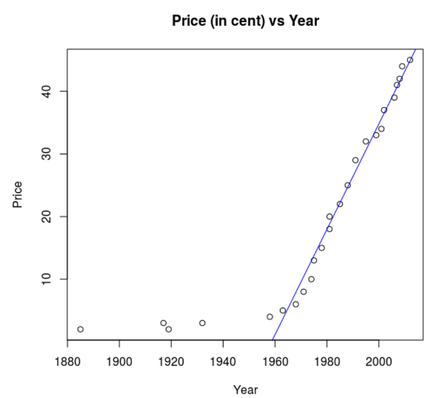
Violet Chang || 5197617 || chan1300

Stat 3022 Homework 2

Problem 1.16

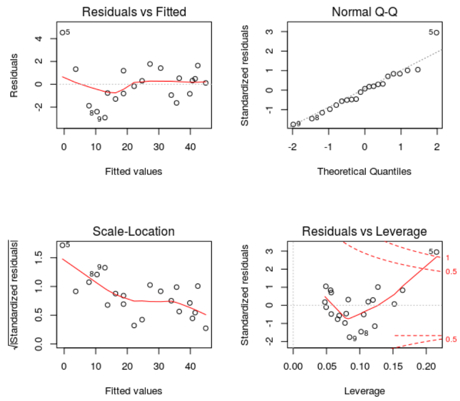
```
> library(Stat2Data)
> data(USstamps)
> help(USstamps)
> head(USstamps)
  Year Price
1 1885     2
2 1917     3
3 1919     2
4 1932     3
5 1958     4
6 1963     5
> plot(Price~Year, data=USstamps, main="Price (in cent) vs Year")
```



1.16 (a): The plot shows the positive linear pattern which indicate that the relationship between price and year of stamps is linearly. However the first four points show the different pattern that indicate they are the noise data of this data set.

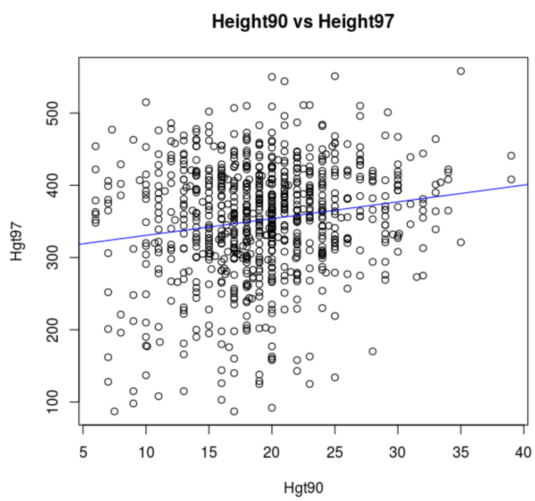
```
> rm4=USstamps[c(-1,-2,-3,-4),]
> lm1=lm(Price~Year, data=rm4)
> abline(lm1,col="blue")
> summary(lm1)

Call:
lm(formula = Price ~ Year, data = rm4)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9232 -0.9478  0.1195  1.1899  4.5325 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.647e+03  4.686e+01  -35.15   <2e-16 ***
Year         8.410e-01  2.357e-02   35.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.737 on 19 degrees of freedom
```

```
Multiple R-squared:  0.9853,    Adjusted R-squared:  0.9845
F-statistic:  1273 on 1 and 19 DF,  p-value: < 2.2e-16
## 1.16 (b): The least squares line is hat(Price) = -1647 + 0.841 * Year.

> par(mfrow=c(2,2))
> plot(lm1)
```



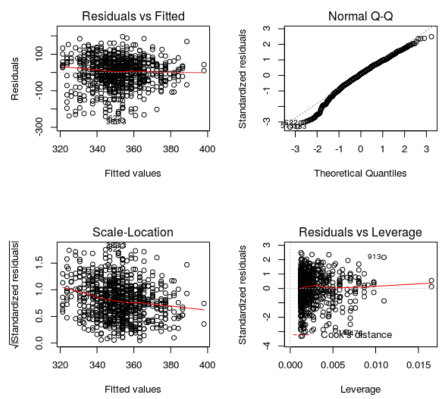

```
## 1.16 (d): The normal Q-Q plot generally shows the linear pattern and the fitted-residuals
plot shows the regular pattern at beginning then shows irregular pattern. So the conditions well
met the regression model.
```

```
## Problem 1.19
> par(mfrow=c(1,1))
> data(Pines)
> help(Pines)
> head(Pines)
  Row Col Hgt90 Hgt96 Diam96 Grow96 Hgt97 Diam97 Spread.97 Needles97 Deer95 Deer97
1   1   1    NA    NA     NA     NA    NA     NA        NA        NA     NA     NA
2   1   2    14   284    4.2     96   362    6.6       162        66      0      1
3   1   3    17   387    7.4    110   442    9.3       250        77      0      0
4   1   4    NA    NA     NA     NA    NA     NA        NA        NA     NA     NA
5   1   5    24   294    3.9     70   369    7.0       176        72      0      0
6   1   6    22   310    5.6     84   365    6.9       215        76      0      0
  Cover95 Fert Spacing
1       0    0      15
2       2    0      15
3       1    0      15
4       0    0      15
5       2    0      15
6       1    0      15
> plot(Hgt97~Hgt90, data=Pines, main="Height90 vs Height97")
```



