Other Considerations in the Regression Model

Qualitative Predictors

Sometimes the predictors X are qualitative and we need to fit them in our regression model

Predictors with Only Two Levels

If a qualitative predictors only has **two levels** or possible values we can simply create a **Dummy Variable** that takes numerical values :

$$x_i = \begin{cases} 1 \ \text{ if } i\text{th person owns a house} \\ 0 \ \text{ if } i\text{th person does not own a house} \end{cases}$$

and use the new **Dummy variable** as a predictor in the regression equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

If person owns a house

$$y_i = \beta_0 + \beta_1 + \varepsilon_i$$

if person doesn't own a house

$$y_i = \beta_0 + \varepsilon_i$$

We can also create another **Dummy Variable**:

$$x_i = \begin{cases} 1 \ \text{ if } i\text{th person owns a house} \\ -1 \ \text{ if } i\text{th person does not own a house} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon = \begin{cases} \beta_0 + \beta_1 + \varepsilon \\ \beta_0 - \beta_1 + \varepsilon \end{cases}$$

- at the end the final **prediction** is gonna be the same, the only difference is how we interpret the coefficients

Predictors with More than Two Levels

When a predictors has more than two levels, Here we create additional **Dummy Variables** for example : Region : South, West, East

$$x_{i,1} = \begin{cases} 1 \ \text{ if } i\text{th Person lives in the South} \\ 0 \ \text{ if } i\text{th person does not live in the South} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 \ \text{ if } i\text{th person lives in the West} \\ -1 \ \text{ if } i\text{th person lives in the East} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon = \begin{cases} \beta_0 + \beta_1 + \varepsilon \text{ lives in the South} \\ \beta_0 + \beta_2 + \varepsilon \text{ lives in the West} \\ \beta_0 - \beta_2 + \varepsilon \text{ lives in the East} \end{cases}$$

Coefficient	Interpretation
β_0	The mean balance of all the **Regions**
β_1	How much South differs from the average of West and East
β_2	Half difference between West and East (Subtract two predictions)

- on β_2 its Half cause West add and the East subtract so the gap between them is $2\beta_2$
- If we wanted the full difference we would have coded them as 1 and 0

Extensions of Linear Model

The Standard Linear Regression model provides interpretable results and working solutions, However it puts a lot of restrictions and forces assumptions on the problem nature :

- The Linear relationship between X and Y
- The additive Relationship $\rightarrow$ the association between X_j and Y it doesn't depend on other values of other predictors (Constant increase unite in Y)

Some classical approaches to extend linear model :

1. **Removing the Additive Assumption**

- Sometimes predictors are related while the linear regression assumes that they are independent from each other which is not always the case
- Some predictors increase the prediction and combine well $Synergy$ effect also called $\text{Interaction Effect}$

$$\text{Standard Linear Model } \; Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- The increase in X_1 doesn't alter of effect the increase in β_2 even if the suggested data supports that it effect
  We can add a $interaction\ term$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \varepsilon$$
$$Y = \beta_0 + X_1(\beta_1 + \beta_3 X_2) + \beta_2 X_2$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2$$

  With $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$
- Now the association is no longer constant and independent between the predictors
  Example :

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{ radio } + \beta_3(\text{TV} \times \text{radio}) + \varepsilon$$
$$\text{Sales} = \beta_0 + \text{TV}(\beta_1 + \beta_3 \times \text{ radio}) + \beta_2 \times \text{radio} + \varepsilon$$

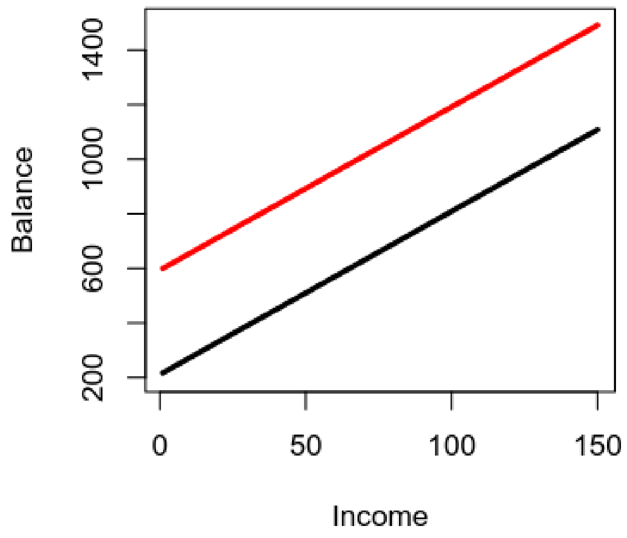
The Hierarchical Principle:

