HW3 - MM2

Ryan Gallagher

2023-03-04

See HW3.SAS and HW3.lst for all SAS code and output that I reference in this document.

(1) Fit a linear regression $Y = \beta_0 + \beta_1 X1 + \cdots + \beta_{15} X15 + e$ using PROC GLM and produce diagnostics plots. What assumptions does it violate?

Start by creating the data:

```
library(MASS)
library(Matrix)

covariance1 = matrix(c(1,0.8,0.8,0.8,1,0.8,0.8,0.8,1),3,3,byrow = TRUE)

covariance = bdiag(covariance1,1,1,1,1,1,1,1,1,1,1,1,1)
mean = rep(0,dim(covariance)[1])
set.seed(1)
n = 100                 #Version with n=400 is made too
x1tox15 = mvrnorm(n,mean,covariance)
betavec = c(0.5,0.6,-0.7,0.4,0.8,-0.6,rep(0,10))
X = cbind(rep(1,n),x1tox15)
y = exp(X%*%betavec+rnorm(n,0,2)) # response variable
Xdesign = x1tox15 # Design matrix

data = cbind(Xdesign, y)
df = as.data.frame(data)
colnames(df) = c('x1', 'x2', 'x3', 'x4', 'x5',
'x6','x7','x8','x9','x10','x11','x12','x13','x14','x15','y')

write.csv(df, 'HW3.csv') #Version of n=400 is made too
```

Then we will import it to SAS:

```
libname lib 'lib';

proc import datafile='HW3.csv' out=lib.HW3
    DBMS=csv REPLACE;
    GETNAMES=YES;
run;

-----For Part 6 --------
proc import datafile='HW3_400.csv' out=lib.HW3_400
```

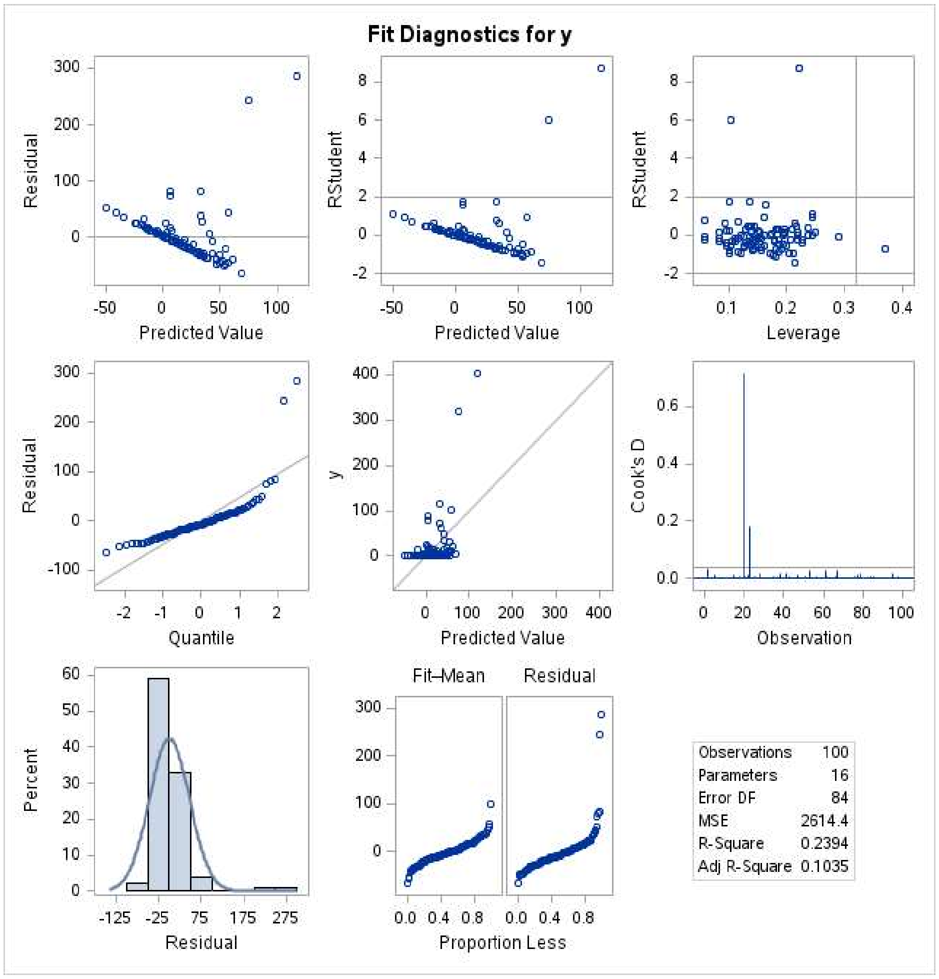
```
    DBMS=csv REPLACE;
    GETNAMES=YES;
run;

data hw3;
    set lib.HW3;
    drop VAR1;
run;
```

