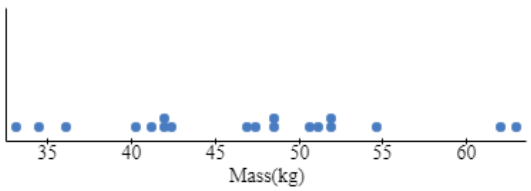
Regression Diagnostics: Motivation

- (3.1-3.7) We wish to assess the appropriateness of a model for a set data.
  -
  -
  - Diagnostic methods:

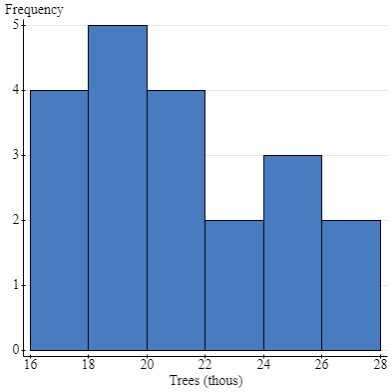
DIAGNOSTICS FOR PREDICTOR VARIABLE

Considerations

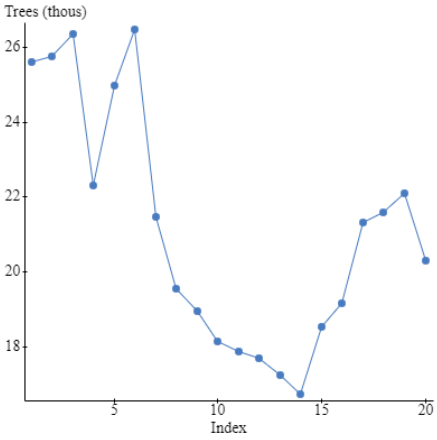
- Are there any outliers (influential or otherwise) in X values?
- Important for determining the scope of the model
- Plots to use:
  - Dot plot



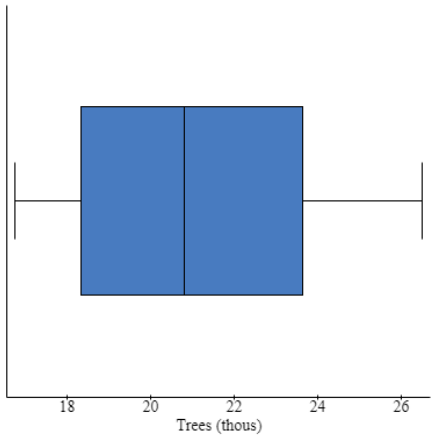
  - Histogram or Stem-and-Leaf Plot



  - Sequence/index plot



  - Box plot



RESIDUAL ANALYSIS

Diagnostics for response variable Y are generally done indirectly ***via the residuals***.

- Recall: $\boldsymbol{e_i} =$ is the **observed error**, called the **residual** for the ith observation.
- A residual is the difference between the observed Y value and predicted Y value.
- There is a residual for each data point in the sample.
- Graphically, a residual measures the vertical distance between a data point and the regression line.
- Residuals must be distinguished from the **true unknown error** terms $\boldsymbol{\varepsilon_i} =$ **.**
    - We are still assuming a Normal error regression model –
    - If the SL regression model is appropriate for a data set, the residuals should also reflect these properties!

Properties of Residuals

- **CENTER (MEAN):** $\bar{e} =$
- **SPREAD (VARIANCE):** $s^2 =$
- So, the distribution of the residuals has a mean of 0 and a standard deviation of $\sqrt{MSE}$, an unbiased estimator of σ.
- **INDEPENDENCE:** $e_i$are ***dependent*** random variables because they are based on fitted values,$\hat{Y}_i$, from the same estimated/fitted regression function and are subject to two constraints:
-
-

(The dependence is not an issue with sufficiently large samples.)

Standardized Form of Residuals

$\boldsymbol{e_i^*} =$ is called a ***semistudentized*** residual.

- This uses the approximated (not estimated) SD of the residuals and thus cannot be called the studentized residual. (R will calculate studentized residuals.)
- Can be very useful in identifying outliers!

Assumptions of Regression (L.I.N.E)

- Linearity
    -
- Independence of Errors
    -
- Normality of Error
    -
- Equal Variance (also called **homoscedasticity**)
    -

Checking for Violations in the Regression Assumptions and Other Issues

What to check for:

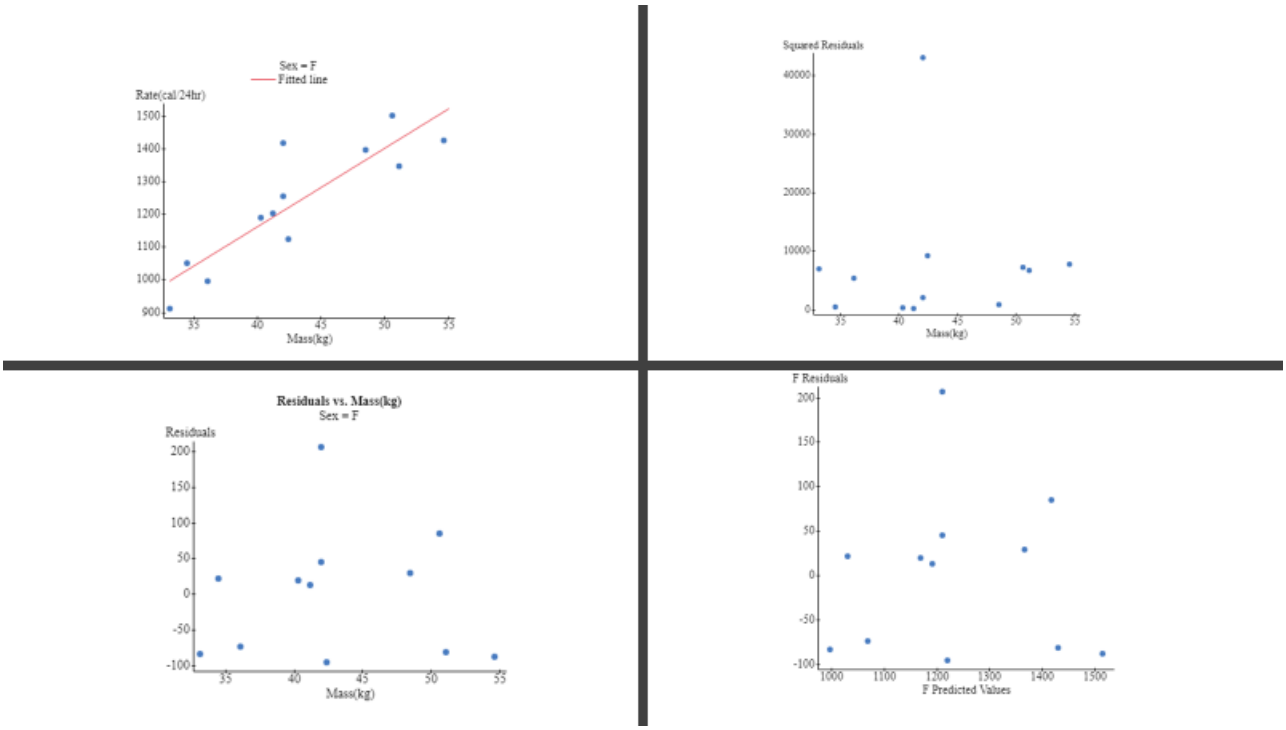
1. The appropriate regression function is not linear.
2. The variance of error terms is not constant.
3. Error terms are dependent.
4. Model fits most values, with the exception of a few outliers.
5. Error terms do not follow a Normal distribution.
6. Other predictor variables have been omitted.

