DSA 8020 R Lab 1: Simple Linear Regression

Meredith Sliger

January 16, 2025

Contents

Leaning Tower of Pisa

The dataset `PisaTower.csv` provides 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
```

```
nrow(PisaTower) # added for myself to see number rows
```

```
## [1] 13
```

Descriptive analysis

Numerical summary

Provide some numerical summaries to describe the response and the predictor variables, respectively, as well as their relationship.

Code:

```
y <- PisaTower$lean; x <- PisaTower$year
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
```

```
var(x) # calculate sample variance of variable x
```

```
## [1] 15.16667
```

```
var(y) # calculate sample variance of variable y
```

```
## [1] 1.333064e-05
```

```
cov(x, y) # calculate covariance
```

```
## [1] 0.01413333
```

```
cor(x, y) # calculate correlation; closer to 1 suggests stronger linear relationship
```

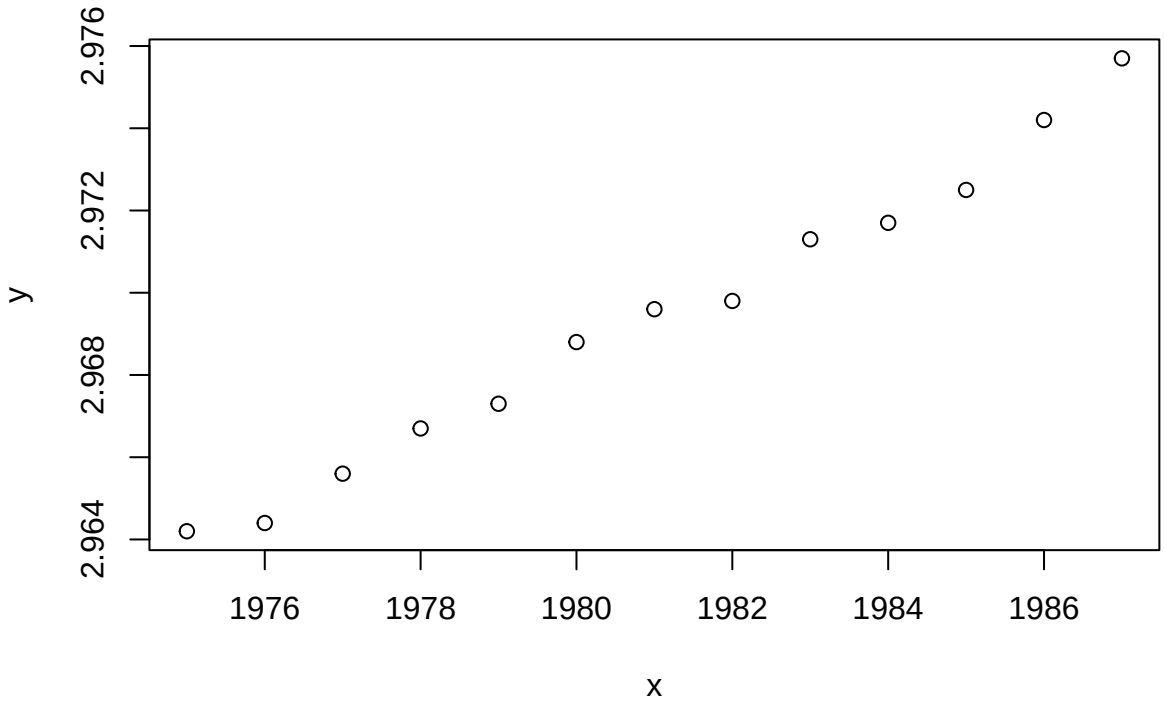
```
## [1] 0.9939717
```

Graphical summary

Provide graphical summaries through plots to describe the response and predictor variables, respectively, as well as their relationship.

Code:

```
plot(x, y)
```



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

Answer: Direction is positive. Strength is strong, supported by the correlation between lean and year being calculated as 0.9939717. Form is linear as the points seem to follow a straight-line.

Simple linear regression

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

Answer: Response Variable: Lean; Predictor Variable: Year; Sample Size: 13.

2. Fit a simple linear regression.

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
```

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

```
##
## 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: lean = 1.123e+00 + 9.319e-04 x year or lean = 1.123 + 0.0009319 x year

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

Answer: 0.0004181

5. Find a 95% confidence interval for β_1.

Code:

```r
confint(lm, level = .95) # adjusted to .95 to reflect 95%
```

```
##                    2.5 %     97.5 %
## (Intercept) 0.9882109317 1.25846599
## year        0.0008636565 0.00100008
```

