LabH_Hw

Kate Brandt

February 22, 2019

Q1. Addressing Heteroskedasticity:

Heteroskedasticity is when the error term is not constant across cases. This makes estimating the confidence intervals and t-statistics unreliable. This means that the statistics we use to test hypothesis under the Gauss-Markov assumptions are no longer valid. So while our beta coefficients are unaffected, the variance of these estimates are not which can have implications for how accurate other assumptions about the model are.

Q2. Addressing Multicollinarity:

Multicollinearity is a problem because it forces the coefficients for correlated data to be derived from a subset of the data in a dataset, ultimately decreasing significance/reliability. If two variables are correlated, the correlated cases are not independent. Non-independent cases cannot be used to calculated the beta coefficient because there is a not enough variation within the data. So, if for example, 70% of cases in x and w are correlated, only 30% of the data points in x and w are being used to calculated the coefficient. This is a problem because it makes the calculation weaker.

Q3. Diagnosis for Dataset A

```r
# Load data
library(readstata13)
library(olsrr)
a <- read.dta13("./data/labwk5_a.dta")

# Create model and summarize
mod_a <- lm(y~x+z+q+w, a)
ols_regress(mod_a)
```

```
##                         Model Summary
## ---------------------------------------------------------------
## R                       1.000       RMSE                8.737
## R-Squared               0.999       Coef. Var          43.006
## Adj. R-Squared          0.999       MSE                76.339
## Pred R-Squared          0.999       MAE                 7.051
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                     ANOVA
## -------------------------------------------------------------------------------
##                    Sum of
##                   Squares          DF     Mean Square         F            Sig.
## -------------------------------------------------------------------------------
## Regression    49824149.482          4    12456037.371    163168.239     0.0000
## Residual         37787.614        495          76.339
## Total         49861937.097        499
## -------------------------------------------------------------------------------
```

```
##
##                              Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model        Beta    Std. Error    Std. Beta       t        Sig      lower      upper
## ------------------------------------------------------------------------------------------
## (Intercept)       0.354         0.392                  0.903    0.367     -0.416      1.124
##           x      25.030         0.031       1.003    806.484    0.000     24.969     25.091
##           z     -22.538         0.396      -0.071    -56.942    0.000    -23.315    -21.760
##           q       3.345         0.534       0.011      6.259    0.000      2.295      4.395
##           w      -0.309         0.536      -0.001     -0.577    0.564     -1.362      0.744
## ------------------------------------------------------------------------------------------
```

Initial analysis shows that the intercept and coefficient for w are not significant. There may be issues with these data. However, the R-squared is nearly 1, so these data are almost perfectly explanatory of y. Notice that the coefficients for x and z are rather large.

Next, we will use the Breusch-Pagan test to check for heteroskedasticity.

```
# Test for heteroskedasticity using Breusch-Pagan test
ols_test_breusch_pagan(mod_a, rhs = TRUE, multiple = TRUE)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##        Data
##  ------------------
##  Response : y
##  Variables: x z q w
##
##         Test Summary (Unadjusted p values)
##  ----------------------------------------------
##   Variable           chi2       df        p
##  ----------------------------------------------
##   x                0.7424041     1    0.3888918
##   z                0.8627134     1    0.3529805
##   q                1.1697741     1    0.2794477
##   w                2.0117363     1    0.1560866
##  ----------------------------------------------
##   simultaneous     3.2918332     4    0.5102273
##  ----------------------------------------------
```

None of the explanatory variables exhibit p-values less than 0.05 in the Breusch-Pagan test, which means we can conclude that there is no heteroskedasticity in this data. Check this again using robust regression. To calculate the estimate of variance for each case, we will use the Huber-White robust standard errors test:

```
library(foreign)
library(sandwich)
library(lmtest)

# Estimate standard errors with robust regression
coeftest(mod_a, vcov = vcovHC(mod_a, "HC1"))
```

```
##
## t test of coefficients:
```

```
##
##                Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)   0.354093   0.391646    0.9041    0.3664
## x            25.030284   0.014385 1740.0042 < 2.2e-16 ***
## z           -22.537628   0.372649  -60.4795 < 2.2e-16 ***
## q             3.344992   0.506251    6.6074 1.013e-10 ***
## w            -0.309311   0.539233   -0.5736    0.5665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No significant difference between the standard errors in the robust regression. The same variables are not significant as in the original regression (intercept and w).

We know there is no heteroskedasticity in this model. Next, test for multicollinearity

