

Lab 1

Yunting Chiu

2021-01-22

Example 1: US population

```
# read a dataset
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.2      v purrr 0.3.4
## v tibble 3.0.3       v dplyr 1.0.2
## v tidyr 1.1.2        v stringr 1.4.0
## v readr 1.3.1        v forcats 0.5.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()

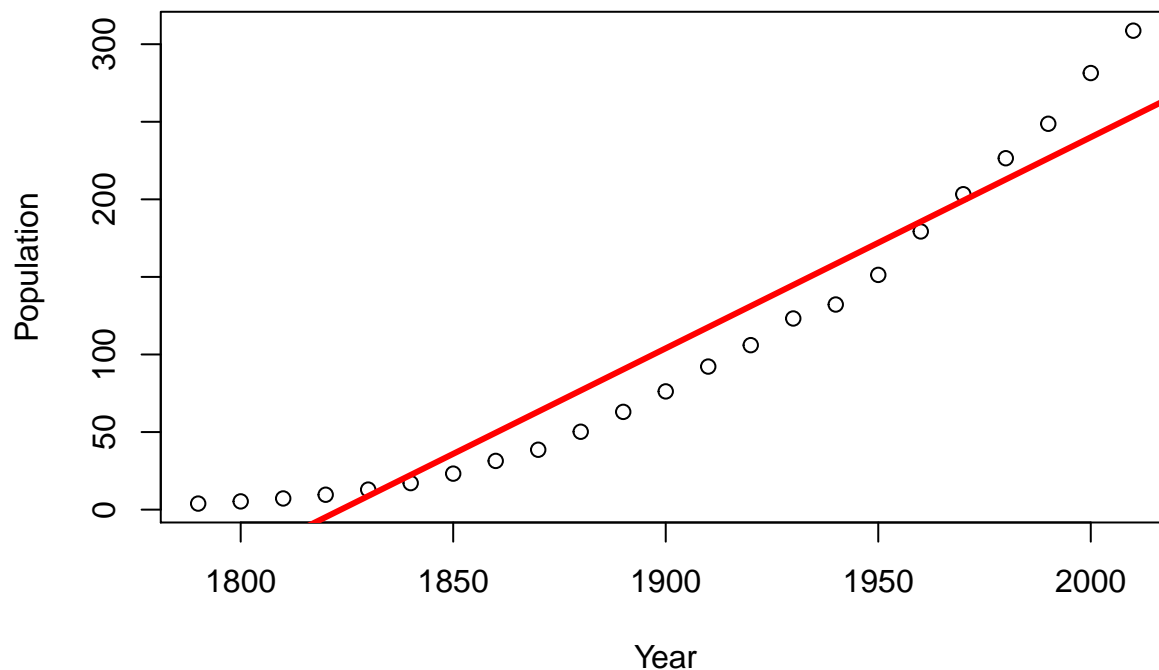
USpop <- read_csv("./data/USpop.csv")

## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Population = col_double()
## )

USpop

## # A tibble: 23 x 2
##   Year Population
##   <dbl>      <dbl>
## 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
## # ... with 13 more rows

attach(USpop) #replace %>%
plot(Year,Population)
regr = lm(Population ~ Year) # lm(y~x)
abline(regr, col="red", lwd=3) #fit a linear regression line
```



- According to the above plot, some outliers can be found at the right top, these observations can be defined as potential outliers, and the population does not grow linearly. We are therefore considering changing another model for this study case.

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

```
##      1
## 267.2166
```

```
summary(regr)
```

```
##
## Call:
## lm(formula = Population ~ Year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.774 -24.872  -6.295  18.374  55.087
##
## 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
```

- With a small p-value, the summary of current simple linear model indicates intercept and slope (Year) are reject the null hypothesis. Also, the Multiple R-squared is 91.93 % of the total variation (greater than 50 %). If we only focus on the result of R-squared, which is a good model, but the prerequisite is that we need to check the plot of x and y variables whether they are linear.

Quadratic model

- We consider changing a `Year` variable to the quadratic transformation. That is, we need a quadratic term in our model.

