DASC7011 Statistical Inference for Data Science

Chapter 3 Least Squares Estimation

Department of Statistics and Actuarial Science
Department of Computer Science
The University of Hong Kong

Contents

§3.1 Introduction

- Least squares estimations are frequently used in neural network models.

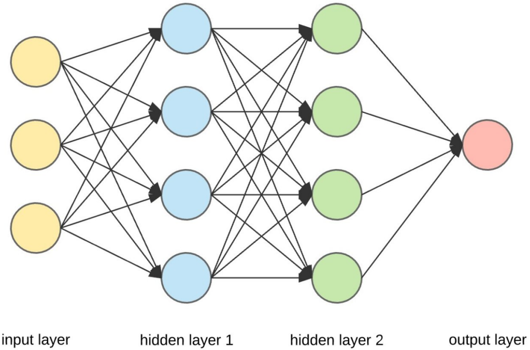

Figure 3.1: A neural network model.

Estimate the neural network model

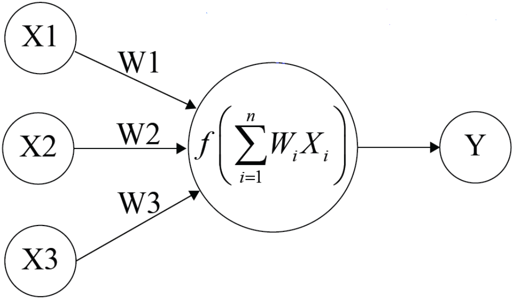

Figure 3.2: Estimating/Fitting the neural network model

- $f(\cdot)$ is called the *activation function*, usually chosen from several candidates based on the nature of Y.
- $\sum_{i=1}^{n} w_i X_i$ is a linear combination of X_i's with weights w_i's to be estimated.

Ideas of the estimation

- Expected (replace n with k):

$$\mathbb{E}\big[f^{-1}(Y)|\{X_i\}\big] = w_0 + w_1 X_1 + \cdots + w_k X_k.$$

- Observed:

$$f^{-1}(Y_j) = w_0 + w_1 X_{1j} + \cdots + w_k X_{kj} + e_j, \qquad 1 \le j \le n,$$

  where e_j's are errors or deviations (from expectations).

- The smaller the errors, the better the model.
- Rule(s) are needed for fitting, prediction, and/or interpretation.

Ideas of the estimation

- Errors/Deviations:

$$e_j = f^{-1}(Y_j) - w_0 - w_1 X_{1j} - \cdots - w_k X_{kj}, \qquad 1 \leq j \leq n.$$

- Possible rules:
  (1) Minimize the sum of absolute errors $\sum_{i=1}^{n} |e_i|$.
  (2) Minimize the sum of squared errors $\sum_{i=1}^{n} e_i^2$.
  (3) Minimize the sum of 4th order errors $\sum_{i=1}^{n} e_i^4$.
  (4) ……

- Among all, rule (2) is the most frequently used.

- The rule "minimizing the sum of squared errors" is usually referred to as **Least-Squares** (**LS**), which was initially developed to estimate *regression models*.

Least Squares Estimation (LSE)

- A (general) regression model:

$$Y_i = f(X_{1i}, X_{2i}, \cdots, X_{pi}; \boldsymbol{\theta}) + \varepsilon_i, \qquad 1 \le i \le n, \tag{3.1}$$

  where

  - Y is the **response variable**,
  - $\{X_{k\cdot} : 1 \le k \le p\}$ are **independent variables** or **regressors**,
  - $f(\cdot)$ is a pre-specified function, or one from a known class of functions,
  - $\boldsymbol{\theta}$ is (are) the parameter(s),
  - $\{\varepsilon_i : 1 \le i \le n\}$ are random errors, usually assumed to be i.i.d. and follow certain distribution F *with mean 0*.

Least Squares Estimation (LSE)

- Suppose $f(\cdot)$ is fixed, and there is no unknown features or structures apart from parameter(s) $\boldsymbol{\theta}$.
- The LS algorithm estimates $\boldsymbol{\theta}$ with the value(s) that minimize the following **Error Sum of Squares** (**ESS** or **SSE**) function,

$$S(\boldsymbol{\theta}; f) = \sum_{i=1}^{n} \left[Y_i - f(X_{1i}, X_{2i}, \cdots, X_{pi}; \boldsymbol{\theta})\right]^2. \tag{3.2}$$

  We include f in the notation to indicate it depends on f too.
- In other words, the **least squares estimator** (**LSE**) of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}_f = \arg\min_{\boldsymbol{\theta}} \Big\{ S(\boldsymbol{\theta}; f) \Big\}. \tag{3.3}$$

- If f is allowed to be selected from a class $\mathcal{C}$, it could be selected/tuned as follows.

$$\widehat{f} = \arg\min \Big\{ S(\widehat{\boldsymbol{\theta}}_f; f) : f \in \mathcal{C} \Big\}. \tag{3.4}$$

§3.2 Ordinary Least Squares Estimation

3.2.1. Simple linear regression models.

- Let $(\boldsymbol{X}, \boldsymbol{Y}) = \{(X_i, Y_i) : 1 \le i \le n\}$ be a random sample from a **simple linear regression model**:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad 1 \le i \le n, \tag{3.5}$$

  where $\{\varepsilon_i\} \overset{i.i.d.}{\sim} F(0, \sigma^2)$. Parameters are $\boldsymbol{\theta} = (\beta_0, \beta_1)'$ and σ^2.

