STAT 425 Homework 4

Giang Le

Multiple Linear Regression

Part II: Homework Questions – to be submitted

The `whitewines.csv` data set contains information related to white variants of the Portuguese "Vinho Verde" wine. Specifically, we have recorded the following information:

(a) fixed acidity, (b) volatile acidity, (c) citric acid, (d) residual sugar,
(b) chlorides , (f) free sulfur dioxide, (g) total sulfur dioxide,
(c) density, (i) pH, (j) sulphates, (k) alcohol, (l) quality (score between 0 and 10)

In this homework, our goal is to explain the relationship between alcohol level (dependent variable) and residual sugar, pH, density and fixed acidity.

Identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State decision rule and conclusion.

First I fit a linear regression model with alcohol as the dependent variable and residual sugar, pH, density, fixed acidity as independent variables.

```
wines = read.csv('whitewines.csv',header=TRUE,sep=";")
dim(wines)
```

```
## [1] 4898   12
```

```
wines.new = wines[,c("alcohol","residual.sugar","pH","density","fixed.acidity")]
wines.reg = lm(alcohol ~ residual.sugar + pH + density + fixed.acidity, data = wines.new)
summary(wines.reg)
```

```
##
## Call:
## lm(formula = alcohol ~ residual.sugar + pH + density + fixed.acidity,
##     data = wines.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3867 -0.2735 -0.0334  0.2200 16.9366
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.790e+02  4.540e+00  149.56   <2e-16 ***
## residual.sugar  2.367e-01  2.702e-03   87.58   <2e-16 ***
## pH              2.535e+00  5.281e-02   48.01   <2e-16 ***
## density        -6.858e+02  4.664e+00 -147.05   <2e-16 ***
## fixed.acidity   5.352e-01  9.858e-03   54.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4702 on 4893 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.854
## F-statistic:  7164 on 4 and 4893 DF,  p-value: < 2.2e-16
```

Perform a Bonferroni outlier test procedure to identify outliers

```r
sr.ex=rstudent(wines.reg);
n = dim(wines.new)[1]
p = dim(wines.new)[2]
sort(sr.ex, decreasing=TRUE)[1:5]
```

```
##      2782      3902      1654      1664      1418
## 43.513422  8.425736  6.464412  6.464412  4.103559
```

```r
qt(0.05/(n*2), n-p-1)
```

```
## [1] -4.417336
```

We have four observations with absolute residual values larger than 4.417336 so these are the outliers. The outliers have indices #2782, #3902, #1654, and #1664.

Obtain the diagonal elements of the hat matrix and identify any high leverage points. If any, are they good or bad?

