DSA 8020 R Lab 1: Simple Linear Regression

Whitney

February 17, 2021

Contents

Leaning Tower of Pisa

The dataset `PisaTower.csv` provides the annual measurements of the lean (the difference between where a point on the tower would be if the tower were straight and where it actually is) from 1975 to 1987. We would like to characterize lean over time by fitting a simple linear regression.

Load the dataset

Code:

```
PisaTower <- read.csv("PisaTower.csv")
head(PisaTower)
```

```
##     lean year
## 1 2.9642 1975
## 2 2.9644 1976
## 3 2.9656 1977
## 4 2.9667 1978
## 5 2.9673 1979
## 6 2.9688 1980
```

Descriptive analysis

Numerical summary

Code:

```
summary(PisaTower)
```

```
##       lean            year
##  Min.   :2.964   Min.   :1975
##  1st Qu.:2.967   1st Qu.:1978
##  Median :2.970   Median :1981
##  Mean   :2.969   Mean   :1981
##  3rd Qu.:2.972   3rd Qu.:1984
##  Max.   :2.976   Max.   :1987
```

```
cor(PisaTower)
```

```
##           lean      year
## lean 1.0000000 0.9939717
## year 0.9939717 1.0000000
```

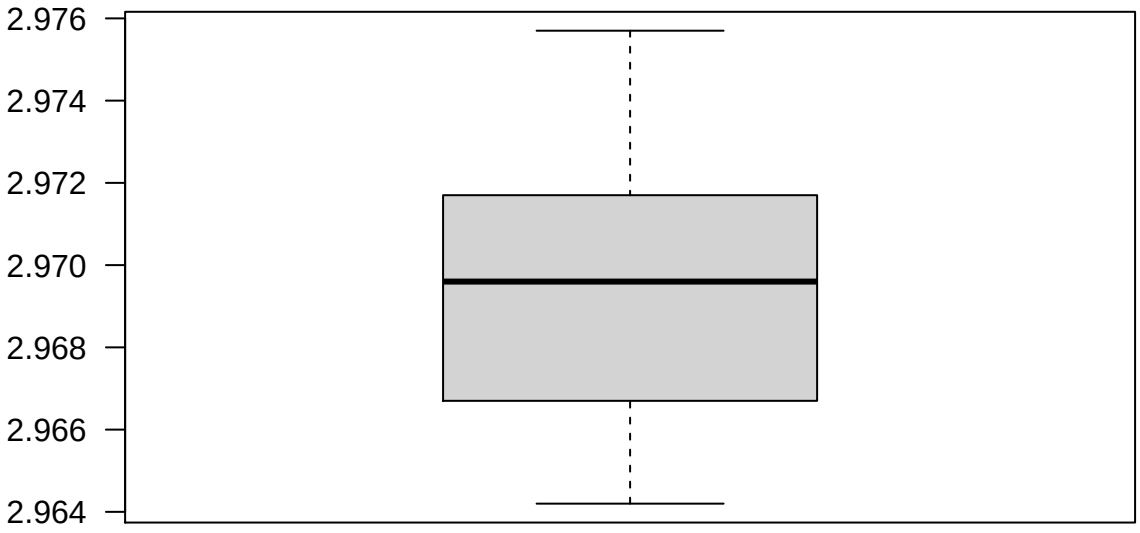
```
sd(PisaTower$lean)
```

```
## [1] 0.003651115
```

