HW - 4 : M (2-12) | Quiz - 3 : M (2-12) | HW - 5 : M (2-18) | Quiz - 4 : M (2-18) | Class-5

Remedial Measures

If the SLR model is not appropriate for data, there are two basic choices.

1) Using a more general regression model (multiple linear, Polynomial) (Later)

2) Using some transformations on X or Y or both such that the SLR model is appropriate for transform data.

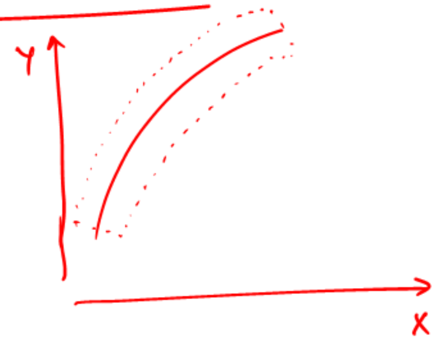
Transformations

* For non-linear relations

When the relation between X and Y is not linear (but the distribution of the error term is normal with constant error variance), X should be transformed.

Basic Rule: Pattern | Transformation

a)



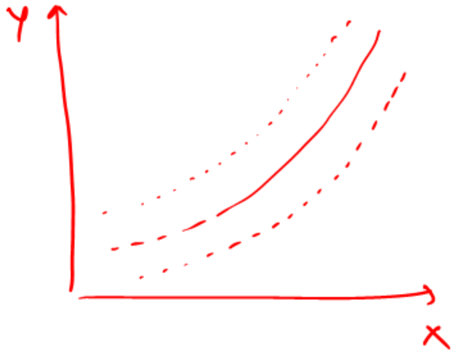
$x' = \log_{10} X$ OR

$x' = \sqrt{X}$ OR

$x' = \ln(X)$.

$Y = \beta_0 + \beta_1 X'$

b)



$x' = x^2$

$x' = \exp(X)$

c)

$x' = 1/x$

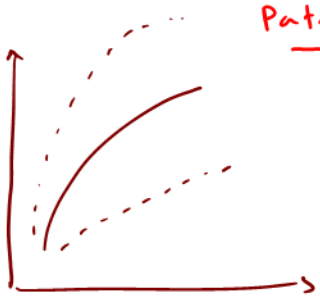
$x' = \exp(-x)$

* Non-normality and unequal error variance

If error term is not normal and variance is not constant (usually appear together), the response Y should be transformed.

Pattern — transformation — $Y = \beta_0 + \beta_1 X + \varepsilon$

a)

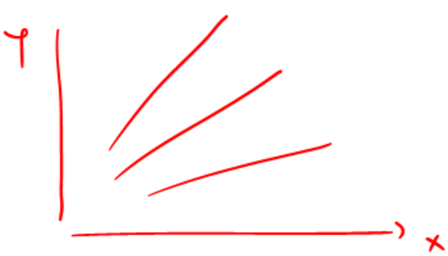


$Y' = \sqrt{Y}$

b)

$Y' = \log_e Y$

c)



$Y' = 1/Y$

Note:

* Some times we may have to transform both X and Y.
* Several transformation should be tried, then choose the better one comparing scatter plot and the residual plots.
* Some times it is difficult to determine transformation from the plots. Box-cox transformation can be used for those casses.

Box-cox transformation

Transformation $Y' = Y^{\lambda}$, where λ is a parameter to be determined from data.

The following are transformations based on the value of λ.

λ	transformation
$\lambda = 2$	$Y' = Y^2$
$\lambda = 0.5$	$Y' = \sqrt{Y}$
$\lambda = 0$	$Y' = \log_e Y = \ln Y$ ← (by the definition)
$\lambda = -0.5$	$Y' = \frac{1}{\sqrt{Y}}$
$\lambda = -1$	$Y' = \frac{1}{Y}$.

New regression model

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

How estimate λ?

