Simple Linear Regression

Andrew Levine

October 2021

```r
DrivingSight <- read.csv(file = "/Users/andrewlevine/Downloads/Statistics II/DrivingSight.csv")
```

Response variable: Maximum Distance one can read a highway sign, Quantitative Variable
Explanatory variable: Age, Quantitative Variable

Correlation Coefficient:

```r
cor(DrivingSight$Age, DrivingSight$Distance)
```

```
## [1] -0.7709635
```

r = -0.7709635
indicates strong, negative, linear relationship between age and max distance

Scatter Plot:

```r
lm_info_DS <- lm(Distance ~ Age, data = DrivingSight)
plot(DrivingSight$Age, DrivingSight$Distance,
     main = "Age vs. Maximum Driving Distance Sight",
     xlab = "Age",
     ylab = "Maximum Distance"
)
abline(a = coef(lm_info_DS)[1], b = coef(lm_info_DS)[2])
abline(a = (lm_info_DS$coefficients)[1], b = (lm_info_DS$coefficients)[2])
```

There do not seem to be any potential regression outliers, high leverage points, or influential observations. Every observation follows the overall trend of the data, and no point impacts the LSRL drastically.

Regression Model:

```
summary(lm_info_DS)
```

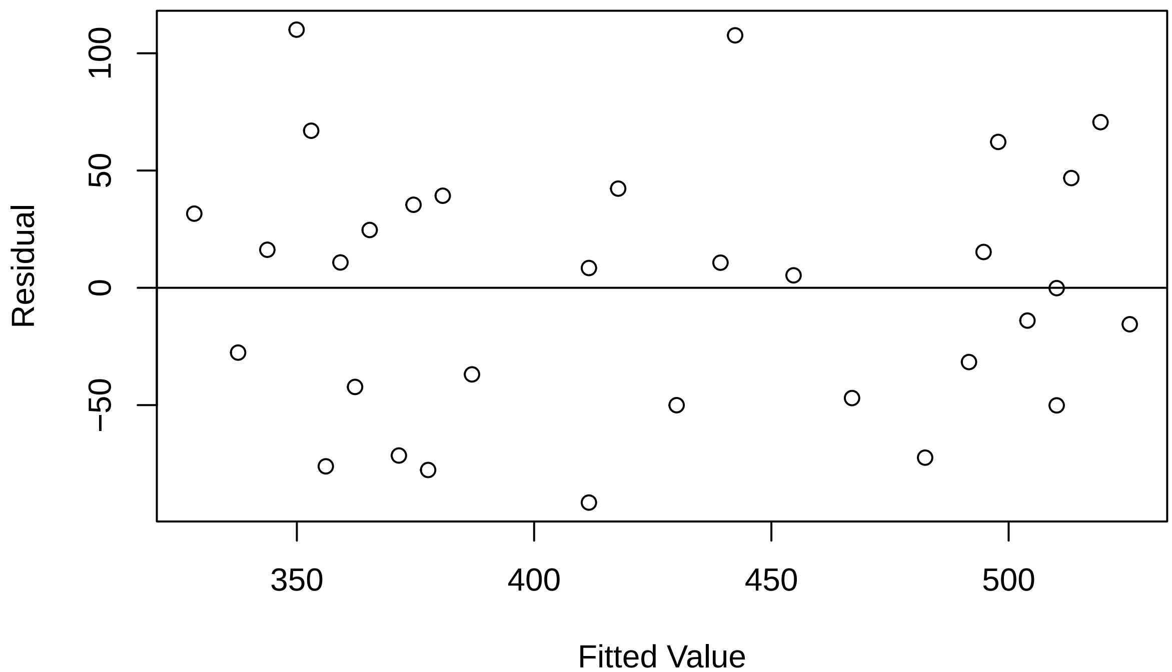
```
##
## Call:
## lm(formula = Distance ~ Age, data = DrivingSight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.548 -43.445   6.889  36.377 110.060
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 580.9719    25.5569   22.73  < 2e-16 ***
## Age          -3.0804     0.4646   -6.63 2.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.59 on 30 degrees of freedom
## Multiple R-squared:  0.5944, Adjusted R-squared:  0.5809
## F-statistic: 43.96 on 1 and 30 DF,  p-value: 2.43e-07
```

