Homework7

Devin Etcitty

4/18/2017

STAT4205

dce2108

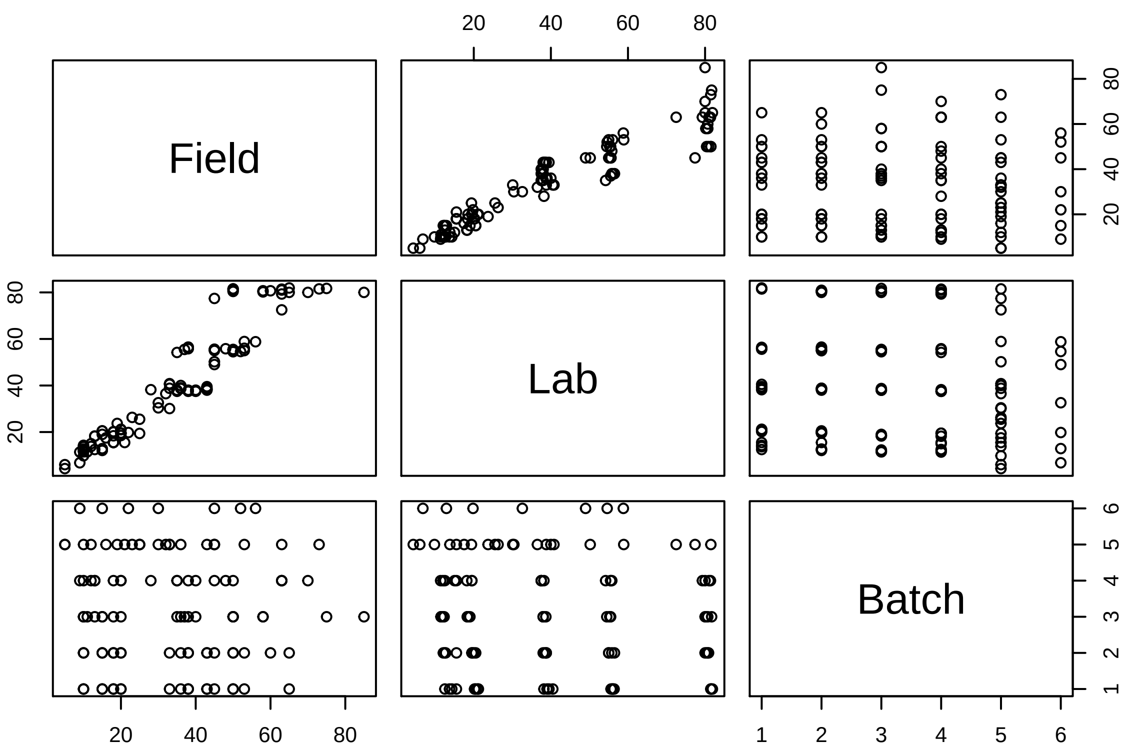
```r
library(alr4)
library(ggplot2)
library(tidyverse)
```

Problem 9.3

```r
head(pipeline)
```

```
##   Field  Lab Batch
## 1    18 20.2     1
## 2    38 56.0     1
## 3    15 12.5     1
## 4    20 21.2     1
## 5    18 15.5     1
## 6    36 39.0     1
```

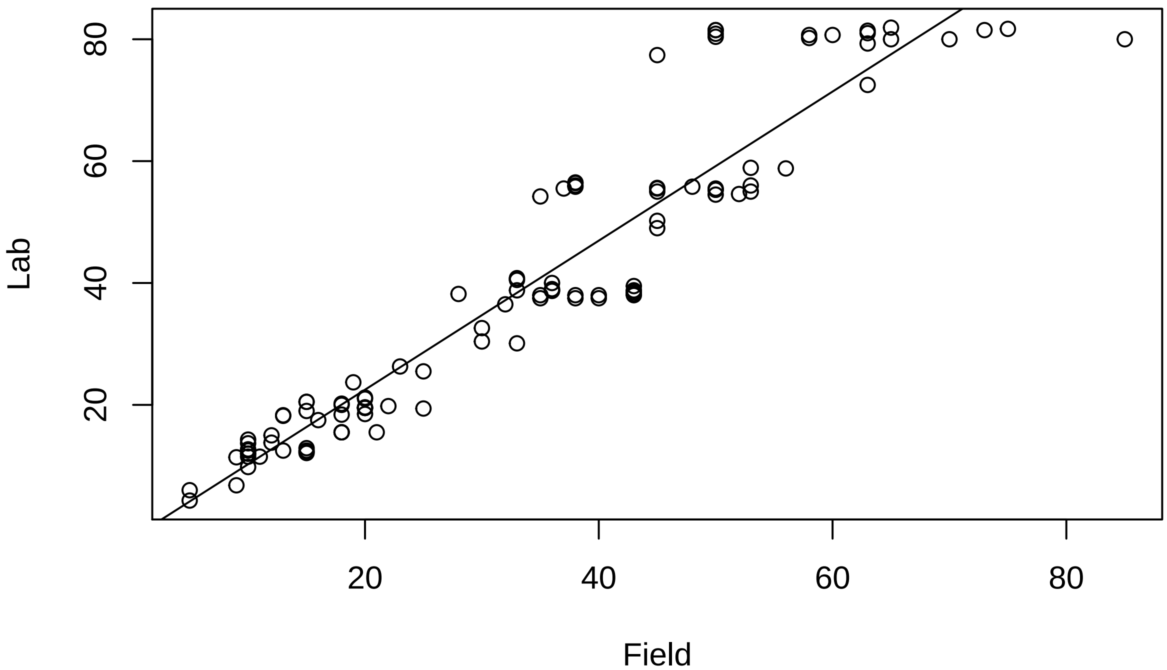
```r
plot(pipeline)
```



```r
labfield <- lm(Lab ~ Field, data=pipeline)
lab.line <- labfield$coefficients
lab.line
```

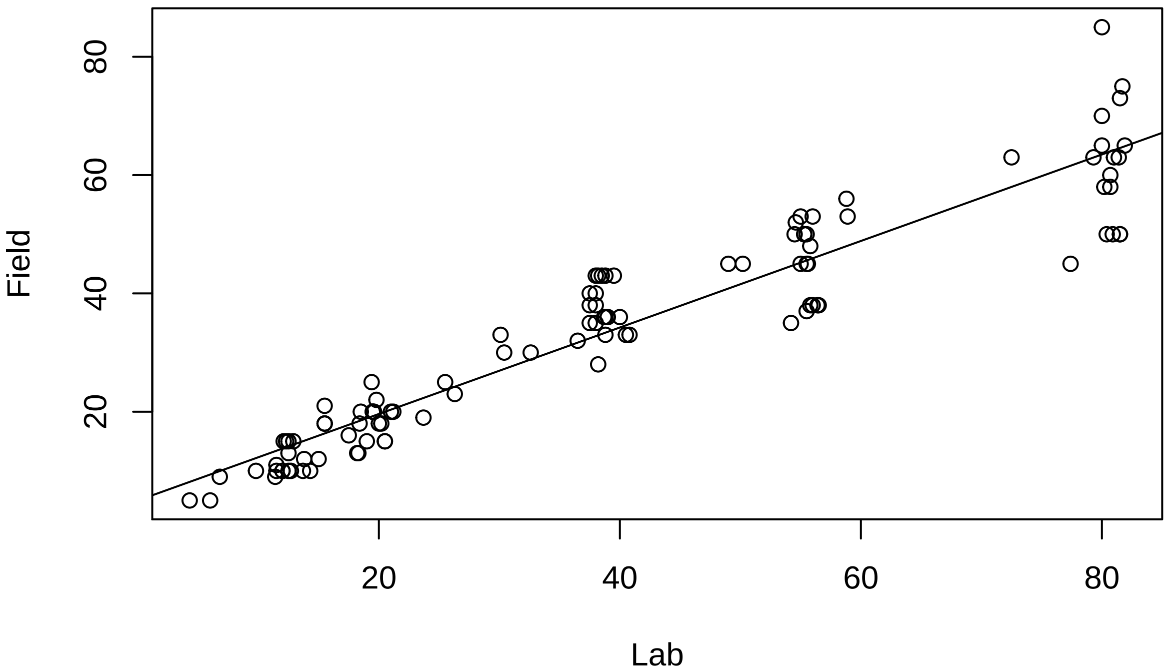
```
## (Intercept)       Field
##   -1.967500    1.222968
```

```r
plot(pipeline[1:2])
abline(lab.line[1], lab.line[2])
```

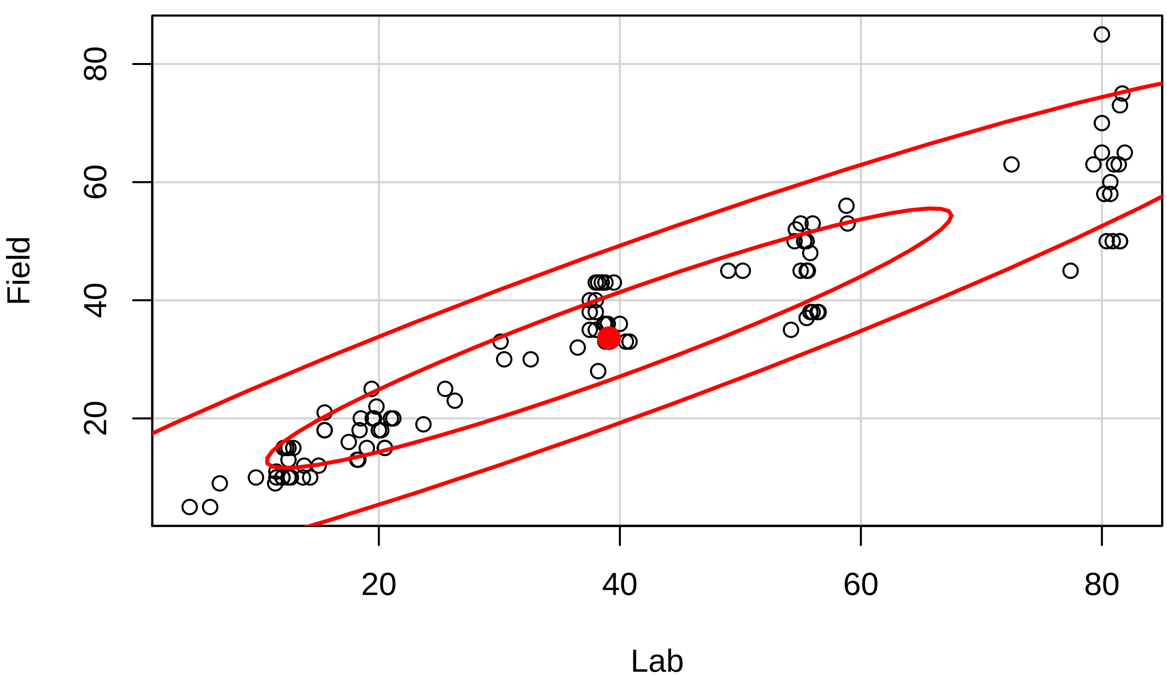



```r
m1.res <- labfield$residuals
```

