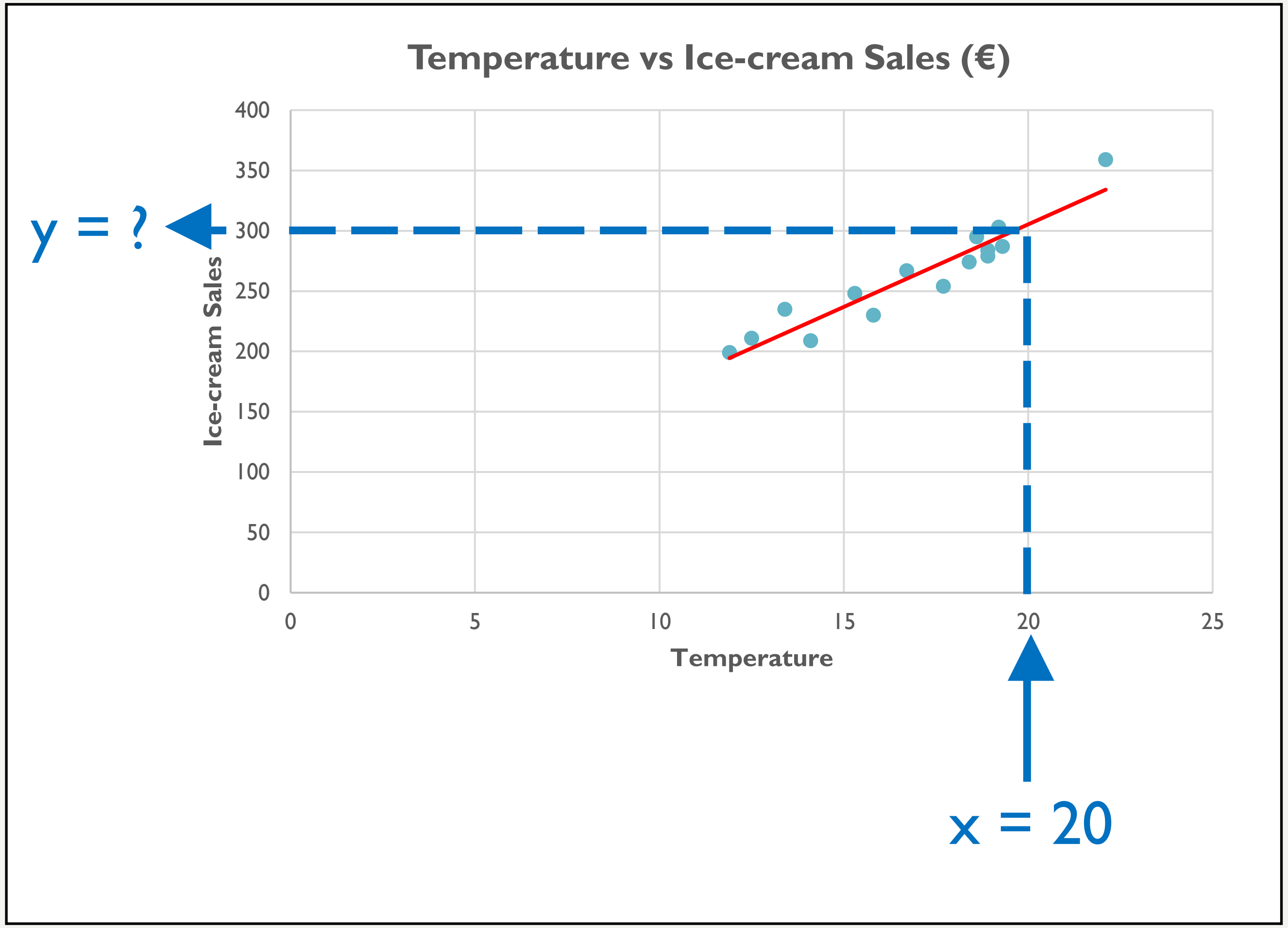
MULTIPLE LINEAR REGRESSION

WHAT IS PREDICTION ALL ABOUT?

