CH4- Model Assessment

Amin Shirazi

2/11/2020

Assessing models

When modeling, it's important to assess the (1) **validity** and (2) **usefulness** of your model. These are model assumptions and we need to check them before making any decisions. If these assumptions are violated, we need to go back and review things and maybe refit the data using other methods to have a valid fit on the data.

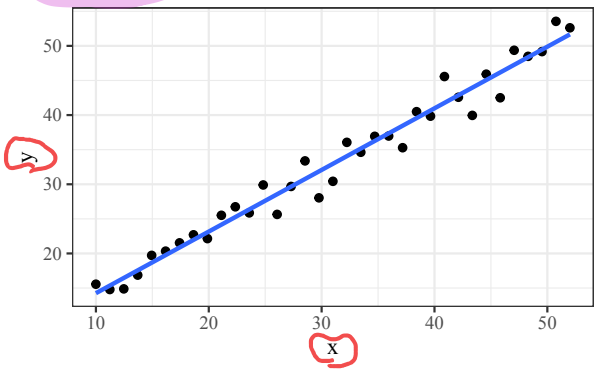
To assess the validity of the model (i.e. to see if model assumptions are met), we will look to the residuals. If the fitted equation is the good one, the residuals will be:

- Ptternless (cloud like, random scatter)
- Centered at zero
- Bell shaped distribution

To check if these three things hold, we will use two plotting methods.

> A **residual plot** is a plot of the residuals, $e = y - \hat{y}$ vs. x (or $\hat{y}$ in the case of multiple regression, Section 4.2).

(y vs. x) scatter plot

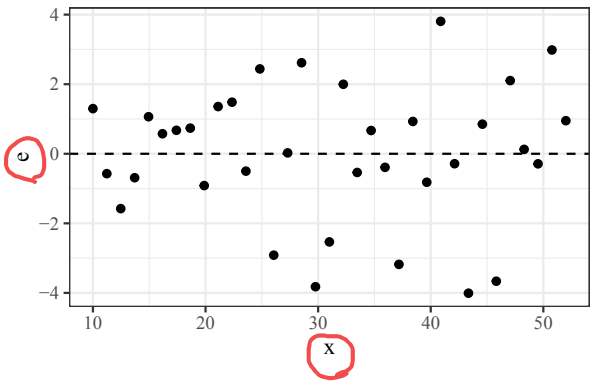


- Looks Linear.

- ideal scatterplot & its ideal Residual plot.

(linear Fit is appropriate as seen in both plots)

Residual plot (e vs. x)



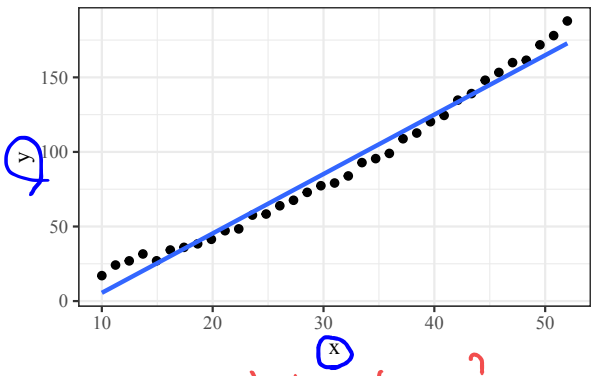
1, centered at zero

2, Patternless (randomly scattered)

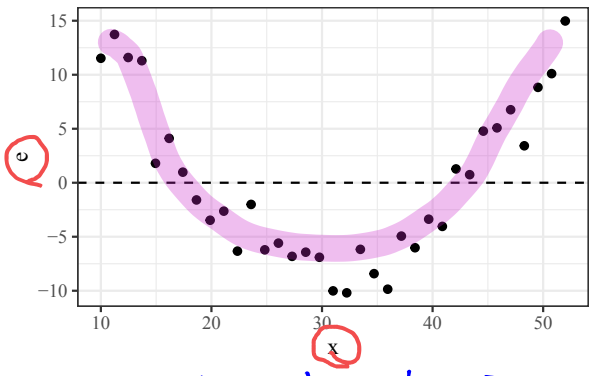
when fitting curve or surface, because there are multiple x's, we can plot residuals vs. each x, or plot residuals vs. $\hat{y}$.