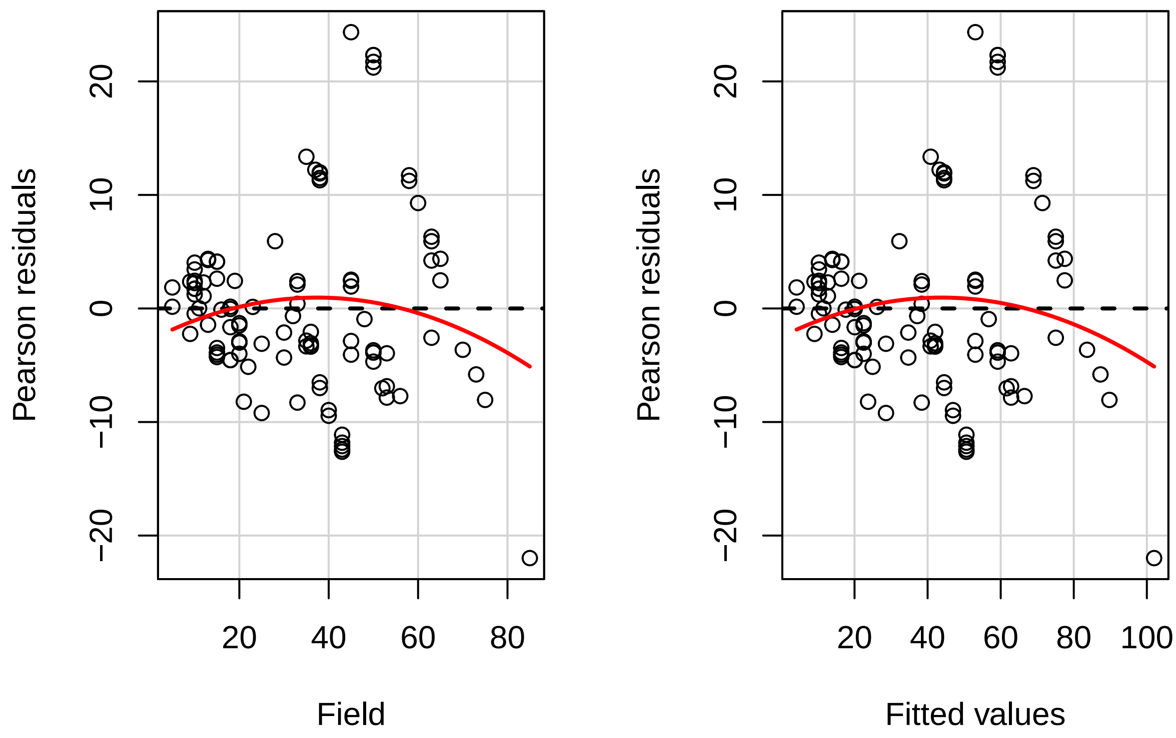
```r
fieldlab <- lm(Field ~ Lab, data=pipeline)
field.line <- fieldlab$coefficients
#field.line
plot(pipeline[2:1])
abline(field.line[1], field.line[2])
```




```
n <- dim(pipeline)[1]
levs <- pf(99*c(.5), 2, n) - 1/n
with(pipeline, dataEllipse(Lab, Field),
     levels=levs - 1/n, xlim=c(-50, 200), ylim=c(-50, 120))
```




```
residualPlots(labfield)
```




```
##              Test stat Pr(>|t|)
## Field           -1.303    0.196
## Tukey test      -1.303    0.193
```

The plot suggests non-constant variance since the curvature of the fitted values versus the residuals is quadratic.

```r
ncvTest(labfield)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 29.58568    Df = 1     p = 5.349868e-08
```

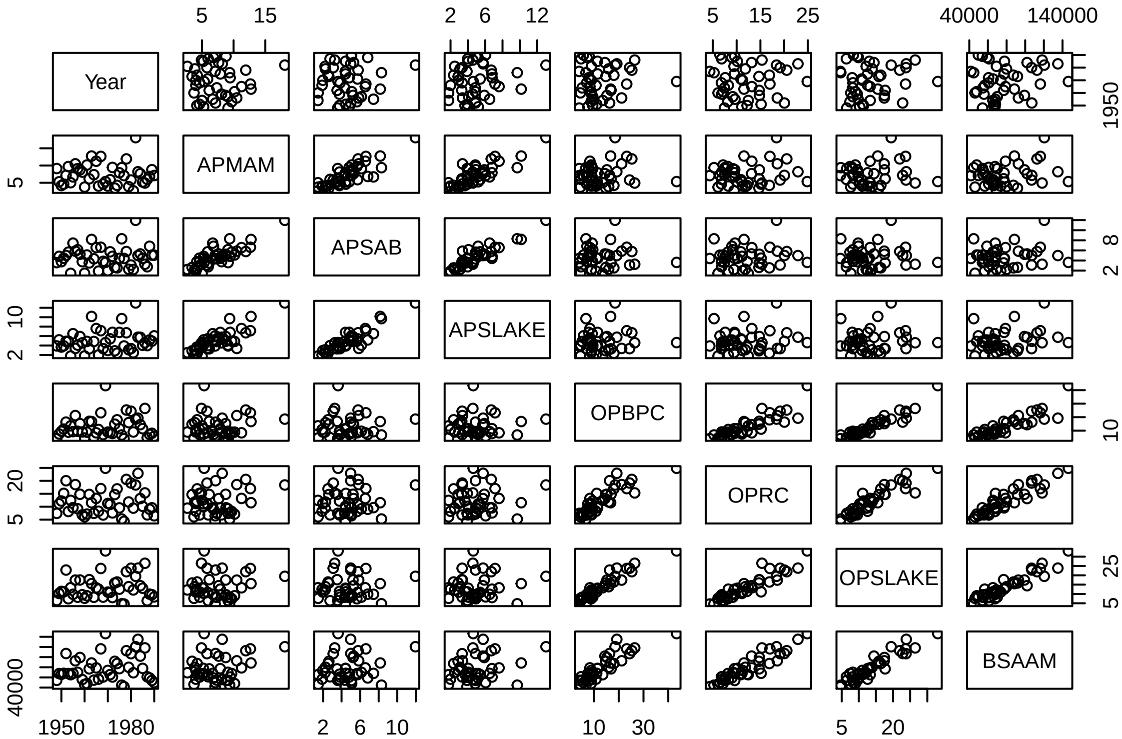
There is a lot p-value This means that the model is non-constant

Problem 9.8

```r
head(water)
```

```
##   Year APMAM APSAB APSLAKE OPBPC  OPRC OPSLAKE  BSAAM
## 1 1948  9.13  3.58    3.91  4.10  7.43    6.47  54235
## 2 1949  5.28  4.82    5.20  7.55 11.11   10.26  67567
## 3 1950  4.20  3.77    3.67  9.52 12.20   11.35  66161
## 4 1951  4.60  4.46    3.93 11.14 15.15   11.13  68094
## 5 1952  7.15  4.99    4.88 16.34 20.05   22.81 107080
## 6 1953  9.70  5.65    4.91  8.88  8.15    7.41  67594
```

```r
plot(water)
```



The problem asks:

“Draw residual plots for the mean function described in Problem 8.3.4 for the California water data, and comment on your results.”

“Test for curvature as a function of fitted values”

Q: What is the model?

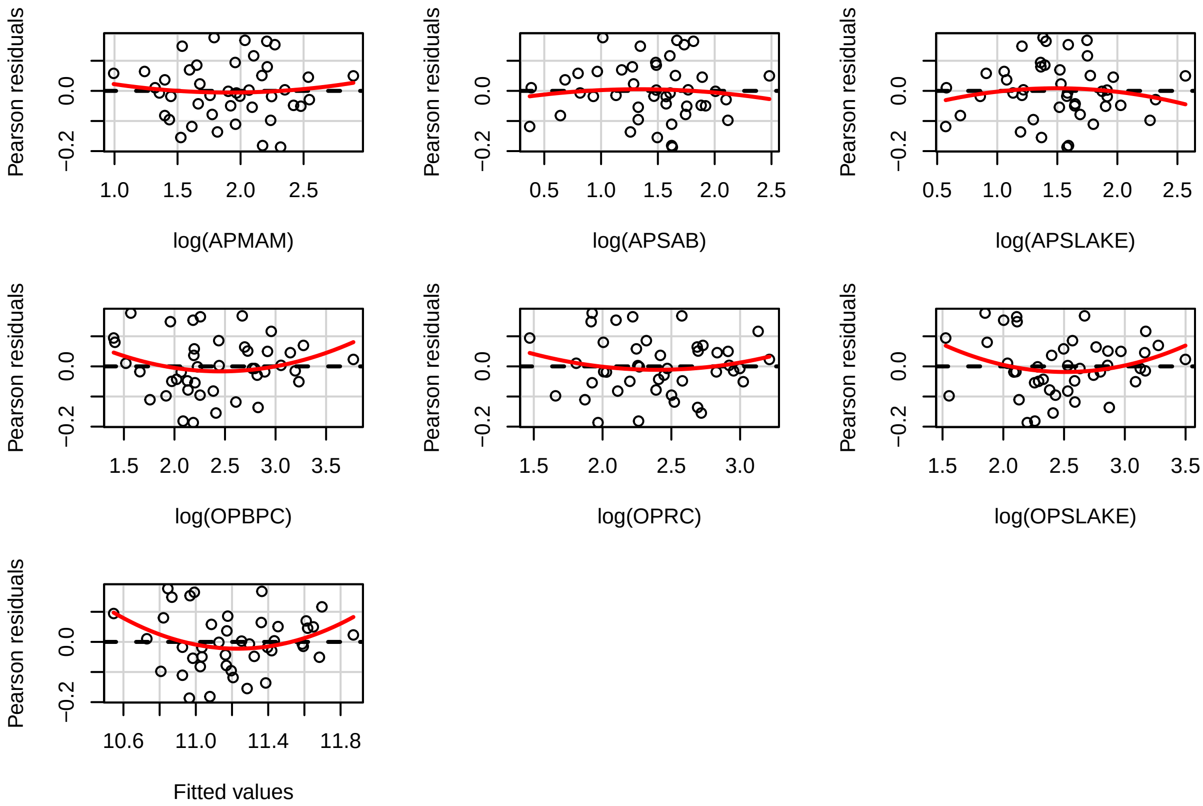
log(BSAMM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) + log(OPBPC) + log(OPRC) + log(OPSLAKE)

```
prob9.3 <- lm(log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) + log(OPBPC) + log(OPRC) + log(OPSLAKE
summary(prob9.3)
```

```
##
## Call:
## lm(formula = log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +
##     log(OPBPC) + log(OPRC) + log(OPSLAKE), data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18671 -0.05264 -0.00693  0.06130  0.17698
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.46675    0.12354  76.626  < 2e-16 ***
## log(APMAM)   -0.02033    0.06596  -0.308  0.75975
## log(APSAB)   -0.10303    0.08939  -1.153  0.25667
## log(APSLAKE)  0.22060    0.08955   2.463  0.01868 *
## log(OPBPC)    0.11135    0.08169   1.363  0.18134
## log(OPRC)     0.36165    0.10926   3.310  0.00213 **
## log(OPSLAKE)  0.18613    0.13141   1.416  0.16524
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1017 on 36 degrees of freedom
## Multiple R-squared:  0.9098, Adjusted R-squared:  0.8948
## F-statistic: 60.54 on 6 and 36 DF,  p-value: < 2.2e-16
```