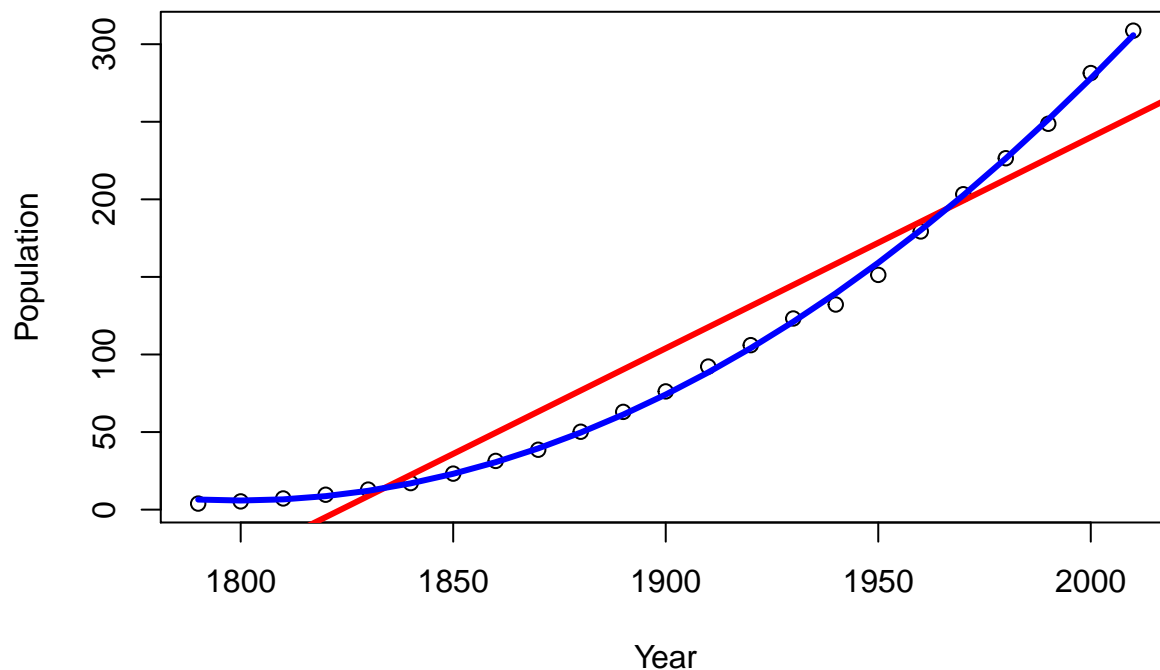
```
quad <- lm(Population ~ poly(Year,2))
summary(quad)

##
## Call:
## lm(formula = Population ~ poly(Year, 2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.8220 -0.7130  0.5961  1.8344  3.7487
##
## 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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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)

# recall and compare it
attach(USpop)

## The following objects are masked from USpop (pos = 3):
##
##      Population, Year
plot(Year,Population)
abline(regr, col="red", lwd = 3) #fit a linear regression line
lines(Year, Yhat, col = "blue", lwd = 3)
```



- Firstly, we saw 99.9% of the total variation in this model. Secondly, The `Yhat` is quadratic polynomial of the `Year` from 1790 to 2010. Based on the plot above, it is obvious that the blue curve fits the data points better, meaning that the quadratic model predicts the US population better than the linear model.

```
predict(quad, data.frame(Year=c(2020,2030,2040)))
```

```
##          1          2          3
## 334.9518 365.4891 397.3812
```

Example 2: US Presidents

```
#read a dataset
pres <- read_csv("../data/presidents.csv")
```

```
## Parsed with column specification:
## cols(
##   name = col_character(),
##   expected = col_double(),
##   actual = col_double()
## )
```

