TRANSFORMATIONS AND WEIGHTING TO CORRECT MODEL INADEQUACIES

- Variance Stabilizing Transformations
- Transformations to linearize the model
- Analytical models to select a Transformation.
- Generalized & weighted Least squares.

The usual approach to deal with inequality of variance is to apply suitable transformation to the response variable or regressor variable, i.e., the case when $V(\epsilon_i) \neq \sigma^2$

Variance Stabilizing Transformation:-

[can be identified from scatter plot of e_i vs. $\hat{y}_i$] (residual)

$V(\epsilon) = \sigma^2 \leftrightarrow$ constant variance assumption.

If the constant variance assumption is violated, the cause is often that the response variable y does not follow a Normal distribution.

Ex.1. $Y \sim \text{Poisson}(\lambda)$

$E(Y) = V(Y) = \lambda$

$Y' = \sqrt{Y}$, then you regress $Y' = \sqrt{Y}$ on x

$V(\sqrt{Y})$ is independent of mean λ.

Ex.2. Y is a proportion $0 < Y < 1$

$Y' = \sin^{-1}(\sqrt{Y})$

Constant variance assumption is violated (double-bow pattern in residual plot)

Y has mean μ and variance σ^2

Situation:- $g(\mu) = \sigma^2$ [Variance depends on mean (not constant)]

Take the following transformation:- $U = f(Y) = f(\mu) + \frac{f'(\mu)}{1!}(Y - \mu)$ [From Taylor's series (next pg.)]

$$V(U) = V(f(Y)) = [f'(\mu)]^2 V(Y) = [f'(\mu)]^2 g(\mu)$$

If we choose the function f such that [since $g(\mu) = \sigma^2$]

$$[f'(\mu)]^2 = \frac{1}{g(\mu)} \Rightarrow f'(\mu) = [g(\mu)]^{-\frac{1}{2}}$$

Then $V(U) = V(f(Y)) = 1$.

$\therefore f(Y)$ is a transformation of Y for which variance becomes constant and this solves the problem on deciding about suitable variance stabilizing transformation for a given problem.

Example:

$$\sigma^2 \propto k\mu^q$$

$$g(\mu) = k\mu^q$$

We want $f'(y) = [g(y)]^{-1/2}$

$$\propto \mu^{-q/2}$$

Thus $f(y) \propto y^{1-q/2}$ if $q \neq 2$

$\log y$ if $q = 2$

Commonly used transformations:

Relationship of σ^2 and $E(Y)$		Transformation to make constant variance
$\sigma^2 \propto$ constant		No transformation
$\sigma^2 \propto E(y)$	$[q=1]$	$f(y) = \sqrt{y}$ [$Y \sim$ Poisson]
$\sigma^2 \propto [E(y)]^2$	$[q=2]$	$f(y) = \log y$ [$Y \sim$ exponential]
$\sigma^2 \propto [E(y)]^3$	$[q=3]$	$f(y) = y^{-1/2}$
$\sigma^2 \propto [E(y)]^4$	$[q=4]$	$f(y) = 1/y$
$\sigma^2 \propto E(y)(1-E(y))$ [when Y is a proportion between 0 and 1]		$f(y) = \sin^{-1}(\sqrt{y})$

[Recall Taylor Series expansion:

Taylor series of a real/complex valued function $f(x)$ that is infinitely differentiable function in a neighborhood of a real/complex number 'a' is

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots]$$

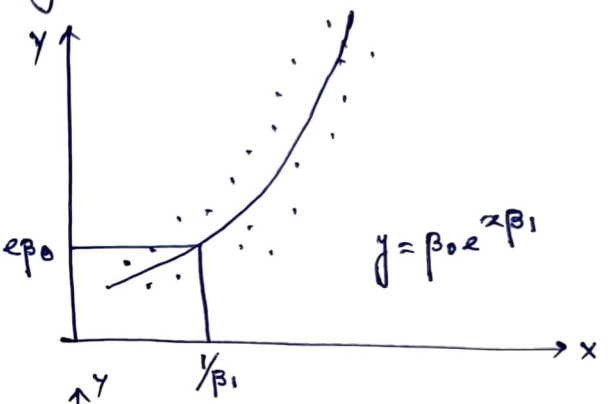
If the relationship between response variable and the regressor variables is NOT linear, we apply transformation to linearize the model:

Transformations to linearize the model:

- Nonlinearity may be detected via scatter plot / residual plot.
- A function can be linearized by using a suitable transformation is called linearizable function.