- Stats if we include an interaction in a model, we should also include the main effects, even if the $p - value$ associated with their coefficients are not significant
  But why?
- if the interaction between predictor X_1 and X_2 is important then we should include both even if their $p - value$ is high
- The logic behind it is if X_1 and X_2 relate to the Response, their coefficients β_1, β_2 being close or $zero$ doesn't matter
- X_1 and X_2 are correlated so leaving them out will cause misinterpretation
  Example :

$$\text{Balance} \approx \beta_0 + \beta_1 \times \text{income} + \begin{cases} \beta_2 \text{ if the person is a student} \\ \beta_0 \text{ if the person is not a student} \end{cases}$$

$$\text{Balance} \approx \beta_1 \times \text{income} + \begin{cases} \beta_0 + \beta_2 \text{ if the person is a student} \\ \beta_0 \text{ if the person is not a student} \end{cases}$$

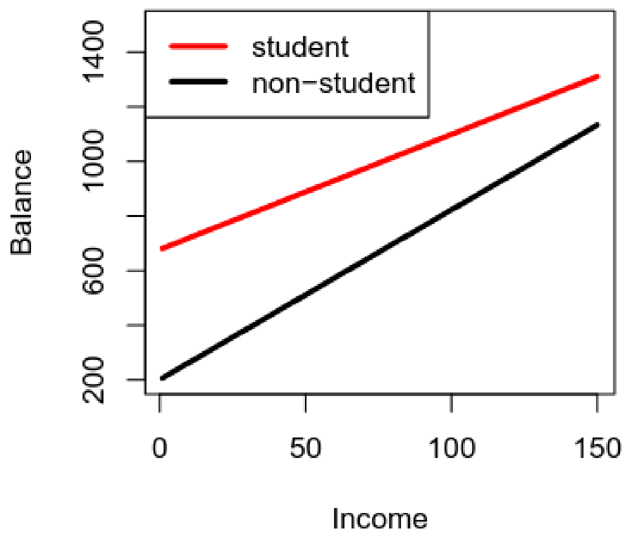
- They have the same slop value $\beta_1 \times \text{income}$
- and different intercept values β_0 vs $\beta_0 + \beta_2$



- Them being parallel lines means that being a student or not doesn't really effect the balance
- While in fact a change in income will have an impact on the balance of a student vs none- student
  Here where adding an interaction variable is important following the **The Hierarchical Principle**

$$\text{Balance} \approx \beta_0 + \beta_1 \times \text{income} + \begin{cases} \beta_2 + \beta_3 \times \text{income } \text{ if the person is a student} \\ 0 \text{ if the person is not a student} \end{cases}$$

$$\text{Balance} \approx \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income } \text{ a student} \\ \beta_0 + \beta_1 \times \text{income } \text{ not a student} \end{cases}$$



- The intercept here is different between a student and non student $\beta_0 + \beta_2 \; vs \; \beta_0$
- Same for the slop $(\beta_1 + \beta_3) \times \text{income} \; vs \; \beta_1 \times \text{income}$

