# HANDOUT: Stats 230 - Added Variable Plots

P.B. Matheson adapted from Shu-Min Liao

Blood Pressure Example

**Review of Models Being Considered for Blood Pressure (BP)**

When last we met, we fitted four models (named `fm1`, `fm.full`, `fm2`, and `fm3`) using the BP (BloodPressure) data and compared them via nested F-tests. Below is a reenactment.

fm1 <- lm(BP ~ Weight, data = BPdata)
fm.full <- lm(BP ~ ., data = BPdata)
fm2 <- lm(BP ~ Weight + Age + Dur + Stress, data = BPdata)
fm3 <- lm(BP ~ Weight + Age, data = BPdata)

.
.

```
anova(fm1, fm3, fm2, fm.full)     #Nested F tests comparing multiple models at once
```

```
## Analysis of Variance Table
##
## Model 1: BP ~ Weight
## Model 2: BP ~ Weight + Age
## Model 3: BP ~ Weight + Age + Dur + Stress
## Model 4: BP ~ Age + Weight + BSA + Dur + Pulse + Stress
##   Res.Df  RSS Df Sum of Sq      F  Pr(>F)
## 1     18 54.5
## 2     17  4.8  1      49.7 299.72 2.3e-10 ***
## 3     15  4.5  2       0.3   0.84  0.4536
## 4     13  2.2  2       2.4   7.20  0.0078 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1) How do we interpret this output (review!)?

**Evaluate BSA (Body Surface Area) as an added predictor**

We are now asking is it helpful to add `BSA` as a predictor to the original model `fm3`. We want to understand what additional information in Y is captured knowing `BSA` but NOT in the others (i.e. `Weight` and `Age`). We can create an **added variable plot** to see this. To do this we do 2 regressions and an error by error plot.

- Step 1: Regress `BP` on `Age` and `Weight`. This is the original model: $Y \sim X_1 + X_2$ and record the residuals $e_1$

```
fm.e1 <- lm(BP ~ Age + Weight, data = BPdata)     #yes, this is the same as `fm3`
```

The residuals in Step 1 (denoted as $e_1$) represent the amount of variation in `BP` that is leftover (unaccounted for) after accounting for `Age` and `Weight`. These y residuals are from the original model that you already recognize as residuals and show us the information in BP that isn't being accounted for by knowing Age and Weight . Referred to here as y residuals.

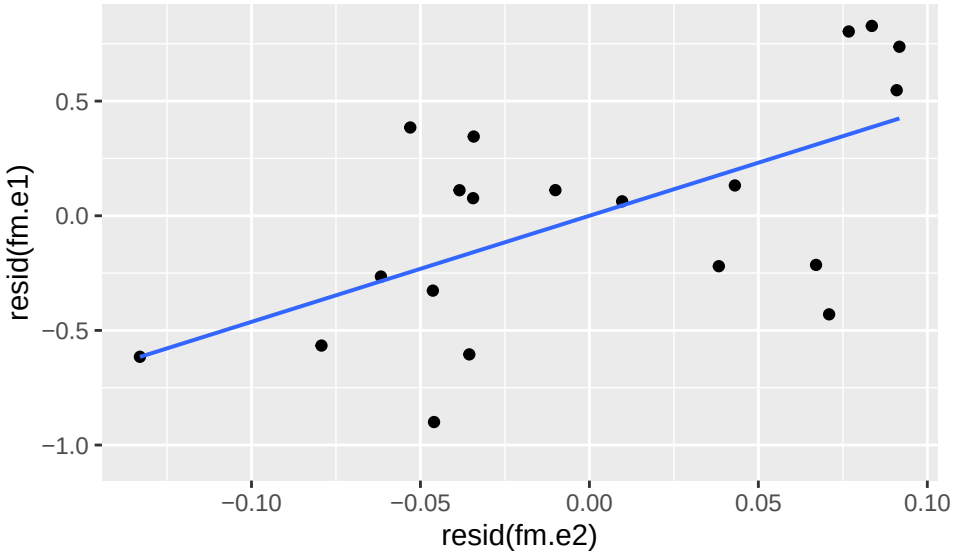
- Step 2: Regress `BSA` on `Age` and `Weight`. Regressing new predictor on old predictors: $X_3 \sim X_1 + X_2$ and record the residuals $e_2$

```
fm.e2 <- lm(BSA ~ Age + Weight, data = BPdata)
```

The residuals in Step 2 (denoted as $e_2$) represent the variation in `BSA` that is NOT captured by the variation in `Age` and `Weight`. Think about these residuals (called x residuals) as the unique information contained in BSA that is not contained in Age and Weight.

- Step 3: Plot the residuals of the two models against each other with $e_1$ as the y-axis and $e_2$ as the x-axis.

