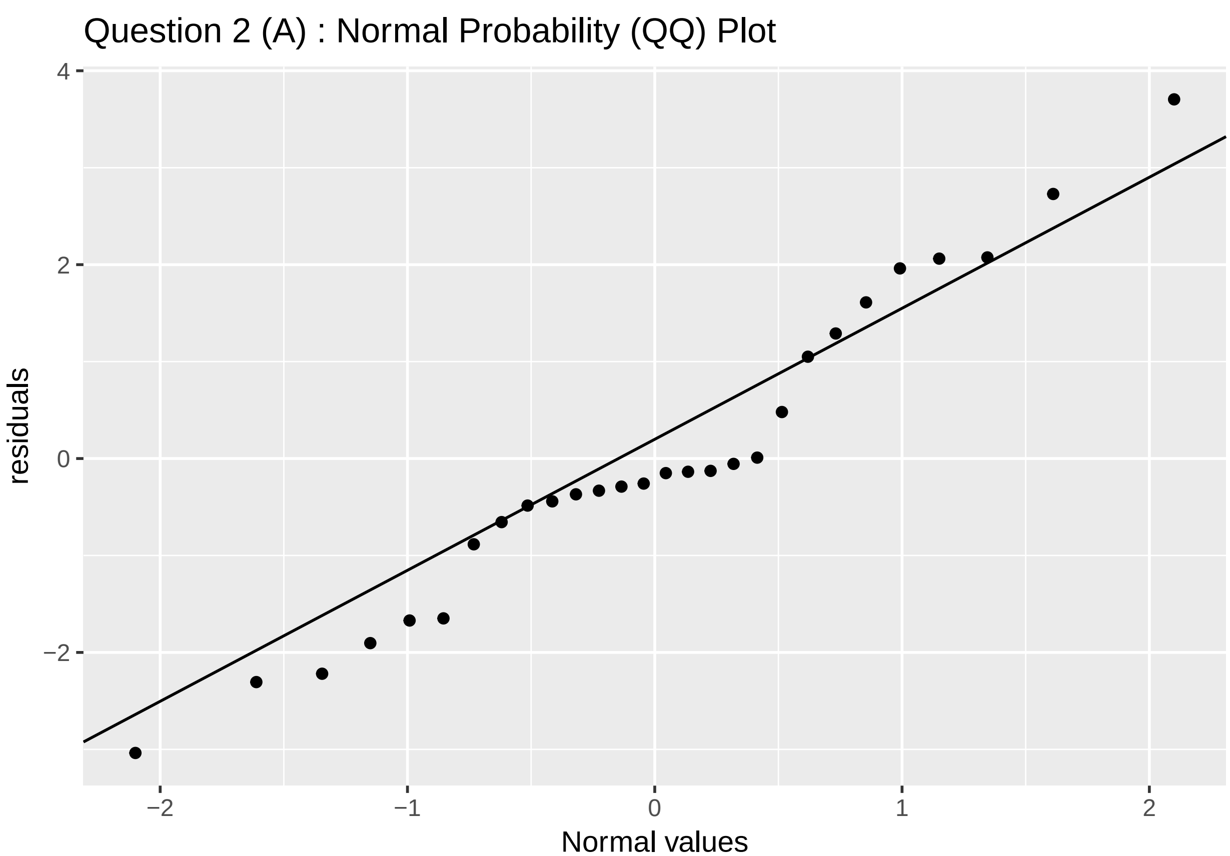
Stat 230: Linear Models
Homework 4
Professor Ding
Lev Golod

QUESTION 2

```
fit1 <- lm(y ~ x2 + x7 + x8, data = table.b1)


### ### ### ### ### ###
## 2 A - Normal Probability (QQ) plot of RAW (NOT STANDARDIZED) residuals
### ### ### ### ### ###
y <- quantile(fit1$residuals, c(0.25, 0.75))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int <- y[1L] - slope * x[1L]
