STAT850 HW1

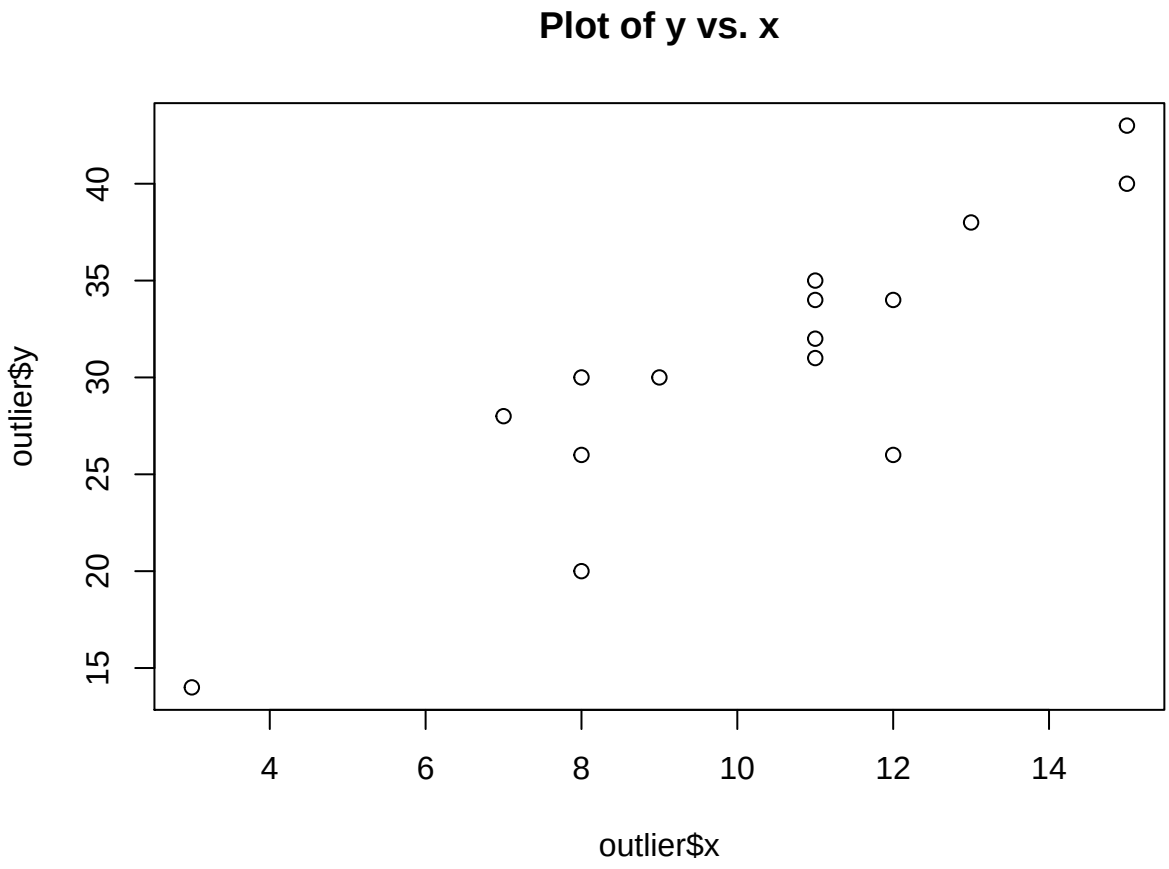
Stewart Kerr

January 25, 2019

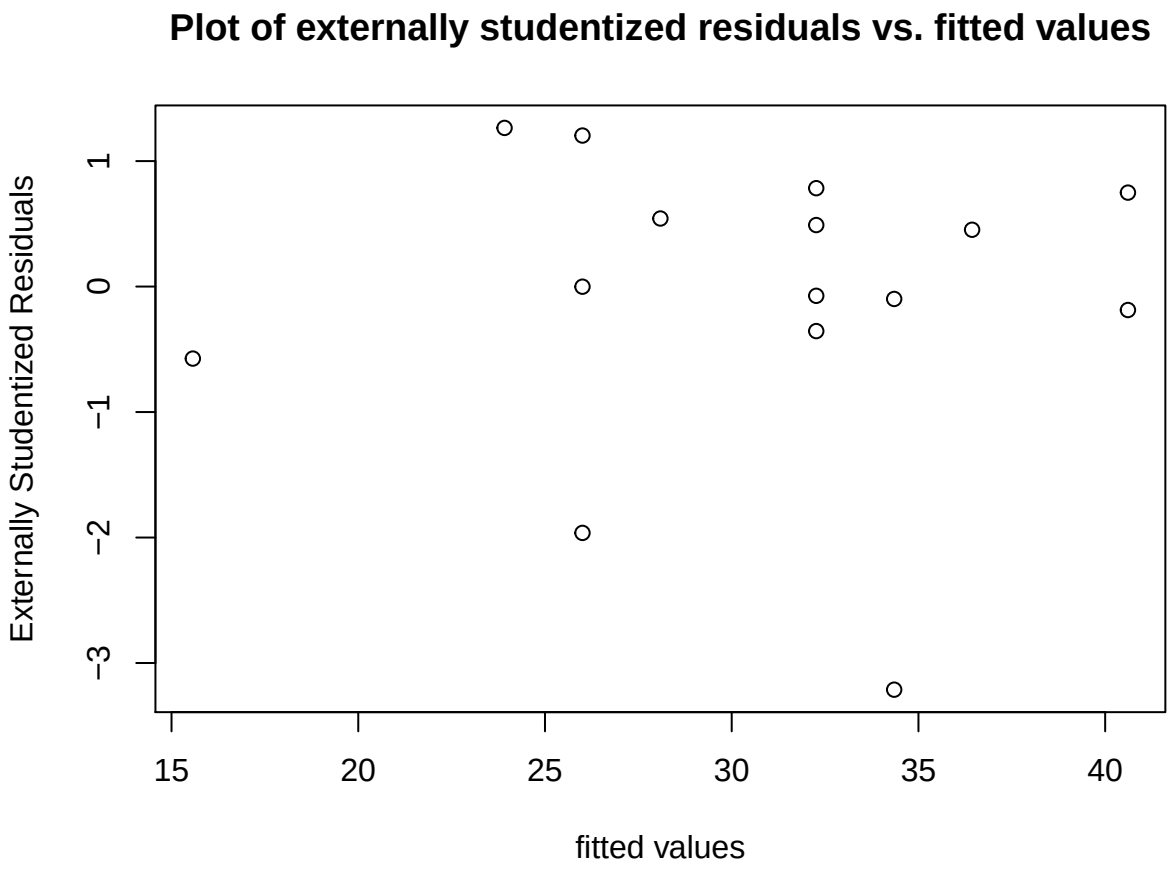
Problem 1

a.

```
#Problem 1a
#Read in the data and perform the regression
outlier <- read.csv("~/2019spring/STAT850/hw1/outlier.txt", sep="")
outlier_lm <- lm(outlier$y ~ outlier$x, qr = T)
plot(outlier$y ~ outlier$x, main = "Plot of y vs. x")
```




```
#Perform diagnostics by looking at externally studentized residuals
outlier_lm_diag <- ls.diag(outlier_lm)
plot(outlier_lm_diag$stud.res ~ outlier_lm$fitted.values, ylab = "Externally Studentized Residuals", xla
```



The externally studentized residuals are t-distributed with n-p-2 degrees of freedom. For this problem, n = 15 and p = 1, so there are 12 degrees of freedom. To formally determine whether or not any of our observations are outlying, I will compare the residuals to a t-distribution with 12 degrees of freedom. The 95% critical values for a t-distribution with 12 degrees of freedom are ±2.179. Therefore, if any of our externally studentized residuals fall outside of this range, they may be outlying in y. For observation 9 (x = 12, y = 26), the externally studentized residual, $t_i = -3.2125$, thus this point is likely outlying in y.

