MA 354: Data Analysis I – Fall 2019
Homework 3:

Complete the following opportunities to use what we've talked about in class. These questions will be graded for correctness, communication and succinctness. Ensure you show your work and explain your logic in a legible and refined submission.

0. **Complete weekly diagnostics.**

1. Plankton samples are typically collected using fine mesh nets towed from research vessels, and larval fish are removed then stored after sample preservation. Samples are frequently collected using the paired bongo net, which consists of two usually round net frames joined at a central point, and towed either obliquely or vertically through the water column. Mesopelagic fish families such as the Myctophidae are some of the most specious and abundant in the worlds oceans. We expect the number of Myctophidae in the left and right side of each net to be highly correlated, but we want to quantify this relationship.

   Muhling et al. (2012) provide data on a total of 261 paired samples from the Gulf of Mexico. Myctophidae counts from the left and right sides of each bongo net are a result of over years (1987-2008) of sampling. We define X and Y to be the count of myctophid larvae in the left and right side of the bongo net, respectively.

   Research Question: Does the data we have confirm our expectation that the number of Myctophidae in the left and right side of each net tend to be highly correlated?

   The data can be loaded as follows. The data for the number of fish caught in the left and right net are in the columns labeled `Left` and `Right`, respectively.

```
> dat.bongo<-read.csv(file = "https://cipolli.com/students/data/BongoNetData.txt",
+                         header = TRUE,sep = ",")
```

   Soluion: First, I load the ggplot2 (Wickham, 2016) and gridExtra (Auguie, 2017) libraries for plotting in `R`.

```
> library("ggplot2")
> library("gridExtra")
```

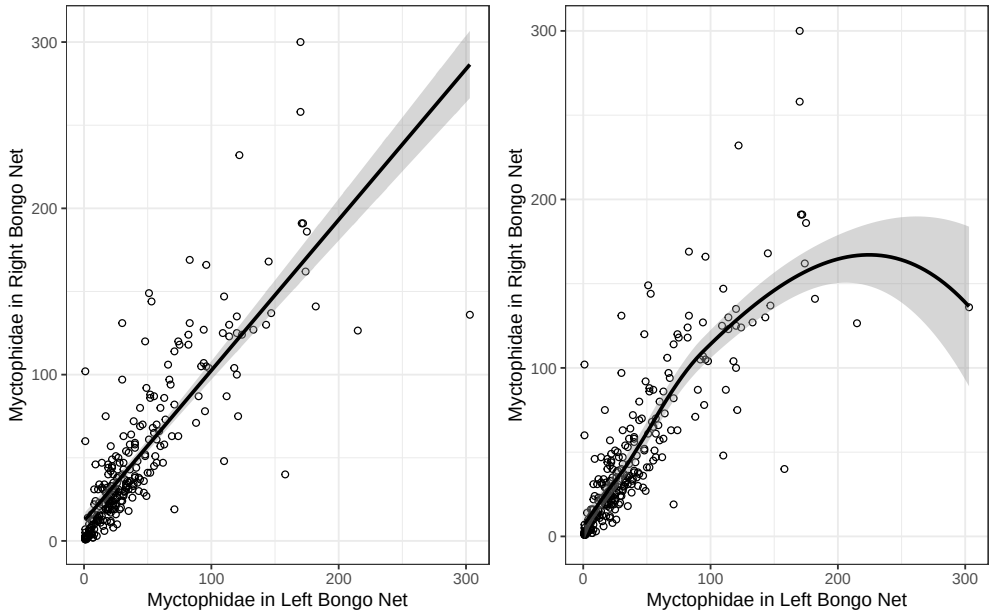


Figure 1: Bongo Data. Scatterplots for the number of Myctophidae caught in the Left and Right nets with a linear regression model (left) and loess curve fit (right) superimposed.

Since the data is discrete, and the data appears to be increasing at a decreasing rate we look to a rank-based correlation. Both Kendall's Tau-b and Spearman's rank indicate a moderate to strong monotone increasing relationship.

```
> cor(dat.bongo$Right,dat.bongo$Left,method = "kendall")

[1] 0.7105615

> cor(dat.bongo$Right,dat.bongo$Left,method = "spearman")

[1] 0.8639953
```

Note that the increased variability (and decreased number of points) show a reasonably similar fit for a linear model. This is confirmed numerically below where Pearson's correlation indicates a moderate to strong linear relationship.

```
> cor(dat.bongo$Right,dat.bongo$Left,method = "pearson")

[1] 0.8232054
```

By any definition, these data are highly dependent (as expected)

2. **(Working with Data)** Hepatitis C is a disease that affects the liver. The virus that causes hepatitis C is spread through blood or bodily fluids of an infected person. The virus is often difficult to diagnose because there are few unique symptoms. Those infected, however, sometimes experience jaundice – a condition that causes yellowing of the skin or eyes, as the liver is infected.

   Bracht et al. (2016) consider the human microfibrillar-associated protein 4, or MFAP4, and its role in disease-related tissue. Stage 0–no fibrosis; Stage 1–enlarged, fibrotic portal tracts; Stage 2–periportal fibrosis or portal-portal septa, but intact architecture; Stage 3–fibrosis with architectural distortion, but no obvious cirrhosis; and Stage 4–probable or definite cirrhosis.

   Previously, it has been shown that MFAP4 is a biomarker candidate for hepatic fibrosis and cirrhosis in hepatitis C patients. The analysis of Bracht et al. (2016) aimed to consider the ability of MFAP4

to differentiate between stages of the disease – fibrosis stages (0-2) and cirrhosis (3-4) based on the Scheuer scoring system.

Below, I load the data from the web.

```
> fn<-"http://cipolli.com/students/data/biomarker.csv"
> dat <- read.csv(file=fn, header=TRUE, sep=",")
> head(dat)

  Patient.ID Year.of.Birth Gender Date.of.sampling Fibrosis.Stage HCV.Genotype
1       1112          1958 female         2/1/2005              0            1
2       3403          1946 female        1/18/2005              2
3       2841          1954 female         1/3/2005              3            1
4        654          1958   male         2/1/2005              3            1
5       2788          1960   male        12/9/2004              0            3
6       2242          1954 female        5/12/2004              0            1
  MFAP4.U.mL
1        5.1
2        5.3
3       12.9
4        6.2
5        3.3
6        7.5
```

In homework 0, you recreated Table 1 in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4932744/. Now, recreate Table 2.

Solution:

```
> log2.MFAP4<-log(dat$MFAP4.U.mL,2)
> model<-aov(log2.MFAP4~as.factor(Fibrosis.Stage), data = dat)
> summary(model)

                          Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Fibrosis.Stage)   4  124.5  31.113   40.38 <2e-16 ***
Residuals                 537  413.8   0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> TukeyHSD(model, conf.level = 0.95)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log2.MFAP4 ~ as.factor(Fibrosis.Stage), data = dat)

$`as.factor(Fibrosis.Stage)`
         diff         lwr       upr     p adj
1-0 0.1923949 -0.11143830 0.4962281 0.4144134
2-0 0.4385611  0.11875460 0.7583677 0.0018040
3-0 1.1167632  0.73508734 1.4984391 0.0000000
4-0 1.4276545  1.04597862 1.8093304 0.0000000
2-1 0.2461662 -0.02872003 0.5210525 0.1036317
3-1 0.9243684  0.57945859 1.2692781 0.0000000
4-1 1.2352596  0.89034988 1.5801694 0.0000000
3-2 0.6782021  0.31914173 1.0372625 0.0000033
4-2 0.9890934  0.63003301 1.3481538 0.0000000
4-3 0.3108913 -0.10422874 0.7260113 0.2438166
```

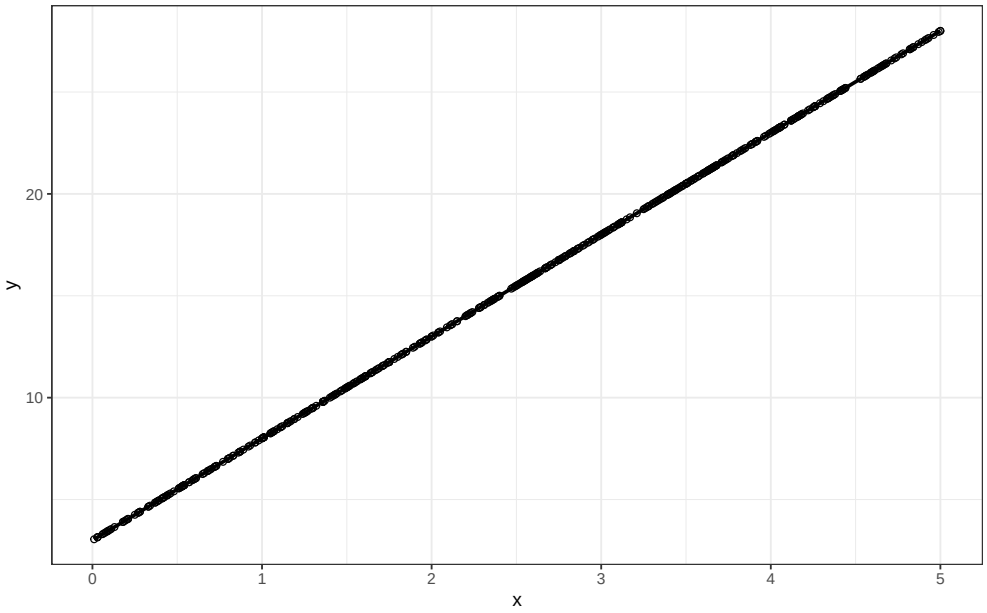
Comparison	Difference	Lower Bound	Upper Bound	p value
F1-F0	0.1924	-0.1114	0.4962	0.4144
F2-F0	0.4386	0.1188	0.7584	0.0018
F3-F0	1.1168	0.7351	1.4984	<0.0001
F4-F0	1.4277	1.0460	1.8093	<0.0001
F2-F1	0.2462	-0.0287	0.5211	0.1036
F3-F1	0.9244	0.5795	1.2693	<0.0001
F4-F1	1.2353	0.8903	1.5802	<0.0001
F3-F2	0.6782	0.3191	1.0373	<0.0001
F4-F2	0.9891	0.6300	1.3482	<0.0001
F4-F3	0.3109	-0.1042	0.7260	0.2438

Table 1: Results of the pairwise comparisons of individual hepatic fibrosis stages with respect to log2 MFAP4 values after significant ANOVA result

3. Complete the following parts. This will lead you through the simulation of data, fitting regression lines and evaluating the assumptions.

   (a) Fit a model to the following simulated data. Make observations about the model equation and the Pearson correlation.
   Solution:

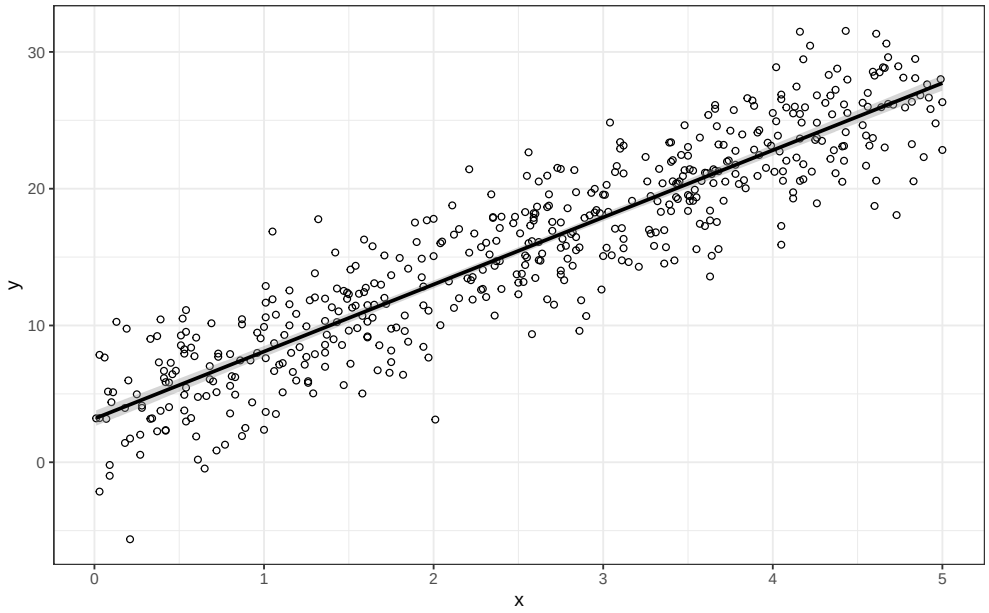
```
> n=500
> x<-sample(x = seq(0,5,0.01),size = n,replace = T)
> y<-5*x + 3
> cor(x,y)

[1] 1

> mod<-lm(y~x)
> summary(mod)

Call:
lm(formula = y ~ x)

Residuals:
       Min         1Q     Median         3Q        Max
-2.440e-13  1.350e-16  7.770e-16  1.633e-15  4.838e-15

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 3.000e+00  1.298e-15 2.311e+15   <2e-16 ***
x           5.000e+00  4.493e-16 1.113e+16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.39e-14 on 498 degrees of freedom
Multiple R-squared:      1,        Adjusted R-squared:      1
F-statistic: 1.238e+32 on 1 and 498 DF,  p-value: < 2.2e-16
```



Here, we see that the data is created to be a line with no error term. The equation of that line is returned exactly in the OLS regression function as calculated in R. The Pearson correlation is one and indicates that the relationship between x and y is a perfect, positive, linear relationship. Additionally, the R^2 value is 1 indicating that all of the variability in y has been explained by x; i.e., x perfectly predicts y.

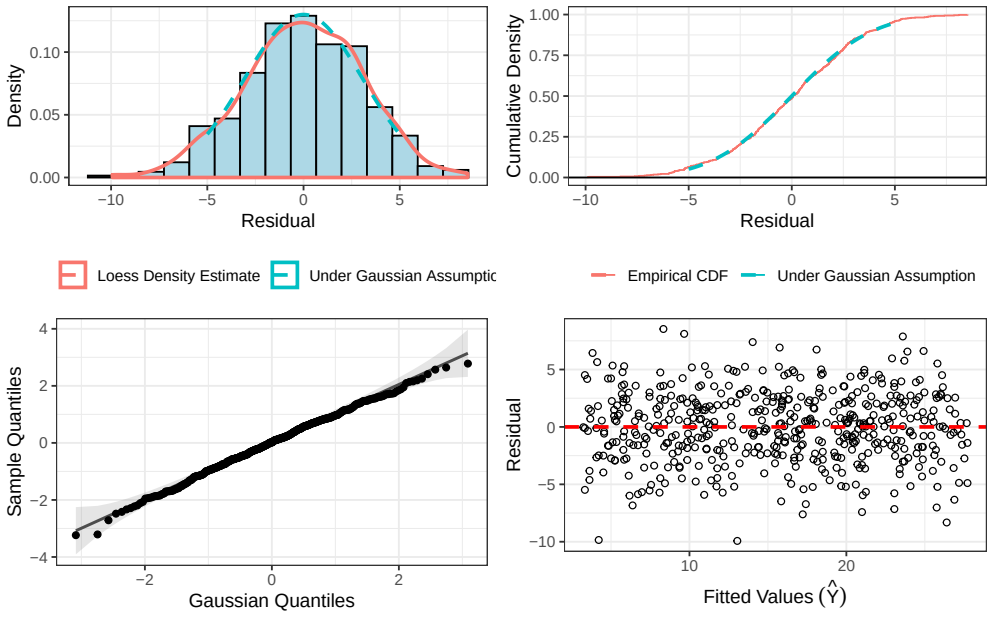
(b) Fit a model to the following simulated data, now with added Normal error. Make observations about the model equation and the Pearson correlation in relation to (a).
Solution:

```
> e<-rnorm(n=n,mean=0,sd=3)
> y2<-5*x + 3 + e
> cor(x,y2)
[1] 0.91144
> mod2<-lm(y2~x)
> summary(mod2)
Call:
lm(formula = y2 ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9283 -1.9871  0.0536  2.1795  8.5296

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.18673    0.28681   11.11   <2e-16 ***
x            4.90665    0.09925   49.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.071 on 498 degrees of freedom
Multiple R-squared:  0.8307,        Adjusted R-squared:  0.8304
F-statistic:  2444 on 1 and 498 DF,  p-value: < 2.2e-16
```



Here, we see that the relationship is no longer perfect since the random error was added. This is visible in the scatterplot where the points no longer lay perfectly on the line. The OLS regression line has coefficient estimates that are close to the line used to simulate the data but not exactly the same; this is due to the added error. Every time this code is run we yield a slightly different line do to the randomness of our simulated data. The correlation is slightly decreased, but still indicates that there is a strong positive, linear relationship between x and y. The R^2 values ranged between 0.80 and 0.90 indicating that a majority of the variation of y was explained by x.