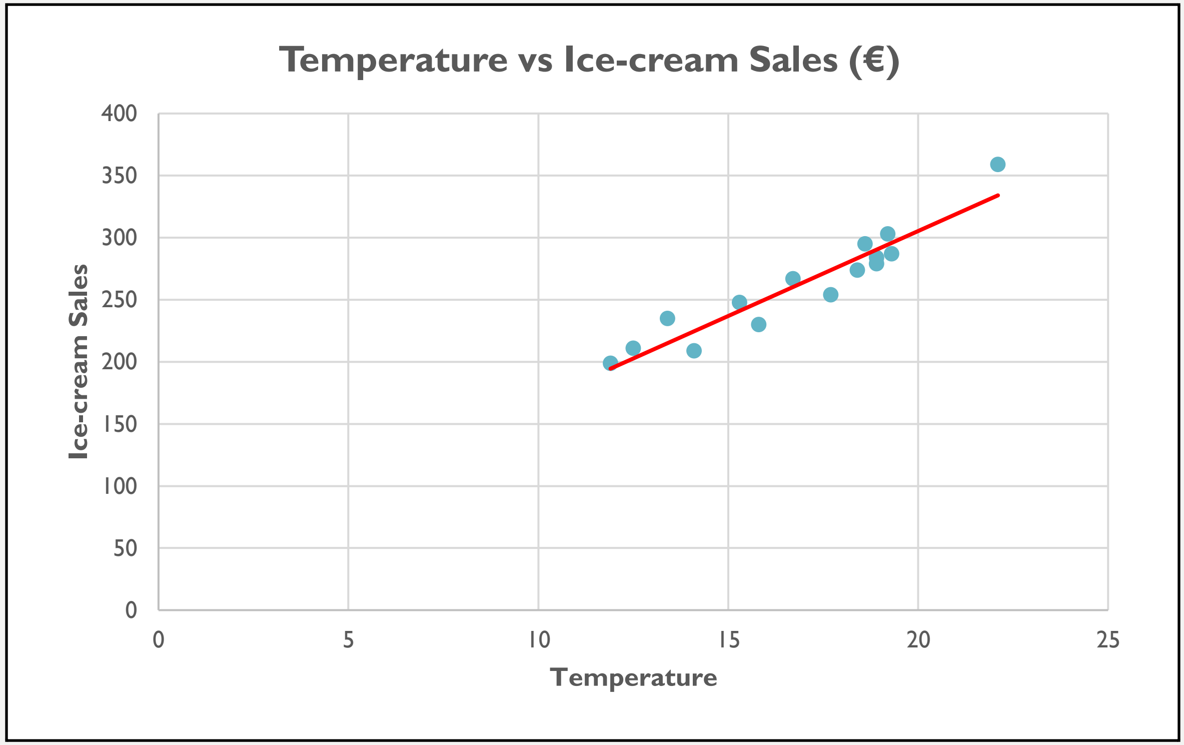


- Correlations
  - Predict the value of one variable based on the value of another
- Basic idea
  - Use a set of previously collected data (such as data on variables X and Y)
  - Calculate how correlated these variables are with one another
  - Use that correlation and the knowledge of X to predict Y.

REGRESSION

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

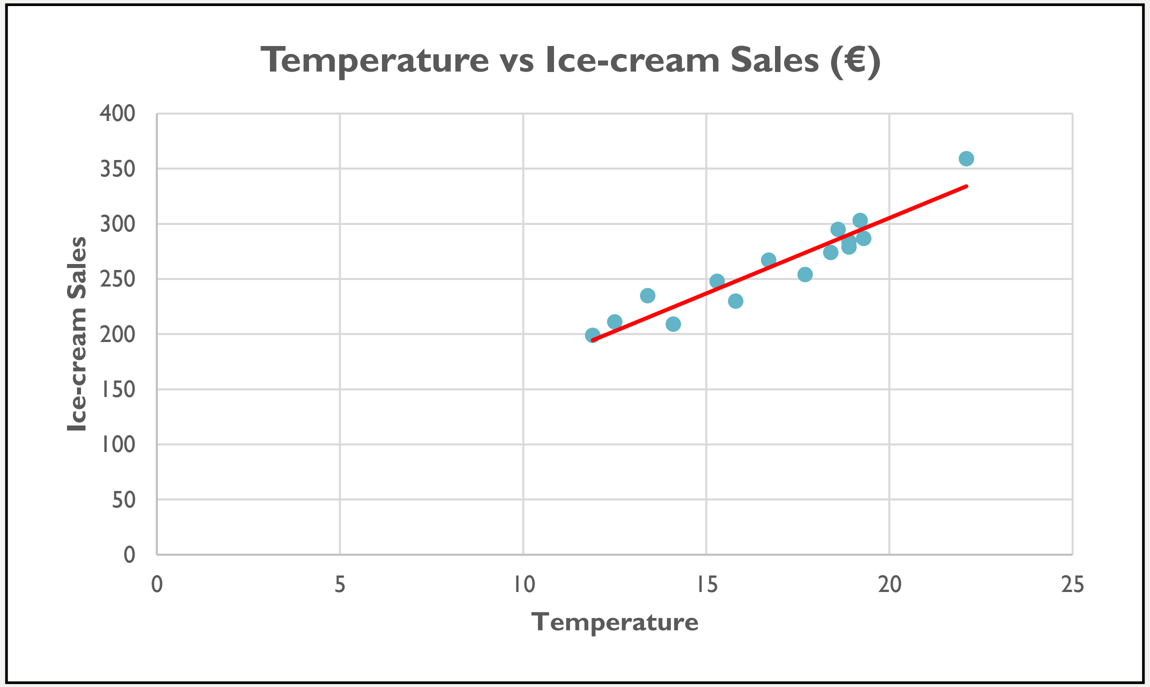
- Dependent variable (y):
  - the variable we wish to predict or explain
- Independent variable (x):
  - the variable used to predict or explain the dependent variable



SIMPLE LINEAR REGRESSION MODEL

- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be related to changes in X

- Predict ice-cream sales for a given temperature



SIMPLE LINEAR REGRESSION MODEL

Linear Regression Function	$y = a + bx$
Slope (b) of Regression Line	$b = r \frac{s_y}{s_x}$
Y-Intercept (a) of Regression Line	$a = \overline{y} - b\overline{x}$

