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

Call:

lm(formula = total ~ takers + ratio + salary + expend, data = sat)

Residuals:

Min 1Q Median 3Q Max

-90.531 -20.855 -1.746 15.979 66.571

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

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.7 on 45 degrees of freedom

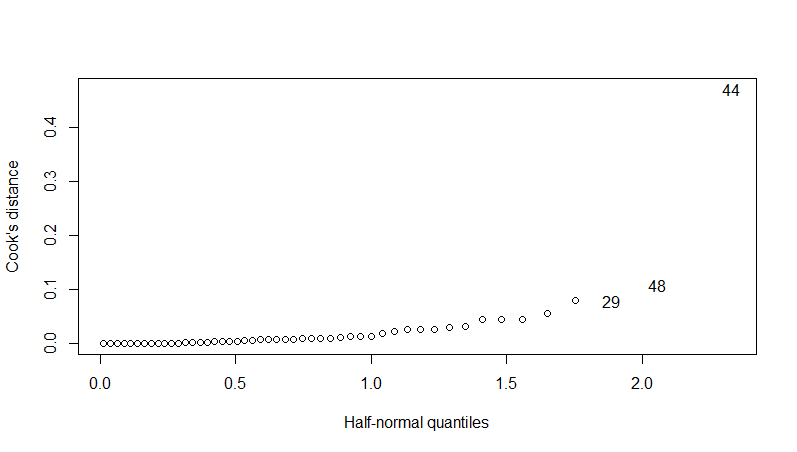
Multiple R-squared: 0.8246, Adjusted R-squared: 0.809

F-statistic: 52.88 on 4 and 45 DF, p-value: < 2.2e-16

Among the four predictors, takers reached significant.

Least squares works well when there are normal errors but performs poorly for long-tailed errors.

As we calculate the Cook’s distance, we found 3 significant observations, which are 29, 44, 48.



After remove these significant observations, the result is as follows.

Call:

lm(formula = total ~ takers + ratio + salary + expend, data = sat2)

Residuals:

Min 1Q Median 3Q Max

-48.848 -13.169 -1.669 13.500 62.144

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

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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 removing those significant observations, among the four predictors, takers and ratio reached significant.

1. Least absolute deviations

Call: rq(formula = total ~ takers + ratio + salary + expend, data = sat)

tau: [1] 0.5

Coefficients:

coefficients lower bd upper bd

(Intercept) 1090.89886 920.17149 1151.85075

takers -3.13961 -3.38485 -2.66479

ratio -7.26632 -10.73796 1.62341

salary 3.18313 -0.15788 5.41909

expend -0.79753 -8.88001 20.92522

Zero lies outside the interval for takers. So that the effect of takers reached significant.

1. Huber’s robust regression

Call: rlm(formula = total ~ takers + ratio + salary + expend, data = sat)

Residuals:

Min 1Q Median 3Q Max

-92.510 -17.701 -1.002 15.015 77.058

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

From the t value distribution we find that to reach significant with 45 degrees of freedom, |t value| has to be larger than 2.014. So takers reached significant.

1. Least trimmed squares

round(lts.reg$coef,3)

(Intercept) takers ratio salary expend

1090.082 -3.043 -8.443 1.124 14.625

With 1000 bootstrap resamples:

apply(bcoef,2,function(x) quantile(x, c(0.025,0.975)))

