Rexample4

Lingxiao Zhou

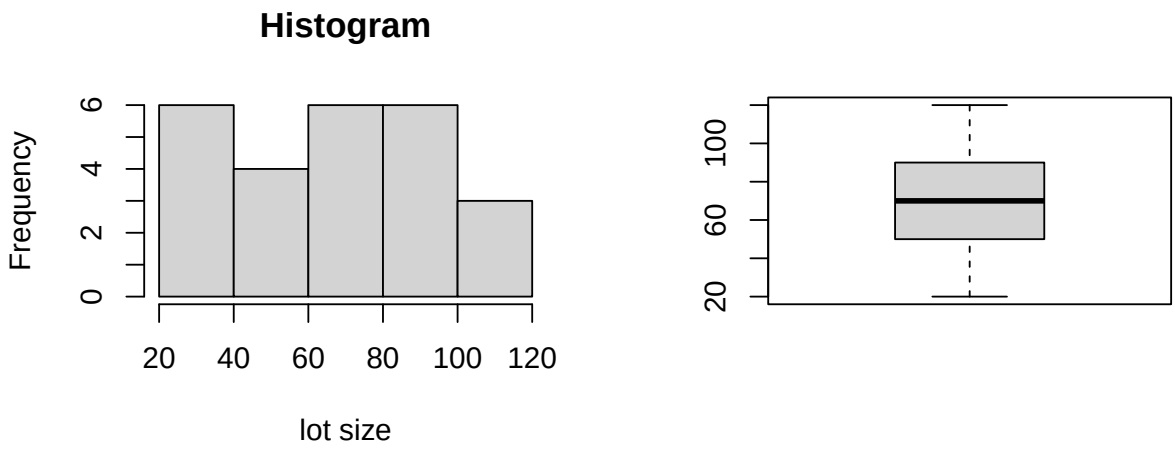
Diagnostic on Predictor

Goal: Identify any outlying values in predictor, X that could affect the appropriateness of the linear model.

Two main issues

- Outliers (Histogram and/or Boxplot)
- Levels of predictor associated with the run order when experiment is run sequentially (Sequence/ Time Series Plot)

```
toluca <- read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData
    col.names = c("lotsize", "workhrs"))
toluca.reg <- lm(workhrs ~ lotsize, data = toluca)


par(mfrow = c(1, 2))
hist(toluca$lotsize, xlab = "lot size", main = "Histogram")
boxplot(toluca$lotsize)
```



- Do not appear to be any outliers

```
plot(toluca$lotsize, type = "b", xlab = "Run order", ylab = "Lot size",
    main = "Sequence Plot")
```

- No discernible pattern/depenency of the values and run order.

Checking Model Assumptions Graphically

Inference are only valid if the model assumptions are valid.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$.

- Normality: The errors are normally distributed
- Independence: The errors are independent of each other.
- Homogeneity of Variance (Homoscedasticity): The variance of the residuals is the same for all values of X.
- Linearity: The relationship between the independent variable (x) and the mean of the dependent variable (Y) is linear.

Residuals

- Assumptions can be checked with residuals $e_i := Y_i - \hat{Y}_i$, or better yet, the standardized/studentized residuals
- Standardized residuals have mean 0 and standard deviation of 1
- Studentized residuals are standardized residuals that use the leave-one-out approach. The $i-$th observation is omitted when fitting thr model is then used to predict (without bias) the response for the $i-$th observation. More on this later
- In R, use `rstandard()` and `rstudent()` function to obtain the standardized and studentized residual.

Normality

Empirical CDF

Def: CDF associated with the smpirical measuare of the sample. Assigns equal probability $1/n$ to each point and is a step function

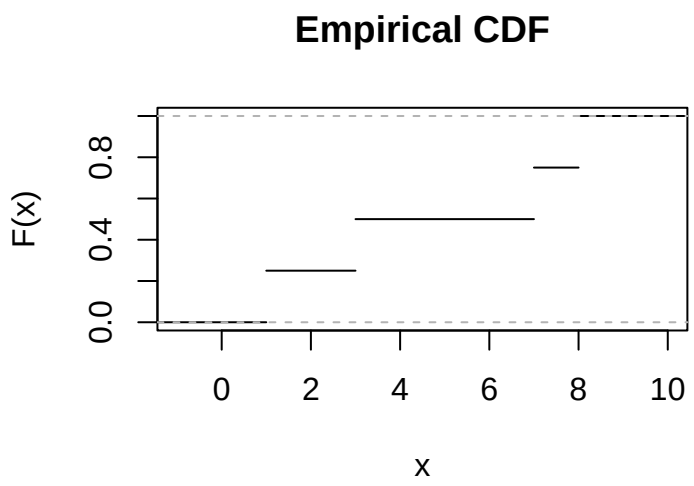
$$\hat{F}_n(x) = \frac{\text{number of elements} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} I\{x_i \leq x\}$$

where $I(\cdot)$ is the indicator function.