Model

LINEAR REGRESSION FUNCTION

Slope of Regression Line

Dependent Variable

$$y = a + bx$$

Y intercept

Independent Variable

MULTIPLE LINEAR REGRESSION

- Predict an outcome from two or more independent variables
- Warning:
  - Add additional variables
    only under certain conditions
  - Any new independent variable has to
    make a **unique** contribution to understanding the dependent variable

MULTIPLE LINEAR REGRESSION

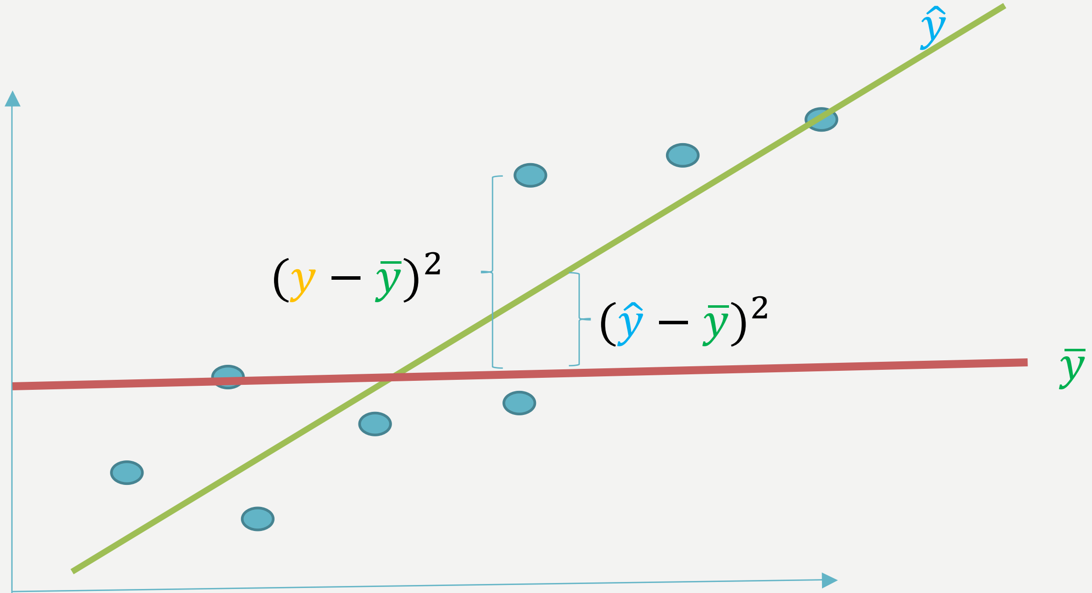
- Interpretation of coefficients: Amount by which Y changes when that particular x increases by one unit with the values of all the other independent variables held constant

$$Y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

EVALUATION OF MLR MODEL-R^2

- COEFFICIENT OF MULTIPLE DETERMINATION R2: The proportion of variation in the dependent variable that is explained by the independent variables, Ranges from 0 to 1.

- R^2=$\frac{Explained\ variance}{Total\ Variance}$ where Explained Variance = $(\hat{y}-\bar{y})^2$, Total Variation = $(y-\bar{y})^2$



ADJUSTED R2

- The number of independent variables in a multiple regression equation makes the R2 larger. Adjusted R2 is used instead to balance the effect that the number of independent variables has.
- The adjusted R2 increases only if the new term improves the model more than would be expected by chance.
- Concept of PARASIMONY: The principle that the most acceptable explanation of an occurrence, phenomenon, or event is the simplest, involving the fewest entities, assumptions, or changes[1].

[1] https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100346221#:~:text=The%20principle%20that%20the%20most,entities%2C%20assumptions%2C%20or%20changes.

GENERALIZED MODEL

- When we run regression, we hope to be able to generalize the sample model to the entire population.
- To do this, several assumptions must be met.
- Violating these assumptions stops us from generalizing conclusions about our target population.

BLUE

- Our main concern when running a linear regression is that our Ordinary Least Squares (OLS) estimates are BLUE, which stands for best linear unbiased estimator.
- This means we want estimates that minimise the error variance (best or efficient) and estimates that on average yield the true population parameter (unbiased); and we are using linear regression (so, linear).
- The OLS estimates are BLUE if they meet the Gauss–Markov Assumptions

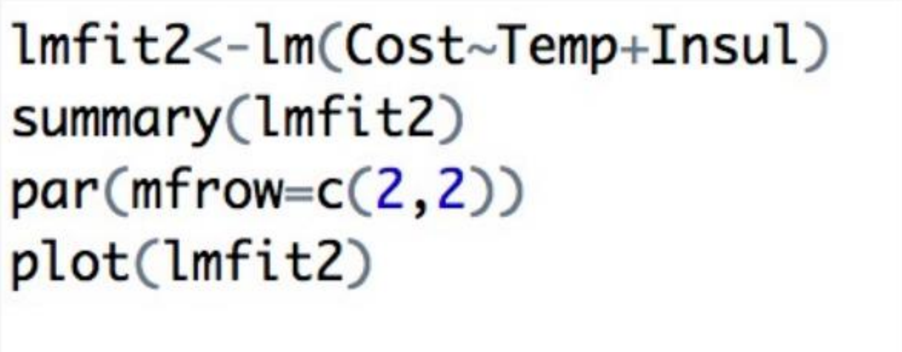
GAUSS MARKOV ASSUMPTION

- We have the correct functional form for our model
- Errors have constant variance, which is known as homoscedasticity
- There is no autocorrelation between errors
- Predictor variables must be independent of the error term (Omitted variable bias!)