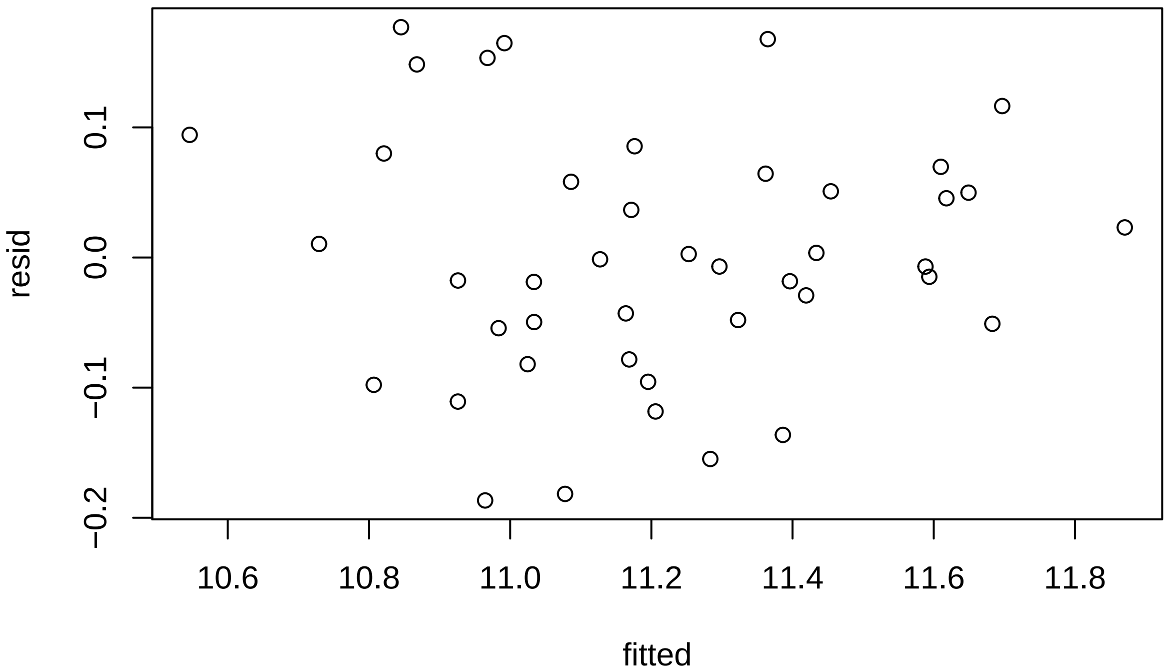
```
residualPlots(prob9.3)
```




```
##                Test stat Pr(>|t|)
## log(APMAM)         0.450    0.656
## log(APSAB)        -0.465    0.645
## log(APSLAKE)      -0.852    0.400
## log(OPBPC)         1.385    0.175
## log(OPRC)          0.839    0.407
## log(OPSLAKE)       1.630    0.112
## Tukey test         1.839    0.066
```

```
resid <- prob9.3$residuals
fitted <- prob9.3$fitted.values

plot(fitted, resid)
```

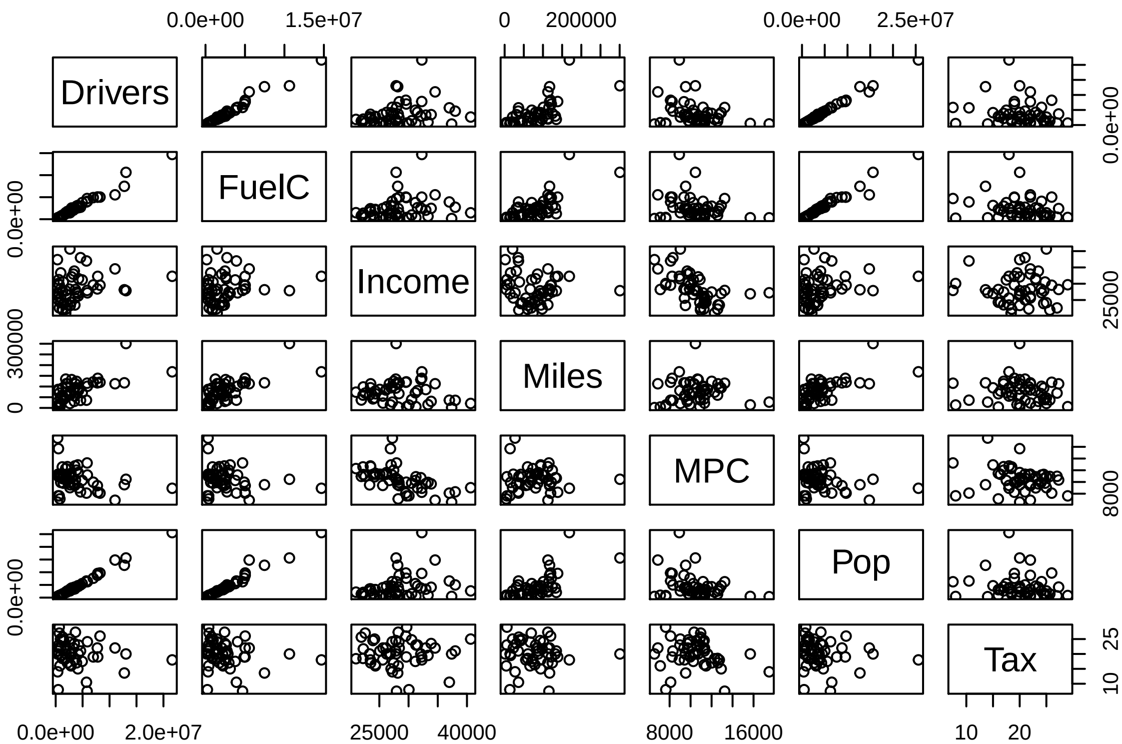


Problem 9.11

```r
head(fuel2001)
```

```
##     Drivers    FuelC Income  Miles      MPC      Pop  Tax
## AL  3559897  2382507  23471  94440 12737.00  3451586 18.0
## AK   472211   235400  30064  13628  7639.16   457728  8.0
## AZ  3550367  2428430  25578  55245  9411.55  3907526 18.0
## AR  1961883  1358174  22257  98132 11268.40  2072622 21.7
## CA 21623793 14691753  32275 168771  8923.89 25599275 18.0
## CO  3287922  2048664  32949  85854  9722.73  3322455 22.0
```

```r
plot(fuel2001)
```




```r
names(fuel2001)
```

```
## [1] "Drivers" "FuelC"   "Income"  "Miles"   "MPC"     "Pop"     "Tax"
```

Linear Model

$$E(fuel|X) = B_0 + B_1 Tax + B_2 Dlic + B_3 Income + B_4 log(miles)$$

Transform data

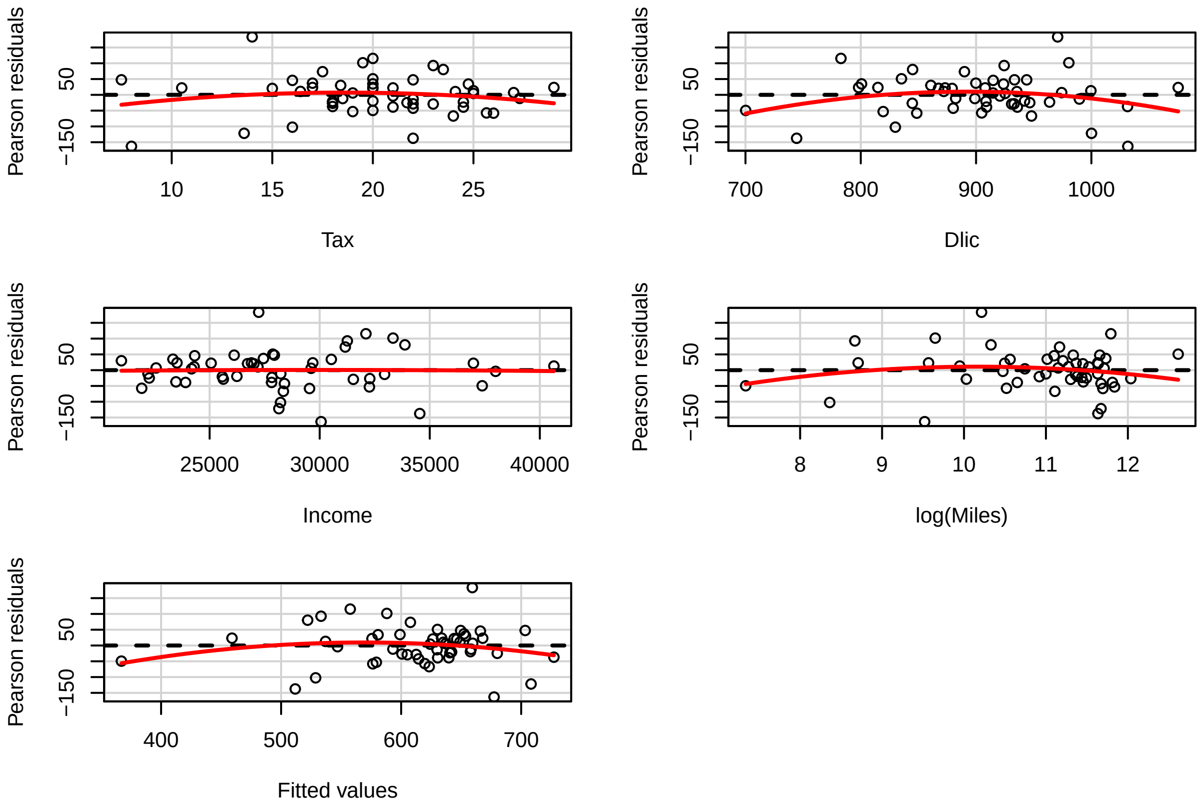
```r
fuel2001 <- transform(fuel2001,
Dlic=1000 * Drivers/Pop,
Fuel=1000 * FuelC/Pop)
```

```r
model9.11 <- lm( Fuel ~ Tax + Dlic + Income + log(Miles), data=fuel2001)
summary(model9.11)
```

```
##
## Call:
## lm(formula = Fuel ~ Tax + Dlic + Income + log(Miles), data = fuel2001)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.145  -33.039    5.895   31.989  183.499
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.192845 194.906161   0.791 0.432938
```

```
## Tax           -4.227983   2.030121  -2.083 0.042873 *
## Dlic           0.471871   0.128513   3.672 0.000626 ***
## Income        -0.006135   0.002194  -2.797 0.007508 **
## log(Miles)    26.755176   9.337374   2.865 0.006259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.89 on 46 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.4679
## F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