GRAPHICAL DIAGNOSTICS: RESIDUAL PLOTS

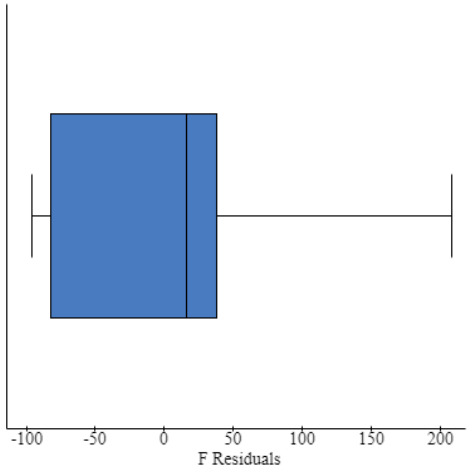
Types of Residual Plots

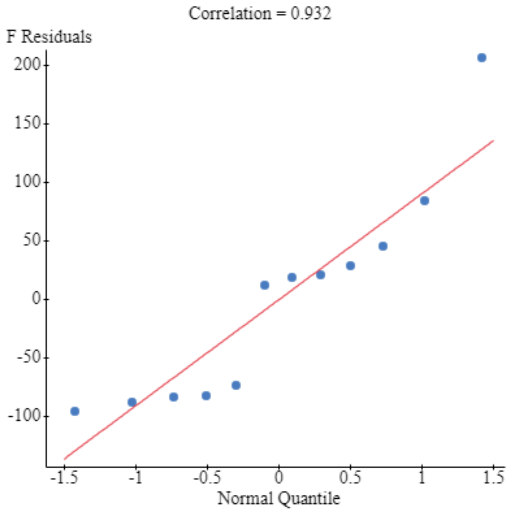
Let's start with:

- Residual vs X plots or, equivalently,
- Residual vs $\hat{Y}$ values
- Absolute value or square of residuals vs. X



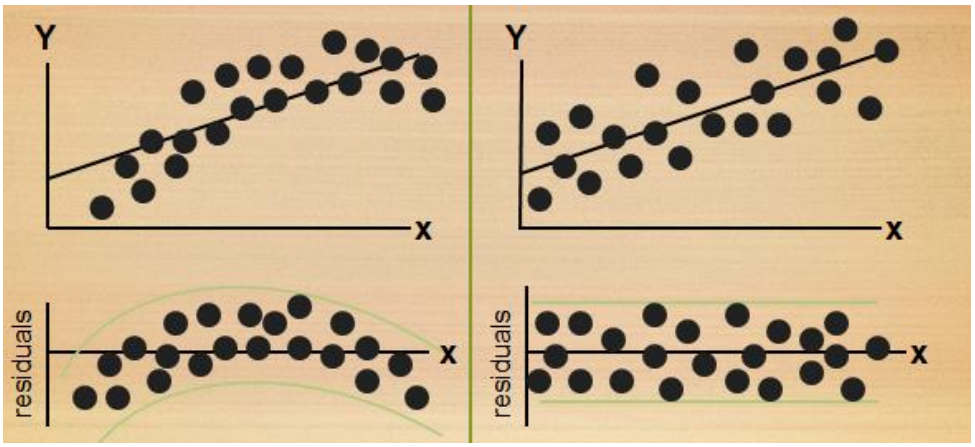
Other residual plots: Boxplot (or histogram) of residuals & QQ (Normal quantile or Normal probability) plots

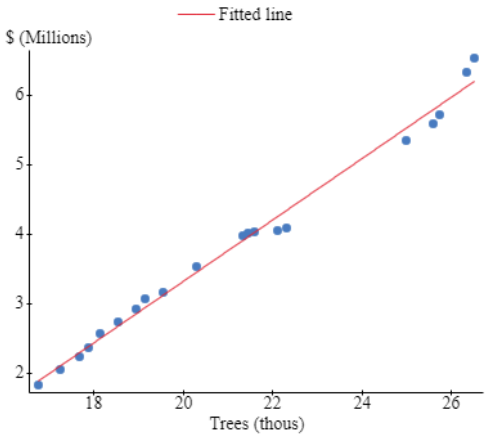


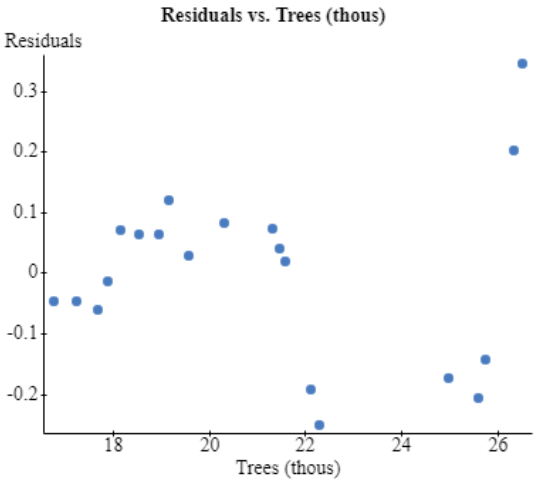


Checking for Nonlinearity: Is a linear regression function appropriate?