```
gf_point(resid(fm.e1) ~ resid(fm.e2)) %>%
  gf_lm()
```




The plot from Step 3 is assessing if the leftover variability in BP after using Age and Weight aka: `fm3` can be further explained by the information *exclusively* contained in `BSA`.

2) What do you think? What does this AV plot tell us?

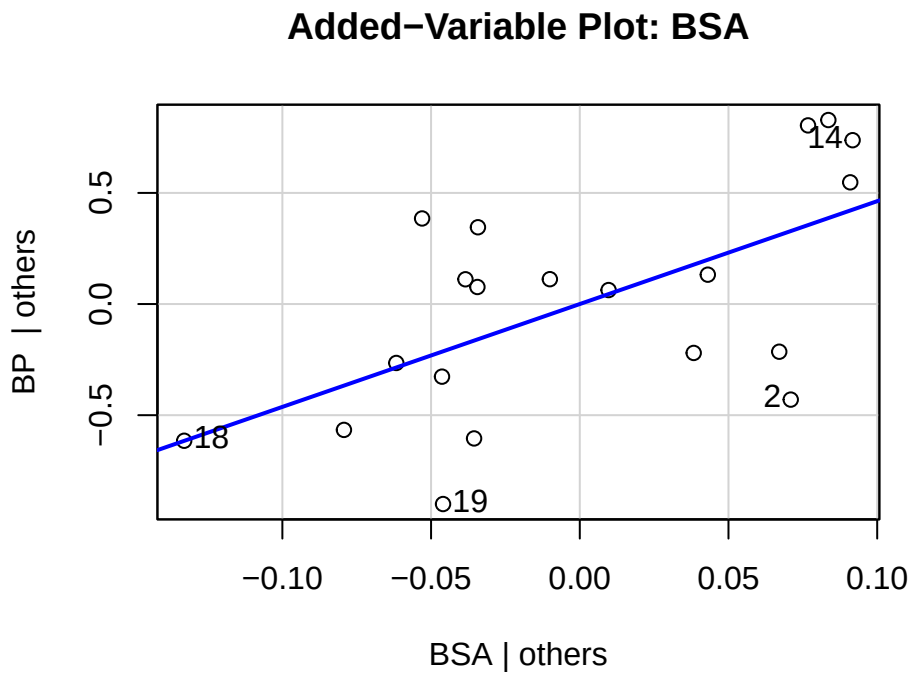
Let's go ahead and create the new and improved model with BSA:

