STAT 5302 Applied Regression Analysis
Course Project

By Chunni Zhao

December 5, 2017

1 DATA PREPARATION

First transform variable Use into factor. From the summary statistics in Table 4.1, we can see that all of the regressors have large range. Besides, in the regressor Use, no category has extremely small number of observations. From the scatter plot matrix Figure 4.1 we can see that there might be a point with high regressor values but relatively low response value. By checking the point with $FTE > 2000$ $ImprV > 1.5 * 10^7$, $LandV > 2 * 10^6$ $Size > 6 * 10^5$, we can find out that it is a single point, the case 77. The case 77 has $Wst = 109.2$, which is relatively small compared with the regressors.
We may use power transformation to make regressors linearly related. From the output in Table 4.2 we can see that all of the confidence intervals include 0. And the likelihood ratio tests in Table 4.3 shows that at the significance level 0.05, we may reject the null hypothesis of making no power transformation , but can not reject making all log transformations. As for a comparison, I have also tried using the power transformation for the data after dropping the case 77. The results are in Table 4.4 and Table 4.5. We can see that the result after dropping case 77 is in accordance with the result using the full data. From the scatter plot matrix Figure 4.2 for the response and log-transformed regressors, we can see that the regressors are generally linearly related. However, the relations between response and regressors are not linear. The scatter plot matrix after dropping case 77 (Figure 4.3) has similar pattern.

2 MODELING

We may first consider fitting the regression of the Wst on the transformed regressors. Since for the linearly related regressors and assuming that $E(Y|X = x) = g(\beta' x)$, then the OLS estimation β is a consistent estimate of $c\beta$ with a constant c[1].
In model1, I first considered all two way interactions of the transformed regressors and categorical variable Use. Then I use backward selection with AIC for variables screening. The

selected model using wilkinson-rogers notation is:

$$Wst \sim log(FTE) + log(ImprV) + log(LandV) + log(Size) + Use + log(FTE) : log(ImprV) + log(FTE) : log(LandV) + log(ImprV) : log(LandV) + log(ImprV) : log(Size) + log(ImprV) : Use$$

Figure 4.4 shows the scatter plot of the response on the fitted value. We can see that the data points are concentrated on the lower left but spreads out on the right.
For the next step, I considered transforming the fitted value to get the function $g(\beta' x)$. From the resulf from Figure 4.5 and Table 4.6, I choose 4 as the power and fit the polynomial model:

$$E(Wst|fitted) = \beta_0 + \beta_1 fitted + \beta_2 fitted^2 + \beta_3 fitted^3 + \beta_4 fitted^4 \tag{2.1}$$

The residual plot for the fitted model is shown in Figure 4.6. We can see that there is still a cluster on the left. For the non-constant variance test, it rejects the null hypothesis of constant variance with $\chi^2 = 74.46437$, $df = 1$ and $p = 6.2e-18$.
To solve this, I considered the weighted regression. I have tried to fit the squared residual from Equation 2.1 on the fitted value to get a weight:

$$resid^2 = \theta_0 + \theta_1 fitted$$

The summary table Table 4.7 shows that we can take

$$\frac{1}{72.6325 + 25.4773 fitted}$$

as the weight. Refit the model and from the Pearson residual plot Figure 4.6 for the weighted regression, we can see that the small trend in the cluster on the left gets smaller compared with Figure 4.7. Besides, non-constant variance test gives $\chi^2 = 0.4139518$, $df = 1$ and $p = 0.51$. We can not reject the null hypothesis of constant variance at the significance level 0.05 this time.
The outlier test in Table 4.8 shows that there are two possible outliers: case 81 and case 61. However, the Cook's distance in Figure 4.8 only shows case 41 is the most influential. The confidence region scatter plot matrix of the model using original data(using solid line) and the model removing case 41(using dashed line) is shown in Figure 4.9. We can see that although some points are outside the sold ellipse, the coefficients before and after dropping case 41 all have the same sign and close to each other to some extent. So that we may claim that the finally fitted model is kind of robust.

