HW 4 | Shawn Xiang Li

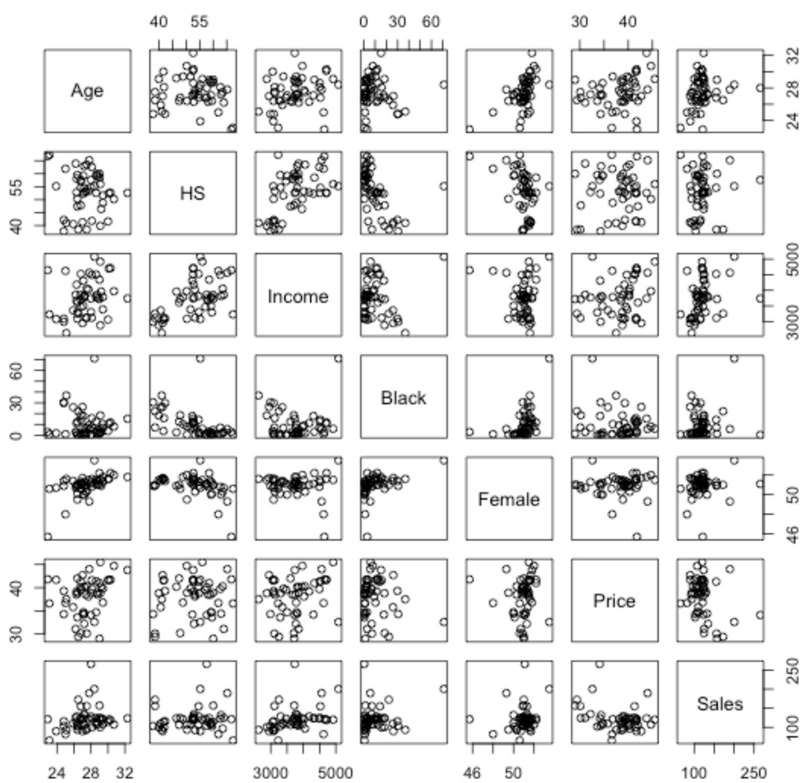
Problem 1 (4.7)

(a) **Price** has a negative relationship with Sales since demand decreases while price increases.
Age has a positive relationship with Sales since older people more like to buy cigarette.
Female has a negative relationship with Sales since females are less likely to smoke
Income has a positive relationship with Sales since there is more money to buy cigarette.
HS has a positive relationship with Sales since higher educated people are more likely to smoke
Black has a positive relationship with Sales since more black people in low-income level so that they are more likely to smoke

(b) The pairwise correlation coefficients matrix:

```
> cor(cigratte_sale)
               Age          HS      Income       Black      Female       Price       Sales
Age     1.00000000 -0.09891626  0.25658098 -0.04033021  0.55303189  0.24775673  0.22655492
HS     -0.09891626  1.00000000  0.53400534 -0.50171191 -0.41737794  0.05697473  0.06669476
Income  0.25658098  0.53400534  1.00000000  0.01728756 -0.06882666  0.21455717  0.32606789
Black  -0.04033021 -0.50171191  0.01728756  1.00000000  0.45089974 -0.14777619  0.18959037
Female  0.55303189 -0.41737794 -0.06882666  0.45089974  1.00000000  0.02247351  0.14622124
Price   0.24775673  0.05697473  0.21455717 -0.14777619  0.02247351  1.00000000 -0.30062263
Sales   0.22655492  0.06669476  0.32606789  0.18959037  0.14622124 -0.30062263  1.00000000
```

The corresponding scatter plot matrix:



HW 4 | Shawn Xiang Li

(c) There is a disagreement between pairwise correlation coefficients and the corresponding scatter plot matrix. The relationship between sales and other variables match from two results. However, from the scatter plot matrix, I can see there is no much linear relationship among Sales vs. Income, Sales vs. Black, Sales vs. Female and Sales vs. HS; because the sales are almost constant with these variables and there are many outliers in these plots.