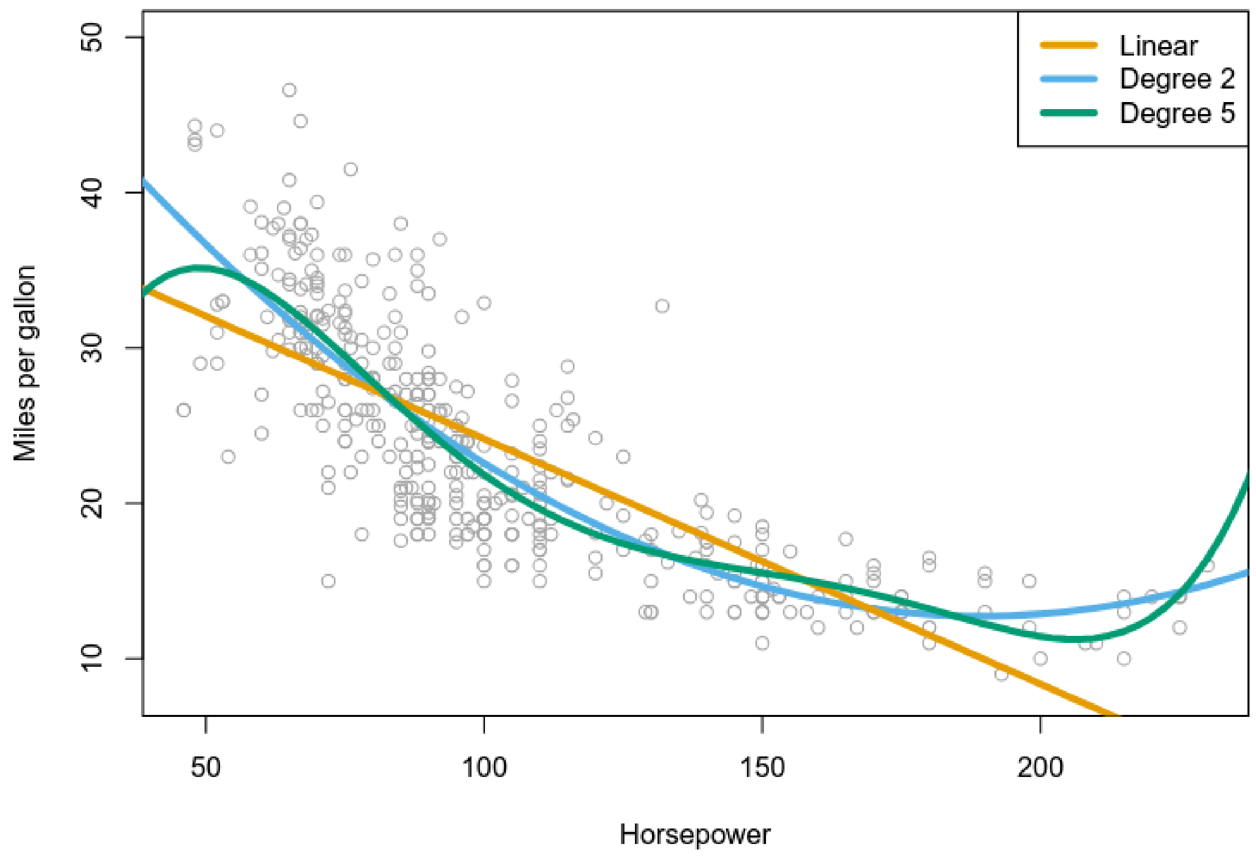
Non-linear Relationships

- Linear Regression assumes that the turn relationship between the Predictors X and Response Y is Linear
- Which is not the case most of the time
  A simple way to extend the linear model is $\text{Polynomial Regression}$

Example :

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$$

- This is still a linear model cause the coefficients are still linear in nature



- The $Degree^2$ fits better than the **Linear** Model
  This approach is called $\text{Polynomial Regression}$ to accommodate the non-linear relationships

