Linear Regression

Mohammad Arshad Rahman
(webpage: https://www.arshadrahman.com)

Course: Econometrics
Chennai Mathematical Institute

Classical Linear Regression Model

A regression model specifies the relationship between dependent variable y (continuous for CLRM) and one or more independent variables: $(x_1 \equiv 1, x_2, \ldots, x_k)$.

Example 1: Regression is fundamental to the Capital Asset Pricing Model (CAPM). The CAPM equation determines the relationship between the expected return of an asset and market risk premium.

Example 2: When forecasting financial statements for a company, multiple regression analysis is employed to determine how changes in drivers of the business (such as number of employees, revenue, etc.) will impact revenue or expenses in the future.

Classical Linear Regression Model

For a single observation i (=person, firm, household, etc.), the multiple linear regression model can be written as,

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k + \epsilon_i.$$

The β terms are fixed unknown population parameters, also known as regression coefficients. β_1 is the intercept or constant term, and $(\beta_2, \ldots, \beta_k)$ are slope coefficients.

The term $(\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k)$ is deterministic, while ϵ_i is the stochastic or random component.

Classical Linear Regression Model

If we stack the model for all individuals, then the linear regression model can be expressed in matrix formulation as:

$$y = X\beta + \epsilon, \tag{1}$$

where $y = (y_1, ..., y_n)'$ is a column vector of dimension $n \times 1$, X is a matrix of dimension $n \times k$ defined as,

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

Classical Linear Regression Model

$\beta = (\beta_1, \cdots, \beta_k)'$ is a $k \times 1$ column vector of unknown parameters, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)'$ is an $n \times 1$ vector of errors.

The component $X\beta$ is deterministic, while ϵ is random. Consequently, y is a random variable.

The objective of regression is to explain the variation in dependent variables (y) with respect to the variation in independent variables (x's).

Assumptions

The CLRM rests on 4 key assumptions. An optional 5th assumption (normality of the error term) can be added in cases where a fully parametric approach is desired or needed.

1) The stipulated relationship is linear in parameters (i.e., in β). This does not mean that the elements of the data matrix X have to be linear. It is okay to have log-terms, squared terms, etc. in the X matrix. Certain nonlinear forms of y are also permissible, as long as they can be transformed to linearity. For example, the following common model is permissible in CLRM:

$$y_i = \exp(x_i'\beta + \epsilon_i) \longleftrightarrow \ln(y_i) = x_i'\beta + \epsilon_i.$$

Assumptions

The log-transformation on both sides preserves linearity. If all x_i's are logged as well, the model is called "log-log" or "log-linear".

If all x_i's are linear, the model is called "semilog". A mix of logged and linear regressors is also possible.

Note, logging either sides changes the interpretation of marginal effects.

Assumptions

2) No perfect multicollinearity. The data matrix X has dimension $n \times k, (n > k)$ and has full column rank. In other words, all variables in X are linearly independent. This is necessary for estimation of the parameters, else there would be an infinite number of solutions for the estimate of β.

3) Expectation of all error terms, conditional on X is zero, i.e.

$$E(\epsilon|X) = \begin{bmatrix} E(\epsilon_1|X) \\ E(\epsilon_2|X) \\ \vdots \\ E(\epsilon_n|X) \end{bmatrix} = 0,$$

Assumptions

In words, Assumption 3 implies that no element of the data matrix contains any information about the expectation of any error term. So, the independent variables are exogenous and ϵ is a true vector of "unknown" or "unobservable" effects.

If this assumption is violated, our estimated parameters will "pick up" undesired effects from "the error term", thus producing misleading results. This is the infamous "omitted variable (OV)" problem.

Assumptions

This assumption also implies the following relationships:
a) $E(\epsilon_i) = 0$,
b) $Cov(\epsilon_i, X) = 0$, and
c) $E(y|X) = X\beta$.

a) Unconditional Expectation: By Assumption 3 (A3), we have $E_{\epsilon_i}(\epsilon_i|x_i) = 0$ for all i's. So, by the law of iterated expectation we have the unconditional expectation as,

$$E(\epsilon_i) = E_{x_i}\left[E_{\epsilon_i}(\epsilon_i|x_i)\right] = E_{x_i}(0) = 0.$$

b) Zero Covariance: A3 is often interpreted as "ϵ_i and x_i" are uncorrelated i.e. have a covariance of zero. Lets check if this correct i.e. if $E_{\epsilon_i}(\epsilon_i|x_i) = 0 \rightarrow cov(\epsilon_i, x_i) = 0$.