(d) YES. I assume the relationship between Sales and Female would be negative since I think Females are less likely to smoke so that less likely to purchase cigarettes. But correlation matrix and scatter plot shows there is a positive relationship between these two variables.

(e) From the regression model, the coefficient of HS is negative, which is different from my expectation in (a). The coefficient of Female is negative and the coefficient of HS is negative. They are inconsistent with results from pairwise correlation matrix.

```
Call:
lm(formula = Sales ~ Age + Black + Female + HS + Income + Price)

Residuals:
   Min      1Q  Median      3Q     Max
-48.398 -12.388  -5.367   6.270 133.213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.34485  245.60719   0.421  0.67597
Age           4.52045    3.21977   1.404  0.16735
Black         0.35754    0.48722   0.734  0.46695
Female       -1.05286    5.56101  -0.189  0.85071
HS           -0.06159    0.81468  -0.076  0.94008
Income        0.01895    0.01022   1.855  0.07036 .
Price        -3.25492    1.03141  -3.156  0.00289 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.17 on 44 degrees of freedom
Multiple R-squared:  0.3208,	Adjusted R-squared:  0.2282
F-statistic: 3.464 on 6 and 44 DF,  p-value: 0.006857
```

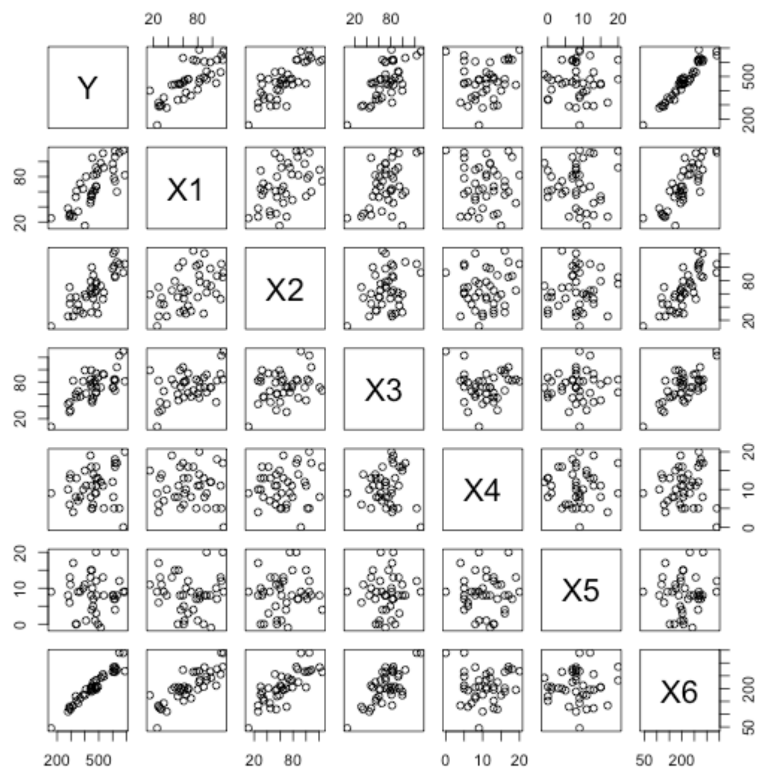
(f) As I mentioned in (c), there are no much linear relationship between Sales and other variables, except Price. Therefore, Age, Black, Female, HS, Income are not good to use in a linear regression to predict Sales. We can prove it by looking at their corresponding P-values: P-value of them are larger than 0.05, which are insignificant in the model. Especially for HS and Female, their P-value is very high, which means they are highly insignificant and needed to remove from the model, so their estimated coefficients, which are inconsistent with results from (b), do not mean anything.

(g) No. In the test I made from 3.15, the Female and HS were still insignificant and excluded from the model.