b.

```
#Problem 1b
#i. Remove observation 9 from the data set
outlier_mod <- filter(outlier, indic == 0)

#ii. Regress the new data set
outlier_mod_lm <- lm(y~x ,data = outlier_mod)
summary(outlier_mod_lm)
```

```
##
## Call:
## lm(formula = y ~ x, data = outlier_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3589 -1.3112  0.1432  1.3932  3.8402
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7656     2.4623   3.560  0.00392 **
## x             2.1992     0.2319   9.482 6.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.722 on 12 degrees of freedom
## Multiple R-squared:  0.8822, Adjusted R-squared:  0.8724
## F-statistic: 89.91 on 1 and 12 DF,  p-value: 6.343e-07
```

```
#iii. Predict value when x = 12 and calculate standard error
y_pred = 8.7656 + 2.1992*12
mse = sum((outlier_mod_lm$residuals^2)/12)
ssqx = sum((outlier_mod$x-mean(outlier_mod$x))^2)
xbar = mean(outlier_mod$x)
se_pred = sqrt(mse*(1+(1/14)+(((12-xbar)^2)/ssqx)))
paste("The standard error for prediction is:  ",round(se_pred,3))
```

```
## [1] "The standard error for prediction is:   2.85"
```

```
#iv. Test whether the questionable observation is within sampling error
t = (26 - y_pred)/(se_pred)
paste("The observed T statistic is ", round(t,3), "with 12 degrees of freedom")
```

```
## [1] "The observed T statistic is  -3.213 with 12 degrees of freedom"
```

```
#v. Compute the p-value and apply Bonferroni
paste("The p-value for this test statistic is: ",round(2*pt(t,12),4))
```

```
## [1] "The p-value for this test statistic is:  0.0075"
```

As seen in the output above, this procedure yields the same t-value as before - the externally studentized residual for observation 9. Again, this statistic is T-distributed with 12 degrees of freedom, so it falls outside of sampling error for this regression (± 2.179). However, after applying a Bonferroni correction, the p-value does not fall below the critical value of $\alpha = 0.05/14 = 0.0036$.

c.

```
#Problem 1c
#i. Perform a multiple linear regression of y on x and indic
outlier_mlm <- lm(y ~ x + indic, data = outlier)
summary(outlier_mlm)
```

```
##
## Call:
## lm(formula = y ~ x + indic, data = outlier)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.359 -1.259  0.000  1.344  3.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7656     2.4623   3.560  0.00392 **
## x             2.1992     0.2319   9.482 6.34e-07 ***
## indic        -9.1556     2.8500  -3.212  0.00746 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.722 on 12 degrees of freedom
## Multiple R-squared:  0.8859, Adjusted R-squared:  0.8669
## F-statistic: 46.57 on 2 and 12 DF,  p-value: 2.209e-06
```

From the output above, we see that the T-statistic and p-value for the indic variable are the same as t-value/p-value produced in parts a and b for previous outlier tests. That is, the t-value for indic is the same as the externally studentized residual for observation 9 and the t-value for the prediction for observation 9.

ii. For the outlier test using the indic variable, we are testing the null hypothesis that the regression coefficient, B_2 for indic is equal to 0 versus the alternative hypothesis that it is not 0. The indic variable is set to 1 only for observation 9 (which we suspect might be an outlier) - thus, if we fail to reject the null hypothesis then we are saying that observation 9 follows the same model as the other observations and is not an outlier.