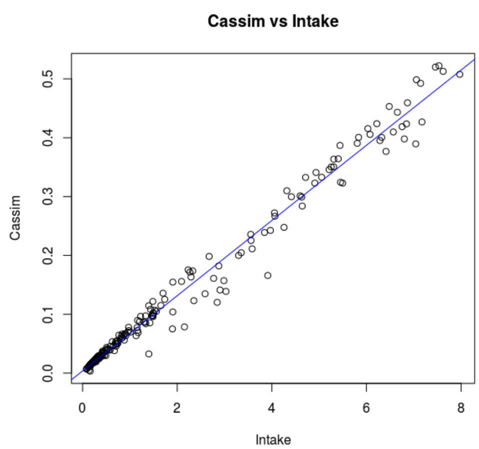

```
## 1.19 (a): The plot shows the positive linear pattern which indicate that the relationship
between height of 1990 and 1997 is linearly.

> lm2=lm(Hgt97~Hgt90, data=Pines)
> abline(lm2, col="blue")
> summary(lm2)

Call:
lm(formula = Hgt97 ~ Hgt90, data = Pines)

Residuals:
     Min       1Q   Median       3Q      Max
-261.886  -44.343    7.308   55.114  196.114

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  307.439      9.841  31.239  < 2e-16 ***
```

```
Hgt90          2.322      0.492   4.721 2.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78.79 on 807 degrees of freedom
  (191 observations deleted due to missingness)
Multiple R-squared:  0.02687,   Adjusted R-squared:  0.02567
F-statistic: 22.28 on 1 and 807 DF,  p-value: 2.772e-06
## 1.19 (b): The least squares line is hat(Hgt97) = 307.439 + 2.322 * Year.

> par(mfrow=c(2,2))
> plot(lm2)
```



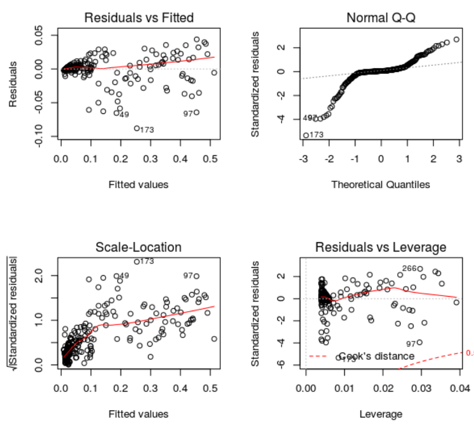

```
## 1.19 (c): The normal Q-Q plot generally shows the linear pattern with a little bit curve.
Generally, the conditions and normality are met and fit the linear model in an acceptable way.
```

```
## Problem 1.21
> par(mfrow=c(1,1))
> data(Caterpillars)
> help(Caterpillars)
> head(Caterpillars)
  Instar ActiveFeeding Fgp Mgp     Mass   LogMass   Intake  LogIntake WetFrass
1      1             Y   Y   Y 0.002064 -2.685290 0.165118 -0.7822056 0.000241
2      1             Y   N   N 0.005191 -2.284749 0.201008 -0.6967867 0.000063
3      2             N   Y   N 0.005603 -2.251579 0.189125 -0.7232511 0.001401
4      2             Y   N   N 0.019300 -1.714443 0.283280 -0.5477841 0.002045
5      2             N   Y   Y 0.029300 -1.533132 0.259569 -0.5857472 0.005377
6      3             Y   Y   N 0.062600 -1.203426 0.327864 -0.4843063 0.029500
  LogWetFrass DryFrass LogDryFrass     Cassim LogCassim   Nfrass LogNfrass      Nassim
1   -3.617983 0.000208   -3.681937 0.01422378 -1.846985 6.61e-06 -5.179510 0.001858999
2   -4.200659 0.000061   -4.214670 0.01739189 -1.759653 1.03e-06 -5.986783 0.002270091
3   -2.853562 0.000969   -3.013676 0.01639923 -1.785177 2.78e-05 -4.555794 0.002302210
4   -2.689307 0.001834   -2.736601 0.02392468 -1.621154 4.64e-05 -4.333480 0.003041352
5   -2.269460 0.003523   -2.453087 0.02122857 -1.673079 9.97e-05 -4.001301 0.002791898
6   -1.530178 0.000789   -3.102923 0.02836365 -1.547238 1.84e-05 -4.735567 0.003627464
  LogNassim
1 -2.730721
2 -2.643957
3 -2.637855
4 -2.516933
5 -2.554100
6 -2.440397
> plot(Cassim~Intake, data=Caterpillars, main="Cassim vs Intake")
```