```
residualPlots(model9.11)
```




```
##             Test stat Pr(>|t|)
## Tax            -1.077    0.287
## Dlic           -1.922    0.061
## Income         -0.084    0.933
## log(Miles)     -1.347    0.185
## Tukey test     -1.446    0.148
```

```
hat_9.11 <- hatvalues(model9.11)
hat_9.11
```

```
##         AL         AK         AZ         AR         CA         CO
## 0.09767199 0.25609617 0.03339132 0.06348242 0.08668537 0.09825485
##         CT         DE         DC         FL         GA         HI
## 0.26953288 0.12350939 0.41491327 0.10088750 0.18246091 0.20572692
##         ID         IL         IN         IA         KS         KY
## 0.06429223 0.08496379 0.05534860 0.03644005 0.03632232 0.06000338
##         LA         ME         MD         MA         MI         MN
```

```
## 0.10182252 0.05552660 0.06745367 0.12277212 0.03872358 0.10573833
##         MS         MO         MT         NE         NV         NH
## 0.09716522 0.04125576 0.11430006 0.06085993 0.04945254 0.09306405
##         NJ         NM         NY         NC         ND         OH
## 0.18395399 0.06650900 0.16237155 0.04755207 0.03267468 0.03689858
##         OK         OR         PA         RI         SC         SD
## 0.10086971 0.04489247 0.07853836 0.21620511 0.05408396 0.03504750
##         TN         TX         UT         VT         VA         WA
## 0.03062388 0.09460381 0.06589521 0.17117439 0.03779511 0.05252210
##         WV         WI         WY
## 0.10375252 0.08213607 0.08378222
```

```
states <- c("AK", "NY", "HI", "WY", "DC")
hat_fivestates <- hat_9.11[states]
hat_fivestates
```

```
##         AK         NY         HI         WY         DC
## 0.25609617 0.16237155 0.20572692 0.08378222 0.41491327
```

```
fuel <- c(514.279, 374.164, 426.349, 842.792, 317.492)
ehat_fuel <- c(-163.145, -137.599, -102.409, 183.499, -49.452)
hat_fuel <- c(0.256, 0.162, 0.206, 0.084, 0.415)
states_resid <- model9.11$residuals[states]

r_hat <-   ehat_fuel / ( states_resid * sqrt(1 - hat_fuel))



r_hat
```

```
##       AK       NY       HI       WY       DC
## 1.159347 1.092395 1.122253 1.044847 1.307444
```

```
n <- 51
p_prime <- 5

t <- r_hat * ((n - p_prime - 1) / (n - p_prime - r_hat^2))^(1/2)
t
```

```
##       AK       NY       HI       WY       DC
## 1.163805 1.094749 1.125502 1.045914 1.317873
```

```
D_i <- (1/p_prime) * r_hat^2 * (hat_fuel /(1 - hat_fuel) )
names(D_i) <- states
D_i
```

```
##         AK         NY         HI         WY         DC
## 0.09249625 0.04613814 0.06535186 0.02002255 0.24253174
```

```
dl <- cooks.distance(model9.11)
dl
```

```
##           AL           AK           AZ           AR           CA
## 7.800459e-03 5.850260e-01 7.188822e-04 2.111673e-03 3.540623e-03
##           CO           CT           DE           DC           FL
## 1.019614e-03 4.054545e-03 6.568633e-02 1.407798e-01 8.807427e-02
##           GA           HI           ID           IL           IN
## 2.932058e-02 1.624367e-01 6.391753e-05 1.358427e-02 1.203642e-03
##           IA           KS           KY           LA           ME
```

```
## 7.891835e-04 2.797952e-03 3.743808e-04 7.223696e-03 2.392345e-03
##           MD           MA           MI           MN           MS
## 2.368748e-02 1.073745e-04 6.916698e-05 8.353715e-02 5.043638e-03
##           MO           MT           NE           NV           NH
## 2.912985e-03 3.408831e-04 1.727086e-03 3.028305e-03 5.546151e-02
##           NJ           NM           NY           NC           ND
## 6.298652e-03 4.988883e-04 2.081099e-01 2.620893e-04 8.025976e-04
##           OH           OK           OR           PA           RI
## 3.407617e-03 3.214596e-03 1.050371e-02 1.479644e-02 9.046794e-03
##           SC           SD           TN           TX           UT
## 6.004272e-03 4.029751e-03 5.884541e-04 1.415694e-02 5.442837e-03
##           VT           VA           WA           WV           WI
## 6.294248e-03 1.042558e-02 2.349349e-03 2.014122e-02 5.942758e-04
##           WY
## 1.596169e-01
```

```r
r <- rstudent(model9.11)
r
```

```
##          AL          AK          AZ          AR          CA          CO
## -0.59604249 -3.19302217 -0.31940486 -0.39101498 -0.42802758 -0.21405044
##          CT          DE          DC          FL          GA          HI
##  0.23197297  1.54976004 -0.99621024 -2.04875497  0.80740453 -1.81436531
##          ID          IL          IN          IA          KS          KY
##  0.06745827 -0.85273394  0.31734297  0.31984853 -0.60502152  0.16942711
##          LA          ME          MD          MA          MI          MN
##  0.56022239 -0.44712723  1.28876755 -0.06126147  0.09165130  1.93472315
##          MS          MO          MT          NE          NV          NH
##  0.48000082  0.57755619  0.11368351 -0.36157527  0.53528241  1.67591799
##          NJ          NM          NY          NC          ND          OH
##  0.37025366 -0.18513742 -2.43822460  0.16028647  0.34135279 -0.66279376
##          OK          OR          PA          RI          SC          SD
##  0.37495905 -1.05843156 -0.93030491  0.40123926  0.72082107  0.74115757
##          TN          TX          UT          VT          VA          WA
## -0.30215152  0.82013190 -0.61691403  0.38673812  1.15620603 -0.45635442
##          WV          WI          WY
## -0.93135606 -0.18029570  3.24608995
```

```r
r.newModel <- r[states]
dl.newModel <- dl[states]




n <- 51
p_prime <- 5

t <- r.newModel * ((n - p_prime - 1) / (n - p_prime - r.newModel^2))^(1/2)
t
```

```
##         AK         NY         HI         WY         DC
## -3.5796348 -2.5843496 -1.8624129  3.6566188 -0.9961265
```

```r
r.newModel
```

```
##         AK         NY         HI         WY         DC
## -3.1930222 -2.4382246 -1.8143653  3.2460899 -0.9962102
```

```r
dl.newModel
```

```
##        AK        NY        HI        WY        DC
## 0.5850260 0.2081099 0.1624367 0.1596169 0.1407798
```

```r
p.new <- 2*pt(-abs(t),df=n-1)

p.adjust(p.new, method = "bonferroni")
```

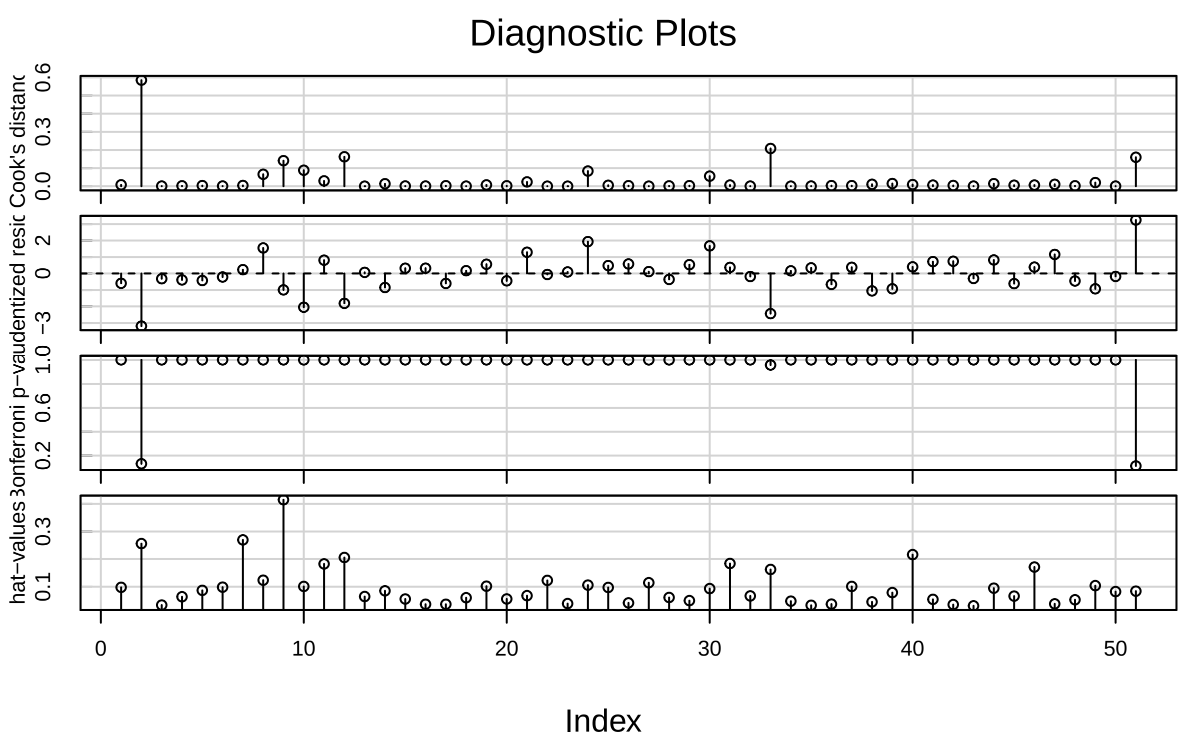
```
##          AK          NY          HI          WY          DC
## 0.003883200 0.063615878 0.342129610 0.003069248 1.000000000
```