Example 1: If the scatter plot of y on x suggest an exponential relationship between them, then the appropriate model will be

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



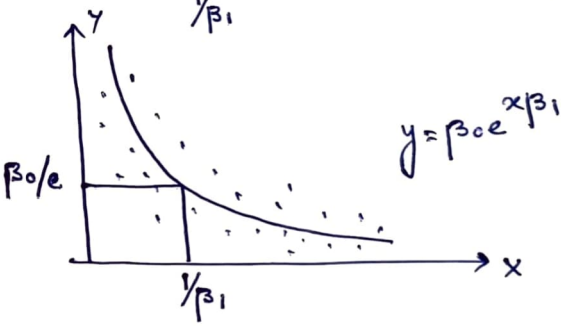
- This model is linear because this is equivalent to the model

$$\log y = \log\beta_0 + \beta_1 x$$

- Transformation: $y' = \log y$

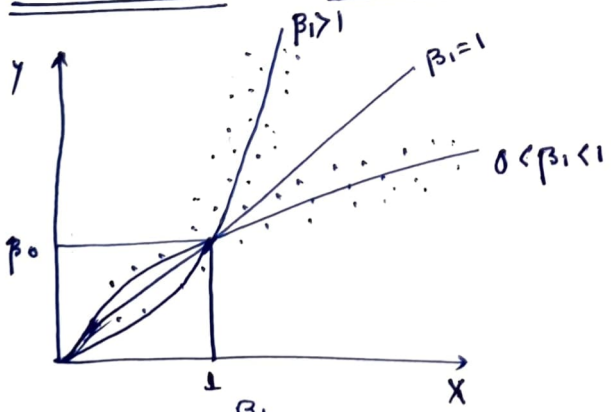
$$y' = \beta_0' + \beta_1 x$$

- Fit a linear model using $(\log y, x)$ transformed data.



Other examples:

Example 2:



- Model: $y = \beta_0 x^{\beta_1}$

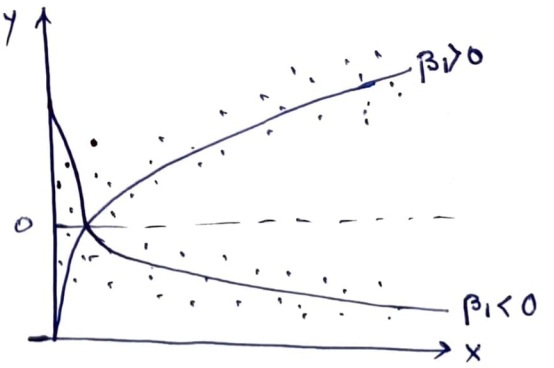
$$\log y = \log\beta_0 + \beta_1 \log x$$

- Transformation: $y' = \log y$ and $x' = \log x$

$$y' = \beta_0' + \beta_1 x'$$

- Fit a linear model using $(\log y, \log x)$ transformed data.

Example 3:

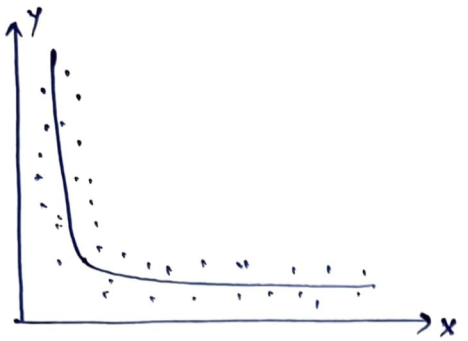


- $y = \beta_0 + \beta_1 \log x \quad (\beta_1 \gtrless 0)$
- Transformation: $x' = \log x$