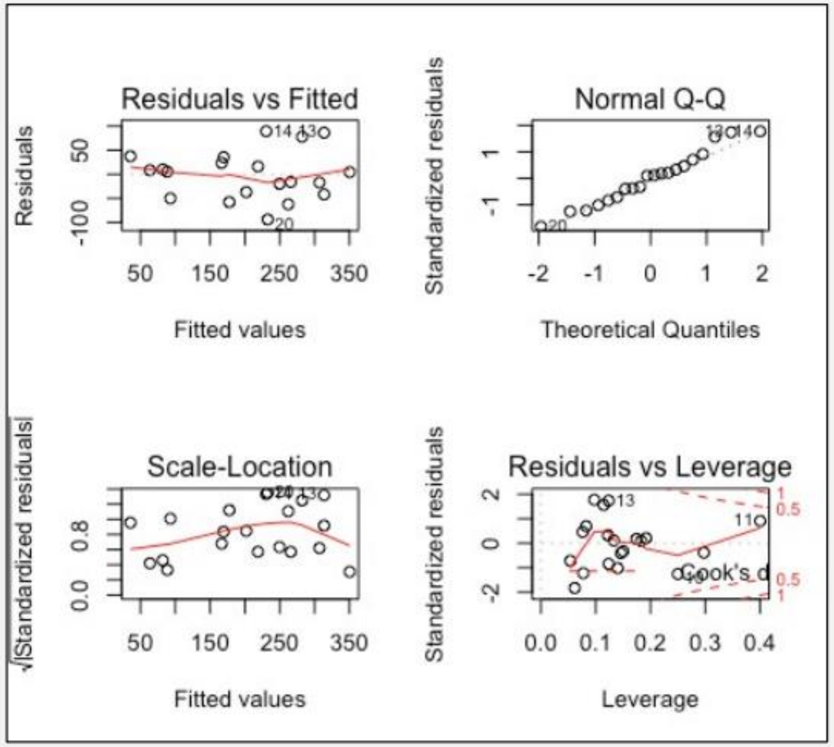
ADDITIONAL ASSUMPTION

- Errors are normally distributed.
- no multicollinearity between predictors
- no influential data points:

DIAGNOSTIC PLOTS IN R



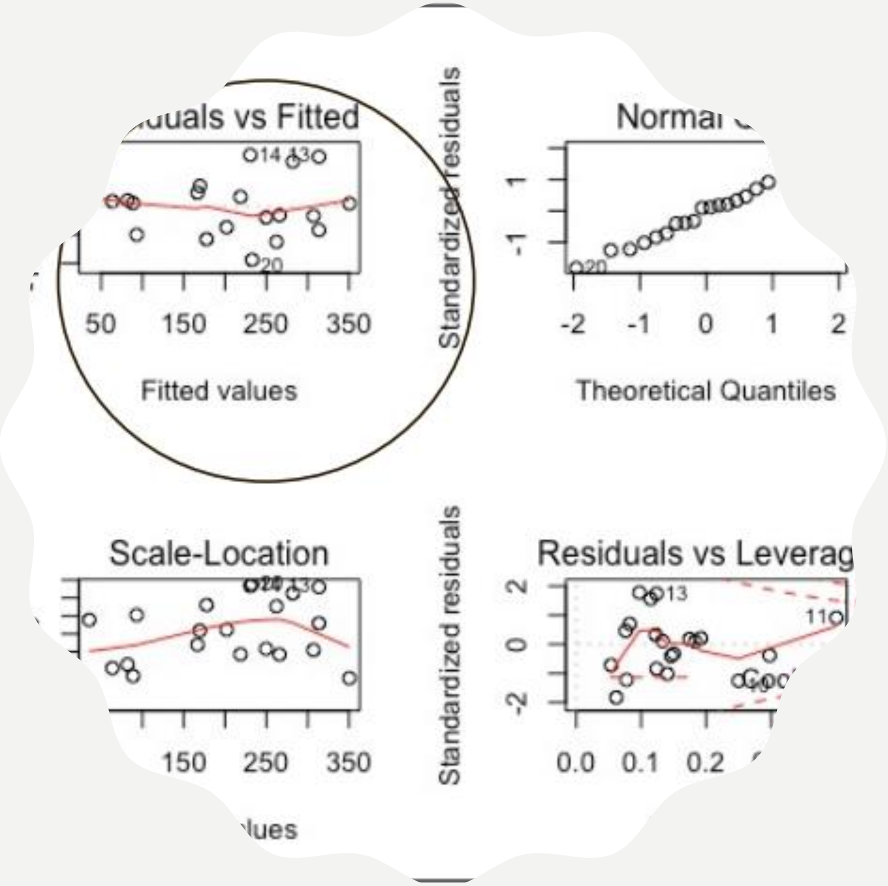
```
lmfit2<-lm(Cost~Temp+Insul)
summary(lmfit2)
par(mfrow=c(2,2))
plot(lmfit2)
```