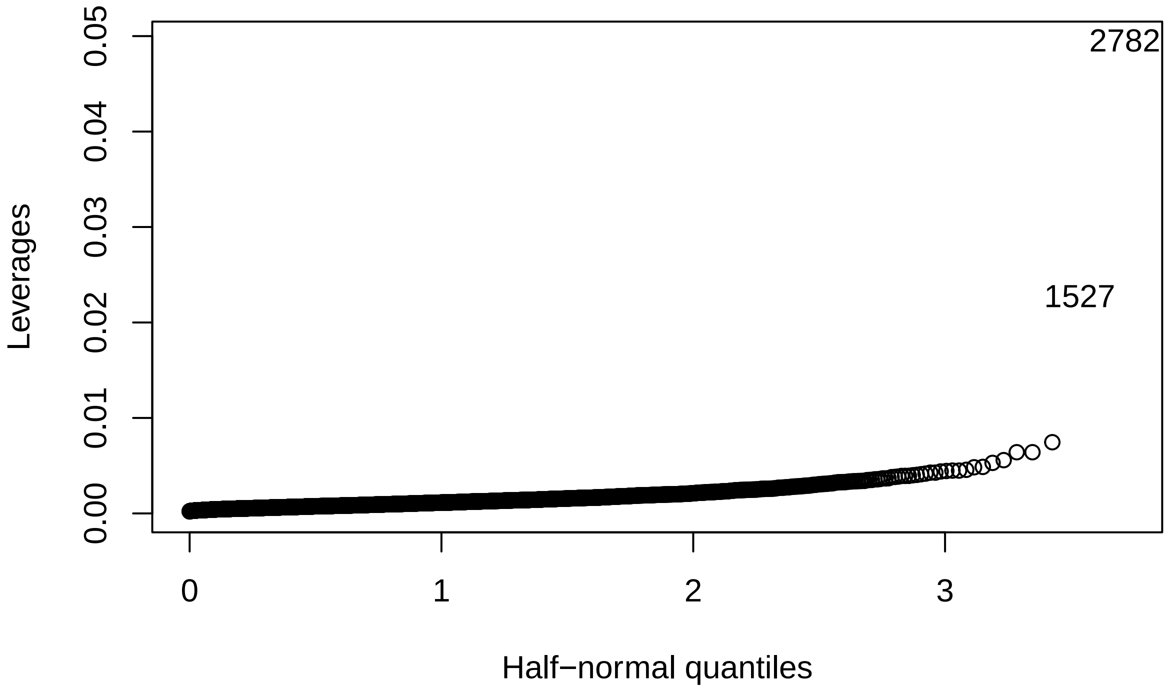
```r
diagonal_hat = lm.influence(wines.reg)$hat
head(diagonal_hat[diagonal_hat>(2*p/n)])
```

```
##          15          32          73          99         116         170
## 0.002069166 0.002601884 0.002970866 0.003260806 0.002446706 0.003052161
```

```r
install.packages("faraway",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/3_mgdyf12z7dvb8rt4d60nt80000gn/T//RtmpVRs7fs/downloaded_packages
```

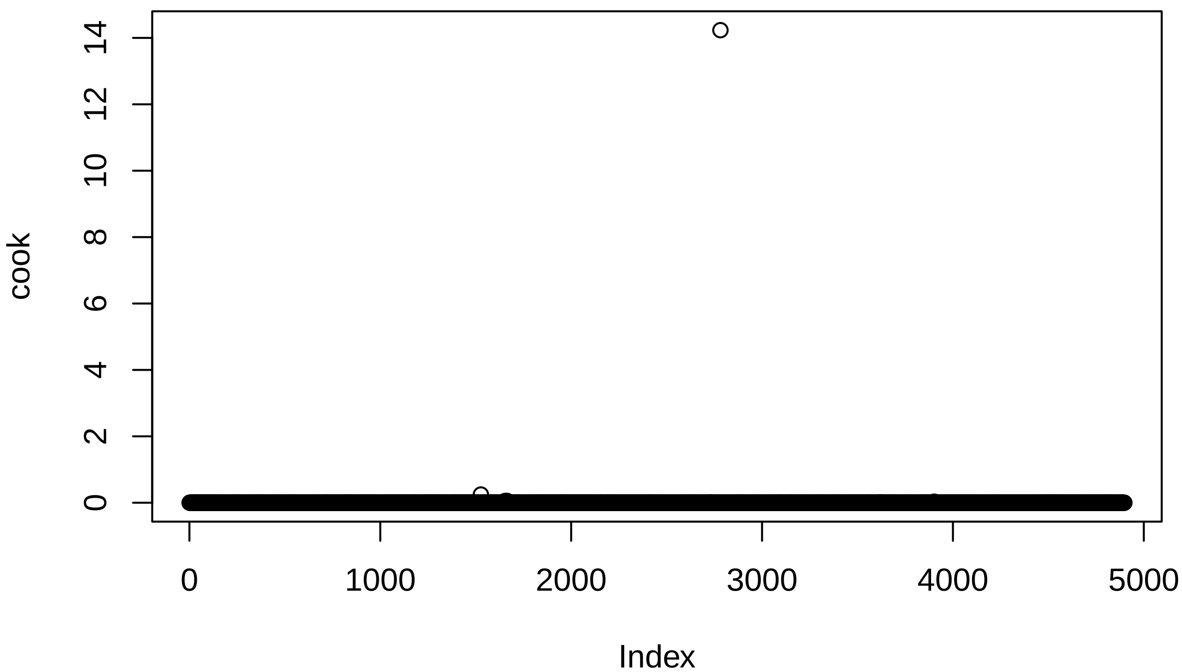
```r
library("faraway")
halfnorm(diagonal_hat, 2, ylab="Leverages")
```