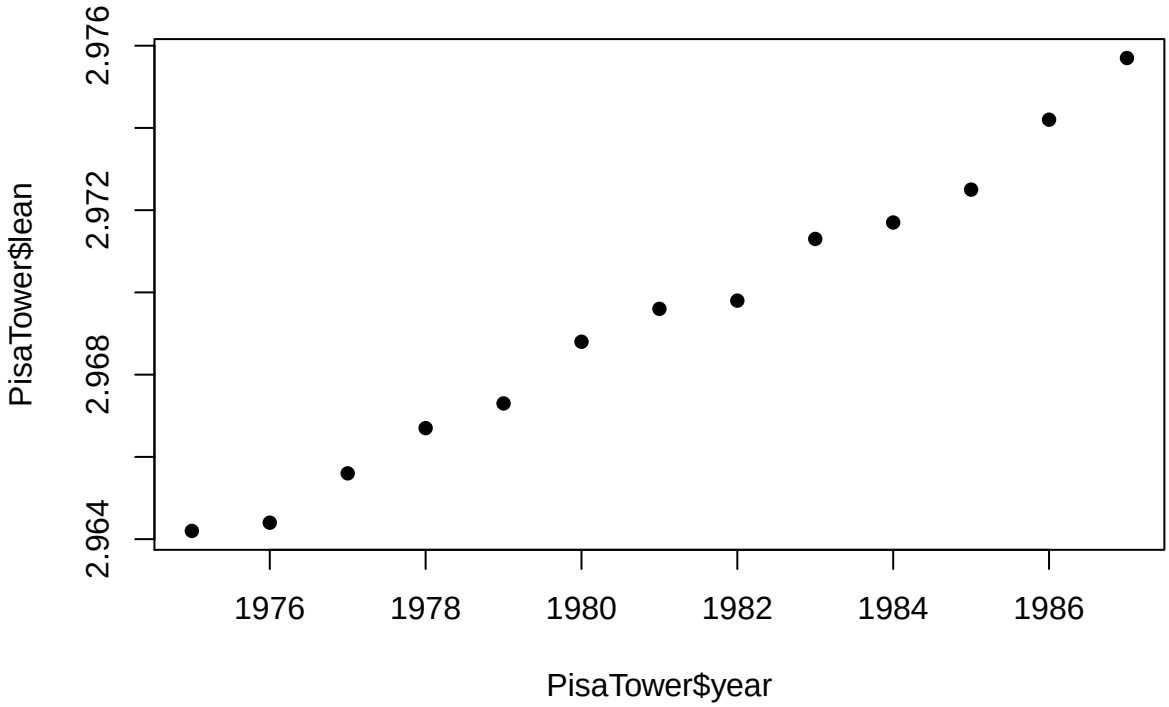
Graphical summary

Code:

```
boxplot(PisaTower$lean, las = 1)
```




```
plot(PisaTower$year, PisaTower$lean, pch = 16)
```



Question: Describe the direction, strength, and the form of the relationship.

Answer:

There is a strong, postive, linear relationship between lean and year.

Simple linear regression

1. Identify the response variable, the predictor variable, and the sample size.

Answer: Response variable: `lean`. Predictor variable: `year`. Sample size: $n = 13$

2. Fit a simple linear regression

Code:

```
lmfit <- lm(lean ~ year, data = PisaTower)
summary(lmfit)
```

```
##
## Call:
## lm(formula = lean ~ year, data = PisaTower)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.967e-04 -3.099e-04  6.703e-05  2.308e-04  7.396e-04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.123e+00  6.139e-02   18.30 1.39e-09 ***
```

```
## year        9.319e-04  3.099e-05   30.07 6.50e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0004181 on 11 degrees of freedom
## Multiple R-squared:  0.988,  Adjusted R-squared:  0.9869
## F-statistic: 904.1 on 1 and 11 DF,  p-value: 6.503e-12
```

3. Write down the fitted linear regression model

Answer:

$\hat{\text{lean}} = 1.123 + 9.319 \times 10^{-4} \times \text{year}$

4. What is $\hat{\sigma}$, the estimate of σ

Answer:

$\hat{\sigma} = 4.1809711 \times 10^{-4}$

5. Find a 95% confidence interval for β_1

Code:

```
confint(lmfit)[2,]
```

```
##        2.5 %       97.5 %
## 0.0008636565 0.0010000798
```

6. Test the following hypothesis: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ with $\alpha = 0.05$

Answer:

Test statistic: $t = 30.07 \Rightarrow$ p-value $= 6.50 \times 10^{-12}$ Reject H_0 and conclue $\beta_1 \neq 0 \Rightarrow$ evidence suggests a postive linear relationship between `lean` and `year`.