Potential Problems

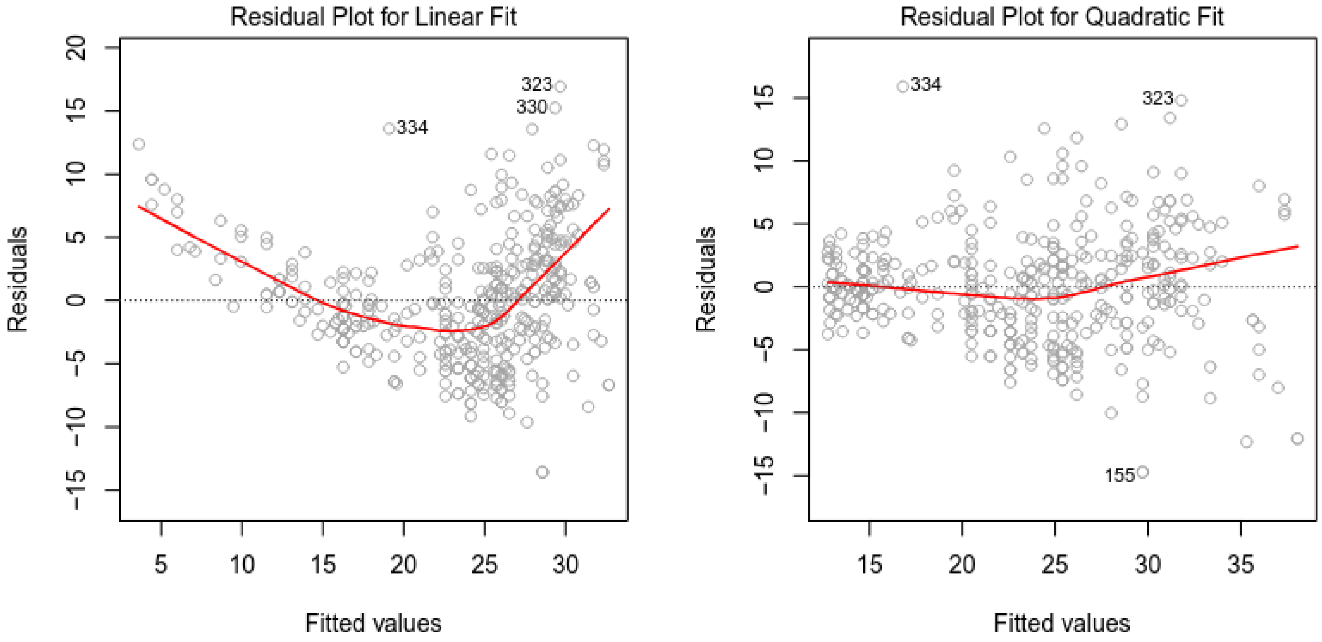
The most common problems when we fit a linear regression model to a data set are :

- **Non-Linearity** of the $\text{Response-Predictor}$ relationships
- **Correlation** of error terms
- Non-constant variance of error terms
- Outliers
- **High-leverage** points
- **Collinearity**

Non-Linearity of the Data

Knowing rather the the relationship between the response and the predictor is linear or not will help a lot when it comes to drawing conclusion for Inference or Prediction.

For that **Residual Plots** are useful graphical way to identify the non-linearity in a data set



- In Simple Linear Regression we plot the residuals $e_i = y_i - \hat{y}_i$ vs predictor x_i
- in Multiple Linear Regression we plot the residuals vs the fitted values $\hat{y}_i$

The **Residual Plot for Linear Fit**

- Exhibit a clear U-shape, which provides a strong evidence of non-linearity
  The **Residual Plot of Quadratic Fit**
- There appear to be a little pattern in the data which shows a better fit

What do we look for in the plot?

1. Random scatter around 0
2. Curved pattern $\rightarrow$ U-shape or n-shape a bad sign and indicates non-linearity
3. Funnel shape $\rightarrow$ The variance isn't constant
4. Clusters or repeating patterns $\rightarrow$ might indicate missing variables or poor modeling

If the **Residual Plot** indicate non-linearity the most simple approach is to use a non linear transformations of the predictors as:

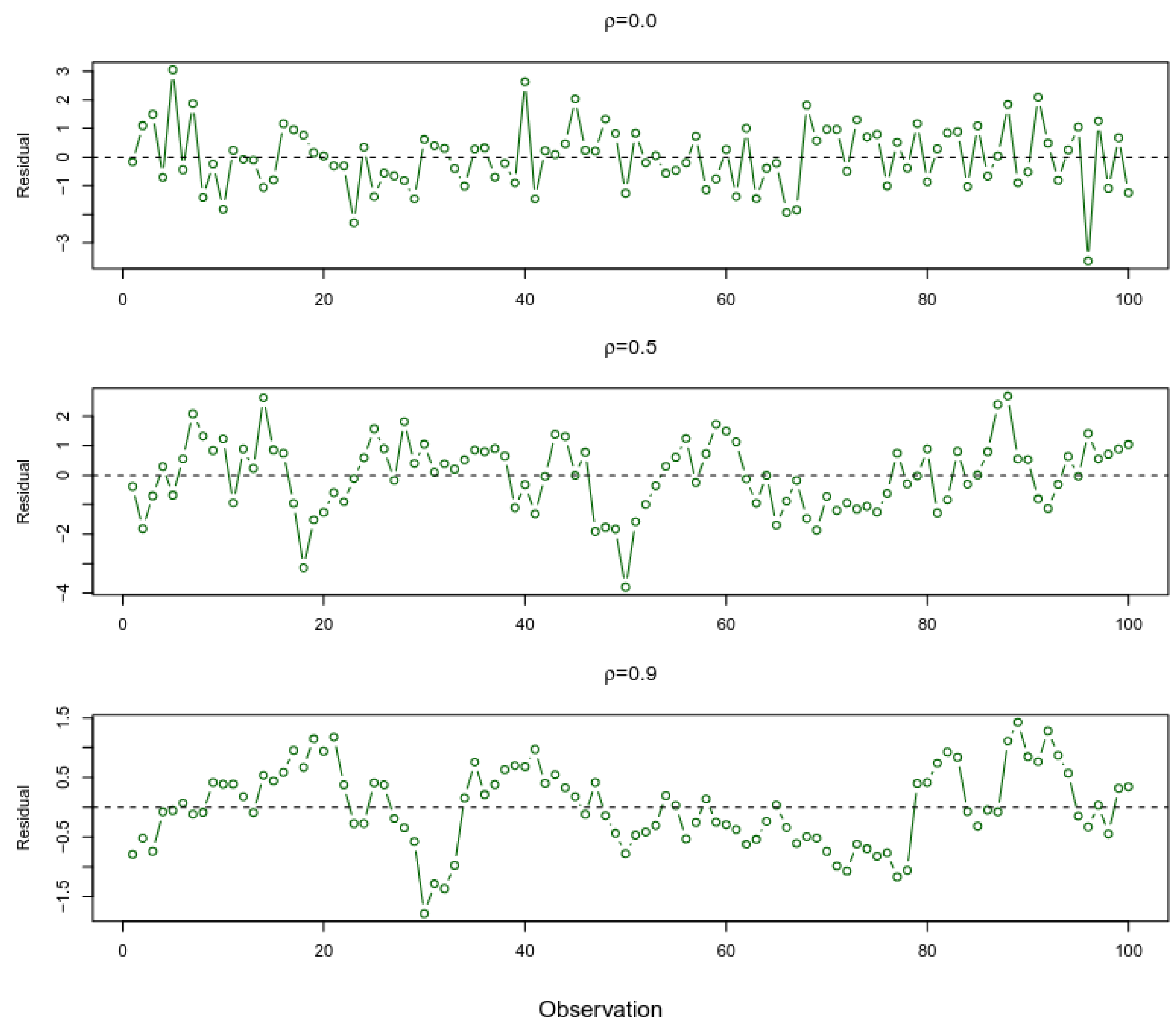
$$\log X, X^2, \sqrt{X}$$

Correlation of Error Terms

Another thing the Linear Regression assumes is that the error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are uncorrelated, Which means that each error terms is independent off the other terms.
Why it matters ?

- If the error terms are correlated, our standard error will underestimate
- The interval Confidence of 95% in fact might be much less then 95%
- The $p - values$ would be much less then expected

To determine rather the errors are correlated or not we plot the **Residuals** of the model and look for patterns that keeps occurring



