Assignement #2 SEIS 763_02

Josh Janzen

2/23/2017

1 & 2: read in data

```
library(car) # companian applied regression package
# unable to load from C:\tmp as that directory doesn't exist on Mac
setwd("/Users/a149174/UST_GPS/seis_763/r/seis_763_machine_learning/assignments")

data <- read.csv("patients.csv", head=T, sep=',', skip = 0)
```

3: Build Linear Model

```
model <- lm(Systolic ~ Age + Gender + Height + Location + SelfAssessedHealthStatus + Smo
ker + Weight, data=data)
```

4: Thetas

```
##                                        Estimate  Std. Error
## (Intercept)                         88.65811329 18.22461158
## Age                                  0.08025966  0.06699892
## Gender'Male'                        -1.47939073  3.26574545
## Height                               0.46962059  0.25390819
## Location'St. Mary's Medical Center' -0.85650078  1.29798791
## Location'VA Hospital'               -1.73484051  1.13322534
## SelfAssessedHealthStatus'Fair'      -2.75096823  1.51063322
## SelfAssessedHealthStatus'Good'       0.58637873  1.17832929
## SelfAssessedHealthStatus'Poor'       0.45934283  1.67618555
## Smoker                               9.67308711  1.04590413
## Weight                              -0.01341834  0.05837056
```