3 FINAL MODEL

The final model fitted is:

$$
\begin{aligned}
fitted(FTE, ImprV, LandV, Size, Use) = {} & 1564.6 + 3.8log(FTE) - 163log(ImprV) - 311.2log(LandV) \\
& + 174.2log(Size) + 942.9Use3 + 674.2Use4 + 483.8Use5 \\
& + 547.8Use6 + 11.8log(FTE) : log(ImprV) \\
& - 13log(FTE) : log(LandV) + 28.8log(ImprV) : log(LandV) \\
& - 14.3log(ImprV) : log(Size) - 78.6log(ImprV) : Use3 \\
& - 55.1log(ImprV) : Use4 - 40.4log(ImprV) : Use5 \\
& - 46.2log(ImprV) : Use6
\end{aligned}
$$

$$
\widehat{E}(Wst|fitted) = 25.6 + 567.3fitted + 307.4fitted^2 + 164.6fitted^3 + 89.1fitted^4
$$

$$
Weight(fitted) = 72.6 + 25.5fitted
$$

This model might be hard for interpretation, but it provides an estimation of conditional mean for the waste given the regressors(without transforming the response). So it might be competent for our purpose of setting the tax which is a monotonically increasing function of the waste.

4 APPENDIX

4.1 TABLES

Table 4.1: Summary statistics for regressors

	FTE	ImprV	LandV	Size	Use	Wst
1	Min. : 1.00	Min. : 11700	Min. : 6300	Min. : 913	2:24	Min. : 0.09
2	1st Qu.: 5.00	1st Qu.: 84700	1st Qu.: 29650	1st Qu.: 3664	3:27	1st Qu.: 2.60
3	Median : 10.00	Median : 200600	Median : 54100	Median : 6680	4:14	Meddian : 9.70
4	Mean : 49.86	Mean : 542341	Mean : 127215	Mean : 19407	5:39	Mean : 25.73
5	3rd Qu.: 26.00	3rd Qu.: 393700	3rd Qu.: 123850	3rd Qu.: 14732	6:43	3rd Qu.: 20.80
6	Max. :3000.00	Max. :22996600	Max. :2530900	Max. :788876		Max. :520.00

Table 4.2: Power transformation for regressors

	Est.Power	Std.Err.	Wald Lower Bound	Wald Upper Bound
FTE	-0.03	0.04	-0.11	0.05
ImprV	0.03	0.03	-0.03	0.10
LandV	-0.10	0.06	-0.22	0.02
Size	-0.04	0.04	-0.12	0.05

Table 4.3: Likelihood ratio tests for the power for regressors

	LRT	df	pval
LR test, lambda = (0 0 0 0)	7.98	4	0.09
LR test, lambda = (1 1 1 1)	1712.24	4	0.00

Table 4.4: Power transformation for regressors after dropping case 77

	Est.Power	Std.Err.	Wald Lower Bound	Wald Upper Bound
FTE	-0.02	0.05	-0.11	0.08
ImprV	0.06	0.04	-0.01	0.13
LandV	-0.11	0.07	-0.24	0.03
Size	-0.03	0.05	-0.12	0.07

Table 4.5: Likelihood ratio tests for the power for regressors after dropping case 77

	LRT	df	pval
LR test, lambda = (0 0 0 0)	7.47	4	0.11
LR test, lambda = (1 1 1 1)	1298.50	4	0.00

Table 4.6: Power transformation on the fitted value

	lambda	RSS
1	4.050555	124577.3
2	-1.000000	561501.6
3	0.000000	467963.1
4	1.000000	231108.9

Table 4.7: Summary table of fitting the squared residual on fitted

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	72.6325	208.8363	0.35	0.7285
fitted(s1)	25.4773	3.9088	6.52	0.0000

Table 4.8: result for outlier test

	rstudent	unadjusted p-value	Bonferonni p
81	5.965307	1.8742e-08	2.7551e-06
61	5.958074	1.9417e-08	2.8542e-06