(c) In the model of part (b), test for normality and constance of error terms. Note that we know both of these items to be true since we've taken $\epsilon \sim \text{N}(\mu = 0, \sigma = 3)$.
Solution: First, we can check the residual versus predicted plot. As expected, we have a nice and even-width band around the line at $y = 0$ (the mean of the residuals). This indicates constant variance of the residual terms.



Alternatively, we can use one of the hypothesis tests. Since we have good reason to believe (we know) that the residuals are normal, we can use the Breusch-Pagan test or the Koenker's studentized Breusch-Pagan test for heteroskedasticity. Lyon and Tsai (1996) suggests that Koenker's studentized Breusch-Pagan test works best for long-tailed or contaminated error distributions; e.g., non-normal and a test called Verbyla's residual likelihood ratio test is best overall, however it is not implemented in R.

```
> library("lmtest")
> ###Koenker Studentized Breusch-Pagan test
> bptest(mod2,studentize=TRUE)

        studentized Breusch-Pagan test

data:  mod2
BP = 0.20552, df = 1, p-value = 0.6503
```

This test indicates that there is little concern of heteroskedasticity as we fail to reject the null hypothesis that errors are homoskedastic (usually the test statistic and p-value go here, but the data changes every time we compile this file).

To assess normality we can consider the following plots. As expected the plots indicate normality – the histogram is matched by the normal PDF, the empirical CDF and normal CDF practically overlap and the qqplot is well fit to the 45 degree line.

Alternatively, we can use one of the hypothesis tests. Since we have a smaller sample size ($n = 500$) we can use the Shapiro Wilkes test or (preferred) Kolmogorov Smirnov.

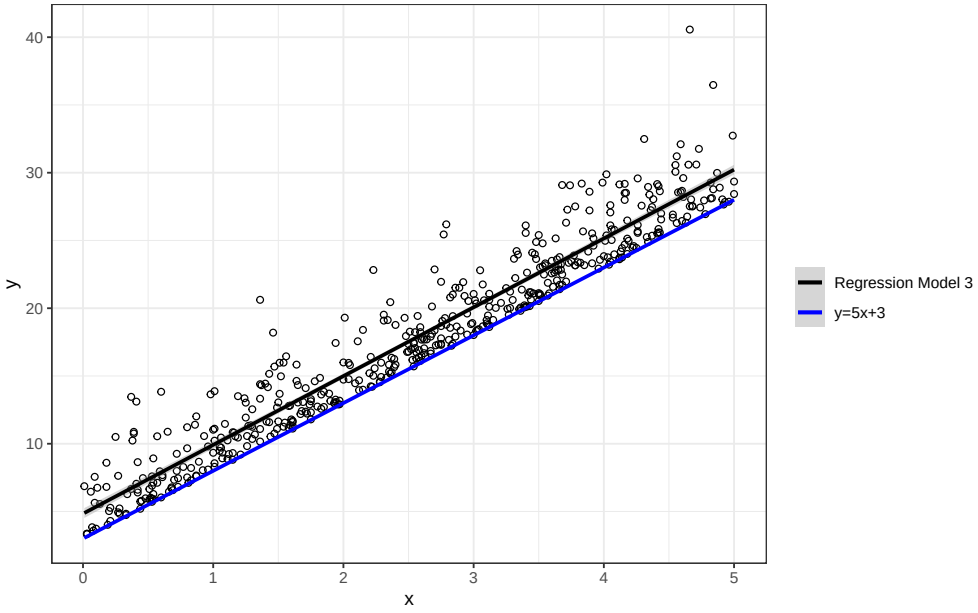
```
> shapiro.test(mod2$residuals)

        Shapiro-Wilk normality test

data:  mod2$residuals
W = 0.99732, p-value = 0.5987

> ks.test(mod2$residuals,pnorm,mean=0,sd=summary(mod2)$sigma)

        One-sample Kolmogorov-Smirnov test

data:  mod2$residuals
D = 0.025206, p-value = 0.9085
alternative hypothesis: two-sided
```

This test indicates that there is little concern about the errors being non-normal as we fail to reject the null hypothesis that errors are normal (usually the test statistic and p-value go here, but the data changes every time we compile this file).

(d) Fit a model to the following simulated data, now with added exponential error. Make observations about the model equation and the Pearson correlation in relation to the model of part (b).
Solution:

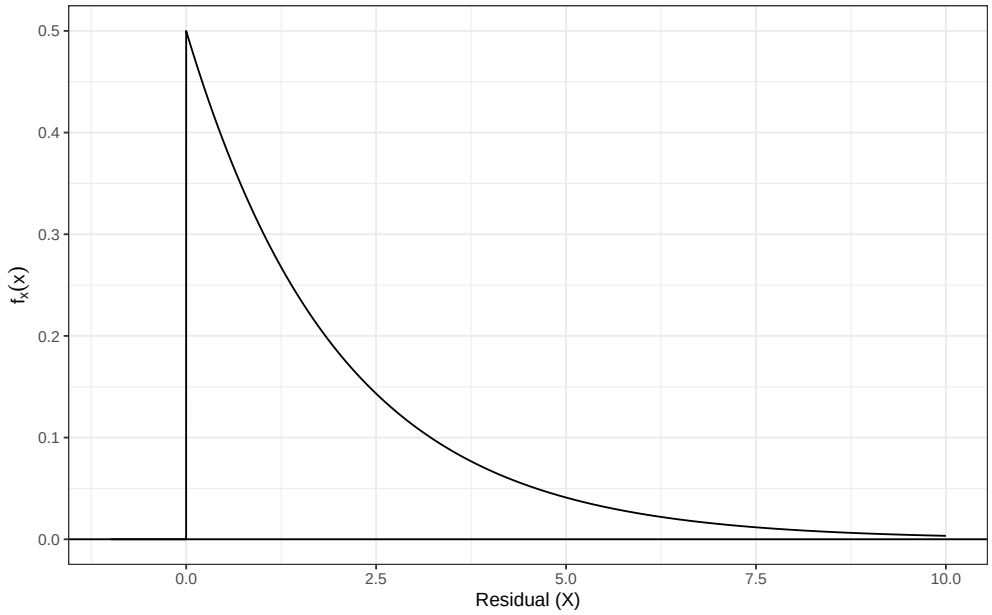
```
> e<-rexp(n=n,rate = 1/2)
> y3<-5*x + 3 + e
> cor(x,y3)

[1] 0.9620663

> mod3<-lm(y3~x)
> summary(mod3)

Call:
lm(formula = y3 ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-2.1644 -1.5059 -0.5823  0.8631 12.0653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.83847    0.18643   25.95   <2e-16 ***
x            5.07703    0.06451   78.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.996 on 498 degrees of freedom
Multiple R-squared:  0.9256,        Adjusted R-squared:  0.9254
F-statistic:  6193 on 1 and 498 DF,  p-value: < 2.2e-16
```

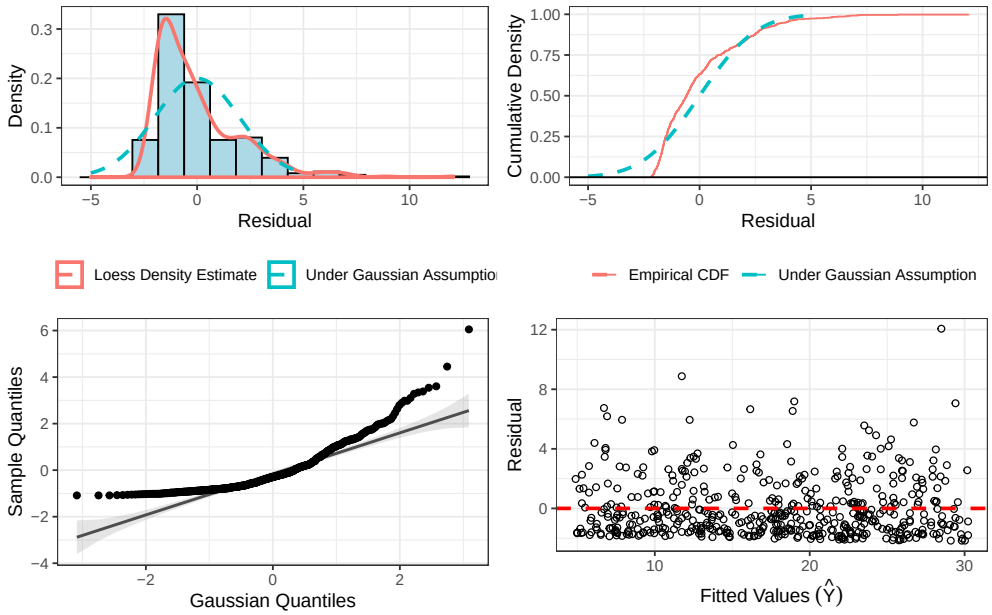


Here, we see a similar linear patten. The correlation is still high, but we notice that the errors above the curve are further than those below. I added the line we're using to simulated the data to show you what's happening. The errors that we added to simulate y were positive; I plotted the distribution that we drew the errors from – the support is positive which is why all of the points above the $y = 5x + 3$ line. Note that the residuals with respect to the regression line are both positive and negative. The OLS regression line has a reasonable slope estimate which is close to 5, but the intercept is biased on the high side because of the mispecificaiton of the error distribution. Although the correlation still indicates that there is a strong positive, linear relationship between x and y and the R^2 value indicates that a majority of the variation of y was explained by x, the coefficient estimates are biased and can cause problems as we interpret them for meaning.

```
> errors<-seq(-1,10,0.001)
> fe<-dexp(errors, rate = 1/2)
```



(e) In the model of part (d), test for normality and constance of error terms. Note that we know that common variance is true but we've taken $\epsilon \sim \exp(\beta = 2)$.
Solution: First, we can check the residual versus predicted plot. As expected, we have an even-width band as the variance is constant; recall, $var(X) = \beta^2 = 4$ for $X \sim$ Exponential$(\beta = 2)$. We note, however, that the band of errors isn't balanced around the line at $y = 0$ (the mean of the residuals which is $E(X) = \beta = 2$ here). This indicates constant variance of the residual terms.



Alternatively, we can use one of the hypothesis tests. Again, we employ the Koenker's studentized Breusch-Pagan test for heteroskedasticity.

```
> library("lmtest")
> ###Koenker Studentized Breusch-Pagan test
> bptest(mod3,studentize=TRUE)
```

```
        studentized Breusch-Pagan test

data:  mod3
BP = 0.79054, df = 1, p-value = 0.3739
```

This test indicates that there is little concern of heteroskedasticity as we fail to reject the null hypothesis that errors are homoskedastic (usually the test statistic and p-value go here, but the data changes every time we compile this file).

As expected the plots indicate non-normality – the histogram is not well matched by the normal PDF, the empirical CDF and normal aren't even close to overlapping and the qqplot is not well fit by the 45 degree line.

Alternatively, we can use one of the hypothesis tests. Since we have a smaller sample size ($n = 500$) we can use the Shapiro Wilkes test.

```
> shapiro.test(mod3$residuals)

        Shapiro-Wilk normality test

data:  mod3$residuals
W = 0.84357, p-value < 2.2e-16

> ks.test(mod3$residuals,pnorm,mean=0,sd=summary(mod3)$sigma)

        One-sample Kolmogorov-Smirnov test

data:  mod3$residuals
D = 0.14467, p-value = 1.629e-09
alternative hypothesis: two-sided
```

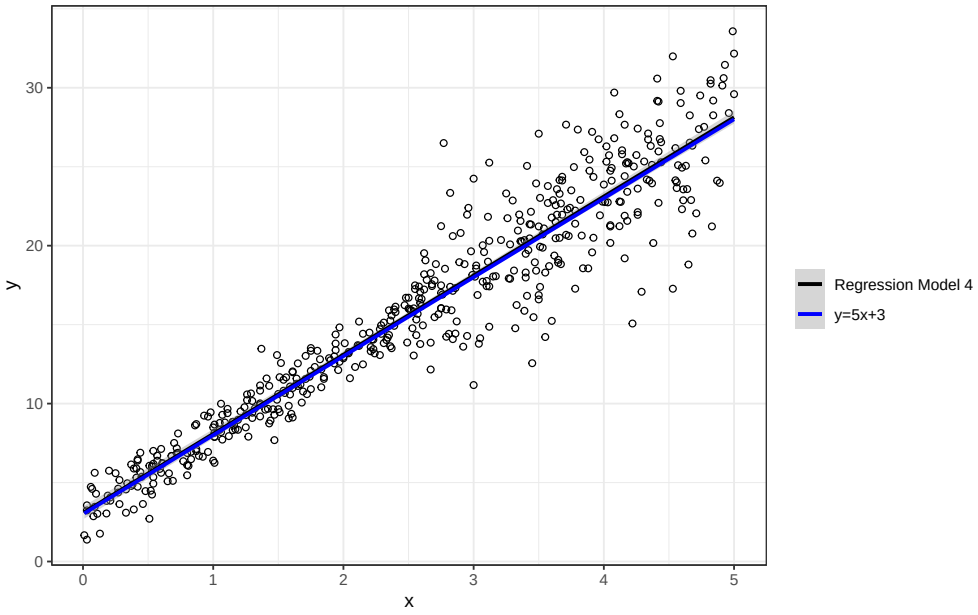
These test indicate that there is concern about the errors being non-normal since our p-value is small ($< \alpha = 0.05$). We reject the null hypothesis that errors are normal (usually the test statistic and p-value go here, but the data changes every time we compile this file) indicating that they likely are drawn from some other distribution, here the exponential.

(f) Fit a model to the following simulated data, now with added Heteroskedastic normal error. Make observations about the model equation and the Pearson correlation in relation to the model of part (b).