- **Best Plots to Use:**
- **What to look for?**
  - If a **linear regression function** is **appropriate**:
    -
    -
  - If a **linear regression function** is **inappropriate** (either no fitting well or not the best choice):
    -
      - For example,

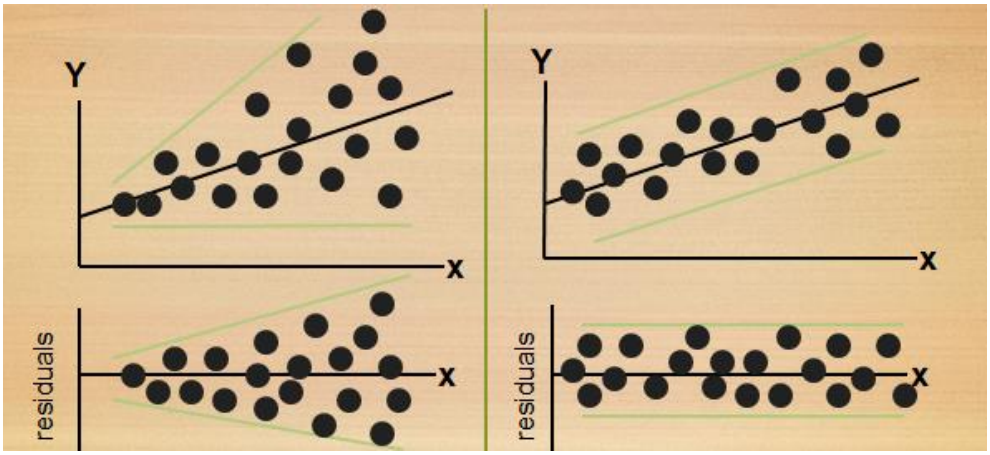




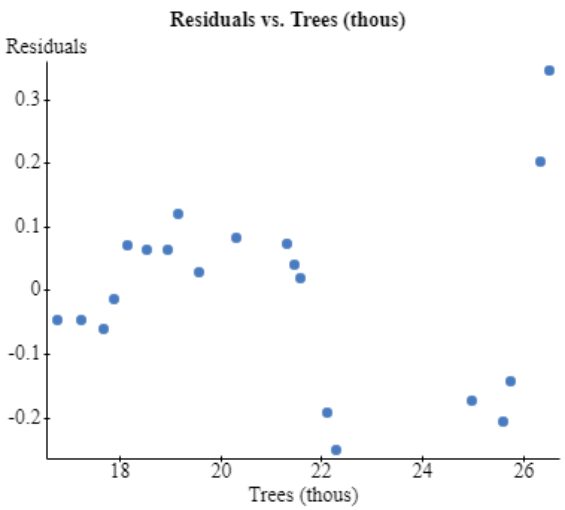


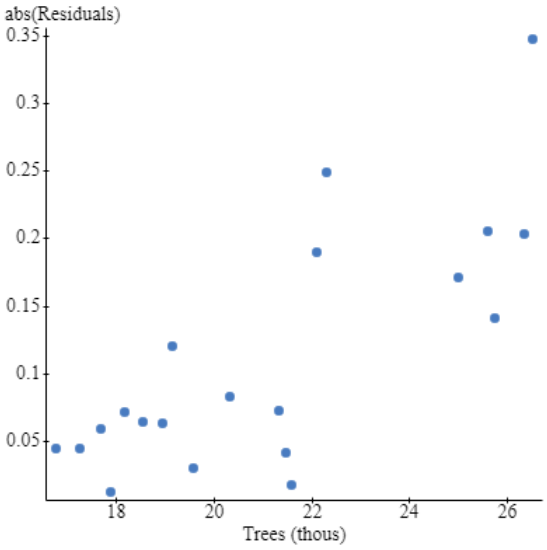
<u>Checking for Nonconstancy of Error Variance:</u> Can we assume $\sigma^2(\varepsilon_i) = \sigma^2$ or not?

- **Best Plots to Use:**
- If $\boldsymbol{\sigma^2(\varepsilon_i)}$ is <u>not</u> constant:
  - **Residual vs X plots (or $\widehat{Y}$)**
    -



  - The **absolute value or square of residuals vs. X (or $\widehat{Y}$)** is also helpful
    -



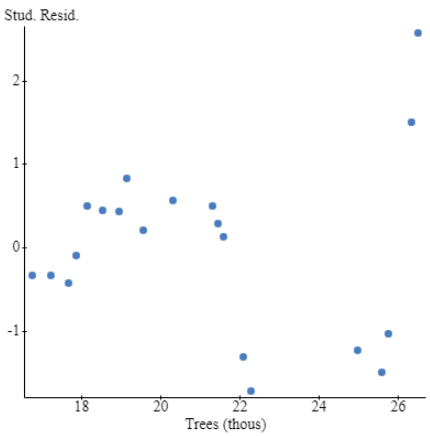


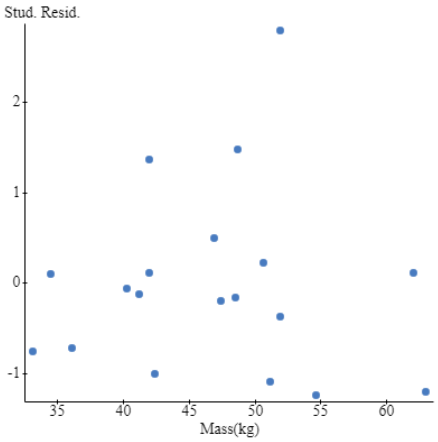
Checking for Presence of Outliers (Extreme Observations)