4.2 GRAPHS

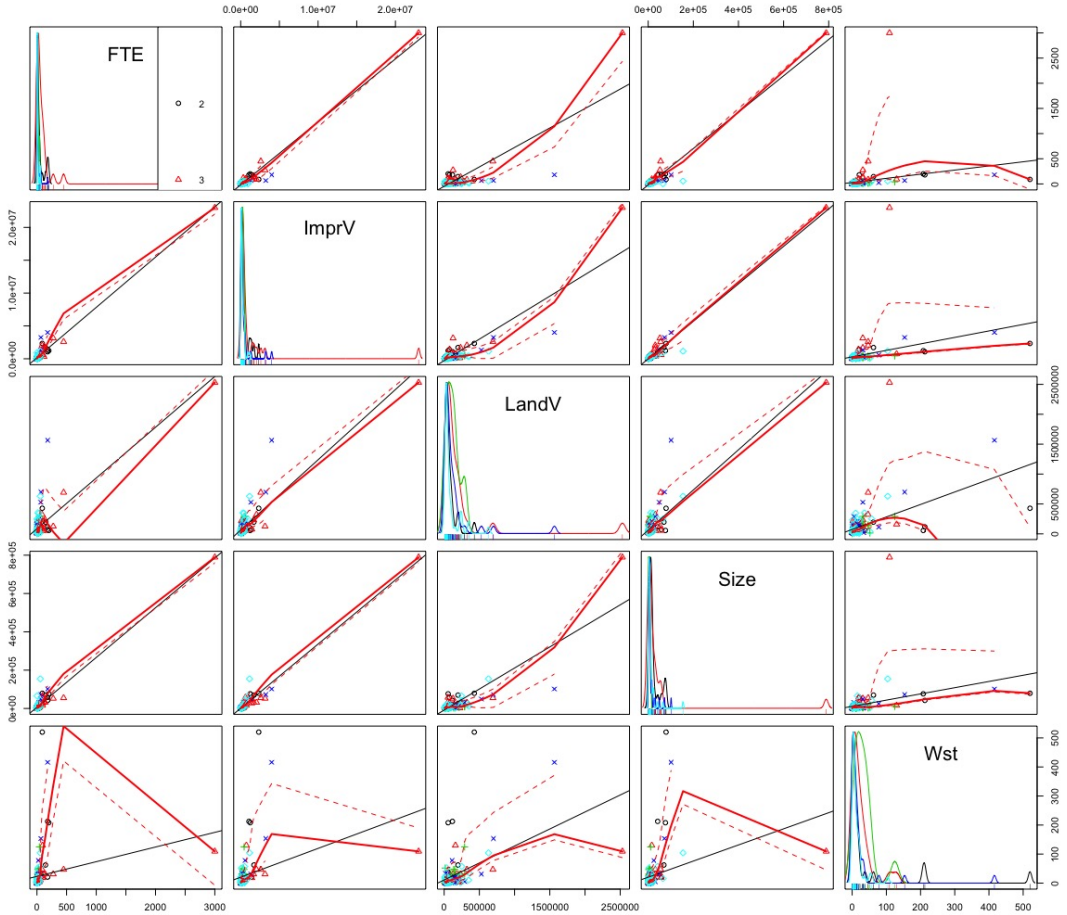

Figure 4.1: scatter plot matrix of original response and regressors