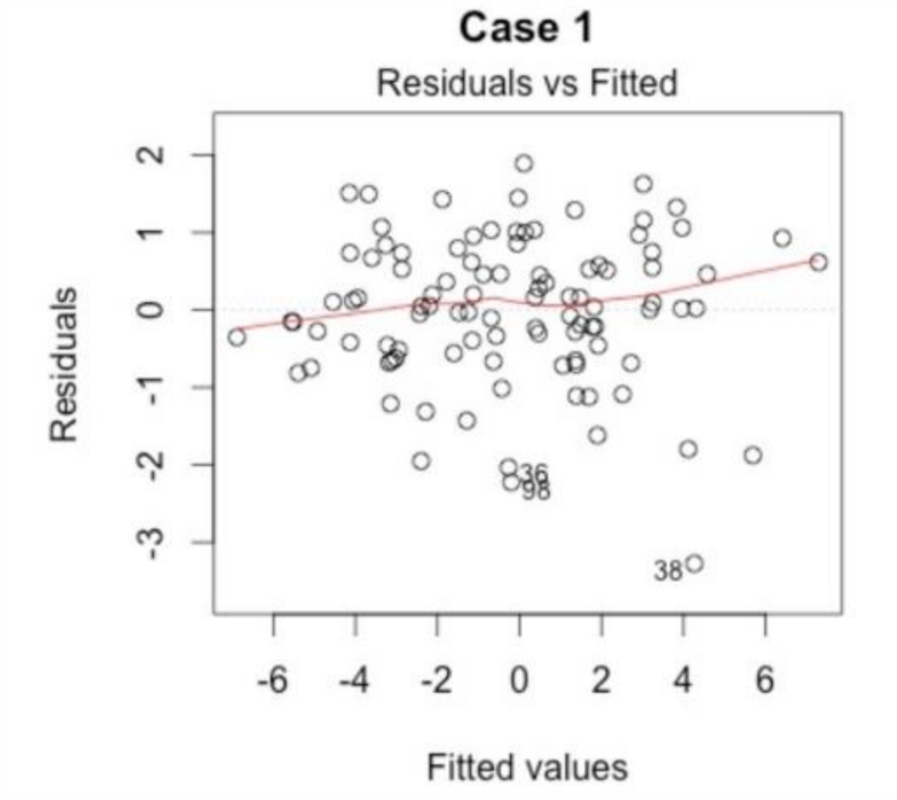
LINEARITY

- If the dependent variable is linearly related to the independent variables, there should be no systematic relationship between the residuals and the predicted (that is, fitted) values. In other words, the model should capture all the systematic variance present in the data, leaving nothing but random noise.

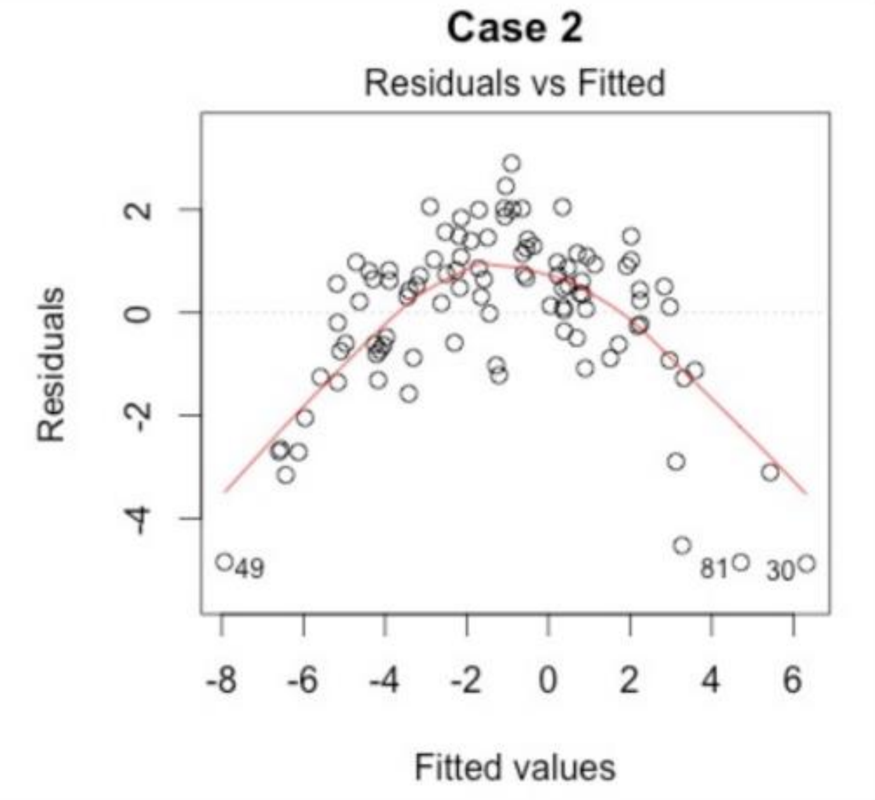
- Check the In the Residuals vs. Fitted graph in R. You don't want to see clear evidence of a curved relationship.



LINEARITY ASSUMPTION



Good model



Bad model

ABSENCE OF LINEARITY – APPROACH TO FIX

- If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to try non-linear transformations of the predictors, such as logX, √X, and X2, in the regression model.