- **Plots to Start With:**

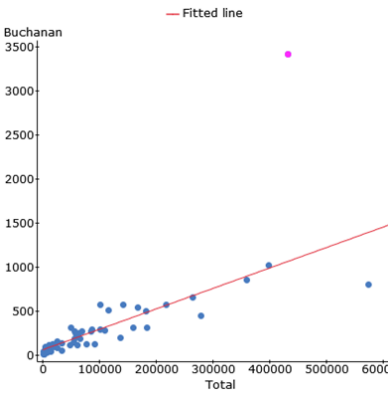
❖ **Best Plots to Use:**

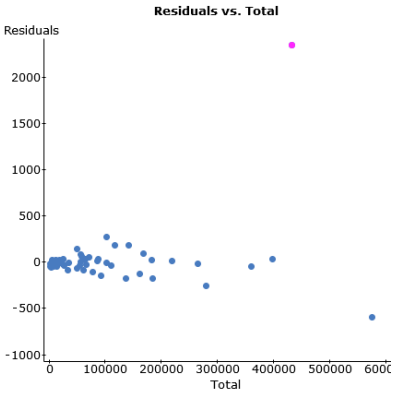
- This allows us to examine: Are there any observations that lie far away (in terms of SDs) from zero (the center)?
    - For **large samples**:

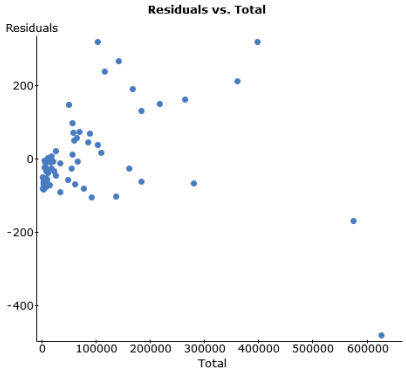




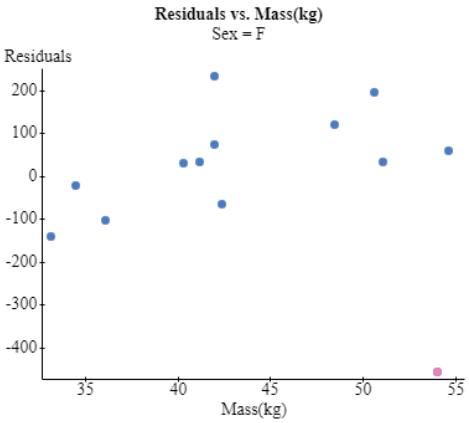
- Values identified as outliers should not necessarily be removed.
    - LSR (least-squares regression) seeks to minimize the sum of squared residuals.
        -
            - Only remove the outlier if you are sure that the value is due to some type of error.

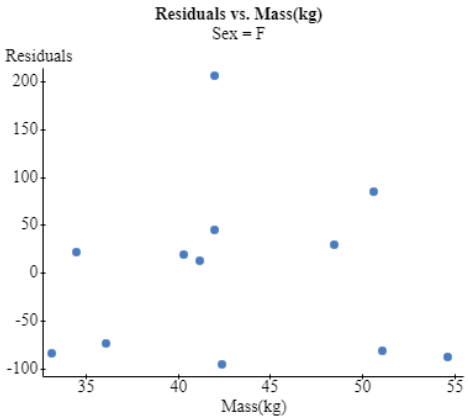






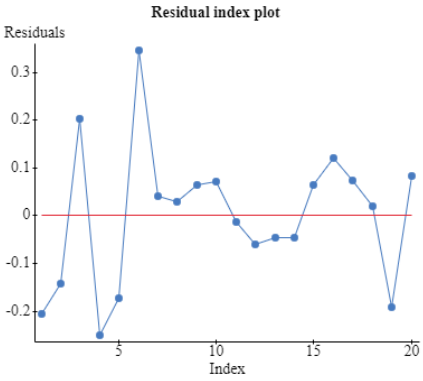
- **Outliers** can have a **huge effect** on data sets with **small sample sizes**.
  -
  -
  -

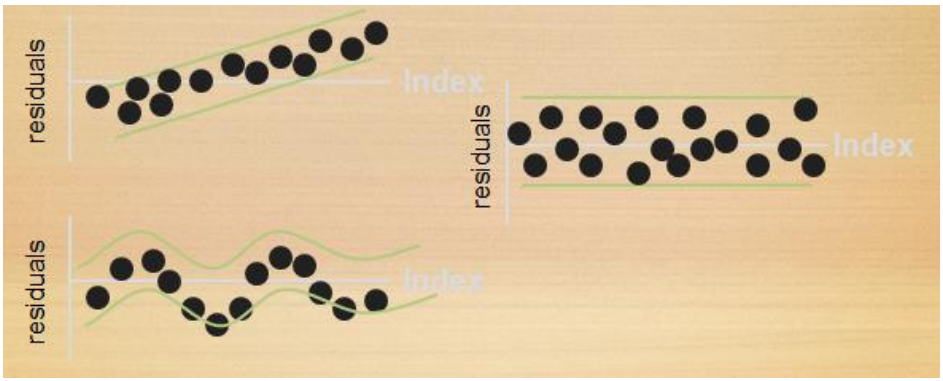




Checking for Dependence of Error Terms

- This often occurs due to an omitted variable, such as time.
  - If **data is collected in sequence** (based on time or something else), use a **sequence plot (residual index plot).**
    - **Ideal:** If ε_i are independent,
    -
    -





Checking for Nonnormality of Error Terms

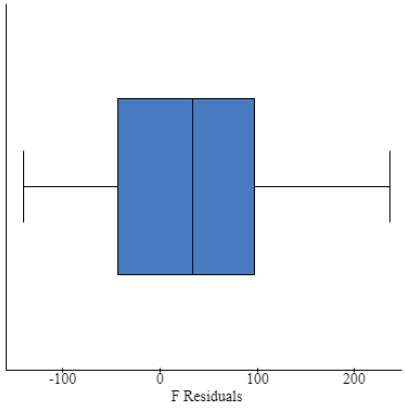
In a Normal error regression model $\boldsymbol{\varepsilon_i}$ terms are assumed to follow a **Normal** distribution ($\boldsymbol{\varepsilon_i \sim N(0, \sigma)}$) .