$$
\begin{aligned}
Cov(\epsilon_i, x_i) &= E_{\epsilon_i, x_i}\left[(\epsilon_i - E(\epsilon_i)) * (x_i - E(x_i))\right] \\
&= E_{\epsilon_i, x_i}\left[\epsilon_i x_i\right] - E_{\epsilon_i, x_i}\left[E(\epsilon_i)x_i\right] - E_{\epsilon_i, x_i}\left[\epsilon_i E(x_i)\right] + E(\epsilon_i)E(x_i) \\
&= E_{x_i}\left[x_i E_{\epsilon_i}(\epsilon_i|x_i)\right] - E_{x_i}\left[E(\epsilon_i)x_i\right] - E_{x_i}\left[E(x_i)E_{\epsilon_i}(\epsilon_i|x_i)\right] + E(\epsilon_i)E(x_i) \\
&= E_{x_i}\left[x_i E_{\epsilon_i}(\epsilon_i|x_i)\right] - E_{x_i}\left[E(\epsilon_i)x_i\right] - E_{x_i}\left[E(x_i)E_{\epsilon_i}(\epsilon_i|x_i)\right] + E(\epsilon_i)E(x_i) \\
&= E_{x_i}\left[x_i E_{\epsilon_i}(\epsilon_i|x_i)\right] - E(\epsilon_i)E(x_i) - E_{x_i}\left[E(x_i)E_{\epsilon_i}(\epsilon_i|x_i)\right] + E(\epsilon_i)E(x_i) \\
&= E_{x_i}\left[x_i E_{\epsilon_i}(\epsilon_i|x_i)\right] - E_{x_i}\left[E(x_i)E_{\epsilon_i}(\epsilon_i|x_i)\right] \\
&= E_{x_i}\left[x_i E_{\epsilon_i}(\epsilon_i|x_i)\right] - E_{x_i}\left[E(x_i)E_{\epsilon_i}(\epsilon_i|x_i)\right] \\
&= E_{x_i}\left[x_i E_{\epsilon_i}(\epsilon_i|x_i) - E(x_i)E_{\epsilon_i}(\epsilon_i|x_i)\right] \\
&= E_{x_i}\left[E_{\epsilon_i}\left((\epsilon_i|x_i)(x_i - E(x_i))\right] = 0,
\end{aligned}
$$

since $E_{\epsilon_i}(\epsilon_i|x_i) = 0$.

Assumptions

c) Regression assumption: From A3 it also follows that,

$$\begin{aligned} E(y_i|x_i) &= E\left[x_i\beta + \epsilon_i|x_i\right] \\ &= E(x_i\beta) + E(\epsilon_i|x_i) \\ &= E(x_i\beta) + 0 = x_i\beta, \end{aligned}$$

i.e., we have indeed a true regression - relationship. For this reason A3 is often referred to as the "regression" assumption.

Assumptions

4) Homoscedasticity implies that the errors ϵ_i have equal variance i.e, $V(\epsilon_i) = \sigma^2$, for all $i = 1, 2, \cdots, n$.

So we assume all n error terms are "drawn" from the same distribution with mean 0 and variance σ^2.

Non-autocorrelation: Error terms are uncorrelated which implies, $Cov(\epsilon_i, \epsilon_j) = E\left[(\epsilon_i - E(\epsilon_i))(\epsilon_j - E(\epsilon_j))\right] = E[\epsilon_i \epsilon_j] = 0$.

Violation of these assumptions are called "heteroskedasticity" and "autocorrelation", respectively.

Assumptions

For the full model $y = X\beta + \epsilon$, Assumption 4 can be written as,

$$V(\epsilon) = E\begin{bmatrix} \epsilon_1\epsilon_1 & \epsilon_1\epsilon_2 & \dots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2\epsilon_2 & \dots & \epsilon_2\epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \dots & \epsilon_n\epsilon_n \end{bmatrix} = E\begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}.$$

Disturbances that satisfy this property are called "spherical".

For the dependent variable, this implies:

$$\begin{aligned} V(y|X) &= V(X\beta + \epsilon) = V(X\beta) + V(\epsilon) + 2Cov(X\beta, \epsilon) \\ &= V(\epsilon) = \sigma^2 I. \end{aligned}$$