Solution:

```
> #order X to simulate non constant error
> x4<-x[order(x)]
> e<-rnorm(n=n,mean=0,sd=c(rep(1,n/2),rep(3,n/2)))
> y4<-5*x4 + 3 + e
> cor(x4,y4)

[1] 0.9520182

> mod4<-lm(y4~x4)
> summary(mod4)

Call:
lm(formula = y4 ~ x4)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1510 -1.0242  0.0324  1.1752  9.5404

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.07760    0.20858   14.76   <2e-16 ***
x4           5.01056    0.07218   69.42   <2e-16 ***
```

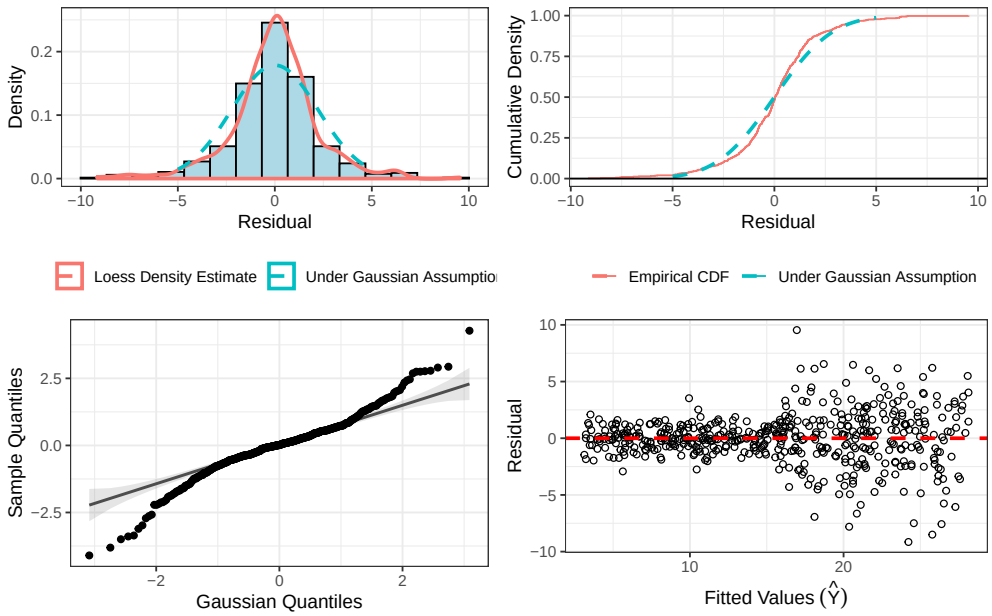
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.233 on 498 degrees of freedom
Multiple R-squared:  0.9063,        Adjusted R-squared:  0.9062
F-statistic:  4819 on 1 and 498 DF,  p-value: < 2.2e-16
```



Here, we see a similar linear patten. The correlation is still high, but we notice that after the sample median the width of the band of points around the regression line widens – this is due to the increased variance on the error terms for observations greater than the sample median. Despite this issue in model assumptions, the OLS regression line has reasonable estimates of both the slope and the intercept. Although the correlation still indicates that there is a strong positive, linear relationship between x and y and the R^2 value indicates that a majority of the variation of y was explained by x, heteroskedasticity is present. When heteroscedasticity is present, the observations with larger residuals can have more of an effect on the OLS fit of the regression line.

(g) In the model of part (f), test for normality and constance of error terms. Note that we know that normality of error terms is true, but $\epsilon \sim \mathrm{N}(\mu = 0, \sigma = 1)$ for $x < \widehat{m}$ and $\epsilon \sim \mathrm{N}(\mu = 0, \sigma = 3)$ for $x > \widehat{m}$.

Solution: First, we can check the residual versus predicted plot. As expected, we don't have an even-width band as the variance is non-constant – the variance of residuals increases at the sample median.



Alternatively, we can use one of the hypothesis tests. Again, we employ the Koenker's studentized Breusch-Pagan test for heteroskedasticity.

```
> library("lmtest")
> ###Koenker Studentized Breusch-Pagan test
> bptest(mod4,studentize=TRUE)

        studentized Breusch-Pagan test

data:  mod4
BP = 58.338, df = 1, p-value = 2.207e-14
```

This test indicates that there is a strong concern of heteroskedasticity as we reject the null hypothesis that errors are homoskedastic (usually the test statistic and p-value go here, but the data changes every time we compile this file).

To assess normality we can consider the following plots. As expected the plots indicate approximate normality but the $RMSE$ doesn't approximate the variance well due to the heteroskedasticity. Looking at the histogram of residuals we see that the estimated density has slightly fatter tails than the superimposed normal distribution – this is seen in the CDF and qqplot as well.

Alternatively, we can use one of the hypothesis tests. Since we have a smaller sample size ($n = 500$) we can use the Shapiro Wilkes test.

```
> shapiro.test(mod4$residuals)

        Shapiro-Wilk normality test

data:  mod4$residuals
W = 0.95747, p-value = 7.969e-11
> ks.test(mod4$residuals,pnorm,mean=0,sd=summary(mod2)$sigma)

        One-sample Kolmogorov-Smirnov test

data:  mod4$residuals
D = 0.14513, p-value = 1.424e-09
alternative hypothesis: two-sided
```

This test indicates that there is concern about the errors being non-normal since our p-value is small ($< \alpha = 0.05$). We reject the null hypothesis that errors are normal (usually the test statistic

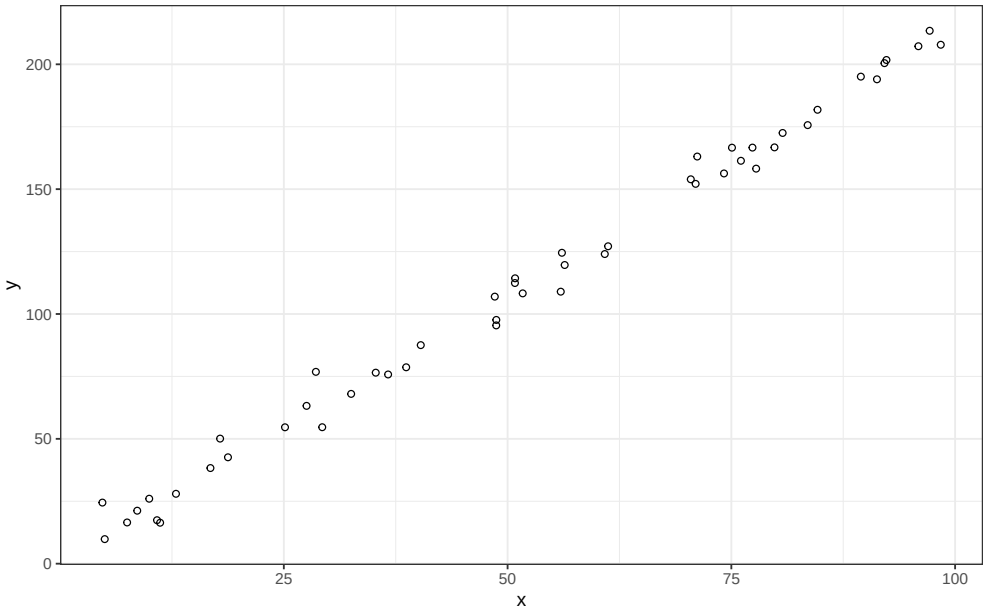
and p-value go here, but the data changes every time we compile this file) indicating that they likely are drawn from some other distribution. This should raise a concern – our errors were drawn from the normal distribution. This error is due to the heteroskedastic error which yields a distribution of residuals that is more peaked and has fatter tails than the normal distribution.

4. Consider the following simulation. This looks intimidating, but it's a fairly simple exploration about what we'll talk about in class. This walks you through "seeing" what's going on in the background.

   (a) Plot the data simulated below. Assess the linear relationship.

```
> set.seed(50)
> x_1<-sample(x=seq(0,100,0.01),size=50,replace=TRUE)
> e_1<-rnorm(n=50,mean=0,sd=5)
> y_1<-3.5+2.1*x_1 + e_1
> cor(x_1,y_1)
[1] 0.9960008
```

   Solution: The correlation indicates what we see in the graph, that there is a strong linear relationship.



   i. Write out the population model.

$$Y_i = \beta_0 + \beta_1 X i + \epsilon_i$$

   Solution:

$$Y_i = 3.5 + 2.1 X_i + \epsilon_i$$

   ii. Fit the model based on the sample data and write out the sample model below.
   Solution: The sample model is calculated in R as follows.

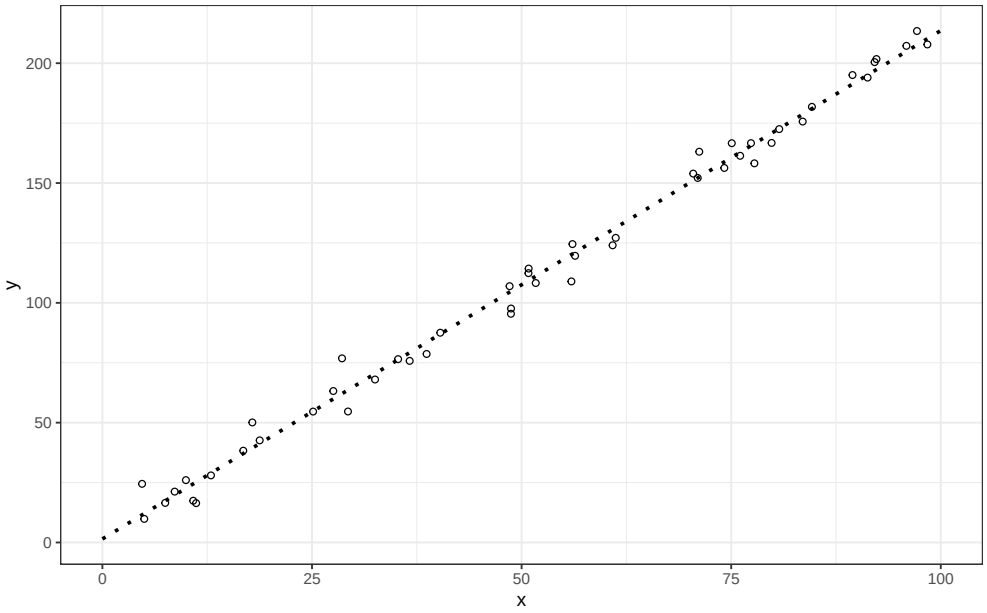
```
> mod1<-lm(y_1~x_1)
> summary(mod1)
Call:
lm(formula = y_1 ~ x_1)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-11.2041  -2.9853  -0.0717   3.2625  14.7368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.46858    1.62986   0.901    0.372
x_1          2.12145    0.02747  77.235   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.646 on 48 degrees of freedom
Multiple R-squared:  0.992,        Adjusted R-squared:  0.9919
F-statistic:  5965 on 1 and 48 DF,  p-value: < 2.2e-16
```

We write the sample model out as

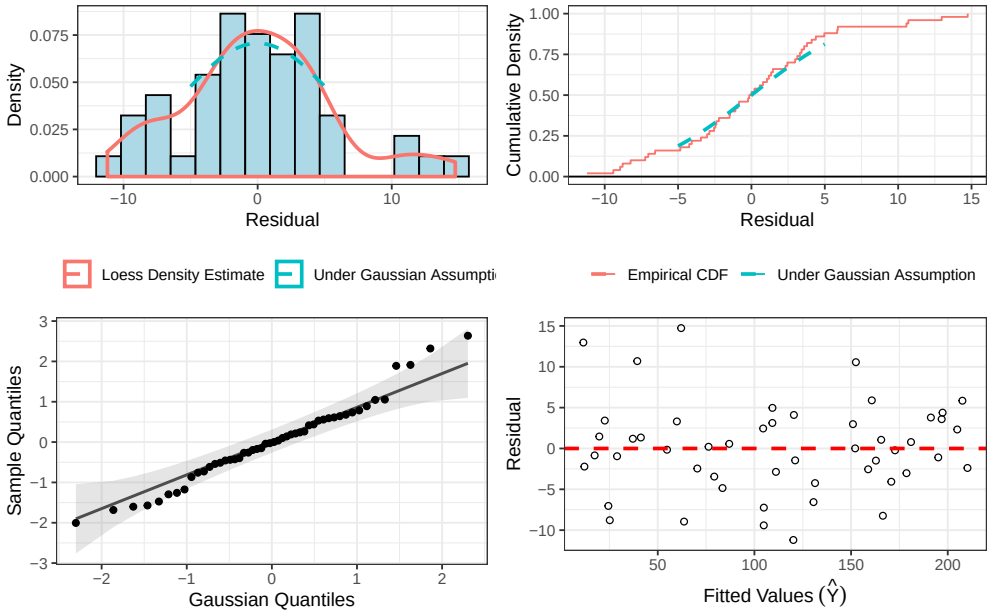
$$\widehat{y_i} = 1.46858 + 2.12145x_i + e_i.$$

iii. Add the regression line to the plot in black, with lwd=2 and lty=3.



iv. Check the assumptions of OLS for this model.
Solution: Below, we see that the residuals have an even pattern around zero and the residuals appear to be normal in the qqplot. There are no apparent outliers; checking for standardized residuals with magnitude greater than four confirms our visual assessment.

```
> length(which(rstandard(mod1)>4)) #check for outliers
[1] 0
```



v. Interpret the R^2 of the model.
Solution: The adjusted R^2 indicates that about 99% of the variability in y is explained by x.

```
> summary(mod1)$adj.r.squared
[1] 0.9918514
```

vi. Interpret the overall F test of the model. Report all 5 steps.
Solution: An overall F test suggests that our model fits better than the intercept only model ($F_{1,48} = 5965$,p value< 0.0001). This test is appropriate because the assumptions of our model are met and it suggests the coefficient for our only coefficient is non-zero.

```
> summary(mod1)

Call:
lm(formula = y_1 ~ x_1)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2041  -2.9853  -0.0717   3.2625  14.7368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.46858    1.62986   0.901    0.372
x_1          2.12145    0.02747  77.235   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.646 on 48 degrees of freedom
Multiple R-squared:  0.992,        Adjusted R-squared:  0.9919
F-statistic:  5965 on 1 and 48 DF,  p-value: < 2.2e-16
```

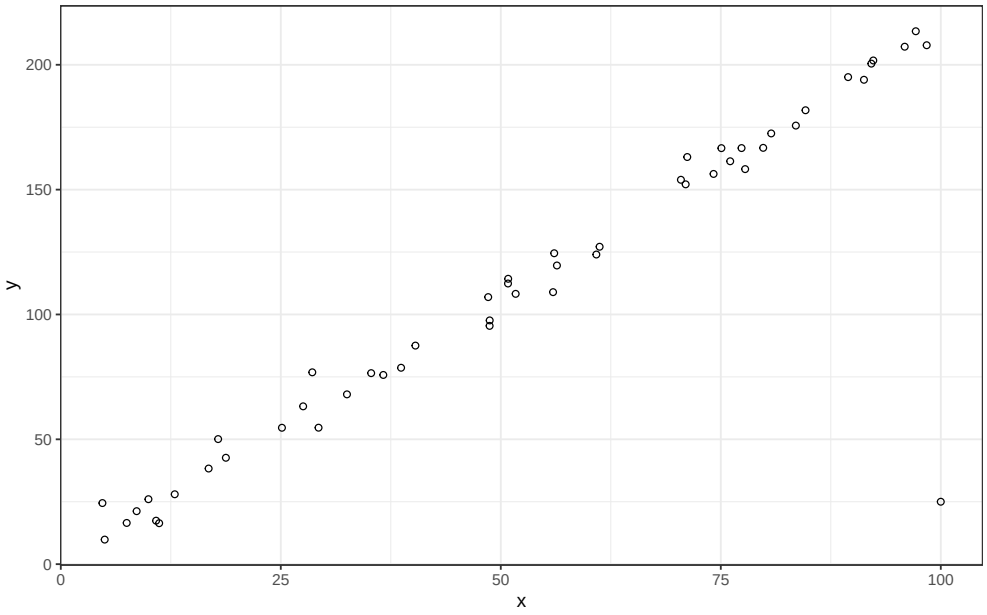
vii. Interpret the coefficients of the model; are they what you would expect?
Solution: The expected y for an observation of $x = 0$ is 1.46858. For every unit increase in x, we expect a 2.12145 increase in y on average. We expect this because the population coefficient is 2.1 – this OLS methodology has worked quite well, although we recognize the intercept estimate is inaccurate.

(b) Now, let's add a bad datapoint to the data created in Question a.

i. Plot the data simulated below; ensure to plot the Assess the linear relationship.