```
fm4 <- lm(BP ~ Weight +  Age + BSA, data = BPdata)    #fm4 = fm3 + BSA
msummary(fm4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.667      2.647   -5.16  9.4e-05 ***
## Weight         0.906      0.049   18.49  3.2e-12 ***
## Age            0.702      0.044   15.96  3.0e-11 ***
## BSA            4.627      1.521    3.04   0.0078 **
##
## Residual standard error: 0.437 on 16 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.994
## F-statistic:  972 on 3 and 16 DF,  p-value: <2e-16
```

Of course there is an easier way to obtain added variable plots in R - we can simply use the `avPlot()` function from the `car` package to do so. This one line of code assumes you have made a new model `fm4` with Age, Weight and BSA as predictors.

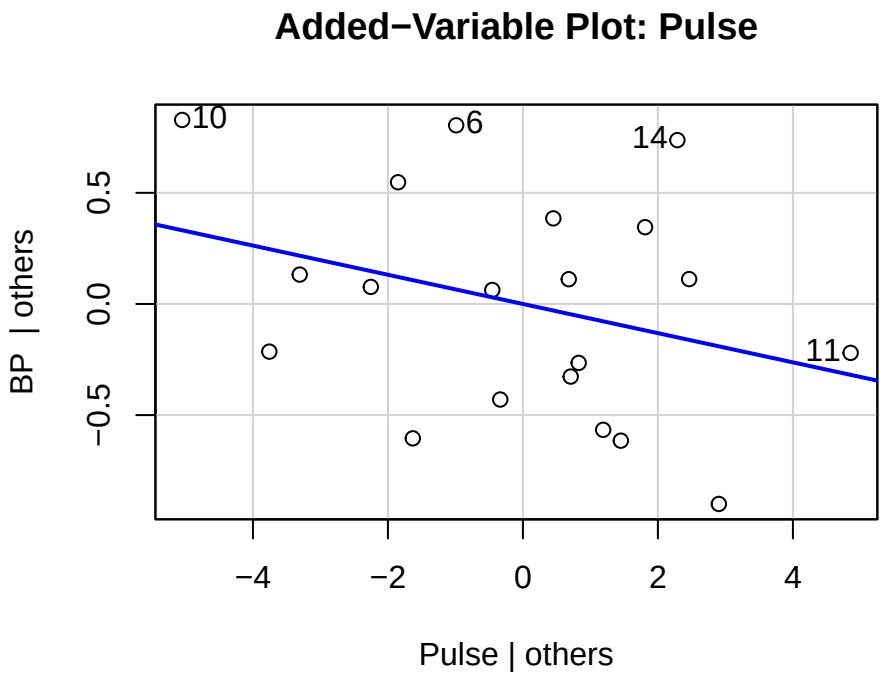
```
car::avPlot(fm4, "BSA") # produces the AVplot we did manually in Steps 1-3 to examine the value of addi
```



### Evaluate PULSE as an added predictor

Lets consider the other possible variable, `Pulse` we were considering to add to the model `fm3`.
Now we are creating a new model `fm5` to with `Weight`, `Age`, and `Pulse` as predictors, so we can use the avPlot function.

```
fm5 <- lm(BP ~ Weight +  Age + Pulse, data = BPdata)     #fm5 = fm3 + Pulse
car::avPlot(fm5, "Pulse")
```




3) What can we learn from this output?

4) Based on the two AV-plots above, which predictor, `BSA` or `Pulse`, would you consider adding to the model `fm3` (which has two predictors, `Weight` and `Age`).

IMPORTANT NOTE: The added variable plot can make it easy to see some potential issues with regression conditions (e.g. nonlinearity or heteroskedasticity) with respect to a particular predictor in a multiple regression model. These issues might not be obvious from the usual residuals vs. fitted plot (i.e. the `which = 1` plot), which does NOT control for the influence of the other variables. Thus, it can be helpful to check the added variable plots for each variable in a model. Examples are given below but not run. They are easy to get, examples below.

car::avPlot(fm4, “Age”)
car::avPlot(fm4, “Weight”)

or even better

car::avPlots(fm4)

## YOUR TURN TO PRACTICE: Attempt the following on your own and let’s review any questions next class

. . ### Evaluate your decision by looking at other information for concordance (vif, Adj Rsquared, residual standard error, nested F tests)

Run for fm4 (BSA)

```
msummary(fm4)
```

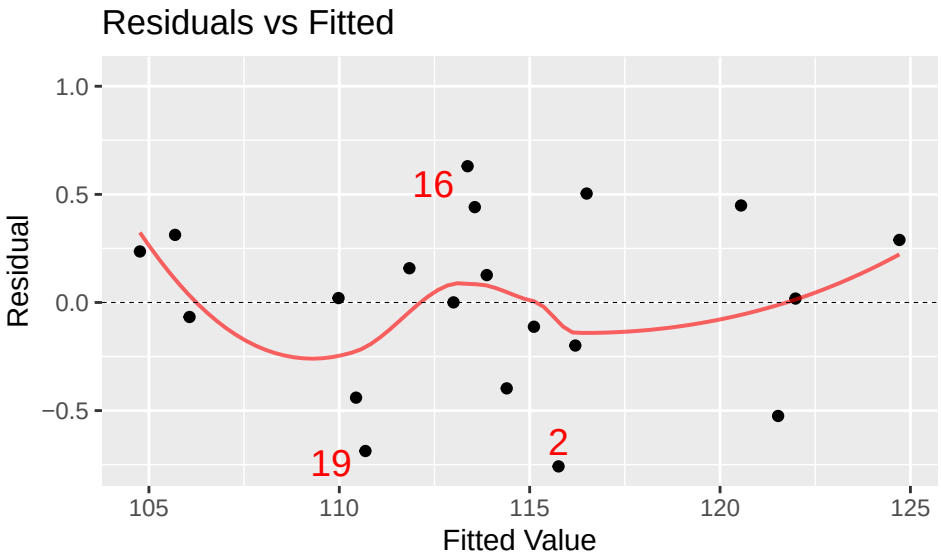
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.667      2.647   -5.16  9.4e-05 ***
## Weight         0.906      0.049   18.49  3.2e-12 ***
## Age            0.702      0.044   15.96  3.0e-11 ***
## BSA            4.627      1.521    3.04   0.0078 **
##
## Residual standard error: 0.437 on 16 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.994
## F-statistic:  972 on 3 and 16 DF,  p-value: <2e-16
```

```
car::vif(fm4)
```