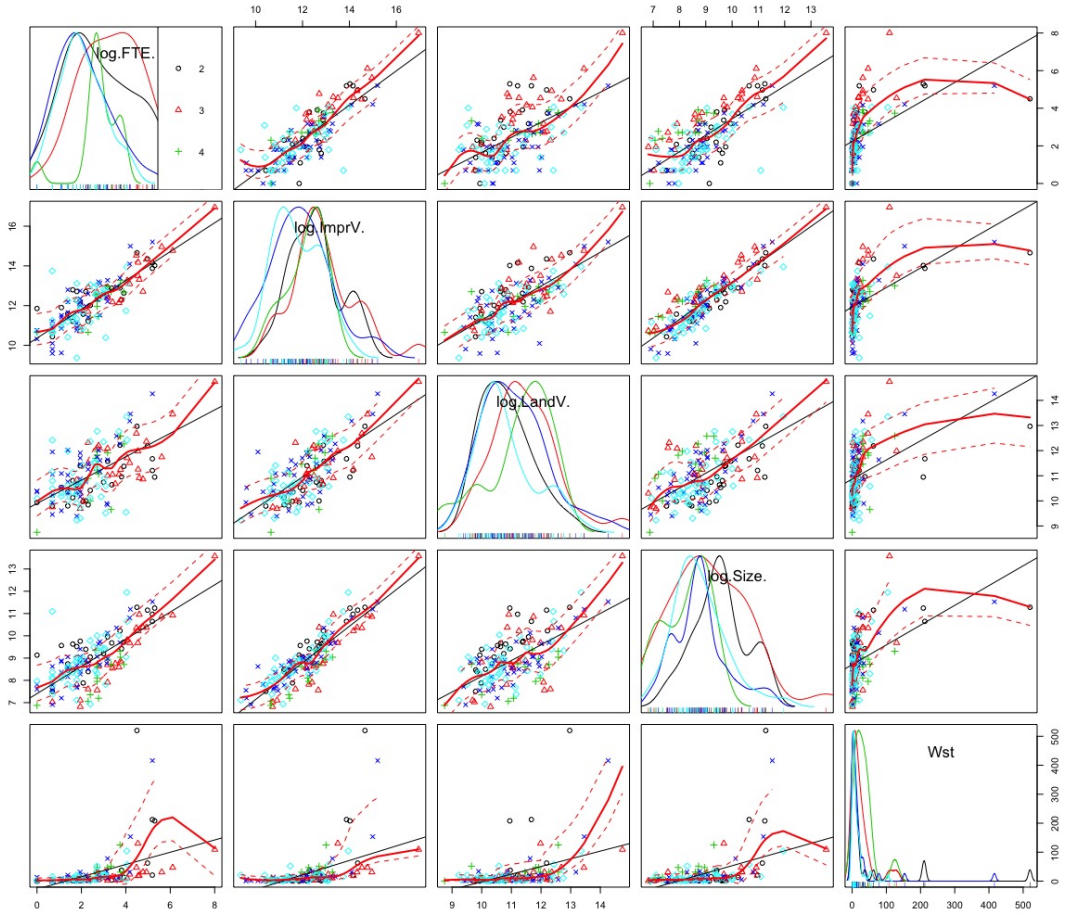

Figure 4.2: scatter plot matrix of response and regressors after using log-transformation on the regressors