Figure 1: Homoskedasticity

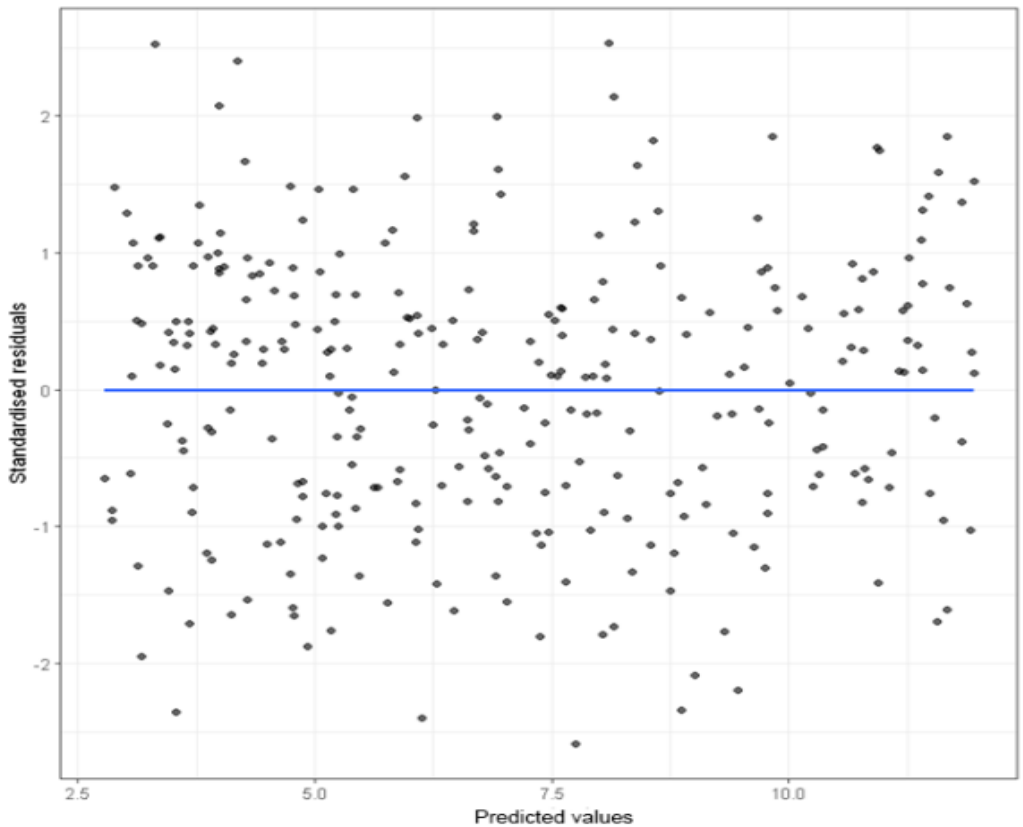
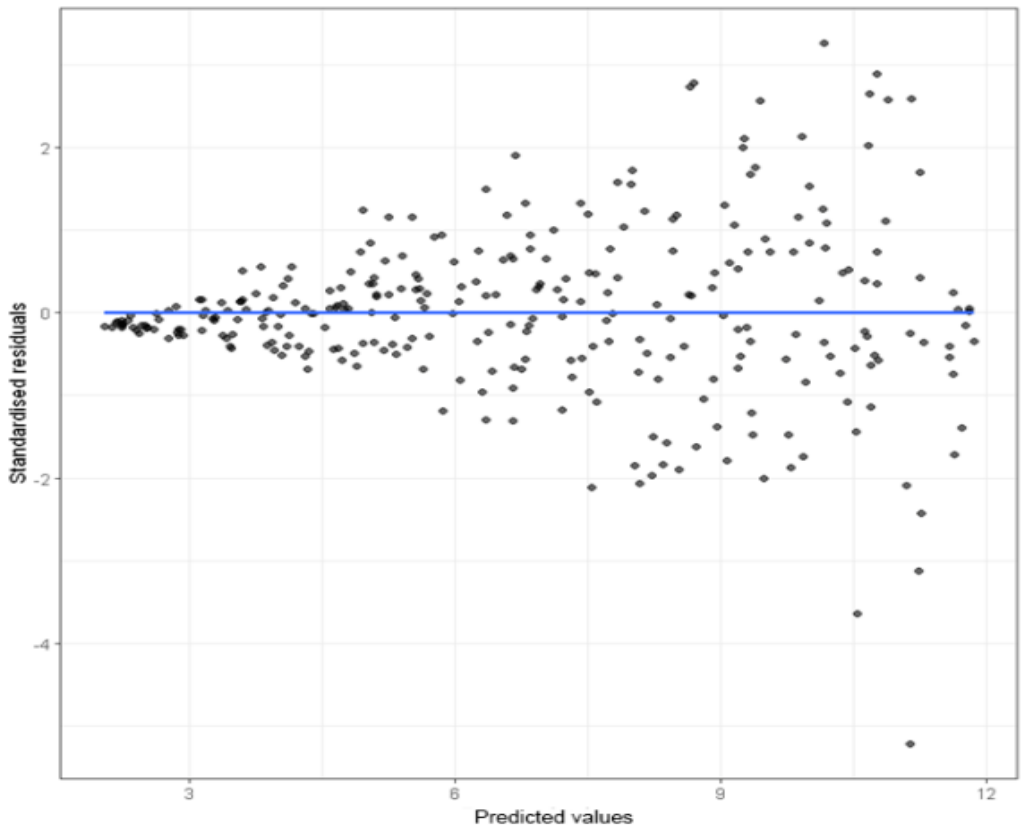


Figure 2: Heteroscedasticity



Assumptions

5) The errors follow a normal distribution, i.e. $\epsilon_i \sim N(0, \sigma^2)$ or in vector form $\epsilon \sim N_n(0, \sigma^2 I)$. For the dependent variable, this implies $y \sim N_n(X\beta, \sigma^2 I)$.

The normality assumption is not necessary to derive the estimates for β, but is needed when a fully parametric approach is desired, for deriving some exact statistical results for estimators and to construct certain test statistics.

Gauss-Markov Theorem

Gauss-Markov Theorem: In the linear regression model (1), if the errors satisfy Assumptions 1-4, then the least squares estimator is the best linear unbiased estimator (BLUE).

