Simple Linear Regression - Lab

Blake Pappas

8/24/2021

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:

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

```
##     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 Statistics
y <- pisa$lean; x <- pisa$year
summary(pisa)
```

```
##       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  
```

```r
# Finding the Variance in year and lean
var(x); var(y)
```

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

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

```r
# Finding the Covariance Between year and lean
cov(x, y)
```

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

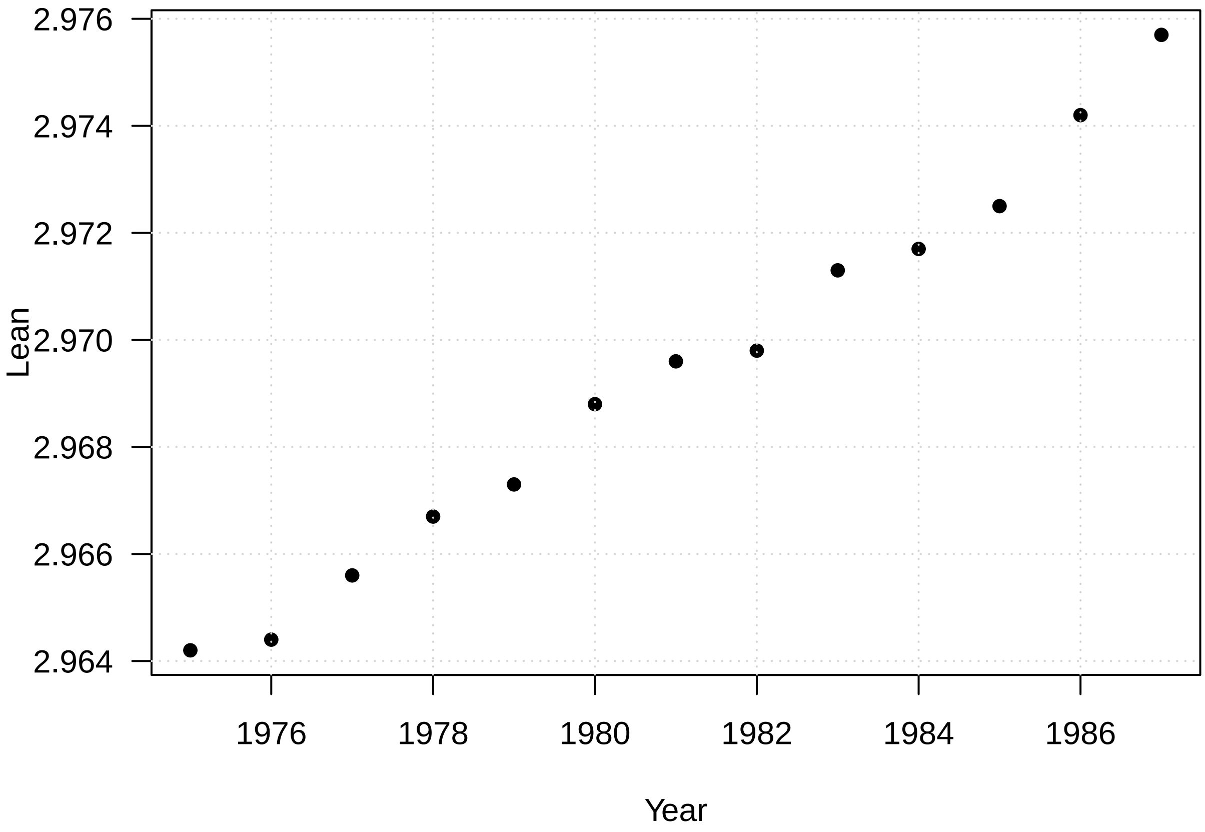
```r
# Finding the Correlation Between year and lean
cor(x, y)
```

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

Graphical Summary

Code:

```r
par(las = 1, mar = c(4.1, 4.1, 1.1, 1.1))
plot(x, y, pch = 16, xlab = "Year", ylab = "Lean")
grid()
```



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

Answer: The variables `lean` and `year` have a strong, positive, linear relationship, with a correlation of approximately 0.9939717. For each additional year in time, the lean of the tower increases.