Example: consider the sample $1, 3, 7, 8$

$$\hat{F}_4(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.25 & \text{if } 1 \leq x < 3 \\ 0.5 & \text{if } 3 \leq x < 7 \\ 0.75 & \text{if } 7 \leq x < 8 \\ 1 & \text{if } x \geq 8 \end{cases}$$

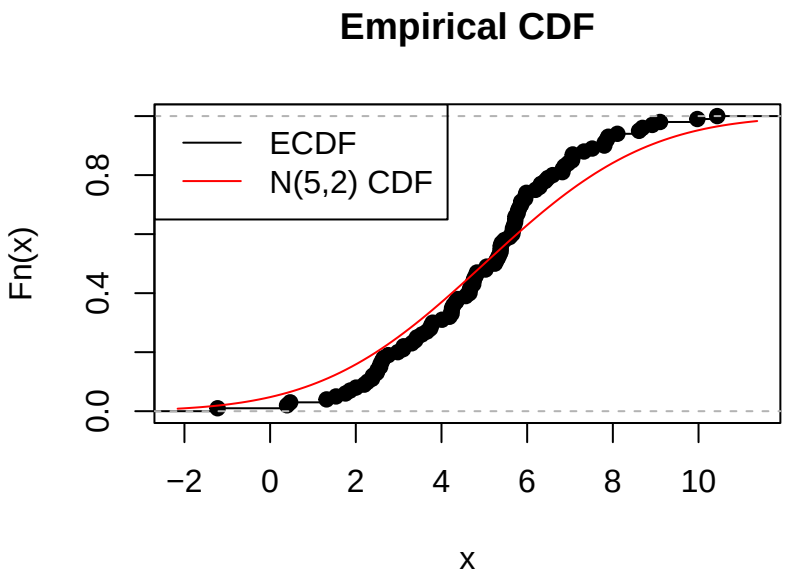
```
sample_data <- c(1, 3, 7, 8)
edf <- ecdf(sample_data)

# Plot the empirical distribution function
plot(edf, main = "Empirical CDF", xlab = "x", ylab = "F(x)",
    verticals = FALSE, do.points = FALSE)
```



- The more data points the smaller the jumps and the step function begins to look like an "S" curve

Example: Generate 100 observations from a N(5,2) and plot the empirical CDF and overlay the theoretical CDF from N(5,2)

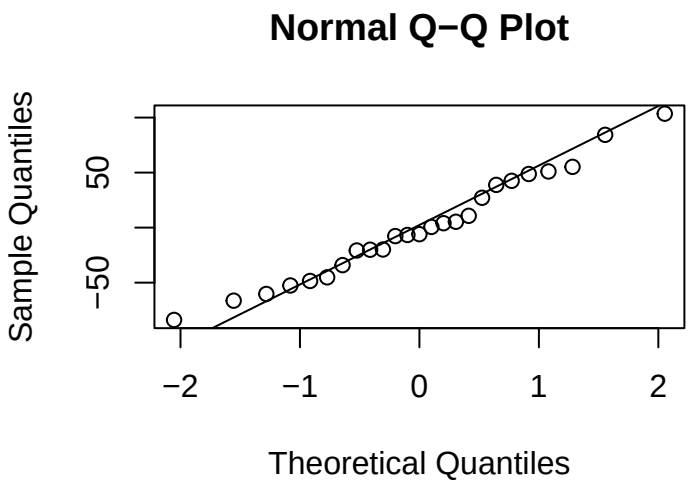
```r
set.seed(111)
y <- rnorm(100, 5, 2)
edf <- ecdf(y)
plot(edf, main = "Empirical CDF")
curve(pnorm(x, 5, 3), add = TRUE, col = "red")
legend("topleft", legend = c("ECDF", "N(5,2) CDF"), lty = 1,
    col = c("black", "red"))
```



Q-Q plot

Let $G(\cdot)$ denotes the theoretical normal CDF
Q-Q plot plots the quantile function $\hat{F}_n^{-1}(x)$ versus $G^{-1}(x)$

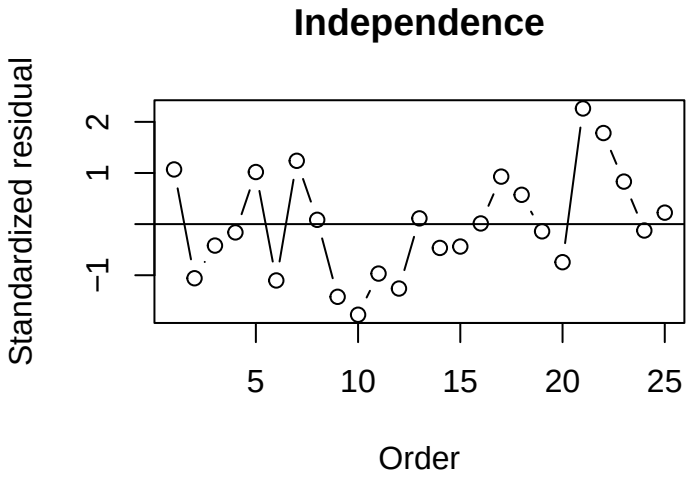
```r
qqnorm(toluca.reg$residuals)
qqline(toluca.reg$residuals)
```



