R Lab 1. Introduction and Regression Examples

Example 1. US population.

Set a working directory. The name of yours will be different from mine. Then read the CSV (comma-separated values) file.

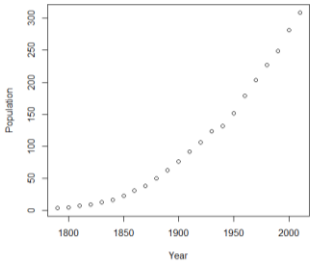
```
> setwd("C:\\Users\\baron\\Documents\\Teach\\615 Regression\\Data")
> read.csv("USpop.csv")
```

	Year	Population
1	1790	3.9
2	1800	5.3
3	1810	7.2
4	1820	9.6
5	1830	12.9
6	1840	17.1
7	1850	23.2
8	1860	31.4
9	1870	38.6
10	1880	50.2
11	1890	63.0
12	1900	76.2
13	1910	92.2
14	1920	106.0
15	1930	123.2
16	1940	132.2
17	1950	151.3
18	1960	179.3
19	1970	203.3
20	1980	226.5
21	1990	248.7
22	2000	281.4
23	2010	308.7

```
> P = read.csv("USpop.csv")
> plot(Year,Population)
Error in plot(Year, Population) : object 'Year' not found
```

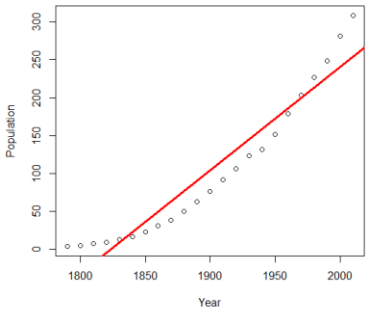
Oh, yes. We need to attach the dataset, so R knows which one we are working with.

```
> attach(P)
> plot(Year,Population)
```



This is called a *scatterplot*. Now fit a *linear regression line*, the closest to the given data points among all straight lines. Then draw this line on our scatterplot and predict the US population for year 2020 using this *linear* model.

```
> regr = lm(Population ~ Year)
> abline(regr, col="red", lwd=3)
```




```
> predict(regr, data.frame(Year=2020))
       1
267.2166
```

Clearly, linear regression is not the best choice here because apparently, the population does not grow linearly. Our linear model underestimates population in 1800s and 2000s and overestimates in the middle of the graph. As a result, the prediction of 267 million people two years from now is ridiculous because during the latest US Census in 2010, the US population was estimated to be 308 million.

At the same time, we see from the summary of our model (below) that both the slope and the intercept of our regression line are *significant*, and overall, the model explained 91.93% of the total variation, which is generally considered rather good.

```
> summary(regr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.481e+03  1.672e+02  -14.84 1.33e-12 ***
Year         1.360e+00  8.794e-02   15.47 5.93e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.97 on 21 degrees of freedom
Multiple R-squared:  0.9193,    Adjusted R-squared:  0.9155
F-statistic: 239.3 on 1 and 21 DF,  p-value: 5.927e-13
```

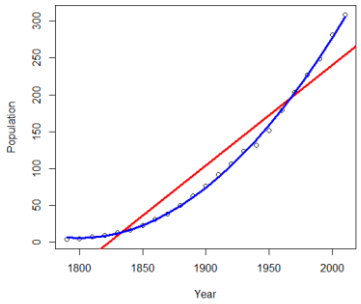
Let's try a more advanced *quadratic* model. Among all quadratic models (that is, all parabolas in the world), this one is the closest to our data points. Then, we calculate *predicted values* "Yhat" that are obtained by plugging all years 1790, 1800, …, 2010 into the resulting quadratic polynomial and plot this curve in blue. Wow, it fits the data really well! Also, now we explained 99.9% of the total variation. Only twice, during the WWII and right after, the US population appeared just a little below the curve. However, the "baby boom" came, and it quickly caught up with the quadratic trend.

```
> quad = lm(Population ~ poly(Year,2))
> summary(quad)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   103.9739     0.6304  164.94   <2e-16 ***
poly(Year, 2)1 432.7557     3.0231  143.15   <2e-16 ***
poly(Year, 2)2 127.4790     3.0231   42.17   <2e-16 ***

Residual standard error: 3.023 on 20 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.999
F-statistic: 1.113e+04 on 2 and 20 DF,  p-value: < 2.2e-16

> Yhat = predict(quad)
> lines(Year, Yhat, col="blue", lwd=3)
```




```
> predict(quad, data.frame(Year=c(2020,2030,2040)))
    1     2     3
334.9518 365.4891 397.3812
```

Our quadratic model predicts the US population of 335 mln. people in year 2020, 365 mln. in 2030, and 397 mln. in 2040. Let's wait and see. The current official estimate, as of August 27, 2018, is 327 mln.