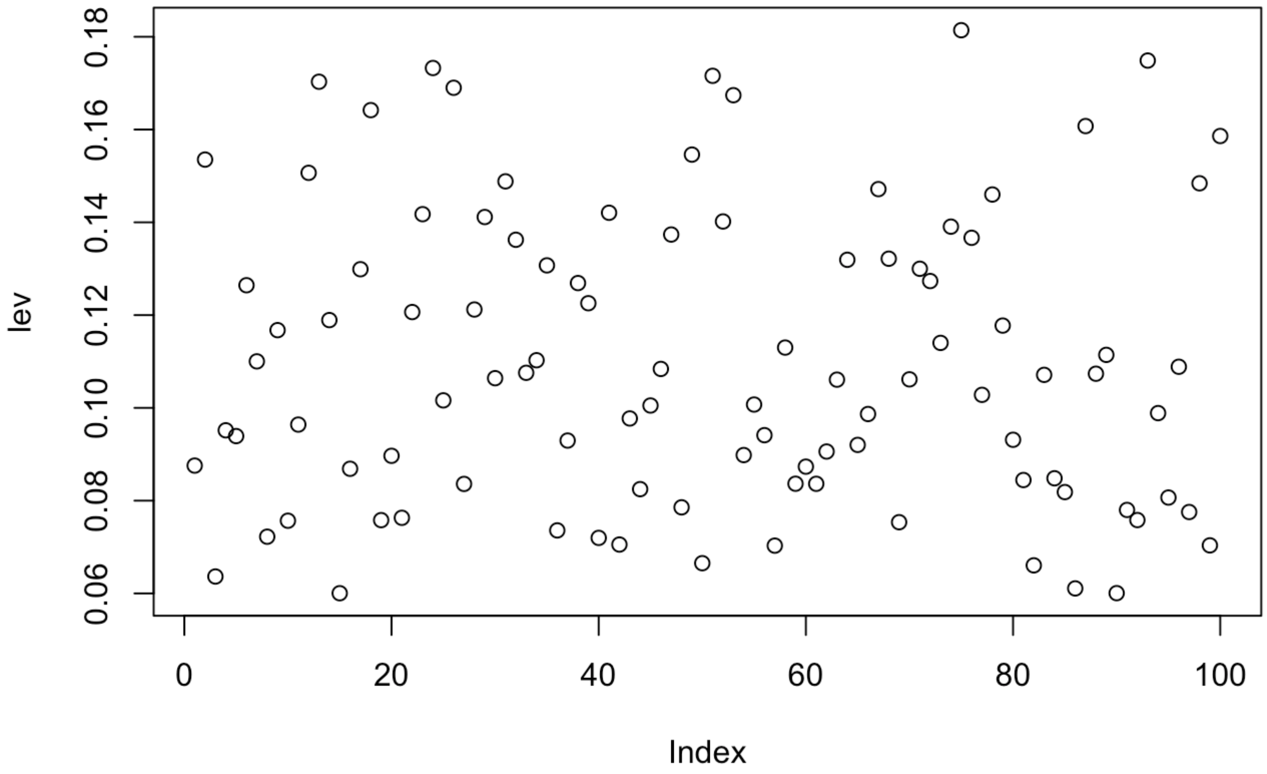
5: Theta Interpretation

- For continuous predictors (Age, Height, Weight), if all other variables are held constaint, 1 unit of a theta will result in that thetas value change in Systolic.
- For categorical predictors (Gender, Location, SelfAssessedHealthStatus, Smoker), if all other variables are held constaint, a change between in Y will be average difference accross category values.

6: Identifying Outlier