$$\Rightarrow y = \beta_0 + \beta_1 x'$$

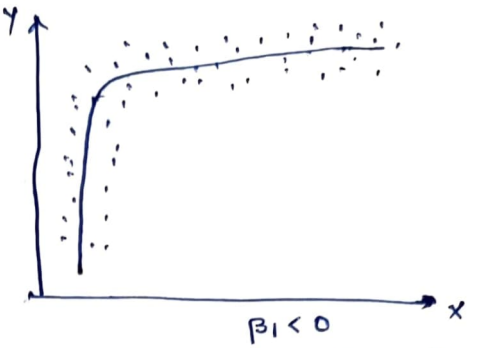
- Fit a linear model with $(\log x, y)$.

Example4:



$\beta_1 > 0$

Example5:



$\beta_1 < 0$

$$y = \frac{x}{\beta_0 x - \beta_1}$$

$$\frac{1}{y} = \frac{\beta_0 x - \beta_1}{x} = \beta_0 - \frac{\beta_1}{x}$$

Tranformation:

$y' = 1/y$, $x' = 1/x$

$\Rightarrow y' = \beta_0 - \beta_1 x'$

Thus, in both the examples, we can fit a linear model using $(1/x, 1/y)$ tranformed data.

Box-Cox METHOD: A usaful class of tranformation is power tranformation Y^λ, where λ is a parameter to be determined.

Major disadvantage of this method is: as $\lambda \to 0$, $Y^\lambda \to 1$, all the response values are equal to 1.

To solve this difficulty, Box-cox method is used:

$$W = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log Y & \text{for } \lambda = 0 \end{cases}$$

[Assuming Y values to be positive]

$\lambda \uparrow$, $W \uparrow$ thus it is impractical to compare reg. models

$\lambda \simeq$ large $\not\Rightarrow W \simeq$ large

To solve this

$$V = \begin{cases} \frac{(Y+\lambda_1)^{\lambda_2} - 1}{\lambda_2} & \text{if } \lambda_2 \neq 0 \\ \log(Y+\lambda_1) & \text{if } \lambda_2 = 0 \end{cases}$$

[Assuming Y values to be negative]

Geometric mean: $\tilde{Y} = \left(\prod_{i=1}^{n} Y_i\right)^{1/n}$ is used as normalization factor.

$$W' = \begin{cases} \frac{Y^\lambda - 1}{\lambda \tilde{Y}^{\lambda-1}} & \text{for } \lambda \neq 0 \\ \tilde{Y} \log Y & \text{for } \lambda = 0 \end{cases}$$

Given data: (x_i, Y_i)

$(Y_1, \ldots, Y_n) \to (W_1', W_2', \ldots, W_n')$

and use $W' = X\beta + \epsilon$ by OLS for any specified value of λ.

Maximum Likelihood Method for estimating λ: →

- Choose a value of λ from $[-2, 2]$ at first and extend the range later, if necessary.
- For each chosen λ value, estimate Y and compute $SS_{Res}(\lambda)$ for the regression model $W' = X\beta + \epsilon$.
- The MLE of λ corresponds to the value of λ for which $SS_{Res}(\lambda)$ is minimum.

WEIGHTED LEAST SQUARES METHOD:

Linear regression model with "non-constant variance" can be fitted by the method of weighted least squares (WLS).

Simple linear regression model: $Y = \beta_0 + \beta_1 x + \epsilon$ $\quad [Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_i^2)]$

The weighted least square function is $S = \sum w_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum w_i e_i^2$

and $w_i \propto \frac{1}{\sigma_i^2}$

Normal equations are:

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum w_i y_i = \hat{\beta}_0 \sum w_i + \hat{\beta}_1 \sum w_i x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum w_i y_i x_i = \hat{\beta}_0 \sum w_i x_i + \hat{\beta}_1 \sum w_i x_i^2$$

Question: Why $w_i \propto \frac{1}{\sigma_i^2}$ and how to know σ_i^2? [NEXT PG.]