```r
# correlation matrix
cor(a[,(c(1,2,3,4,6))])
```

```
##            x          z          w          q          y
## x 1.00000000 0.08922922 0.01918552 0.04649155 0.99711301
## z 0.08922922 1.00000000 0.07084426 0.06565461 0.01920997
## w 0.01918552 0.07084426 1.00000000 0.69592903 0.02073744
## q 0.04649155 0.06565461 0.69592903 1.00000000 0.05207843
## y 0.99711301 0.01920997 0.02073744 0.05207843 1.00000000
```

There is high correation between x and y. All other relationships appear uncorrelated.

Next, calculate the VIF.

```r
# VIF check

ols_vif_tol(mod_a)
```

```
## # A tibble: 4 x 3
##   Variables Tolerance   VIF
##   <chr>         <dbl> <dbl>
## 1 x             0.990  1.01
## 2 z             0.987  1.01
## 3 q             0.514  1.94
## 4 w             0.515  1.94
```

There are no significantly high VIF values (VIF > 10). This isn't surprising since it seems like the main correlation is between x and y. Multicollinearity is not really a problem in this data, but there is probably something going on with the x data. Even if there is correlation between the x and y, the lack of a significant VIF means that we don't have to worry much until we have more information.

Next, we will check for outliers and influential cases.

```r
# Start by looking at the basic stats of the data
summary(a)
```

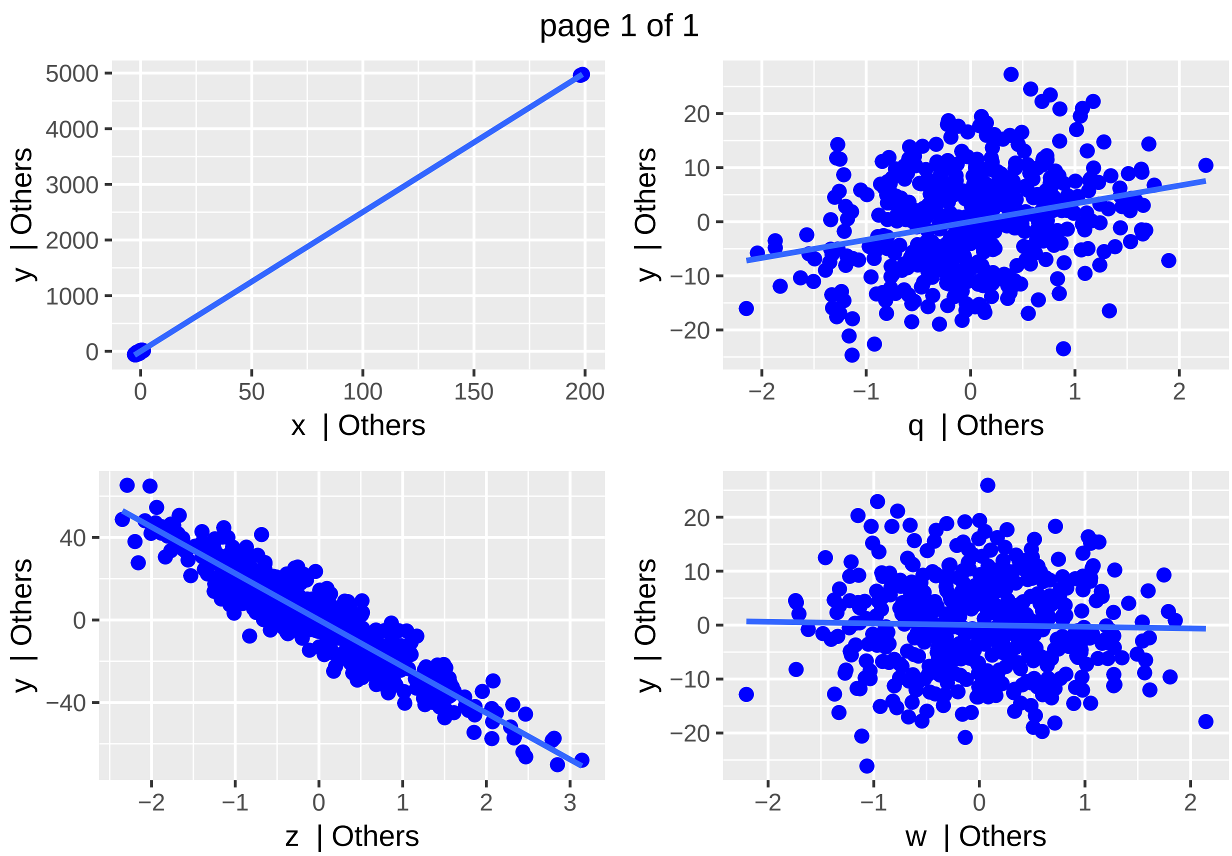
```
##        x                   z                   w
##  Min.   : -2.82066   Min.   :-2.32055   Min.   :-3.12846
##  1st Qu.: -0.61709   1st Qu.:-0.63376   1st Qu.:-0.76479
##  Median : -0.02613   Median :-0.01503   Median :-0.01857
##  Mean   :  0.81725   Mean   : 0.01656   Mean   :-0.03512
##  3rd Qu.:  0.67864   3rd Qu.: 0.63459   3rd Qu.: 0.68867
##  Max.   :200.20560   Max.   : 3.08475   Max.   : 2.76435
##        q                  n               y                   d
```

