Linear Regression Model

In a linear regression model, a response variable is modeled as a linear function of one or several explanatory variables (predictors). We assume that predictors can be measured precisely and so they are not random and can be treated as constants. We also assume that the response variable cannot be measured precisely and it equals the true value plus a measurement error which is a random variable. So the response variable is also random and its distribution can be used to estimate parameters of the linear model using the maximum likelihood approach.

Nonlinear relationships between the response and predictors can be modeled by including high-order powers of these predictors.

Example 1: Consider the following model:

$$Y_i = \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad \epsilon_1, \ldots, \epsilon_n \text{ are independent and have } N(0, \sigma^2) \text{ distribution.}$$

This is a model with two predictors, x_i and x_i^2, and with no intercept. Here we assume a quadratic relationship between Y_i and x_i. The values x_i are not random; these are constants and Y_i is a random variable because it depends on the random measurement error ϵ_i.

We have three unknown parameters in this model: β_1, β_2 and σ^2. These parameters can be estimated using the maximum likelihood approach.

Step 1. Denote $\mu_i = \beta_1 x_i + \beta_2 x_i^2$. We have $Y_i = \mu_i + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$ and therefore $Y_i \sim N(\mu_i, \sigma^2)$. It implies that the pdf of Y_i is $f_i(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i-\mu_i)^2}{2\sigma^2}\right\}$. Note that Y_i are not identically distributed but these variables are independent and therefore the likelihood function is the product of marginal pdfs:

$$L(\beta_1, \beta_2, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i-\mu_i)^2}{2\sigma^2}\right\} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu_i)^2\right\}.$$

We take the logarithm and replace $\mu_i = \beta_1 x_i + \beta_2 x_i^2$:

$$\ell(\beta_1, \beta_2, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_1 x_i - \beta_2 x_i^2)^2 .$$

Step 2. We maximize the log-likelihood function $\ell(\beta_1, \beta_2, \sigma^2)$ with respect to unknown parameters by taking derivatives:

$$\begin{aligned}
\frac{\partial \ell(\beta_1, \beta_2, \sigma^2)}{\partial \beta_1} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2x_i(Y_i - \beta_1 x_i - \beta_2 x_i^2) = 0, \\
\frac{\partial \ell(\beta_1, \beta_2, \sigma^2)}{\partial \beta_2} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2x_i^2(Y_i - \beta_1 x_i - \beta_2 x_i^2) = 0, \\
\frac{\partial \ell(\beta_1, \beta_2, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \beta_1 x_i - \beta_2 x_i^2)^2 = 0.
\end{aligned}$$

From the first two equations, we get:

$$\begin{cases}
\beta_1 \sum_{i=1}^n x_i^2 + \beta_2 \sum_{i=1}^n x_i^3 = \sum_{i=1}^n x_i Y_i, \\
\beta_1 \sum_{i=1}^n x_i^3 + \beta_2 \sum_{i=1}^n x_i^4 = \sum_{i=1}^n x_i^2 Y_i.
\end{cases}$$

Let $m_2 = \sum_{i=1}^n x_i^2$, $m_3 = \sum_{i=1}^n x_i^3$ and $m_4 = \sum_{i=1}^n x_i^4$. Solving this system of equations, we get:

$$\hat{\beta}_1 = \frac{m_4 \sum_{i=1}^n x_i Y_i - m_3 \sum_{i=1}^n x_i^2 Y_i}{m_2 m_4 - m_3^2}, \quad \hat{\beta}_2 = \frac{m_2 \sum_{i=1}^n x_i^2 Y_i - m_3 \sum_{i=1}^n x_i Y_i}{m_2 m_4 - m_3^2}.$$

From the third equation, we get:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)^2.$$