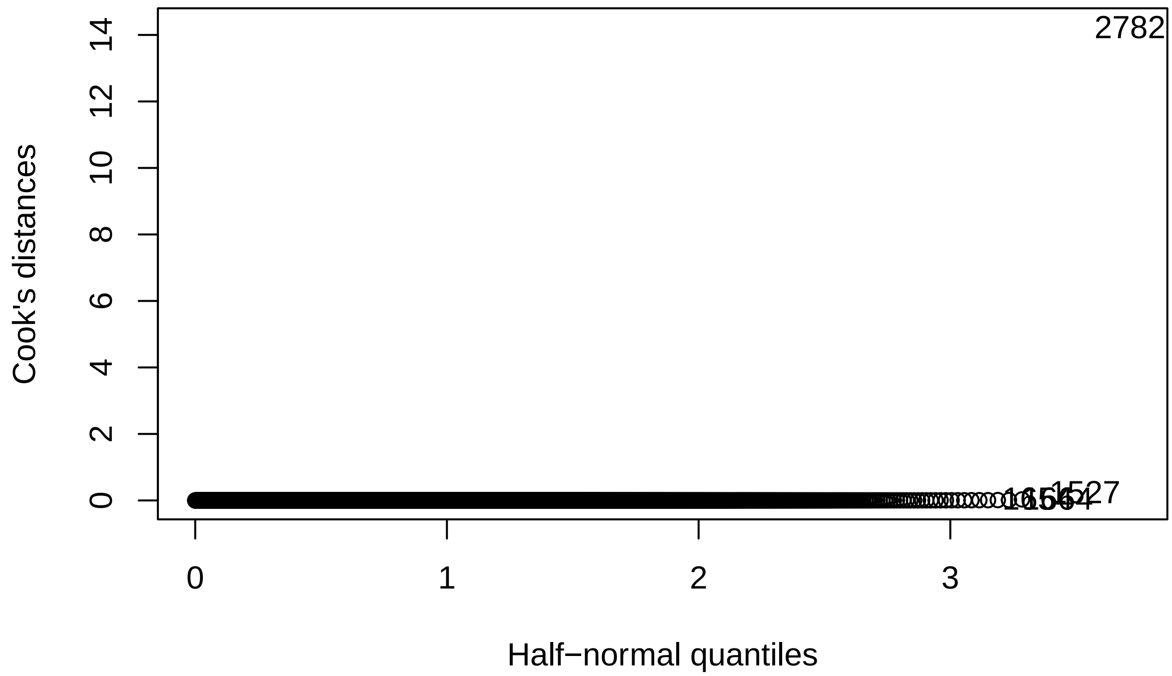
According to the plot, observations #1527 and #2782 are likely to be bad high leverage points.

Use Cook's distance to investigate whether there are any high influential points. What do you conclude?

```
cook = cooks.distance(wines.reg)
plot(cook)
```




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




```
max(cook)
```

```
## [1] 14.23251
```

```
which.max(cook)
```

```
## 2782
## 2782
```