GENERALISED LEAST SQUARES METHOD:

- Weighted least square for multiple regression model.
- Consider the same model $Y = X\beta + \epsilon$ with $E(\epsilon) = 0$ and $V(\epsilon) = V\sigma^2$ and $V(\epsilon) \neq I\sigma^2$ (constant variance), where V is a positive definite matrix.
- This happens when: — observations Y have unequal variances and/or
  — observations are correlated.
- In either case, the condition of Gauss-markov theorem are violated. So, $\hat{\beta} = (X'X)^{-1}X'Y$ is NOT BLUE.
- However, it is possible to find BLUE of β for arbitrary positive definite V by suitable linear transformation on the model.

$$(X, Y) \rightarrow (GX, GY)$$

$$GY = GX\beta + G\epsilon$$

$V(G\epsilon) = \sigma^2 GVG'$ since $V(\epsilon) = V\sigma^2$.

$V(G\epsilon)$ is constant if $GVG' = I$. Then $V(G\epsilon) = \sigma^2 I$.

Therefore if we choose G such that $GVG' = I$ then the transformed data satisfy Gauss-Markov conditions and the BLUE of β is obtained by OLS estimation of transformed data.

$GVG' = I \Leftrightarrow V^{-1} = G'G$

$\Leftrightarrow V = G^{-1}G'^{-1}$

$(X, Y) \rightarrow (GX, GY)$

$\hat{\beta} = (X'G'GX)^{-1}X'GGY$

$= (X'V^{-1}X)^{-1}X'V^{-1}Y$

$V(\hat{\beta}) = \sigma^2 (X'G'GX)^{-1} = \sigma^2 (X'V^{-1}X)^{-1}$.

- It is always possible to find a symmetric G_1 by using the orthogonal decomposition of positive definite V

$$V = U' \Lambda^{-1} U,$$

where Λ is the diagonal matrix of eigenvalues and U is the matrix of eigenvectors.

$$V^{-1} = U^{-1} \Lambda U'^{-1}$$

$$\Leftrightarrow V^{-1} = G_1' G_1 \text{ ; where } G_1 = \Lambda^{\frac{1}{2}} U'^{-1} \quad [\text{choice of } G_1]$$

- Thus, if the data does not satisfy the assumption of constant variance and if the errors/observations are correlated then $V(\epsilon) = V\sigma^2$, then we can find a transformation of the data $(G_1 X, G_1 Y)$ which satisfy all the assumptions and we can apply OLS technique on the transformed data.

- WLS technique as a particular case of GLS :→

— Observations are uncorrelated and have unequal variance.

$$V(\epsilon) = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & & \sigma_n^2 \end{pmatrix} = V$$

$$V^{-1} = G'G \Rightarrow G_1 = V^{-1/2} = \begin{pmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n \end{pmatrix}$$

$$= \text{Diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n}\right); \quad W = V^{-1}$$

$$w_i \propto \frac{1}{\sigma_i^2}$$

$$\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}Y$$

$$= (X'WX)^{-1}(X'WY)$$

$$\text{Var}(\hat{\beta}) = (X'V^{-1}X)^{-1} = (X'WX)^{-1}.$$

- GLS technique is used to deal with non-constant variance in the response variable or observations are correlated.

EXAMPLE:

Suppose we have n observations of variables $X_1, X_2, \ldots, X_k, Y$, where X_i's are predictor variables and Y is a response variable. Suppose we are told that observations Y_i are uncorrelated but the last observation has variance $4\sigma^2$ rather than σ^2. Find the best linear unbiased estimator (BLUE) of β using weighted least square?

Solution:

$$V(Y_i) = V(\epsilon_i) = \sigma^2, \quad i = 1(1)n-1, \quad V(Y_n) = V(\epsilon_n) = 4\sigma^2$$

$$V(\epsilon) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 4\sigma^2 \end{pmatrix} = V(\sigma^2) \text{ where}$$

$$V = \text{diag}(1, 1, \ldots, 1, 4).$$

$$Y = X\beta + \epsilon \quad \text{where } \text{Var}(Y_i) = V\sigma^2$$

$$GY = GX\beta + G\epsilon$$

$$V(G\epsilon) = \sigma^2 GVG' = \sigma^2 I$$

$$GVG' = I$$

$$\Rightarrow V^{-1} = G'G.$$

$$\hat{\beta} = (X'G'GX)^{-1} X'G'GY \text{ is BLUE}$$

$$= (X'V^{-1}X)^{-1} X'V^{-1}Y$$

$$= \left(X' \,\text{diag}(1, \ldots, 1, 1/4)\, X\right)^{-1} \left(X' \,\text{diag}(1, 1, \ldots, 1/4)\, Y\right).$$