plt1a_qqplot <- ggplot(data.frame(fit1$residuals), aes(sample=fit1.residuals)) +
  stat_qq() +
  geom_abline(slope = slope, intercept = int) +
  ylab('residuals') +
  xlab('Normal values') +
  ggtitle('Question 2 (A) : Normal Probability (QQ) Plot')
plt1a_qqplot
```

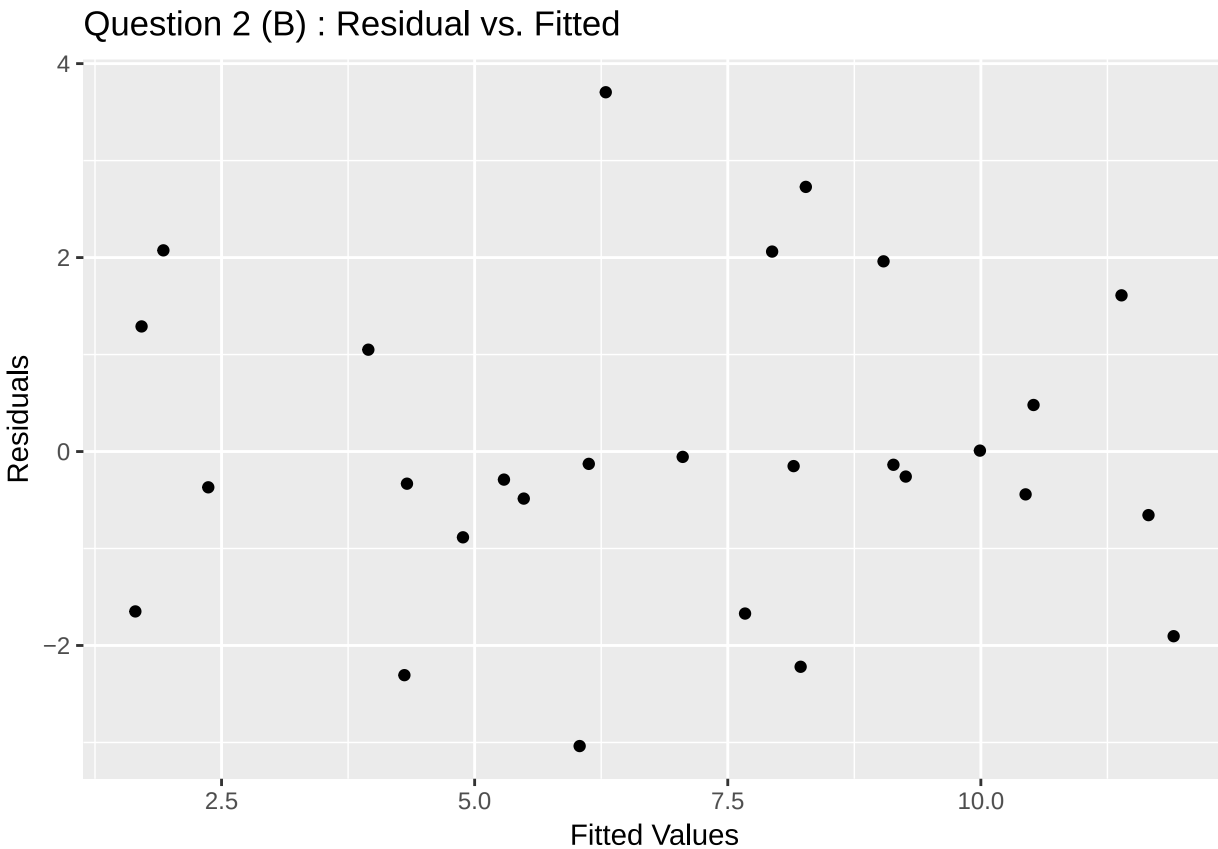



```
cat('The residuals lie nearly in a straight line, which is good.