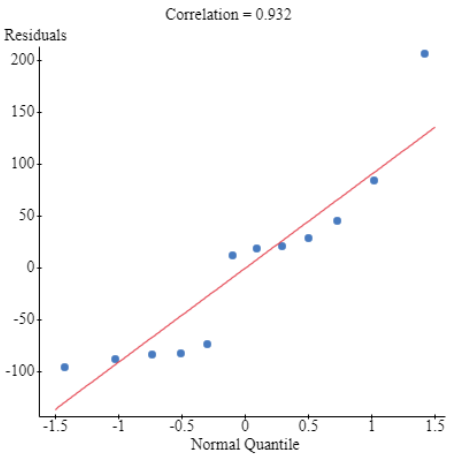
- As discussed previously, small departures from normality are generally fine, but large/clear departures are worrisome.

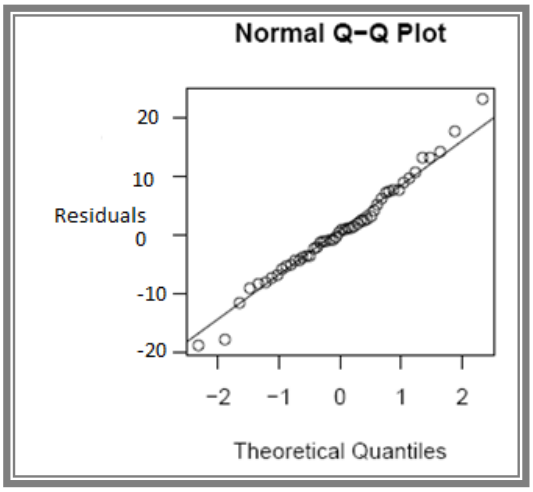
Best plots to check normality:

-

(If n is sufficiently large, we can also look at a histogram.)



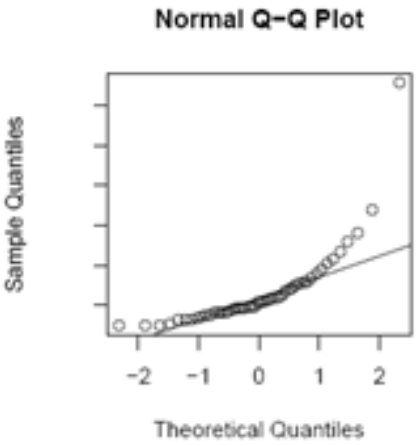


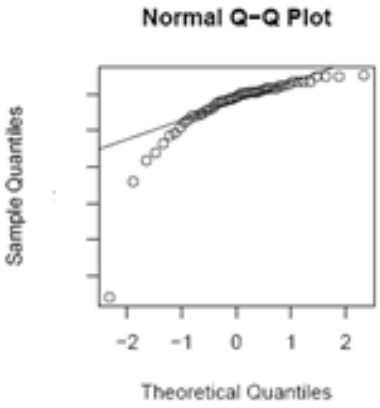


-

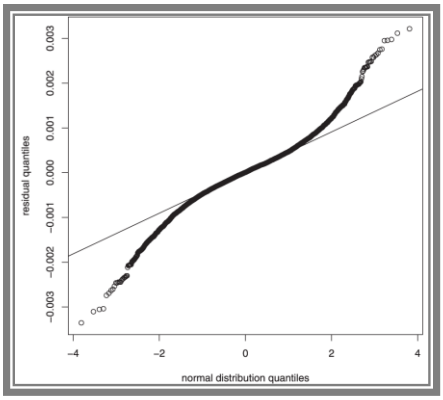
  - Compares residual frequencies to expected frequencies based on a Normal distribution (using percentiles and estimated SD).
  - A clearly nonlinear pattern suggests

- A curved pattern suggests a skewed distribution.





- A heavy-tailed distribution is symmetric and bell-shaped, but has more values in the tails than a Normal distribution (a t-distribution is a good example).