Note that the errors need not be normal nor do they need to *independently and identically distributed*.

If we make the normality assumption, then the OLS estimator is minimum variance unbiased estimator (MVUE), which is stronger than BLUE.

Statistical Properties of X

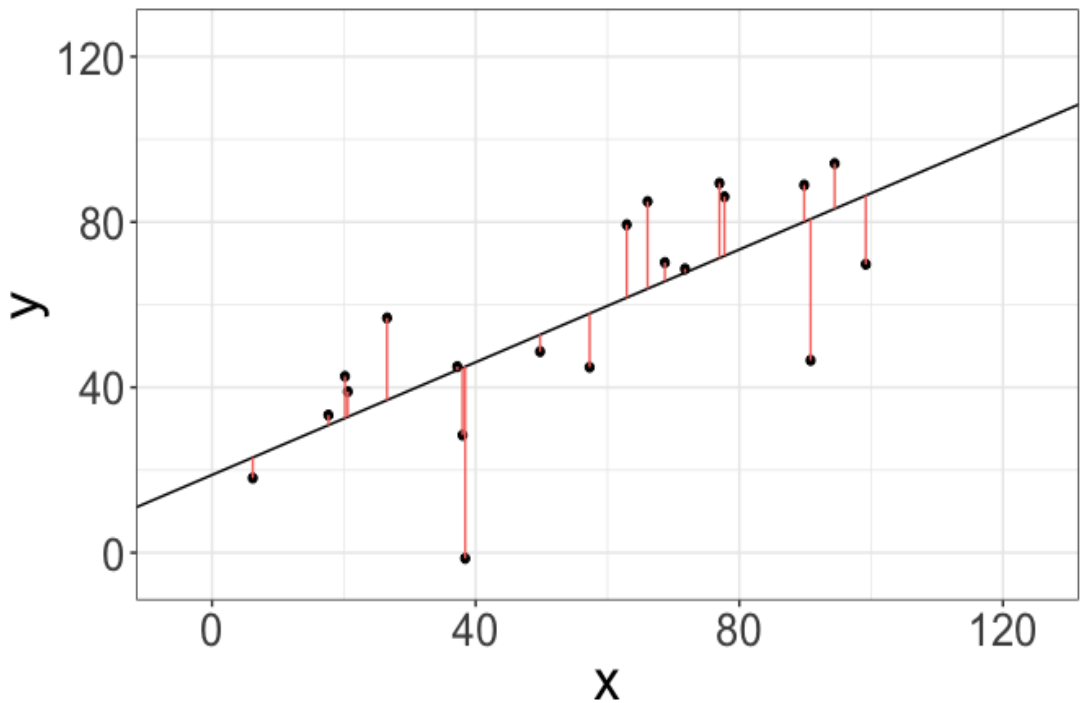
In some applications, the elements of observed data X can be viewed as fixed (e.g. in experimental settings where all "data" are perfectly controlled), but in most cases X itself will follow some distribution in the underlying population.

If so, we view the entire analysis as "conditional on X", which means we essentially "freeze" X at the observed values. A different X might produce different results, but we hope that these differences would vanish with (hypothetical) collection of more and more data.

What's important is Assumption 3: Whichever mechanism generates X is unrelated to the mechanism that generates the error terms.

Estimation

Figure 3: Ordinary least squares



Estimation of the regression model (1) via ordinary least squares (OLS) requires minimization of the sum of squared errors with respect to β.

$$S = \epsilon'\epsilon = (y - X\beta)'(y - X\beta).$$

First order condition $\frac{\partial S}{\partial \beta} = 0$, yields the following estimator,

$$\hat{\beta} = (X'X)^{-1}(X'y). \tag{2}$$

Second order condition requires that $\frac{\partial^2 S}{\partial \beta \partial \beta'} > 0$.

We can express the OLS estimator $\hat{\beta}$ as follows:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}(X'y) \\ &= (X'X)^{-1}(X'[X\beta + \epsilon] \\ &= \beta + (X'X)^{-1}(X'\epsilon) \end{aligned} \tag{3}$$

We can also express the residuals as follows:

$$\begin{aligned} \hat{\epsilon} &= y - \hat{y} = y - X\hat{\beta} \\ &= y - X(X'X)^{-1}(X'y) \\ &= [I - X(X'X)^{-1}X']y \\ &= My, \end{aligned} \tag{4}$$

where $M = [I_n - X(X'X)^{-1}X']$ is an idempotent matrix (i.e, M'M = M).

Estimation

Based on equation (3), we can clearly see $E(\hat{\beta}) = \beta$, and we derive the covariance as below:

$$\begin{aligned} Cov(\hat{\beta}) &= E[\{\hat{\beta} - E(\beta)\}\{\hat{\beta} - E(\beta)\}'] \\ &= E[(X'X)^{-1}(X'\epsilon)][(X'X)^{-1}(X'\epsilon)]' \\ &= (X'X)^{-1}\sigma^2. \end{aligned}$$

To derive an estimator of σ^2, we note that

$$E(\hat{\epsilon}'\hat{\epsilon}) = E[\mathtt{tr}(\epsilon' M \epsilon)] = E[\mathtt{tr}(M\epsilon\epsilon')],$$

where the last equality hold because both `trace` and `expectation` are linear operators.

