L9 Modeling Problems - Other Diagnostics

- Non-Normality Residuals
- False assumption of linearity between Y and any predictor.

1. When Residuals suggest error term does not following Normal Distribution:

 - OLS method still has "BLUE" estimates for parameters

(a) problem: $\boxed{\text{• Normality is needed to perform } t \text{ test, C.I. and anova}}$

 - But when sample size is large, since $\hat{\beta}_K$ is a weighted sum (linear combination) of $y_1, \cdots, y_n$, C.L.T. guarantees that it's approximately Normal distributed.

$n > 30p$

More commonly, a Normality assumption is presented but is described as less important than other assumptions of regression models :

"only extreme departures of the distribution of Y from normality yield spurious results"

— Kleinbaum et al.

(b) Detection:

(i) Normal - Probability Plot (Q-Q plot)

- resulting plot is approximately linear on diagonal of the plot suggests normality

(ii) Tests for Normality.

There are many tests addressing normality using different hypothesis: Anderson-Darling, Shapiro-Wilk, Ryan-Joiner, Kolmogorov-Smirnov, etc.

we will briefly discuss several things given in the summary table that indicate the normality / skewness of the data :

- Skew: skewness of residuals — measure for symmetry.

  sample estimate of $E\left[\left(\frac{e_i}{\sigma}\right)^3\right]$: 3rd moment

  A distribution that is perfectly symmetric around the mean will have zero skewness

- Kurtosis : tailshape of residuals — measure for clustering

  sample estimate of $E\left[\left(\frac{e_i}{\sigma}\right)^4\right]$

  Normal distribution has kurtosis = 0.

  Higher peaks leads to bigger kurtosis value.

- Omnibus K-squared Normality Test : combine Skewness and kurtosis.

$$K_{stat} = \underbrace{Z_1(g_1)^2 + Z_2(g_2)^2}_{\text{Transformation functions}}$$

$Z_1(g_1)^2$: skewness variable

$Z_2(g_2)^2$: kurtosis variable

under H_0:

$K_{stat} \sim \chi^2(2)$

H_0: Residuals are normally distributed

H_1: Residual are not normally distributed

Reject H_0: suggest residuals are not normally distributed.

- Jarque-Bera Test: similar to Omnibus, it's a test combining skewness and kurtosis

H_0: Residuals are approx. normal

H_1: Residuals are not close to normal

$$JB_{stat} = \frac{n-p}{6}\left[\underbrace{(g_1)^2}_{\text{skewness}} + \frac{1}{4}\underbrace{(g_2)^2}_{\text{kurtosis}}\right]$$

JB stat $\sim \chi^2(2)$ under H_0

rejection suggest residuals are not normally distributed.

Note*: JB test is very sensitive to outliers

(c) solutions:

(i) Natural-log transformation on y: $y_{new} = \ln y$

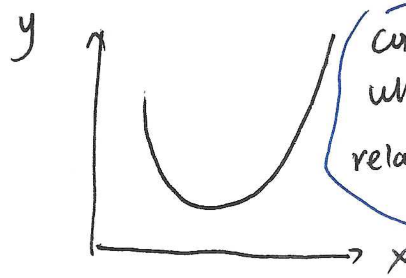
(ii) Box-cox transformation on y: $y\text{-}new = \dfrac{y^\lambda - 1}{\lambda}$

λ can be estimated using boxcox function in Python.

(iii) When sample size is large, the skewness of data doesn't impact the model estimation or inference very much.

II. False assumption of linearity between Y and any predictor.

When the linearity between Y and X don't seem valid, there've ways to adjust but we need to be extremely careful:



consider when y is relate to x nonlinearly by nature

(i) go with nonlinear approach:
very complicated.

adding polynomial terms: in MLR part, it adds # of parameters and adds multicollinearity. so the trade-off is all the testing methods become unstable. — CH 8.1: SLR case. optional
CH 13. optional. MLR case

Estimation doesn't have analytical expressions anymore.

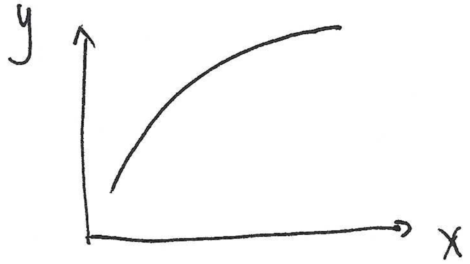
(ii) log-log transformation is powerful without losing inference.

y_new = ln y

X_new = ln X

Ex: KC_house_data.