However they are not quite exactly straight, which means it is possible
```

```
that the errors might NOT be IID normal.')
```

```
## The residuals lie nearly in a straight line, which is good.
## However they are not quite exactly straight, which means it is possible
## that the errors might NOT be IID normal.
```

```
### ### ### ### ### ###
## 2 B - Plot residuals vs predicted values
### ### ### ### ### ###
plt1b_resid <- ggplot(data.frame(x=fit1$fitted.values, y=fit1$residuals),
                      aes(x,y)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Question 2 (B) : Residual vs. Fitted')
plt1b_resid
```

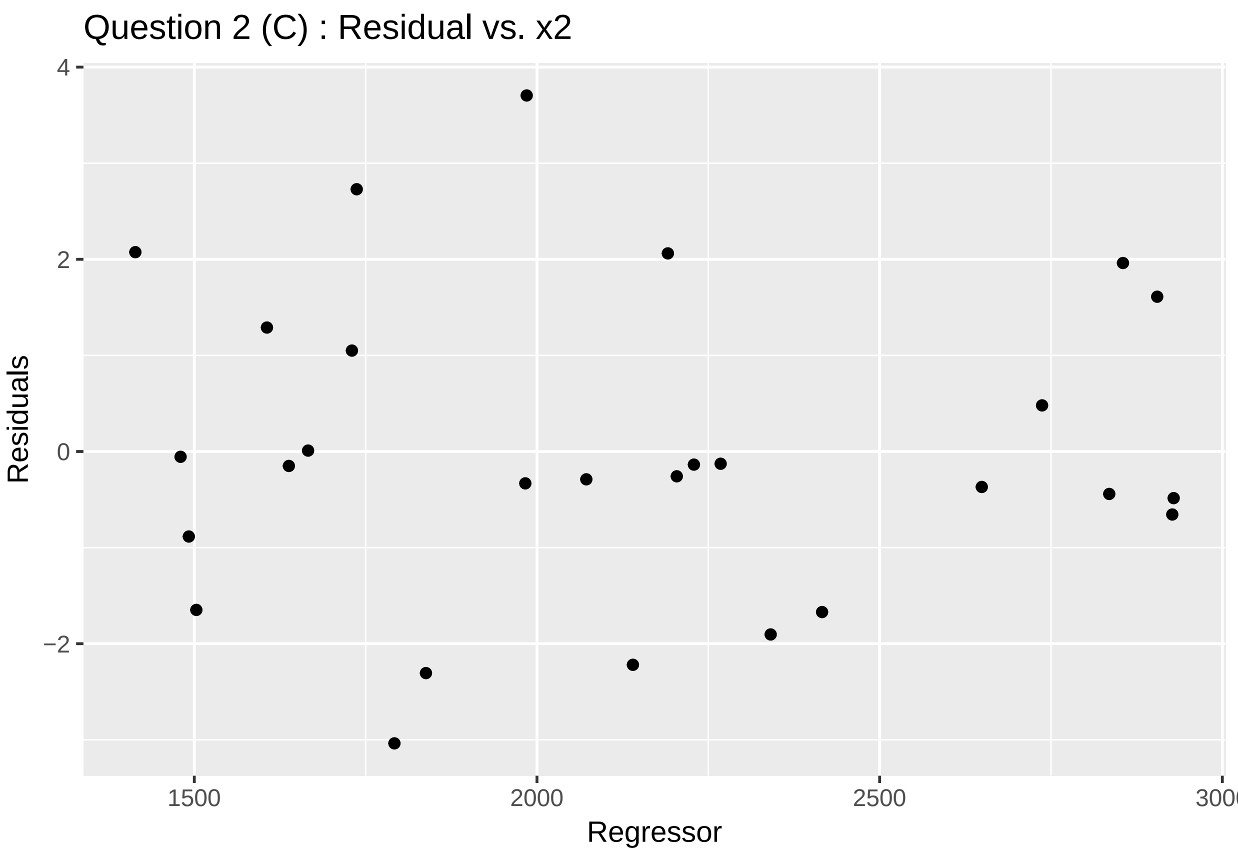


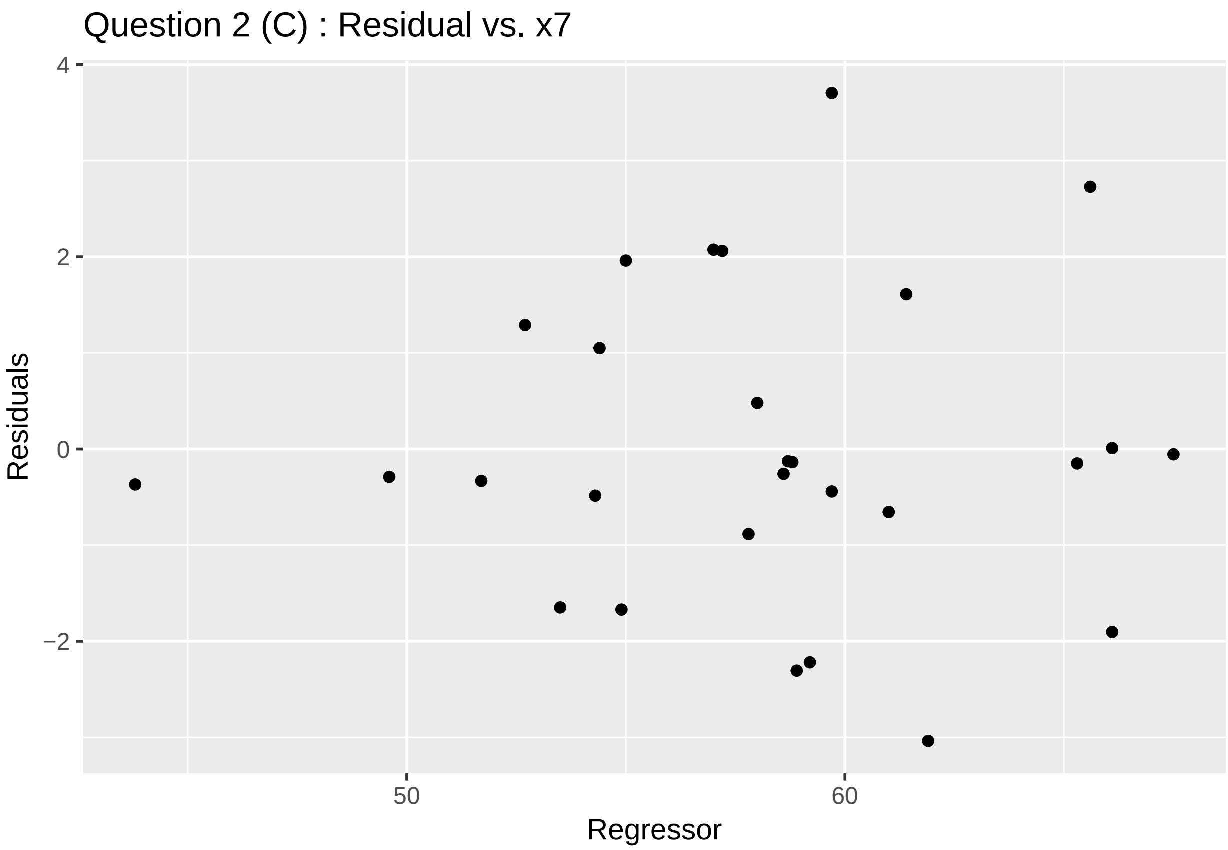

```
cat('There is no strong, clearly visible pattern to the residual plot.
This gives evidence that the errors are IID.')