```
> x_2<-c(x_1,100)
> y_2<-c(y_1,25)
> cor(x_2,y_2)
[1] 0.9088512
```

Solution: The correlation indicates what we see in the graph, that there is a strong linear relationship.



ii. Fit the model based on the sample data and write out the sample model below.

Solution: The sample model is calcualted in R as follows.

```
> mod2<-lm(y_2~x_2)
> summary(mod2)
Call:
lm(formula = y_2 ~ x_2)

Residuals:
     Min       1Q   Median       3Q      Max
-175.431   -2.876    2.983    9.398   18.459

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.328      7.607   1.095    0.279
x_2            1.921      0.126  15.252   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.58 on 49 degrees of freedom
Multiple R-squared:  0.826,        Adjusted R-squared:  0.8225
F-statistic: 232.6 on 1 and 49 DF,  p-value: < 2.2e-16
```

We write the sample model out as

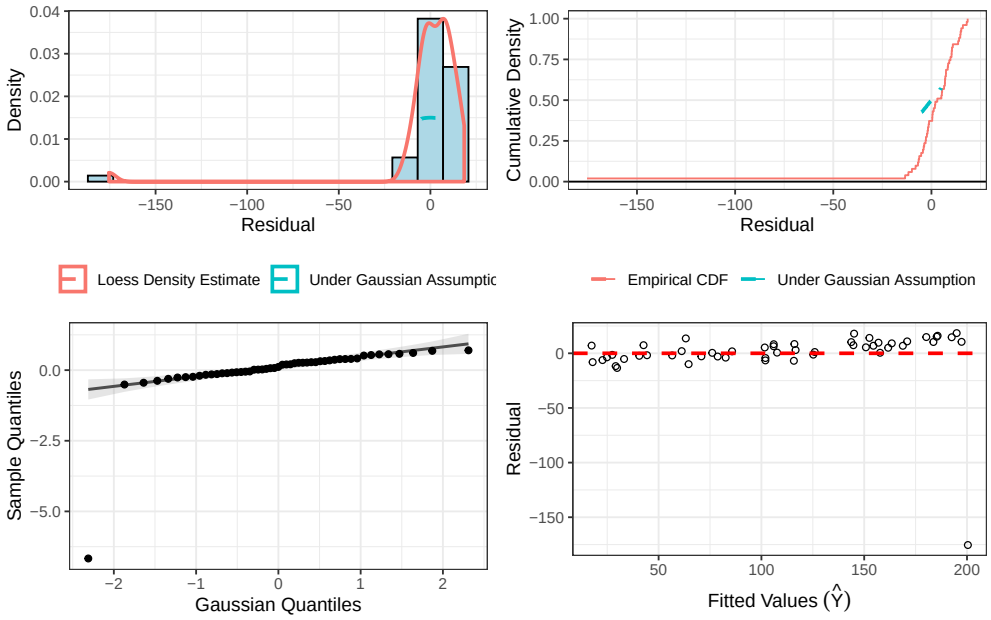
$$\widehat{y_i} = 8.328 + 1.921x_i + e_i.$$

We note that adding one poorly behaved point decreases the R^2 by approximately 10%!

iii. Check the assumptions of OLS for this model.

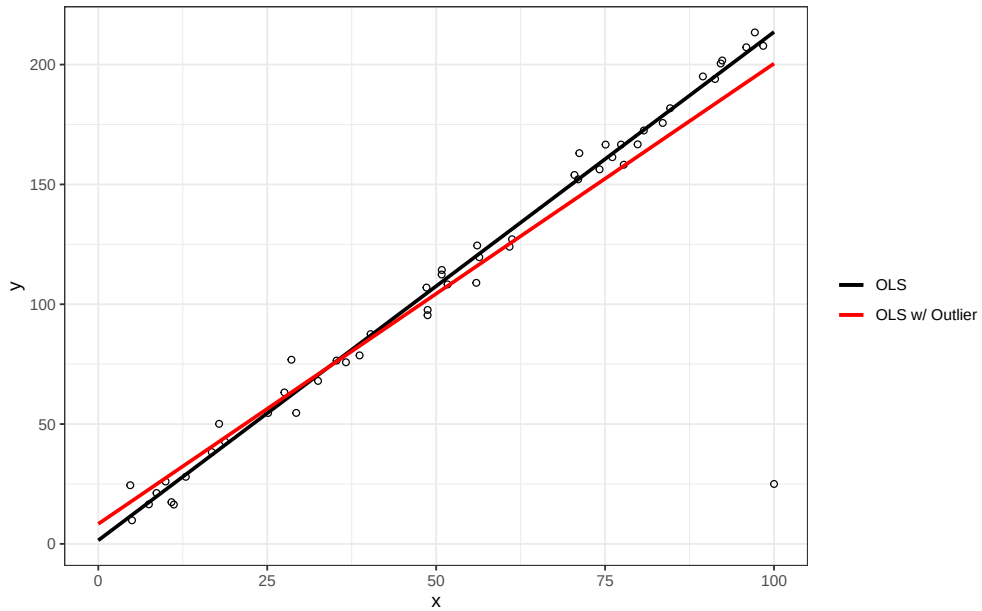
Solution: Below, we see that the residuals do not have an even pattern around zero, but an increasing trend across fitted values. There is a clear outlying value marked 51 – this is the added observation. However, the residuals appear to be normal in the qqplot. The outlying observation is confirm by checking standardized residuals.

```
> length(which(abs(rstandard(mod2))>4)) #check for outliers
[1] 1
```



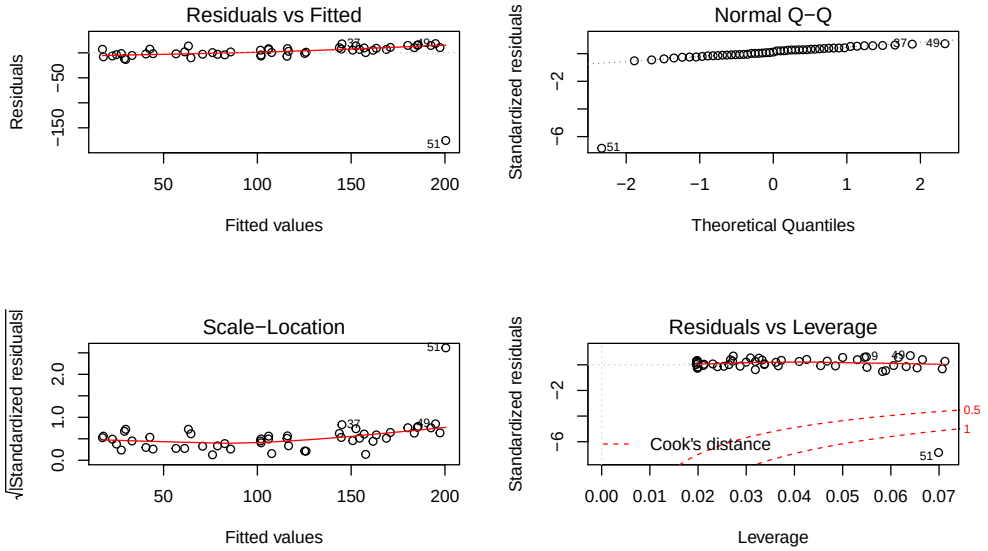
iv. Plot the data and regression lines from Questions a-b. Add the regression line to the plot in red.

Solution:



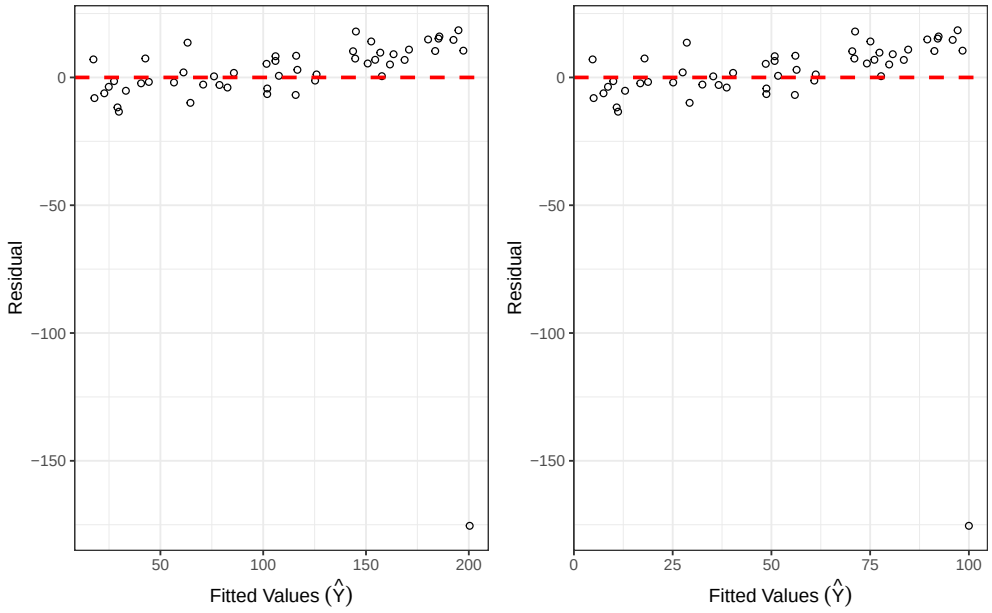
v. Interpret the coefficients of the model; are they what you would expect?
Solution: The expected y for an observation of $x = 0$ is 8.328 For every unit increase in x, we expect a 1.921 increase in y on average. Due to the outlier, the coefficients β_0 and β_1 are biased – the outlier tilts the OLS regression line toward it.
This tilting occurs in a way that is similar to a seesaw, noting we sit on the ends of the seesaw which have the most leverage. Because the outlier is on the end, it has more leverage than it would if it were closer to the middle. The visualization of leverage can be retrieved in `R` as follows. The point labeled 51 is the outlier we've added to the model. It is beyond the Cook's distance lines indicating that it has a "leveraging" effect on the OLS line.

```
> par(mfrow=c(2,2))
> plot(mod2)
```



(c) Continue with the data from Question b.

i. Plot the residuals of your model against x_2 as well as the residuals against predicted.
Solution:



ii. Fit the appropriate model for estimating the weights for weighted least squares regression (See Lecture 17.)
Solution: The e_i versus x plot indicates that we should estimate the weights using the following regression function,

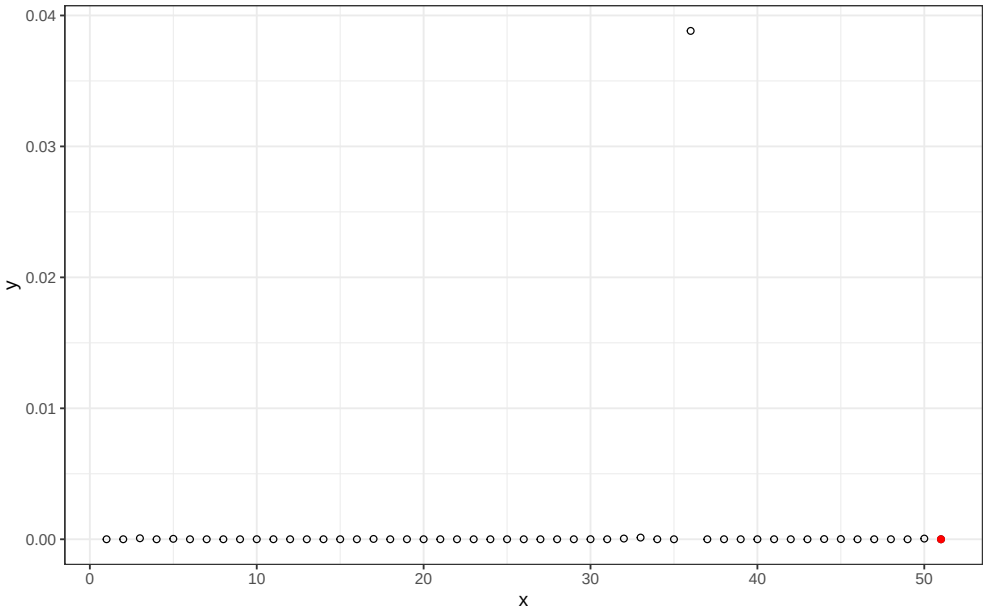
$$\epsilon_i^2 = \beta_0 + \beta_1 x.$$

```
> mod_weights<-lm(residuals(mod2)^2~x_2)
> weights<-1/(fitted(mod_weights)^2)
```

iii. Provide a summary of the weights. How does the weight for the observation (100,25) compare to the other observations?
Solution: A summary and plot of the weights show that the outlier has a low weight – the lowest overall, in fact.

```
> summary(weights)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
1.900e-07 4.300e-07 1.500e-06 7.701e-04 3.400e-06 3.882e-02
```



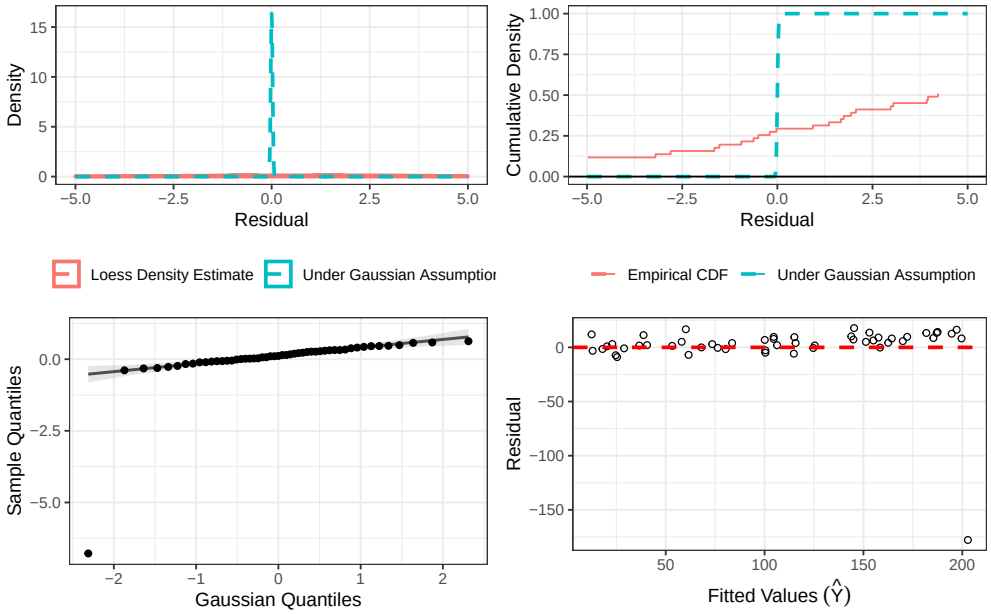
iv. Use the weights from the previous part to fit a weighted least squares regression.
Solution: Using the weights from part c(ii), we fit the model as follows.