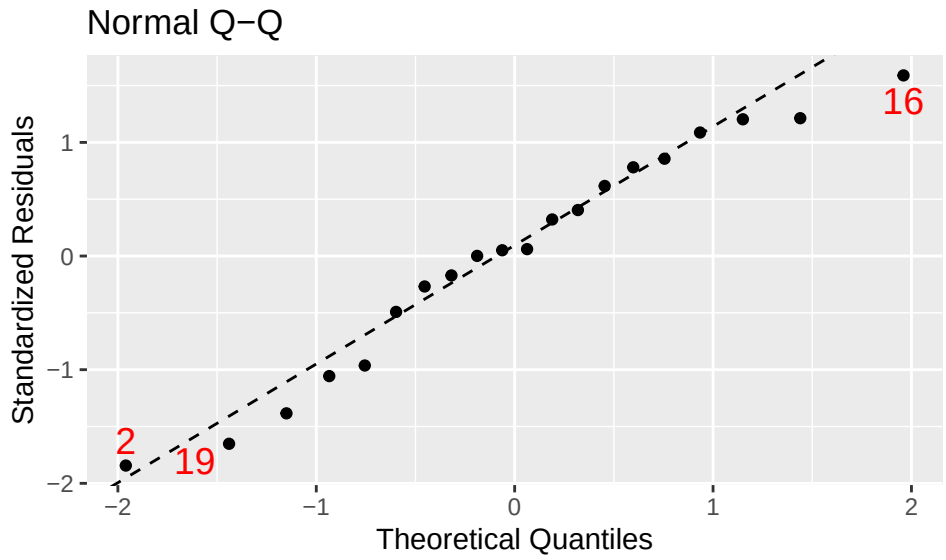
```
## Weight    Age    BSA
## 4.4036 1.2019 4.2869
```

```
mplot(fm4, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
mplot(fm4, which = 2)
```




```
anova(fm3, fm4)
```

```
## Analysis of Variance Table
##
## Model 1: BP ~ Weight + Age
## Model 2: BP ~ Weight + Age + BSA
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     17 4.82
## 2     16 3.06  1      1.77 9.25 0.0078 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Run for fm5 (Pulse)

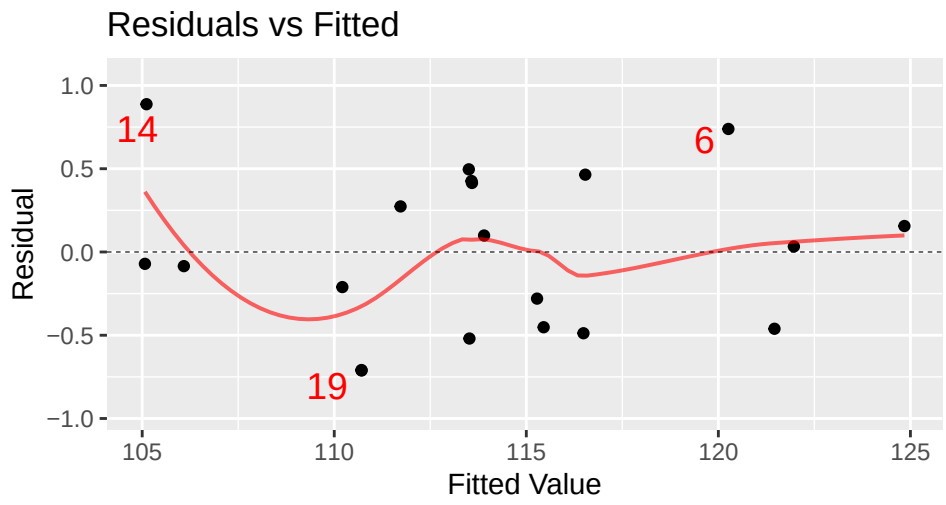
```
msummary(fm5)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.6900     2.9376   -5.68  3.4e-05 ***
## Weight        1.0614     0.0370   28.72  3.4e-15 ***
## Age           0.7502     0.0607   12.35  1.4e-09 ***
## Pulse        -0.0657     0.0485   -1.35     0.19
##
## Residual standard error: 0.52 on 16 degrees of freedom
## Multiple R-squared:  0.992,  Adjusted R-squared:  0.991
## F-statistic:  685 on 3 and 16 DF,  p-value: <2e-16
```

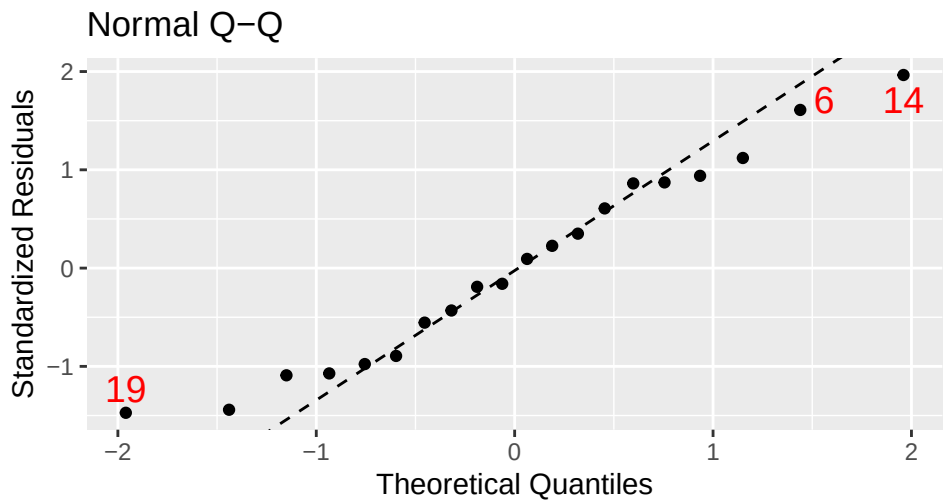
```
car::vif(fm5)
```