```

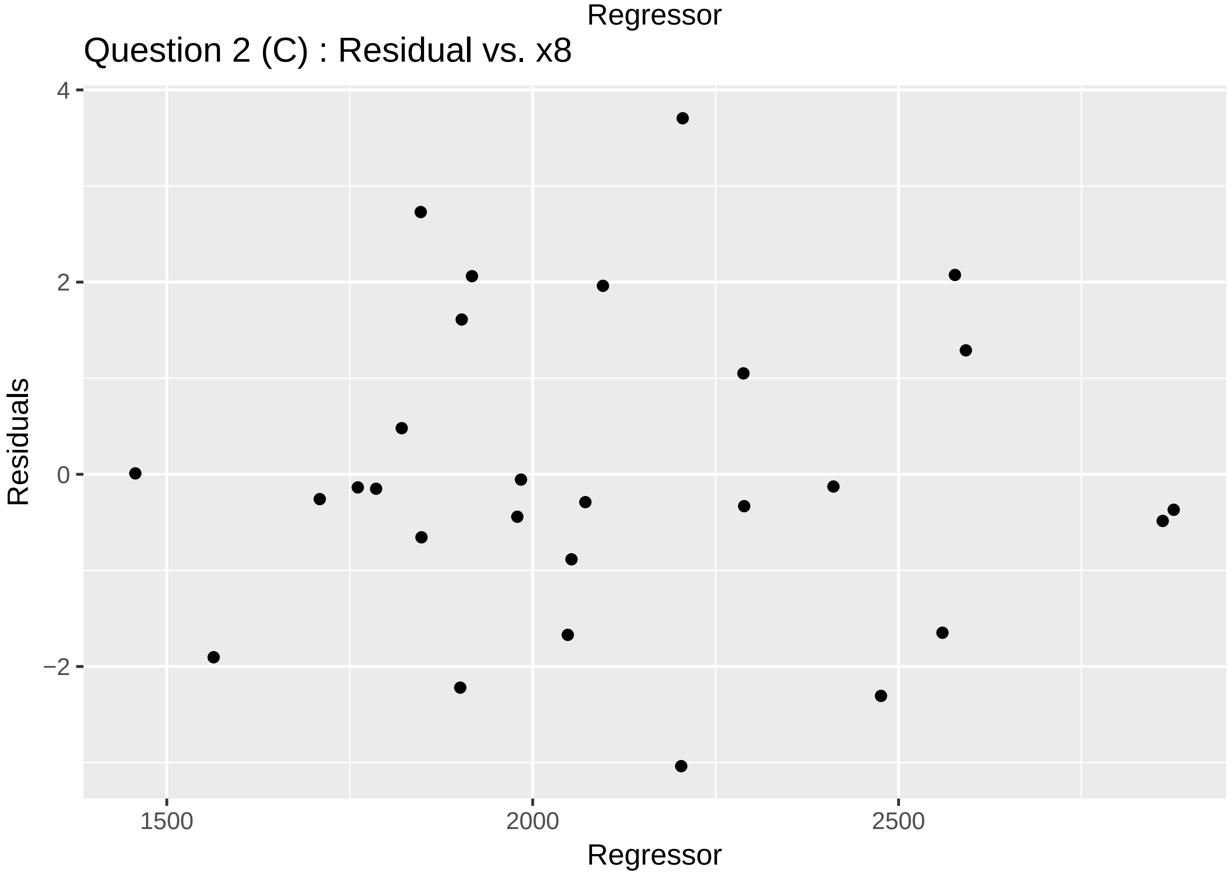
```
## There is no strong, clearly visible pattern to the residual plot.
## This gives evidence that the errors are IID.
```

```
### ### ### ### ### ###
## 2 C - Plot residuals vs regressors
### ### ### ### ### ###
# plt1c_x2 <- ggplot(data.frame(x=table.b1$x2, y = fit1$residuals), aes(x,y)) +
#   geom_point() +
```

```r
#   xlab('Regressor') +
#   ylab('Residuals') +
#   ggtitle('Question 1 (C) : Residual vs. x2')
regs <- c('x2', 'x7', 'x8')
for (var in regs){