Problem 2 (4.12 a, b)

(a) There are 3 assumptions of least squares regression:

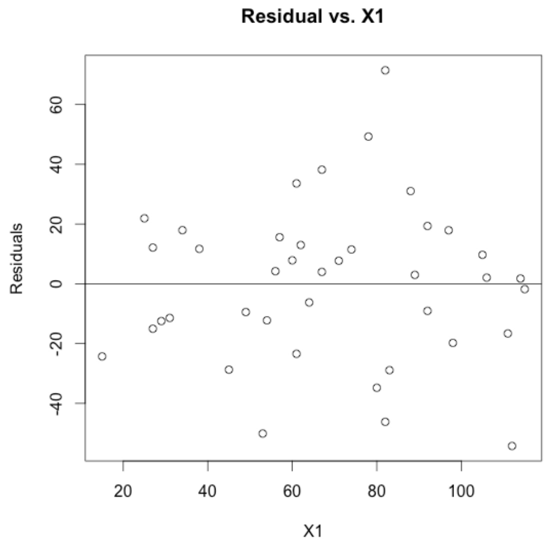
1. Model is good (the relationship is linear and not quadratic, exponential or others).

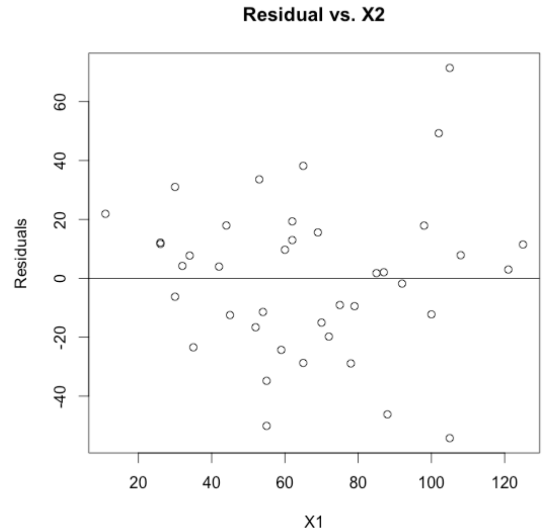


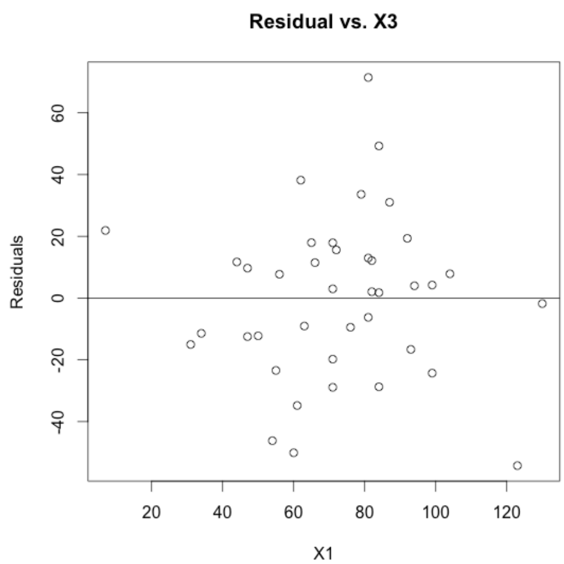
Look at the first row of scatter plot matrix; we can see the relationship between Y and other predictive variables. Except X5, there are linear trends between Y and X's variables.