- The *ordinary least squares* (OLS) algorithm estimates $\boldsymbol{\theta}$ by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \Big\{ S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \Big\}. \tag{3.6}$$

Ordinary Least Squares Estimation (OLSE)

- Solving (3.6),

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0, \\ \dfrac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0, \end{cases}$$

gives

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}. \tag{3.7}$$

- The OLS estimate of σ^2 is (chosen to be)

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2. \tag{3.8}$$

Example 3.1: simple LM

- **Example 3.1**: (*Simple linear regression model*) Read the data set saved in file `Stories.txt` into the R program by

```
mydata <- read.table('C:/Teaching/.../Stories.txt',
                     header = TRUE)
```

- Suppose we want to regress the heights of buildings (variable `HGHT`) against the number of stories (variable `STORIES`) using a simple linear regression model $\texttt{HGHT} = \beta_0 + \beta_1 \texttt{STORIES} + \varepsilon$. We may do some preparing work to simplify coding:

```
# Preparing work
View(mydata)
Y <- mydata[,2]     # variable HGHT
X <- mydata[,3]     # variable STORIES
n <- nrow(mydata)   # sample size
```

Example 3.1: simple LM

- Find the OLS estimates using formulae (3.7) and (3.8):

```
> # Find OLS estimates by formulae
> y <- Y-mean(Y)    # deviation form of Y
> x <- X-mean(X)    # deviation form of X
> beta1 <- sum(y*x)/sum(x^2)
> beta0 <- mean(Y)-beta1*mean(X)
> sigma <- sqrt(sum((Y-beta0-beta1*X)^2)/(n-2))
> beta0; beta1; sigma
[1] 90.3096
[1] 11.29237
[1] 58.32593
```

- We can alternatively fit the model using the built-in function/algorithm `lm()`:

```
# Fit the model by lm()
fit <- lm(Y~X)    # Fit the model
summary(fit)      # Summary of the fitted model
```

Example 3.1: simple LM

- The summary of the fitted model:

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-156.759  -33.239    5.995   28.450  167.487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  90.3096    20.9622   4.308 6.44e-05 ***
X            11.2924     0.4844  23.310  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.33 on 58 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.9019
F-statistic: 543.4 on 1 and 58 DF,  p-value: < 2.2e-16
```

3.2.2. Multiple linear regression models.

- If there are more than one regressors, then we have a **multiple linear regression model**.

- A multiple linear regression model can be defined as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon, \tag{3.9}$$

where $\beta_0, \beta_1, \cdots, \beta_k$ are unknown parameters, Y is the dependent variable, and $X_1, X_2, \cdots, X_k$ are k independent variables/predictors/regressors.

- For observations $(Y_i, X_{1i}, X_{2i}, \cdots, X_{ki})_{i=1,2,\cdots,n}$, the model becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \cdots + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, \cdots, n. \tag{3.10}$$

Multiple linear regression models

- The multiple linear regression model (3.10) can be re-written into the following matrix form:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

For convenience, we write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.11}$$

- We call matrix $\mathbf{X}$ the **design matrix**.
- Generally, we may assume that $k \ll n$, or at least, $k < n$.

Basic assumptions for multiple LM

(A1) The relationship between $\mathbf{Y}$ and $\mathbf{X}$ is *linear* and is given by Eq. (3.11).

(A2) The design matrix $\mathbf{X}$ is *non-stochastic*. In addition, no exact linear relationship exists between the independent variables, and hence, $\mathbf{X}$ is *fully ranked* ($\text{Rank}(\boldsymbol{X}) = k + 1 \leq n$).

(A3) The error term has *zero mean* for all observations.

(A4) The error term has *constant variance* σ^2 for all observations.

(A5) Errors corresponding to different observations are mutually *independent* and therefore uncorrelated.

(A6) The error term is *normally distributed*.

OLSE for multiple LM

- The **sum of squared errors/residuals (SSE)** function is defined as

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_k X_{ki})^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{Y}. \end{aligned}$$

- Differentiate is with respective to $\boldsymbol{\beta}$,

$$S'(\boldsymbol{\beta}) = 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{Y}.$$

- Equating it to zero, we obtain the OLSE of $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.12}$$

Properties of OLSE

- Under assumptions (A1) through (A5), $\widehat{\boldsymbol{\beta}}$ is **unbiased** since

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

- Under assumptions (A1) through (A5),

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

- Under assumptions (A1) through (A6), $\widehat{\boldsymbol{\beta}}$ is normal,

$$\widehat{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \tag{3.13}$$

- Under assumptions (A1) through (A5), $\widehat{\boldsymbol{\beta}}$ is *asymptotically normal*.

Estimate of error variance

- The population variance, or the error variance, σ^2, is estimated by the **mean square error**:

$$\widehat{\sigma}^2 = s^2 = \frac{\sum e_i^2}{n-k-1}, \tag{3.14}$$

where $e_i = Y_i - \widehat{Y}_i$ is the ith regression residual.

- The mean square error s^2 is an *unbiased* estimator of the error variance σ^2, which has $(n-k-1)$ degrees of freedom.
- The non-negative square-root of the residual variance s^2, denoted by s, and sometimes SER, is called the **standard error of the regression**.

Gauss-Markov Theorem

Gauss-Markov Theorem: *Under assumptions (A1) through (A5), the OLS estimators $\widehat{\boldsymbol{\beta}}$ defined by Eq. (3.12) are the* **best linear unbiased estimators (BLUE)** *of $\boldsymbol{\beta}$.*

§3.3 Statistical inferences with OLSE

3.3.1. Inferences on a single parameter.

- Denote the OLSE of a single parameter β (we drop the index i for simplicity) by $\widehat{\beta}$, and the estimated standard error (e.s.e) by s_β.
- By Property (3.13), the credible interval of β at the confidence level $(1-\alpha)$ is

$$\big(\widehat{\beta} - s_\beta \cdot t_{\alpha/2}(n-k-1),\ \widehat{\beta} + s_\beta \cdot t_{\alpha/2}(n-k-1)\big), \tag{3.15}$$

  where $t_{\alpha/2}(n-k-1)$ is the $\frac{\alpha}{2}$-th upper quantile of the $t(n-k-1)$ distribution.