  text1 <- paste0('plt1c_', var, ' <- ggplot(data.frame(x=table.b1$', var, ",y = fit1$residuals), aes(x
  eval(parse(text=text1))
  text2 <- paste0('print(plt1c_', var,')')
  eval(parse(text=text2))
}
```




Question 2 (C) : Residual vs. x7
Residuals
Regressor
4
2
0
−2
50
60


Question 2 (C) : Residual vs. x8
Residuals
Regressor
4
2
0
−2
1500
2000
2500

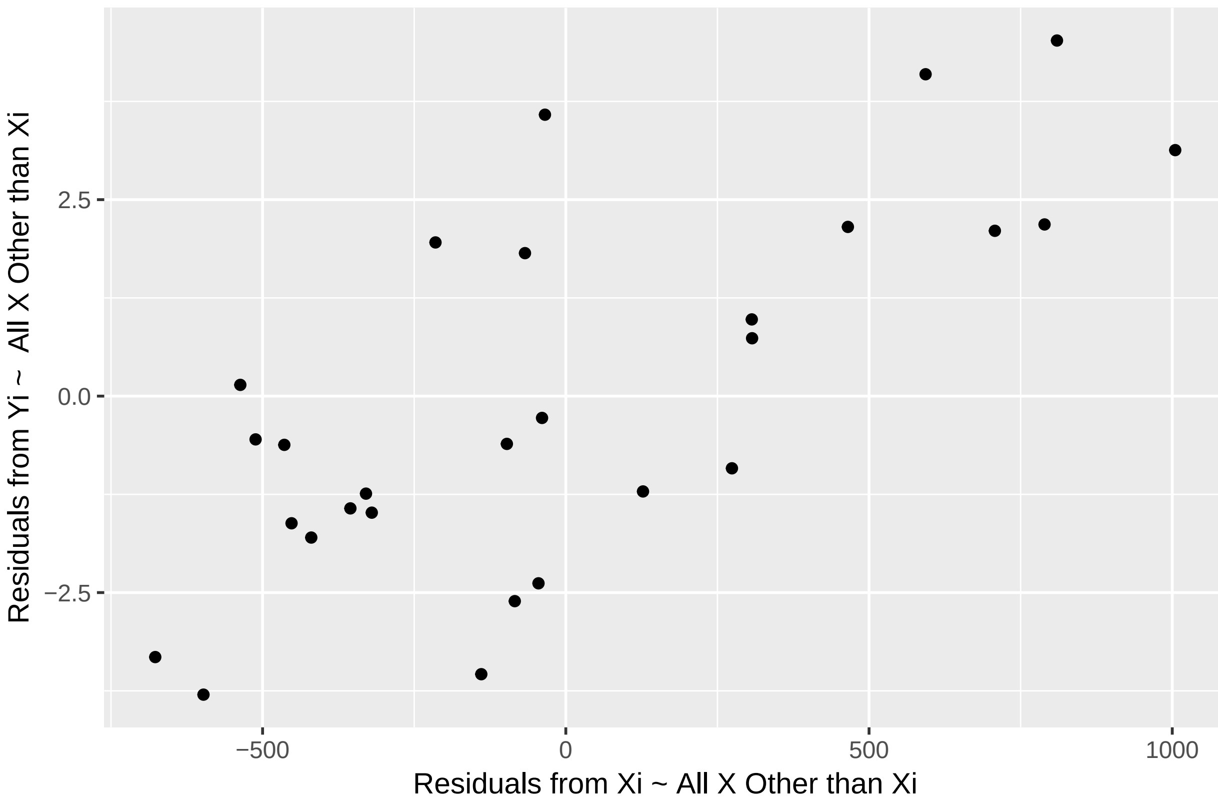
```r
cat('These plots do NOT provide evidence of constant variance, since they do
not look like random scatters that have no pattern.