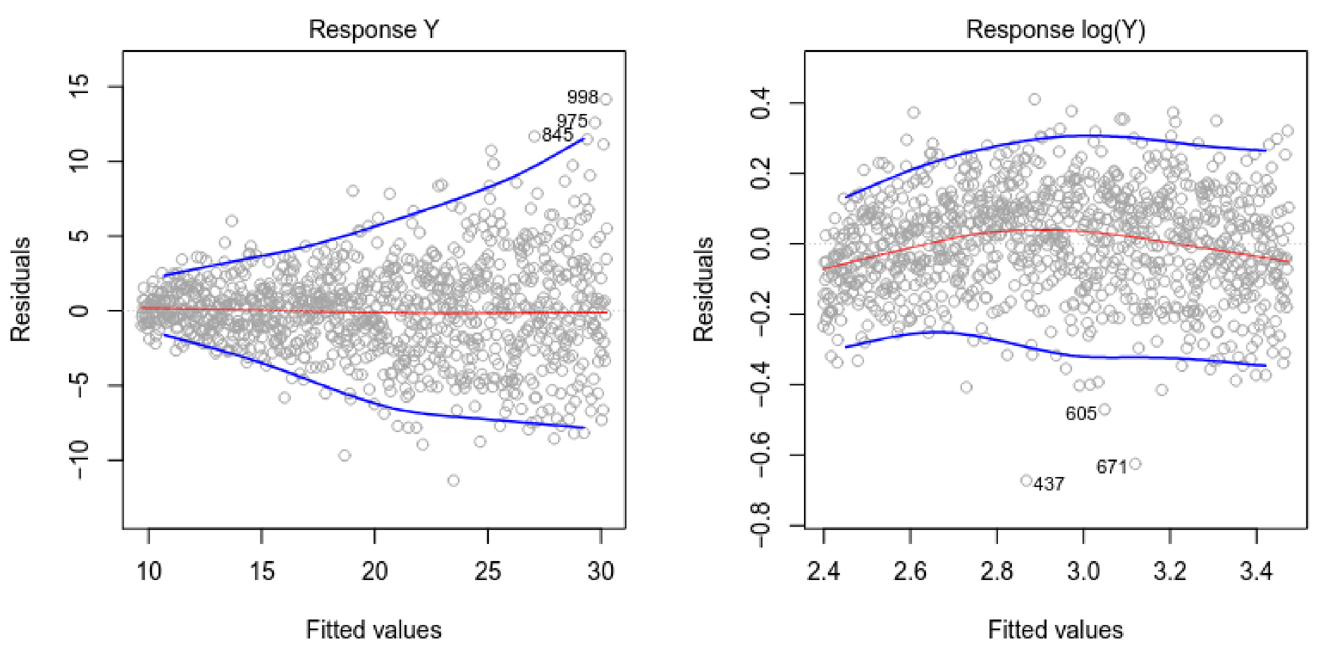
Non-constant Variance Of Error Terms

Another important assumption of the linear regression model is that the error terms are constant in their variance

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

- Confidence interval
- Hypothesis testing
- Standard error
  All of there rely on the assumption that the variance is constant

Hetroscedasticity or non constant variance can be identified by **Funnel shape**



- The left figure **Funnel Shape** which indicate that the variance in the error term isn't constant and keeps increasing

One simple solution is to transform the response Y using **Concave** function such as $\log Y, \sqrt{Y}$

- When variance is all equal for all observation its called **heteroscedasticity** as shown in $\log()$
- In the $\mathrm{Resonse}\ Y$ heteroscedasticity doesn't apply to it, we can notice unevenness in the variance
  To fix this :
- We use Weighted Least Squares **WLS**

- If we know or estimate the variance of each y_i instead of treating them equally we give **weight** to more reliable ones (the ones with smaller variance)

$$w_i = \frac{1}{\mathrm{Var}(y_i)} = \frac{1}{\frac{\sigma^2}{n_i}} = \frac{n_i}{\sigma^2}$$

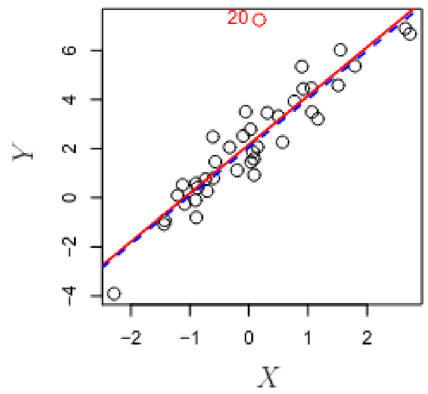
Since σ^2 is constant it cancels out in practice

$$w_i = n_i$$

Outliers

An outlier is a point for which y_i is far from the value predicted by the model, It can be cause by different reasons :