- It’s also possible that the issue is an important missing variable

CONSTANT VARIANCE (HOMOSCEDASTICITY)

- Another important assumption of the linear regression model is that the error terms have a constant variance.
- The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption.
- When heteroscedasticity is present standard errors are biased downwards (are smaller than they should be).
- If they are smaller than they should be, we may find statistically significant predictors which, in fact, are really not because $t = \beta$ / s.e.

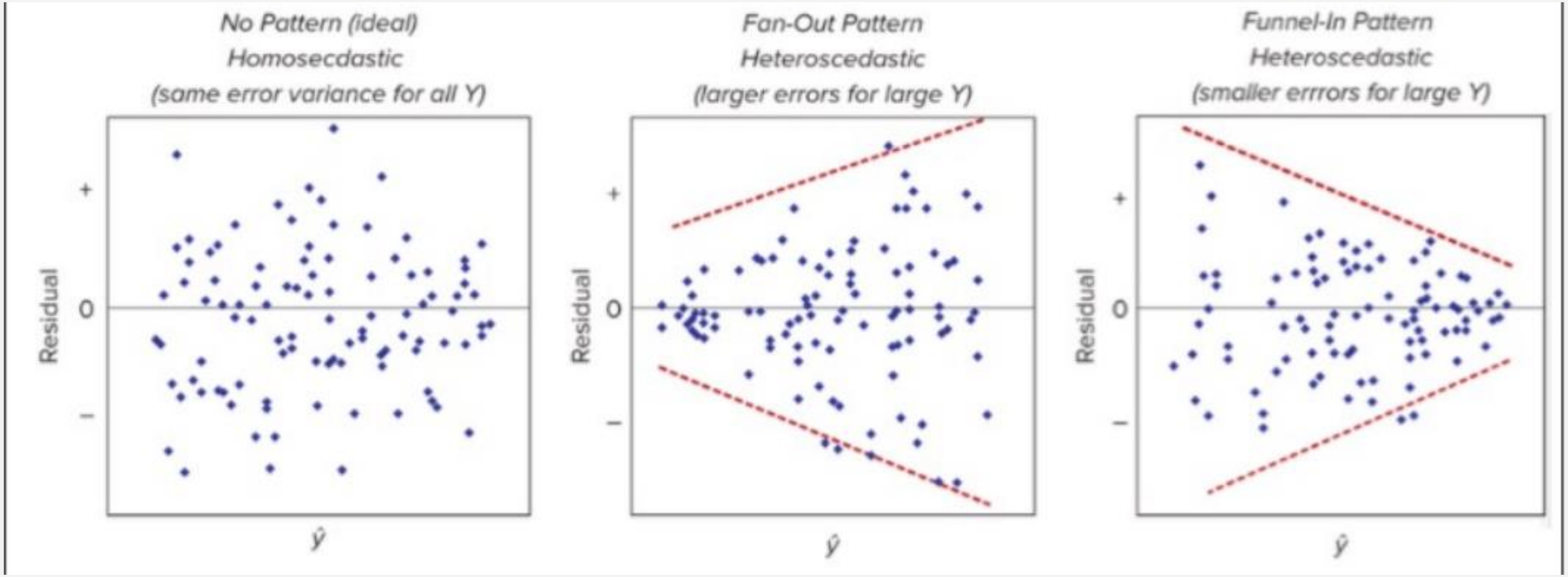
HOMOSCEDASTICITY

Testing for heteroscedasticity:

- Plot the residuals against the fitted values. - If there is a pattern then we probably have heteroscedasticity. If the residuals just look like noise with no obvious pattern, then we probably do not have heteroscedasticity.