For x2 we see something like a weak bow-tie shape.
For x7 we see a very strong funnel shape.
For x8 we see a weak football-like shape: narrow at the ends, fat in the middle.
These plots do NOT suggest that the relationship between the regressors and the
response is non-linear. That is, the plots DO imply the regressors are all
correctly specified.')
```

```
## These plots do NOT provide evidence of constant variance, since they do
## not look like random scatters that have no pattern.
## For x2 we see something like a weak bow-tie shape.
## For x7 we see a very strong funnel shape.
## For x8 we see a weak football-like shape: narrow at the ends, fat in the middle.
## These plots do NOT suggest that the relationship between the regressors and the
## response is non-linear. That is, the plots DO imply the regressors are all
## correctly specified.
```

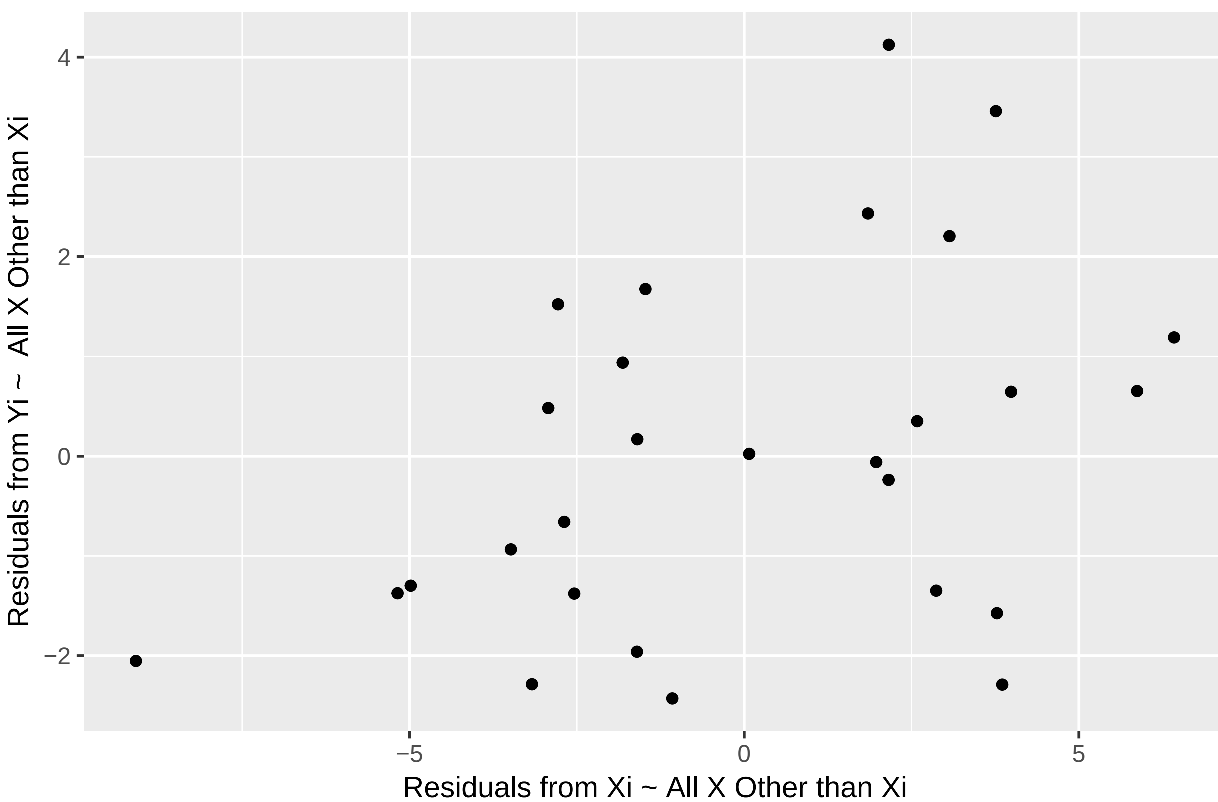
```r
### ### ### ### ### ###
## 2 D - Partial Regression Plots
### ### ### ### ### ###
#http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/partregr.htm

# myregs <- list(table.b1$x2,table.b1$x7,table.b1$x8)
for (var in regs){
  other_regs <- regs[regs != var]
  text1 <- paste
  # y_resid <- lm(y ~ x7 + x8, data = table.b1)$residuals
  text1 <- paste0('y_resid <- lm(y ~ ',
                  paste0(other_regs, collapse='+'),
                  ', data = table.b1)$residuals')
  eval(parse(text=text1))
  text2 <- paste0('x_resid <- lm(', var, ' ~ ',
                  paste0(other_regs, collapse='+'),
                  ', data = table.b1)$residuals')
  eval(parse(text=text2))
  text3 <- paste0('plt1d_', var)
  assign(text3,
         ggplot(data.frame(x=x_resid,y=y_resid), aes(x,y)) +
           geom_point() +
           xlab('Residuals from Xi ~ All X Other than Xi') +
           ylab('Residuals from Yi ~  All X Other than Xi') +
           ggtitle(paste0('Question 2 (D) : Partial Residual Plot for ', var)))
  print(eval(parse(text=text3)))
}
```

Question 2 (D) : Partial Residual Plot for x2



Question 2 (D) : Partial Residual Plot for x7




```r
cat('A Partial Regression plot for Xi allows us to the effect of adding Xi
to the model, removing the influence of all of the other regressors. The y axis
represents Y*, the information in Y that cannot be explained be the other
regressors. The x axis represents Xi*, the information contained in Xi that is not