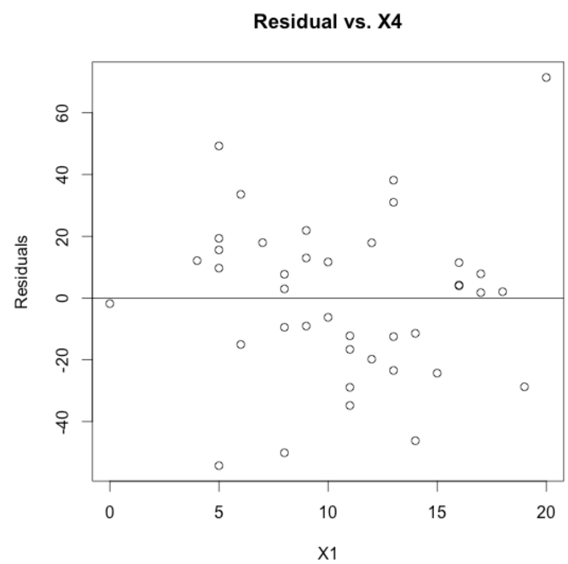
Then I plot residual and each X's to further explore Y and X's relationship.

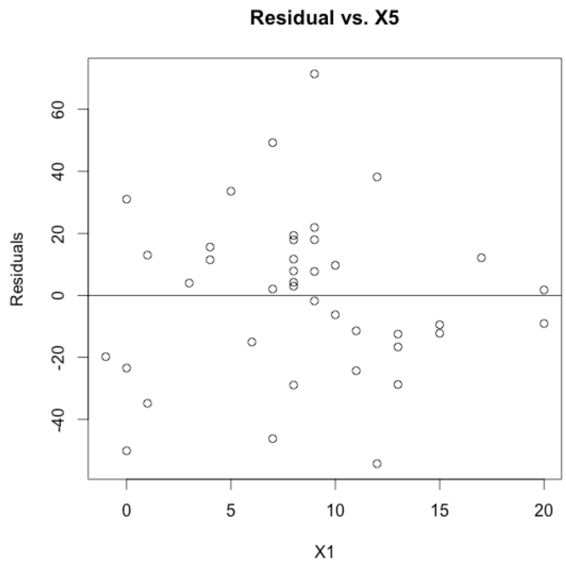
HW 4 | Shawn Xiang Li

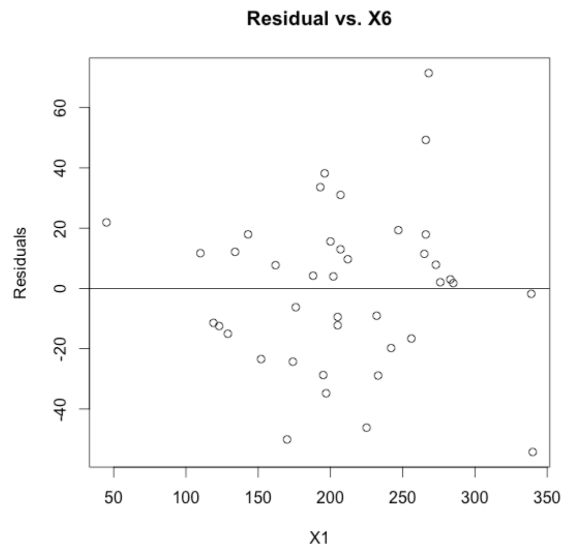






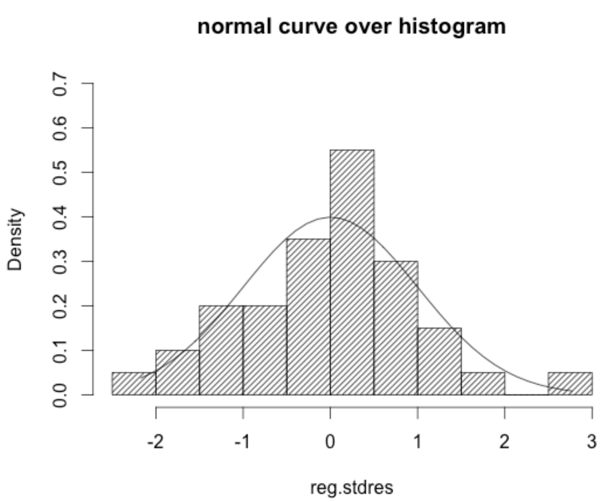


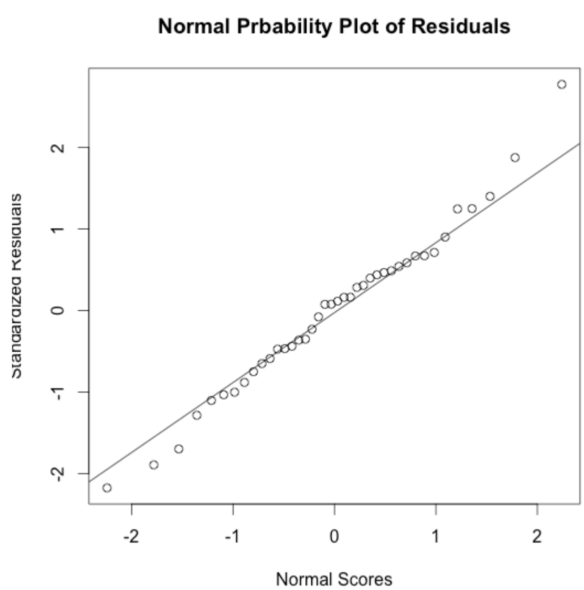




From all Residuals vs. X's plot, we can see there is no any pattern. Data are randomly distributed above and below X-axis. The result shows all data satisfy the linear model.