Problem 2

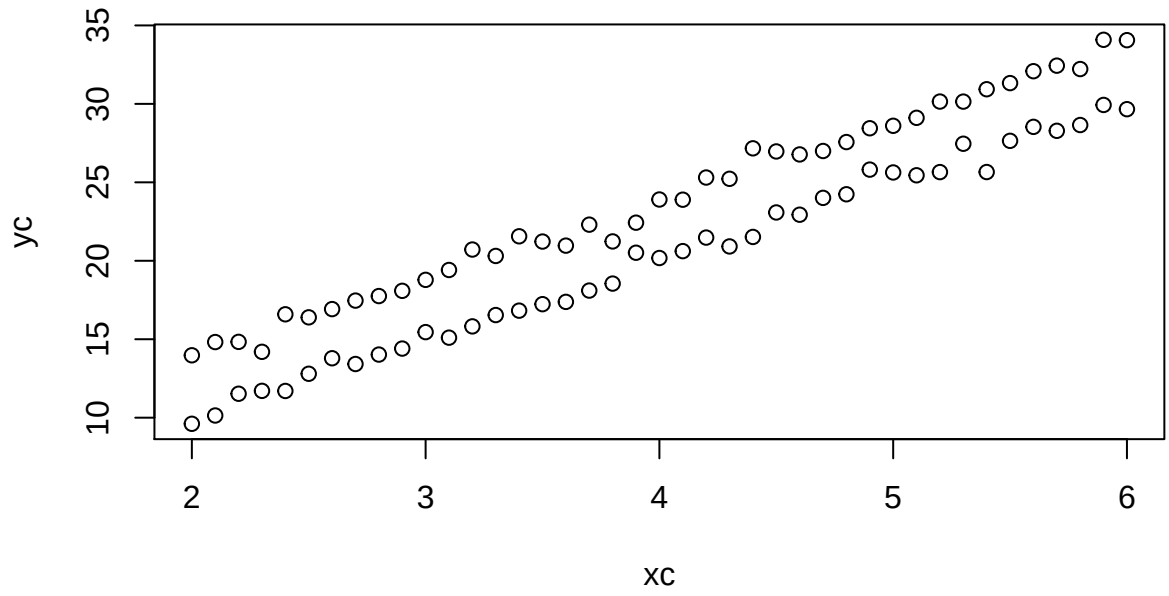
a. This statement is *false*. The assumption of normality for linear regression is that the error term, ϵ is normally distributed. This doesn't necessarily mean that all of the Y values are normally distributed - only that, for a given x-value, repeated observations of y are assumed to be normally distributed. Thus, if we draw the best-fit line between the observed (x,y) points, then at each x the observed response variables should be centered on the line of best fit and normally distributed. For this data, that appears to be true.

b. This statement is *false*. There are four assumptions to linear regression, the first is that a linear model is appropriate. Based on this graph, it appears that we have a regression line through data points clustered around x = 15 and x = 30. While it would be better to have additional data points between x = 15 and x = 30, the residuals do not indicate that a linear model is inappropriate. Just because the data is clumped does not imply that a linear model is inappropriate. Next, we assume that each residual is indepdenent. This appears to be true because there is no clear pattern in the data. Next we assume that the residuals are normally distributed and have equal variance. While the residuals around 30 seem to have a slightly higher variance, it's reasonable to assume that both residuals are normally distributed. Therefore, I conclude that there are no clear violations of any of the linear regression assumptions.

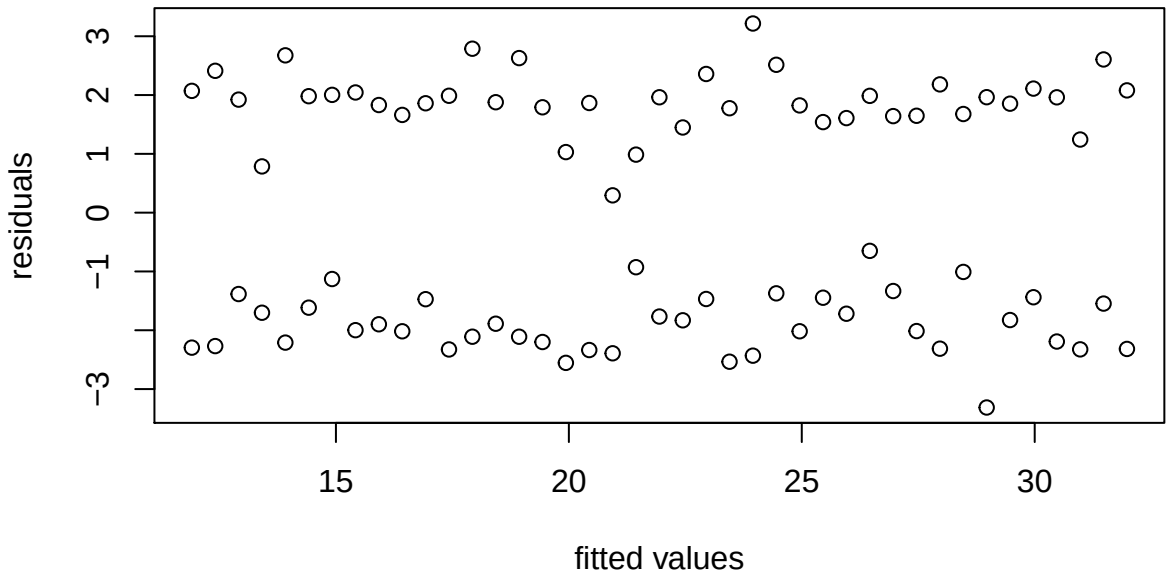
c. The assumption of normality is clearly violated. We expect that the residuals will be normally distributed and centered on zero. This plot has no residuals between -1 and 1 and is bimodal, so clearly the residuals are not normally distributed. I will provide two scenarios where this could happen. First, if there are two separate populations that should have separate but paralell regression lines but are instead grouped into one population then we will observe a plot of this form. An example is given in the R code below.

```r
#Problem 2c
#Generate data from two parallel lines
x = seq(2,6, by = 0.1)
error1 = rnorm(41,0,0.5)
error2 = rnorm(41,0,0.5)
y1 = 5*x + 4 + error1
y2 = 5*x + error2
xc = c(x,x); yc = c(y1,y2)

#Plot the example data
plot(yc~xc)
```