- The hypothesis $H_0 : \beta = b$ can be tested using the following T-test (statistic),

$$T = \frac{\widehat{\beta} - b}{s_\beta} \sim t(n-k-1). \tag{3.16}$$

Coefficient of determination

3.3.2. Inferences on multiple parameter.

- Inferences involving in multiple parameters needs more properties of the LSE.
- Define sums of squares (squared deviations):

$$\begin{aligned}
\mathrm{SST} &= \sum (Y_i - \overline{Y})^2 = \mathbf{Y}'\mathbf{Y} - n\overline{Y}^2, \\
\mathrm{SSR} &= \sum (\widehat{Y}_i - \overline{Y})^2 = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} - n\overline{Y}^2, \\
\mathrm{SSE} &= \sum (Y_i - \widehat{Y}_i)^2 = \sum e_i^2 = \boldsymbol{e}'\boldsymbol{e}.
\end{aligned}$$

- It is not difficult to see that

$$\begin{aligned}
\mathrm{SST} &= \sum (Y_i - \overline{Y})^2 = \sum (Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})^2 \\
&= \sum (\widehat{Y}_i - \overline{Y})^2 + \sum e_i^2 + 2\sum e_i(\widehat{Y}_i - \overline{Y}) \\
&= \mathrm{SSR} + \mathrm{SSE}.
\end{aligned}$$

Coefficient of determination

- The last equation holds because
  - The residual vector $\boldsymbol{e}$ is orthogonal to each column of $\mathbf{X}$, including the constant column/vector $\mathbf{1} = (1, \cdots, 1)'$, since

$$\begin{aligned}\mathbf{X}'\boldsymbol{e} &= \mathbf{X}'(\mathbf{Y} - \widehat{\mathbf{Y}}) \\ &= \mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} \\ &= [\mathbf{X}' - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} \\ &= (\mathbf{X}' - \mathbf{X}')\mathbf{Y} = \mathbf{0}.\end{aligned}$$

  - Denote $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\widehat{\mathbf{Y}} = \mathbf{RY}$. The residual vector $\boldsymbol{e}$ is orthogonal to the fitted response vector $\widehat{\mathbf{Y}}$ since $\mathbf{R}' = \mathbf{R}$, $\mathbf{R} = \mathbf{R}^2$, and

$$\widehat{\mathbf{Y}}'\boldsymbol{e} = \mathbf{Y}'\mathbf{R}'(\mathbf{I} - \mathbf{R})\mathbf{Y} = \mathbf{Y}'\mathbf{R}\mathbf{Y} - \mathbf{Y}'\mathbf{R}^2\mathbf{Y} = \mathbf{0}.$$

Coefficient of determination

- The sums of squares are measures of *total variation* of Y, variations *explained* by the model, and variations *not explained* by the model (of residuals), respectively.

- The **coefficient of determination** is defined as the ratio of the sum of squares due to regression to the total sum of squares:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}. \tag{3.17}$$

- By definition, R^2 is a measure (in percentage) of how much the total variation of Y is accounted for, or explained, by the model. Or equivalently, how well the independent variables can explain Y in terms of its variation.

- Roughly speaking, the larger is R^2, the better is the fitted model.

Adjusted coefficient of determination

- The **adjusted R-square**, denoted as R_a^2, considers the ratio of means of squares:

$$R_a^2 = 1 - \frac{S_e^2}{S_Y^2} = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)}. \tag{3.18}$$

- Notice that

$$\begin{aligned} 1 - R_a^2 &= \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = \frac{\text{SSE}}{\text{SST}} \cdot \frac{n-1}{n-k-1} \\ &> \frac{\text{SSE}}{\text{SST}} = 1 - R^2. \end{aligned}$$

  Therefore,

$$0 \le R_a^2 < R^2 \le 1.$$

- R^2 and R_a^2 are measures of the **goodness-of-fit (GOF)** of the model.

Inferences on multiple parameters

- Suppose that we want to test the following null hypothesis

$$H_0: \ \beta_k = \beta_{k-1} = \cdots = \beta_{k-p+1} = 0$$

against alternative

$$H_1: \ \text{not all of } \beta_k, \beta_{k-1}, \cdots, \beta_{k-p+1} \text{ are zeros,}$$

where $1 \leq p \leq k$ are two integers.

- Under the null hypothesis, the model becomes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-p} X_{k-p} + \varepsilon. \tag{3.19}$$

- We call model (3.19) the **restricted model**, in contrast to the **unrestricted model** or **full model** defined by Eq. (3.9).

Inferences on multiple parameters

- Here an F-test is applied. The test statistic is

$$F = \frac{(\mathrm{SSE}_R - \mathrm{SSE}_F)/p}{\mathrm{SSE}_F/(n-k-1)} \sim F(p, n-k-1) \quad \text{under } H_0, \quad (3.20)$$

  where SSE_R and SSE_F are the sum of squared residuals functions for the restricted model and the full model, respectively.

- For a given significance level $\alpha > 0$, the critical region is $F > F_\alpha(p, n-k-1)$, the α-th upper quantile of the $F(p, n-k-1)$ distribution.