Generalized Linear Model (GLM)

- GLM analysis comes into account when error distribution is not normal (response is not normal) and/or (but must be a member of exponential family) when a vector of non-linear functions of the responses $\eta(Y) = (\eta(Y_1), \eta(Y_2), \ldots, \eta(Y_n))'$ are not Y itself, has expectation $X\beta$.

In standard MLR :- $Y = X\beta + \epsilon$

$E(Y) = X\beta$, $E(\epsilon) = 0$

$\therefore$ Here we assume $\boxed{E(\eta(Y)) = X\beta}$

- In GLM, the response variable distribution must be a member of the exponential family.
- **The exponential family of distribution :-**

A random variable u that belongs to exponential family with single parameter θ has a pdf

$$f(u,\theta) = s(u)\, t(\theta)\, e^{a(u)\, b(\theta)},$$

where s, t, a, b are all known functions.

Rewrite: $f(u,\theta) = \exp\{a(u)b(\theta) + d(u) + c(\theta)\}$,

where $d(u) = \ln(s(u))$, $c(\theta) = \ln(t(\theta))$ and when $a(u) = u$, the distn is said to be in canonical form.

Note: $b(\theta)$ is called natural parameter, parameters other than the parameter of interest (θ) are called nuisance parameter.

- **Some member of exponential family :-**

1 **Normal distribution :-** $N(\mu, \sigma^2)$ and parameter of interest : μ. (natural parameter)

$$f(u,\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2}; \quad -\infty \le u \le \infty$$

$$= \exp\left\{u \cdot \frac{\mu}{\sigma^2} + \left[-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2\right] - \frac{u^2}{2\sigma^2}\right\}$$

where, $a(u) = u$, $b(\theta) = \frac{\mu}{\sigma^2}$, $c(\theta) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2$, $d(u) = -\frac{u^2}{2\sigma^2}$.

2 **Binomial distribution :-** $Bin(n,p)$ and p is the parameter of interest and n is the nuisance parameter

$$f(u,p) = \binom{n}{u} p^u (1-p)^{n-u}, \quad u = 0, 1, \ldots, n$$

$$= \binom{n}{u}\left(\frac{p}{1-p}\right)^u (1-p)^n$$

$$= \exp\left\{u \ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln\binom{n}{u}\right\}$$

where, $a(u) = u$, $b(\theta) = \ln\left(\frac{p}{1-p}\right)$, $c(\theta) = n\ln(1-p)$, $d(u) = \ln\binom{n}{u}$.

Similarly, we can show Poisson distribution, Gamma distribution, exponential distribution, Negative Binomial distn., etc. belong to exponential family of distribution.

Expected value and variance of $a(u)$:-

$$E(a(u)) = -\frac{c'(\theta)}{b'(\theta)}, \quad V(a(u)) = \frac{b''(\theta)\,c'(\theta) - c''(\theta)\,b'(\theta)}{[b'(\theta)]^3} \quad (*)$$

- **Fitting GLM :-** Suppose we have a set of independent obsn.s. $(Y_i, \tilde{x}_i')$, $i = 1(1)n$, $\tilde{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ from some exponential type distribution of canonical form [i.e., $a(Y) = Y$]. the joint pdf is

$$f(Y_1, Y_2, \ldots, Y_n, \theta, \phi) = \exp\left\{\sum_{i=1}^{n} Y_i b(\theta) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(Y_i)\right\},$$

where ϕ is a vector of a nuisance parameters that occur with in $b(\cdot)$, $c(\cdot)$, & $d(\cdot)$.

Here $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$! vector of parameters of interest.

