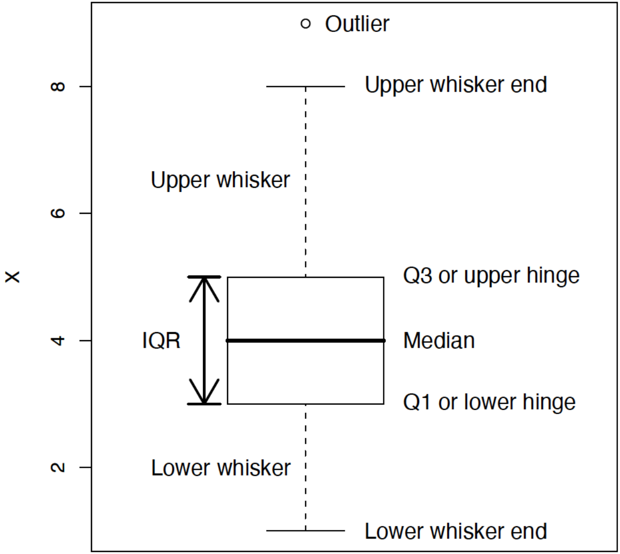
Open Review Session

Roberta De Vito

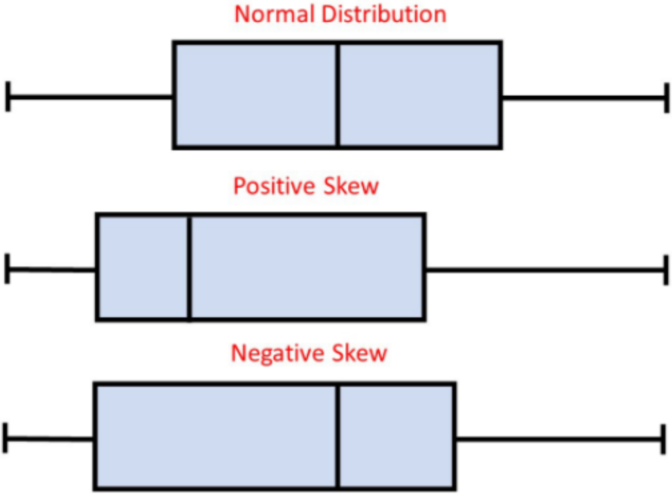
BROWN
Public Health

1 Asymmetry of a distribution

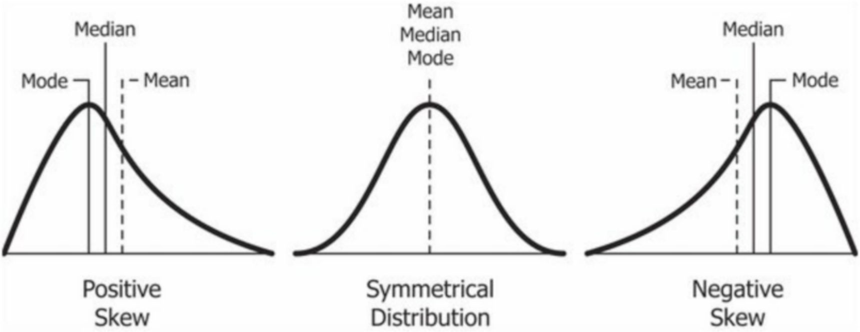
Boxplots


Outlier
Upper whisker end
Upper whisker
Q3 or upper hinge
IQR
Median
Q1 or lower hinge
Lower whisker
Lower whisker end
8
6
4
2
x