```r
beta1_hat <- lm$coefficients[2]
se_beta1_hat <- se_beta1 <- summary(lm)[["coefficients"]][, 2][2]

n <- dim(PisaTower)[1]
alpha <- 0.05 # changed from 0.10 to 0.05 to reflect 95%

t <- qt(1 - alpha / 2, n - 2)

beta1_hat + c(-1, 1) * t * se_beta1_hat
```

```
## [1] 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: Since the 95% confidence interval for β_1 (0.0008636565, 0.0010000798) excludes 0, we reject the null hypothesis at the 5% significance level. This provides strong evidence of a significant relationship between lean and year.

7. Construct a 90% confidence interval for E[`lean`] in year 1984

Code:

```
year_1984 <- data.frame(year = 1984)
hat_Y <- lm$coefficients[1] + lm$coefficients[2] * 1984
hat_Y
```

```
## (Intercept)
##    2.972165
```

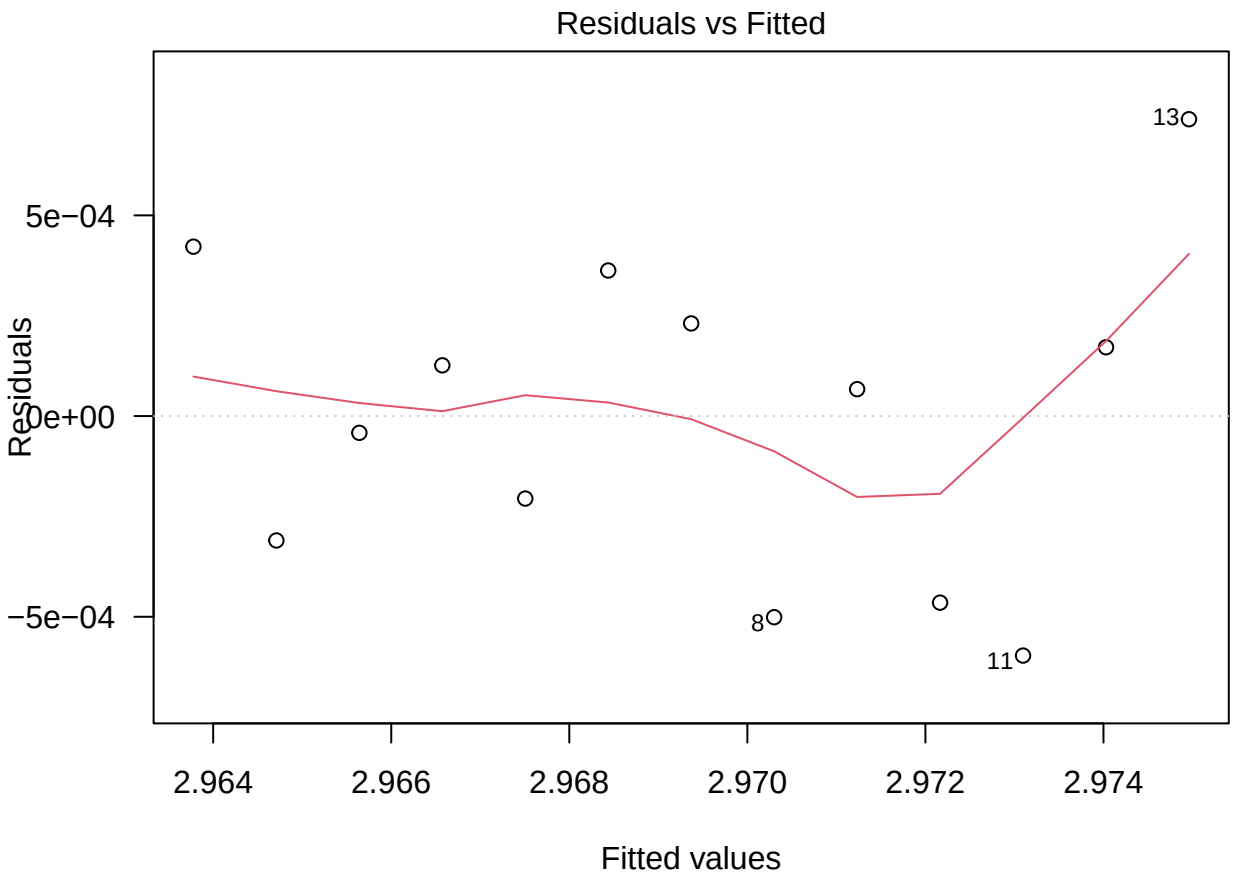
```
predict(lm, year_1984, interval = "confidence", level = 0.90)
```