```
##  Min.   :-3.0749   Min.   :  1.0   Min.   : -18.960   Min.   :0.0000000
##  1st Qu.:-0.6942   1st Qu.:125.8   1st Qu.:  -3.676   1st Qu.:0.0001465
##  Median :-0.1107   Median :250.5   Median :   0.390   Median :0.0006241
##  Mean   :-0.0394   Mean   :250.5   Mean   :  20.316   Mean   :0.0019112
##  3rd Qu.: 0.6962   3rd Qu.:375.2   3rd Qu.:   4.402   3rd Qu.:0.0019800
##  Max.   : 2.8090   Max.   :500.0   Max.   :5004.394   Max.   :0.1688666
```

It appears as though there may be something unusual about the x data. The maximum is much higher than the other variables, and is very distant from the 3rd quartile value. It would make sense if there are some outlier cases in the x data at the upper limit.

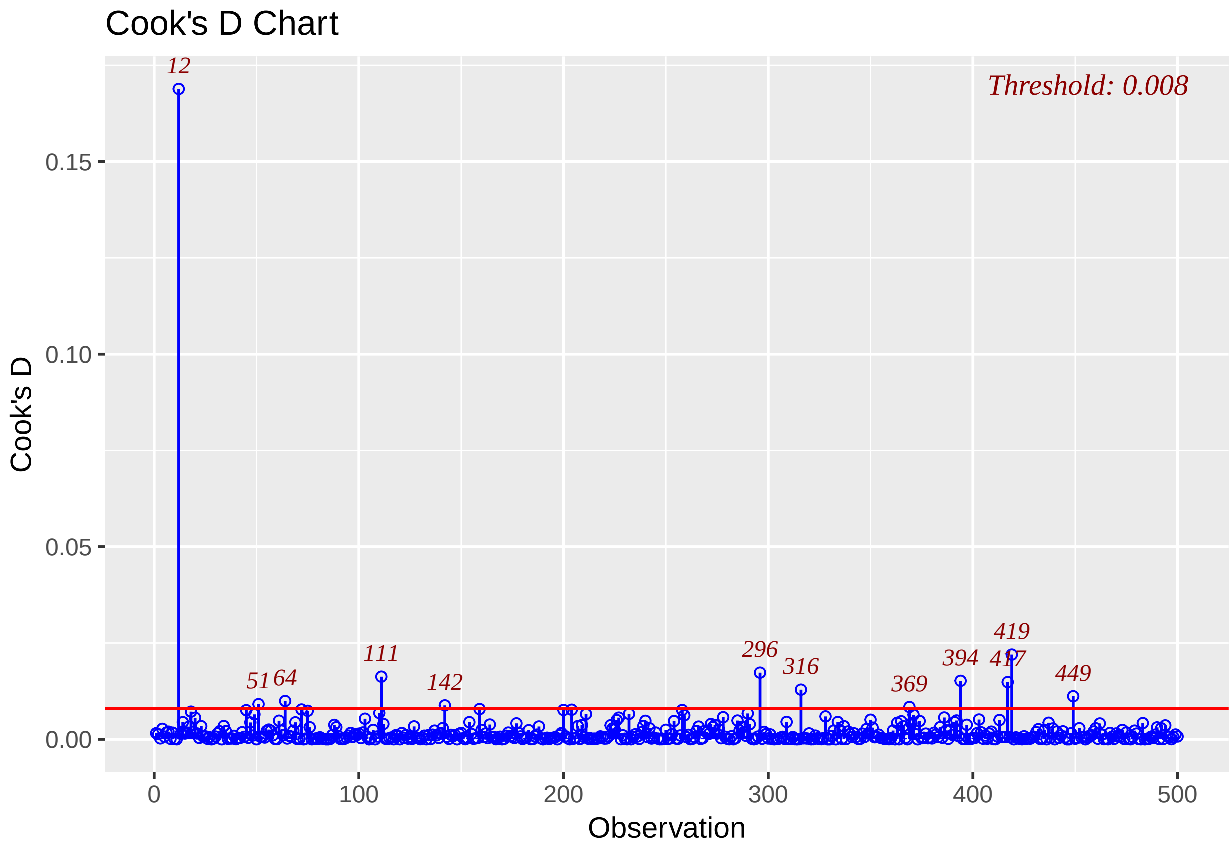
Next, added-variable plots:

```r
ols_plot_added_variable(mod_a)
```



These av-plots indicate there is something wrong with the x data. There is likely at least one outlier case in the data. Next, we will identify this case using Cook's D:

```r
# identify influential case
ols_plot_cooksd_chart(mod_a)
```



This chart shows us exactly which cases are the problem (n = 11 and 12). It also shows us the threshold to classify cases as outliers. In this model, case 12 is far, far over the threshold. There are other outlier cases in these data, but this is likely normal variation. It would probably not be ethical to exclude the cases that are somewhat above the threshold, but it seems obvious that there was a coding error in case 12.

(Note: I know for some reason it's difficult to see the actual cases from the chart so that case 11 seems to not be identifiable here. I couldn't get it to plot properly.)

Next, let's exclude this outlier case and regress.

```
mod_a_fixed <- lm(y~x+w+z+q,a[-c(11,12),])
ols_regress(mod_a_fixed)
```

```
##                          Model Summary
## ----------------------------------------------------------------
## R                          0.626       RMSE                4.733
## R-Squared                  0.391       Coef. Var        1540.379
## Adj. R-Squared             0.386       MSE                22.397
## Pred R-Squared             0.379       MAE                 3.753
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ----------------------------------------------------------------------
##                  Sum of
##                 Squares          DF    Mean Square      F          Sig.
```

```
## --------------------------------------------------------------------
## Regression     7096.715          4       1774.179    79.217    0.0000
## Residual      11041.513        493         22.397
## Total         18138.228        497
## --------------------------------------------------------------------
##
##                                Parameter Estimates
## ------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ------------------------------------------------------------------------------------
## (Intercept)     0.408         0.212                  1.923    0.055    -0.009     0.825
##           x     1.079         0.694        0.182    1.556    0.120    -0.284     2.442
##           w     0.480         0.292        0.081    1.646    0.100    -0.093     1.053
##           z     0.797         0.709        0.131    1.125    0.261    -0.595     2.190
##           q     2.747         0.290        0.464    9.475    0.000     2.178     3.317
## ------------------------------------------------------------------------------------
```

The coefficient for x is no longer highly inflated, nor is the coefficient for z. However, both of these variables' cofficients are no longer significant.

There was an issue with correlation in the first model, so let's check if this is still an issue.