- This F test applies for testing hypothesis involving a single parameter too, and is equivalent to the t test.

A goodness-of-fit test

- Another special example of the F test is the case that $p = k$. In this case, the restricted model becomes a *null model*

$$Y = \beta_0 + \varepsilon.$$

- In other words, the null hypothesis suggests that all predictors in model (3.9) are superfluous.
- If the null hypothesis is rejected, then we say that the regression model (3.9) is *significant* (significantly better than the null model). Otherwise, the model is *insignificant* or *trivial*.
- Such an F test is usually referred to as the *goodness-of-fit test*.

3.3.3. Predictions.

- Consider the multiple regression model (3.9) with OLSE (and BLUE) $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k)'$.
- Given a special observation of predictors $(X_{10}, X_{20}, \cdots, X_{k0})$, the corresponding $\mathbb{E}(Y_0)$ is predicted by

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_{10} + \cdots + \widehat{\beta}_k X_{k0}. \tag{3.21}$$

- If we denote the new observation as a (column) vector $\mathbf{X}_0 = (1, X_{10}, \cdots, X_{k0})'$, then

$$\widehat{Y}_0 = \mathbf{X}_0' \widehat{\boldsymbol{\beta}}.$$

- The prediction $\widehat{Y}_0$ is actually an estimate of the *conditional expected value* of Y given $(X_1, \cdots, X_k) = (X_{10}, \cdots, X_{k0})$, i.e., $\mathbb{E}(Y|\boldsymbol{X}_0)$. We use the notation $\mathbb{E}(Y_0)$ for simplicity.

Prediction or Forecast

- The prediction $\widehat{Y}_0$ is a linear combination of the BLUE $\widehat{\boldsymbol{\beta}}$, and thus a BLUE of $\mathbb{E}(Y_0)$.

- Moreover, $\widehat{Y}_0$ is normally distributed with variance

$$\begin{aligned}\sigma^2_{Y_0} &= \mathrm{Var}(\mathbf{X}_0'\widehat{\boldsymbol{\beta}}) = \mathrm{Var}[\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \\ &= \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \\ &= \sigma^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0.\end{aligned}$$

- The estimated value of this variance is

$$s^2_{Y_0} = s^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0,$$

where s^2 is the estimated residual variance.

Confidence interval

- Therefore, the **confidence interval** at the γ-level of confidence (e.g., $\gamma = 0.95$) for $\mathbb{E}(Y_0)$ is

$$CI_\gamma = \widehat{Y}_0 \pm t_{(1-\gamma)/2}(n-k-1)s_{Y_0}. \tag{3.22}$$

- Interpretation:

$$\mathbb{P}\Big(\mathbb{E}(Y) \in CI_\gamma \Big| \boldsymbol{X}, \boldsymbol{X}_0\Big) = \gamma.$$

Prediction interval

- The **prediction**/**forecast error** of $\widehat{Y}_0$ is

$$\begin{aligned} e_0 &= Y_0 - \widehat{Y}_0 \\ &= \mathbf{X}_0'\boldsymbol{\beta} + \varepsilon_0 - \mathbf{X}_0'\widehat{\boldsymbol{\beta}} \\ &= \mathbf{X}_0'(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \varepsilon_0. \end{aligned}$$

- The variance of the prediction error is

$$\begin{aligned} \sigma^2_{F_0} &= \mathrm{Var}[\mathbf{X}_0'(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})] + \mathrm{Var}(\varepsilon_0) \\ &= \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\sigma^2\mathbf{X}_0 + \sigma^2 \\ &= \sigma^2[1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0]. \end{aligned}$$

Prediction interval

- The estimated variance of the prediction error is

$$s_{F_0}^2 = s^2[1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0].$$

- Therefore, the **prediction interval** at the γ-level of confidence is

$$PI_\gamma = \widehat{Y}_0 \pm t_{(1-\gamma)/2}(n-k-1)s_{F_0}. \tag{3.23}$$

- Understanding:

$$\mathbb{P}\Big(Y \in PI_\gamma \Big| \boldsymbol{X}, \boldsymbol{X}_0\Big) = \gamma.$$

- Since $s_{F_0}^2 > s_{Y_0}^2$, the prediction interval is always *wider* than the corresponding (at the same confidence level) confidence interval.

Example 3.2: Crime Data

- **Example 3.2**: Analyze the Freedman data from R package car (Companion to Applied Regression).
- Load data: `library(car); data(Freedman); Freedman`.
- Observations from 110 U.S. metropolitan areas with 1968 populations of 250,000 or more, with some missing data.
- Four variables:
  - `population`: Total 1968 population, in thousands.
  - `nonwhite`: Percent nonwhite population, 1960.
  - `density`: Population per square mile, 1968.
  - `crime`: Crime rate per 100,000, 1969.

Example 3.2: Crime Data

- **Purpose**: investigate effects of other three variables on the crime rate.
- First model named as `fit0` in R:
  - R coding:

```
fit0 <- lm(crime ~ ., data = Freedman)
summary(fit0)
```

  - Summary of the fit, `summary(fit0)`:

Coefficients	Estimate	Std.Error	t-value	$Pr(>\lvert t\rvert)$
(Intercept)	2193.70088	143.04566	15.336	$< 2e{-16}$
population	0.24495	0.06095	4.019	0.000116
nonwhite	26.03770	8.76746	2.970	0.003764
density	-0.02145	0.06578	-0.326	0.745045

Example 3.2: Crime Data

- Effect of `density` is insignificant by the t-test. Remove it from the model.
  - R coding:

```
fit1 <- lm(crime ~ population + nonwhite,
           data = Freedman)
```

  - Summary:

(Intercept)	2.185e+03	1.398e+02	15.630	< 2e−16
population	2.376e−01	5.639e−02	4.214	5.63e−05
nonwhite	2.611e+01	8.724e+00	2.993	0.0035
Multiple R-squared: 0.2279 (R^2)				
Adjusted R-squared: 0.212 (R_a^2)				
F-statistic: 14.32 on 2 and 97 DF, p-value: 3.56e−06				

Example 3.2: Crime Data

- All coefficients are significant at the 5% level.
- The summary also provides the overall GOF test results: $F = 14.32$, null distribution $F(2, 97)$, p-value 3.56e−06.
- To find SST, SSR and SSE, we look at the ANOVA table produced by `avona(fit1)`:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
population	1	15000803	15000803	19.678	2.421e−05
nonwhite	1	6828890	6828890	8.958	0.003504
Residuals	97	73945387	762324		

- It is not difficult to check that

$$F = \frac{(15000803 + 6828890)/2}{73945387/97} = 14.31787.$$

- Conclusion: the model is significant.

Example 3.2: Crime Data

- To conduct the F-test for `fit1` against `fit0`, we need the ANOVA table of `fit0`.
  - ANOVA of `fit0`:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
population	1	15000803	15000803	19.678	2.421e−05
nonwhite	1	6828890	6828890	8.958	0.003504
density	1	81829	81829	0.1064	0.74504
Residuals	96	73863558	769412		

  - F-test by manual calculation:

$$F = \frac{(73945387 - 73863558)/1}{73863558/96} = \frac{81829/1}{73863558/96} = 0.1063526.$$

  - The p-value is `pf(F.stat, 1, 96, low.tail=F)` $=0.7450453$. We cannot reject `fit1` in favor of `fit0`.

$\square$

§3.4 Generalized Least Squares Estimation[1]

- Good properties of OLSEs, especially the (asymptotical) normality and consistency, make statistical inferences relatively simple.
- These good properties are ensured by the basic assumptions, especially the following three.

  (A2) Explanators are not random.

  (A4) Errors $\{X_i\}$ have constant variance (homoscedastic).

  (A5) Errors $\{X_i\}$ are serially uncorrelated.

- However, in practice, these assumptions are frequently disrupted. Consequently, OLSEs could be inefficient, inconsistent, and/or non-normal.
- **Generalized Least Squares Estimation (GLSE)** is needed.

[1] More details are introduced in STAT8007: *Statistical methods in economics and finance.*

3.4.1: GLSE for LM with heteroscedastic errors

- For illustration, consider the following rent-income data.

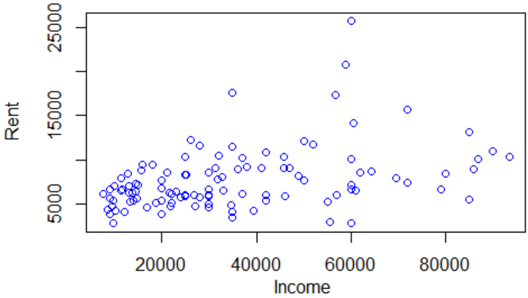

Figure 3.3: Annual rents and incomes for a sample of New Yorkers.

- We are estimating a simple LM for this,

$$Rent_i = \beta_0 + \beta_1 Income_i + \varepsilon_i.$$

Heteroscedasticity

- **Observation**: data with larger incomes are more diversified.
- **Check understanding**: Consumers with low values of income have little scope for varying their rent expenditures, and hence $\mathrm{Var}(\varepsilon_i)$ is low. On the other hand, wealthy consumers can choose to spend a lot of money on rent, or to spend less, depending on tastes, as a result, $\mathrm{Var}(\varepsilon_i)$ is high.
- **Heteroscedasticity** presents in the model: the variance of ε_i, and hence of $Rent_i$, is NOT a constant σ^2.
- In general, we have

$$\mathrm{Var}(\varepsilon_i) = \sigma_i^2, \qquad i = 1, 2, \cdots, n.$$

- In other words, basic assumption (A4) is destroyed.

Problems with heteroscedasticity

- Under heteroscedasticity, OLSEs are still unbiased and consistent.
- However, OLSEs are inefficient — they are no longer the BLUEs.
- More importantly, SER in formulas (3.8) or (3.14) are WRONG! There is NOT any unified σ^2 to be estimated.
- Consequently,
  - T-tests based on OLS SER are WRONG.
  - F-tests (need homoscedasticity for the F-distribution) are WRONG.
  - Confidence intervals and prediction intervals are invalid.

GLSE for LM with heteroscedastic errors

- For simplicity, rewrite the model as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \mathrm{Var}(\varepsilon_i) = \sigma_i^2, \qquad 1 \le i \le n.$$

- Basic assumptions other than (A4) are satisfied.

- Taking into account our comprehension of the data's characteristics, we may *assume* that error variances are proportional to incomes, that is,

$$\sigma_i^2 = \sigma^2 X_i \quad \text{for some constant } \sigma^2 > 0.$$

- Based on this assumption, we may *transform* the data as follows.

$$\widetilde{Y}_i = \frac{Y_i}{\sqrt{X_i}}, \quad \widetilde{X}_{1i} = \frac{1}{\sqrt{X_i}}, \quad \widetilde{X}_{2i} = \sqrt{X_i}, \quad \widetilde{\varepsilon}_i = \frac{\varepsilon_i}{\sqrt{X_i}}.$$

GLSE for LM with heteroscedastic errors

- Then, the model can be rewritten into

$$\widetilde{Y}_i = \beta_0 \widetilde{X}_{1i} + \beta_1 \widetilde{X}_{2i} + \widetilde{\varepsilon}_i. \tag{3.24}$$

- Notice that the transformed model (3.24) is a multiple linear model without intercept (or, *through the origin*).
- It is not difficult to check that all basic assumptions (A1) through (A6) are satisfied.
- Finally, apply OLS to the transformed model (3.24).
- Statistical inferences introduced in §3.3 become valid.
- We call the above process of estimation the **Generalized Least Squares Estimation (GLSE)**.

Weighted Least Squares Estimation (WLSE)

- Notice that the SSE functions for OLSE (S_o) and GLSE (S_g) are

$$S_o(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2,$$

$$S_g(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(\widetilde{Y}_i - \beta_0 \widetilde{X}_{1i} - \beta_1 \widetilde{X}_{2i}\right)^2,$$

$$= \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{X_i}.$$

- Comparing these two SSEs, we see that a sequence of *weights* $\{W_i = X_i^{-1}\}$ are added to terms in the summation.

- Therefore, such a GLSE is usually referred to the **Weighted Least Squares Estimation (WLSE)**.

Feasible GLSE (FGLSE)

- The aforesaid GLSE is based on our assumption of the heteroscedasticity: $\sigma_i^2 \propto X_i$.
- We are not quite sure whether the transformed errors $\{\widetilde{\varepsilon}_i\}$ are homoscedastic. The **Breusch-Pagan test**, including the **White's test** as a special case, can be used for this problem. The test should be applied to the original errors $\{\varepsilon_i\}$ before conducting the GLSE process.
- If we assume $\sigma_i^2 \propto X_i^2$, another sequence of weights, $\{W_i = X_i^{-2}\}$.
- We can make a weaker but more feasible assumption: $\sigma_i^2 \propto X_i^d$, where $d > 0$ is an unknown parameter to be estimated or tuned. In this case, we call this the **Feasible Generalized Least Squares Estimation (FGLSE)**.

Example 3.3: Rent-Income

- **Example 3.3**: Consider the rent-income data and suppose that we want to estimate the relationship

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where Y_i and X_i stand for rent expenditure and income, respectively.

- The OLSEs of coefficients are summarized as follows.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.455e+03  6.028e+02   9.051 7.67e-15 ***
x           6.357e-02  1.439e-02   4.418 2.42e-05 ***
```