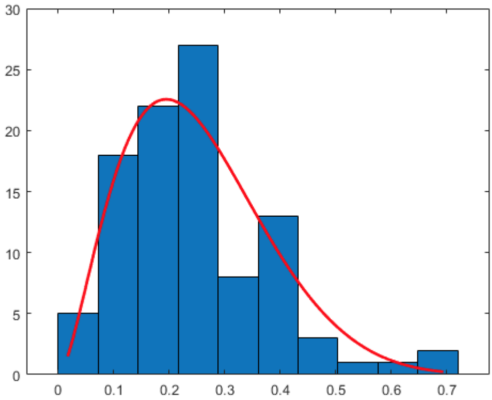
Boxplots interpretation



Mean Median and Mode



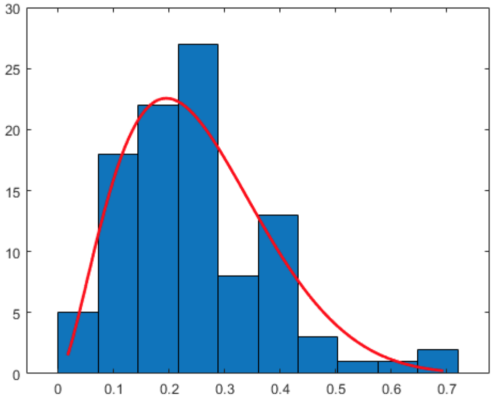
Questions



Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Questions

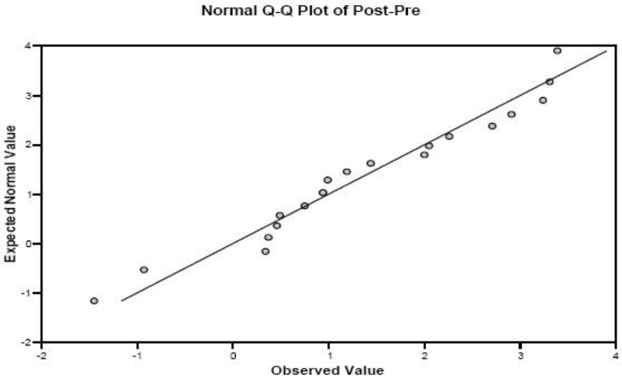


Can you write down the order of the mode, mean and median?

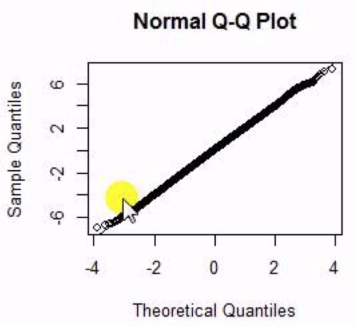
What is a QQ plot?

1. Let $\varepsilon_{(1)}, \ldots, \varepsilon_{(n)}$ be the ordered residuals with $\varepsilon_{(1)} \leq \varepsilon_{(2)} \leq \ldots \leq \varepsilon_{(n)}$.
2. Assume the ε are standardized by subtracting mean and dividing by standard error. This ensures that they have mean zero and variance one. Then, the distribution to compare to is a $\mathcal{N}(0, 1)$.
3. If the ε-s come from a $\mathcal{N}(0, 1)$ distribution, we expect $\varepsilon_{(k)}$ to be approximately equal to the $\frac{k}{n}$-th quantile of the $\mathcal{N}(0, 1)$.
4. A qq-plot plots the observed quantiles vs the theoretical quantiles. If points fall on a straight line, indication of the sample coming from a normal distribution.

QQ plot in practice



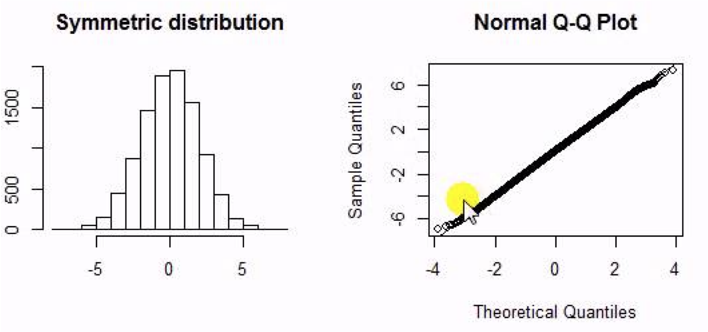
Question I QQplot



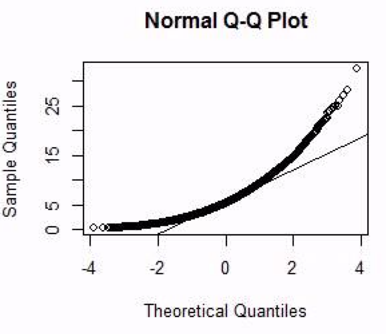
Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Question I QQplot