```r
outlierTest(model9.11)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## WY  3.24609           0.002212      0.11281
```

Wyoming has the largest influence on the regression

```r
influenceIndexPlot(model9.11)
```



#Extra Practice Problems

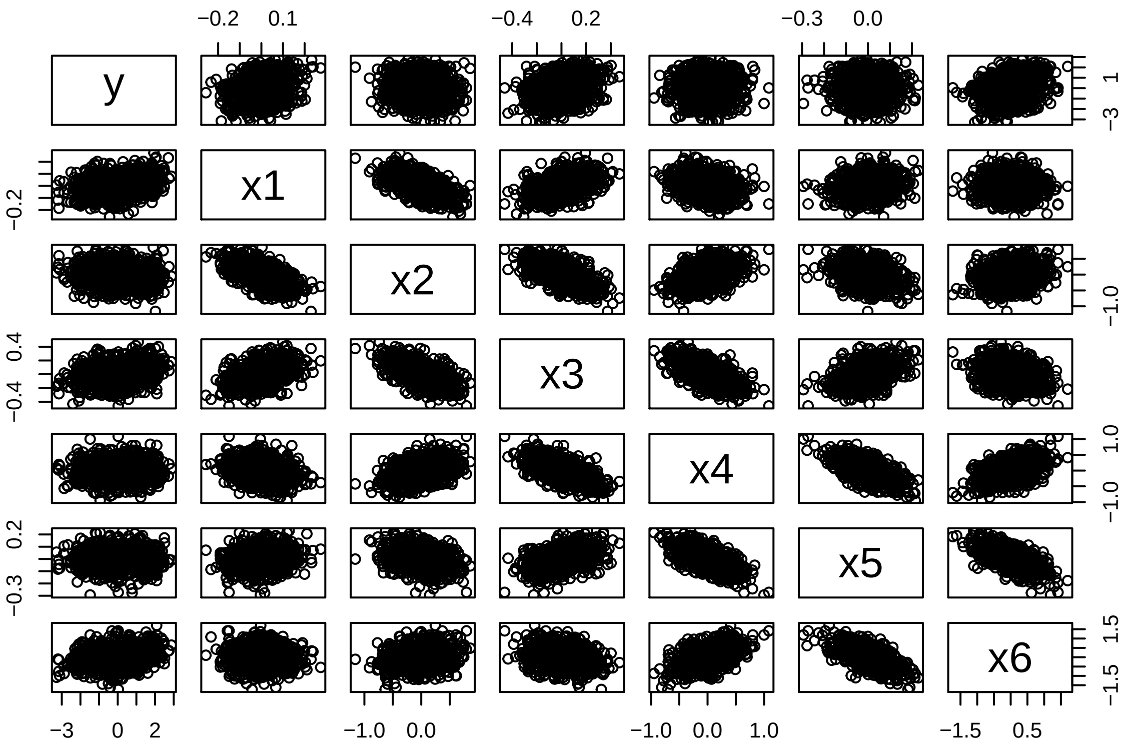
Problem 9.1

```r
head(Rpdata)
```

```
##            y         x1          x2          x3         x4          x5        x6
```

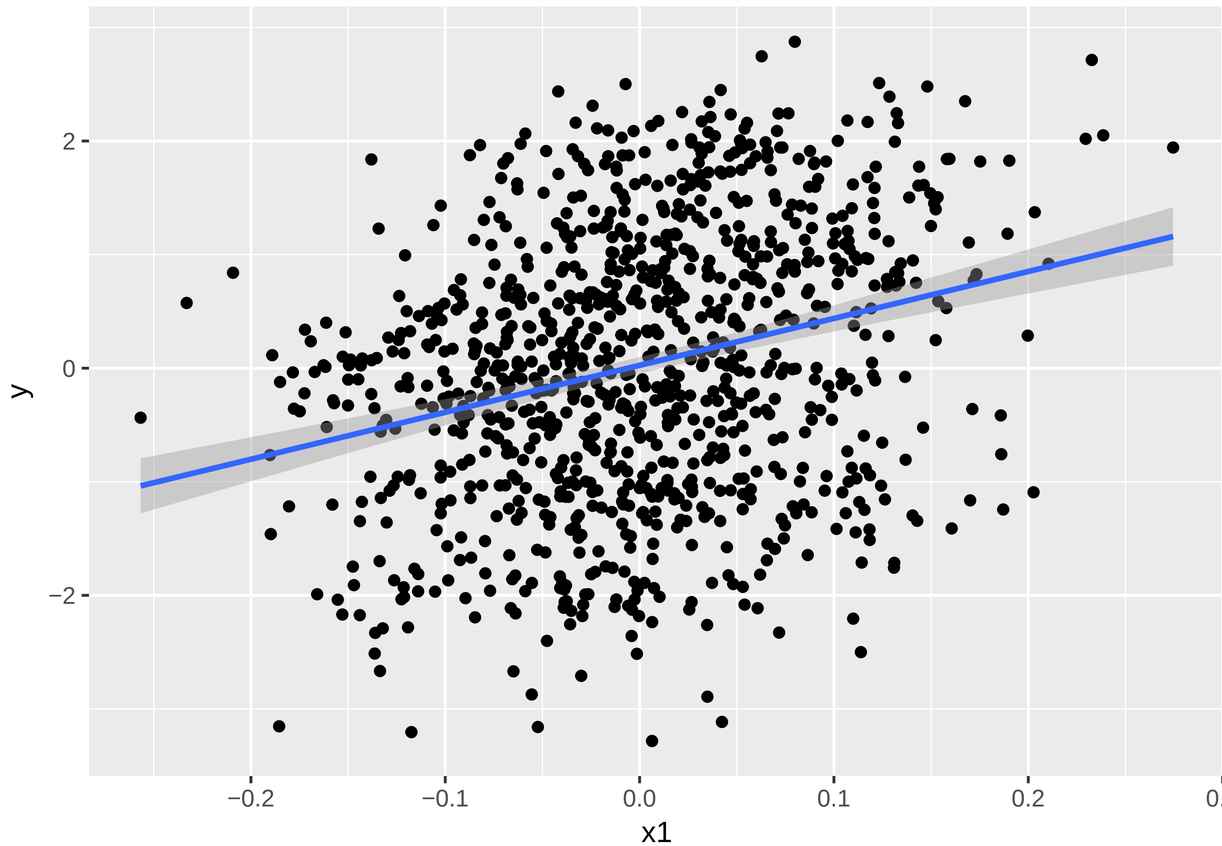
```
## 1  0.31680 -0.034611 -0.0635220  0.0437570 -0.030642  0.1303200 -0.56870
## 2 -0.58091 -0.028196  0.2368700 -0.1884600  0.190510 -0.0818980  0.29123
## 3 -0.38789  0.059727  0.0884460 -0.0946590  0.022736  0.0073608  0.40923
## 4 -0.31245  0.052236  0.0032441 -0.1222200  0.127780  0.0471430 -0.64441
## 5  0.96025 -0.057986  0.0659260 -0.0070171  0.107700  0.1265700 -0.71757
## 6  1.18400  0.189340 -0.4765500  0.1547300 -0.321130  0.0726280 -0.13887
```