Estimation

$$
\begin{aligned}
E(\hat{\epsilon}'\hat{\epsilon}) &= \mathtt{tr}\{ME[\epsilon\epsilon']\} = \mathtt{tr}[M(\sigma^2 I_n)] = \sigma^2 \mathtt{tr}(M) \\
&= \sigma^2 \,\mathtt{tr}[I_n - X(X'X)^{-1}X'] \\
&= \sigma^2 \,[\mathtt{tr}(I_n) - \mathtt{tr}(X(X'X)^{-1}X')] \\
&= \sigma^2 \,[\mathtt{tr}(I_n) - \mathtt{tr}(X'X(X'X)^{-1})] \\
&= \sigma^2 \,[\mathtt{tr}(I_n) - \mathtt{tr}(I_k)] \\
&= \sigma^2 (n-k).
\end{aligned}
$$

Consequently, $E[\frac{\hat{\epsilon}'\hat{\epsilon}}{n-k}] = \sigma^2$ and hence $\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-k}$ is an unbiased estimator of σ^2.

Estimation

With the OLS estimator available, the resulting estimate for $E[y_i|x_i]$ is $\hat{y}_i$, i.e.

$$E\left[y_i|x_i\right] = \hat{y}_i = x_i'\hat{\beta}.$$

The term $\hat{y}_i$ is called the "fitted value". The difference between observed y_i and estimated (or "predicted") $\hat{y}_i$ is called "residual" i.e.

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - x_i'\hat{\beta}.$$

In passing, we note that this implies the following inequality,

$$x_i'\beta + \epsilon_i = x_i'\hat{\beta} + \hat{\epsilon}_i, \qquad \text{and} \qquad X\beta + \epsilon = X\hat{\beta} + \hat{\epsilon}.$$

Goodness of Fit

Recall, the aim of regression analysis is to explain the variation in y via observed variation in X. So, it is natural to calculate the goodness-of-fit measures to quantify the explained variation.
1) The *Coefficient of Determination* or R^2 is defined as,

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\hat{\beta}' X' M_0 X \hat{\beta}}{y' M_0 y} = 1 - \frac{\hat{\epsilon}' \hat{\epsilon}}{y' M_0 y},$$

where $\hat{\epsilon} = y - \hat{y}$ and $M_0 = I - \iota(\iota'\iota)^{-1}\iota$. Here I is the $n \times n$ identity matrix and ι is an $n \times 1$ vector of ones. $R^2 \in [0, 1]$. A value of 0 implies that X has nothing to say about y, while a value of 1 implies a perfect fit.

Goodness of Fit

2) The R^2 measure doesn't penalize for superfluous regressors, or, alternatively, doesn't reward for parsimony. We therefore prefer "Adjusted R^2 defined as,

$$\bar{R}^2 = 1 - \frac{\hat{\epsilon}'\hat{\epsilon}/(n-k)}{y'M_0y/(n-1)}.$$

As the number of regressors (including intercept) k increases relative to sample size n, the last term increases and the adjusted fit deteriorates.

Goodness of Fit

If we wish to use $\bar{R}^2$ to choose between two models, the following must hold:

a) The dependent variable y must be same for both models. So, you can't compare a model that is linear in y to one that uses, say, $\log y$. Also, sample size must be identical.

b) The models have to be linear in parameters. For nonlinear regression models use AIC, BIC, or related measures.

c) Both models must have a constant term, and the mean of the error term in the population model must be zero.

Goodness of Fit

3) If the model does not include an intercept term, then the interpretation of R^2 and adjusted-R^2 becomes ambiguous.

In such cases, we use *Akaike Information Criterion* (AIC) or *Bayesian Information Criterion* (BIC). Both these measures work for linear models without constant terms and non-linear regression models. They also reward parsimony.

$$\begin{aligned} \ln AIC &= \ln\left(\frac{\hat{\epsilon}'\hat{\epsilon}}{n}\right) + \frac{2k}{n} \\ \ln BIC &= \ln\left(\frac{\hat{\epsilon}'\hat{\epsilon}}{n}\right) + \frac{k\ln(n)}{n}. \end{aligned}$$

Both AIC and BIC decline as model fit improves.

Data

Let's consider a salary dataset where we regress the dependent variable (salary) on the independent variable (years of experience).