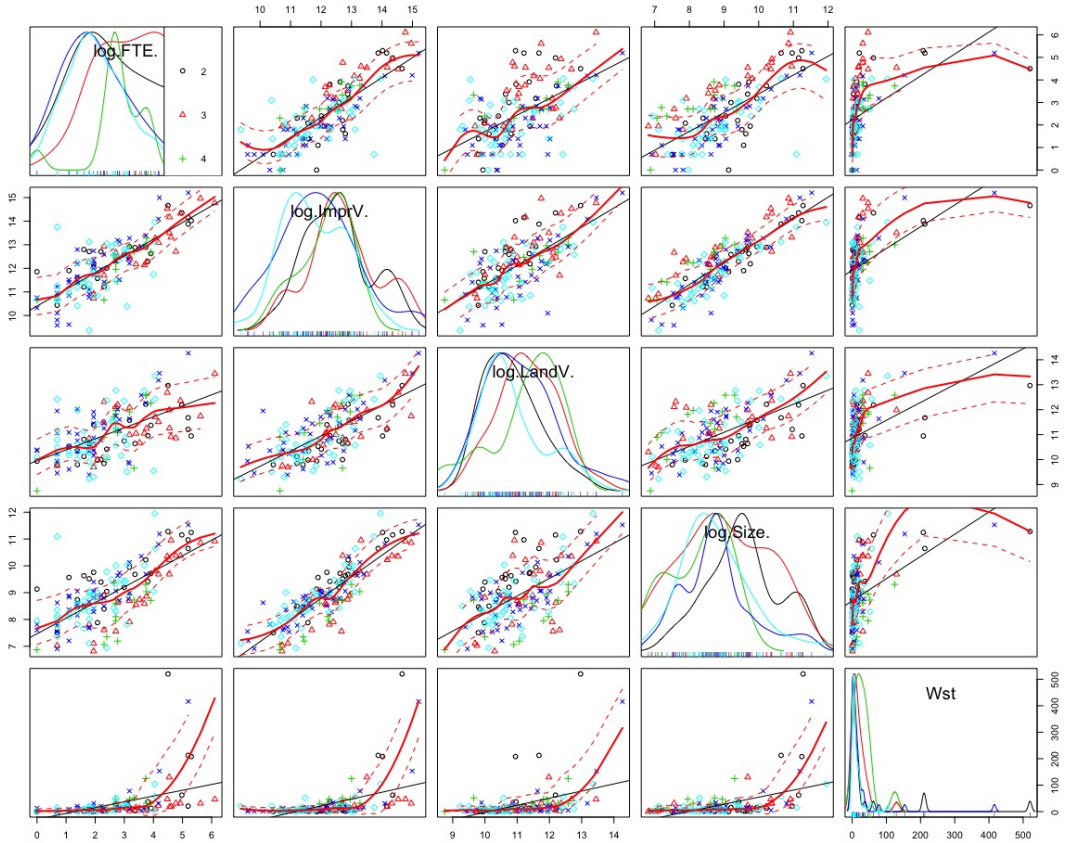

Figure 4.3: scatter plot matrix of response and regressors after using log-transformation on the regressors after dropping case 77

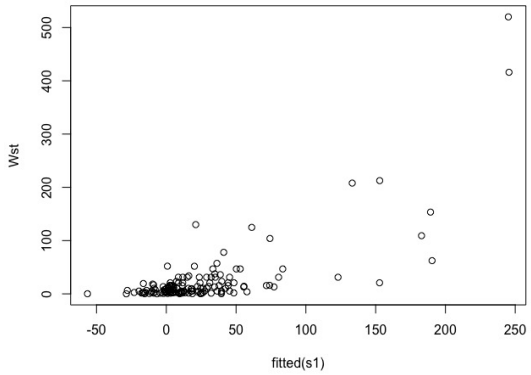

Figure 4.4: scatter plot between the response Wst and the fitted value from model selected by stepwise regression

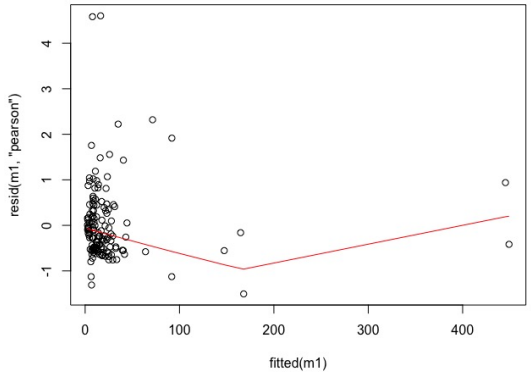

Figure 4.5: Power transforming the fitted value

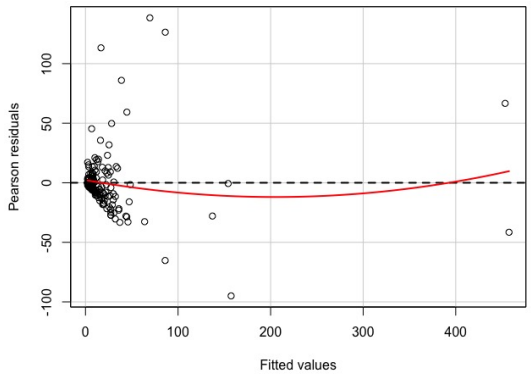

Figure 4.6: the residual plot of the model using 4th order polynomial of fitted value

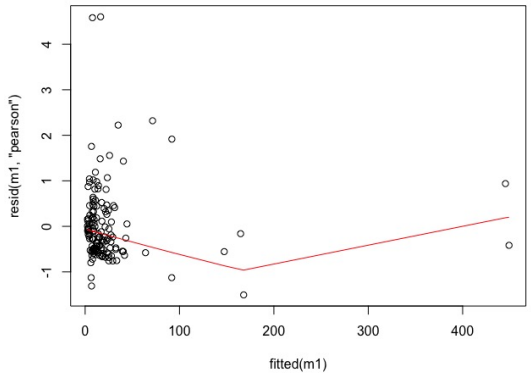

Figure 4.7: the residual plot of the weighted regression model using 4th order polynomial of fitted value