```r
pairs(Rpdata)
```

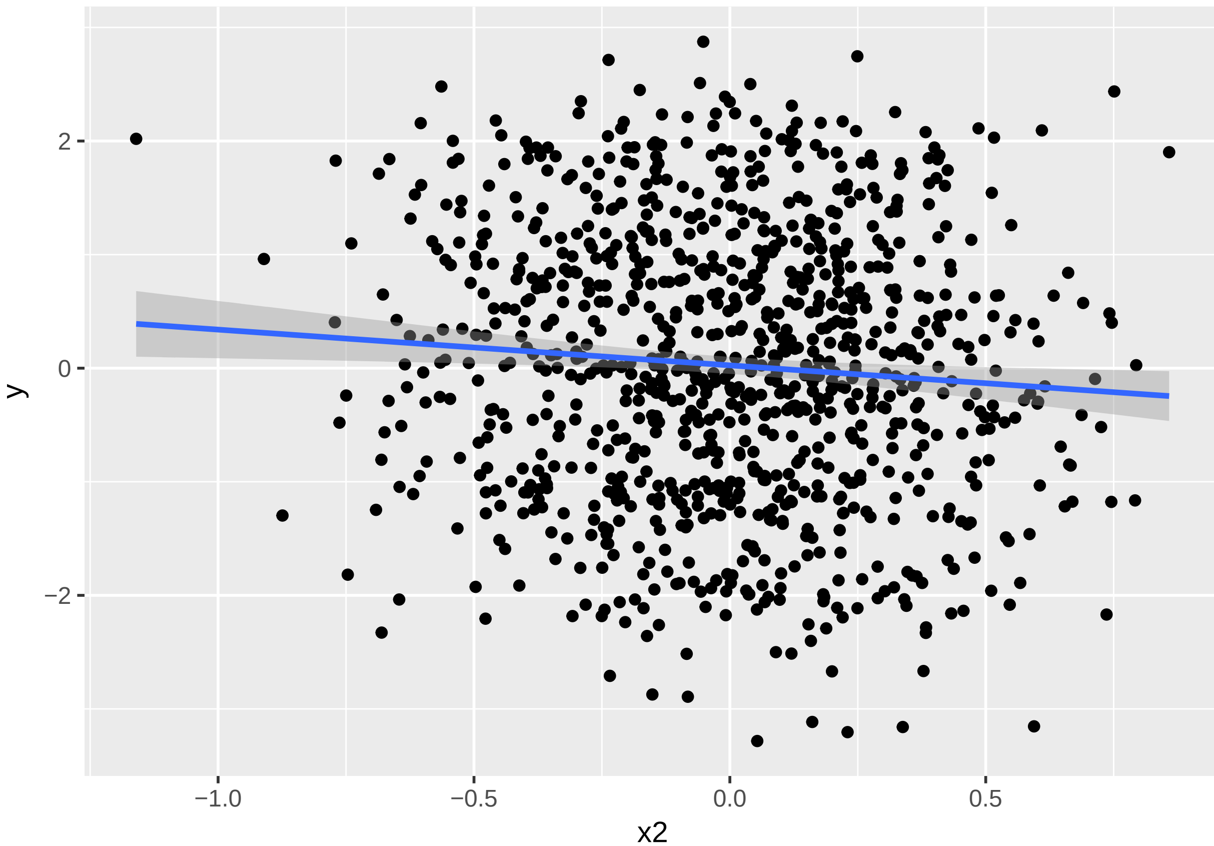


The variables next to each other seem strongly correlated, except with x1 and y. x1 neighbors is x2.

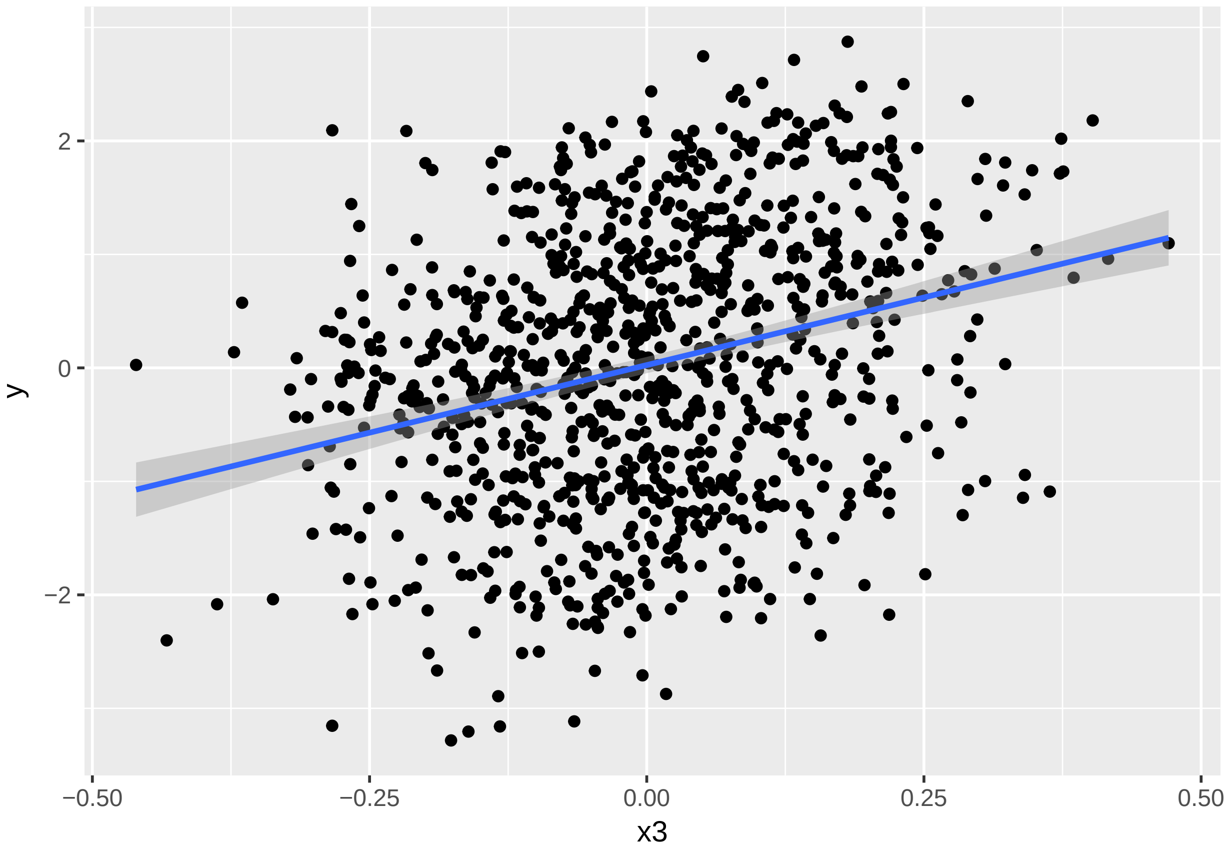
```r
ggplot(Rpdata, aes(x1, y)) +
  geom_point() +
  geom_smooth(method=lm)
```



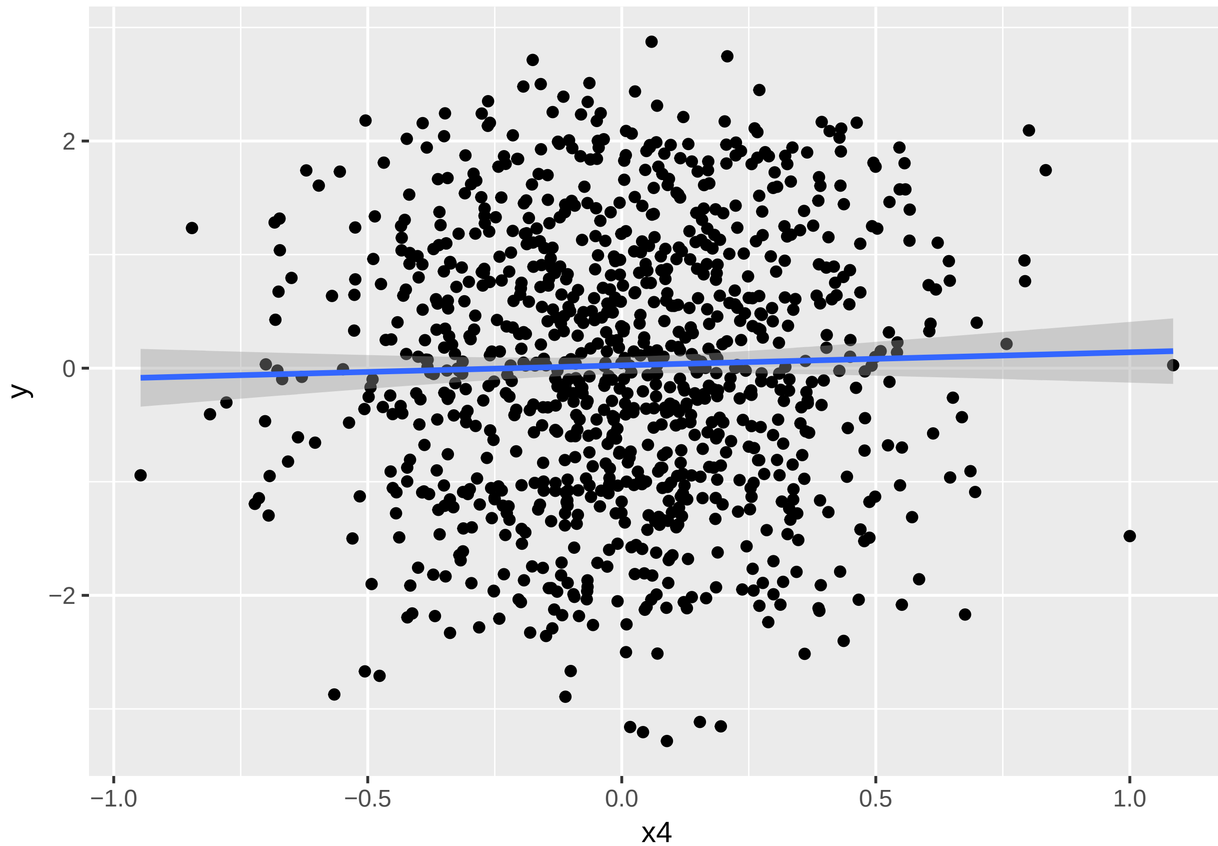

```r
ggplot(Rpdata, aes(x2, y)) +
  geom_point() +
  geom_smooth(method=lm)
```



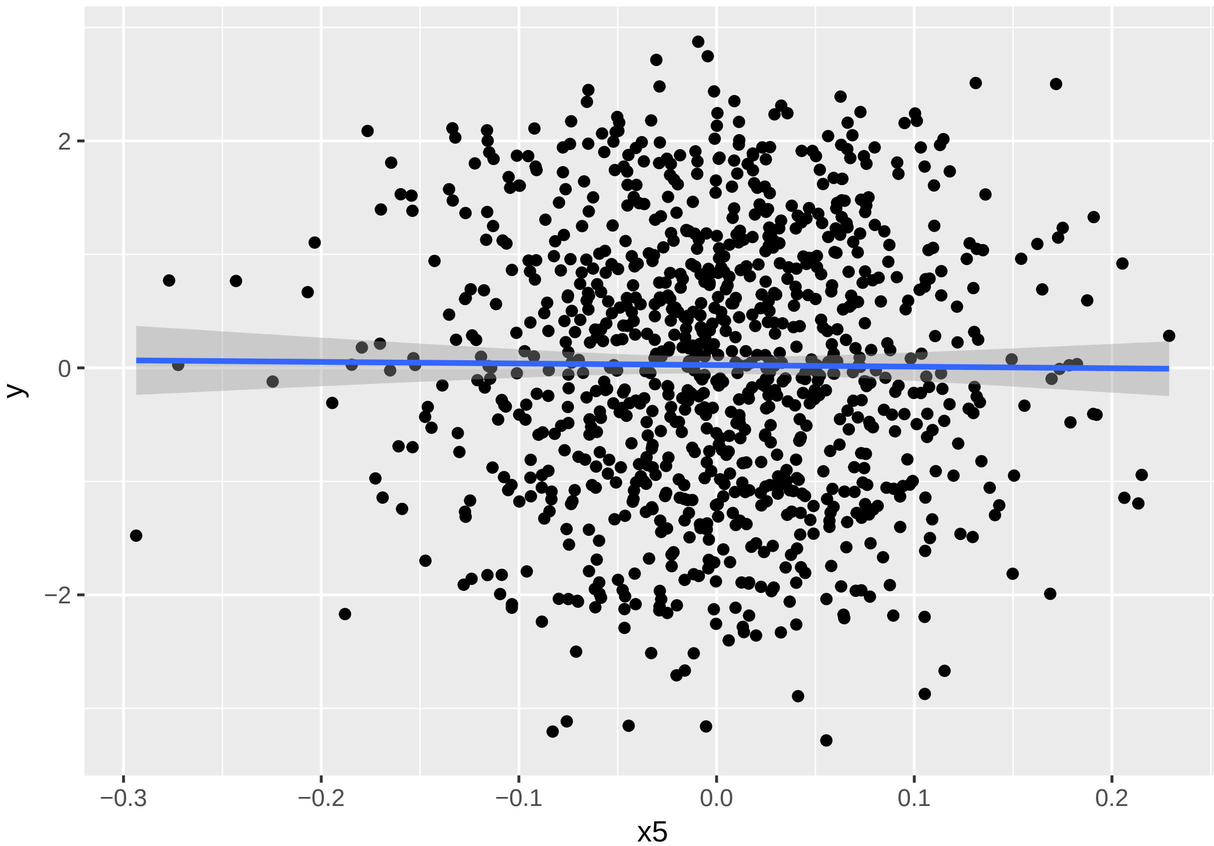