7. Construct a 90% confidence interval for E[*Lean*] in year 1984

Code:

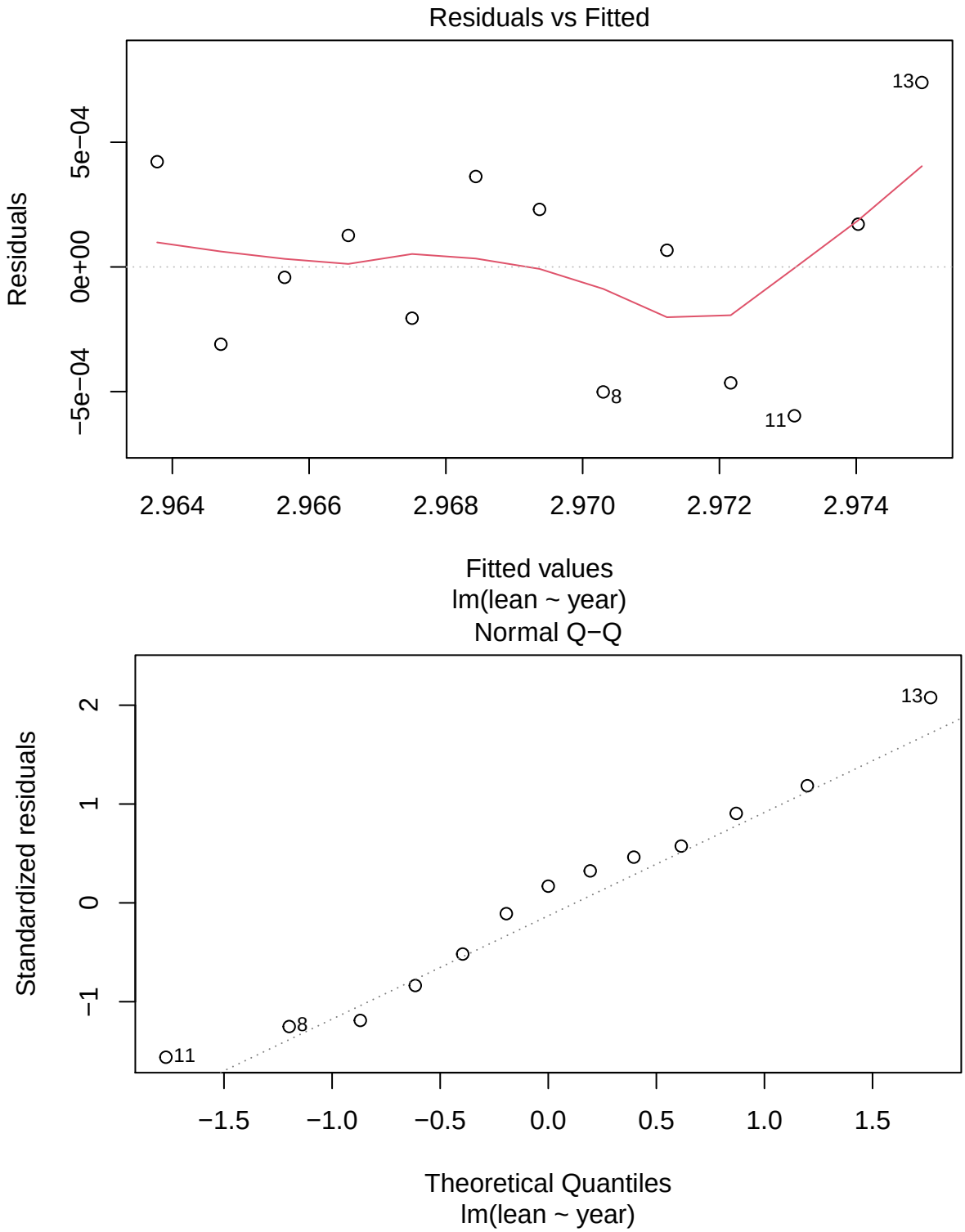
```
predict(lmfit, newdata = data.frame(year = 1984), interval = "confidence", level = 0.9)
```

```
##        fit      lwr      upr
## 1 2.972165 2.971898 2.972432
```