Polynomial Fit: $y = \beta_0 + \beta_1 X + \beta_2 X^2$

Scatter plot (y vs. X)



Residual plot (e vs. X)

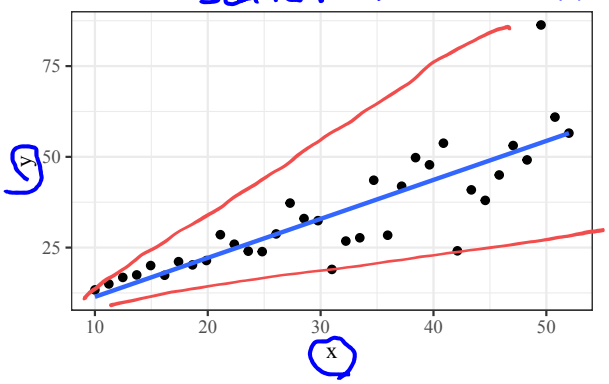


- The scatterplot doesn't seem Linear. Checking residual plot helps to see if a linear Fit is appropriate.

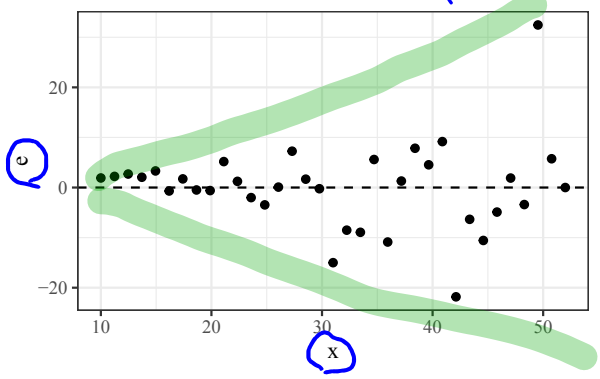
1- centered at zero.
2- not patternless. there's pattern among the residuals.

=> linear Fit isn't appropriate!

Scatterplot Y vs. X



Residual plot (e vs. X)



- variation between y_i's increases as the value of x increases.

1- centered around zero!
2- There is Pattern in the residual plot.

(This is called "heteroscedasticity" or "megaphone" shape)

(As x increases, so does the variance giving a "megaphone" shape")

=> Linear Fit doesn't seem appropriate.

Example:

When the linear model doesn't? fit the data well, there might be other type of mathematical relationship between the data.

(This is something you need to check before running your analysis.)

say the relationship between x, y is described as

$$\underbrace{y = \beta_0 x^{\beta_1}}$$

non-linear relationship between x & y.

Remedy: sometimes, we just need a transformation

$$\xrightarrow{\log} \underbrace{\log y}_{\tilde{y}} = \underbrace{\log \beta_0}_{\beta_0'} + \beta_1 \underbrace{\log x}_{\tilde{x}}$$

Now, we can fit a linear model on the transformed data:

$$\tilde{y} = \beta_0' + \beta_1 \tilde{x}$$

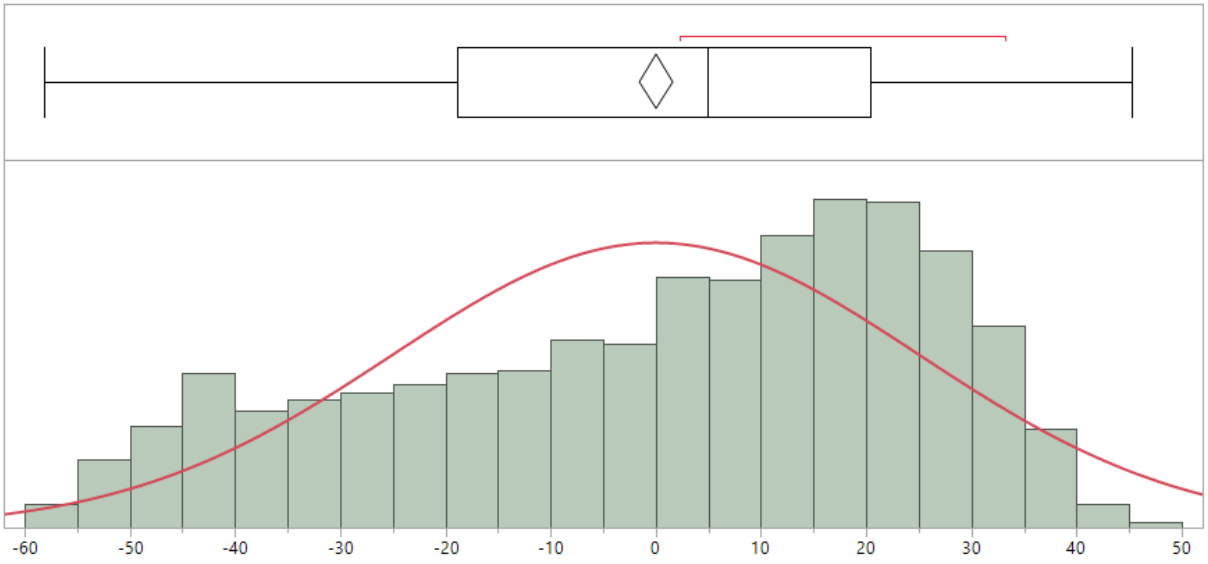
Takeaway: when the model assessment fail, try to
① re-investigate the measurement process &
② transform the data!

Normality of residuals

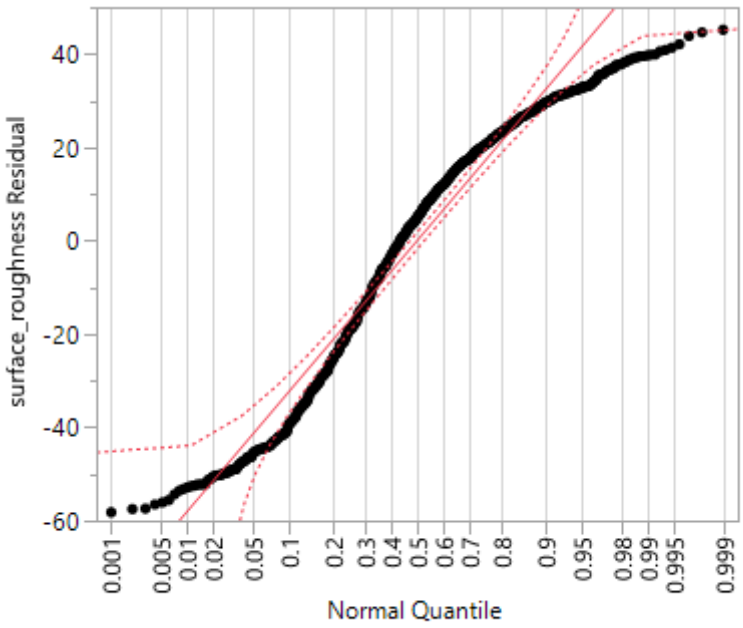
- In addition to the residual versus predicted plot, there are other residual plots we can use to check regression assumptions.
- A **histogram of residuals** and a **normal probability plot (QQ-plot)** of residuals can be used to evaluate whether our residuals are approximately normally distributed.
  - However, unless the residuals are far from normal or have an obvious pattern, we generally don't need to be overly concerned about normality.
- Note that we check the residuals for normality. We don't need to check for normality of the raw data. Our response and predictor variables do not need to be normally distributed in order to fit a linear regression model.

How to check the Normality assumptions of residuals? We are already familiar with theoretical Normal QQ-plots and histograms. So, in order to check normality assumptions of residuals, we need to see if (1) the histogram of the residuals are bell shaped and (2) check their normal QQ-plot to see if the residuals lie around a straight line.

Draw a histogram of the residuals (review the JMP toturial for histograms)



It seems the residuals are not normaly distributed in this example.The residuals have a left skewed distirbution.



- Again, the QQ-plot also confirms that the assumption of Normal distribution of residuals is violated to some extend in this example.
- More examination is required to fix the issue or to find the problem.