- The variation in response variable (Y_i) can be explained in terms of $\tilde{x}_i$ values, i.e., $\tilde{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$
- Consider the set of parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ (vector of regression coefficients)
  We find some suitable link function $g(\cdot)$ such that

$$g(\mu_i) = \tilde{x}_i'\beta \quad | \quad y_i = \tilde{x}_i'\beta + \epsilon \quad \left(\begin{array}{l}\text{standard MLR}\\ Y_i \sim \text{Normal}\\ g(\theta) = \theta\end{array}\right)$$
$$g(E(y_i)) = \tilde{x}_i'\beta \quad | \quad E(y_i) = \tilde{x}_i'\beta$$

- A link function that is often regarded as sensible one is natural parameter.

Example :- Binomial Distribution

Suppose we have data $(Y_i, \tilde{x}_i')$ from a binomial distn $Bin(n_i, P_i)$ The single observation Y_i is of the form $\frac{n_i}{n_i}$, where n_i is the no. of success in n_i trials, each having prob. P_i of success (parameter of interest) and $\tilde{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a set of observations of p regressors associated with Y_i. Binomial distn is a member of exp. family. (n_i is the nuisance parameter)

[(*) Example: For Binomial distribution: $a(\theta) = u$, $b(\theta) = \ln\left(\frac{p}{1-p}\right)$, $c(\theta) = n\ln(1-p)$

$$E(a(u)) = E(u) = -\frac{c'(\theta)}{b'(\theta)} = \frac{n}{1-p} \times p(1-p) = np.$$

$$\text{Var}(a(u)) = V(u) = \frac{b''(\theta)\,c'(\theta) - c''(\theta)\,b'(\theta)}{[b'(\theta)]^3} = np(1-p).]$$

Joint pdf $= f(Y_1, Y_2, \ldots, Y_n)$

$$= \prod_{i=1}^{n} \binom{n_i}{Y_i} p_i^{Y_i} (1-p_i)^{n_i - Y_i}$$

$$= \prod_{i=1}^{n} \exp\left\{ Y_i \ln\left(\frac{p_i}{1-p_i}\right) + n_i \ln(1-p_i) + \ln\binom{n_i}{Y_i} \right\}$$

$$= \exp\left\{ \sum_{i=1}^{n} Y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} n_i \ln(1-p_i) + \sum_{1}^{n} \ln\binom{n_i}{Y_i} \right\}$$

Natural parameter $= \ln\left(\frac{p_i}{1-p_i}\right)$.

We would hope that the variation in the $Y_i / E(Y_i) = p_i$ could be explained in terms of $\tilde{x}_i'$ values, i.e., we could hope that we could find a suitable link function $g(\cdot)$ such that $g(p_i) = \tilde{x}_i' \beta$.

we fit the model $\ln\left(\frac{p_i}{1-p_i}\right) = \tilde{x}_i' \beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

$$E(y_i) = p_i = \frac{\exp(\tilde{x}_i' \beta)}{1 + \exp(\tilde{x}_i' \beta)}$$

Instead of fitting $y_i = \tilde{x}_i' \beta + \epsilon$ we fit here

$y_i = p_i + \epsilon$,

where $\tilde{x}_i' \beta = \beta_1 + \beta_2 x_{i2}$ ——— (*) is called the logistic function.
$[x_{i1} = 1]$

Estimation via ML function:- To estimate β, we use maximum likelihood method:

Likelihood funct; $L = \exp\left\{ \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} n_i \ln(1-p_i) + \sum_{i=1}^{n} \ln\binom{n_i}{y_i} \right\}$

$$\ln L = \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} n_i \ln(1-p_i) + \sum_{i=1}^{n} \ln\binom{n_i}{y_i}$$

$$= \sum y_i \tilde{x}_i' \beta - \sum_{1}^{n} n_i \ln(1 + \exp(\tilde{x}_i' \beta)) + \sum_{1}^{n} \ln\binom{n_i}{y_i}$$

Maximize $\ln L$ w.r.t. β, use numerical search/iteratively reweighted least square (IRLS) could be used to compute MLE of β.