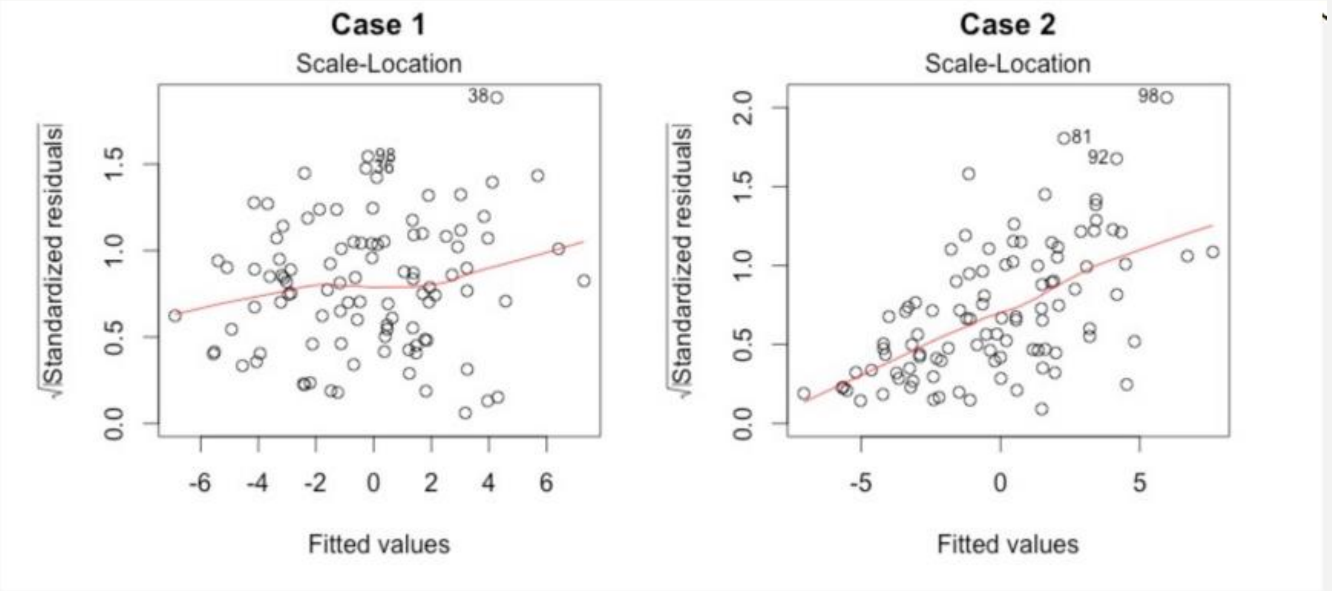
- Plot ZRESID against ZPRED

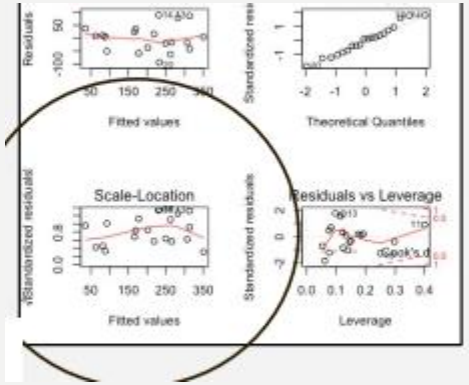
HOMOSCEDASTICITY



HOMOSCEDASTICITY - DIAGNOSTIC PLOT IN R

- To meet the constant variance assumption, the points in the Scale-Location graph of the R Diagnostic plots (bottom left) should be a random band around a horizontal line.



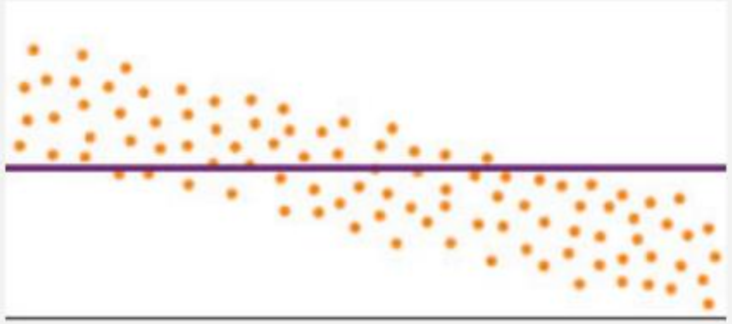


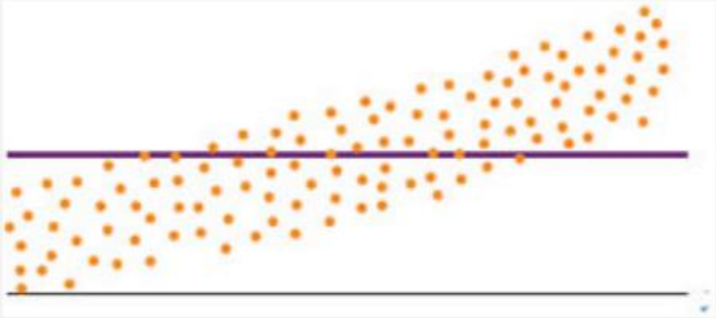
CORRECTING FOR HETEROSCEDASTICITY

- When faced with this problem, one possible solution is to transform the response Y using a function such as log Y or √Y .
- Often heteroscedasticity indicates that a variable is missing

ASSUMPTION NO AUTOCORRELATION BETWEEN ERRORS / INDEPENDENCE OF ERRORS

- An important assumption of the linear regression model is that the error terms, ε1,ε2,...,εn, are uncorrelated.
- You should use your understanding of how the data was collected
- The Durbin–Watson statistic also informs us about whether the assumption of independent errors is tenable. Durbin-Watson values less than 1 or greater than 3 would raise alarm bells.
- The closer to 2 that the value is, the better



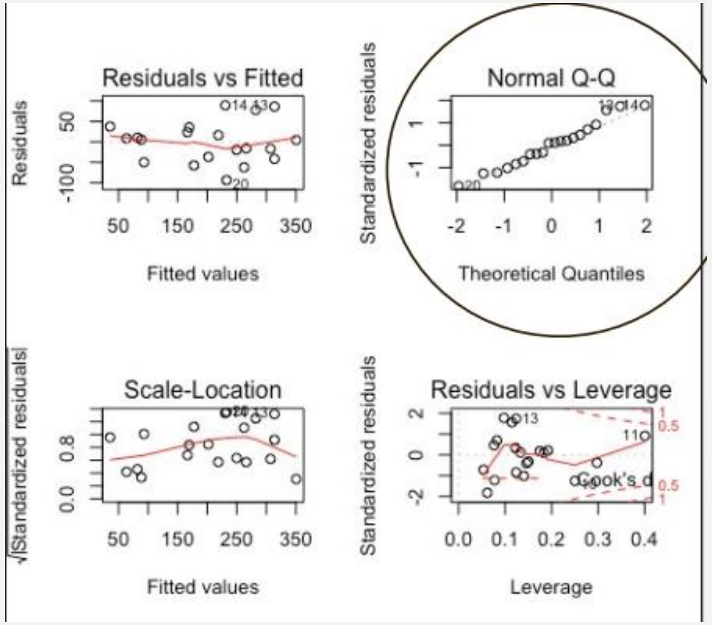


ASSUMPTION ERRORS ARE NORMALLY DISTRIBUTED

- If the dependent variable is normally distributed for a fixed set of predictor values, then the residual values should be normally distributed with a mean of 0