Simple Linear Regression

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

Answer: The response variable for this data set is `lean`. The predictor variable for this data set is `year`. The sample size of this data set is 13.

2. Fit a simple linear regression.

Code:

```r
# Solving for Slope
y_diff <- y - mean(y)
x_diff <- x - mean(x)
beta_1 <- sum(y_diff * x_diff) / sum((x_diff)^2)
beta_1
```

```
## [1] 0.0009318681
```

```r
# Solving for Intercept
beta_0 <- mean(y) - mean(x) * beta_1
beta_0
```

```
## [1] 1.123338
```

```r
# Solving for the Fitted Values
y_hat <- beta_0 + beta_1 * x
y_hat
```

```
##  [1] 2.963778 2.964710 2.965642 2.966574 2.967505 2.968437 2.969369 2.970301
##  [9] 2.971233 2.972165 2.973097 2.974029 2.974960
```

```r
# Solving for Variance and Standard Deviation
sigma2 <- sum((y - y_hat)^2) / (length(y) - 2)
sqrt(sigma2)
```

```
## [1] 0.0004180971
```

3. Write down the fitted linear regression model.

Answer: LEAN = 1.1233385 + 0.0009319 x YEAR + ϵ, where Y = `lean`, X = `year`, $\hat{\beta}_0$ = 1.1233385, $\hat{\beta}_1$ = 0.0009319, and ϵ = stochastic error.

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

Answer: The $\hat{\sigma}$ is approximately 0.0004180971.

5. Find a 95% confidence interval for β_1.

Code:

```
fit <- lm(lean ~ year, data = pisa)

alpha = 0.05
beta1_hat <- summary(fit)[["coefficients"]][, 1][2]
se_beta1 <- summary(fit)[["coefficients"]][, 2][2]
CI_beta1 <- c(beta1_hat - qt(1 - alpha / 2, 11) * se_beta1,
beta1_hat + qt(1 - alpha / 2, 11) * se_beta1)
CI_beta1
```

```
##         year         year
## 0.0008636565 0.0010000798
```

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

```
beta1_null <- 0
t_star <- (beta1_hat - beta1_null) / se_beta1
p_value <- 2 * pt(t_star, 11, lower.tail = F)
p_value
```

```
##         year
## 6.503367e-12
```

Answer: The calculated p-value of 6.503367e-12 is less than the α of 0.05. Therefore, we reject H_0. There is statistically significant evidence to suggest that $\beta_1 \neq 0$.

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

Code:

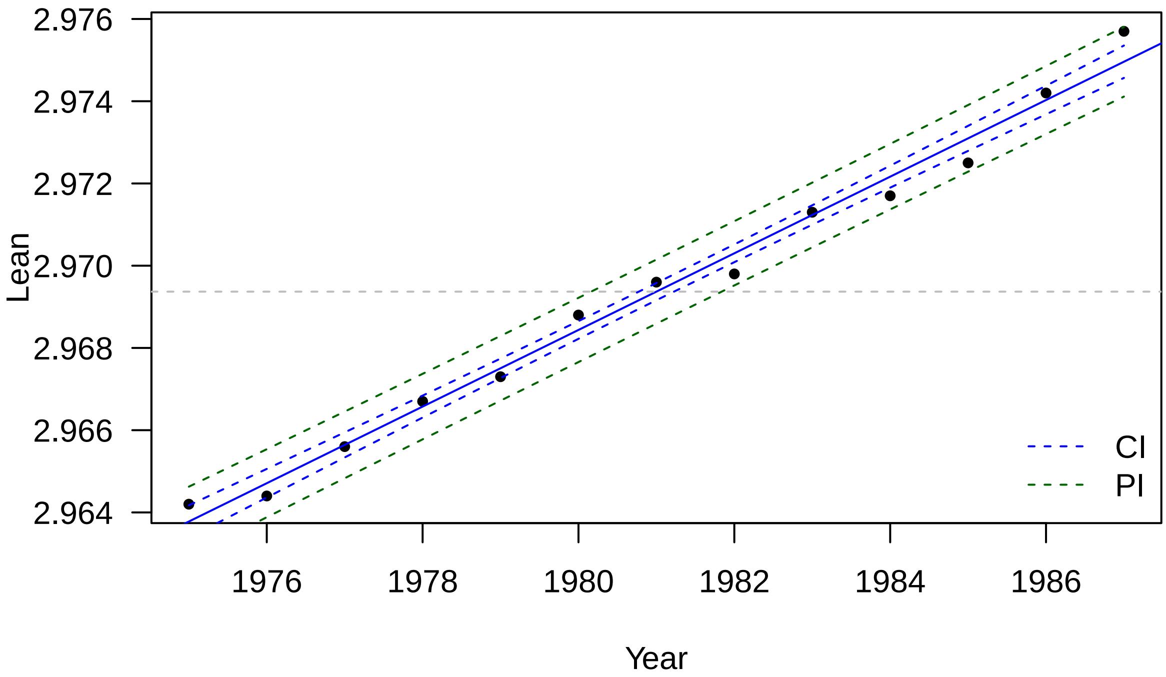
```
year_new = data.frame(year = 1984)
hat_Y <- fit$coefficients[1] + fit$coefficients[2] * 1984
hat_Y
```