contained in the other regressors.
We see moderately string linear relationships for x2 (positive) and x8
(negative). We see a distinct cone shape for x7, which means the variability
in Y* increases as Xi* increases')
```

```
## A Partial Regression plot for Xi allows us to the effect of adding Xi
## to the model, removing the influence of all of the other regressors. The y axis
## represents Y*, the information in Y that cannot be explained be the other
## regressors. The x axis represents Xi*, the information contained in Xi that is not
## contained in the other regressors.
## We see moderately string linear relationships for x2 (positive) and x8
## (negative). We see a distinct cone shape for x7, which means the variability
## in Y* increases as Xi* increases
```

```r
### ### ### ### ### ###
## 2 E - Studentized and R-Studentized Residuals
### ### ### ### ### ###
student_resid <- rstandard(fit1)
rstudent_resid <- rstudent(fit1)
summary(student_resid)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.87000 -0.45200 -0.12700 -0.00377  0.68800  2.23000
```

```r
summary(rstudent_resid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.9800 -0.4440 -0.1250  0.0033  0.6810  2.4500
```

```r
cat('The Studentized Residuals are normalized using a function of Hii - the
hat value for the ith point. This means their variance doesn not depend on the
value of xi. As a result, they are good at detecting outliers.
The R-Studentized Residuals have the further advantage that the estimate of
sigma-hat-squared is made without the ith point. This means they are even
better at detecting outliers.
In our specific case, both the Studentized and R-Studentized Residuals take
values from about -2 to 2.5, which suggests we do not have extreme outliers.')
```

```
## The Studentized Residuals are normalized using a function of Hii - the
## hat value for the ith point. This means their variance doesn not depend on the
## value of xi. As a result, they are good at detecting outliers.
## The R-Studentized Residuals have the further advantage that the estimate of
## sigma-hat-squared is made without the ith point. This means they are even
## better at detecting outliers.
## In our specific case, both the Studentized and R-Studentized Residuals take
## values from about -2 to 2.5, which suggests we do not have extreme outliers.
```