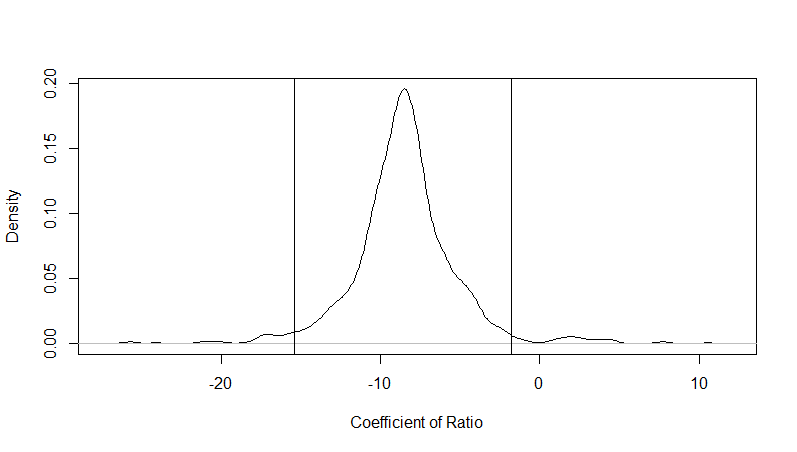
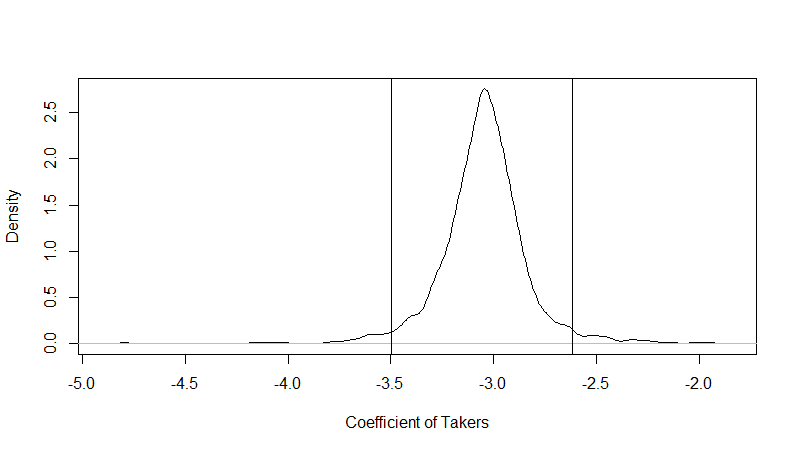
(Intercept) takers ratio salary expend

2.5% 947.2875 -3.498019 -15.411433 -3.645088 -7.886509

97.5% 1210.4065 -2.615203 -1.766441 5.976657 37.077860

Zero lies outside the confidence interval of takers and ratio. We are confident that there is a significant effect for takers and ratio.

The bootstrap distribution is shown below.



Bootstrap distribution of β^takers andβ^ratio with 95% confidence intervals

Results:

The Least Trimmed Squares Method is the most robust method, it is regardless of outliers. However, bootstrap has slightly difference each time applied.

The robust methods have the same significance for takers, but only LTS method revealed the significance of ratio.

Rcode used in this homework:

library(faraway)

library(quantreg)

library(MASS)

library(dplyr)

data(sat)

attach(sat)

olsreg<-lm(total~takers+ratio+salary+expend,data=sat)

cook<-cooks.distance(olsreg)

halfnorm(cook,nlab=3,ylab="Cook's distance")

sat2<-filter(sat,sat$total!='935')

sat2<-filter(sat2,sat2$total!='1076')

sat2<-filter(sat2,sat2$total!='932')

olsreg.re<-lm(total~takers+ratio+salary+expend,data=sat2)

ladreg<-rq(total~takers+ratio+salary+expend,data=sat)

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

lts.reg<-ltsreg(total~takers+ratio+salary+expend,data=sat)

x<-sat[,c(4,2,3,1)]

bcoef<-matrix(0,nrow=1000,ncol=5)

for(i in 1:1000){

newy<-lts.reg$fit+lts.reg$resid[sample(30,rep=T)]

bcoef[i,]<-ltsreg(x,newy,nsamp="best")$coef

}

colnames(bcoef)=names(coef(lts.reg))

apply(bcoef,2,function(x) quantile(x,c(0.025,0.975)))

round(lts.reg$coef,3)

plot(density(bcoef[,3]),xlab="Coefficient of Ratio",main="")

abline(v=quantile(bcoef[,3],c(0.025,0.975)))

plot(density(bcoef[,2]),xlab="Coefficient of Takers",main="")

abline(v=quantile(bcoef[,2],c(0.025,0.975)))