Huiwen Chen ID: 02156341 unique name:huiwenc

STATS 500 - Homework 5

Using the sat data, fit a model with total as the response and takers, ratio, salary and expend as predictors using the following methods:

1. Ordinary least squares

##Ordinary least squares

library(faraway)

##read in the data

data(sat)

attach(sat)

g1 <- lm (total~ takers+ratio+salary+expend, sat)

summary(g1, cor= T)

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) 1045.9715 52.8698 19.784 < 2e-16 takers -2.9045 0.2313 -12.559 2.61e-16 ratio -3.6242 3.2154 -1.127 0.266 salary 1.6379 2.3872 0.686 0.496 expend 4.4626 10.5465 0.423 0.674 Residual standard error: 32.7 on 45 degrees of freedom Multiple R-squared: 0.8246,Adjusted R-squared: 0.809 F-statistic: 52.88 on 4 and 45 DF, p-value: < 2.2e-16

Correlation of Coefficients: (Intercept) takers ratio salary

takers 0.04 ratio -0.80 0.25 salary 0.29 -0.35 -0.72 expend -0.53 0.09 0.75 -0.91

When fitting the model with ordinary least squares, we noticed that takers is very significant, while salary and expend is not really significant. And the correlation for ratio and salary is typically negative, the correlation for expend and salary is also strongly negative. The Residual standard error is so large.

##check for outlier-leverage points

>ti <- rstudent(g1)

> which.max(ti)

Utah

44

> 2\*(1-pt(m,df = 50-4-1))

[1] 0.003114645

> 0.05/50

[1] 0.001

## Compute Cook’s distance

cook <- cooks.distance(g1)

halfnorm (cook, nlab=4, ylab = "Cook's distances")

## Compute changes in coefficients

result.inf <- lm.influence(g1)

plot(result.inf$coef[,2], result.inf$coef[,3], xlab="Change in takers", ylab="Change in ratio")

## interactive tool to identify points by clicking

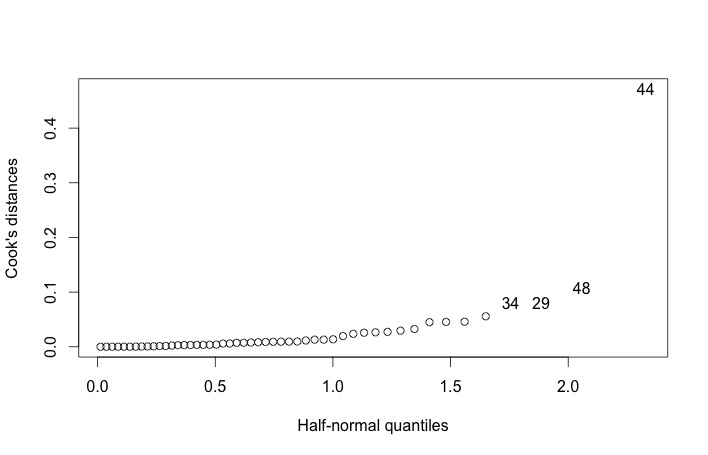
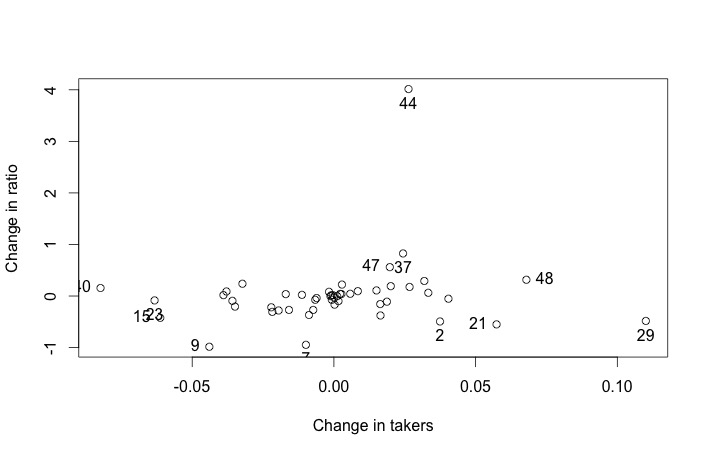
identify (result.inf$coef[, 2], result.inf$coef[, 3])

##romove #29,#44,#48 points

df1 = df[-c(29,44,48),]

result <- lm(total~ takers+ratio+salary+expend, data = df1)

summary(result)



First we check the p-value of the largest (externally) studentized residual is 0.0031, which is larger than level 0.001, we conclude that the point is not an outlier, Then no outlier can be seen in the regression model.

Then we calculate the cook’s distance of each point, in the left plot #29, #34, #44 and #48 points have larger cook’s distance. The right plot shows the leaveout-one differences in the coefficients related to takers and ratio. We find that lots of points stick out on the plot. Then we examine the effects of removing #29, #44 and #48 points below.