	Experience	Salary
1	1.1	39343
2	1.3	46205
3	1.5	37731
4	2.0	43525
5	2.2	39891
6	2.9	56642
7	3.0	60150
8	3.2	54445
9	3.2	64445
10	3.7	57189

Scatter Plot

```
# Create the scatter plot

plot(data$Years_Exp, data$Salary,
xlab = "Years Experienced",
ylab = "Salary",
main = "Scatter Plot of Years Experienced vs Salary")
```

Scatter Plot

Regression

```
modelSLR <- lm(formula = Salary ~ Years_Exp,data = data)
summary(modelSLR)

# Output
Call:lm(formula = Salary ~ Years_Exp, data = data)
Residuals:
Min       1Q       Median   3Q      Max
-8171.3  -3695.9   -717.2   4219.7  7362.1

Coefficients:  Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)    28217      5130        5.501     0.000573 ***
Years_Exp      9021       2003        4.503     0.001995 **

---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regression

```
# Output
Residual standard error: 5482 on 8 degrees of freedom
Multiple R-squared:  0.7171, Adjusted R-squared:  0.6817
F-statistic: 20.28 on 1 and 8 DF,  p-value: 0.001995

fitted(modelSLR)

#Output
1         2         3         4         5
38139.57  39943.69  41747.82  46258.14  48062.27

6         7         8         9         10
54376.71  55278.78  57082.90  57082.90  61593.22
```

Regression

```
# plot a scatter plot
plot(data$Years_Exp,data$Salary,
main='Regression of Experience and Salary',
xlab='Years of Experience',ylab='Salary')