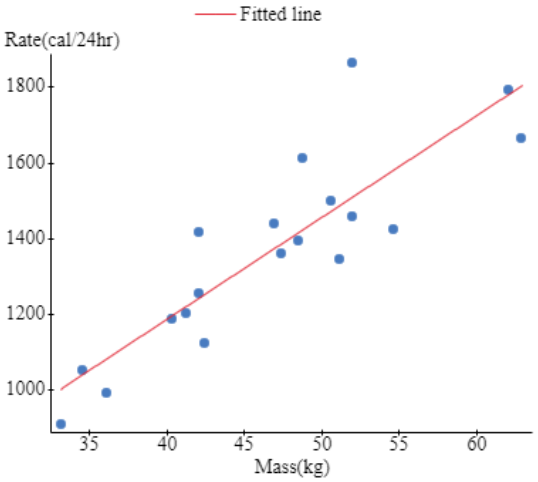


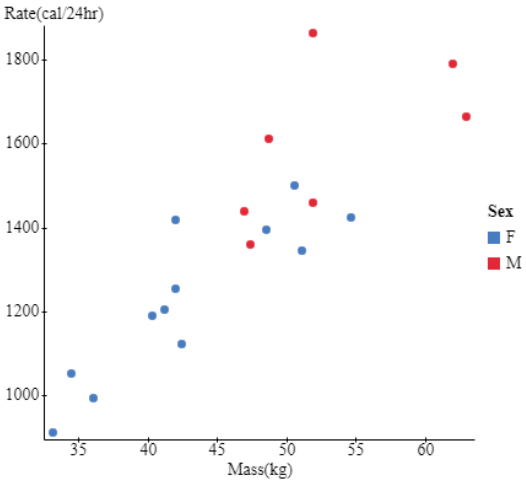
Difficulties

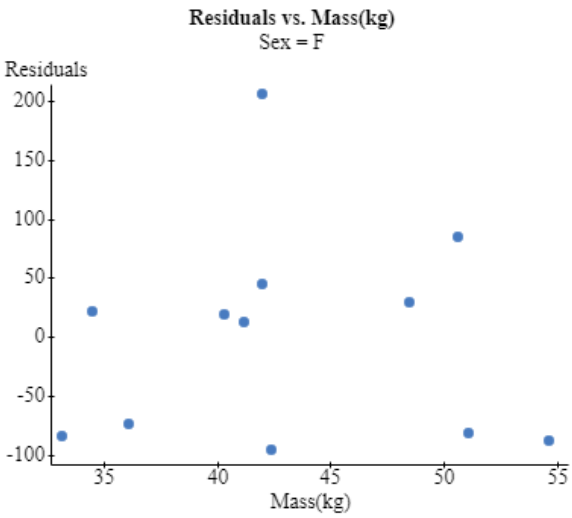
- Be careful in assuming nonnormality with small sample sizes.
- Nonnormality may also be due to an inappropriate regression model or nonconstant variance of ε_i.
  - Check for these violations first!

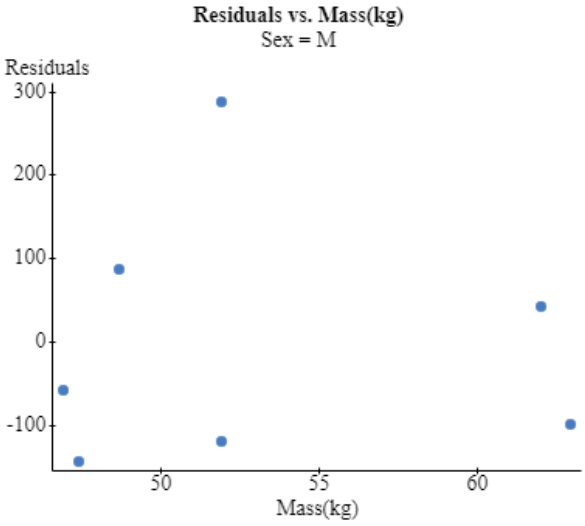
Omission of Important Predictor Variables

-
  -
-



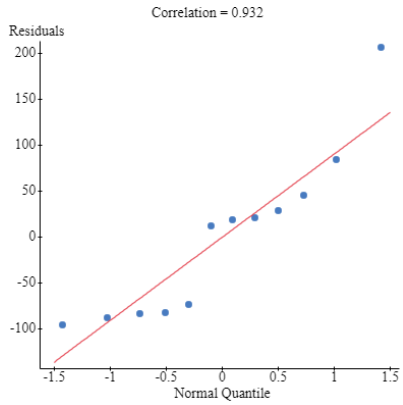






TESTS INVOLVING RESIDUALS: CORRELATION TEST FOR NORMALITY

- Start by looking at the QQ plot of the residuals.
- Calculate the correlation between the residuals, e_i, and their expected values on a Normal distribution.
    - Is the correlation close enough to 1 for ε_i to be considered Normal?



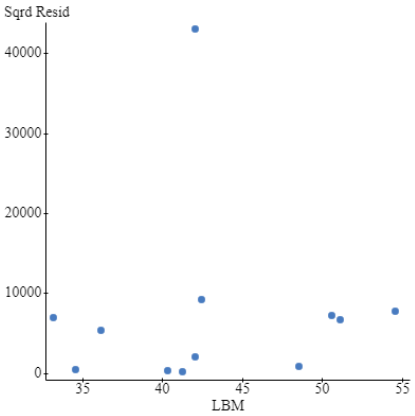
- Use **Shapiro-Wilk test** in **R**.
    - The test-statistic $\boldsymbol{W}$ is approximated based on the correlation value, measuring how close the graph (QQ plot) is to a straight line. It can be thought of as approximately based on studentized residuals and expected values assuming Normality.
    - $\boldsymbol{H_0}$:
-

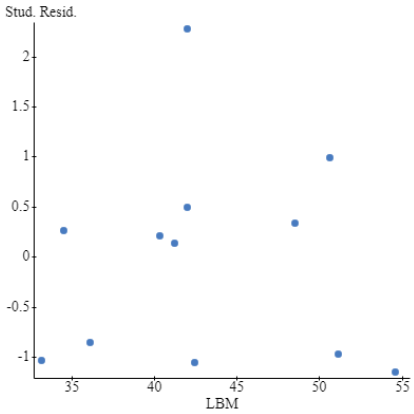
shapiro.test(rstudent(reg))

Shapiro-Wilk normality test; data: rstudent(reg); W = 0.83199, p-value = 0.02216

TESTS INVOLVING RESIDUALS: TEST FOR CONSTANCY OF ERROR VARIANCE (HOMOSCEDASTICITY)

- Start by plotting the squared residuals ${e_i}^2$ vs. X (and semistudentized or studentized residuals vs X).