```r
cor(a[-c(11,12),(c(1,2,3,4,6))])
```

```
##            x          z          w          q         y
## x 1.00000000 0.95349564 0.08099674 0.05888374 0.3411657
## z 0.95349564 1.00000000 0.06801159 0.06330737 0.3398196
## w 0.08099674 0.06801159 1.00000000 0.69487565 0.4267677
## q 0.05888374 0.06330737 0.69487565 1.00000000 0.5390274
## y 0.34116567 0.33981965 0.42676768 0.53902735 1.0000000
```

Removing the two influential-outlier cases did fix the problem of correlation between x and y. However, there seems to be a problem of multicollinearity between z and x now. Let's check the VIF:

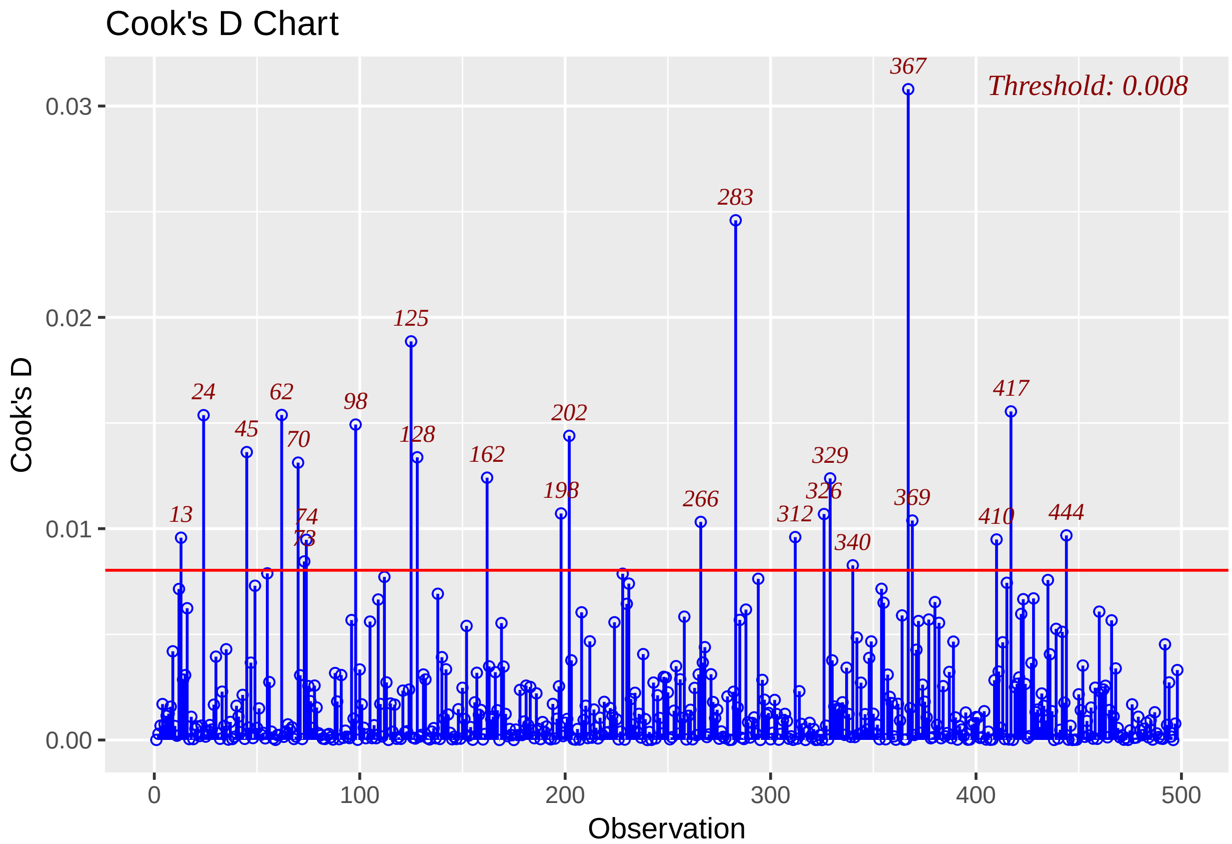
```r
ols_vif_tol(mod_a_fixed)
```