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept)1091.5571 45.6268 23.924 <2e-16 takers -3.1062 0.1902 -16.334 <2e-16 ratio -7.4880 2.9037 -2.579 0.0135 salary 2.4870 1.9702 1.262 0.2138 expend 3.7532 8.6665 0.433 0.6672

Residual standard error: 26.07 on 42 degrees of freedom

Multiple R-squared: 0.8902, Adjusted R-squared: 0.8797

F-statistic: 85.12 on 4 and 42 DF, p-value: < 2.2e-16

After removal, the absolute value of coefficient of ratio increases more than 100%. Also predictor ratio, salary and expend all become more significant, especially for ratio. The residual standard error becomes smaller and R-squared increases slightly.

1. Least absolute deviations

>library(quantreg) > glad <- rq(total~ takers+ratio+salary+expend, data = sat) > summary(glad)

Coefficients:

coefficients lower bd upper bd

(Intercept) 1090.89886 920.17149 1151.85075 takers -3.13961 -3.38485 -2.6647 ratio -7.26632 -10.73796 1.62341 salary 3.18313 -0.15788 5.41909 expend -0.79753 -8.88001 20.92522

Compared to least squares, most of the coefficients, except for coefficient of expend, are more close to those in OLS method after removing influential points. So identifying and removing bad and unusual points when applying OLS method can achieve a better fit as in LAD method. It also shows that only takers is a significant predictor, while in least squares method takers and ratio can be considered significant.

1. Huber’s robust regression

> library(MASS)

> gr <- rlm(total~ takers+ratio+salary+expend, data = sat)

> summary(gr)

Coefficients:

Value Std. Error t value

(Intercept) 1060.2074 49.8845 21.2533 takers -2.9778 0.2182 -13.6470 ratio -5.1254 3.0339 -1.6894 salary 2.0933 2.2525 0.9293 expend 3.9158 9.9510 0.3935 Residual standard error: 25.58 on 45 degrees of freedom

The numerical values of coefficients have changed a small amount when compared to those in OLS without unusual points, but in general they are close to each other and the general significance of the variables is almost the same.

4. Least trimmed squares

set.seed(123) glts <- ltsreg(total~ expend+ratio+salary+takers, data = sat) x <- df [, 1:4] bcoef <- matrix(0, nrow = 1000, ncol = 5) for (i in 1:1000){

newy <- glts$fit + glts$resid[sample(20, rep =T)] bcoef[i,] <- ltsreg(x, newy, nsamp = "best")$coef

} colnames(bcoef) <- names(coef(glts)) apply(bcoef, 2, function(x) quantile(x, c(0.025,0.975)))

library(ggplot2)

bcoef <- data.frame(bcoef)

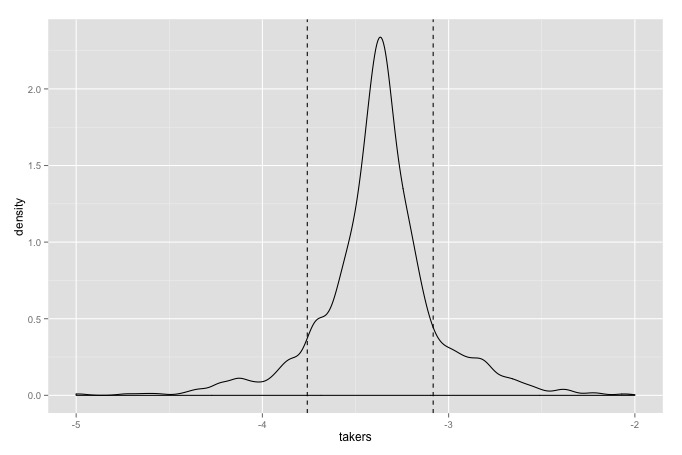
p1 <- ggplot(bcoef, aes(x= takers))+geom\_density()+xlim(-5,-2)

p1 + geom\_vline(xintercept = c(-3.759,-3.083),linetype= "dashed")

(Intercept) expend ratio salary takers

2.5% 1080.546 -4.536154 -19.413370 -0.8484121 -3.759272

97.5% 1261.343 26.047694 -8.614763 5.7348908 -3.083768



By LTS method, it is clear that both takers and ratio are significant, since 0 is outside confidence intervals of the two’s coefficients. The fitting result is much close to that of OLS method without unusual points, as ratio becomes much more significant after removing some influential points. But in OLS, the predictor salary also show significance, though not that obvious. Maybe it is because some bad or unusual points have not been removed, and LTS excluded these points.

Also, from the density distribution of coefficient of takers, we see that the distribution has longish tails, which suggests that the error may not be normally distributed. Robust regression like LAD, Buber’s regression and LTS can help solve the problem of lacking of fit, or we can use regression diagnostics in conjunction with least squares to identify bad and unusual points.