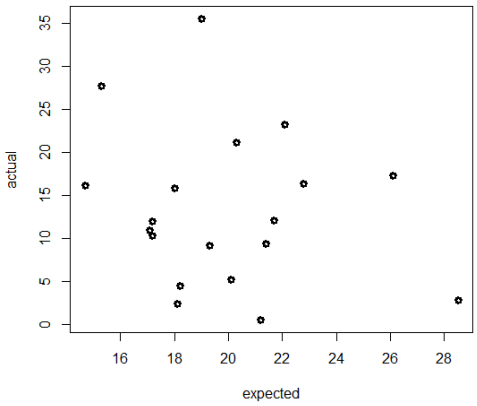
Example 2. US Presidents.

The given data set contains Presidents from Andrew Johnson to Lyndon Johnson. It compares the actual number of years each President lived after his first inauguration against the average number of years that a man of the same age would live during the same time.

```
> Pres = read.csv("presidents.csv")
> Pres
                  name expected  actual
```

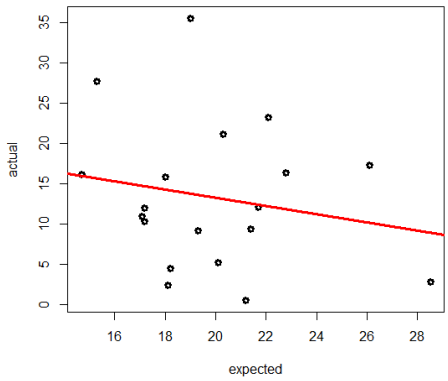
1	ANDREW JOHNSON	17.2	10.3
2	ULYSSES S. GRANT	22.8	16.4
3	RUTHERFORD B. HAYES	18.0	15.9
4	JAMES A. GARFIELD	21.2	0.5
5	CHESTER A. ARTHUR	20.1	5.2
6	GROVER CLEVELAND	22.1	23.3
7	BENJAMIN HARRISON	17.2	12.0
8	WILLIAM MCKINLEY	18.2	4.5
9	THEODORE ROOSEVELT	26.1	17.3
10	WILLIAM H. TAFT	20.3	21.2
11	WOODROW WILSON	17.1	10.9
12	WARREN G. HARDING	18.1	2.4
13	CALVIN COOLIDGE	21.4	9.4
14	HERBERT C. HOOVER	19.0	35.6
15	FRANKLIN D. ROOSEVELT	21.7	12.1
16	HARRY S. TRUMAN	15.3	27.7
17	DWIGHT D. EISENHOWER	14.7	16.2
18	JOHN F. KENNEDY	28.5	2.8
19	LYNDON B. JOHNSON	19.3	9.2

```
> attach(Pres)
> plot(expected, actual, lwd=3)
```



If presidents are just ordinary people, and the fact of presidency does not affect their lifetime, then the points should be close to the bisector line $y = x$. But surprisingly, here we see that this is not the case. OK, let's fit a linear regression line.

```
> reg = lm(actual ~ expected)
> abline(reg, col="red", lwd=3)
```



> reg

```
Coefficients:
(Intercept)     expected
    23.3867      -0.5061
```

What do we see? The slope is negative! It means that presidents who were expected to live longer actually passed away sooner. Perhaps, it is because three of the listed presidents were assassinated – Garfield, McKinley, and Kennedy (#4, 8, 18 in the data set). One may argue that they should be excluded from modeling (although others may disagree with this referring to the US history - the risk of being assassinated is much higher when you become a US President, so these three may be a part of the trend and not an outlier). But the whole picture is not that different even without the assassinated presidents.

> Z = c(4,8,18)
> reg = lm(actual ~ expected, data=Pres[-Z,])
> summary(reg)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.384876  14.970195   1.028    0.322
expected    -0.003409   0.763337  -0.004    0.997
```

The slope is still negative. We can see from a high p-value actually, it is not significant, based on our data. That is, there is no evidence that the slope is not 0. In other words, knowing the age of a US President at his or her first inauguration and the life expectancy of a person of the same age does not really help to predict the life expectancy of the President. A surprising conclusion, I think.