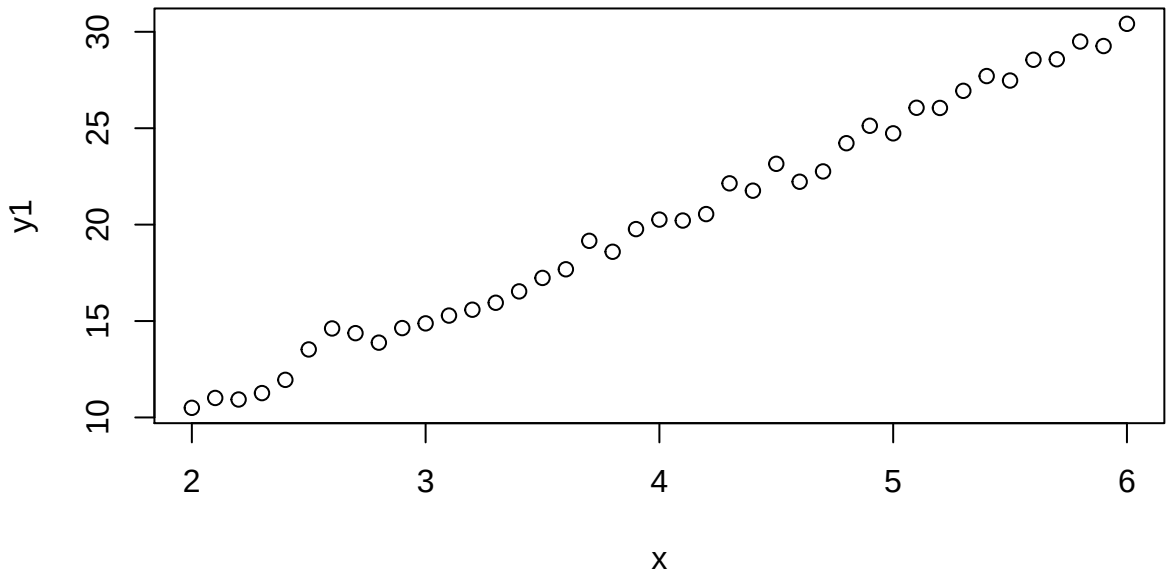



```
lm_example = lm(yc~xc)
plot(lm_example$residuals ~ lm_example$fitted.values, ylab = "residuals", xlab = "fitted values")
```