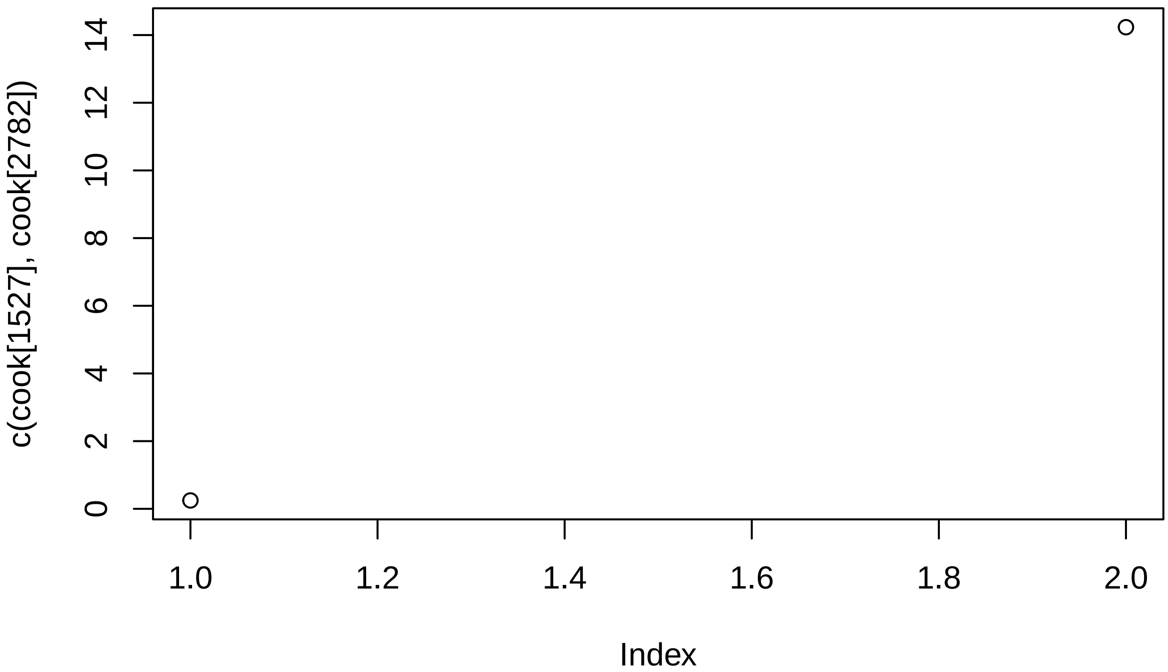
From there plots, the high influential points are likely to be observations #2782 and #1527.

Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

```
c(cook[1527], cook[2782])
```

```
##      1527      2782
##  0.247249 14.232508
```

```
plot(c(cook[1527], cook[2782]))
```



Their Cook's values are less than 1 in one case and more than 1 in other case. From this, we can evaluate that observation #2782 is an influential data point.

Predict the amount of alcohol of a white wine with residual.sugar = 1.7 , pH = 3, density = 1, fixed.acidity = 6.3 with an appropriate 95% confidence interval.

Below is the 95% confidence interval for alcohol with the given data.

```
new_input = data.frame(residual.sugar=1.7, pH=3, density=1, fixed.acidity=6.3)
predict(wines.reg, new=new_input, interval="confidence")
```

```
##        fit      lwr      upr
## 1 4.533544 4.440764 4.626325
```

Predict the amount of alcohol of a white wine with residual.sugar = 67 , pH = 4, density = 1.1, fixed.acidity = 15 with an appropriate 95% prediction interval.

Below is the 95% prediction interval for alcohol with the given data.

```
new_input1 = data.frame(residual.sugar=67, pH=4, density=1.1, fixed.acidity=15)
predict(wines.reg, new=new_input1, interval="prediction")
```

```
##         fit       lwr       upr
## 1 -41.40047 -42.52562 -40.27533
```