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

```
predict(fit, year_new, interval = "confidence", level = 0.9)
```

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

```
year_grid = data.frame(year = 1975:1987)
CI_band <- predict(fit, year_grid, interval = "confidence", level = 0.9)
PI_band <- predict(fit, year_grid, interval = "predict", level = 0.9)
plot(pisa$year, pisa$lean, pch = 16, cex = 0.75,
xlab = "Year", ylab = "Lean", las = 1)
abline(fit, col = "blue")
abline(h = mean(pisa$lean), lty = 2, col = "gray")
lines(1975:1987, CI_band[, 2], lty = 2, col = "blue")
lines(1975:1987, CI_band[, 3], lty = 2, col = "blue")
lines(1975:1987, PI_band[, 2], lty = 2, col = "darkgreen")
lines(1975:1987, PI_band[, 3], lty = 2, col = "darkgreen")
legend("bottomright", legend = c("CI", "PI"), col = c("blue", "darkgreen"), lty = 2, bty = "n")
```

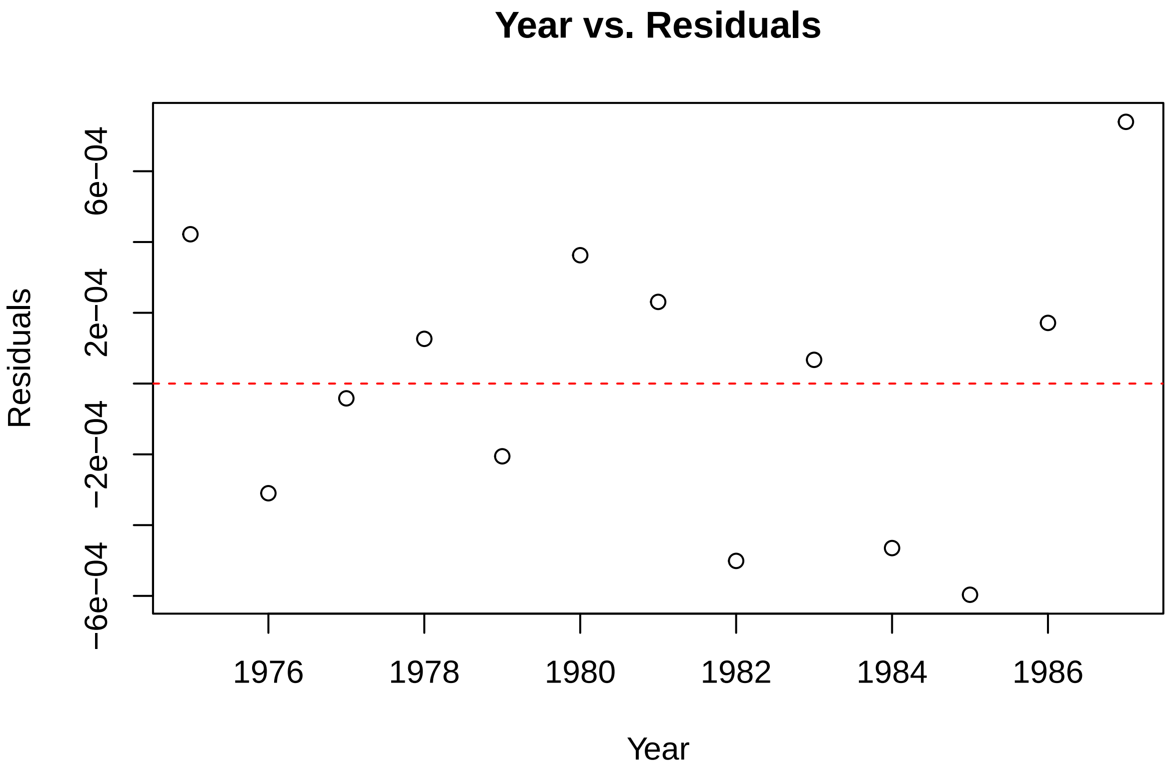


Answer: The 90% confidence interval for $\mathrm{E}[Lean]$ in year 1984 is (2.971898, 2.972432).

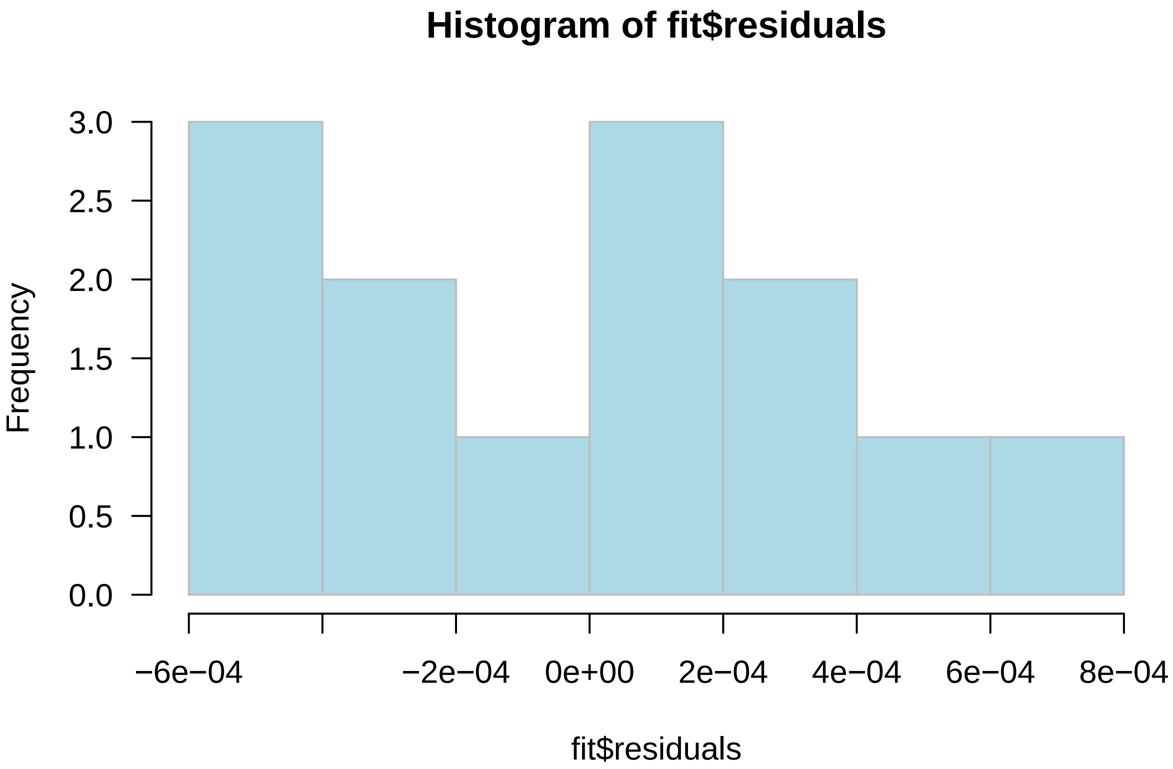
8. Use residuals to check model assumptions.