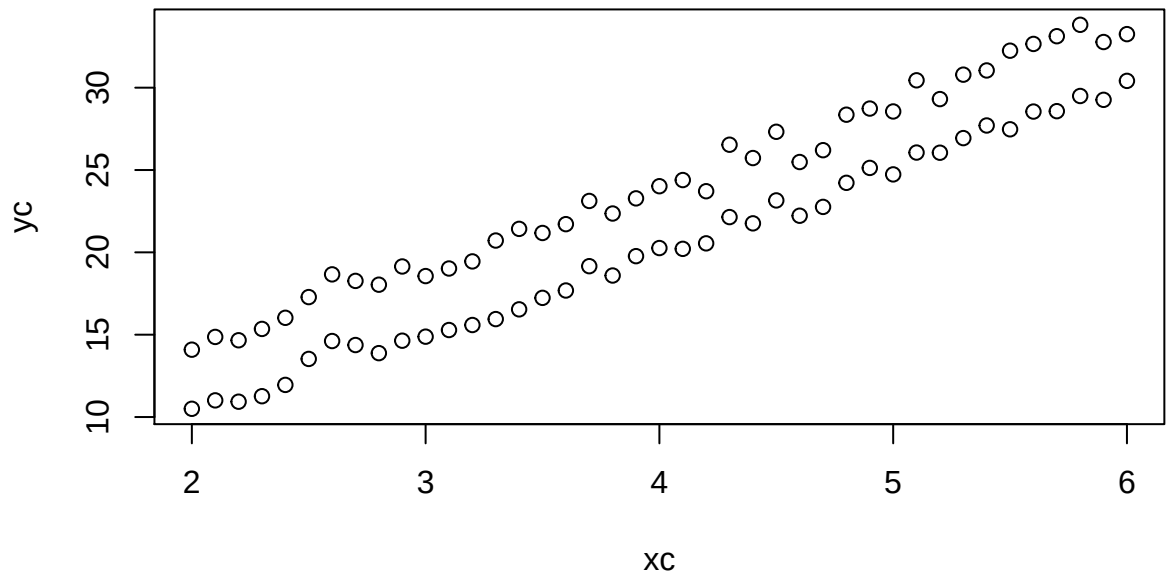
Next, if there is some dependency in the data, then we would observe a plot of this form. Specifically, if two data points (y1,y2) at the same x are related by y2 = y1 + 4 then we would observe this plot. This is illustrated using the R code below.

```
#Generate a data from a line
x = seq(2,6, by = 0.1)
error1 = rnorm(41,0,0.5)
error2 = rnorm(41,0,0.5)
y1 = 5*x + error1
plot(y1~x)
```

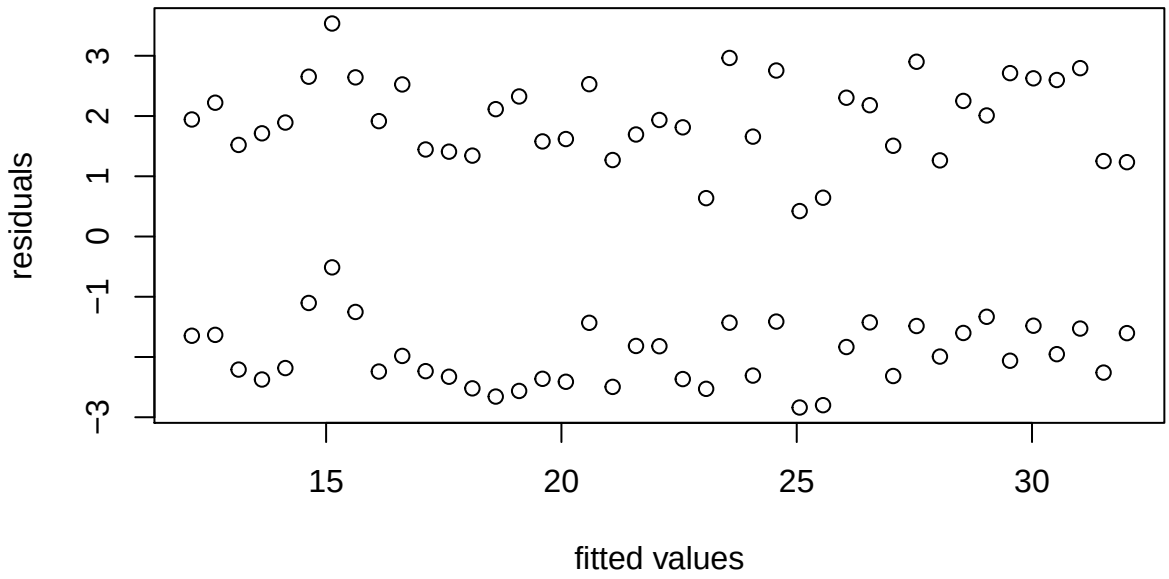



```
#Introduce dependent data points
y2 = y1 + 4 + error2
xc = c(x,x); yc = c(y1,y2)

#Plot dependent data
plot(yc~xc)
```




```
lm_example = lm(yc~xc)
plot(lm_example$residuals ~ lm_example$fitted.values, ylab = "residuals", xlab = "fitted values")
```