Construct a 95% confidence region for the slope coefficients of pH and density. What do you conclude about the statistical significance of β_{pH} and $\beta_{density}$?

```
install.packages("ellipse",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/3_mgdyf12z7dvb8rt4d60nt80000gn/T//RtmpVRs7fs/downloaded_packages
```

```
library("ellipse")
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

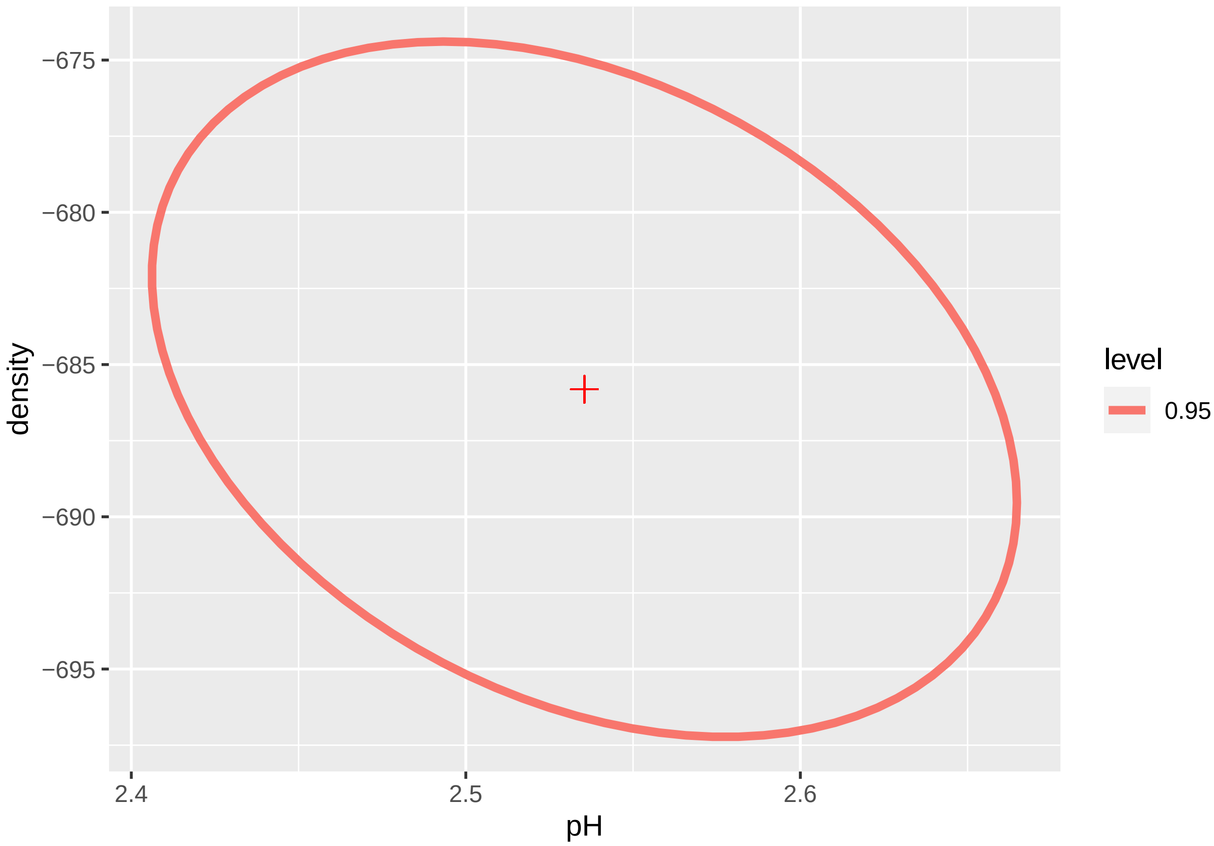
```
betas = ellipse(wines.reg, c(3,4))
head(betas)
```

```
##             pH   density
## [1,] 2.610530 -679.1795
## [2,] 2.603703 -678.6032
## [3,] 2.596600 -678.0560
## [4,] 2.589252 -677.5399
## [5,] 2.581686 -677.0572
## [6,] 2.573935 -676.6097
```

```
names(betas) = c("pH","density");
betas = data.frame(betas);
betas[, 'level']=as.factor(c(rep(0.95, dim(betas)[1])));

install.packages("ggplot2",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/3_mgdyf12z7dvb8rt4d60nt80000gn/T//RtmpVRs7fs/downloaded_packages
```

```
library("ggplot2")
ggplot(data=betas, aes(x=pH, y=density, colour=level)) +
  geom_path(aes(linetype=level), size=1.5) +
  geom_point(x=coef(wines.reg)[3], y=coef(wines.reg)[4], shape=3, size=3, colour='red') +
  geom_point(x=0, y=0, shape=1, size=3, colour='red')
```



As (0,0) is not in the ellipsoid, we can conclude that both β_{pH} and $\beta_{density}$ are statistically significant.

Regress `alcohol` against `fixed acidity` and construct a 95% simultaneous confidence band for the fitted regression line.