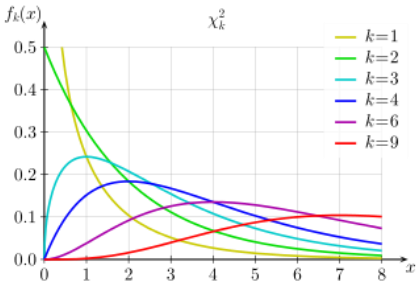




- **Breusch-Pagan Test** (ncvTest () in *"car"* package in **R**)
  - For sufficiently large sample sizes.
  - Assumes ε_i are independent and Normally distributed.
    - $\sigma^2(\varepsilon_i) = \sigma^2$ is related to the level of X by a natural logarithmic function.
      - $ln(\sigma^2) =$

- **Breusch-Pagan Test** (ncvTest () in *"car"* package in **R**)
  - Hypotheses:
    - $\boldsymbol{H_0}$:
    - $\boldsymbol{H_A}$:
  - Carried out by regressing ${e_i}^2$ on X. The test statistic is a Chi-square statistic:

$\chi^2{}_{BP} =$



  - $SSR^* =$
  - $SSE =$

Example: Lean body mass (LBM) and calorie rate

The Regression Sum of Squares for e_i^2 :

$$\boldsymbol{SSR}^* =$$

Call: lm(formula = residsqrd ~ LBM)… var(Women$LBM)=

Coefficients: Estimate

(Intercept) 4331.66 ; LBM 74.41

The Error Sum of Squares:

$$\boldsymbol{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} {e_i}^2 =$$

Call:

lm(formula = RATE ~ LBM)…Residual standard error: 95.08 on 10 degrees of freedom

$$\chi^2{}_{BP} =$$

In R:

ncvTest(reg)

Non-constant Variance Score Test

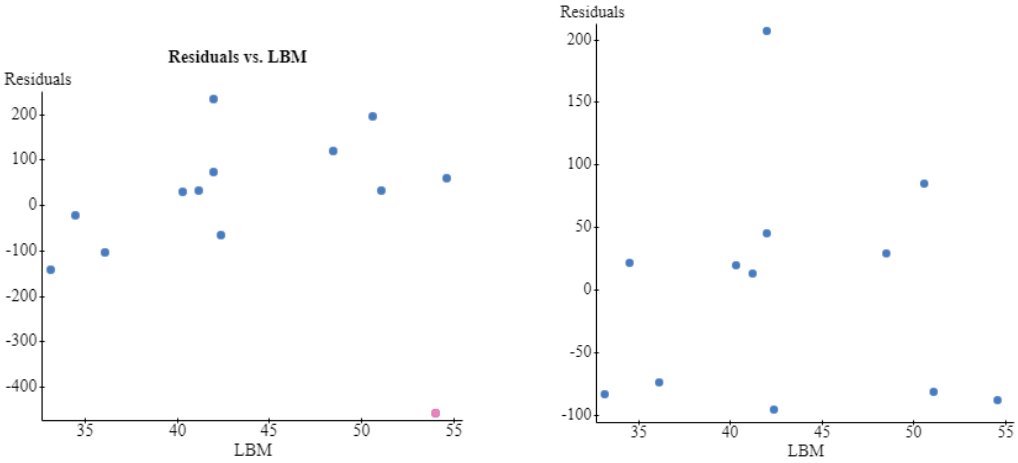
Variance formula: ~ fitted.values

Chisquare = 0.02530998, Df = 1, p = 0.8736

TESTS INVOLVING RESIDUALS: OUTLIERS

- Try fitting a new regression line without outlier.to examine the effect.

Calorie data with an additional data value-(X=LBM=54, Y=RATE=900). Is this an outlier?



- Outlier test (**ncvTest ()** in *"car"* package in **R**)
    - Based on studentized (t) residuals with Bonferonni correction
        - Requires large t value with small adjusted p-value to support that an observation is an outlier.
            - (More details to come about Bonferonni corrections.)

outlierTest(reg2)

rstudent unadjusted p-value Bonferroni p

TESTS INVOLVING RESIDUALS: LACK OF FIT TEST

- Lack of Fit tests are used to see if a regression function fits the data or not.

Here will examine the approach with a linear regression function.

Assumptions

- Assumes $\boldsymbol{Y}$ observations are:
  - **Independent** for given level of X
  - **Normal**
- Assumes $\boldsymbol{Y_i}$ have **constant variance** σ^2.
-
  -
  -

Modified Notation:

- Different $\boldsymbol{X}$ levels:
- **Number** of **replicates** for j^{th} level of X:
  - **Total sample size**: $\boldsymbol{n} =$
- **Observed value of response** (Y) for i^{th} replicate for the j^{th} level of X:
  - $i = 1,\ 2, \cdots, n_j$
  - j$= 1,\ 2, \cdots, C$
- **Mean of the $\boldsymbol{Y}$ observations** at level $X = X_j$:

Full Model

Assumptions are **identical** to those for **SLR** (Simple Linear Regression) model, **except** that we will **not assume that the relation is linear**.

$$\boldsymbol{Y_{ij}} =$$

Expected Value (Mean) **response** when $X = X_j$ **:** $\boldsymbol{E}(\boldsymbol{Y_{ij}}) =$

To **fit the full model to data the LS estimators** (or maximum likelihood estimators) are **assumed:**

- **Estimated Mean** for Y_{ij} **:** $\widehat{\boldsymbol{\mu}}_{\boldsymbol{j}} =$
- **Error Sum of Squares (Estimated Variance):** $\boldsymbol{SSPE} =$

(Pure error sum of squares)

- Composed of **sums of squared deviations** at **<u>each</u>** $\boldsymbol{X}$ **level**, $\sum_i \left(\boldsymbol{Y_{ij}} - \overline{\boldsymbol{Y}}_{\boldsymbol{j}}\right)^{\boldsymbol{2}}$.
  -
- **Degrees of Freedom for SSPE:**

$$\boldsymbol{df_F} =$$

<u>Reduced Model</u>

In testing whether or not a linear regression function is appropriate, we muse assume that $\boldsymbol{\mu_j}$ is linearly related to $\boldsymbol{X_j}$**:**

$$\boldsymbol{\mu_{ij}} =$$

So, the reduced model is:

$$\boldsymbol{Y_{ij}} =$$

- **Estimated Mean:** $\widehat{\boldsymbol{Y}}_{\boldsymbol{ij}} =$
- **Error Sum of Squares (Estimated Variance):** $\boldsymbol{SSE}(\boldsymbol{R}) =$