2. The residual does not have a normal distribution.



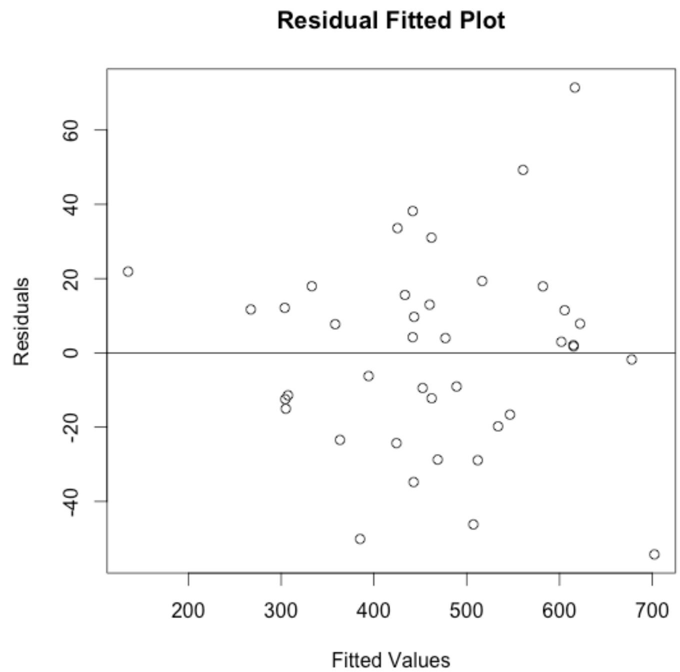


For the assumption, we draw the Normal Probability plot and see whether the dots form a straight line. QQ-plot is not good since there are some outliers. And, histogram tells the residuals are NOT normal distributed.

3. Residuals have equal variance

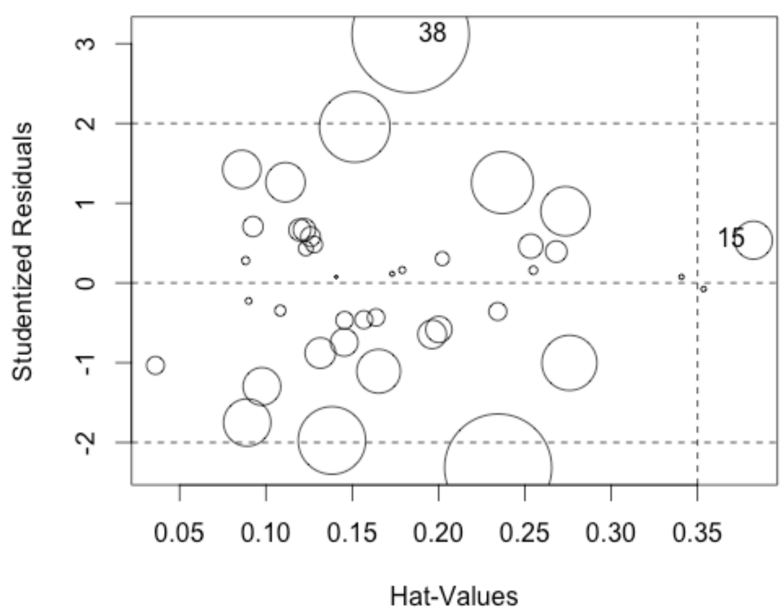
Then, we look at each Residual vs. Fitted plot and check any patterns.