Choice of Link function:-

Normal :	$g(\mu) = \mu$	(Identity link)
Binomial :-	$g(p) = \ln\left(\frac{p}{1-p}\right)$	(logistic link)
Poisson :-	$g(\mu) = \ln \mu$	(log link)
Gamma/Exponential :	$g(\mu) = \frac{1}{\mu}$	(reciprocal link)

Example (Pneumoconiosis Data)
Lung disease result from breathing & dust in coalmines

Number of years of exposure (X_i)	Number of severe cases	Total number of miners	Proportion of severe cases (Y_i)
5.8	0	98	0
15.0	1	54	0.0185
21.5	3	43	0.0698
27.5	8	48	0.1667
33.5	9	51	0.1765
39.5	8	38	0.2105
46.0	10	28	0.3571
51.5	5	11	0.4545

Check the variation in Y_i can be explained in terms of the number of years of exposure?

Sol. Probability distn. for the number of severe cases is binomial.
We will fit a logistic regression model to the data.

$$p_i = E(Y_i) = \frac{\exp(\tilde{x}_i'\beta)}{1+\exp(\tilde{x}_i'\beta)}$$

Here $\tilde{x}_i'\beta = \beta_1 + \beta_2 x$, where x: # years of exposure.

$$\hat{y}_i = \frac{\exp(4.79 - 0.0935x)}{1+\exp(4.79-0.0935x)}.$$

Poisson Regression Model:

Data $(Y_i, \tilde{x}_i')$ from $P(\mu_i)$, $E(Y_i) = \mu_i$.

$$f(y,\mu) = \exp\{y\ln\mu - \mu - \ln y!\}$$ where $\ln\mu$ is the natural parameter, i.e., $g(\mu) = \ln\mu$ is the link function.

The variation in Y_i could be explained in terms of the $\tilde{x}_i'$ values. We fit the model: $g(\mu_i) = \tilde{x}_i'\beta$

$$\ln\mu_i = \tilde{x}_i'\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

$\mu_i = e^{\tilde{x}_i'\beta}$; Final model to fit.

$E(Y_i) = e^{\tilde{x}_i'\beta}$. Same as writing $Y_i = e^{\tilde{x}_i'\beta} + \epsilon$ [Final model]

In SLR: $Y = \beta_0 + \beta_1 X + \epsilon$; Assumption: $\epsilon_i \overset{IID}{\sim} N(0, \sigma^2)$

Y: random variable
X_i: controlled variable/deterministic variable.

Two variations of situation could be:
(a) Response and the regression variables are jointly distributed RVs.
(b) there are measurement errors in regressors.

Measurement errors & Calibration Problem

- Case where Response & Regressors are jointly distributed RVs.
- Measurement Errors in Regressors
- The Calibration problem (inverse problem)

① X & Y are jointly normally distributed:

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y-\mu_1}{\sigma_1}\right)\left(\frac{x-\mu_2}{\sigma_2}\right)\right]\right\}$$

$E(Y) = \mu_1$, $V(Y) = \sigma_1^2$, $E(X) = \mu_2$, $V(X) = \sigma_2^2$

and $\rho = \frac{E[(Y-\mu_1)(X-\mu_2)]}{\sigma_1\sigma_2} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ is correlation coefficient between Y and X.

∴ The conditional distn. of Y given X is

$$Y|X \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x-\mu_2),\ (1-\rho^2)\sigma_1^2\right).$$

$$E(Y|X) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x-\mu_2)$$

$\boxed{E(Y|x) = \beta_0 + \beta_1 x}$; where $\beta_0 = \mu_1 - \rho\mu_2\frac{\sigma_1}{\sigma_2}$, $\beta_1 = \rho\frac{\sigma_1}{\sigma_2}$ → when both X and Y are RVs.