- incorrect recording of the observation

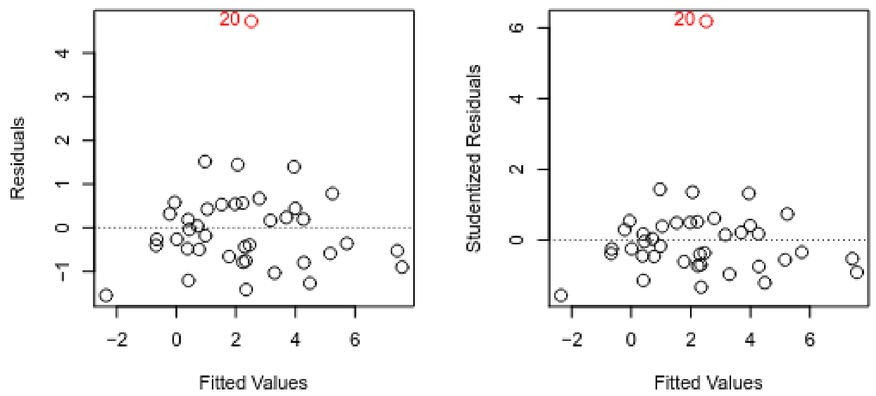


Most of the time removing the **Outlier** have little to no effect on the fitted regression line

- The red line is the fitted line before removing the outlier
- The dashed blue line is after removing the outlier

However The RSE saw a huge drop when removing the outliers which we use to compute Confidence And Prediction Intervals Derivations and Hypothesis Testing.
So its very important to know if its an incorrect recording of the observation or deficiency with the model
Studentized Residual

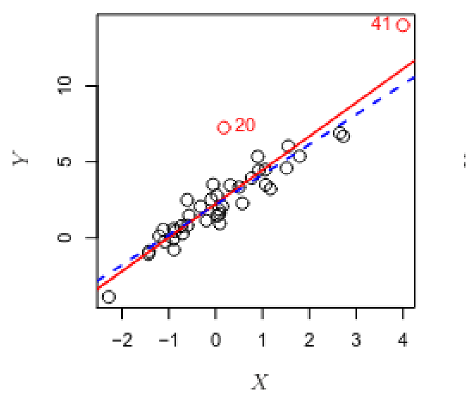
- Dividing each residual e_i by its estimated standard error



- Most data points falls between $-2, 2$
- While the outlier is over 6

High Leverage Points

Outliers are observations that have unusual y_i values while High leverage Points are Observation that have an unusual x_i value

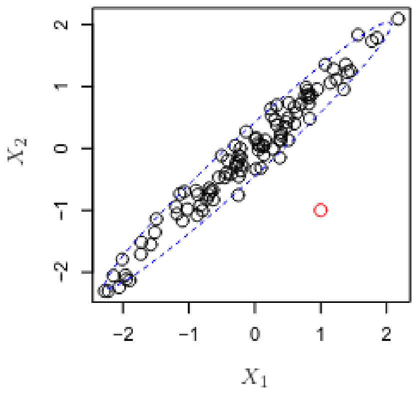


- The data points 41 is a High leverage point
- The red line represent the regression line with the data point 41
- The dashed blue line represent the regression line after deletion of 41

As we can see removing the high leverage points have much more substantial impact on the least squares line (Regression fitted line), For that its important to identify these points cause the can invalidate the whole model

In Simple Linear Regression its easy to spot High Leverage Points with just plotting the least squares line and noticing the observations with high X_i values

In Multiple Linear Regression Its much more tricky to spot cause its possible to have an observation that is pretty usual and in range of other predictors values but its **unusual** in terms of full set of predictors