```r
ggplot(Rpdata, aes(x3, y)) +
  geom_point() +
  geom_smooth(method=lm)
```



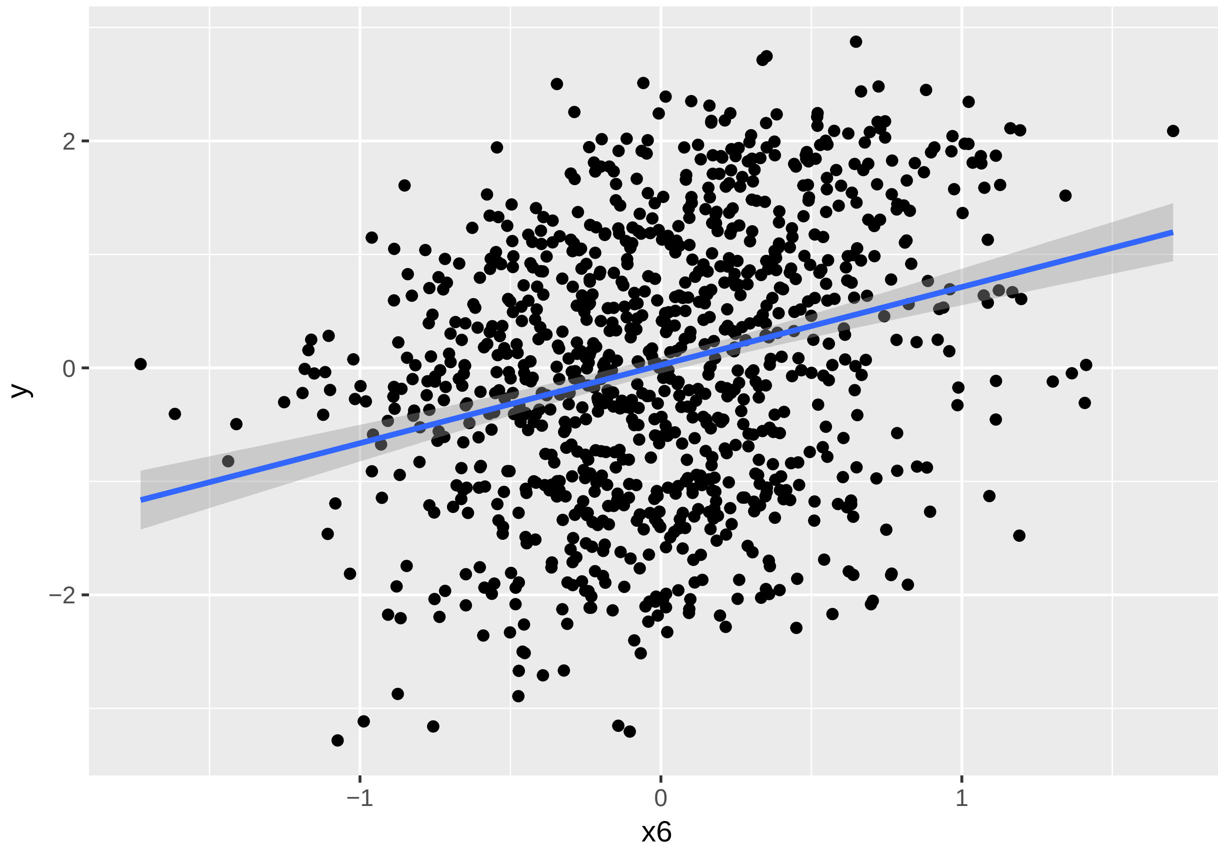

```
ggplot(Rpdata, aes(x4, y)) +
  geom_point() +
  geom_smooth(method=lm)
```




```r
ggplot(Rpdata, aes(x5, y)) +
  geom_point() +
  geom_smooth(method=lm)
```




```
ggplot(Rpdata, aes(x6, y)) +
  geom_point() +
  geom_smooth(method=lm)
```




```r
pr9.1 <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
summary(pr9.1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.1977 -0.7631  0.1729  0.8851  1.6359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02481    0.03188   0.778    0.437
## x1           4.14061    0.50954   8.126 1.32e-15 ***
## x2           1.01233    0.15522   6.522 1.11e-10 ***
## x3           3.99614    0.32663  12.234  < 2e-16 ***
## x4           0.96045    0.16657   5.766 1.09e-08 ***
## x5           3.75122    0.64726   5.796 9.17e-09 ***
## x6           0.95390    0.08561  11.142  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.003 on 983 degrees of freedom
## Multiple R-squared:  0.3112, Adjusted R-squared:  0.307
## F-statistic: 74.03 on 6 and 983 DF,  p-value: < 2.2e-16
```

All the p-values are signficant. This might mean that the model is wrong.

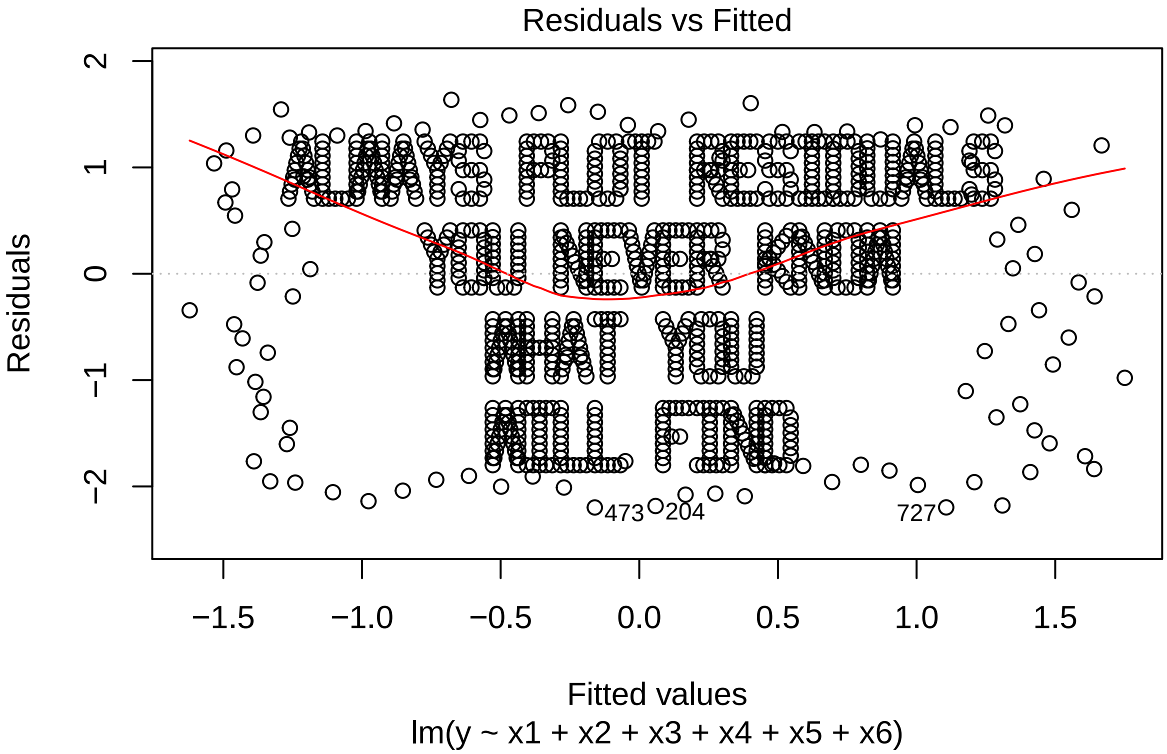
```r
pr9.1.1 <- lm(x1 ~ x2,  data = Rpdata)
summary(pr9.1)
```

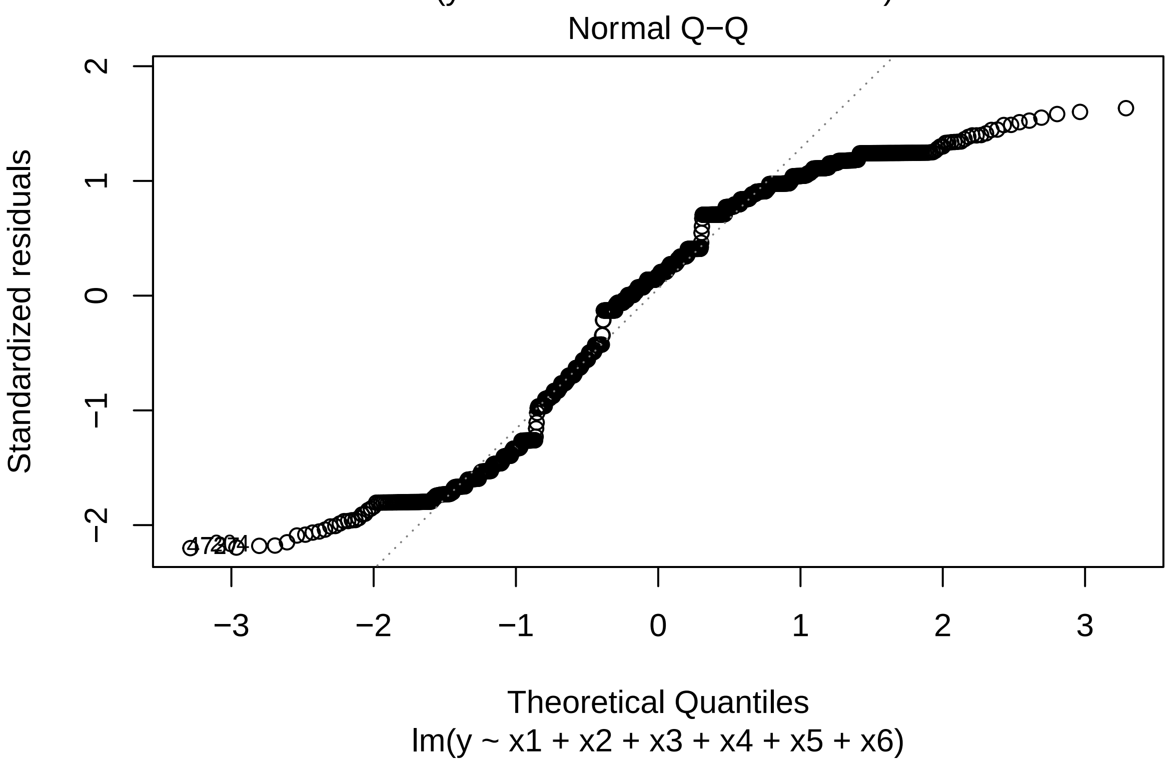
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1977 -0.7631  0.1729  0.8851  1.6359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02481    0.03188   0.778    0.437
## x1           4.14061    0.50954   8.126 1.32e-15 ***
## x2           1.01233    0.15522   6.522 1.11e-10 ***
## x3           3.99614    0.32663  12.234  < 2e-16 ***
## x4           0.96045    0.16657   5.766 1.09e-08 ***
## x5           3.75122    0.64726   5.796 9.17e-09 ***
## x6           0.95390    0.08561  11.142  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.003 on 983 degrees of freedom
## Multiple R-squared:  0.3112, Adjusted R-squared:  0.307
## F-statistic: 74.03 on 6 and 983 DF,  p-value: < 2.2e-16
```