```
> mod3<-lm(y_2~x_2,weights=weights)
> summary(mod3)

Call:
lm(formula = y_2 ~ x_2, weights = weights)

Weighted Residuals:
      Min        1Q    Median        3Q       Max
-0.077841 -0.001035  0.004332  0.007662  0.120191

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0696     3.8060   0.807    0.424
x_2           1.9975     0.1169  17.083   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02441 on 49 degrees of freedom
Multiple R-squared:  0.8562,        Adjusted R-squared:  0.8533
F-statistic: 291.8 on 1 and 49 DF,  p-value: < 2.2e-16
```

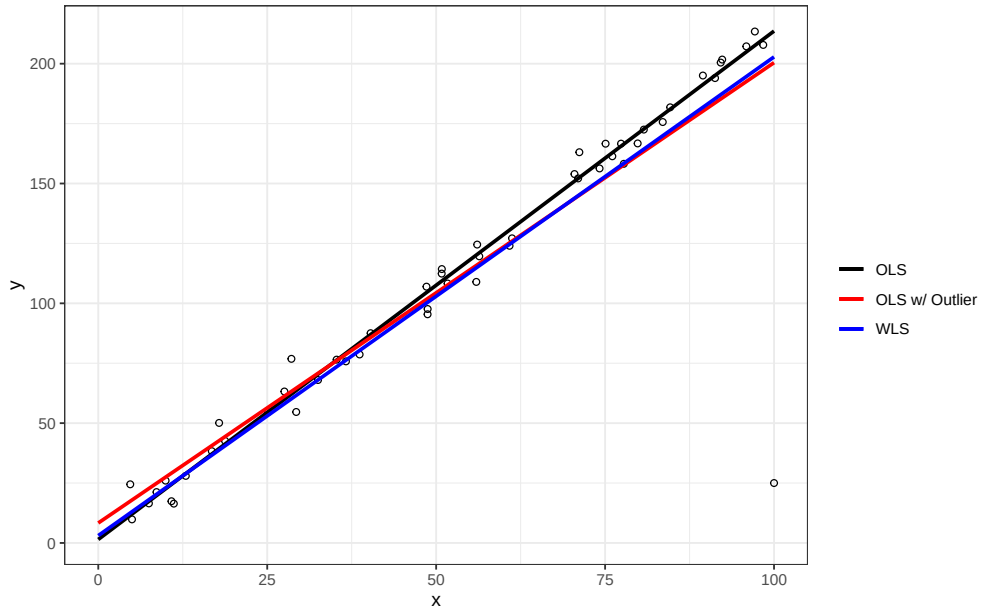
We write the sample model out as

$$\widehat{y_i} = 3.0696 + 1.975x_i + e_i.$$

v. Check the assumptions of OLS for this model.
Solution: Below, we see that the residuals have a similar increasing trend in the residual plot but with a decreased slope. Using weighted least squares doesn't better fit the outlier but the rest of the data. We should also note that the normality diagnostic plots (histogram and ECDF) are comical, while we see normality qqplot shows roughly Gaussian errors with exception to the outlier.



vi. Plot the data and regression lines from Questions a-b. Add the regression line to the plot in blue.



vii. Interpret the coefficients of the model; are they what you would expect?
Solution: The expected y for an observation of $x = 0$ is 3.0696. For every unit increase in x, we expect a 1.9975 increase in y on average. Giving the outlying observation less weight than the other observations yields coefficients β_0 and β_1 close to the original OLS estimates.

(d) Continue with the data from Question c.

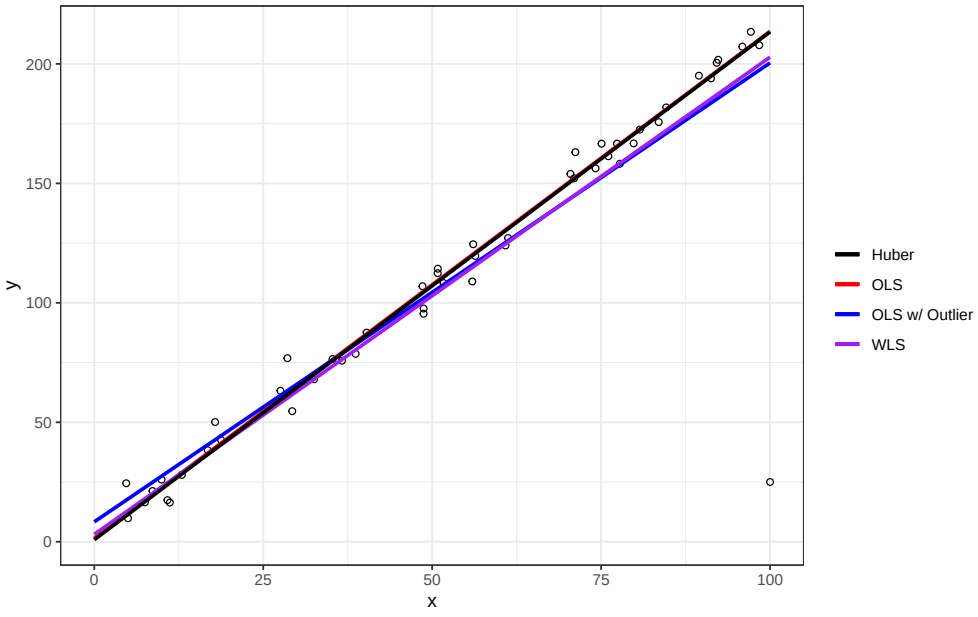
i. Fit a robust regression using Huber-weighted iterated reweighted least squares and write out the sample model below.
Solution: We fit the Huber IRLS sample model as follows in R.

```
> library("MASS")
> mod4<-rlm(y_2~x_2,psi=psi.huber)
> summary(mod4)
Call: rlm(formula = y_2 ~ x_2, psi = psi.huber)
Residuals:
      Min        1Q    Median        3Q       Max
-188.4031   -2.8540    0.3427    3.6831   15.2490

Coefficients:
            Value   Std. Error t value
(Intercept)  0.8357  1.6047     0.5208
x_2          2.1257  0.0266    80.0091

Residual standard error: 5.248 on 49 degrees of freedom
```

We write the sample model out as

$$\widehat{y_i} = 0.8357 + 2.1257x_i + e_i.$$

ii. Plot the data and regression lines from Questions a-c. Add the regression line to the plot in purple.
Solution:



iii. Interpret the coefficients of the model; are they what you would expect?
Solution: The expected y for an observation of $x = 0$ is 0.8357. For every unit increase in x, we expect a 2.1357 increase in y on average. Giving the outlying observation less weight than the other observations yields the coefficient β_1 close to the original OLS estimate and β_0 to be the closest intercept estimate yet.

(e) Continue with the data from Question d.

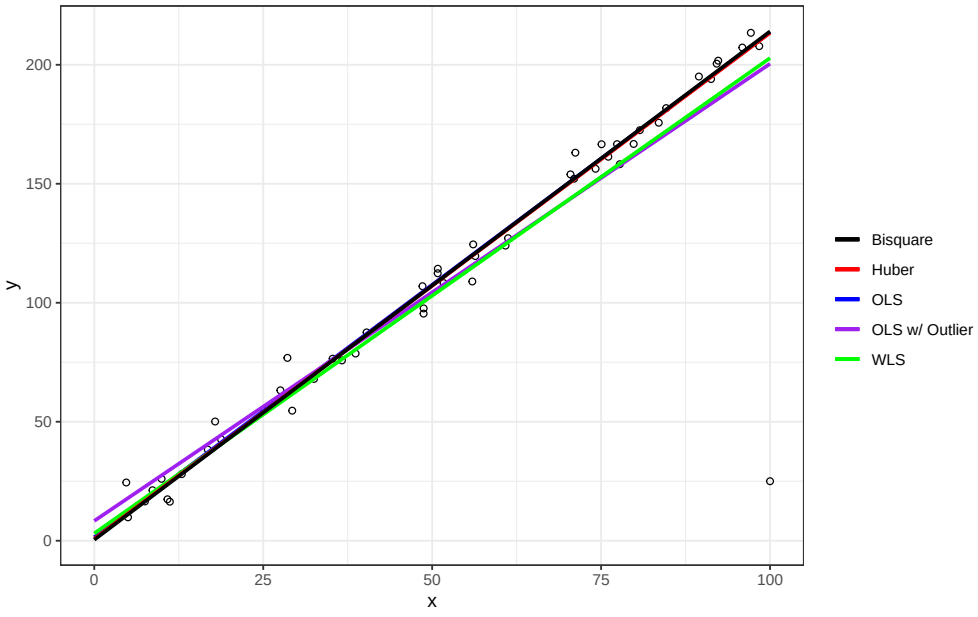
i. Fit a robust regression using Bisquare-weighted iterated reweighted least squares and write out the sample model below.
Solution: We fit the Tukey Bisquare IRLS sample model as follows in R.

```
> library("MASS")
> mod5<-rlm(y_2~x_2,psi=psi.bisquare)
> summary(mod5)
Call: rlm(formula = y_2 ~ x_2, psi = psi.bisquare)
Residuals:
       Min         1Q     Median         3Q        Max
-189.03084   -3.06115    0.09745    3.35943   15.37392

Coefficients:
            Value   Std. Error t value
(Intercept)  0.4097  1.6045     0.2553
x_2          2.1362  0.0266    80.4142

Residual standard error: 4.876 on 49 degrees of freedom
```

We write the sample model out as

$$\widehat{y_i} = 0.4097 + 2.1362 + e_i.$$

ii. Plot the data and regression lines from Questions 1-4. Add the regression line to the plot in green.
Solution:



iii. Interpret the coefficients of the model; are they what you would expect?
Solution: The expected y for an observation of $x = 0$ is 0.4097. For every unit increase in x, we expect a 2.1362 increase in y on average. Giving the outlying observation less weight than the other observations yields coefficients β_0 and β_1 close to the original OLS estimates. Note the β_0 estimate is further from the true value than when using Huber IRLS.

(f) Continue with the data from Question e.

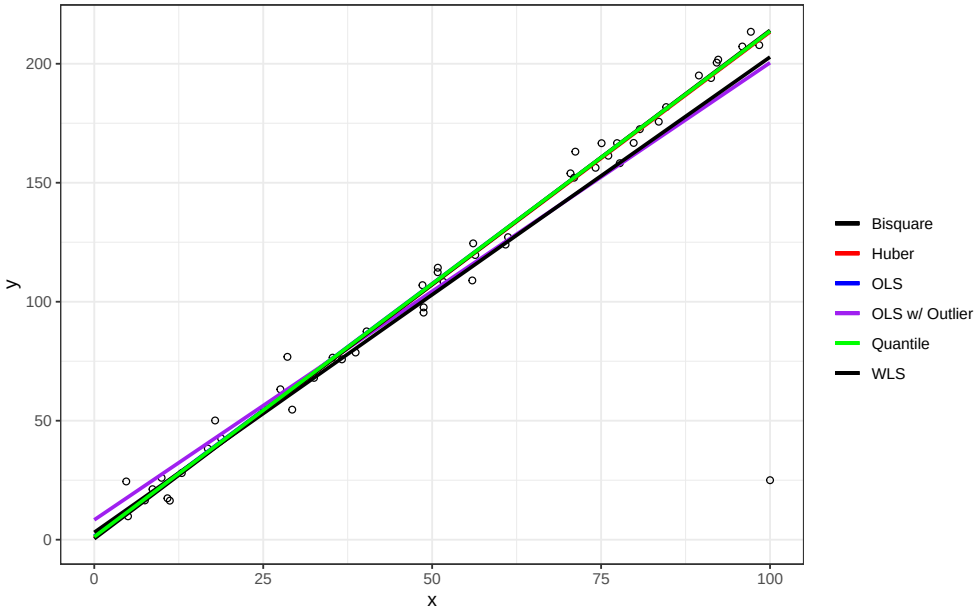
i. Fit a quantile regression and write out the sample model below.
Solution: We calculate the sample model via quantile regression in R as follows.

```
> library(quantreg)
> mod6<-rq(y_2~x_2,tau=0.50) #median quantile regression
> summary(mod6)
Call: rq(formula = y_2 ~ x_2, tau = 0.5)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept)  1.22668     -0.88641  3.26448
x_2          2.12499      2.07830  2.15902
```

We write the sample model out as

$$\widehat{y_i} = 1.22668 + 2.12499 x_i + e_i.$$

ii. Plot the data and regression lines from Questions a-e. Add the regression line to the plot in black.
Solution:



iii. Interpret the coefficients of the model; are they what you would expect?
Solution: The expected y for an observation of $x = 0$ is 1.22668 For every unit increase in x, we expect a 2.12499 increase in the median y. By modelling the median, instead of the mean, our approach is robust to outliers. This flexibility yeilds a estimated coefficient β_1 close to the original OLS estimate and β_0 to be the closest intercept estimate yet.

(g) Reflect on Questions a-f.

i. Which model is "right"? If there's no one model that is "right", which one is "best"? You might consider the AIC, BIC, Log Likelihood, cross validation etc.

Solution: There are many ways to evaluate model fit as listed above. In general, I prefer to use cross validation as it evaluates how well our model predicts new data while methods like AIC, BIC, and Log Likelihoods tell us how well we fit the data we already have. In general, if we're building a predictive model we should consider cross validation and if we're simply fitting a model to describe data we can settle for Log Likelihoods, BIC or AIC measurements (in order of my preference).

We see that the log likelihood, AIC and BIC are similar for all of the models, noting the that OLS model has the highest loglikelihood, lowest BIC and lowest AIC. Furthermore, when we complete cross validation without the outlier, the OLS model has the lowest averaged squared error indicating that it predicts better than the robust models. However, when we include the outlier the IRLS models soundly trounce the OLS and WLS models. Note that quantile regression has an averaged squared error close to that of the IRLS models.

Remark: It should be stated that quantile regression yielded coefficient estimates closer to the true model. This is important to note because of how we interepret the coefficents to provide meaning to our results in the context of a research question.

```
> c(logLik(mod2),logLik(mod3),logLik(mod4),logLik(mod5),logLik(mod6))
[1] -238.6342 -217.2693 -240.3547 -240.5202 -191.3860
> c(BIC(mod2),BIC(mod3),BIC(mod4),BIC(mod5),BIC(mod6))
[1] 489.0638 446.3341 492.5049 492.8359       NA
> c(AIC(mod2),AIC(mod3),AIC(mod4),AIC(mod5),AIC(mod6))
[1] 483.2683 440.5387 486.7094 487.0404 386.7721
```

According to the AIC and log likelihood (BIC is not available for quantile regression) the quantile regression model fits the data best followed by weighted least squares. This addresses model fit.

We can also look at the cross validation error. In doing this, we build the model with subsets of the data and see how close our predictions are for the rest of the data. To do this we split the data into ten subsets and iteratively test our model's predictions for each subset when the regression model is built using the other nine.