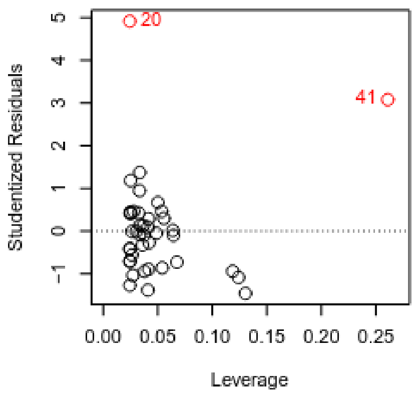


- The red line is neither value for X_1 nor X_2
  In order to quantify an observation's leverage, we compute $\text{Leverage Statistic}$:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_{i'} - \bar{x})^2}$$

- h_i increases as the distance of x_i from the mean $\bar{x}$
  The formula for Multiple Linear Regression is:

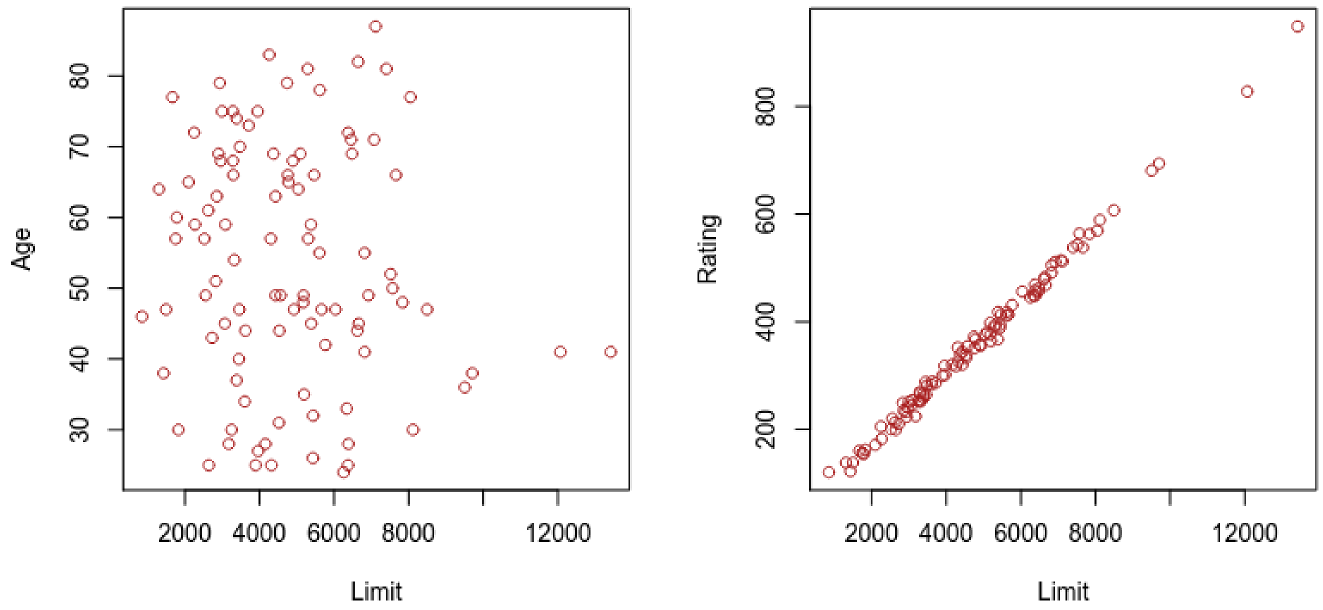
- Its always between $\frac{1}{n}$ and 1
- the average leverage for all the observations is always equal to $\frac{p+1}{n}$
- If an observation have a higher leverage statistic than average we suspect its a high leverage point



- point 20 is an **Outlier** but within the average Leverage statistic range
- point 41 is both an Outlier and high Leverage point
- Priority to remove point 41 cause having both is a very dangerous combination that might effect the model largely

Collinearity

Collinearity refers to the situation in which two or more predictor variables p are closely related to one another



- The **Age** and **Limit** variables plotted graph doesn't show any relationship
- Unlike **Rating** and the **Limit** are very Correlated with each other $collinear$

The presence of collinearity can cause problems in regression cause

- Its hard to separate out the individual effects
- We cant know how much **separately** each one effect the Response (They increase and decrease together)