The Box-cox transformation procedure uses MLE of λ as well as other parameters β_0, β_1 and σ^2.

* Simple way to find $\hat{\lambda}_{MLE}$.

Steps:

1) Choose the several values for λ in the interval $(-2, 2)$

Eg:- $-2, -1.75, -1.5, \ldots, 2$

2) For each λ, standardize Y_i^{λ} observations as follows,

$$W_i = \begin{cases} K_1 (Y_i^{\lambda} - 1) & . \ \lambda \neq 0, \\ K_2 (\log_e Y_i) & : \ \lambda = 0, \end{cases}$$

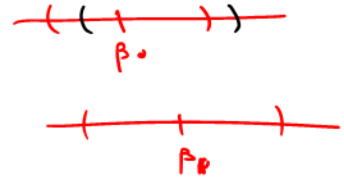
where $K_1 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n}$ and $K_2 = \dfrac{1}{\lambda K_1^{\lambda - 1}}$.

3) For each λ, fit a SLR model for W and X and calculate SSE.

(Here the value of SSE does not depend on λ, because we use standardize responses).

4) Choose the value of λ for which SSE is minimum. That is the MLE of λ.

Chapter - 4: Simultaneous Inferences and other topics in Regression Analysis.

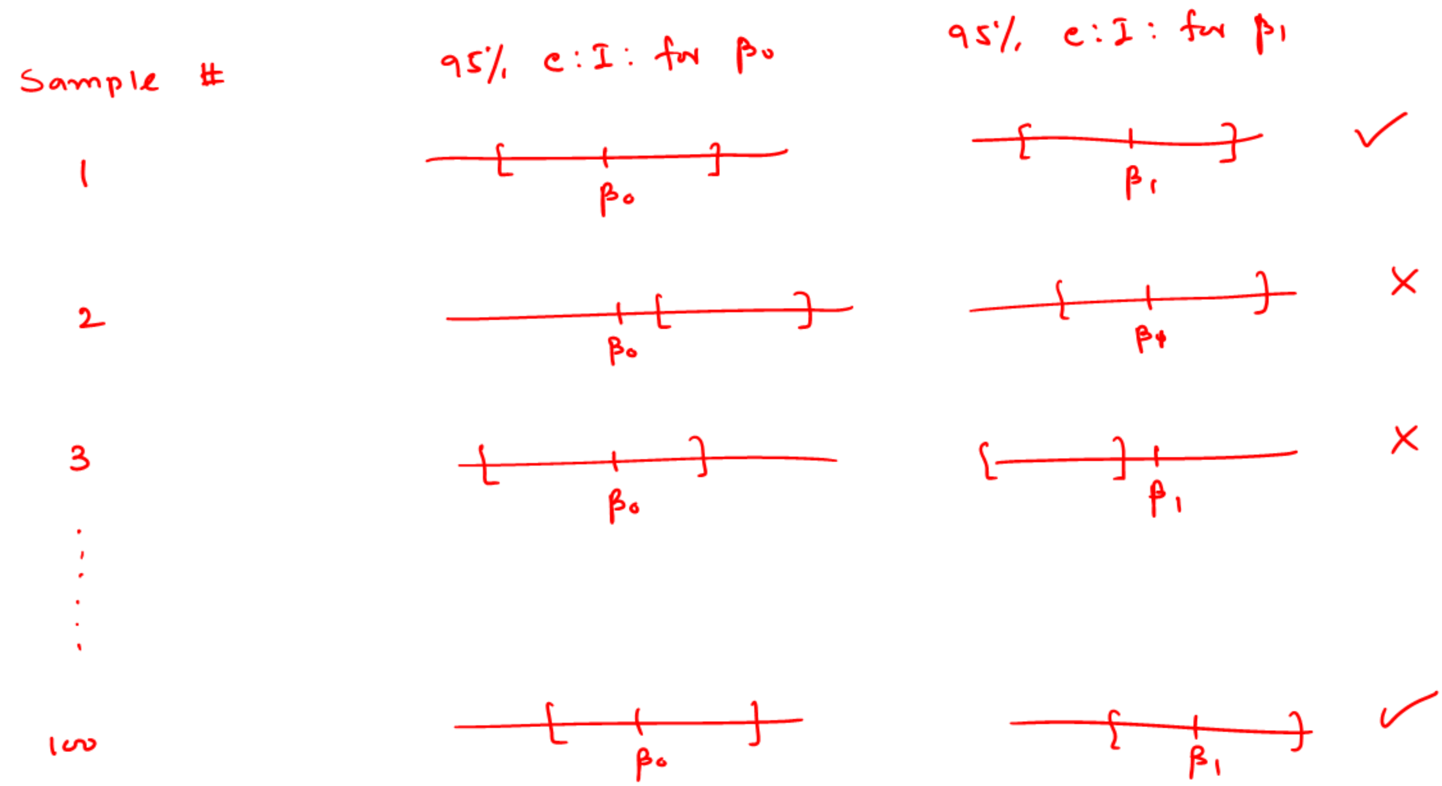


Joint estimation of β_0 and β_1

Joint estimation is needed when we want $1-\alpha$ (eg - 95%) confidence that the conclusions for both β_0 and β_1, are correct.

idea:

consider 95% c:I:s for β_0 and β_1 (seperatly)



So if we consider both β_0 and β_1 together, the confidence coefficient may be less than 95%.