```
## Weight    Age  Pulse
## 1.7691 1.6204 2.3909
```

```
mplot(fm5, which = 1)
```




```
mplot(fm5, which = 2)
```




```
anova(fm3, fm5)
```

```
## Analysis of Variance Table
##
## Model 1: BP ~ Weight + Age
## Model 2: BP ~ Weight + Age + Pulse
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     17 4.82
## 2     16 4.33  1     0.496 1.83   0.19
```

5) After looking at the conditions (not so good for fm5 - pulse model), compare other information. Is there additional evidence to support our decision to chose fm4 (adding BSA) over fm5 (adding PULSE)?

**3 RULES or NOTE to REMEMBER:**

In fact, a model for $e_1 \sim e_2$ has several useful properties.

```
fm.res <- lm(fm.e1$residuals ~ fm.e2$residuals) #here is the regression of $e_1 \sim e_2$  (our plot)
```

First, this fitted line in the AVplot should go through the origin (0,0) as both sets of residuals, $e_1$ and $e_2$, have a mean 0; in other words, the intercept of this fitted line should be 0.

```
fm.res$coefficients       #the intercept here is practically 0
```

```
##     (Intercept) fm.e2$residuals
##     -1.6489e-17      4.6274e+00
```

Second, the residuals of this model $e_1 \sim e_2$ should be equal to the residuals of the new model of $BP \sim Age + Weight + BSA$.

```
fm4 <- lm(BP ~ Weight +  Age + BSA, data = BPdata)    #fm4 = fm3 + BSA
sum(fm.res$residuals - fm4$residuals)
```

```
## [1] 1.3184e-16
```

```
# this should be 0, just due to rounding if a little off
```

Third, the slope of this fitted line should be equal to the estimated coefficient of `BSA` in the full model `fm4`.

```
fm4$coefficients["BSA"]   #nice technique to pull out an individual coefficient from a model.
```

```
##    BSA
## 4.6274
```