Example 3.3: Rent-Income

- Heteroscedasticity is tested to be significant.

```
> library(lmtest)
> bptest(fit0)

        studentized Breusch-Pagan test

data:  fit0
BP = 5.607, df = 1, p-value = 0.01789
```

- The GLSE in Eq. (3.24) can be manually conducted using the following R codes.

```
# Manual GLSE: transform the data and OLS
x1 <- 1/sqrt(x)
x2 <- sqrt(x)
y1 <- y/sqrt(x)
fit1 <- lm(y1 ~ x1+x2-1)   # LM without intercept
```

Example 3.3: Rent-Income

- The GLSEs of coefficients are given below.

```
> summary(fit1)$coefficients
       Estimate   Std. Error   t value     Pr(>|t|)
x1 5.085513e+03 411.42410002 12.360755 2.851614e-22
x2 7.396252e-02   0.01426932  5.183325 1.049165e-06
```

- Heteroscedasticity becomes insignificant in the transformed model.

```
> bptest(fit1)

        studentized Breusch-Pagan test

data:  fit1
BP = 1.7005, df = 1, p-value = 0.1922
```

Example 3.3: Rent-Income

- The GLSE/WLSE can be automatically conducted using the following codes.

```r
# WLSE: Define weights and WLS
w <- 1/x
fit2 <- lm(y ~ x, weights = w)
summary(fit2)$coefficients
```

- The same estimations, as by manual GLSE, are obtained.
- However, when we apply the BP test to `fit2`, the same testing results, as by manual OLSE, will be obtained.
- This is because we modified the SSE function but didn't do any transformation to the data. □

3.4.2: GLSE for LM with serial correlation

- **Serial correlation** occurs in a time series data $\{X_t\}$ when X_t is correlated with some lagged version of itself, e.g., X_{t-1}.
- When considering regression models with time series data such as annual GDPs, monthly inflations, etc., we have to be aware of possible serial correlation in the data.
- For example, we are regression monthly CPIs Y_t against monthly Inflations X_t,

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \qquad 1 \le t \le T.$$

  Both $\{X_t\}$ and $\{Y_t\}$, and hence $\{\varepsilon_t\}$, are (common sense) serially correlated.
- OLSE for such an LM with serial correlation will be unbiased, consistent, but inefficient. More importantly, estimated SER is invalid.

- Define a null hypothesis:

$$H_0 : \{\varepsilon_t\} \text{ is serially uncorrelated.}$$

- The **Durbin-Watson test** can be used to test the *presence of lag 1 serial correlation*, i.e., to test H_0 against

$$H_a : \mathbb{E}(\varepsilon_t \varepsilon_{t-1}) \neq 0.$$

- The **Breusch-Godfrey test** tests H_0 against

$$H_a : \text{ disturbances are serially correlated among the first } L \text{ lags (up to lag } L\text{),}$$

where $L \geq 1$ is a pre-specified upper bound of lags.

GLSE for LM with serial correlation

- When serial correlation presents, like for heteroscedasticity, we need an assumption on the structure of the serial correlation to conduct the GLSE.
- The most commonly used structure is: assume $\{\varepsilon_t\}$ are firstly order autoregressive,

$$\varepsilon_t = \rho\varepsilon_{t-1} + \widetilde{\varepsilon}_t,$$

  where $|\rho| < 1$ is an unknown constant, and $\{\widetilde{\varepsilon}_t\}$ are i.i.d with mean 0 and variance σ^2.
- *If ρ is known*, we may transform the data as follows,

$$\widetilde{Y}_t = Y_t - \rho Y_{t-1}, \quad \widetilde{X}_t = X_t - \rho X_{t-1}, \qquad 2 \leq t \leq T.$$

- The transformed model becomes

$$\widetilde{Y}_t = \widetilde{\beta}_0 + \beta_1 \widetilde{X}_t + \widetilde{\varepsilon}_t, \qquad 2 \leq t \leq T,$$

  where $\widetilde{\beta}_0 = \beta_0(1 - \rho)$.