First, I want to look at Leverage, by plotting it, and then sorting for highest values

```
lev <- hat(model.matrix(model))
plot(lev)
```



```
# sort with give values sorted with highest leverage
head(sort(lev, decreasing=T))
```

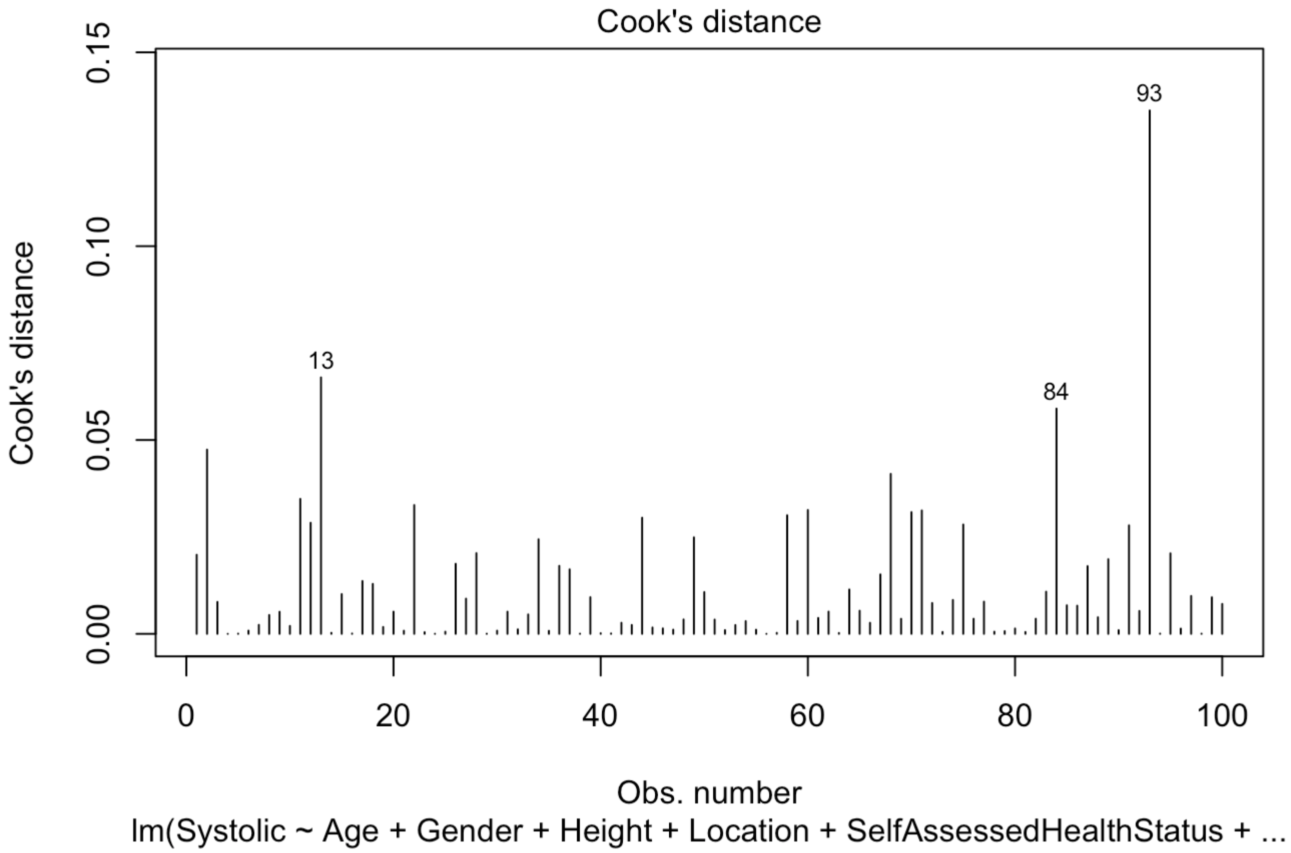
```
## [1] 0.1814159 0.1748754 0.1732668 0.1715823 0.1703090 0.1690049
```

```
# order will provide the index of the obs.  1: obs 75, 2: 93..
head(order(lev, decreasing=T))
```

```
## [1] 75 93 24 51 13 26
```