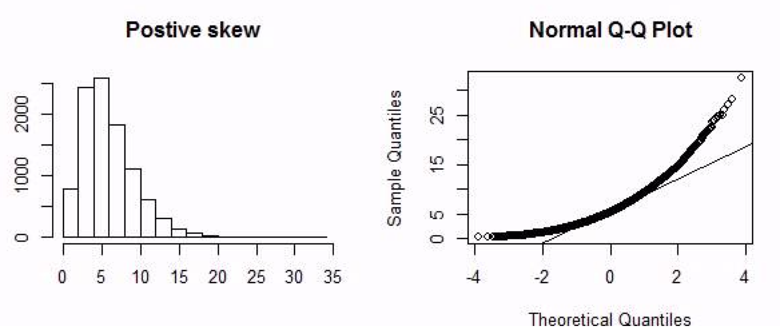
Question II QQplot



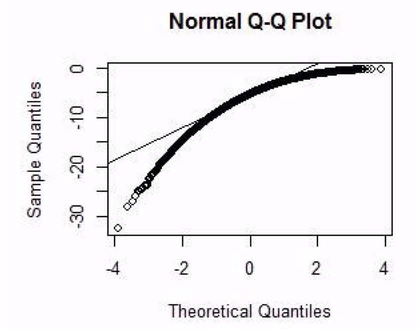
Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Question II QQplot



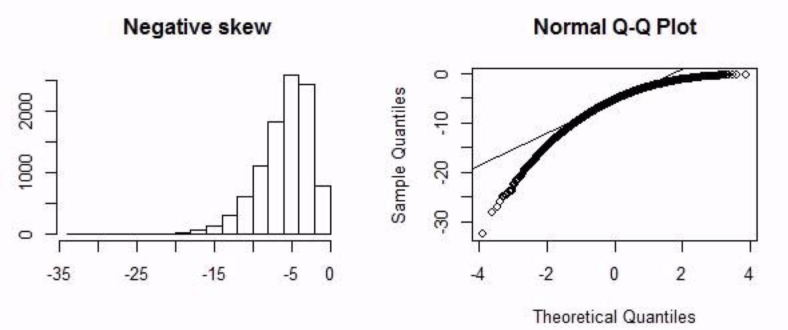
Question III QQplot



Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Question III QQplot



2 Linear regression model

The linear regression

- $y_i = f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$
- $f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

$$\Downarrow$$

Matrix form: $Y = X\beta + \epsilon$

Model Assumption

1. $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$
2. $\epsilon \sim N(0, \sigma^2)$
3. Error term is independent of (uncorrelated with) covariate(s)

$$Corr(X, \epsilon) = 0$$

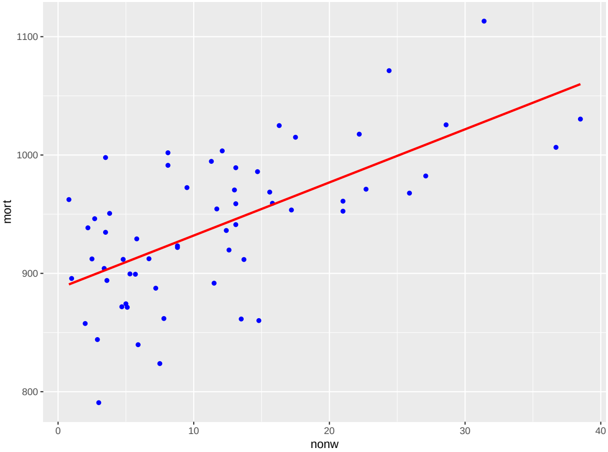
4. Variance of error term is same, regardless of value of x (homoscedasticity)

$$Var(\epsilon) = \sigma^2$$

Example: Pollution data set

`ID`	code for the identification of the sample
`OVR65`	% of 1960 SMSA population aged 65 or older
`EDUC`	Median school years completed by those over 22
`HOUS`	% of housing units with all facilities
`DENS`	Population per sq. mile in urbanized areas, 1960
`NONW`	% non-white population in urbanized areas, 1960
`WWDRK`	% employed in white collar occupations
`POOR`	% of families with income < 3000
`HC`	Relative hydrocarbon pollution potential
`NOX`	Same for nitric oxides
`SO2`	Same for sulphur dioxide
`HUMID`	Annual average % relative humidity at 1pm
`MORT`	Total age-adjusted mortality rate per 100,000
`PREC`	Average annual precipitation in inches

How do we find regression line that fits best?