<u>Hypotheses</u>

- $\boldsymbol{H_0}$: **(linear regression model is appropriate)**
- $\boldsymbol{H_A}$: **(a nonlinear regression model may be appropriate)**

<u>Test Statistic</u>

$$\boldsymbol{F_0} =$$

where SSLF=SSE-SSPE (Lack of fit sum of squares)

- $\boldsymbol{MSLF}$=
- $\boldsymbol{MSPE}$=

**As usual, large F values and small p-values lead to supporting H_A.

$$\boldsymbol{SSE} = \boldsymbol{SSPE} + \boldsymbol{SSLF}$$

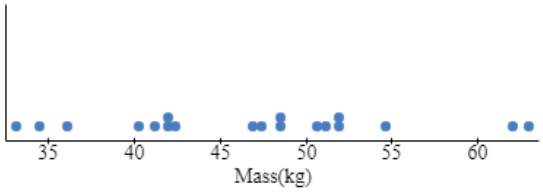
Follows from:

$Y_{ij} - \hat{Y}_{ij} = \left(Y_{ij} - \bar{Y}_j\right) + \left(\bar{Y}_j - \hat{Y}_{ij}\right)$

ANOVA Table for Lack of Fit Test

Source of Variation	Sum of Squares (SS)	df	Mean Squares (MS)	$F-statistic$
REGRESSION	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR =$	$F_0 =$
ERROR	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE =$	
LACK OF FIT	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF =$	
PURE ERROR	$SSPE = \sum_j\sum_i(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE =$	
TOTAL	$SST = \sum(Y_{ij} - \bar{Y})^2$	$n-1$		

Example: Lean body mass (LBM) and calorie rate



Example: TESTS INVOLVING RESIDUALS: LACK OF FIT TEST

Lack of Fit tests are used to see if a regression function fits the data or not. Here we will examine the approach with a linear regression function. What we are testing:

Hypotheses

H_0: $E(Y) = \beta_0 + \beta_1 X$ **(A linear regression model is appropriate)**

H_A: $E(Y) \neq \beta_0 + \beta_1 X$ **(A nonlinear regression model may be appropriate)**

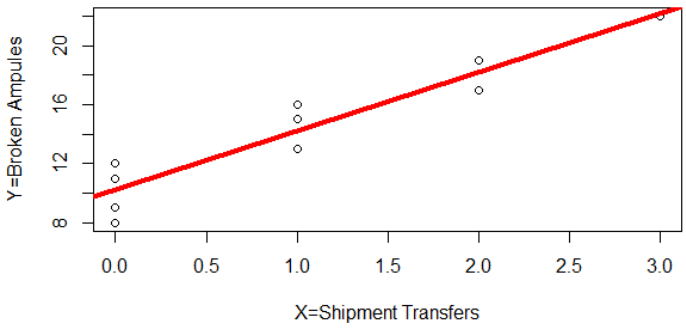
In using the Lack of Fit test in ***R*** we compare the fit of two models, Model 1 being the reduced model (the normal linear regression model) and Model 2 being the full model (treating X as a categorical variable and finding the mean at each level of X), by using **anova(***reduced,full***).**

To look at some of the details, let's look at this small data set from the text (***CH01PR21.txt***)

Airfreight breakage. A substance used in biological and medical research is shipped by air-freight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model (1.1) is appropriate.

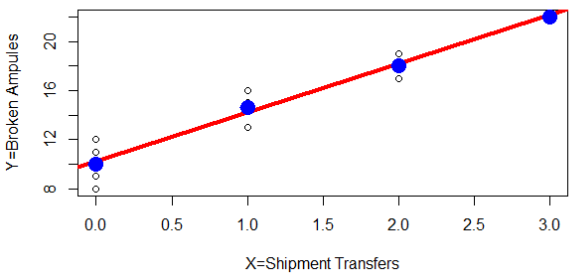
i:	1	2	3	4	5	6	7	8	9	10
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11

We can see here that we do have four different X levels (the number of times the carton was transferred), with all but one having replicates. Fitting a normal regression model (this will be **Model 1**):



So, we will define this as `reduced<-lm(Y~X)`.

Defining Model 2, treating the four X levels as categorical variables (factors):
`full<-lm(Y~as.factor(X)).`
Note: In the plot below, the blue points correspond to the means at $X = 0, 1, 2, 3$.



Now running the ANOVA test (`anova(reduced,full)` we have:

```
Analysis of Variance Table

Model 1: Y ~ X
Model 2: Y ~ as.factor(X)
     Res.Df    RSS   Df  Sum of Sq      F   Pr(>F)
1       8     17.600
2       6     16.667    2   0.93333    0.168  0.8492
```

For comparison purposes, let's also run an ANOVA test with just the reduced model (like we would have done in Chapter 2)- `anova(reduced)`:

```
Analysis of Variance Table

Response: Y
          Df Sum Sq    Mean Sq   F value    Pr(>F)
X          1  160.0     160.0    72.727     2.749e-05 ***
Residuals  8   17.6      2.2
```

One sum of squares is the same in both tables: it is the error sum of squares $SSE(R) = SSE = 17.6$, measuring variation that cannot be explained by the linear model, with $df = 10 - 2 = 8$. It is composed of the other two sums of squares displayed: the lack of fit sum of squares $SSLF = 0.93333$, measuring the variation due to a lack of fit of the linear model (and what can be improved by fitting another) , with $df = c - 2 = 4 - 2 = 2$ and the pure error sum of squares $SSPE = 16.667$, measuring variation that cannot be explained by any model, with $df = n - c = 10 - 4 = 6$. ($SSE = 17.6 = 16.667 + 0.93333 = SSPE + SSLF$)

Note that the amount of error due to the lack of fit is very small leading to a test statistic of:

$$F_0 = \frac{MSLF}{MSPE} = \frac{(0.93333/2)}{(16.667/6)} = 0.168$$

With the small F-value and large p-value (0.8492) we fail to reject H_0 and, thus, do not have enough evidence to support a lack of fit of a linear model. We assume that a linear model is appropriate.