Model when X is a deterministic variable:
$E(Y) = \beta_0 + \beta_1 x$
$Y = \beta_0 + \beta_1 x + \epsilon$; $\epsilon \sim N(0, \sigma^2)$

$$Y_i|x_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 x_i,\ (1-\rho^2)\sigma_1^2) \quad [\text{Model: } E(Y|x) = \beta_0 + \beta_1 x]$$

MLE:-
$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}(y_i - \beta_0 - \beta_1 x_i)^2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}}\right)^n e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}\sum(y_i - \beta_0 - \beta_1 x_i)^2}$$

[Max L ≡ Min $\sum(y_i - \beta_0 - \beta_1 x_i)^2$]

We find β_0 & β_1 ∋ $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ is minimum.

We know (from OLS) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ and $\hat{\beta}_1 = \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \sqrt{\frac{SS_T}{S_{xx}}}.r$

identical to those given by LSE in case where x is a controlled variable.

$\rho = \frac{Cov(X,Y)}{\sigma_1\sigma_2}$; estimate of ρ is $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ = sample correlation coefficient.

To test $H_0: \rho = 0$ Vs $H_1: \rho \neq 0$:

Test statistic is $t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$, under H_0.

We Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$. When $\rho = 0$, there is no relationship between X and Y.

[$S_{yy} = SS_T$]

Measurement errors in Regressions :-

We wish to fit the simple linear reg. model, but the regressors is measured with error.

$$X_i = x_i + a_i \quad ; \quad E(X_i) = x_i \quad ; \quad E(a_i \epsilon_i) = 0$$

X_i : Observed value, x_i : true value, a_i : measurement errors

Assume: $E(a_i) = 0$ and $V(a_i) = \sigma_a^2$. Also, a_i and ϵ_i are independent.

The response variable is subject to the usual errors ϵ_i

The reg. model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $(i = 1(1)n)$. [Given data: (y_i, X_i)]

$$= \beta_0 + \beta_1 (X_i - a_i) + \epsilon_i$$

$$= \beta_0 + \beta_1 X_i + (\epsilon_i - \beta_1 a_i)$$

$$\boxed{y_i = \beta_0 + \beta_1 X_i + \gamma_i} \quad ; \quad \gamma_i = \epsilon_i - \beta_1 a_i$$

Since X_i and γ_i both are RVs, then we need to compute $Cov(X_i, \gamma_i)$

$$Cov(X_i, \gamma_i) = E\left[(X_i - E(X_i))(\gamma_i) \right]$$

$$= E\left[(X_i - x_i)(\epsilon_i - \beta_1 a_i) \right]$$

$$= E\left[a_i (\epsilon_i - \beta_1 a_i) \right]$$

$$= -\beta_1 E(a_i^2)$$

$$= -\beta_1 \sigma_a^2$$

If we apply standard LSM to the data, the estimates of the model parameters are no longer unbiased.

Usual case : $\hat{\beta}_1 = \dfrac{\sum (y_i - \bar{y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$

$$E(\hat{\beta}_1) = \frac{\beta_1}{1+\theta}$$

where $\theta = \dfrac{\sigma_a^2}{\sigma_x^2}$, [Hence in the presence of measurement errors]

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$\hat{\beta}_1$ is a biased estimator of β_1, unless $\sigma_a^2 = 0$, that is there is no measurement errors in regressors.

- If σ_a^2 is small relative to σ_x^2, the bias will be small [$\theta \equiv$ small].
- If variability in the measurement errors is small relative to the variability of the x's; then measurement error can be ignored & OLS method can be applied directly.

Example (Measurement Error Problem)

X : current flow in an electric circuit

Current flow is measured with an ammeter which is not completely accurate. Here, a measurement error is experienced.

$$X = x + a$$

x ↓ true current flow

The Calibration Problem:→

- Given an observed value of y, say y_0, determine the x value corresponding to it. (Inverse problem)
- **Example:** We know the temperature reading given by thermocouple is a linear function of the actual temperature.

  Observed temp $= \beta_0 + \beta_1$ (actual temp) $+ \epsilon$
- Solution:- Suppose we have (y_i, x_i), $i = 1(1)n$.

  First, fit the model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

  Let y_0 be the observed value of y. A natural point estimation of the corresponding value of x is $\hat{x}_0 = \dfrac{y_0 - \hat{\beta}_0}{\hat{\beta}_1}$, assuming that $\hat{\beta}_1 \neq 0$.