```
wines.small = wines[,c("alcohol","fixed.acidity")]
wines.reg.small = lm(alcohol ~  fixed.acidity, data = wines.small)
summary(wines.reg.small)
```

```
##
## Call:
## lm(formula = alcohol ~ fixed.acidity, data = wines.small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9823 -1.0768 -0.1356  0.8699  3.6056
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.72264    0.14289  82.041   <2e-16 ***
## fixed.acidity -0.17628    0.02069  -8.521   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.222 on 4896 degrees of freedom
## Multiple R-squared:  0.01461,    Adjusted R-squared:  0.01441
## F-statistic:  72.6 on 1 and 4896 DF,  p-value: < 2.2e-16
```

I construct a 95% simultanous confidence band for the fitted regression line.

```r
install.packages("ALSM", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/3_mgdyf12z7dvb8rt4d60nt80000gn/T//RtmpVRs7fs/downloaded_packages
```

```r
library("ALSM")
```

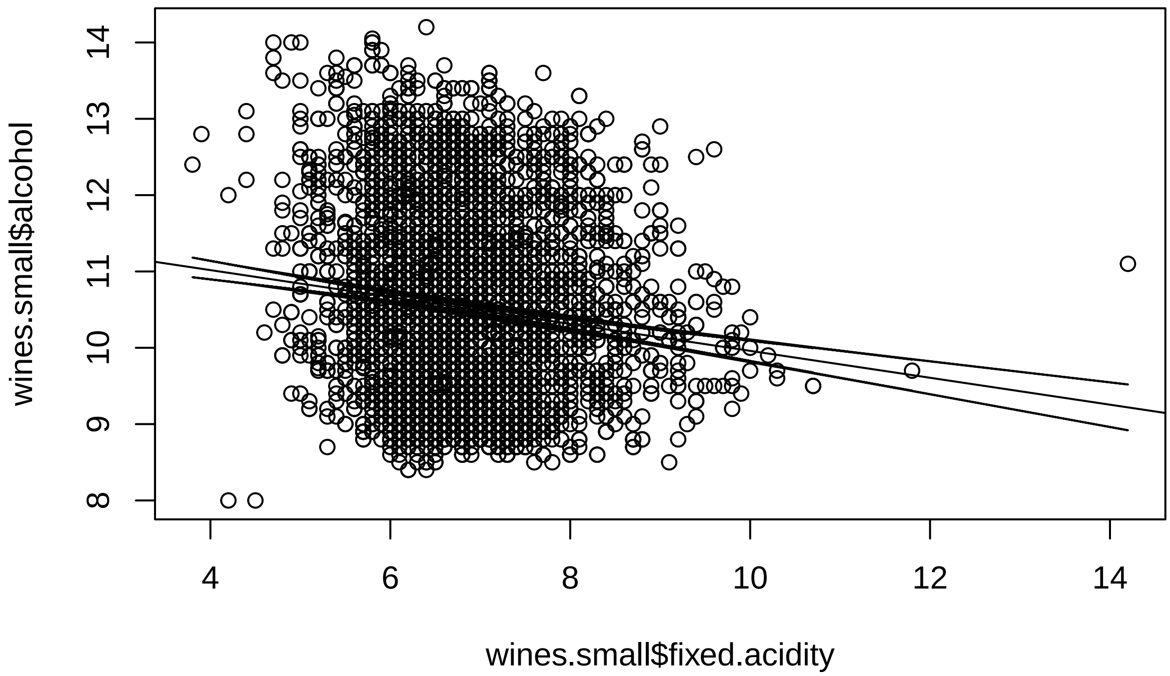
```
## Loading required package: leaps

## Loading required package: SuppDists

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:ellipse':
##
##     ellipse

## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```r
conf_band <- ci.reg(wines.reg.small, newdata=wines[,c("fixed.acidity")], type = c("b"), alpha = 0.05)
head(ci.reg(wines.reg.small, newdata=wines[,c("fixed.acidity")], type = c("b"), alpha = 0.05))
```