Regression Equation: $\hat{y}_i = 580.9719 - 3.0804x_i$
slope: $b_1 = 3.0804x_i$; for every 1 year increase in age, distance decreases by 3.0804 ft

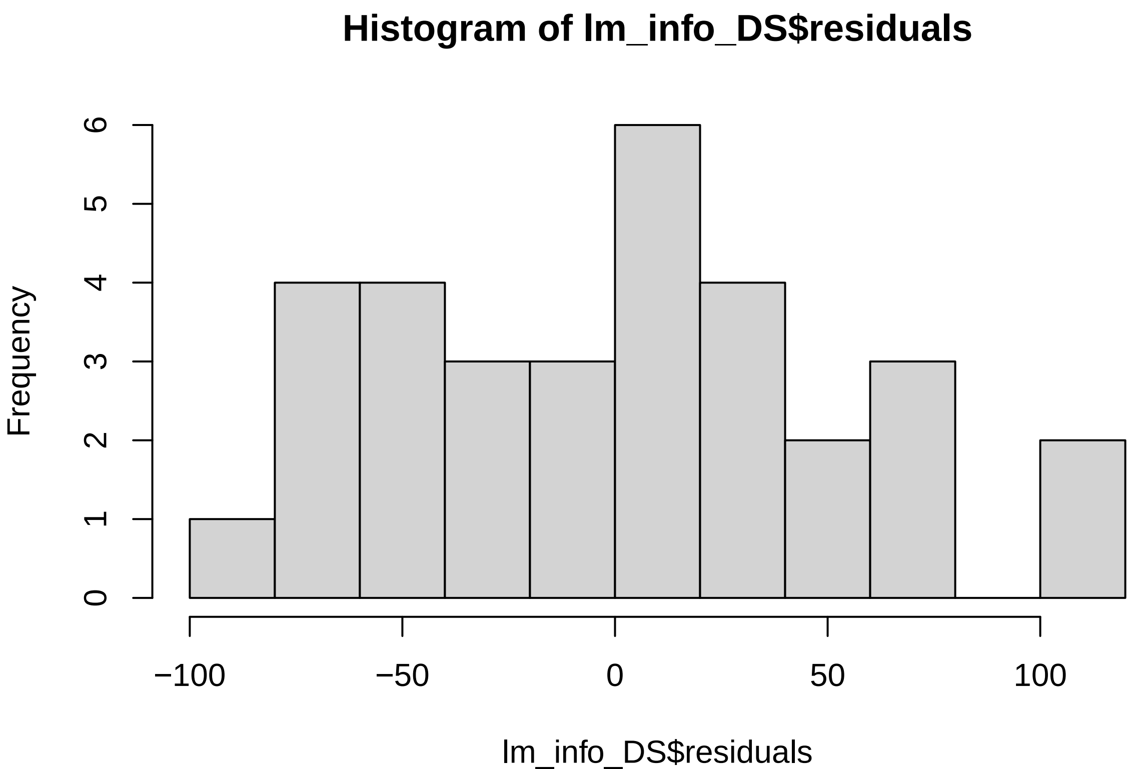
y-int: $b_0 = 580.9719$; not appropriate to interpret; no 0 year olds in the data set
$R^2 = .5944$; About 59.44% of the variation in the maximum distance at which a driver in the dataset can read a highway sign can be explained by the linear relationship between maximum distance and age.
$s = 54.59$; The typical deviation between the maximum distances at which the drivers in the dataset can read a highway sign and the corresponding predicted maximum distances is 54.59 feet.