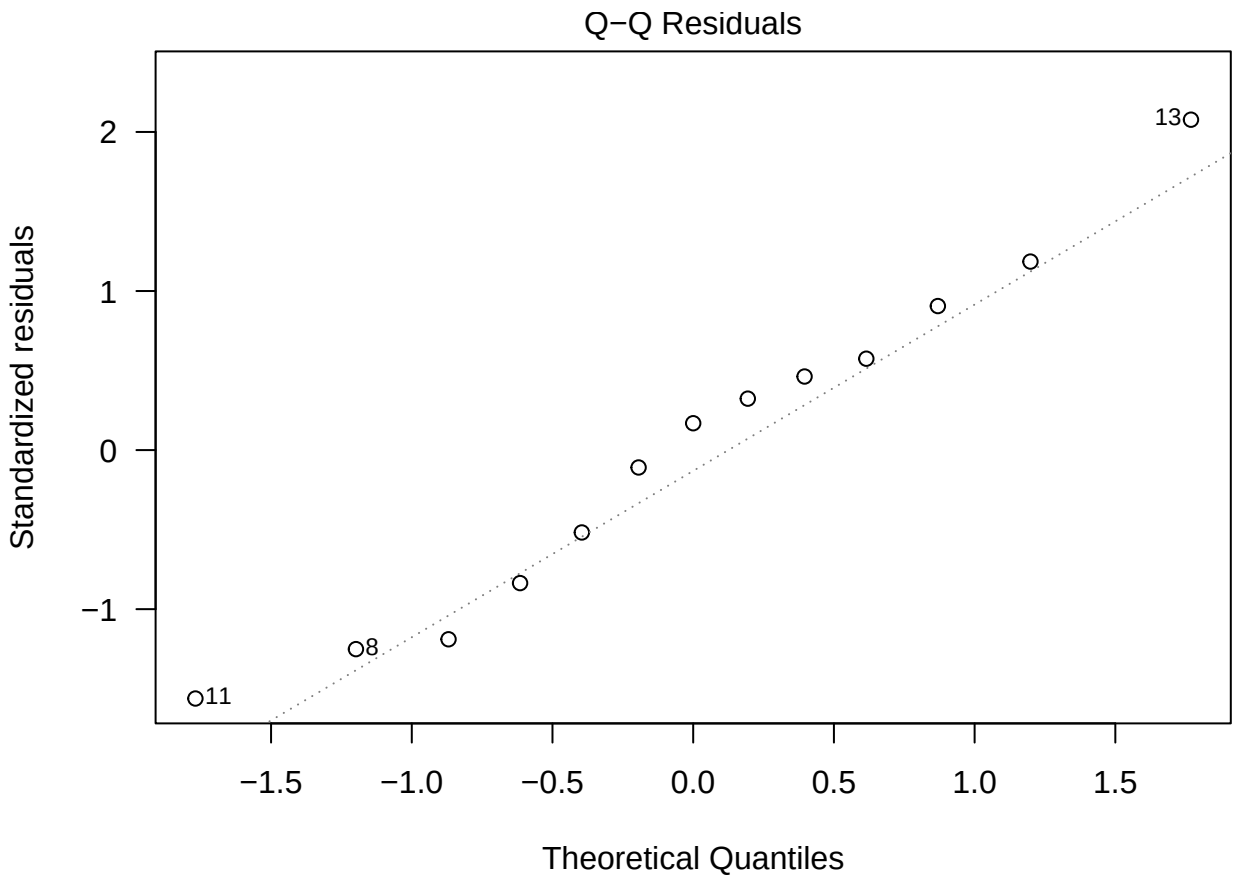
```
##        fit      lwr      upr
## 1 2.972165 2.971898 2.972432
```