Next, I want to look at Cook’s Distance, by plotting it, and then sorting for highest values

```
# Cook's Distance plot
cutoff <- 4/((nrow(model)-length(model$coefficients)-2))
plot(model, which=4, cook.levels=cutoff)
```



```
# sort with give values sorted with highest distance
head(sort(cooks.distance(model), decreasing=T))
```

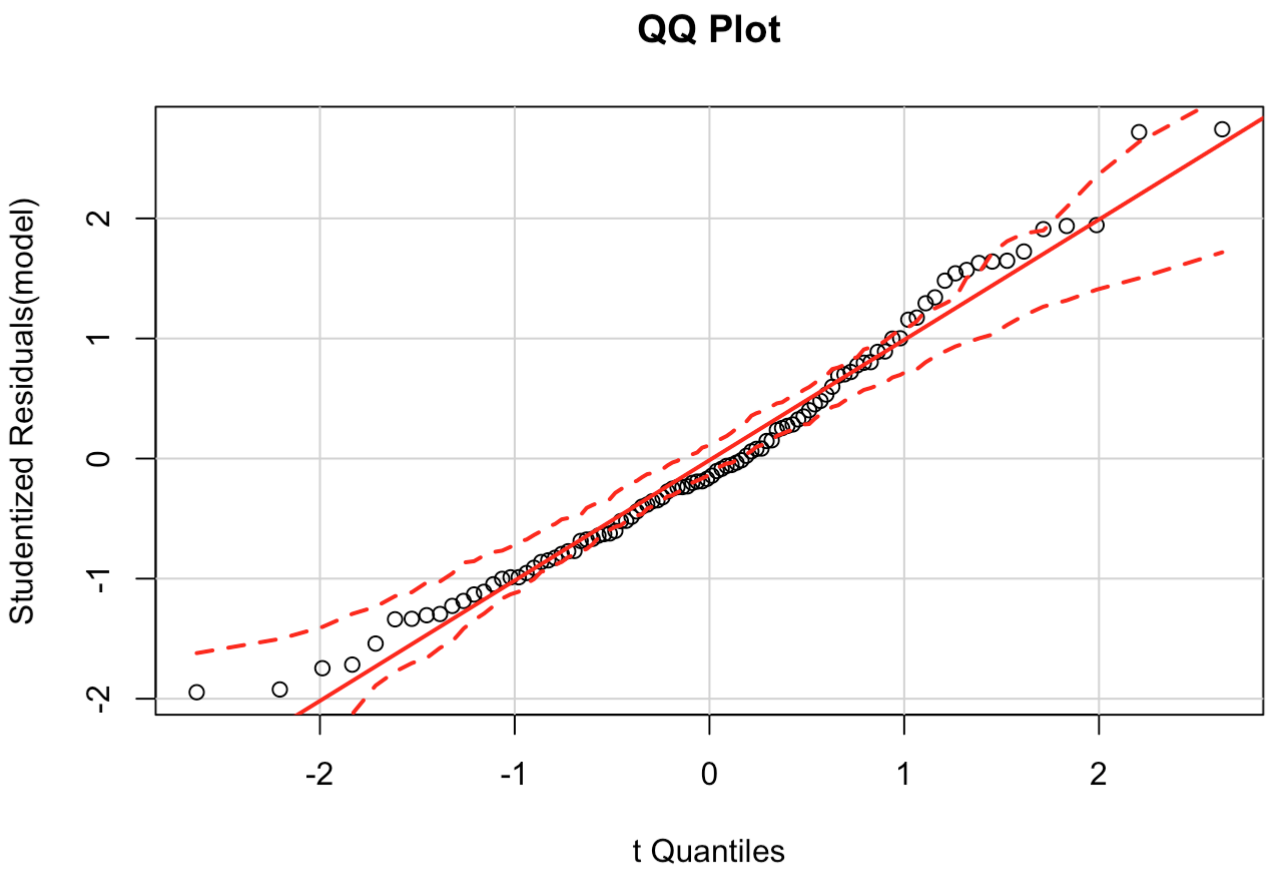
```
##         93         13         84          2         68         11 
## 0.13500953 0.06614635 0.05810115 0.04751771 0.04125499 0.03480392
```

```
# order will provide the index of the obs.  1: obs 93, 2: 13...
head(order(cooks.distance(model), decreasing=T))
```

```
## [1] 93 13 84  2 68 11
```