This is called the family confidence coefficient.

Problem: How to find C.I.s for both β_0 and β_1 with family confidence interval $1-\alpha$.

Answer: use Bonferroni method.

* Bonferroni Method

Idea:

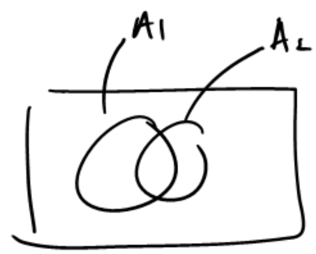
consider $1-\alpha$ C.I. for β_0 and β_1 (Seperatly)

$$b_0 \pm t_{1-\alpha/2 : n-2}\, s\{b_0\} \longrightarrow (1)$$

$$b_1 \pm t_{1-\alpha/2 : n-2}\, s\{b_1\} \longrightarrow (2)$$

then,

$$P(\underbrace{(1) \text{ is not correct}}_{A_1}) = P(A_1) = \alpha$$



$$P(\underbrace{(2) \text{ is not correct}}_{A_2}) = P(A_2) = \alpha$$

$$P(\text{at least one of (1) OR (2) is not correct}) = P(A_1 \cup A_2)$$
$$= P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

$$\Rightarrow P(\text{both correct}) = P(\bar{A}_1 \cap \bar{A}_2)$$
$$= P(\overline{A_1 \cup A_2}) \quad (\text{De Morgan's Law})$$
$$= 1 - \underbrace{P(A_1 \cup A_2)}$$

$$\Rightarrow P(\bar{A}_1 \cap \bar{A}_2) = 1 - P(A_1) - P(A_2) + \underbrace{P(A_1 \cap A_2)}$$

But $P(A_1 \cap A_2) \geq 0$,

$$P(\bar{A}_1 \cap \bar{A}_2) \geq 1 - P(A_1) - P(A_2) \quad (\text{Bonferroni Inequality})$$

$$\Rightarrow P(\text{Both correct}) \geq 1 - P(A_1) - P(A_2)$$
$$= 1 - \alpha - \alpha = 1 - 2\alpha.$$

(lower bound for the joint (family) confidence coefficient).