Then I'll fit a linear regression and produce a diagnostic plot:

```
title 'part1';
ods graphics on;
proc GLM data=hw3 PLOTS = DIAGNOSTICS;
    model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15;
run;
ods graphics off;
```

with diagnostics plot:

```
knitr::include_graphics("DiagnosticsPanel.png")
```



This data has massive violations all across the board. It violates linearity from the Residual vs. Predicted Value plot. It violates the normality assumption when looking at the QQ plot. It also has a few observations that could be seen as outliers, but I don't think they're too bad.

(2) Conduct a Box-Cox transformation. What transformation do you recommend for this data set?

In SAS:

```
title 'part2';
proc transreg data=hw3;
    model boxcox(y / lambda=-1 to 1 by 0.1) = identity(x1 x2 x3 x4 x5 x6 x7
x8 x9 x10 x11 x12 x13 x14 x15);
run;
```

Where we find that $\lambda = 0.1$ as our recommended transformation. log(y) could also work.

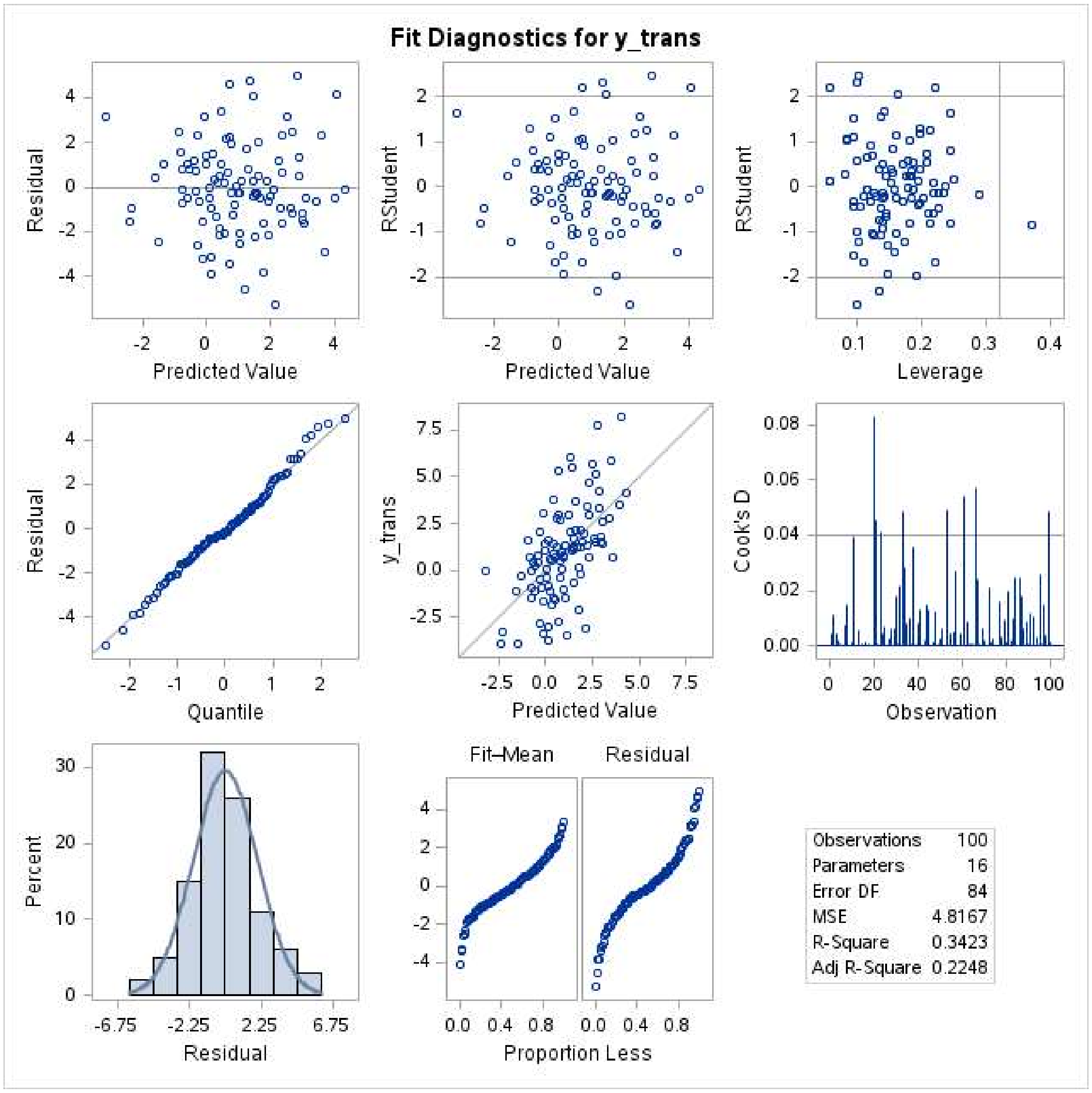
(3) Fit a linear regression using (2) and check the model diagnostics