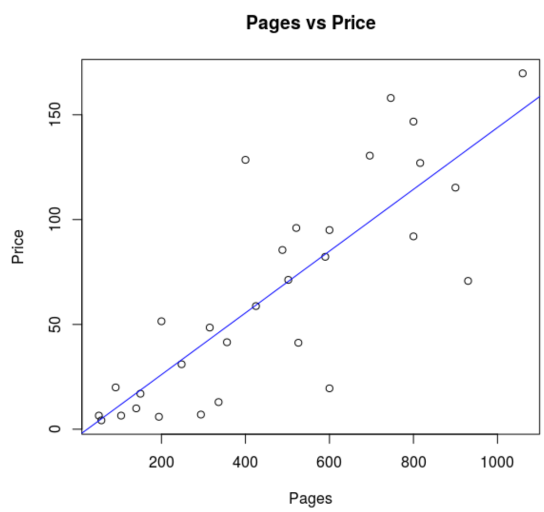

```
## 1.21 (a): The plot shows the positive linear pattern which indicate that the relationship
between Cassim and Intake is linearly.

> lm3=lm(Cassim~Intake, data=Caterpillars)
> abline(lm3,col="blue")
> summary(lm3)

Call:
lm(formula = Cassim ~ Intake, data = Caterpillars)

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

```
-0.087967 -0.000908  0.000927  0.004093  0.043898

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0037867  0.0013171   2.875  0.00438 **
Intake      0.0639029  0.0004908 130.208  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01654 on 252 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.9854,	Adjusted R-squared:  0.9853
F-statistic: 1.695e+04 on 1 and 252 DF,  p-value: < 2.2e-16
## 1.21 (b): The least squares line is hat(Cassim) = 0.00379 + 0.0639 * Intake.

> par(mfrow=c(2,2))
> plot(lm3)
```




```
## 1.21 (c): The conditions for inference not met. The normal Q-Q plot shows an irregular pattern instead of linear and the residuals vs fitted value plot shows that the variances is not consistent. Thus this model is not fit properly for this data set.
```

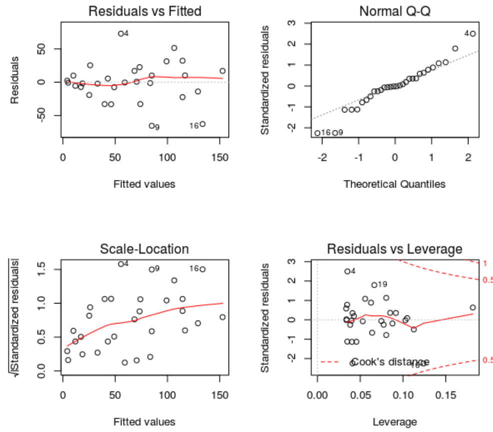
```
## Problem 1.26
> par(mfrow=c(1,1))
> data(TextPrices)
> help(TextPrices)
> head(TextPrices)
  Pages  Price
1   600  95.00
2    91  19.95
3   200  51.50
4   400 128.50
5   521  96.00
6   315  48.50
> plot(Price~Pages, data=TextPrices, main="Pages vs Price")
```




```
## 1.26 (a): The plot shows that when pages get larger, the price is also getting higher, so
there is a potential linear pattern between pages and price.

> lm4=lm(Price~Pages, data=TextPrices)
> abline(lm4,col="blue")
> summary(lm4)

Call:
lm(formula = Price ~ Pages, data = TextPrices)

Residuals:
    Min      1Q  Median      3Q     Max
-65.475 -12.324  -0.584  15.304  72.991

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.42231   10.46374  -0.327    0.746
Pages        0.14733    0.01925   7.653 2.45e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.76 on 28 degrees of freedom
Multiple R-squared:  0.6766,    Adjusted R-squared:  0.665
F-statistic: 58.57 on 1 and 28 DF,  p-value: 2.452e-08
```

```
## 1.26 (b): The least squares line is hat(Price) = -3.4223 + 0.1473 * Pages.
```