```
> ############################################################
> #####Without Outlier
> ############################################################
> #install.packages("caret",repos = "http://cloud.r-project.org/")
> library("caret")
> cverr2<-0;cverr3<-0;cverr4<-0;cverr5<-0;cverr6<-0;
> dat<-data.frame(x_1=x_1,y_1=y_1)
> folds<-createFolds(y=y_1,k=10)
> for(i in 1:10){
+   training<-dat[-folds[[i]],]
+   testing<-dat[folds[[i]],]
+   mod2<-lm(y_1~x_1,data=training)
+   cverr2=cverr2+sum((testing$y_1-predict(mod2,testing))^2)
+   mod_weights<-lm(residuals(mod2)^2~x_1,data=training)
+   weights<-1/(fitted(mod_weights)^2)
+   mod3<-lm(y_1~x_1,weights=weights,data=training)
+   cverr3=cverr3+sum((testing$y_1-predict(mod3,testing))^2)
+   mod4<-rlm(y_1~x_1,psi=psi.huber,data=training)
+   cverr4=cverr4+sum((testing$y_1-predict(mod4,testing))^2)
+   mod5<-rlm(y_1~x_1,psi=psi.bisquare,data=training)
+   cverr5=cverr5+sum((testing$y_1-predict(mod5,testing))^2)
```

```
+   mod6<-rq(y_1~x_1,tau=0.50,data=training) #median quantile regression
+   cverr6=cverr6+sum((testing$y_1-predict(mod6,testing))^2)
+ }
> c(cverr2,cverr3,cverr4,cverr5,cverr6)/nrow(dat)
[1] 32.29460 32.68578 32.40129 32.49706 31.99998
> ############################################################
> #####With the Outlier
> ############################################################
> cverr2<-0;cverr3<-0;cverr4<-0;cverr5<-0;cverr6<-0;
> dat<-data.frame(x_2=x_2,y_2=y_2)
> folds<-createFolds(y=y_2,k=10)
> for(i in 1:10){
+   training<-dat[-folds[[i]],]
+   testing<-dat[folds[[i]],]
+   mod2<-lm(y_2~x_2,data=training)
+   cverr2=cverr2+sum((testing$y_2-predict(mod2,testing))^2)
+   mod_weights<-lm(residuals(mod2)^2~x_2,data=training)
+   weights<-1/(fitted(mod_weights)^2)
+   mod3<-lm(y_2~x_2,weights=weights,data=training)
+   cverr3=cverr3+sum((testing$y_2-predict(mod3,testing))^2)
+   mod4<-rlm(y_2~x_2,psi=psi.huber,data=training)
+   cverr4=cverr4+sum((testing$y_2-predict(mod4,testing))^2)
+   mod5<-rlm(y_2~x_2,psi=psi.bisquare,data=training)
+   cverr5=cverr5+sum((testing$y_2-predict(mod5,testing))^2)
+   mod6<-rq(y_2~x_2,tau=0.50,data=training) #median quantile regression
+   cverr6=cverr6+sum((testing$y_2-predict(mod6,testing))^2)
+ }
> c(cverr2,cverr3,cverr4,cverr5,cverr6)/nrow(dat)
[1] 785.3693 784.1045 739.0540 739.4771 742.1814
```

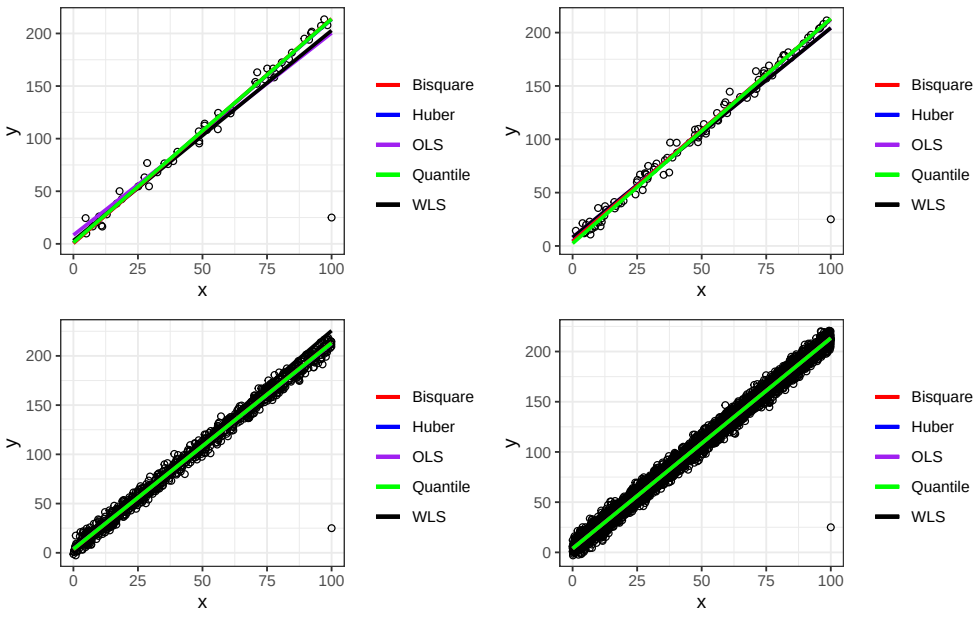
This addresses the predictive ability of the model. Interestingly, they all do about the same when there is no outlier. When there is an outlier, we see that the IRWLS models perform best, followed by quantile regression. These models soundly trounce regular OLS and the WLS models.

ii. Rerun your code for Questions a-f, but change the original sample size from 50 to 1000. There's no need to redo all of the parts from those questions, but discuss the difference. Look at the graph of the data with all the fitted regression models and compare it to that from the original data.

Solution: We see that as the sample size increases, the effect of the outlier fades; when the sample size approaches 1000 and 5000 the regression estimates practically overlap.

```
> par(mfrow=c(2,2))
> sampsize<-c(50,100,1000,5000)
> plots<-list()
> for(i in 1:4){
+   set.seed(50)
+   x_1<-sample(x=seq(0,100,0.01),size=sampsize[i],replace=TRUE)
+   e_1<-rnorm(n=sampsize[i],mean=0,sd=5)
+   y_1<-3.5+2.1*x_1 + e_1
+   x_2<-c(x_1,100)
+   y_2<-c(y_1,25)
+   mod2<-lm(y_2~x_2)
+   mod_weights<-lm(residuals(mod2)^2~x_2)
+   weights<-1/(fitted(mod_weights)^2)
+   mod3<-lm(y_2~x_2,weights=weights)
```

```
+   mod4<-rlm(y_2~x_2,psi=psi.huber)
+   mod5<-rlm(y_2~x_2,psi=psi.bisquare)
+   mod6<-rq(y_2~x_2,tau=0.50) #median quantile regression
+
+   ggdat<-data.frame(x=x_2,
+                     y=y_2)
+   scatterplot<-ggplot(data=ggdat,aes(x=x,y=y))+
+     geom_point(shape=1)+
+     theme_bw()
+
+   new.data<-data.frame(x_2=seq(0,100,1))
+   ggdat<-data.frame(x=new.data$x_2,
+                     y.2=predict(object=mod2, newdata=new.data),
+                     y.3=predict(object=mod3, newdata=new.data),
+                     y.4=predict(object=mod4, newdata=new.data),
+                     y.5=predict(object=mod5, newdata=new.data),
+                     y.6=predict(object=mod6, newdata=new.data))
+
+   scatterplot<-scatterplot+
+     geom_line(data=ggdat,aes(x=x,y=y.2,color="OLS"),size=1)+
+     geom_line(data=ggdat,aes(x=x,y=y.3,color="WLS"),size=1)+
+     geom_line(data=ggdat,aes(x=x,y=y.4,color="Huber"),size=1)+
+     geom_line(data=ggdat,aes(x=x,y=y.5,color="Bisquare"),size=1)+
+     geom_line(data=ggdat,aes(x=x,y=y.6,color="Quantile"),size=1)+
+     scale_color_manual("",values=c("red","blue","purple","green","black"))
+
+   plots[[i]]<-scatterplot
+ }
> grid.arrange(plots[[1]],plots[[2]],plots[[3]],plots[[4]],ncol=2)
```



5. **Case Study** The MASS package in `R` (Venables and Ripley, 2002) provides data about housing values in the Suburbs of Boston. The data provided is described below.

   - **crim** – per capita crime rate by town.
   - **zn** – proportion of residential land zoned for lots over 25,000 sq.ft.
   - **indus** – proportion of non-retail business acres per town.
   - **chas** – Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
   - **nox** – nitrogen oxides concentration (parts per 10 million).
   - **rm** – average number of rooms per dwelling.
   - **age** – proportion of owner-occupied units built prior to 1940.
   - **dis** – weighted mean of distances to five Boston employment centres.
   - **rad** – index of accessibility to radial highways.
   - **tax** – full-value property-tax rate per $10,000.
   - **ptratio** – pupil-teacher ratio by town.
   - **black** – $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.
   - **lstat** – lower status of the population (percent).
   - **medv** – median value of owner-occupied homes in $1000s.

   You can load this data using

```
> library(MASS)
> data(Boston)
```

   Use your tools to build a regression model that predicts the median value of owner-occupied homes in $1000s based on the other variables in the data set.
   Solution: There are many variables that correlate with the median value of homes in the dataset. The two variables with the highest correlation include the number of rooms and the percent of population that has lower status indicating that areas that contain houses with more rooms have higher median home values and areas that contain houses in areas with a higher percentage of lower status people in the population have lower median home values.

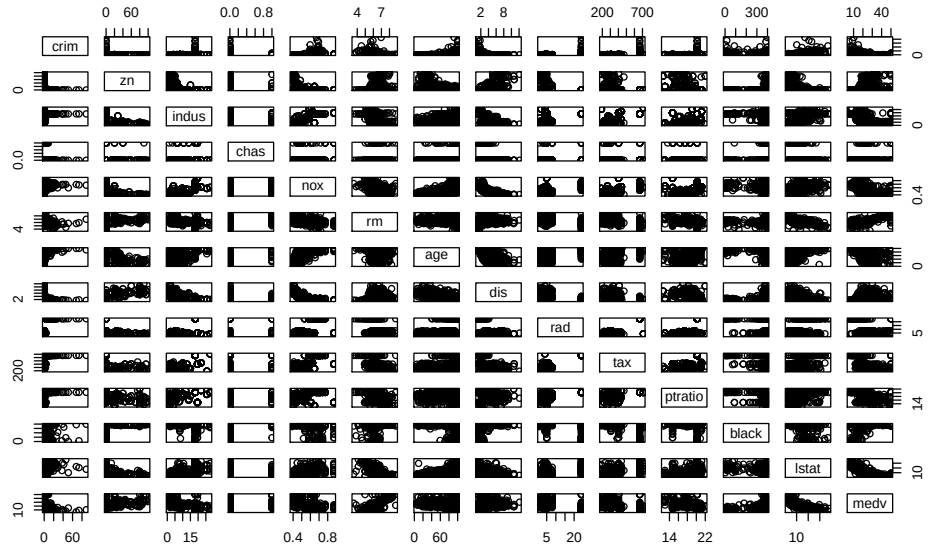


Figure 2: A larger, screen-sized plot of this is easier to read and can be produced using `pairs()`.

```
> round(cor(Boston),2)

         crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
crim     1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29
zn      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39
indus    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38
chas    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12
nox      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19
rm      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36
age      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26
dis     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23
rad      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46
tax      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46
ptratio  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00
black   -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44   -0.18
lstat    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37
medv    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51
        black lstat  medv
crim    -0.39  0.46 -0.39
zn       0.18 -0.41  0.36
indus   -0.36  0.60 -0.48
chas     0.05 -0.05  0.18
nox     -0.38  0.59 -0.43
rm       0.13 -0.61  0.70
age     -0.27  0.60 -0.38
dis      0.29 -0.50  0.25
rad     -0.44  0.49 -0.38
tax     -0.44  0.54 -0.47
ptratio -0.18  0.37 -0.51
black    1.00 -0.37  0.33
lstat   -0.37  1.00 -0.74
medv     0.33 -0.74  1.00
```

The full model, with 91 terms, yields a residual plot that appears to have a balanced line around zero besides a few seemingly outlying points. The sample size of 506 might yield an OLS model that isn't largely affected by the outlier. Later, we will assess the model and compare cross validation error to a robust model.

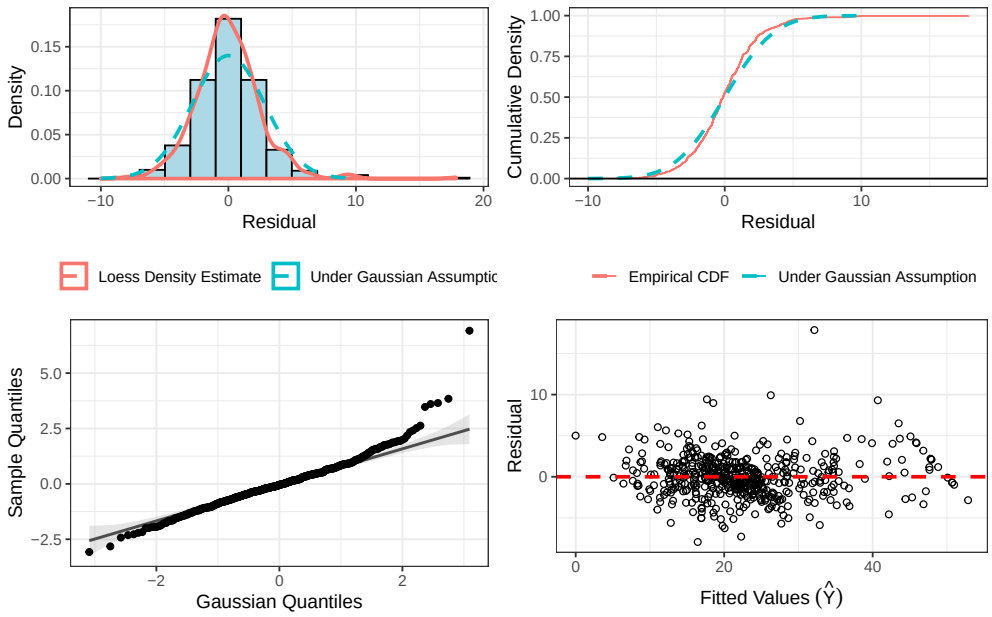
The full model suggests that many of the interactions are significant. For example, per capita crime rate has a negative unique effect on median home values, and the Black-crime interaction shows that the effect of per capita crime rate by town is more negative when a higher proportion of the population is Black. Additionally, the number of rooms has a positive unique effect indicating that areas that contain houses with more rooms have higher median home values but the increase is lower in areas where the proportion of owner-occupied units built prior to 1940 is higher. This indicates that the bigger, older homes aren't valued as highly as bigger, newer homes.

```
> mod.boston<-lm(medv~.^2,data=Boston)
> summary(mod.boston)

Call:
lm(formula = medv ~ .^2, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9374 -1.5344 -0.1068  1.2973 17.8500
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.579e+02  6.800e+01  -2.323 0.020683 *
crim          -1.707e+01  6.554e+00  -2.605 0.009526 **
zn            -7.529e-02  4.580e-01  -0.164 0.869508
indus         -2.819e+00  1.696e+00  -1.663 0.097111 .
chas           4.451e+01  1.952e+01   2.280 0.023123 *
nox            2.006e+01  7.516e+01   0.267 0.789717
rm             2.527e+01  5.699e+00   4.435 1.18e-05 ***
age            1.263e+00  2.728e-01   4.630 4.90e-06 ***
dis           -1.698e+00  4.604e+00  -0.369 0.712395
rad            1.861e+00  2.464e+00   0.755 0.450532
tax            3.670e-02  1.440e-01   0.255 0.798978
ptratio        2.725e+00  2.850e+00   0.956 0.339567
black          9.942e-02  7.468e-02   1.331 0.183833
lstat          1.656e+00  8.533e-01   1.940 0.053032 .
crim:zn        4.144e-01  1.804e-01   2.297 0.022128 *
crim:indus    -4.693e-02  4.480e-01  -0.105 0.916621
crim:chas      2.428e+00  5.710e-01   4.251 2.63e-05 ***
crim:nox      -1.108e+00  9.285e-01  -1.193 0.233425
crim:rm        2.163e-01  4.907e-02   4.409 1.33e-05 ***
crim:age      -3.083e-03  3.781e-03  -0.815 0.415315
crim:dis      -1.903e-01  1.060e-01  -1.795 0.073307 .
crim:rad      -6.584e-01  5.815e-01  -1.132 0.258198
crim:tax       3.479e-02  4.287e-02   0.812 0.417453
crim:ptratio   4.915e-01  3.328e-01   1.477 0.140476
crim:black    -4.612e-04  1.793e-04  -2.572 0.010451 *
crim:lstat     2.964e-02  6.544e-03   4.530 7.72e-06 ***
zn:indus      -6.731e-04  4.651e-03  -0.145 0.885000
zn:chas       -5.230e-02  6.450e-02  -0.811 0.417900
zn:nox         1.998e-03  4.721e-01   0.004 0.996625
zn:rm         -7.286e-04  2.602e-02  -0.028 0.977672
zn:age        -1.249e-06  8.514e-04  -0.001 0.998830
zn:dis         1.097e-02  7.550e-03   1.452 0.147121
zn:rad        -3.200e-03  6.975e-03  -0.459 0.646591
zn:tax         3.937e-04  1.783e-04   2.209 0.027744 *
zn:ptratio    -4.578e-03  7.015e-03  -0.653 0.514325
zn:black       1.159e-04  7.599e-04   0.153 0.878841
zn:lstat      -1.064e-02  4.662e-03  -2.281 0.023040 *
indus:chas    -3.672e-01  3.780e-01  -0.971 0.331881
indus:nox      3.138e+00  1.449e+00   2.166 0.030855 *
indus:rm       3.301e-01  1.327e-01   2.488 0.013257 *
indus:age     -4.865e-04  3.659e-03  -0.133 0.894284
indus:dis     -4.486e-02  6.312e-02  -0.711 0.477645
indus:rad     -2.089e-02  5.020e-02  -0.416 0.677560
indus:tax      3.129e-04  6.034e-04   0.519 0.604322
indus:ptratio -6.011e-02  3.783e-02  -1.589 0.112820
indus:black    1.122e-03  2.034e-03   0.552 0.581464
indus:lstat    5.063e-03  1.523e-02   0.332 0.739789
chas:nox      -3.272e+01  1.243e+01  -2.631 0.008820 **
chas:rm       -5.384e+00  1.150e+00  -4.681 3.87e-06 ***
chas:age       3.040e-02  5.840e-02   0.521 0.602982
chas:dis       9.022e-01  1.334e+00   0.676 0.499143
chas:rad      -7.773e-01  5.707e-01  -1.362 0.173907
```