So if the family confidence limit is $1-\alpha$,

$$1-\underbrace{1-\alpha}_{\text{c: limit}} = \alpha$$

$$\alpha/2 \qquad \alpha/2$$

$$\alpha/2+\beta \qquad \alpha/2-\beta$$

So Bonferroni c:I:s for β_0 and β_1 are the c:Is with confidence coefficient "$1-\alpha/2$".

$$b_0 \pm t_{1-\alpha/4\,:\,n-2}\, S\{b_0\} \quad \text{and}$$

$$b_1 \pm t_{1-\alpha/4\,:\,n-2}\, S\{b_1\}.$$

Note:

More generally $1-\alpha$ family confidence intervals for β_0 and β_1 can be obtained by dividing α into "$\alpha/2-\beta$" and "$\alpha/2+\beta$" $(0 \le \beta \le \alpha/2)$.

ie

$$b_0 \pm t_{1-\frac{(\alpha/2-\beta)}{2}\,:\,n-2}\, S\{b_0\} \quad \text{and}$$

$$b_1 \pm t_{1-\frac{(\alpha/2+\beta)}{2}\,:\,n-2}\, S\{b_1\}$$

Eg:- Suppose family c:c: is 90% then,

$$1-0.9 = 0.1$$

$$\beta_0 \qquad \beta_1$$

$$\alpha/2+\beta = 0.08 \qquad \alpha/2-\beta = 0.02$$

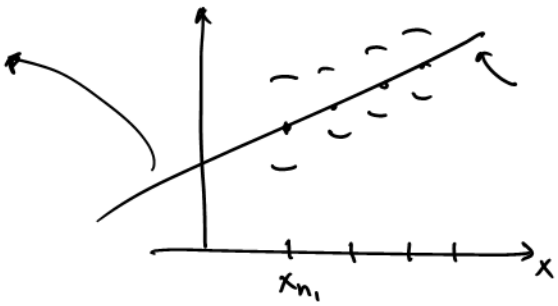
$\beta = 0.03$ (3%).

c:c for β_0 is 92% c:c: for β_1 is 98%.

* This method can be applied for more general regression models. Suppose there are "g" parameters to estimate. If the family confidence coefficient is $1-\alpha$, then confidence coefficient for each parameter is "$1-\alpha/g$".

Simultaneous Estimation for Mean Response

$E[Y] = \beta_0 + \beta_1 X$



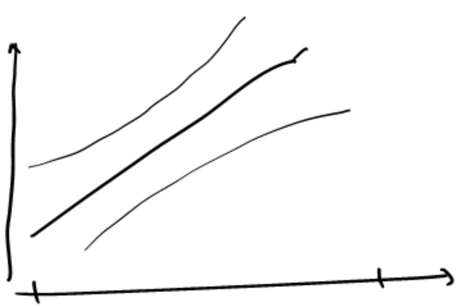
Some times we have to estimate the mean response $E(Y)$ at a numer of X- levels.

Let family confidence coefficient is $1-\alpha$.

There are two methods:

① Working – Hoteling Procedure

This is a more general method and this gives a confidence band for the regression line. (This gives confidence intervals for all the X values in the range with family c:c: $1-\alpha$).



c:I: is given by

$$\hat{Y}_n \pm W S\{\hat{Y}_n\} \text{ for all } X,$$

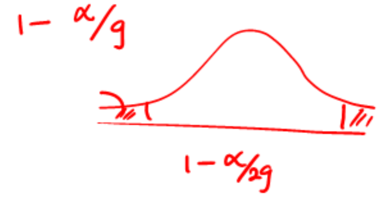
where $W = \sqrt{2F(1-\alpha, 2, n-2)}$.

② Bonferroni Procedure

$1-\alpha$, Bonferroni confidence limits for "g" levels are

$$\hat{Y}_h \pm B\, S\{\hat{Y}_h\},$$

where $B = t\,(1-\alpha/2g;\ n-2)$.



Note:

* Working-Hoteling confidence limits do not change with number of intervals (g). But Bonferroni gets wider with g.

* Both methods provide lower bound for the family confidence coefficients.

