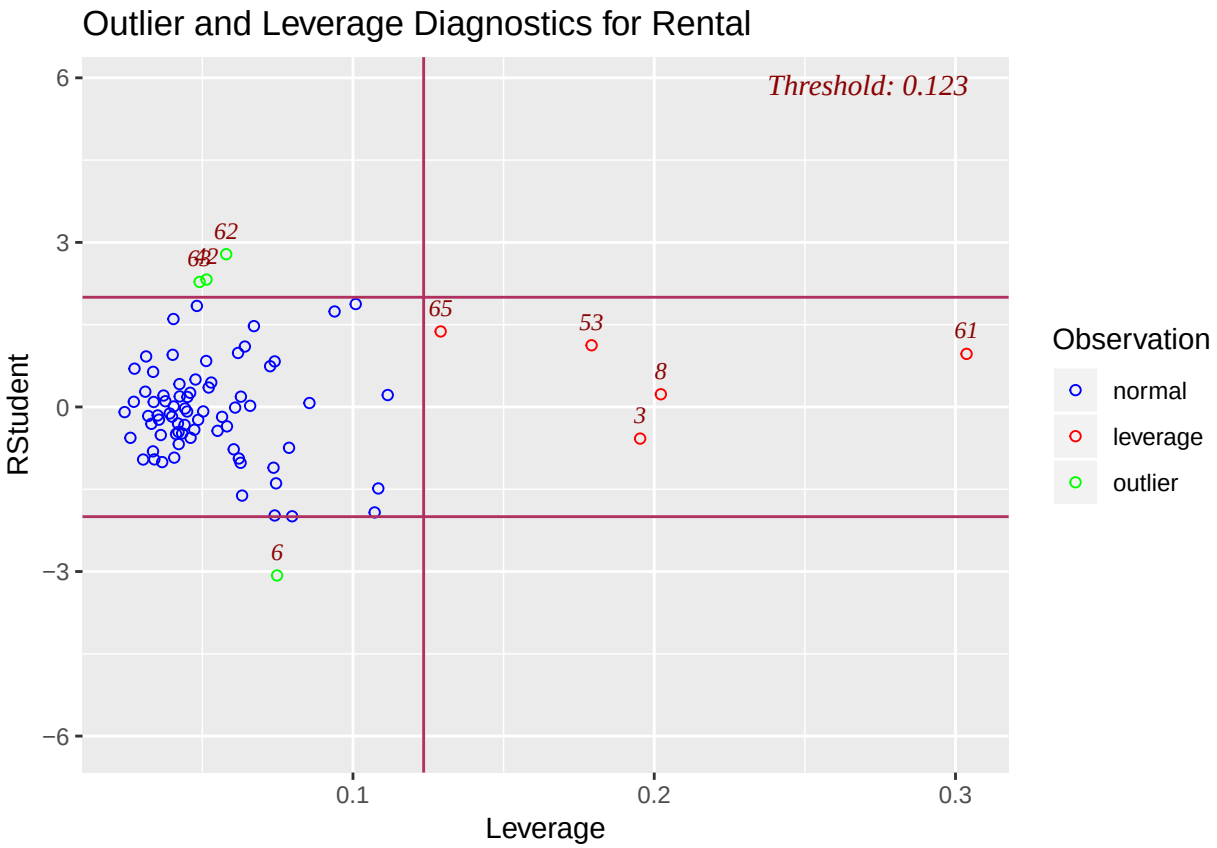
The code above chooses which DFFITS values are $> \sqrt{p/n}$ and we can see that observations 6,53,61,62 are influential.

```
#DFBETAS (For Large Data Set 81 observations)
which(abs(potential.outliers[c(1:5),6]) > 2*sqrt(n))
```

```
## named integer(0)
```

This code shows which DFBETAS values are $> 2\sqrt{n}$ which would indicate these observations are influential, but in this case we do not see that occur.

```
ols_plot_resid_lev(comm.reg.full)
```



In this generalized leverage and outlier plot above, we can see that observation 3,8,53,61, and 65 have significant leverage while 6 and 62 are outliers.

10.26) Prove (9.11) using (10.27) and exercise 5.31.

$$\sum_{i=1}^{n} \sigma^2\{\hat{Y}_i\} = p\sigma^2 \quad (9.11)$$

$$0 \leq h_{ii} \leq 1 \quad \sum_{i=1}^{n} h_{ii} = p \quad (10.27)$$

Pf

$$Recall \quad Var(\hat{\vec{Y}}) = \sigma^2 \vec{H} \implies Var(Y_i) = \sigma^2 h_{ii}$$

$$\begin{aligned} \sum_{i=1}^{n} \sigma^2\{\hat{Y}_i\} &= \sum_{i=1}^{n} \sigma^2 h_{ii} \\ &= \sigma^2 \sum_{i=1}^{n} h_{ii} \\ &= p\sigma^2 \end{aligned}$$