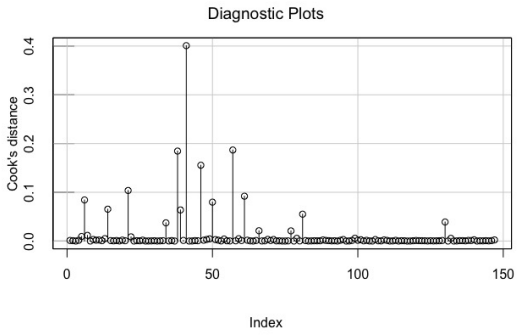

Figure 4.8: Cook's distance of the weighted regression model using 4th order polynomial of fitted value

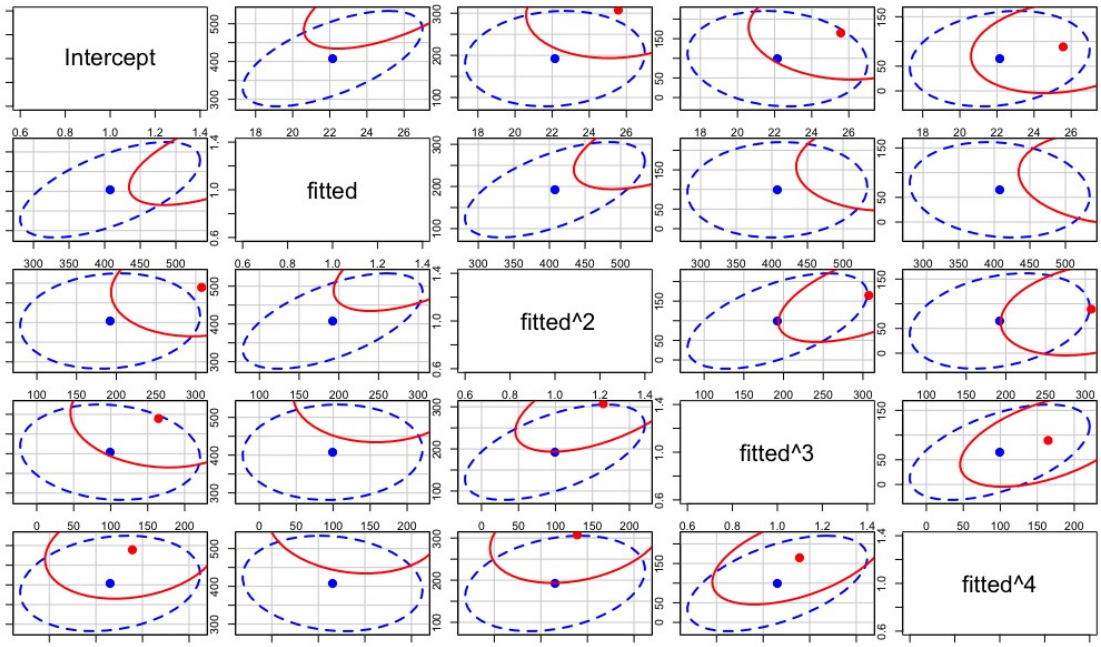

Figure 4.9: The confidence region scatter plot matrix of the model using original data(solid line) and the model removing case 41(dashed line).