```
## # A tibble: 4 x 3
##   Variables Tolerance   VIF
##   <chr>         <dbl> <dbl>
## 1 x            0.0903 11.1
## 2 w            0.513   1.95
## 3 z            0.0904 11.1
## 4 q            0.515   1.94
```

As suspected, there is significant correlation. The best solution to this problem would be to collect more data. Otherwise, we should drop z or x from the model.

One last check: replot Cook's D with our new model that excludes the problematic cases:

```r
# identify influential case
ols_plot_cooksd_chart(mod_a_fixed)
```



Summary: Coding error on cases 11 and 12 seemed to have been the issue in this dataset. By dropping these cases, we were able to fix problems of outlier influence but revealed that the data have problems of multicollinearity.

Q4. Diagnosis for Dataset B

```r
# Load data
library(readstata13)
library(olsrr)
b <- read.dta13("./data/labwk5_b.dta")

# Create model and summarize
mod_b <- lm(y~x+z+q+w, b)
ols_regress(mod_b)
```

```
##                            Model Summary
## ----------------------------------------------------------------
## R                         0.248       RMSE                49.369
## R-Squared                 0.061       Coef. Var           48.743
## Adj. R-Squared            0.057       MSE               2437.307
## Pred R-Squared            0.052       MAE                 39.331
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
```

```
##                                  ANOVA
## ------------------------------------------------------------------------
##                     Sum of
##                    Squares         DF    Mean Square      F         Sig.
## ------------------------------------------------------------------------
## Regression     158291.420          4      39572.855    16.236    0.0000
## Residual      2425120.405        995       2437.307
## Total         2583411.825        999
## ------------------------------------------------------------------------
##
##                                    Parameter Estimates
## ---------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig       lower     upper
## ---------------------------------------------------------------------------------------------
## (Intercept)    -10.501        16.960                -0.619    0.536    -43.782    22.780
##           x     11.194         1.689        0.219    6.627    0.000      7.879    14.509
##           z      2.947         1.633        0.059    1.804    0.072     -0.258     6.152
##           q      0.084         1.627        0.002    0.052    0.959     -3.109     3.277
##           w      1.381         1.579        0.027    0.874    0.382     -1.718     4.480
## ---------------------------------------------------------------------------------------------
```

The R-squared of this model is very low. None of the variables are significant, except for x.

First, test for heteroskedasticity using Breusch-Pagan test:

```r
# Test for heteroskedasticity using Breusch-Pagan test
ols_test_breusch_pagan(mod_b, rhs = TRUE, multiple = TRUE)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  ----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##        Data
##  ------------------
##  Response : y
##  Variables: x z q w
##
##           Test Summary (Unadjusted p values)
##  --------------------------------------------------
##   Variable            chi2       df          p
##  --------------------------------------------------
##   x                33.9761550     1    5.579164e-09
##   z                 6.6203416     1    1.008205e-02
##   q                 0.9957297     1    3.183460e-01
##   w                 0.5720473     1    4.494465e-01
##  --------------------------------------------------
##   simultaneous     36.9753929     4    1.822451e-07
##  --------------------------------------------------
```

There are major issues with heteroskedasticity in these data. All variables are exhibiting p-values < 0.05 in the BP test, so they are all exhibiting heteroskedasticity. Next, let's check this again using robust regression. To calculate the estimate of variance for each case, we will use the Huber-White robust standard errors test:

```r
library(foreign)
library(sandwich)
```

```r
library(lmtest)

# Estimate standard errors with robust regression
coeftest(mod_b, vcov = vcovHC(mod_a, "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -10.501000   0.391646 -26.8125 < 2.2e-16 ***
## x            11.193900   0.014385 778.1547 < 2.2e-16 ***
## z             2.947052   0.372649   7.9084 6.904e-15 ***
## q             0.083897   0.506251   0.1657   0.86841
## w             1.380960   0.539233   2.5610   0.01058 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This seemed to have addressed much of the problem. The standard errors are much smaller and similar to one another, and all but the q coefficient are significant. Heteroskedasticity means that there are issues with the error terms, but not with the estimates of our coefficients, so we shouldn't expect too many issues with outliers and multicollinearity in the rest of the diagnosis.