Example: Pollution data set

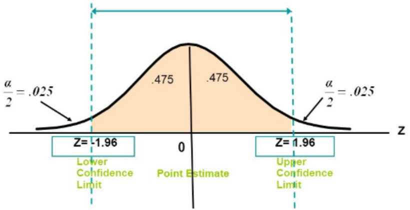
`ID`	code for the identification of the sample
`OVR65`	% of 1960 SMSA population aged 65 or older
`EDUC`	Median school years completed by those over 22
`HOUS`	% of housing units with all facilities
`DENS`	Population per sq. mile in urbanized areas, 1960
`NONW`	% non-white population in urbanized areas, 1960
`WWDRK`	% employed in white collar occupations
`POOR`	% of families with income < 3000
`HC`	Relative hydrocarbon pollution potential
`NOX`	Same for nitric oxides
`SO2`	Same for sulphur dioxide
`HUMID`	Annual average % relative humidity at 1pm
`MORT`	Total age-adjusted mortality rate per 100,000
`PREC`	Average annual precipitation in inches

The lm function in R: what are we looking?

```
Call:
lm(formula = mort ~ nonw + so2 + educ + nonw)

Residuals:
    Min      1Q  Median      3Q     Max
-94.201 -19.410   1.294  16.537  92.986

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1156.06487   71.68018  16.128  < 2e-16 ***
nonw           3.70485    0.58615   6.321 4.55e-08 ***
so2            0.25699    0.08298   3.097 0.003054 **
educ         -24.92413    6.28208  -3.967 0.000209 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.02 on 56 degrees of freedom
Multiple R-squared:  0.6266,  Adjusted R-squared:  0.6066
F-statistic: 31.33 on 3 and 56 DF,  p-value: 5.063e-12
```

Inference

- $H_0 : \beta_1 = 0$
- 95% confidence intervals



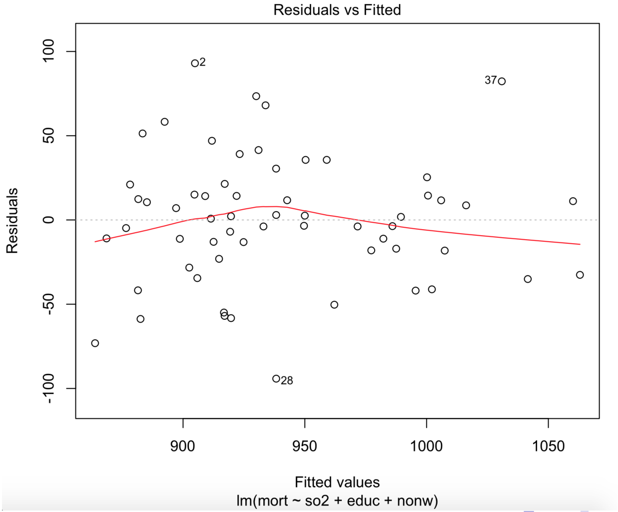
- R^2
- F-statistics: Does the model fit better than a model with only an intercept?

Diagnostics and Assumption Checking

1. is the linear relationship a good assumption?
2. is the error term variance constant?
3. are the error term normally distributed?
4. are there any outliers?
5. do we repeat some information?