Next, Normal Probability of Residuals

```
# Normality of Residuals
qqPlot(model, main="QQ Plot")
```

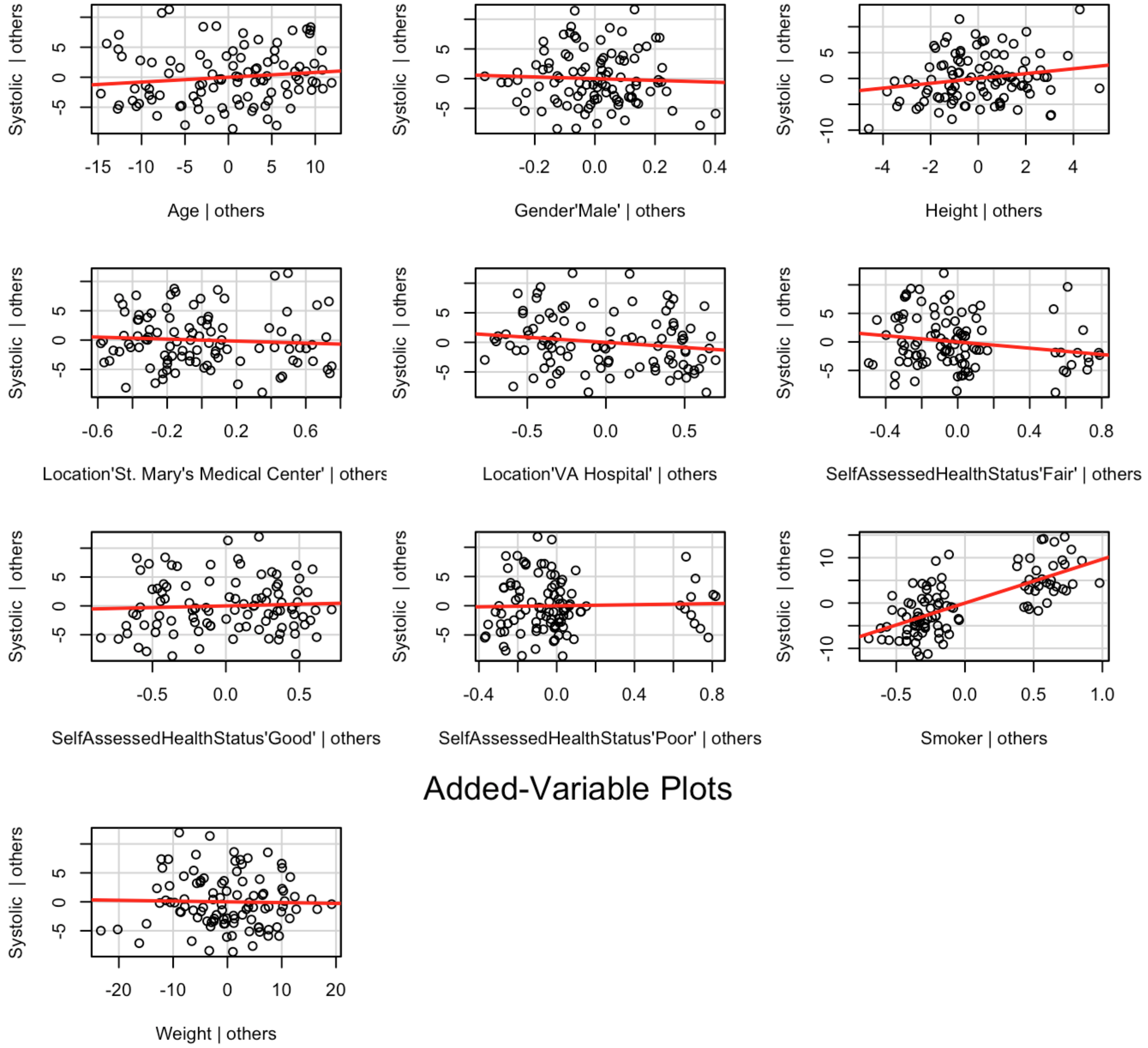


Based that observation 93 is 2nd in Leverage and 1st in Cook’s Distance, I’d recommend removing

7: Identifying Useless Features (Predictor)

Added Variable Plots

```
# added variable plots (if line is near horizontal, then the variable is insignificant)
avPlots(model)
```


Added-Variable Plots

Weight has flattest line for leverage, could be a good candidate for removal

Model Summary

```
##
## Call:
## lm(formula = Systolic ~ Age + Gender + Height + Location + SelfAssessedHealthStatus +

##     Smoker + Weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6277 -3.1293 -0.6898  3.1426 11.8390
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         88.65811   18.22461   4.865 4.92e-06
## Age                                  0.08026    0.06700   1.198   0.2341
## Gender'Male'                        -1.47939    3.26575  -0.453   0.6516
## Height                               0.46962    0.25391   1.850   0.0677
## Location'St. Mary's Medical Center' -0.85650    1.29799  -0.660   0.5110
## Location'VA Hospital'               -1.73484    1.13323  -1.531   0.1293
## SelfAssessedHealthStatus'Fair'      -2.75097    1.51063  -1.821   0.0720
## SelfAssessedHealthStatus'Good'       0.58638    1.17833   0.498   0.6200
## SelfAssessedHealthStatus'Poor'       0.45934    1.67619   0.274   0.7847
## Smoker                               9.67309    1.04590   9.249 1.15e-14
## Weight                              -0.01342    0.05837  -0.230   0.8187
##
## (Intercept)                         ***
## Age
## Gender'Male'
## Height                              .
## Location'St. Mary's Medical Center'
## Location'VA Hospital'
## SelfAssessedHealthStatus'Fair'      .
## SelfAssessedHealthStatus'Good'
## SelfAssessedHealthStatus'Poor'
## Smoker                              ***
## Weight
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.713 on 89 degrees of freedom
## Multiple R-squared:  0.5569, Adjusted R-squared:  0.5071
## F-statistic: 11.19 on 10 and 89 DF,  p-value: 3.894e-12
```

Weight has heighest P-value, then Gender, Location

Based on this, Recommend removing Weight

Extra: Re-run model removed outlier, to see change in Adj R-Squared verus original model

```
## [1] 0.03285454
```

Extra: Re-run model removed outlier and Weight, to see change in Adj R-Squared verus just removing outlier

```
## [1] 0.005097908
```

Both removing outlier and Weight increased Adj R-Squared