```
> par(mfrow=c(2,2))
> plot(lm4)
```




```
## 1.26 (c): The normal Q-Q plot shows a roughly linear pattern with a little bit curve.
Generally, the conditions and normality are met and fit the linear model in an acceptable way.
However, the residuals vs fitted value plot shows that the variability of large predictions is
larger then small predictions. Thus, some conditions are in doubt even though as a whole this is
not a big deal.
```

```
## Problem 2.14
> anova(lm4)
Analysis of Variance Table

Response: Price
           Df Sum Sq Mean Sq F value    Pr(>F)
Pages       1  51877   51877  58.573 2.452e-08 ***
Residuals 28  24799     886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 7.653^2
[1] 58.56841
## 2.14 (a): The hypothesis: H0: beta1 = 0 and H1: beta1 != 0;
From problem 1.26, t value is 7.653 so t² value is 58.57 which is also the value of F
Since p value is 2.452e-08<0.05, we reject the anova (or reject hypothesis 0).

> confint(lm4, level=0.95)
                  2.5 %    97.5 %
(Intercept) -24.8563229 18.011694
Pages         0.1078959  0.186761
##2.14 (b): The true slope of price is a measure of change in price lies between 0.10 and 0.19
with 0.95 confidence.
```

```
## Problem 2.16
> par(mfrow=c(1,1))
> data(Sparrows)
> lm5 = lm(Weight ~ WingLength, data = Sparrows)
> summary(lm5)

Call:
lm(formula = Weight ~ WingLength, data = Sparrows)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5440 -0.9935  0.0809  1.0559  3.4168 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.36549    0.95731   1.426    0.156    
WingLength   0.46740    0.03472  13.463   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.4 on 114 degrees of freedom
Multiple R-squared:  0.6139,	Adjusted R-squared:  0.6105 
F-statistic: 181.3 on 1 and 114 DF,  p-value: < 2.2e-16
## 2.16 (a): The hypothesis is H0: beta1 = 0 and H1: beta1 != 0. From the summary of linear model, the t value is 13.463, p value is 2.2e-16 which is much smaller than 0.05 so H0 is rejected. Thus, the slope of the least squares regression line for predicting Weight from Wing length is different from zero.

> confint(lm5, level=0.95)
                 2.5 %    97.5 %
(Intercept) -0.5309316 3.2619109
WingLength   0.3986288 0.5361792
## 2.16 (b): The 95% confidence interval for slope of regression line is 0.397 and 0.536. In other words, the true slope of price is a measure of change in weight lies between 0.397 and 0.536 with 0.95 confidence.

## 2.16 (c): From part (b), the confidence interval lies between 0.397 and 0.536 which does not contains 0. This supports the hypothesis that the slope of regression is not zero in part (a).
```

```
## Problem 2.24
> data(MathEnrollment)
> help(MathEnrollment)
> head(MathEnrollment)
  Ayear Fall Spring
1  2001  259    246
2  2002  301    206
3  2003  343    288
4  2004  307    215
5  2005  286    230
6  2006  273    247
# Remove the data row of Ayear = 2003.
> mathenroll = data.frame(MathEnrollment)
> newMathenroll = subset(mathenroll, Ayear!=2003)
# Set up the linear model for this data set to predict
> lm6 = lm(Spring ~ Fall, data = newMathenroll)
> summary(lm6)

Call:
lm(formula = Spring ~ Fall, data = newMathenroll)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.500 -17.353  -6.058  22.711  29.418 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 548.0094   106.7236   5.135 0.000891 ***
Fall         -1.0483     0.3805  -2.755 0.024870 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.94 on 8 degrees of freedom
Multiple R-squared:  0.4868,	Adjusted R-squared:  0.4227 
F-statistic: 7.589 on 1 and 8 DF,  p-value: 0.02487
# Create new data frame
> new = data.frame(Fall = 290)
> predict(lm6, newdata = new)
       1 
244.0025 
## 2.24 (a): Using linear model to predict the spring enrollment is 244 for 290 of fall enrollment.

> predict(lm6, newdata = new, interval = "confidence", level = 0.95)
       fit     lwr     upr
1 244.0025 223.693 264.312
## 2.24 (b): With 95% confidence, the mean of spring enrollment lies in the interval between 223.693 and 264.312, when fall enrollment is 290

> predict(lm6, newdata = new, interval = "predict", level = 0.95)
       fit      lwr      upr
1 244.0025 183.0076 304.9974
## 2.24 (c): With 95% confidence, the spring enrollment lies in the predict interval between 183
```

and 305, when fall enrollment is 290.

2.24 (d): Use the interval from part c, which is the predict interval because this interval is used to predict a new value for a particular spring enrollment, not an average.