The distribution is pretty random above and below X-axis; there is not trend between fitted value and residual. Therefore, the Var(residual) is constant.



(d) **Checking outliers & leverages:**
We can see **obs. 38** is an outlier and **obs. 15** is leverage

```
> outlierTest(reg)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
38 3.119525          0.0038192      0.15277
> influencePlot(reg)
     StudRes       Hat     CookD
15 0.5357756 0.3823626 0.1610822
38 3.1195250 0.1837930 0.4975412
```



Problem 3 (4.13)

HW 4 | Shawn Xiang Li

First, I ran a model 1 relating Y and first 3 variables. All X's variable in the model 1 are highly significant since their corresponding P-value are very small and less than 0.05. Adjusted R^2 is 93.59%; it tells us model 1 fits Y's data very well.

```
Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
    Min      1Q  Median      3Q     Max 
-73.919 -15.681  -4.493  22.570  99.903 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  61.9253    18.1589   3.410  0.00162 ** 
X1            1.6365     0.2208   7.413 9.50e-09 ***
X2            2.1769     0.2028  10.734 9.05e-13 ***
X3            2.0173     0.2398   8.411 5.10e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.63 on 36 degrees of freedom
Multiple R-squared:  0.9408,	Adjusted R-squared:  0.9359 
F-statistic: 190.7 on 3 and 36 DF,  p-value: < 2.2e-16
```

(a) Added X4 and create Model 2. From the summary, we can see X4 is significant since its P-value= 0.00197 is smaller than 0.05. So, **X4 is indeed added into model**. Moreover, Adjusted R^2= 95% increases after added X4 into the model; it tells us X4 improves the model to fit Y's data better.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
   Min     1Q Median     3Q    Max 
-55.05 -17.03   2.83  17.08  72.40 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.3469    18.9141   1.499  0.14291    
X1            1.7006     0.1958   8.684 2.97e-10 ***
X2            2.0907     0.1809  11.558 1.68e-13 ***
X3            2.0209     0.2117   9.544 2.83e-11 ***
X4            3.2295     0.9654   3.345  0.00197 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.92 on 35 degrees of freedom
Multiple R-squared:  0.9551,	Adjusted R-squared:   0.95 
F-statistic: 186.3 on 4 and 35 DF,  p-value: < 2.2e-16
```

(b) Added X5 create Model 3. From the summary, we can see X4 is insignificant since its P-value= 0.45992 is larger than 0.05. So, **X5 should not be added into model**. Moreover, Adjusted R^2= 94.94% decreases after added X5 into the model; it tells us X5 worse the model to fit Y's data.

HW 4 | Shawn Xiang Li

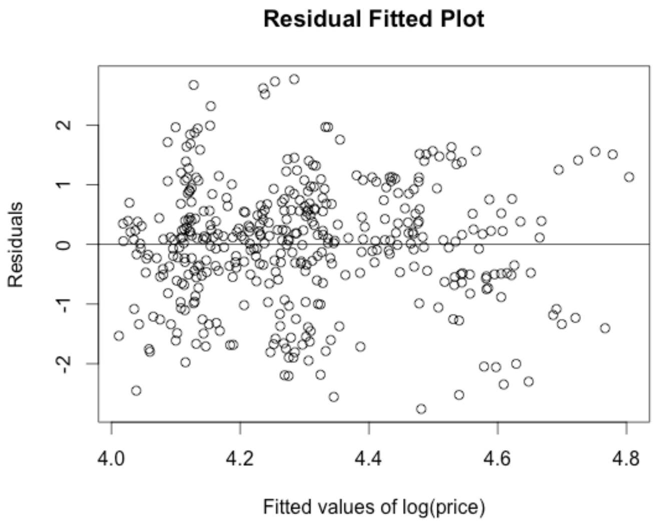
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.236 -14.888   2.002  14.359  72.406 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  33.3688    20.1854   1.653  0.10751    
X1            1.6863     0.1980   8.516 6.01e-10 ***
X2            2.1077     0.1835  11.489 2.97e-13 ***
X3            2.0286     0.2133   9.509 4.17e-11 ***
X4            3.2118     0.9718   3.305  0.00225 ** 
X5           -0.6575     0.8796  -0.747  0.45992    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.1 on 34 degrees of freedom
Multiple R-squared:  0.9559,	Adjusted R-squared:  0.9494 
F-statistic: 147.3 on 5 and 34 DF,  p-value: < 2.2e-16
```