# plot a regression line
abline(lm(Salary~Years_Exp,data=data),col='red')
```

Regression


Regression of Experience and Salary
Salary
40000
45000
50000
55000
60000
65000
1.0
1.5
2.0
2.5
3.0
3.5
Years of Experience

Maximum Likelihood Estimation

If we feel safe in making assumptions on the statistical distribution of the error term, maximum likelihood estimation (MLE) is an attractive alternative to least squares for linear regression models. Even better, MLE can be used for non-linear models. It is thus a more generally applicable estimation strategy than OLS.

Sticking with the classical linear regression model, we have

$$y_i = x_i'\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$
$$y = X\beta + \epsilon, \quad \epsilon \sim N(0_n, \sigma^2 I).$$

We continue to assume that the errors are independent and identically distributed as $N(0, \sigma^2)$.

Maximum Likelihood Estimation

Let $\theta = (\beta, \sigma^2)$, then the density for the entire sample ("the probability of observing this very sample") as a product of individual densities is,

$$\begin{aligned} L(\theta; y) &\equiv f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \right] \\ &= \left[2\pi\sigma^2\right]^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \right]. \end{aligned}$$

Maximum Likelihood Estimation

Note that the likelihood function $L(\theta; y)$ is equivalent to the joint density $f(y|\theta)$, but they are not equal. Although both quantities have the same expression, the interpretations are different.

The likelihood function is a function of the parameter given data. The maximum likelihood method aims to maximize the sample likelihood and attempts to find those parameter values that is most likely to have generated the observed sample.

In contrast, the joint density is a function of the data given the parameter.

Maximum Likelihood Estimation

We take logarithm of the likelihood and maximize,

$$\log L(\theta; y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2,$$

with respect to (β, σ^2). This is done by taking the first partial derivatives and equating them to zero.

Maximum Likelihood Estimation

For the normal regression model, the first order conditions (also called likelihood equations) are:

$$\begin{bmatrix} \frac{\partial \ln L(\theta)}{\partial \beta} \\ \\ \frac{\partial \ln L(\theta)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{X'(y - X\beta)}{\sigma^2} \\ \\ -\frac{n}{2\sigma^2} + \frac{(y-X\beta)'(y-X\beta)}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}.$$

They lead to the following maximum likelihood estimators:

$$\tilde{\beta} = (X'X)^{-1}(X'y) \qquad \text{and} \qquad \tilde{\sigma}^2 = \frac{(y - X\tilde{\beta})'(y - X\tilde{\beta})}{n}. \quad (5)$$

Maximum Likelihood Estimation

Clearly, the MLE solution $\tilde{\beta} = (X'X)^{-1}(X'y)$ is identical to the OLS solution $\hat{\beta} = (X'X)^{-1}(X'y)$.

The ML solution $\tilde{\sigma}^2$ differs slightly from the OLS estimator $\hat{\sigma}^2$ as it does not correct the denominator for degrees of freedom (k).

Note that the OLS estimator $\hat{\sigma}^2$ is unbiased, but the ML estimator $\tilde{\sigma}^2$ is biased but consistent i.e., the bias goes to zero as $n \to \infty$.

Maximum Likelihood Estimation

To assure a maximum, we need to examine the properties of the Hessian matrix of second derivatives. For the normal regression model:

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 \ln L(\theta)}{\partial \beta \, \partial \beta'} & \frac{\partial^2 \ln L(\theta)}{\partial \beta \, \partial \sigma^2} \\ \frac{\partial^2 \ln L(\theta)}{\partial \sigma^2 \, \partial \beta'} & \frac{\partial^2 \ln L(\theta)}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{X'X}{\sigma^2} & -\frac{X'\epsilon}{\sigma^4} \\ -\frac{\epsilon'X}{\sigma^4} & -\frac{2\epsilon'\epsilon - \sigma^2 n}{2\sigma^6} \end{bmatrix}, \quad (6)$$

where $\epsilon = y - X\beta$. Maximization requires the principal minor of the Hessian matrix to alternate in sign starting with negative. Alternatively, once can show that the eigenvalues for $H(\theta)$ is negative definite, and we have indeed a maximum.

Generalized Linear Regression Model

The Generalized Linear Regression Model (GLRM) differs from the CLRM in the structure of the variance-covariance matrix of the error vector ϵ. We no longer have spherical disturbances (independent errors that all share the same variance), but error that may be potentially correlated and/or follow distributions with different variances.

The GLRM in its generic form can be written as:

$$y = X\beta + \epsilon \qquad E(\epsilon) = 0 \qquad E(\epsilon\epsilon') = \Omega \neq \sigma^2 I_n. \tag{7}$$

The error vector still has a mean of zero, but its variance-covariance matrix now takes a general form with theoretically up to $n(n+1)/2$ unknown parameters.

Generalized Linear Regression Model

With only n observations, estimation of that many additional parameters is infeasible. So, we usually follow one of two strategies:

(a) Claim complete ignorance about the structure of Ω and use robust estimation methods to still derive consistent estimates of coefficients.

(b) Assume that Ω has a simple structure with only a few additional parameters (example: clusters of errors have separate variances, but not each individual error term).

Robust Estimation

The general intuition for robust estimation is to use an expression for Ω that does not require knowledge or estimation of additional parameters, but still allows for consistent estimation of $V(\hat{\beta})$.

There exists a variety of robust estimators for GLRM's. Their exact form depends on the nature of generalization and hence there is no "one fits all" robust estimator. Example: for *heteroskedasticity*, White (1980) has shown that under general conditions the term

$$\widehat{V_a(\hat{\beta})} = (X'X)^{-1}(X'EX)(X'X)^{-1},$$

is a consistent estimator of $V(\hat{\beta})$, where

Robust Estimation

$$E = \begin{bmatrix} \hat{\epsilon}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\epsilon}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\epsilon}_n^2 \end{bmatrix}.$$

The estimator $\widehat{V_a(\hat{\beta})}$, which is based on the OLS residuals $\hat{\epsilon}_i$'s, corrects the shortcomings of the naive OLS approach. This works well in a large sample context, but the properties of this estimator under small sample sizes are still disputed.

Robust Estimation

The analogous version of this *robust* or "*White*" estimator (sometimes also called "sandwich estimator" or Huber-White sandwich estimator) for MLE is,