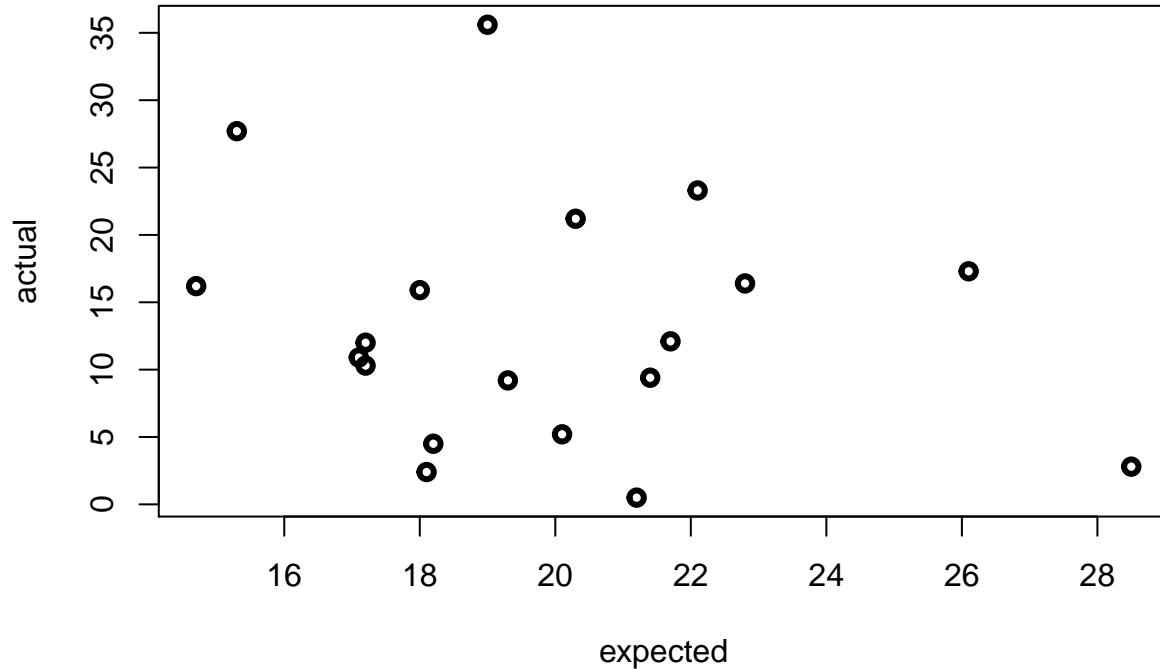
```
pres
```

```
## # A tibble: 19 x 3
##   name                expected actual
##   <chr>                <dbl>   <dbl>
## 1 ANDREW JOHNSON        17.2    10.3
## 2 ULYSSES S. GRANT      22.8    16.4
## 3 RUTHERFORD B. HAYES   18      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
##	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	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

Plot

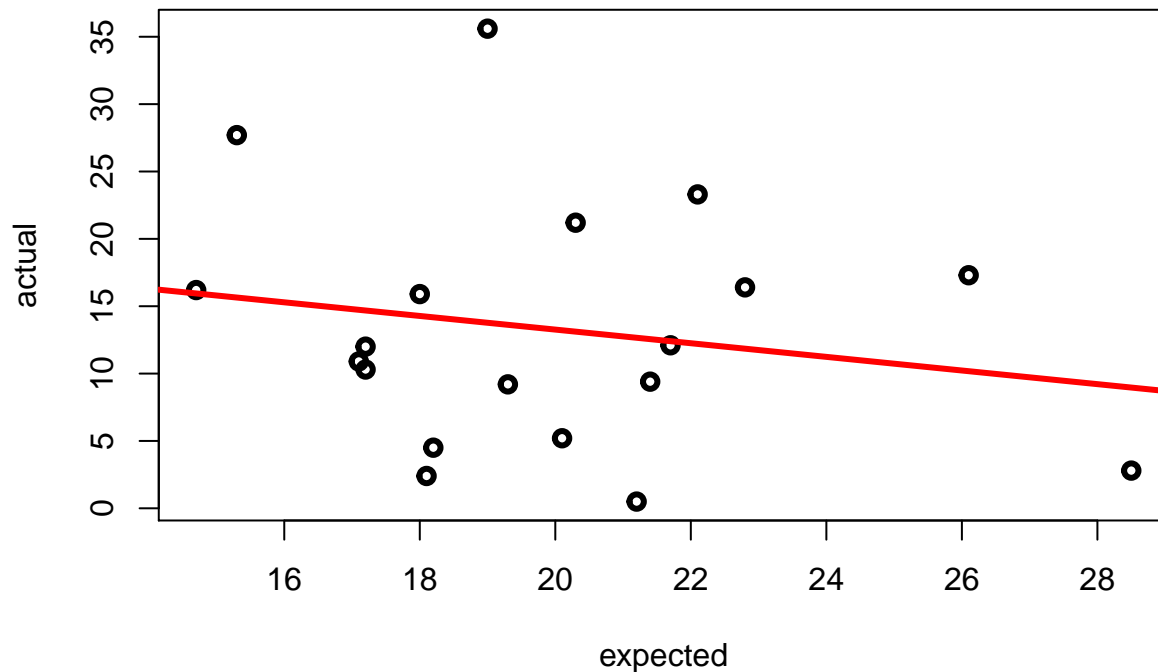
```
attach(pres)
plot(expected, actual, lwd=3)
```



- If the president's position does not interfere with the life expectancy, the points should be **actual** equal to **expected**.

```
# add abline and run again
attach(pres)
```

```
## The following objects are masked from pres (pos = 3):
##
##      actual, expected, name
plot(expected, actual, lwd=3)
reg = lm(actual ~ expected)
abline(reg, col="red", lwd=3)
```



```
Z = c(4,8,18)
```

```
# remove outliers and fit linear regression
reg = lm(actual ~ expected, data = pres[-Z,])
summary(reg)
```

```
##
## Call:
## lm(formula = actual ~ expected, data = pres[-Z, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.923  -5.248  -1.317   2.974  20.280
##
## 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
##
## Residual standard error: 8.771 on 14 degrees of freedom
## Multiple R-squared:  1.424e-06, Adjusted R-squared:  -0.07143
## F-statistic: 1.994e-05 on 1 and 14 DF,  p-value: 0.9965
```

- The negative slope shows we have all overestimated their life expectancy. In this case, we may found out about some outliers, such as 4, 8, 18 observations. So we try to remove then and run the linear regression model.
- In summary, the large p-value 0.997 indicates that we fail to reject the null hypothesis, which is not significant. Also, the slope is still negative that we did not find the slope is equal to zero, meaning that the `expected` variable cannot help to predict the President's life expectancy (`actual`).