Plots for Linear Assumptions:

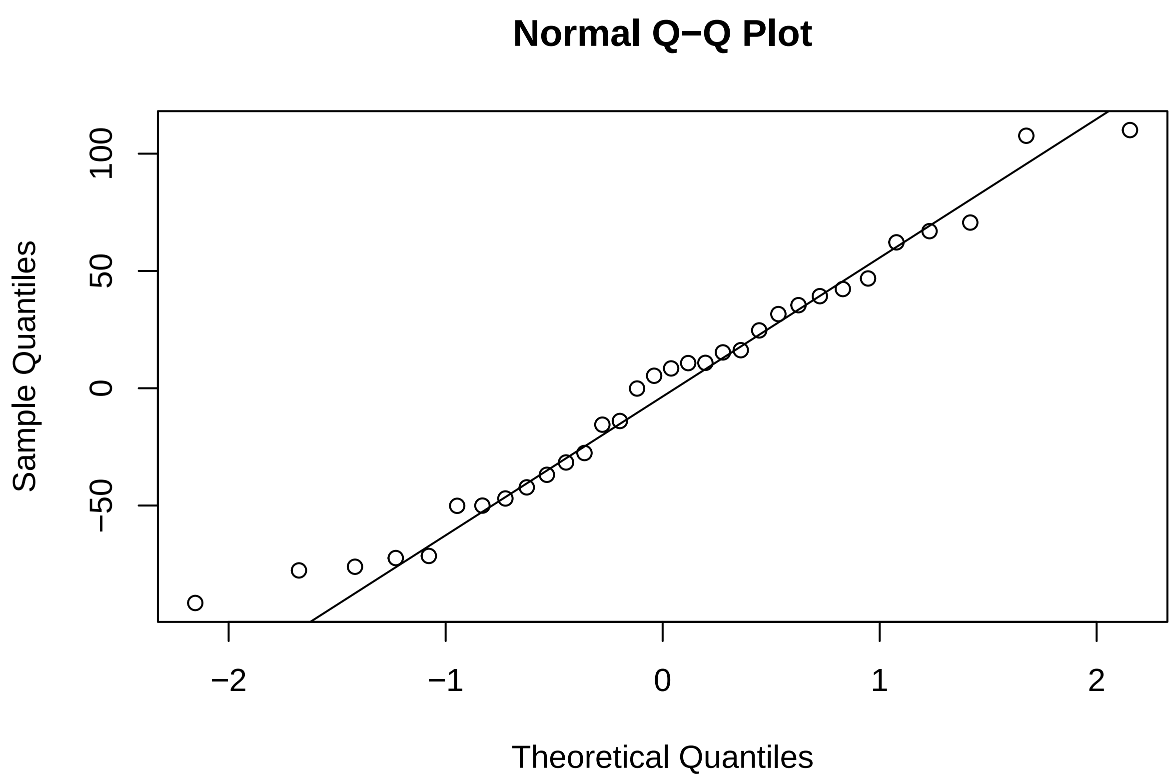
```
#residual plot:
plot(x = lm_info_DS$fitted.values, y = lm_info_DS$residuals,
     xlab = "Fitted Value",
     ylab = "Residual"
)
abline(a = 0, b = 0)
```




```
#histogram of residuals:
hist(lm_info_DS$residuals, breaks = 10)
```




```
qqnorm(lm_info_DS$residuals)
qqline(lm_info_DS$residuals)
```



Hypothesis Test:
H_0: $b_1 = 0$

H_a: $b_1 \neq 0$
α: 0.05
p-value = 2.43 * 10^{-7}
Our p-value of 2.43 * 10^{-7} is less than the α of 0.05, so we reject H_0. There is statistically evidence suggesting a linear relationship between age and max. distance one can read a highway sign.

95% confidence interval for the slope of the relationship between age and distance:

```
confint(lm_info_DS, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 528.777719 633.166119
## Age          -4.029259  -2.131602
```

(-4.02926, -2.1316); 0 is not included, therefore we reject H_0 at this time.
We are 95% confident that as the age of a driver increases by one year, the predicted maximum distance at which the driver can read a highway sign decreases between 2.13 and 4.03 feet.

Predict the average maximum distance for **all** 60 year olds:

```
predict(object=lm_info_DS, newdata=data.frame(Age=60), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 396.1461 282.6138 509.6784
```

We predict the average maximum distance one can read a highway sign for all 60 year olds to be 396.1461 ft.