* Given a data set, calculate C.Is using both methods (i.e. calculate W and B), and then chose the most efficient one.

Simultaneous Prediction Intervals

Consider the simultaneous prediction limits for "g" new observations at g different X-levels with family C.C. $1-\alpha$.

There are two method:

① Scheffe Procedure

$$\hat{Y}_h \pm S\, S\{pred\},$$

where $S = \sqrt{g\, F_{(1-\alpha;\ g;\ n-2)}}$.

② Bonferroni Procedure

$$\hat{Y}_h \pm B\, S\{pred\}$$

where $B = t_{1-\alpha/2g:\, n-2}$,

* calculate both prediction intervals and then choose the most efficient one.

Regression through the Origin

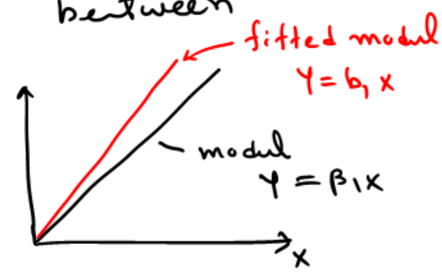
Why important?

$Y = \beta_0 + \beta_1 X$

$Y = \beta_1 X$

There are some applications.

Eg:- Suppose we need to model the association between
X — # of units (output) and
Y — variable cost.



Model:

$$Y_i = \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$\Rightarrow E[Y_i] = \beta_1 X_i$$

Least Square Estimator of β_1

$$Q = \Sigma (Y_i - \beta_1 X_i)^2$$

$$\Rightarrow \frac{\partial Q}{\partial \beta_1} = \Sigma X_i (Y_i - b_1 X_i) \overset{set}{=} 0$$

$$\Rightarrow b_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2}$$

* i^{th} fitted value $= \hat{Y}_i = b_1 X_i$

* i^{th} residual $= e_i = Y_i - \hat{Y}_i = Y_i - b_1 X_i$

* unbiased estimator of error variance

$$S^2 = MSE = \frac{\Sigma (Y_i - \hat{Y}_i)^2}{n-1} = \frac{\Sigma (Y_i - \hat{Y}_i)^2}{n-1}.$$

Note:

* $E(b_1) = \beta_1$

* $var(b_1) = \frac{\sigma^2}{\Sigma X_i^2}$ and $S^2\{b_1\} = \frac{MSE}{\Sigma X_i^2}$

* $E(\hat{Y}_n) = E(Y_n)$

* $\sigma^2\{\hat{Y}_n\} = \frac{X_n^2 \sigma^2}{\Sigma X_i^2}$ and $S^2\{\hat{Y}_n\} = \frac{X_n^2 \, MSE}{\Sigma X_i^2}$

* $\sigma^2\{pred\} = \sigma^2\left\{1 + \frac{X_n^2}{\Sigma X_i^2}\right\}$, and $S^2\{\hat{Y}_n\} = MSE\left(1 + \frac{X_n^2}{\Sigma X_i^2}\right)$.

Proof — HW.

Interval Estimators:

Parameter	Confidence limit
β_1	$b_1 \pm t\, S\{b_1\}$
$E[Y_n]$	$\hat{Y}_n \pm t\, S\{\hat{Y}_n\}$
$\hat{Y}_{n(new)}$	$\hat{Y}_n \pm t\, S\{pred\}$,

where $t = t_{1-\alpha/2 : n-1}$.

Note:

1) $\sum_{i=1}^{n} e_i \neq 0$ (in general).

2) $\sum_{i=1}^{n} X_i e_i = 0$

may be

3) $SSE = \sum e_i^2 \geq SSTO = \sum (Y_i - \bar{Y})^2$

4) $R^2 = 1 - \frac{SSE}{SSTO}$ may be negative.

So residuals can not be used to check the quality of the fit.
We always try to avoid using regression throug the origin.