```
chas:tax       4.627e-02  3.645e-02   1.270 0.204930
chas:ptratio  -6.145e-01  6.914e-01  -0.889 0.374604
chas:black     2.500e-02  1.567e-02   1.595 0.111423
chas:lstat    -2.980e-01  1.845e-01  -1.615 0.107008
nox:rm         5.990e+00  5.468e+00   1.095 0.273952
nox:age       -7.273e-01  2.340e-01  -3.108 0.002012 **
nox:dis        5.694e+00  3.723e+00   1.529 0.126969
nox:rad       -1.994e-01  1.897e+00  -0.105 0.916360
nox:tax       -2.793e-02  1.312e-01  -0.213 0.831559
nox:ptratio   -3.669e+00  3.096e+00  -1.185 0.236648
nox:black     -1.854e-02  3.615e-02  -0.513 0.608298
nox:lstat      1.119e+00  6.511e-01   1.719 0.086304 .
rm:age        -6.277e-02  2.203e-02  -2.849 0.004606 **
rm:dis         3.190e-01  3.295e-01   0.968 0.333516
rm:rad        -8.422e-02  1.527e-01  -0.552 0.581565
rm:tax        -2.242e-02  9.910e-03  -2.262 0.024216 *
rm:ptratio    -4.880e-01  2.172e-01  -2.247 0.025189 *
rm:black      -4.528e-03  3.351e-03  -1.351 0.177386
rm:lstat      -2.968e-01  4.316e-02  -6.878 2.24e-11 ***
age:dis       -1.678e-02  8.882e-03  -1.889 0.059589 .
age:rad        1.442e-02  4.212e-03   3.423 0.000682 ***
age:tax       -3.403e-04  2.187e-04  -1.556 0.120437
age:ptratio   -7.520e-03  6.793e-03  -1.107 0.268946
age:black     -7.029e-04  2.136e-04  -3.291 0.001083 **
age:lstat     -6.023e-03  1.936e-03  -3.111 0.001991 **
dis:rad       -5.580e-02  7.075e-02  -0.789 0.430678
dis:tax       -3.882e-03  2.496e-03  -1.555 0.120623
dis:ptratio   -4.786e-02  9.983e-02  -0.479 0.631920
dis:black     -5.194e-03  5.541e-03  -0.937 0.349116
dis:lstat      1.350e-01  4.866e-02   2.775 0.005774 **
rad:tax        3.131e-05  1.446e-03   0.022 0.982729
rad:ptratio   -4.379e-02  8.392e-02  -0.522 0.602121
rad:black     -4.362e-04  2.518e-03  -0.173 0.862561
rad:lstat     -2.529e-02  1.816e-02  -1.392 0.164530
tax:ptratio    7.854e-03  2.504e-03   3.137 0.001830 **
tax:black     -4.785e-07  1.999e-04  -0.002 0.998091
tax:lstat     -1.403e-03  1.208e-03  -1.162 0.245940
ptratio:black  1.203e-03  3.361e-03   0.358 0.720508
ptratio:lstat  3.901e-03  2.985e-02   0.131 0.896068
black:lstat   -6.118e-04  4.157e-04  -1.472 0.141837
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.852 on 414 degrees of freedom
Multiple R-squared:  0.9212,        Adjusted R-squared:  0.9039
F-statistic: 53.18 on 91 and 414 DF,  p-value: < 2.2e-16
```

Remark: I use a lot of the stuff we talked about the last day of class. If you didn't use the same approach, that is most reasonable. This just helps you see what we were talking about and how helpful it can be.



Due to the number of predictors in the model, and the fact that we're considering all two-way interactions, best subsets regression would require fitting an infeasible amount of models; i.e.,

$$2^{\binom{13}{2}+13} = 2^{91} = 2.47588 \times 10^{27} \text{ models.}$$

Thus, due to the number of predictors (91), we continue with stepwise model selection via BIC to yield a parsimonious. This model selection was conducted in `R` using the "MASS" package (Venables and Ripley, 2002). A likelihood ratio test conducted using the "lmtest" package (Zeileis and Hothorn, 2002) suggests the full model fits significantly better than the model produced via stepwise model selection ($\chi^2_{52} = 70.56$, $p = 0.04425$), but there are less than half the number of original predictors and the significance is marginal so we'll continue with the less complicated model.

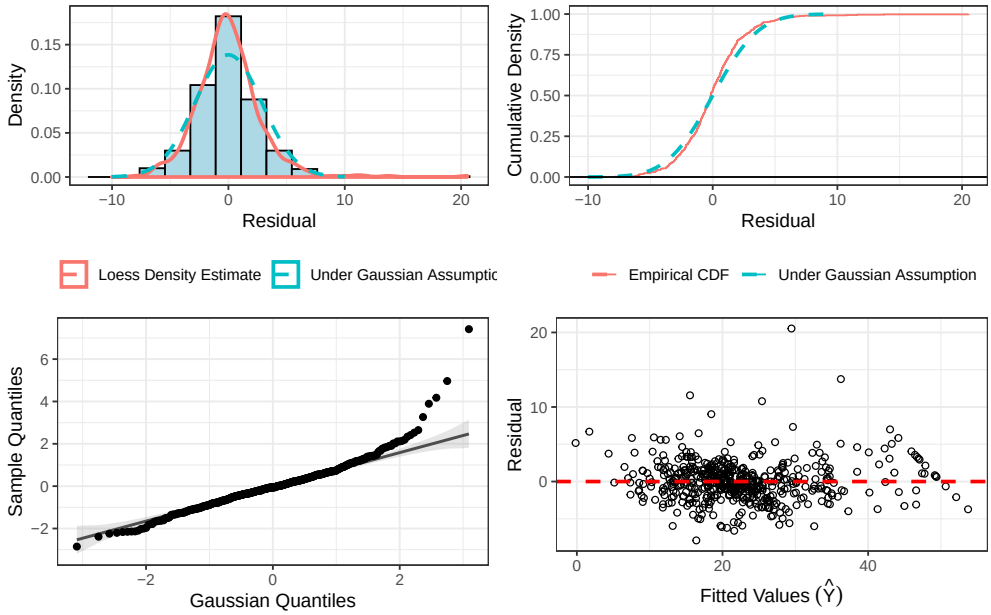
```
> library(MASS)
> mod.boston_stepBIC<-stepAIC(mod.boston,k=log(nrow(Boston)),trace = FALSE)
> summary(mod.boston_stepBIC)

Call:
lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
    dis + rad + tax + ptratio + black + lstat + crim:chas + crim:rm +
    crim:rad + crim:black + crim:lstat + zn:tax + zn:lstat +
    indus:nox + indus:rm + chas:nox + chas:rm + chas:black +
    chas:lstat + nox:age + nox:ptratio + rm:age + rm:tax + rm:ptratio +
    rm:lstat + age:dis + age:rad + age:black + age:lstat + dis:tax +
    tax:ptratio + tax:lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8998 -1.6010 -0.1567  1.4164 20.5308

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.186e+02  2.074e+01 -10.537  < 2e-16 ***
crim        -4.465e-01  6.065e-01  -0.736 0.461902
zn          -9.601e-02  3.605e-02  -2.663 0.008004 **
```

```
indus       -3.562e+00  5.226e-01  -6.815 2.93e-11 ***
chas         3.949e+01  9.726e+00   4.060 5.75e-05 ***
nox          1.173e+02  1.873e+01   6.265 8.51e-10 ***
rm           3.073e+01  2.058e+00  14.931  < 2e-16 ***
age          1.082e+00  1.336e-01   8.099 4.89e-15 ***
dis          1.452e+00  4.326e-01   3.357 0.000853 ***
rad         -2.720e-01  1.368e-01  -1.988 0.047417 *
tax          1.108e-01  3.123e-02   3.548 0.000428 ***
ptratio      4.095e+00  1.078e+00   3.799 0.000165 ***
black        6.656e-02  1.310e-02   5.082 5.43e-07 ***
lstat        2.967e+00  2.617e-01  11.337  < 2e-16 ***
crim:chas    1.999e+00  2.906e-01   6.879 1.95e-11 ***
crim:rm      2.362e-01  4.076e-02   5.796 1.25e-08 ***
crim:rad    -6.525e-02  2.276e-02  -2.867 0.004338 **
crim:black  -3.296e-04  1.321e-04  -2.495 0.012949 *
crim:lstat   2.407e-02  4.250e-03   5.664 2.60e-08 ***
zn:tax       4.689e-04  1.276e-04   3.675 0.000265 ***
zn:lstat    -6.308e-03  2.405e-03  -2.623 0.009001 **
indus:nox    2.483e+00  4.597e-01   5.402 1.05e-07 ***
indus:rm     3.750e-01  6.690e-02   5.606 3.56e-08 ***
chas:nox    -2.362e+01  5.854e+00  -4.035 6.38e-05 ***
chas:rm     -5.419e+00  9.877e-01  -5.486 6.75e-08 ***
chas:black   3.047e-02  1.164e-02   2.617 0.009169 **
chas:lstat  -4.314e-01  1.245e-01  -3.464 0.000581 ***
nox:age     -9.392e-01  1.509e-01  -6.224 1.08e-09 ***
nox:ptratio -5.290e+00  8.033e-01  -6.585 1.23e-10 ***
rm:age      -5.310e-02  1.340e-02  -3.964 8.54e-05 ***
rm:tax      -2.858e-02  3.167e-03  -9.024  < 2e-16 ***
rm:ptratio  -5.952e-01  1.189e-01  -5.006 7.89e-07 ***
rm:lstat    -3.067e-01  3.680e-02  -8.333 8.88e-16 ***
age:dis     -1.580e-02  4.340e-03  -3.641 0.000302 ***
age:rad      7.510e-03  1.594e-03   4.711 3.26e-06 ***
age:black   -6.871e-04  1.504e-04  -4.569 6.27e-06 ***
age:lstat   -4.560e-03  1.523e-03  -2.994 0.002897 **
dis:tax     -5.155e-03  1.224e-03  -4.211 3.05e-05 ***
tax:ptratio  5.528e-03  1.285e-03   4.300 2.08e-05 ***
tax:lstat   -2.795e-03  2.376e-04 -11.763  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.882 on 466 degrees of freedom
Multiple R-squared:  0.9094,        Adjusted R-squared:  0.9018
F-statistic: 119.9 on 39 and 466 DF,  p-value: < 2.2e-16
```

```
> #install.packages("lmtest",repos = "http://cloud.r-project.org/")
> library(lmtest)
> lrtest(mod.boston_stepBIC,mod.boston)

Likelihood ratio test

Model 1: medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + black + lstat + crim:chas + crim:rm + crim:rad +
    crim:black + crim:lstat + zn:tax + zn:lstat + indus:nox +
    indus:rm + chas:nox + chas:rm + chas:black + chas:lstat +
    nox:age + nox:ptratio + rm:age + rm:tax + rm:ptratio + rm:lstat +
```

```
    age:dis + age:rad + age:black + age:lstat + dis:tax + tax:ptratio +
    tax:lstat
Model 2: medv ~ (crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + black + lstat)^2
  #Df  LogLik Df Chisq Pr(>Chisq)
1  41 -1232.7
2  93 -1197.5 52 70.56    0.04425 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual and QQplot look very similar to the full model. The constant variance and normality of error terms approximately hold with the exception of a handful of outliers.



Checking the VIF of this model yields very large values. This is due to including several of the interactions in the model, thus these values can be ignored. Checking the correlation between predictors we see that the index of accessibility to radial highways and the proportion of non-retail business acres per town are highly correlated and the proportion of non-retail business acres per town is moderately associated with weighted mean of distances to five Boston employment centres, full-value property-tax rate per $10,000, nitrogen oxides concentration (parts per 10 million), and the proportion of owner-occupied units built prior to 1940. Despite the high correlations, we see that a model including all the unique effects and no interactions shows only slight to moderate multicollinearity when checked using the "car" package for R (Fox and Weisberg, 2019).

```
> library("car")
> vif(mod.boston_stepBIC)
```

```
       crim          zn       indus        chas         nox          rm
1654.523013   42.973557  781.609074  371.028637  286.443801  127.171973
        age         dis         rad         tax     ptratio       black
 860.422000   50.451933   86.312407 1684.217134  331.125253   86.954799
      lstat    crim:chas     crim:rm    crim:rad  crim:black  crim:lstat
 212.299012    3.286106  267.664846 1364.576909    7.524232   42.394951
     zn:tax    zn:lstat   indus:nox    indus:rm   chas:nox     chas:rm
  51.489529    8.396810  327.957033  458.084872   50.271903  165.729024
```

```
 chas:black  chas:lstat     nox:age nox:ptratio      rm:age      rm:tax
  75.563969   10.528302  655.857589  269.118369  334.369075  615.418665
 rm:ptratio    rm:lstat     age:dis     age:rad   age:black   age:lstat
 196.930601  122.176032   10.708584  112.288566  185.906327   90.207550
    dis:tax tax:ptratio   tax:lstat
  35.480893 1399.398738   88.707203
```

```
> vifmod<-lm(medv~.,data=Boston)
> vif(vifmod)
```

```
    crim       zn    indus     chas      nox       rm      age      dis
1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945
     rad      tax  ptratio    black    lstat
7.484496 9.008554 1.799084 1.348521 2.941491
```