d. In this plot, there are two distinct regions - the sparse region below 2.1 and the dense region mostly above 2.4. There is likely no problem with unequal variance because the bulk of data points in the dense region are close to zero - which is the same in the sparse region. If the residuals are normally distributed and centered on zero, we would expect both regions to have most of their residuals close to zero and some out towards -3/3. We have many data points in the dense region, so we observe both the cluster around 0 and the few data points out towards the tails. In the sparse region, we have few data points, so there is a small likelihood of observing the rare residuals near -3/3. However, that does not imply that the observations in the sparse region have a different variance than observations in the dense region. Therefore, we cannot conclude with certainty that there is a problem with unequal variance.

Code

```
## ----include = FALSE-----------------------------------------------------
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(knitr)

set.seed(1104)              # make random results reproducible

this_file <- "kerr_stat850_hw01.Rmd"  # used to automatically generate code appendix

## ----fig.height = 5------------------------------------------------------
#Problem 1a
#Read in the data and perform the regression
outlier <- read.csv("~/2019spring/STAT850/hw1/outlier.txt", sep="")
outlier_lm <- lm(outlier$y ~ outlier$x, qr = T)
plot(outlier$y ~ outlier$x, main = "Plot of y vs. x")

#Perform diagnostics by looking at externally studentized residuals
outlier_lm_diag <- ls.diag(outlier_lm)
plot(outlier_lm_diag$stud.res ~ outlier_lm$fitted.values, ylab = "Externally Studentized Residuals", xla

## ---- fig.height = 5----------------------------------------------------
#Problem 1b
#i. Remove observation 9 from the data set
outlier_mod <- filter(outlier, indic == 0)

#ii. Regress the new data set
outlier_mod_lm <- lm(y~x ,data = outlier_mod)
summary(outlier_mod_lm)

#iii. Predict value when x = 12 and calculate standard error
y_pred = 8.7656 + 2.1992*12
mse = sum((outlier_mod_lm$residuals^2)/12)
ssqx = sum((outlier_mod$x-mean(outlier_mod$x))^2)
xbar = mean(outlier_mod$x)
se_pred = sqrt(mse*(1+(1/14)+(((12-xbar)^2)/ssqx)))
paste("The standard error for prediction is:  ",round(se_pred,3))

#iv. Test whether the questionable observation is within sampling error
t = (26 - y_pred)/(se_pred)
paste("The observed T statistic is ", round(t,3), "with 12 degrees of freedom")
```

```r
#v. Compute the p-value and apply Bonferroni
paste("The p-value for this test statistic is: ",round(2*pt(t,12),4))


## ---- fig.height=5-------------------------------------------------------
#Problem 1c
#i. Perform a multiple linear regression of y on x and indic
outlier_mlm <- lm(y ~ x + indic, data = outlier)
summary(outlier_mlm)

## ---- fig.height=4-------------------------------------------------------
#Problem 2c
#Generate data from two parallel lines
x = seq(2,6, by = 0.1)
error1 = rnorm(41,0,0.5)
error2 = rnorm(41,0,0.5)
y1 = 5*x + 4 + error1
y2 = 5*x + error2
xc = c(x,x); yc = c(y1,y2)

#Plot the example data
plot(yc~xc)
lm_example = lm(yc~xc)
plot(lm_example$residuals ~ lm_example$fitted.values, ylab = "residuals", xlab = "fitted values")

## ---- fig.height=4-------------------------------------------------------
#Generate a data from a line
x = seq(2,6, by = 0.1)
error1 = rnorm(41,0,0.5)
error2 = rnorm(41,0,0.5)
y1 = 5*x + error1
plot(y1~x)

#Introduce dependent data points
y2 = y1 + 4 + error2
xc = c(x,x); yc = c(y1,y2)

#Plot dependent data
plot(yc~xc)
lm_example = lm(yc~xc)
plot(lm_example$residuals ~ lm_example$fitted.values, ylab = "residuals", xlab = "fitted values")

## ----code = readLines(purl(this_file, documentation = 1)), echo = T, eval = F----
## # this R markdown chunk generates a code appendix
```