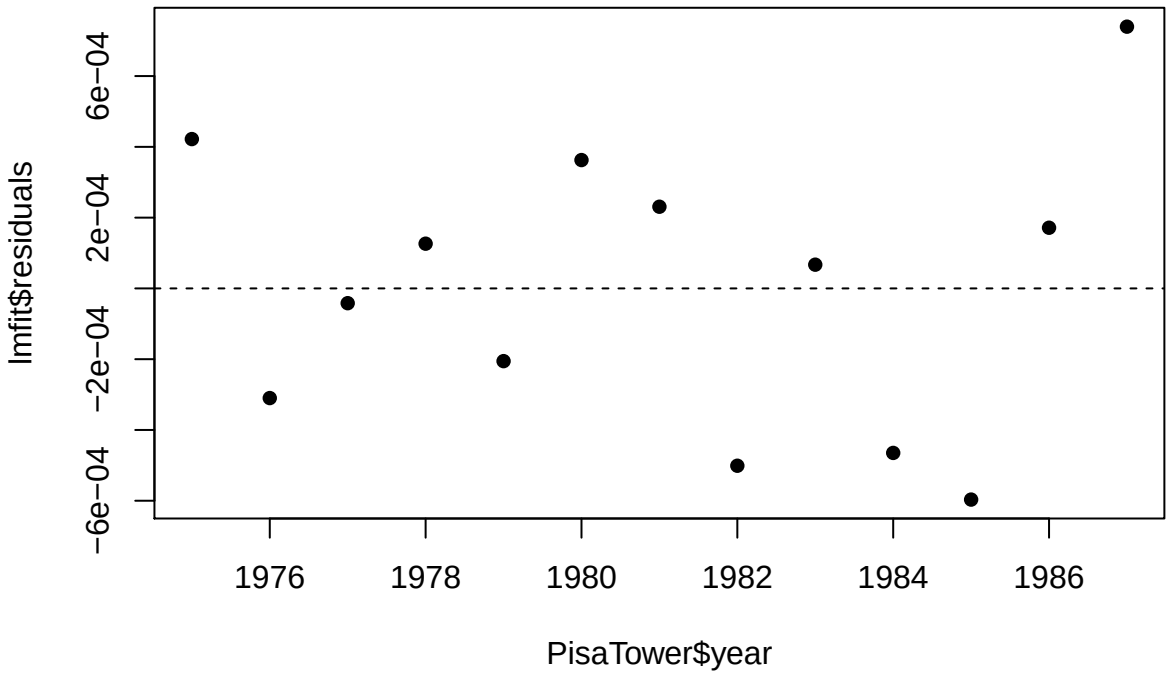
8. Use residuals to check model assumptions

Code:

```r
plot(lmfit, which = 1:2)
```




```r
plot(PisaTower$year, lmfit$residuals, pch = 16)
abline(h = 0, lty = 2)
```



Answer:

There is no major concern about model assumptions, namely linearity, constant variance, normality, and independence given the size of the data. The relatively large deviation of the very last observation may casue some concerns but it is also possible simply just due to random fluctuation.

9. Would it be a good idea to use the fitted linear regression equation to predict `lean` in year 2010? Explain your answer.

Answer:

This is an example of extrapolation in regression and most of the statistics textbooks will tell not to do so, even the fit within the data range is nearly prefect. In this particular example we have a good reason not to carry out such an extrapolation. Please take a look at the third paragrah of the wiki page (link) to find out why.