A likelihood ratio test suggests the removing the full-value property-tax rate per $10,000 due to multicollinearity results in worse model fit ($\chi^2_2 = 43.531$, $p = 3.526 \times 10^{-10}$). While robust regression can be considered here, we continue with an OLS model noting that larger sample sizes erode the benefit in using the biased approach.

```
> Bostonsub<-Boston[,-10]
> mod.boston.sub<-lm(medv~.^2,data=Bostonsub)
> mod.boston_stepBIC2<-stepAIC(mod.boston.sub,k=log(nrow(Bostonsub)),trace = FALSE)
> summary(mod.boston_stepBIC2)
```

```
Call:
lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
    dis + rad + ptratio + black + lstat + crim:zn + crim:chas +
    crim:nox + crim:rm + crim:lstat + zn:dis + zn:lstat + indus:nox +
    indus:rm + indus:rad + chas:nox + chas:rm + chas:ptratio +
    nox:age + rm:age + rm:rad + rm:ptratio + rm:lstat + age:rad +
    age:ptratio + age:black + age:lstat + dis:rad + dis:lstat +
    rad:lstat, data = Bostonsub)

Residuals:
   Min     1Q Median     3Q    Max
-7.481 -1.720 -0.185  1.507 19.953

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.565e+02  1.798e+01  -8.702  < 2e-16 ***
crim         -5.164e-01  4.202e-01  -1.229 0.219722
zn           -4.844e-02  3.237e-02  -1.496 0.135259
indus        -2.231e+00  4.877e-01  -4.574 6.14e-06 ***
chas          7.434e+01  1.119e+01   6.645 8.45e-11 ***
nox          -9.357e+00  1.124e+01  -0.832 0.405568
rm            2.375e+01  2.345e+00  10.127  < 2e-16 ***
age           1.147e+00  1.523e-01   7.535 2.54e-13 ***
dis          -1.256e+00  3.628e-01  -3.461 0.000587 ***
rad           3.701e+00  4.555e-01   8.125 4.01e-15 ***
ptratio       3.565e+00  8.768e-01   4.066 5.62e-05 ***
black         7.920e-02  1.409e-02   5.622 3.25e-08 ***
lstat         2.274e+00  2.800e-01   8.123 4.07e-15 ***
crim:zn       3.343e-01  1.053e-01   3.176 0.001593 **
crim:chas     2.884e+00  3.771e-01   7.650 1.16e-13 ***
```

```
crim:nox     -2.427e+00  6.158e-01  -3.940 9.38e-05 ***
crim:rm       2.173e-01  3.969e-02   5.476 7.12e-08 ***
crim:lstat    3.276e-02  5.089e-03   6.437 3.02e-10 ***
zn:dis        1.337e-02  4.171e-03   3.206 0.001438 **
zn:lstat     -7.872e-03  2.574e-03  -3.058 0.002354 **
indus:nox     2.504e+00  5.315e-01   4.711 3.26e-06 ***
indus:rm      1.668e-01  5.496e-02   3.035 0.002537 **
indus:rad    -3.152e-02  9.949e-03  -3.168 0.001635 **
chas:nox     -4.438e+01  6.201e+00  -7.156 3.22e-12 ***
chas:rm      -4.379e+00  7.658e-01  -5.717 1.93e-08 ***
chas:ptratio -1.368e+00  3.688e-01  -3.709 0.000233 ***
nox:age      -4.330e-01  1.356e-01  -3.194 0.001499 **
rm:age       -5.512e-02  1.439e-02  -3.829 0.000146 ***
rm:rad       -3.853e-01  5.311e-02  -7.254 1.69e-12 ***
rm:ptratio   -4.822e-01  1.255e-01  -3.841 0.000140 ***
rm:lstat     -3.360e-01  3.792e-02  -8.860  < 2e-16 ***
age:rad       5.564e-03  1.999e-03   2.783 0.005604 **
age:ptratio  -1.582e-02  3.258e-03  -4.856 1.63e-06 ***
age:black    -8.678e-04  1.576e-04  -5.507 6.03e-08 ***
age:lstat    -5.228e-03  1.567e-03  -3.336 0.000918 ***
dis:rad      -1.433e-01  3.409e-02  -4.202 3.17e-05 ***
dis:lstat     6.360e-02  2.328e-02   2.733 0.006522 **
rad:lstat    -5.033e-02  5.109e-03  -9.851  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.002 on 468 degrees of freedom
Multiple R-squared:  0.9013,        Adjusted R-squared:  0.8934
F-statistic: 115.4 on 37 and 468 DF,  p-value: < 2.2e-16

> #install.packages("lmtest",repos = "http://cloud.r-project.org/")
> library(lmtest)
> lrtest(mod.boston_stepBIC2,mod.boston_stepBIC)

Likelihood ratio test

Model 1: medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
    ptratio + black + lstat + crim:zn + crim:chas + crim:nox +
    crim:rm + crim:lstat + zn:dis + zn:lstat + indus:nox + indus:rm +
    indus:rad + chas:nox + chas:rm + chas:ptratio + nox:age +
    rm:age + rm:rad + rm:ptratio + rm:lstat + age:rad + age:ptratio +
    age:black + age:lstat + dis:rad + dis:lstat + rad:lstat
Model 2: medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + black + lstat + crim:chas + crim:rm + crim:rad +
    crim:black + crim:lstat + zn:tax + zn:lstat + indus:nox +
    indus:rm + chas:nox + chas:rm + chas:black + chas:lstat +
    nox:age + nox:ptratio + rm:age + rm:tax + rm:ptratio + rm:lstat +
    age:dis + age:rad + age:black + age:lstat + dis:tax + tax:ptratio +
    tax:lstat
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  39 -1254.5
2  41 -1232.7  2 43.531  3.526e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There are 3 outlying observations in the OLS model yielded from stepwise model selection according to BIC.

```
> length(which(abs(rstandard(mod.boston_stepBIC))>4))

[1] 3
```

Due to the outliers, we fit a robust regression with Tukey's bisquare weighting in R using the "MASS" package Venables and Ripley (2002). We see that many of the predictors have similar coefficient estimates – the median absolute difference is 0.054795. The exception is the coefficient for nitrogen oxides concentration which increased by 27.720431.

```
> library(MASS)
> formula<-as.formula(medv ~ crim + zn + indus + chas + nox + rm + age +
+     dis + rad + tax + ptratio + black + lstat + crim:chas + crim:rm +
+     crim:rad + crim:black + crim:lstat + zn:tax + zn:lstat +
+     indus:nox + indus:rm + chas:nox + chas:rm + chas:black +
+     chas:lstat + nox:age + nox:ptratio + rm:age + rm:tax + rm:ptratio +
+     rm:lstat + age:dis + age:rad + age:black + age:lstat + dis:tax +
+     tax:ptratio + tax:lstat)
> mod8robust<-rlm(formula,data=Boston, psi=psi.bisquare)
> summary(mod8robust)

Call: rlm(formula = formula, data = Boston, psi = psi.bisquare)
Residuals:
      Min        1Q    Median        3Q       Max
-16.35805  -1.29085  -0.03483   1.51553  25.39717

Coefficients:
            Value     Std. Error t value
(Intercept) -211.3723   17.0563   -12.3926
crim          -0.2509    0.4986    -0.5031
zn            -0.0284    0.0296    -0.9579
indus         -3.1669    0.4297    -7.3698
chas          36.3254    7.9966     4.5426
nox           89.6186   15.4009     5.8190
rm            30.9962    1.6924    18.3149
age            0.9421    0.1099     8.5737
dis            0.2143    0.3557     0.6026
rad           -0.1587    0.1125    -1.4102
tax            0.1039    0.0257     4.0454
ptratio        4.4938    0.8863     5.0704
black          0.0690    0.0108     6.4096
lstat          2.6081    0.2151    12.1223
crim:chas      3.3108    0.2389    13.8566
crim:rm        0.1943    0.0335     5.7964
crim:rad      -0.0539    0.0187    -2.8811
crim:black    -0.0003    0.0001    -2.8845
crim:lstat     0.0142    0.0035     4.0610
zn:tax         0.0003    0.0001     2.4853
zn:lstat      -0.0067    0.0020    -3.3917
indus:nox      2.1489    0.3780     5.6851
indus:rm       0.3438    0.0550     6.2509
chas:nox     -26.6408    4.8131    -5.5351
chas:rm       -5.3011    0.8121    -6.5275
```

```
chas:black     0.0384    0.0096     4.0139
chas:lstat    -0.3492    0.1024    -3.4098
nox:age       -0.7496    0.1241    -6.0417
nox:ptratio   -3.9470    0.6605    -5.9759
rm:age        -0.0460    0.0110    -4.1750
rm:tax        -0.0264    0.0026   -10.1255
rm:ptratio    -0.6756    0.0978    -6.9101
rm:lstat      -0.2949    0.0303    -9.7465
age:dis       -0.0106    0.0036    -2.9718
age:rad        0.0053    0.0013     4.0666
age:black     -0.0007    0.0001    -5.7094
age:lstat     -0.0052    0.0013    -4.1475
dis:tax       -0.0014    0.0010    -1.3736
tax:ptratio    0.0039    0.0011     3.6890
tax:lstat     -0.0018    0.0002    -9.2289

Residual standard error: 1.979 on 466 degrees of freedom

> summary(abs(coef(mod.boston_stepBIC)-coef(mod8robust)))

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.000016  0.003447  0.054795  1.197074  0.340533 27.720431

> coef(mod.boston_stepBIC)[6]-coef(mod8robust)[6]

     nox
27.72043
```

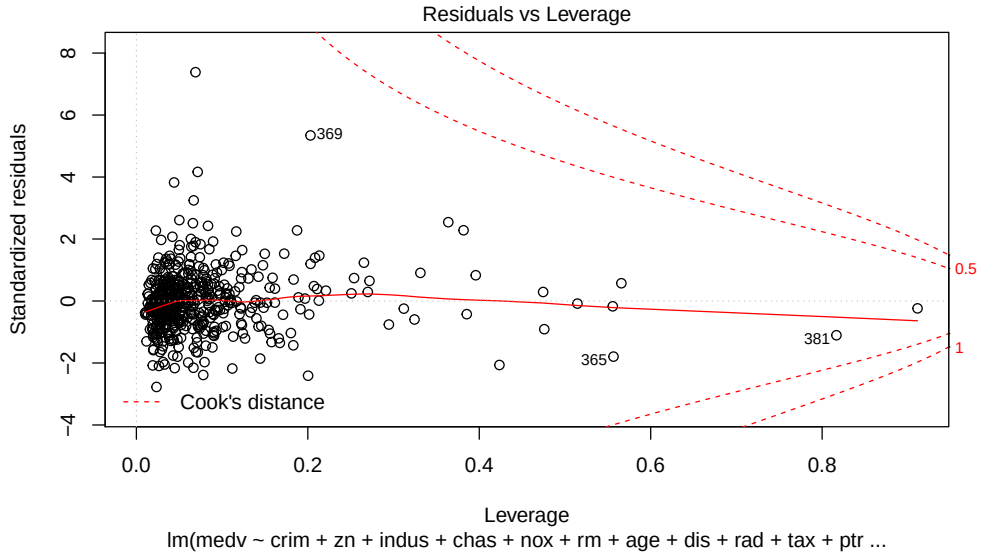
Comparing the robust and the stepwise BIC models shows that the original BIC model fits better by every metric – AIC, BIC, Logliklihood and cross validation. This is in part due to the sample size. The folds for cross validation were randomly generated using the "caret" package for R (from Jed Wing et al., 2019).

```
> c(logLik(mod.boston),logLik(mod.boston_stepBIC),logLik(mod8robust))

[1] -1197.447 -1232.727 -1271.198

> c(BIC(mod.boston),BIC(mod.boston_stepBIC),BIC(mod8robust))

[1] 2973.961 2720.742 2797.684

> c(AIC(mod.boston),AIC(mod.boston_stepBIC),AIC(mod8robust))

[1] 2580.893 2547.454 2624.396

> library("caret")
> library("MASS")
> cverr8<-0;cverr8BIC<-0;cverr8robust<-0;
> folds<-createFolds(y=Boston$medv,k=10)
> for(i in 1:10){
+   training<-Boston[-folds[[i]],]
+   testing<-Boston[folds[[i]],]
+   mod8cv<-lm(medv~.^2,data=training)
+   cverr8=cverr8+sum((testing$medv-predict(mod8cv,testing))^2)
+   mod8BICcv<-lm(formula,data=training)
+   cverr8BIC=cverr8BIC+sum((testing$medv-predict(mod8BICcv,testing))^2)
+   mod8robustcv<-rlm(formula,data=training)
```

```
+   cverr8robust=cverr8robust+sum((testing$medv-predict(mod8robustcv,testing))^2)
+ }
> c(cverr8,cverr8BIC,cverr8robust)/nrow(Boston)

[1] 11.98669  9.61275 10.04880
```

We can also visualize the effect of outliers on our linear model by consulting the leverage plot from R. This plot shows our points and their leverage statistics. Points beyond the red Cook's distance lines are taken to be "leveraging" points. Think back to questions 1-7 where we added an outlier – this point, when the sample size was low, "leveraged" or "tilted" the OLS line.

```
> plot(mod.boston_stepBIC,which=5)
```



To simplify the overall interpretation of the model we employ the "margins" package in R (Leeper, 2018) to calculate the marginal effects of each variable over the interactions.

```
> library("margins")
> margins(mod.boston_stepBIC)

   crim     zn  indus  chas    nox    rm      age    dis      rad      tax
 0.7404 0.0156 0.1727 4.964 -18.66 5.212 -0.06336 -1.736 0.007207 -0.01643
 ptratio   black   lstat
 -0.3237 0.02036 -0.4294
```

The largest positive marginal effect is the average number of rooms per dwelling; for each additional room we expect an increase of $5212 in median home value. This is closely followed by whether or not the area is bounded by the Charles River; we expect the median home value to be $4964 higher in areas bounded by the Charles River than those that don't on average.

The largest negative marginal effect is the nitrogen oxides concentration (parts per 10 million); for parts per 10 million increase in nitrogen oxides concentration we expect an $18660 decrease in median home value. This seems unexpected and it is misleading because of the unit of measurement leads to us never observing a unit increase. A summary of nitrogen oxides concentration is provided below which motivates a z score transformation.

```
> summary(Boston$nox)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3850  0.4490  0.5380  0.5547  0.6240  0.8710

> Boston$nox<-(Boston$nox-mean(Boston$nox))/sd(Boston$nox)
> mod.boston_stepBIC_z<-lm(formula,data=Boston)
> margins(mod.boston_stepBIC_z)

   crim     zn  indus  chas    nox    rm      age    dis      rad      tax
 0.7404 0.0156 0.1727 4.964 -2.163 5.212 -0.06336 -1.736 0.007207 -0.01643
 ptratio   black   lstat
 -0.3237 0.02036 -0.4294
```

Still, nitrogen oxides concentration (parts per 10 million) has the largest negative effect indicating that for a standard deviation increase in nitrogen oxides concentration we expect the median home value to decrease by $2163, on average. This is closely followed by the weighted mean of distances to five Boston employment centres; areas one unit further from these employment centers cost $1736 less on average.

Interestingly, crime has a positive effect on median home values, however when we consider the full output of the model this appears to be driven by the whether or not the area is bound by the Charles River, the amount of low-status people and the number of rooms in a home. A reasonable explanation for this is that regardless of crime, or the amount of low-status people in the population people are willing to pay for larger residences or residences in prime locations; e.g., near the Charles River.

This message seems to confirm the "location, location, location" mantra of real estate agents. While the size of the house certainly matters, closeness to the Charles River has about the same effect. Also, the distance from employment centers matters – people appear to want homes closer to where they work, likely to decrease commuting time. The nitrogen oxides concentration is harder to tie to location. Harrison Jr and Rubinfeld (1978) suggests that much of U.S. nitrous oxide emissions come from agricultural soil managment, but very little agriculture occurs in the city of Boston. It is more likely that areas with higher nitrous oxide concentration are industrial or transportation heavy. It is possible that this effect is driven by areas that are less desirable because they're mainly industrial – note the correlation between these variables from the start.

References

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.

Bracht, T., Molleken, C., Ahrens, M., Poschmann, G., Schlosser, A., Eisenacher, M., Stuhler, K., Meyer, H. E., Schmiegel, W. H., Holmskov, U., Sorensen, G. L., and Sitek, B. (2016). Evaluation of the biomarker candidate mfap4 for non-invasive assessment of hepatic fibrosis in hepatitis C patients. *Journal of Translational Medicine*, 14.

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt., T. (2019). *caret: Classification and Regression Training*. R package version 6.0-84.

Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102.

Leeper, T. J. (2018). *margins: Marginal Effects for Model Objects*. R package version 0.3.23.

Lyon, J. D. and Tsai, C.-L. (1996). A comparison of tests for heteroscedasticity. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45(3):337–349.

Muhling, B. A., Lamkin, J. T., and Richards, W. J. (2012). Decadal-scale responses of larval fish assemblages to multiple ecosystem processes in the northern gulf of mexico. *Marine Ecology Progress Series*, 450:37–53.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7–10.