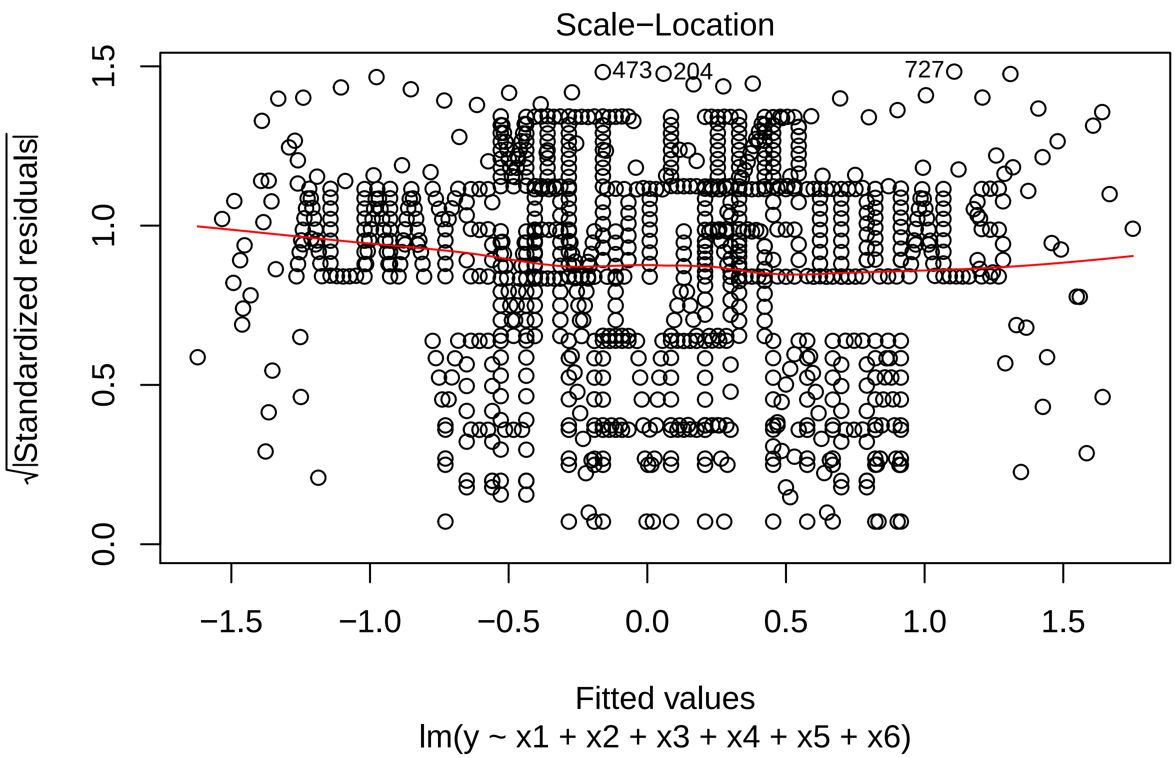
```r
pr9.1.2 <- lm(x2 ~ x1,  data = Rpdata)
summary(pr9.1.2)
```

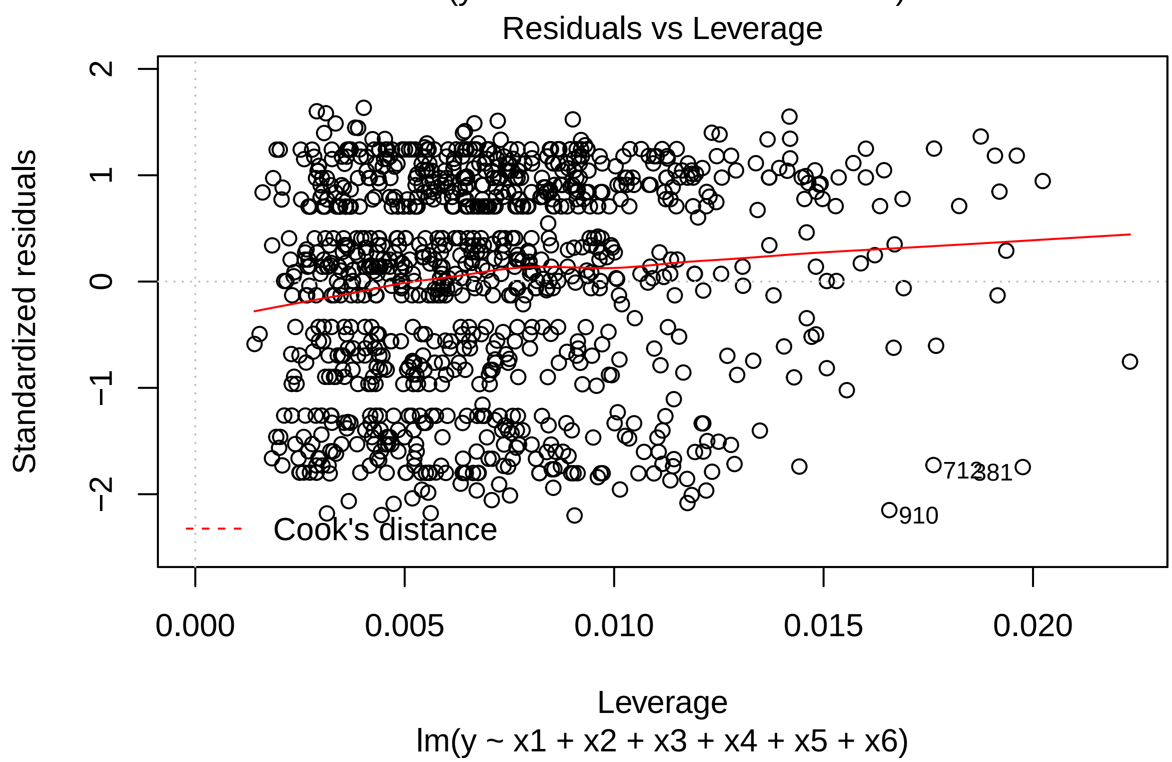
```
##
## Call:
## lm(formula = x2 ~ x1, data = Rpdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73758 -0.16832 -0.00125  0.17387  0.86465
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.036e-07  7.813e-03    0.00        1
## x1          -2.384e+00  9.724e-02  -24.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2458 on 988 degrees of freedom
## Multiple R-squared:  0.3782, Adjusted R-squared:  0.3776
## F-statistic: 600.9 on 1 and 988 DF,  p-value: < 2.2e-16
```

```r
plot(pr9.1)
```


Residuals vs Fitted
ALWAYS PLOT RESIDUALS
YOU NEVER KNOW
WHAT YOU
WILL FIND
473
204
727
Residuals
2
1
0
−1
−2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6)

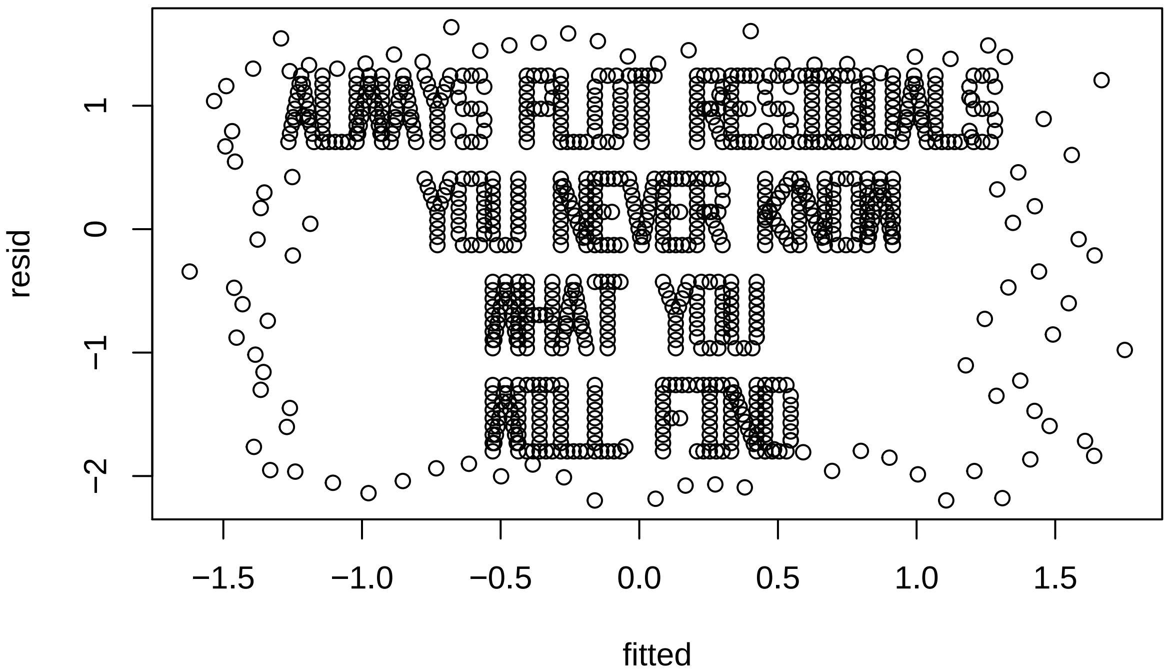

Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6)






```
resid <- pr9.1$residuals
fitted <- pr9.1$fitted.values

plot(fitted, resid)
```


ALWAYS PLOT RESIDUALS
YOU NEVER KNOW
WHAT YOU
WILL FIND
resid
1
0
−1
−2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
fitted