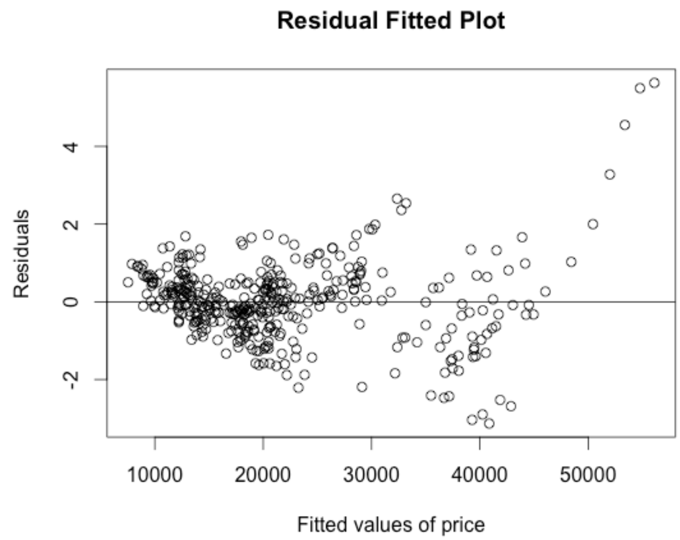
(c) Deleted X5 and Added X6 create Model 4. From the summary, we can see X1, X2 and X3 becomes insignificant after X6 added and X6 itself is insignificant. So, **X6 should not be added into model**.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X6)