- Check the Normality plot

- Alternatively, look at a histogram of the residuals



CORRECTING FOR VIOLATION OF NORMAL DISTRIBUTION OF ERRORS

- If errors are not normally distributed, then we cannot assume that the estimates from our regression apply to the population
- One potential solution is to use logged versions of some variables.

ASSUMPTION ABSENCE OF MULTICOLLINEARITY

- Multicollinearity is when two or more predictors are functions of one other (possibly latently). When multicollinearity is present, standard errors will be larger than they should be.
- An informal test is to check the correlations between the predictor variables. If any variables are correlated at |0.8| or higher
- The second is a formal test known as the variance inflation factor (VIF) test. if a predictor has a VIF of 5 or above then it will likely be collinear with other predictors.
- The most common solution to the problem of mulitcollinearity is to drop the problematic variable
- Another option is to create a new combination variable (if theoretically justified)

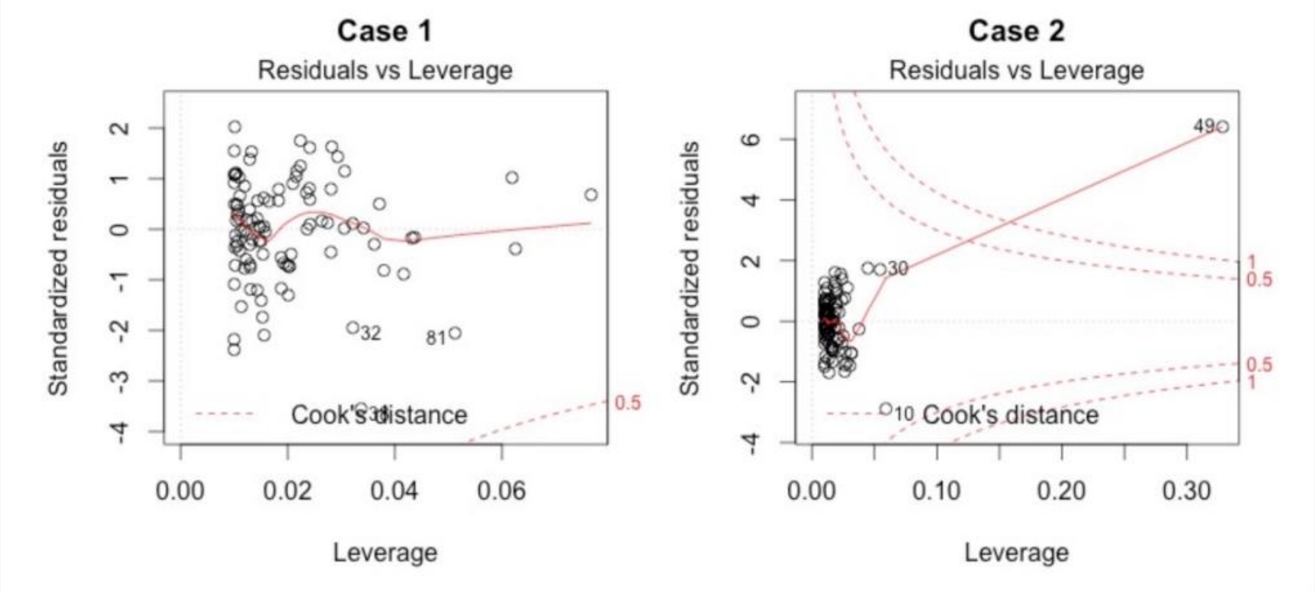
ASSUMPTION NO INFLUENTIAL DATA POINTS

- Outliers are observations that aren't predicted well by the model. They have unusually large positive or negative residuals.

- Observations that have high leverage are outliers with regard to the other predictors. In other words, they have an unusual combination of predictor values. The response value isn't involved in determining leverage. Observations with high leverage are identified through hat values.
- Influential observations have a disproportionate impact on the values of the model parameters. The model changes dramatically with the removal of a single observation. It's this concern that leads you to examine your data for influential points.

ASSUMPTION NO INFLUENTIAL DATA POINTS

- We test for influential data points by looking at Cook's distance (often shortened as Cook's d), which measures the effect of each observation on the regression coefficients. The rule of thumb is that any points that have a Cook's d of 1 or greater are considered to be influent



Good Model

Bad Model

CORRECTING FOR OUTLIERS

- The most common solution is to drop them from the regression
- If we are working with a dataset that has a large number of observations, dropping a few influential data points is likely not going to matter too much. If we are working with data with a relatively small number of observations, dropping a single influential data point may have a large effect