GLSE for LM with serial correlation

- However, ρ is actually unknown. We may *pre-estimate* it using the OLS residuals.
- Such a GLSE is referred to as the **Cochrane-Orcutt** estimation.
- Another slightly different but more efficient estimation is the **Prais-Winsten** estimation. It rescues the first observation ($t = 1$) and uses certain iterative computation to improve the accuracy and efficiency of the estimation.
- A third method is the **Non-linear Least Squares Estimation (NLSE)**, the following non-linear regression model is estimated on the whole.

$$Y_t = \rho Y_{t-1} + \beta_0(1-\rho) + \beta_1(X_t - X_{t-1}) + \widetilde{\varepsilon}_t, \qquad 1 \le t \le T.$$

Consistent estimated standard errors

3.4.3: Consistent estimated standard errors

- GLSEs are not automated.
- Assumptions on the structure of heteroscedasticity and/or serial correlation can be sometime difficult.
- Nevertheless, OLSEs are not that bad — they are unbiased and consistent.
- A *shortcut* frequently in econometrics is: *using OLSE and revising the formulas for estimated standard errors* to ensure valid statistical inferences.
- Notice that the OLSE of each coefficient β is a linear function of Y_i's. Denote it as

$$\widehat{\beta} = \sum_{i=1}^{n} w_i Y_i.$$

Consistent estimated standard errors

- The **White's consistent standard errors** for $\widehat{\beta}$ is defined as

$$e.s.e.(\widehat{\beta}) = \sqrt{\sum_{i=1}^{n} w_i^2 e_i^2},$$

where e_i's are OLS residuals.

- For LM with serial correlation, or both heteroscedasticity and serial correlation, we use the the following **Newey-West consistent standard errors**,

$$e.s.e(\widehat{\beta}) = \sqrt{\sum_{t=1}^{T} \sum_{s=t-L}^{t+L} w_t w_s e_t e_s},$$

where e_i's are OLS residuals, and $L > 0$ is a pre-specified integer standing for the order of serial correlation.

- **Example 3.4**: Consistent estimated standard errors.
- Consider the rent-income data first.
- The White's consistent standard errors, as well as corresponding tests, are conducted using the following R codes.

```
# White's Robust Standard Errors for rent-income model
library(sandwich)
summary.white <- function(model) {
  print(coeftest(model, vcov. = vcovHC))
  print(waldtest(model, vcov = vcovHC))
}                          # Define a function
summary(fit0)
summary.white(fit0)
```

  - R string `vcovHC` stands for **H**eteroscedasticity **C**onsistent variances and covariances.

Example 3.4: the rent-income model

- Results from OLSE with White's consistent standard errors:

```
> summary.white(fit0)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 5.4555e+03 4.0984e+02  13.311 < 2.2e-16 ***
x           6.3568e-02 1.5060e-02   4.221 5.151e-05 ***
---
Wald test

Model 1: y ~ x
Model 2: y ~ 1
  Res.Df Df      F    Pr(>F)
1    106
2    107 -1 17.817 5.151e-05 ***
```

Example 3.4: the rent-income model

- Compared with results from pure OLSE:

```
> summary(fit0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.455e+03  6.028e+02   9.051 7.67e-15 ***
x           6.357e-02  1.439e-02   4.418 2.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3303 on 106 degrees of freedom
Multiple R-squared:  0.1555,	Adjusted R-squared:  0.1475 
F-statistic: 19.51 on 1 and 106 DF,  p-value: 2.417e-05
```

  - Estimates of coefficients are the same.
  - Estimated standard errors and test results are revised.

Example 3.4: the poverty-unemployment model

- Consider one more data: poverty rate (variable `P`) and unemployment rate (variable `U`) for $T = 24$ years (from 1980 to 2003).
- The economists ask: *How much does the poverty rate rise when the unemployment rate rises?*
- The following simple regression model is used to address this question.

$$P_t = \beta_0 + \beta_1 U_t + \varepsilon_t, \qquad 1 \le t \le T.$$

- This is a regression model with time series data.

Example 3.4: the poverty-unemployment model

- Both the **Durbin-Watson test** and the **Breusch-Godfrey test** show that serial correlations are significant.
- Both GLSEs, the **Cochrane-Orcutt Estimation** and the **Prais-Winsten Estimation**, are conducted.
- The NLSE is also provided in the R script.
- We focus on statistical inferences with the Newey-West consistent standard errors given by R string `vcovHAC`.
    - `HAC` stands for Heteroscedasticity and Autocorrelation Consistent.

Example 3.4: the poverty-unemployment model

- Results from OLSE and the Newey-West consistent SE.

```
> summary.nw(fit.ols)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 9.792052   0.638355 15.3395 3.128e-13 ***
U           0.586614   0.096963  6.0499 4.339e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Wald test

Model 1: P ~ U
Model 2: P ~ 1
  Res.Df Df      F    Pr(>F)
1     22
2     23 -1 36.601 4.339e-06 ***
```

Example 3.4: the poverty-unemployment model

- Results from pure OLSE.

```
> summary(fit.ols)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.79205    0.61119  16.021 1.30e-13 ***
U            0.58661    0.09473   6.193 3.12e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6763 on 22 degrees of freedom
Multiple R-squared:  0.6355,    Adjusted R-squared:  0.6189
F-statistic: 38.35 on 1 and 22 DF,  p-value: 3.116e-06
```

□