Next, let's check for multicollinearity:

```r
cor(b[,-5])
```

```
##             x            z           w             q          y
## x  1.00000000  0.3597645606 -0.02180513  0.0748053571 0.23979892
## z  0.35976456  1.0000000000 -0.01469716  0.0003163632 0.13777428
## w -0.02180513 -0.0146971600  1.00000000 -0.0324633568 0.02117784
## q  0.07480536  0.0003163632 -0.03246336  1.0000000000 0.01711003
## y  0.23979892  0.1377742764  0.02117784  0.0171100278 1.00000000
```

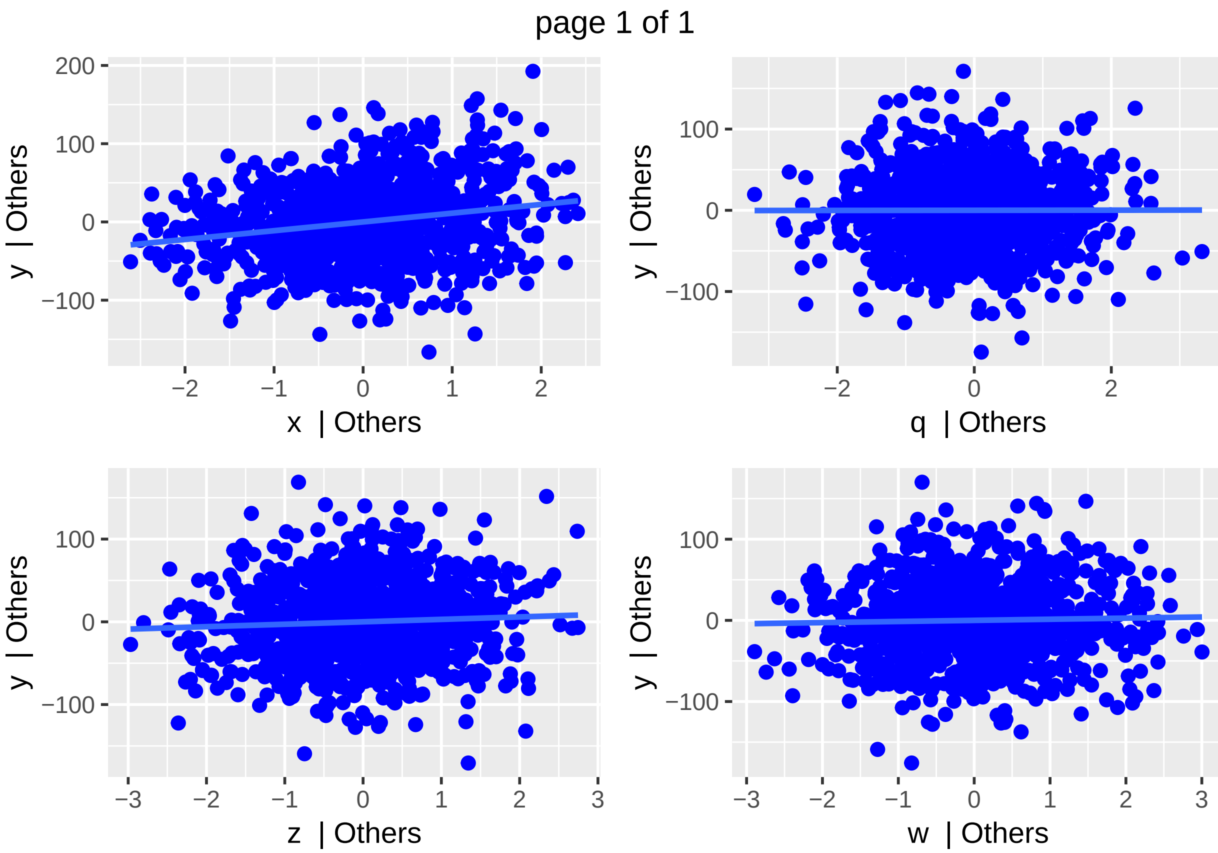
```r
ols_vif_tol(mod_b)
```

```
## # A tibble: 4 x 3
##   Variables Tolerance   VIF
##   <chr>         <dbl> <dbl>
## 1 x             0.865  1.16
## 2 z             0.870  1.15
## 3 q             0.993  1.01
## 4 w             0.999  1.00
```

The correlation matrix indicates no correlation. There are no large VIF values, indicating that we don't have a problem with multicollinearity.