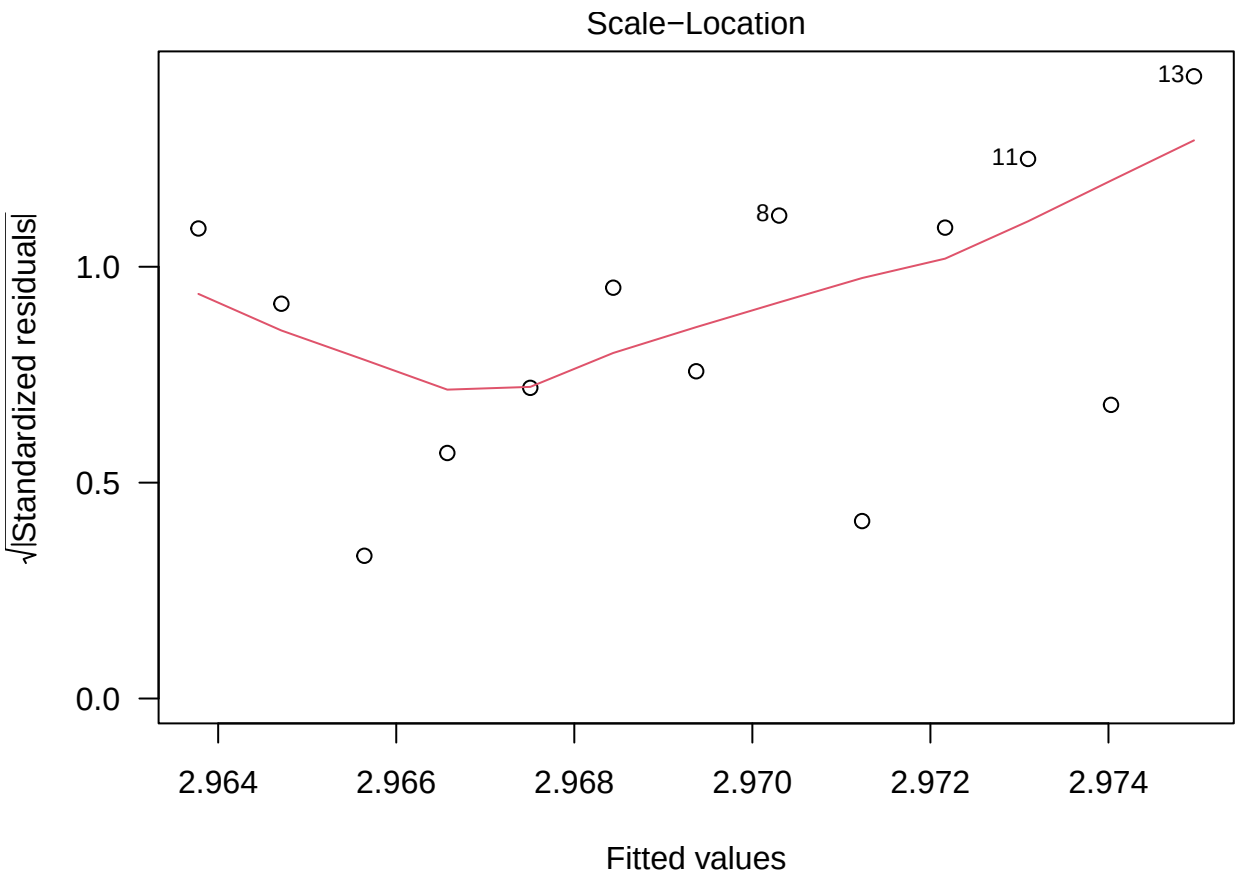
8. Use residuals to check model assumptions.

Code:

```
par(mar = c(4, 4, 1, 0.5), mgp = c(3, 1, 0), las = 1)
plot(lm, which = 1:3)
```


Residuals vs Fitted
Residuals
Fitted values
5e−04
0e+00
−5e−04
2.964
2.966
2.968
2.970
2.972
2.974
13
8
11


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
2
1
0
−1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
13
8
11



Answer: Residuals vs fitted plot shows a slight curve in the red line that shows the trend of the residuals. This indicates that the relationship between year and lean may not be perfectly linear, possibly due to outliers or other external factors influencing the data. The Q-Q plot seems to show a mostly normal distribution. Two points in particular seem to be further away from the line (11, 13), indicating some minor deviation. The Scale-Location plot shows a slight upward trend as the fitted values increase. The trend is not too noticeable and the points are scattered around the graph with no real pattern. Because of the slight upward trend with the red line, there could be a possibility that the variance of the residuals increases as the predicted values get larger but it is not a given.

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

Answer: Personally I would not use the fitted linear regression equation for a prediction in 2010. My main reasoning is due to our data being from the year 1975 to 1987 and nothing beyond that to present day. There is no confirmation that this trend will continue beyond 1987. There are also external factors such as environmental changes or structural changes influencing the lean which would make any predictions outside the observed years' range unreliable. To make a prediction for 2010 would rely on assumptions and extrapolation which would produce untrustworthy results.