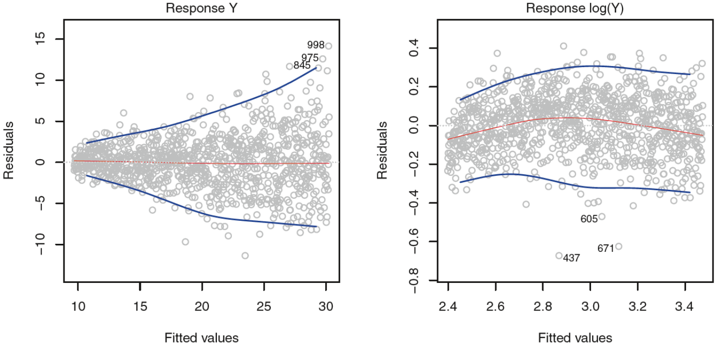
1. Non-linearity of the data

Residual plot of fitted values vs. residuals should
- have no discernible pattern
- be scattered evenly around 0

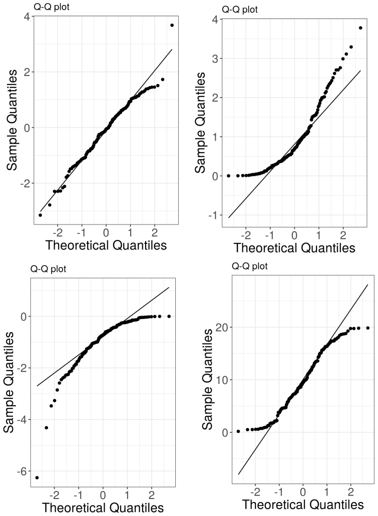


2. Non-constant Variance of Error Terms: Heteroscedasticity

- Patterns might indicate wrong form of model variable
- Funnel shape in the residual plot: transform Y

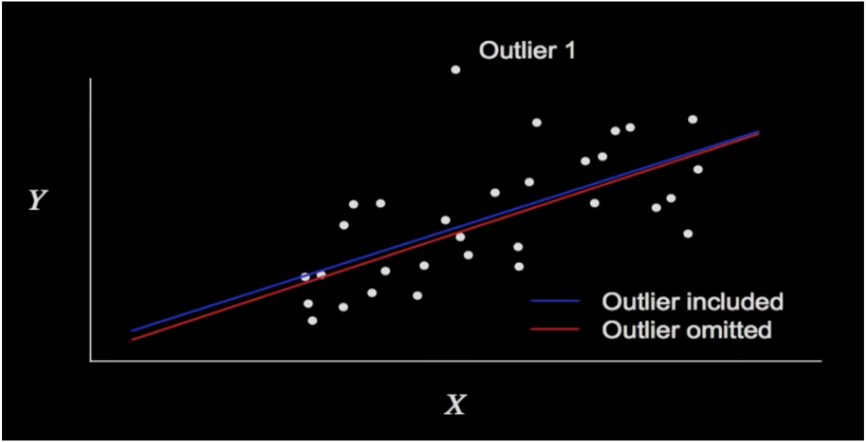


3. Normal distribution



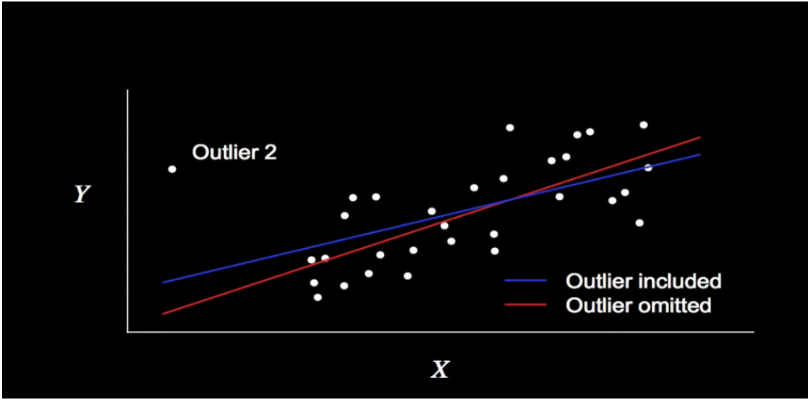
4. Outliers

Outliers for Y



4. Outliers

Outliers for X (High Leverage)



5. Collinearity

- Collinearity refers to when the predictors are highly correlated.
- Repetition of information
- Leads to increased standard errors of the regression coefficients $\rightarrow$ fail to reject $H_0 : \beta_j = 0$
- take a look at the correlation of two covaiates