Independence

- Time series plot of residual vs time order in which it was recorded
- Sorted data may invalidate conclusion
- Independence is graphically check if there is no discernible pattern

```
plot(rstandard(toluca.reg), type = "b", xlab = "Order", ylab = "Standardized residual",
    main = "Independence")
abline(h = 0)
```



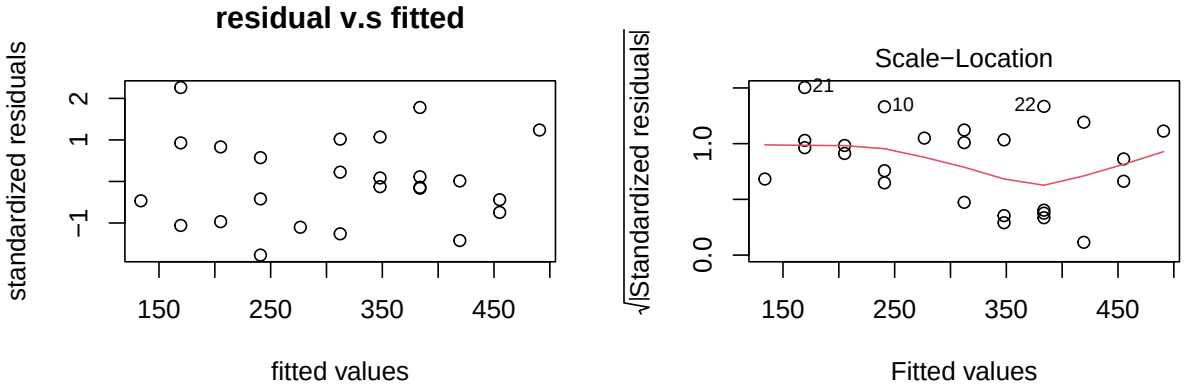
Homogeneity of variance

- Plot of the raw/standardized/studentized residuals versus the fitted values

  - Constant variance: constant spread/distance of the residuals (no funnel shape)

- Scale-Location Plot: The square root of the absolute standardized residuals is plotted against the fitted values
  - Constant variance: the red line is roughly horizontal across the plot and the spread of the points are roughly equal at all fitted values

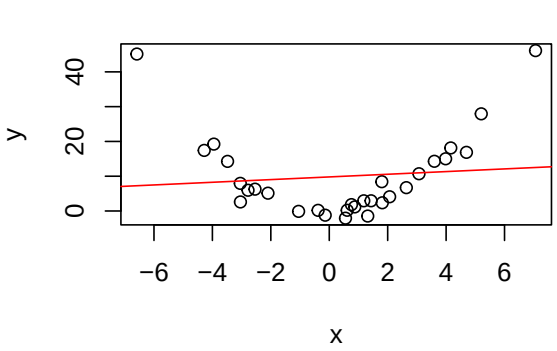
```
par(mfrow = c(1, 2))
plot(toluca.reg$fitted.values, rstandard(toluca.reg), xlab = "fitted values",
    ylab = "standardized residuals", main = "residual v.s fitted")

# plot(toluca.reg) generate several diagnostic plot use
# which = 3 for only generating scale-location plot
plot(toluca.reg, which = 3)
```

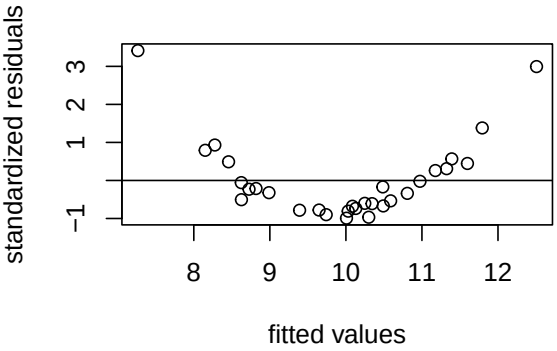


Linearity

- Scatter plot: For good model fit, scatterplot indicate a linear trend.
- Residual vs fitted: For good model fit, residuals should evenly spread on either side of the 0 line.

```
par(mfrow = c(1, 2))

# generate data from a nonlinear model y = x^2+epsilon
x <- rnorm(30, 0, 3)
eps <- rnorm(30, 0, 2)
y <- x^2 + eps
plot(x, y)
fit <- lm(y ~ x)
abline(fit, col = "red")
plot(fit$fitted.values, rstandard(fit), xlab = "fitted values",
    ylab = "standardized residuals")
abline(h = 0)
```





- According to the plots, linearity assumption does not hold.
- We may still fit a linear model by adding or transforming predictors to include higer polynomial terms. For example, $Y_i = \beta_0 + \beta_1 x^2 + \epsilon_i$