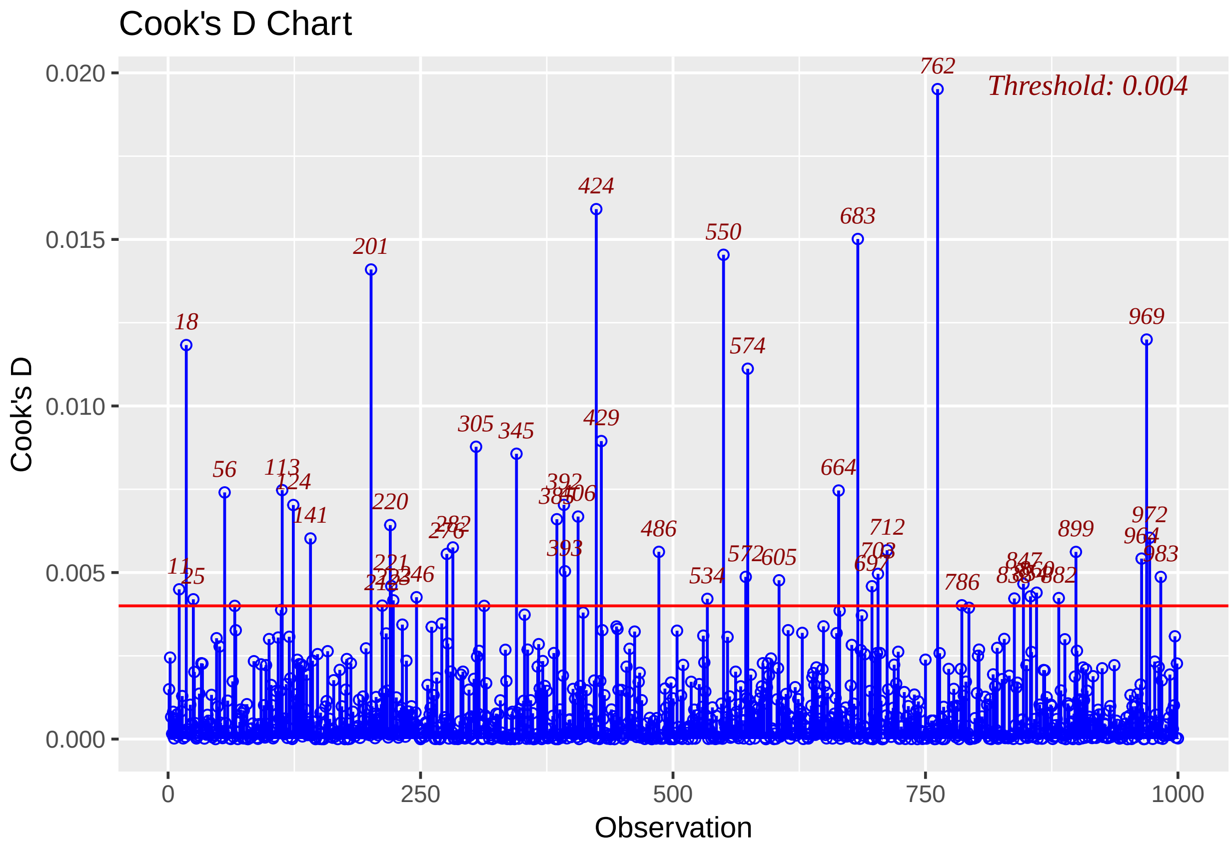
Next, let's check for outliers and influential cases:

```r
#Start with added-variable plots
ols_plot_added_variable(mod_b)
```



These av-plots indicate there is no issue with outliers. To be sure, let's plot Cook's D:

```
# identify influential cases
ols_plot_cooksd_chart(mod_b)
```



While there are some cases that cross the threshold here, this seems like the function of regular variation. There is no need to be concerned about outlier cases and their influence on the model in this data set.

Summary: Data set B has major problems with heteroskedasticity, which isn't necessarily a problem, but something that we as researchers/data scientists should be aware of when drawing conclusions about the estimations of the beta coefficients.