Code:

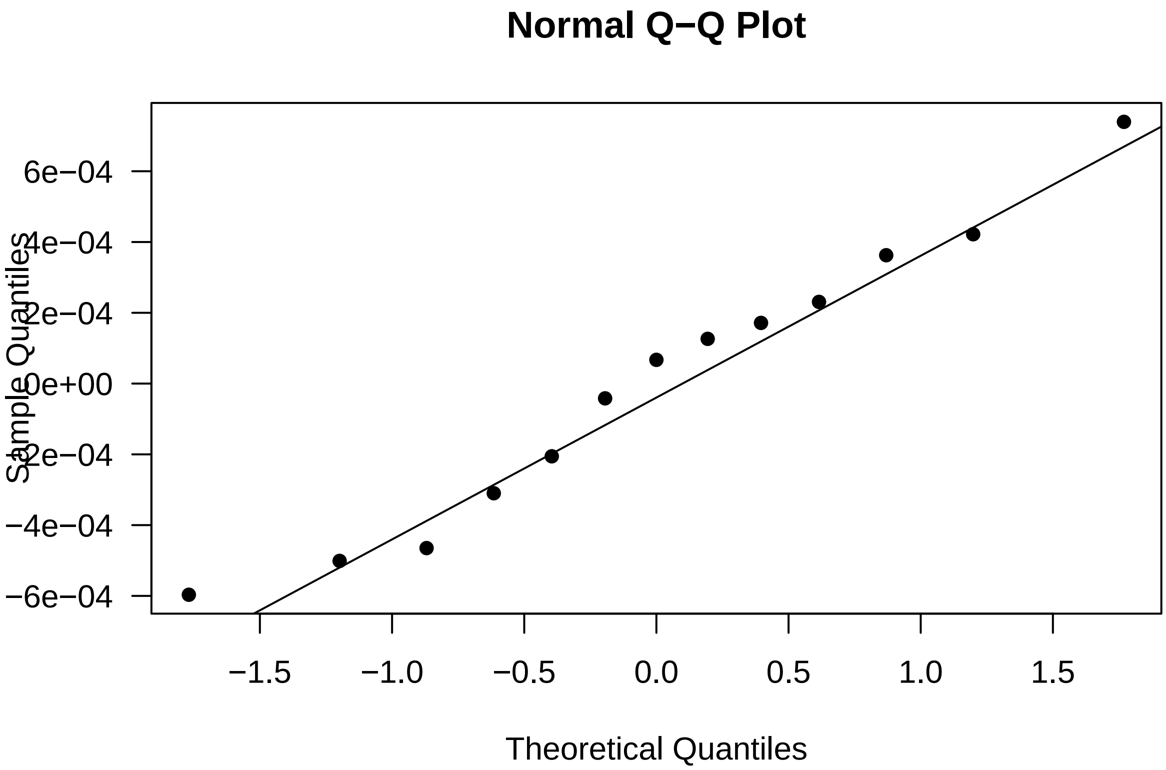
```
pisa_resid <- residuals(fit)
pisa_fitted <- fitted(fit)

plot(pisa$year, pisa_resid, main = 'Year vs. Residuals',
     xlab = 'Year', ylab = 'Residuals')
abline(h = 0, col = "red", lty = 2)
```



```r
hist(fit$residuals, col = "lightblue", border = "gray", las = 1)
```



```r
qqnorm(fit$residuals, pch = 16, las = 1)
qqline(fit$residuals)
```



Answer: Based on the plots and histogram, the assumption of normality appears to be reasonable. The plot of the residuals appears to be random and the Normal Q-Q Plot appears to run closely (in an S-shaped pattern) to the trend line.

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

Answer: It would not be a good idea to use the fitted linear regression equation to predict `lean` in the year 2010 because of danger of extrapolation. This occurs when a regression is used to make a prediction for something that lies beyond the range of data given. Often times, it leads to extremely biased estimates. In this instance, the year 2021 lies greatly outside the data set, which has a maximum year of 1987. It would not be ideal to use 13-years of data to from the late 1900s to predict something nearly 25 years later in the early 2000s.