```
##   fixed.acidity      Fit Lower.Band Upper.Band
## 1           7.0 10.48867   10.41042   10.56691
## 2           6.3 10.61207   10.51978   10.70435
## 3           8.1 10.29476   10.15729   10.43222
## 4           7.2 10.45341   10.37010   10.53673
## 5           7.2 10.45341   10.37010   10.53673
## 6           8.1 10.29476   10.15729   10.43222
```

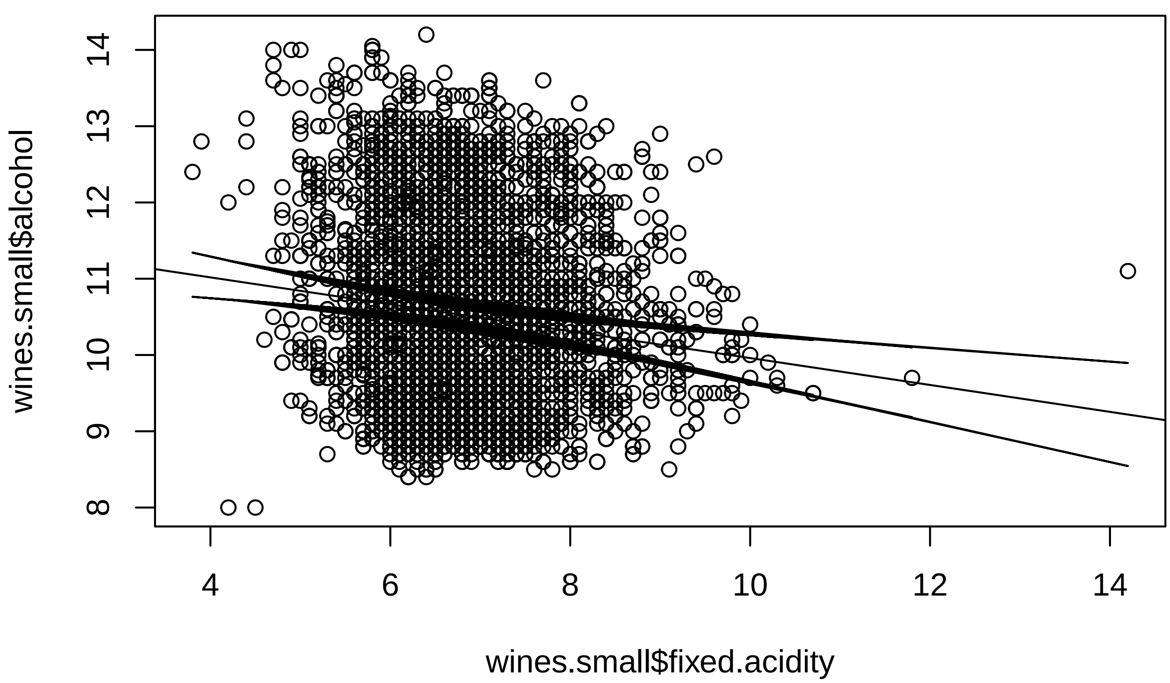
Plot the raw data corresponding to question (h), fitted regression line, 95% point-wise confidence intervals and 95% confidence band calculated in (h). What do you observe?

```r
new_input2 = data.frame(fixed.acidity=wines.small$fixed.acidity)
pointwise <- data.frame(predict(wines.reg.small, new=new_input2, interval="confidence"))
plot(x=wines.small$fixed.acidity, y=wines.small$alcohol)
abline(wines.reg.small)
lines(wines.small$fixed.acidity, pointwise$lwr)
lines(wines.small$fixed.acidity, pointwise$upr)
```



Here is the plot with the 95% confidence band and the regression line.

```
plot(x=wines.small$fixed.acidity, y=wines.small$alcohol)
abline(wines.reg.small)
lines(wines.small$fixed.acidity, conf_band$Lower.Band)
lines(wines.small$fixed.acidity, conf_band$Upper.Band)
```



We observe that the 95% confidence band is wider than the 95% pointwise confidence interval.