Residuals:
   Min     1Q Median     3Q    Max 
-55.92 -16.28   3.27  17.86  71.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  29.7692    20.7200   1.437  0.15993   
X1            2.6612     5.2962   0.502  0.61857   
X2            3.0502     5.2898   0.577  0.56800   
X3            2.9723     5.2466   0.567  0.57476   
X4            3.2100     0.9849   3.259  0.00254 **
X6           -0.9583     5.2798  -0.181  0.85705   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.32 on 34 degrees of freedom
Multiple R-squared:  0.9552,	Adjusted R-squared:  0.9486 
F-statistic: 144.9 on 5 and 34 DF,  p-value: < 2.2e-16
```

(d) I think Model 2, **lm(Y~X1+X2+X3+X4)** is the best possible description of Y. Since:
1. All predictive variables are significant from the model.
2. Adjusted R^2= 95% is highest among these models. It tells Model 2 fit Y's data best.

Problem 4





Left plot is fitted **log(price)** vs. its residuals, the distribution looks very random above and below the X-axis. It shows the residuals' variances are same of the model, which is good.

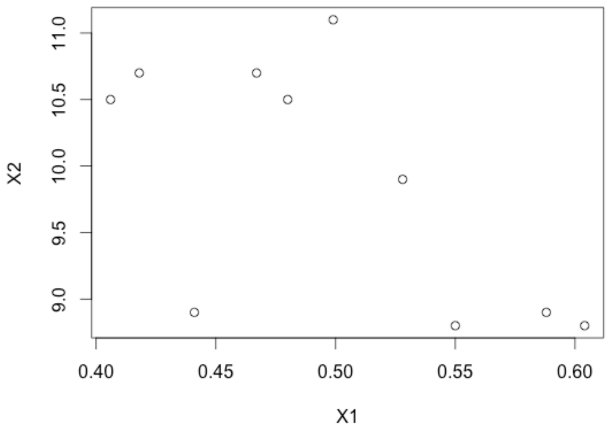
Right plot is fitted **Price** vs. its residuals. The distribution looks like a "U" shape. It tells the variances of residuals are not same. It violated one of the assumptions. The model is not good.

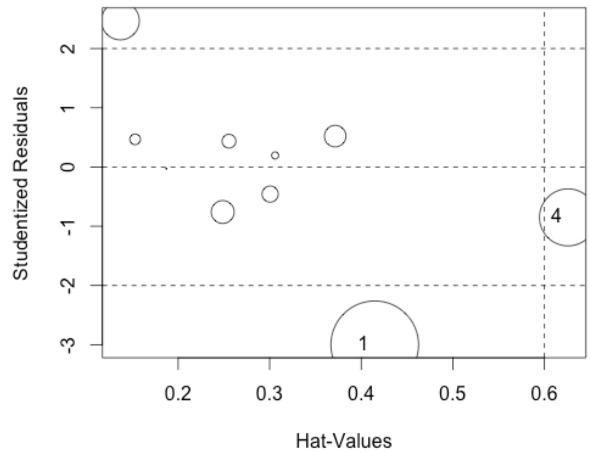
Log-transforming would make a model's residuals with non-constant variances changed to residuals with a constant variance, like price and log(price) here.

HW 4 | Shawn Xiang Li

Problem 5

(a)





From the plot, I can see **obs. 4** where locates at left corner is an outlier since it has both extreme value of x1 and x2

(b) **Obs 4**'s larger than 2(p+1)/n = 0.6; **Obs. 4** is influential.

```
Call: Old Model
lm(formula = Y ~ X1 + X2)

Residuals:
    Min       1Q   Median       3Q      Max
-0.44422 -0.12780  0.05365  0.10521  0.44985

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.3015     1.8965   5.432 0.000975 ***
X1           8.4947     1.7850   4.759 0.002062 **
X2          -0.2663     0.1237  -2.152 0.068394 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2754 on 7 degrees of freedom
Multiple R-squared:    0.9,     Adjusted R-squared:  0.8714
F-statistic:  31.5 on 2 and 7 DF,  p-value: 0.0003163
```

```
Call: New Model
lm(formula = x$Y ~ x$X1 + x$X2)

Residuals:
    Min       1Q   Median       3Q      Max
-0.33339 -0.05037  0.01127  0.05615  0.46579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.4107     2.9071   4.269  0.00527 **
x$X1          6.7992     2.5166   2.702  0.03549 *
x$X2         -0.3905     0.1794  -2.177  0.07237 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.277 on 6 degrees of freedom
Multiple R-squared:  0.9108,	Adjusted R-squared:  0.8811
F-statistic: 30.65 on 2 and 6 DF,  p-value: 0.0007089
```

(c) Left is the result before omitting influential observations. Right is the result after omitting the influential observation, obs. 4.

We can see the fit changes much for the coefficients after omitting influential observations. In the old model, X2 is insignificant with a 0.068394 P-value. In the new model, X2 is still insignificant, but with higher R^2, which shows the model fits data better than old one. Therefore, the new model is much better to predict the wood beam strength.