In SAS:

```
title 'part 3';
data boxcoxhw3;
    set hw3;
    y_trans = (y**(0.1) - 1) / (0.1);
run;

ods graphics on;
proc GLM data=boxcoxhw3 PLOTS = DIAGNOSTICS;
    model y_trans = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15;
    output out=hw3bc;
run;
ods graphics off;
```

With diagnostic plots:

```
knitr::include_graphics('DiagnosticsPanel1.png')
```



Observations	100
Parameters	16
Error DF	84
MSE	4.8167
R-Square	0.3423
Adj R-Square	0.2248

See DiagnosticPanel1.png - the violated assumptions have been remedied. The QQ plot for normaility is now straight, the Residual vs. Predicted is random with no trend, and the highest leverage point is very small.

(4) Select significant variables using stepwise based on BIC

In SAS:

```
title 'part 4';
proc reg data=hw3bc;
    model y_trans = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 / sbc;
run;
```

From this, we find that $x4$, $x5$, and $x13$ are the only significant variables using this method.

(5) Select significant variables using stepwise selection based on alpha=0.05.

In SAS:

```
title 'part 5';
proc reg data=hw3bc;
    model y_trans = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 /
selection=stepwise slentry=0.05;
run;
```

Here, we find that $x4, x5, x9, x13$, and $x14$ qualify via stepwise selection.

(6) Use n=400 instead of n=100. Perform (4) and (5). What is your final model? Is it different from what you found in (4) and (5)?

In SAS:

```
title 'part6';

data n400;
    set lib.HW3_400;
    drop VAR1;
    y=log(y);
    /* BoxCox says labda=0 best -> so we log transform. */
run;

proc reg data=n400;
    model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 / sbc;
run;
```

Here, we only find that $x1, x2, x3, x4$, and $x5$ are significant.

```
proc reg data=n400;
     model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 /
selection=stepwise slentry=0.05;
run;
```

Here we once again find that $x1, x2, x3, x4$, and $x5$ are all significant.

Our final model is log(y) = $0.59x_1 - 0.81x_2 + 0.51x_3 + 0.72x_4 - 0.42x_5 + 0.46 + error$.

These results are certainly different than what we found in (4) and (5). I'm curious as to why when we increased n that some of our variables were less significant and others more. I would guess that higher n would lead to higher accuracy in these sorts of things, so perhaps I should trust the $n = 400$ model more.

Again, look to HW3.lst for the SAS output that confirm my findings.