$$\widehat{V_a(\hat{\beta})}_{mle} = (-\hat{H})^{-1}\hat{G}'\hat{G}(-\hat{H})^{-1},$$

where $\hat{H}$ is the Hessian matrix at the MLE solution, and $\hat{G}$ is the $n \times k$ matrix of individual gradients at the MLE solution.

The White estimator can be used in any context where it is suspected that the spherical distribution assumption of the CLRM is violated. It has been especially popular to control for heterskedastic errors.

GLS Estimator

Let's assume for a moment that Ω is known, then we can derive the Generalized Least Squares (GLS) estimator by a simple extension of the CLRM estimation framework.

Let $\Omega = \sigma^2\Psi$, this is trivial–you can always factor out any scalar from any matrix or vector. Then factor Ψ^{-1} into $P'P$ where P itself is an $n \times n$ symmetric square matrix. Then pre-multiply the CLRM by P – this will reinstate the spherical properties of the disturbances – and use OLS on the transformed model.

GLS Estimator

Formally, the model can be written as follows:

$$\begin{aligned} &y^* = X^*\beta + \epsilon^*, \quad \text{where} \quad y^* = Py, X^* = PX, \epsilon^* = P\epsilon, \\ &E(\epsilon\epsilon') = \Omega = \sigma^2\Psi, \qquad\qquad P'P = \Psi^{-1} \\ &E(\epsilon^*) = PE(\epsilon) = 0, \\ &E(\epsilon^*\epsilon^{*'}) = PE(\epsilon\epsilon')P = \sigma^2 P\Psi P = \sigma^2 PP^{-1}P^{-1}P = \sigma^2 I. \end{aligned} \tag{8}$$

Then the GLS estimator is derived as,

$$\begin{aligned} \hat{\beta}_{GLS} &= (X^{*'}X^*)^{-1}(X^{*'}y^*) = (X'P'PX)^{-1}(X'P'Py) \\ &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y). \end{aligned} \tag{9}$$

GLS Estimator

The GLS estimator has the following properties:

$$\begin{aligned} E(\hat{\beta}_{GLS}) &= E\left[(X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}X)\beta\right] + E\left[(X'\Omega^{-1}X)^{-1}X'\epsilon\right] = \beta \\ V(\hat{\beta}_{GLS}) &= E\left[(X'\Omega^{-1}X)^{-1}X'\epsilon\epsilon'X(X'\Omega^{-1}X)^{-1}\right] = (X'\Omega^{-1}X)^{-1} \\ \hat{\beta}_{GLS} &\overset{a}{\sim} N(\beta, (X'\Omega^{-1}X)^{-1}). \end{aligned} \tag{10}$$

Thus, $\hat{\beta}_{GLS}$ is unbiased and consistent. It is also asymptotically efficient. It is also BLUE for the generalized regression model. Specifically, the correct variance of the OLS estimator $V(\hat{\beta}) = (X'X)^{-1}(X'\Omega^{-1}X)^{-1}(X'X)^{-1}$ will be less eficient than $\hat{\beta}_{GLS}$.

FGLS Estimator

The full content of Ω is rarely known. In practice, we have a general idea about the structure of Ω and assume that it is a function of just a few additional parameters i.e., $\Omega = \Omega(\theta)$. Examples:

1) In time series applications, we specify the Ω matrix as a function of a single additional parameter (say ρ), shown later.

2) Group-wise heteroskedasticity, where a limited number of blocks (or "cluster") of errors share different variances.

3) Alternatively, the elements of Ω may be assumed to be a combination of observed data and a few unknown parameters. For example,

FGLS Estimator

under heteroskedasticity we often assume that the variance of observation i's error term is itself a function of attributes associated with that observation i.e., $\sigma_i^2 = f(z_i, \gamma)$ and

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}.$$

FGLS Estimator

Whatever the assumed structure of Ω, FGLS proceeds in the following 2 steps:

1) Derive a consistent estimator of Ω i.e., $\hat{\Omega} = \Omega(\hat{\theta})$.

2) Use this estimator in the expression for $\hat{\beta}_{GLS}$.

Formally, the FGLS estimator is given by the expression:

$$\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}y), \qquad \hat{\Omega} = \Omega(\hat{\theta}). \tag{11}$$

For most "standard settings" the resulting $\hat{\beta}_{FGLS}$ will have the same asymptotic properties as $\hat{\beta}_{GLS}$.

FGLS Estimator

Alternatively, we can use the *Maximum Likelihood* approach to simultaneously estimate θ and β. Such a "full-information maximum likelihood" (FIML) approach is generally more efficient in a small sample context.

However, if our assumptions regarding the structure of Ω are incorrect, the same problems as presented for the OLS estimator arise. Some researchers therefore prefer not to make any assumption on Ω, but instead use *robust estimation* to derive consistent results.

Non-spherical Disturbances

In general, non-spherical or non-scalar identity covariance matrix may arise due to one of the following.

1) Heteroskedasticity

2) Auto-correlation

3) Equation error related sets of regression equations.

We cover estimation of the first two case and skip the third case.

Heteroskedasticty – Estimation when Ψ is unknown

Model: $y = X\beta + \epsilon$, $E(\epsilon) = 0$, $E(\epsilon\epsilon') = \Omega = \sigma^2\Psi$. The matrix Ψ is typically unknown and so we use the FGLS estimator,

$$\hat{\beta}_{FGLS} = (X'\hat{\Psi}^{-1}X)^{-1}(X'\hat{\Psi}^{-1}Y).$$

The large number of unknown parameters ($\frac{n(n+1)}{2}$) cannot be estimated with n observations. A structure on Ψ is necessary.

$$\Psi = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

Heteroskedasticty – Estimation when Ψ is unknown

- Using least squares find the estimator $\hat{\beta}$ and $\hat{\epsilon} = Y - X\hat{\beta}$.
- Construct the matrix

$$\hat{\Psi} = \text{diag}(\ \hat{\epsilon}\,\hat{\epsilon}'\ ) = \begin{pmatrix} \hat{\epsilon}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\epsilon}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \hat{\epsilon}_n^2 \end{pmatrix}$$

- Calculate $\hat{\beta}_{FGLS} = (X'\hat{\Psi}^{-1}X)^{-1}(X'\hat{\Psi}^{-1}Y)$.

Autocorrelation

$$\begin{aligned} y_t &= x_t'\beta + \epsilon_t, \\ \epsilon_t &= \rho\epsilon_{t-1} + \nu_t, \qquad t = 1(1)T \end{aligned}$$

where, $E(\nu_t) = 0$, $E(\nu_t^2) = \sigma_\nu^2$ and $E(\nu_t\nu_s) = 0$ for $t \neq s$.

$$\Omega = \frac{\sigma_\nu^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \dots & \dots & 1 \end{pmatrix}$$

Earlier notation: $\sigma_\nu^2 = \sigma^2$ and Ψ is remaining expression.

Autocorrelation – Estimation when ρ is unknown

- Using least squares find the estimator $\hat{\beta}$ and $\hat{\epsilon} = Y - X\hat{\beta}$.
- Find the least squares estimator to $\hat{\epsilon}_t = \rho\hat{\epsilon}_{t-1} + \nu_t$, which yields

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum_{t=2}^{T} \hat{\epsilon}_{t-1}^2}.$$

- Use $\hat{\rho}$ in place of ρ to calculate

$$\hat{\beta}_{FGLS} = (X'\hat{\Psi}^{-1}X)^{-1}(X'\hat{\Psi}^{-1}Y).$$

Thank you!