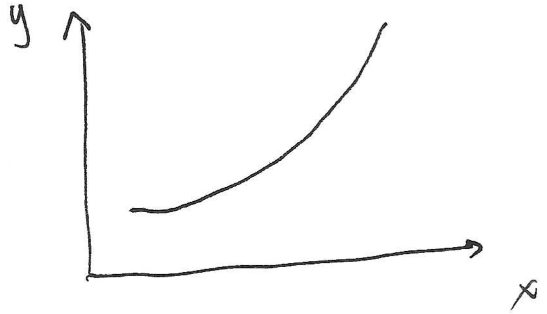
(iii) Only transformation on X if non-linearity seems to be the only problem with the data



X-new = $\log_a X$

or

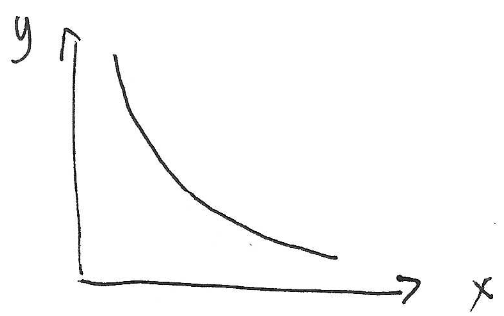
X-new = $\sqrt{X}$



X-new = X^2

or

X-new = $\exp(x)$



X-new = $\frac{1}{X}$

or

X-new = $\exp(-x)$

III. Notes on adding interaction terms in the model

We will start with two predictors:

The regression model for two predictors with interaction effect is defined as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- When X_1 increase by 1 and X_2 stay the same

$$E(Y_i - X_1 = x) = \beta_0 + \beta_1 x + \beta_2 X_{i2} + \beta_3 x \cdot X_{i2} \quad (1)$$

$$E(Y_i \, X_1 (x+1)) = \beta_0 + \beta_1 (x+1) + \beta_2 X_{i2} + \beta_3 (x+1) X_{i2} \quad (2)$$

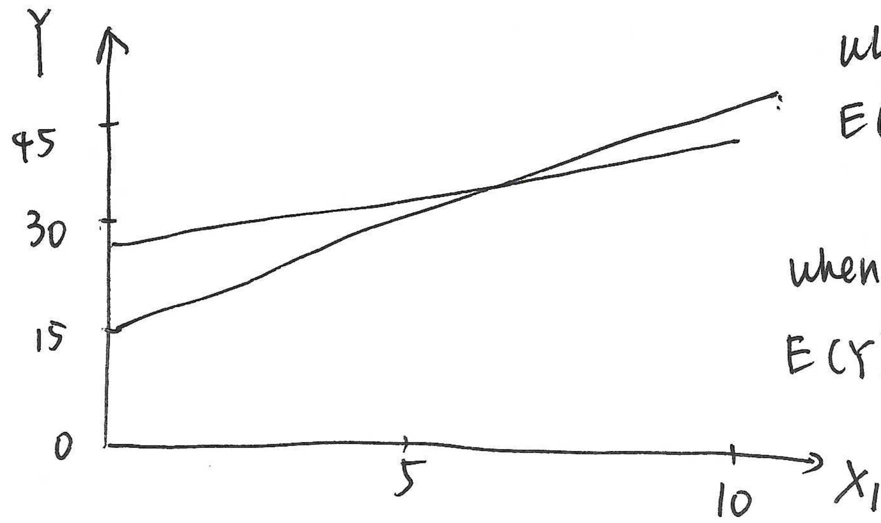
(2) - (1): Change in $E(Y)$ when X_1 increase by 1 unit

$$= \beta_1 + \beta_3 \cancel{X_{i2}} X_2$$

Similarly: Change in $E(Y)$ when X_2 increase by 1 unit

$$= \beta_1 + \beta_3 X_1$$

For ex: $E(y) = 10 + 2X_1 + 5X_2 - 0.5 X_1 X_2$



when $X_2 = 1$:
$E(Y) = 15 + 1.5X_1$

when $X_2 = 3$:
$E(Y) = 25 + 0.5X_1$

~~whith~~ with different values of X_2, the impact of X_1 to Y changes.

- The estimation of $\vec{\beta}$ and testing are constructed the same as OLS for MLR without interactions but again we need to add interaction very causiously:

  (1) Increase high multicolineanity

  (2) When the numeber of predictors is high, the potiential number $\binom{p-1}{2}$ of interaction terms is very high.

(3) In practice, we only add interaction terms if we obtain knowleadge in advance that significant interaction may exist in the study.