4.3 CODES

```r
#read in data
assDat <- read.table("/Users/mikezhang/Documents/åŋęäźă17fall/STAT5302_TA/lab12/Waste.
    txt", header = FALSE)

#name the variables
nam <- c("FTE", "ImprV", "LandV", "Size", "Use", "Wst")
names(assDat) <- nam

#change the variable Use into factor
assDat$Use <- as.factor(assDat$Use)

#get the summary of the data
summary(assDat)

#draw the scatterplot matrix of the data
library(car)
scatterplotMatrix(~ FTE + ImprV + LandV + Size
                  + Wst |Use, data = assDat)

#there is a point that has huge FTE, ImprV, LandV
#and Size but small Wst
which(assDat$FTE > 2000)
which(assDat$ImprV > 1.5*10^7)
which(assDat$LandV > 2*10^6)
which(assDat$Size > 6*10^5)
assDat$Wst[77]

############################################################
#modeling
############################################################


#try the power transformation with that point (77)
pt1 <- powerTransform(cbind(FTE, ImprV, LandV, Size) ~ 1,
                      data = assDat)
#power transformation suggests using all log transformation
summary(pt1)


#try the power transformation without observation 77
pt2 <- powerTransform(cbind(FTE, ImprV, LandV, Size) ~ 1,
                      data = assDat[-77, ])
#It still suggests using all log transformation
summary(pt2)

#scatterplot matrix of variables after log transformation
#with observation 77
scatterplotMatrix(~ log(FTE) + log(ImprV) + log(LandV)
                  + log(Size) + Wst |Use, data = assDat)

#scatterplot matrix of variables after log transformation
#without observation 77
```

```r
scatterplotMatrix(~ log(FTE) + log(ImprV) + log(LandV)
                  + log(Size) + Wst |Use, data = assDat[-77, ])

#comment: we can see that whether dropping the observation
#77 makes little difference in the making the regressors
#linearly related

#variable screening using backward selection
s1 <- step(lm(Wst ~ (log(FTE) + log(ImprV) + log(LandV)
                     + log(Size) + Use)^2, data = assDat), direction = "backward")

#plot the response versus the fitted value
plot(Wst ~ fitted(s1), data = assDat)

#Try to do a power transformation on the fitted value
#the result shows that we may choose 4 as the power
with(assDat, invTranPlot(abs(fitted(s1)), Wst))



#choose 4 as a power and transform the fitted value
s2 <- lm(Wst ~ poly(I(fitted(s1)), 4), data = assDat)

#plot the residual plot
 #there is a huge cluster on the left
residualPlot(s2)

#test nonconstant variance
#the result shows that it is significant that
#the model has a non constant variance
ncvTest(s2)

#we can do the regression of square of residuals on
 #the fitted value to model the variance
we <- lm(resid(s2)^2 ~ fitted(s1), data = assDat)
#the result turns out that it is significant
summary(we)

#fit the model with the reciprocal of fitted we as the weight
ml <- lm(Wst ~ poly(I(fitted(s1)), 4), data = assDat,
         weights = 1/abs(fitted(we)))
#plot the residual plot, we can see that
 #there is still a cluster of points on the left.
plot(resid(ml, "pearson") ~ fitted(ml))
lines(lowess(fitted(ml), f = 1,
             resid(ml, "pearson")),col="red")

#the summary shows that almost all of the terms are very
 #significant
summary(ml)
#ncvTest shows that we can not reject the null hypothesis
 #that there is no nonconstant variance
ncvTest(ml)

```

```r
#the outlier test shows that there are 2 possible outliers
outlierTest(ml)

#plot the values
influenceIndexPlot(ml, vars = c("Cook"))
#find out the observation with largest Cook's distance
which(cooks.distance(ml) == max(cooks.distance(ml)))


##################################################
#plot the confidence region matrix of retaining that
#point and removing that point
##################################################

#create function cE to plot confidence ellipses:
#betas is a vector with length 2 to store the indices
#for betas.
#c is the confidence level
#dp is the indice for the observation to drop
ce <- function(betas, c, dp){

  #confidence region after dropping case dp
  confidenceEllipse(lm(Wst ~ poly(I(fitted(s1)[-dp]), 4),
                       data = assDat[-dp, ],
                       weights = 1/abs(fitted(we)[-dp])),
                    which.coef = betas,
                    levels = c,
                    col = "blue", lty = 2)
  #confidence region before dropping case dp for betaas
  confidenceEllipse(ml, which.coef = betas,
                    levels = c ,add=TRUE)

}


#there are 6 coefficients
d <- 5
#set the arrangement of the plots
par(mfrow = c(d,d))
#change the figure margins setting
par(mar=c(1,1,1,1))

#create a vector to store variable names
vname <- c("Intercept", "fitted"," fitted^2", "fitted^3",
           "fitted^4")
for (i in c(1:d)){
  for (j in c(1:d)){
    if (i == j){
      plot(1, type = "n")
      #text the variable name
      text(1,1, vname[i], cex = 2)
    }else{
      ce(c(i,j), 0.95, 41)
    }
```

```
  }
}

#set the figure margins to the default
par(mar=c(5.1, 4.1, 4.1, 2.1))
#set the figure arrangement to default
par(mfrow = c(1,1))
```

REFERENCES

[1] Weisberg, Sanford. Applied linear regression. Vol. 528. John Wiley & Sons, 2005, p194.