Step 3. We can study the properties of these estimators. We start with $\hat{\beta}_1$ and $\hat{\beta}_2$. Since these estimators are linear functions of independent normal random variables $Y_1, \ldots, Y_n$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are normal random variables. We can find the mean of these estimators using the linearity property of expectation:

$$\begin{aligned}
E(\hat{\beta}_1) &= \frac{m_4 \sum_{i=1}^n x_i E(Y_i) - m_3 \sum_{i=1}^n x_i^2 E(Y_i)}{m_2 m_4 - m_3^2} = \frac{m_4 \sum_{i=1}^n x_i (\beta_1 x_i + \beta_2 x_i^2) - m_3 \sum_{i=1}^n x_i^2 (\beta_1 x_i + \beta_2 x_i^2)}{m_2 m_4 - m_3^2} \\
&= \frac{m_4(\beta_1 m_2 + \beta_2 m_3) - m_3(\beta_1 m_3 + \beta_2 m_4)}{m_2 m_4 - m_3^2} = \beta_1, \\
E(\hat{\beta}_2) &= \frac{m_2 \sum_{i=1}^n x_i^2 E(Y_i) - m_3 \sum_{i=1}^n x_i E(Y_i)}{m_2 m_4 - m_3^2} = \frac{m_2 \sum_{i=1}^n x_i^2 (\beta_1 x_i + \beta_2 x_i^2) - m_3 \sum_{i=1}^n x_i (\beta_1 x_i + \beta_2 x_i^2)}{m_2 m_4 - m_3^2} \\
&= \frac{m_2(\beta_1 m_3 + \beta_2 m_4) - m_3(\beta_1 m_2 + \beta_2 m_3)}{m_2 m_4 - m_3^2} = \beta_2.
\end{aligned}$$

We can see that $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimators of β_1 and β_2, respectively. Now we find the variance of these estimators using the linearity property of variance: if $Y_1, \ldots, Y_n$ are independent, then $Var(\sum_{i=1}^n c_i Y_i) = \sum_{i=1}^n c_i^2 Var(Y_i)$. To apply this property, we combine all terms with Y_i to write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (m_4 x_i - m_3 x_i^2) Y_i}{m_2 m_4 - m_3^2}, \quad \hat{\beta}_2 = \frac{\sum_{i=1}^n (m_2 x_i^2 - m_3 x_i) Y_i}{m_2 m_4 - m_3^2}, \tag{1}$$

$$\begin{aligned}
Var(\hat{\beta}_1) &= \frac{\sum_{i=1}^n (m_4 x_i - m_3 x_i^2)^2 \cdot Var(Y_i)}{(m_2 m_4 - m_3^2)^2} = \sigma^2 \frac{\sum_{i=1}^n (m_4 x_i - m_3 x_i^2)^2}{(m_2 m_4 - m_3^2)^2}, \\
Var(\hat{\beta}_2) &= \frac{\sum_{i=1}^n (m_2 x_i^2 - m_3 x_i)^2 \cdot Var(Y_i)}{(m_2 m_4 - m_3^2)^2} = \sigma^2 \frac{\sum_{i=1}^n (m_2 x_i^2 - m_3 x_i)^2}{(m_2 m_4 - m_3^2)^2}.
\end{aligned}$$

We can simplify these expressions:

$$\begin{aligned}
\sum_{i=1}^n (m_4 x_i - m_3 x_i^2)^2 &= \sum_{i=1}^n m_4^2 x_i^2 - 2 \sum_{i=1}^n m_4 m_3 x_i^3 + \sum_{i=1}^n m_3^2 x_i^4 = m_4^2 m_2 - 2 m_4 m_3 m_3 + m_3^2 m_4 = m_4(m_2 m_4 - m_3^2), \\
\sum_{i=1}^n (m_2 x_i^2 - m_3 x_i)^2 &= \sum_{i=1}^n m_2^2 x_i^4 - 2 \sum_{i=1}^n m_2 m_3 x_i^3 + \sum_{i=1}^n m_3^2 x_i^2 = m_2 m_4^2 - 2 m_2 m_3 m_3 + m_2 m_3^2 = m_2(m_2 m_4 - m_3^2).
\end{aligned}$$

Finally we can write,

$$Var(\hat{\beta}_1) = \frac{m_4 \sigma^2}{m_2 m_4 - m_3^2}, \quad Var(\hat{\beta}_2) = \frac{m_2 \sigma^2}{m_2 m_4 - m_3^2}, \tag{2}$$

and hence

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{m_4\sigma^2}{m_2m_4 - m_3^2}\right), \quad \hat{\beta}_2 \sim N\left(\beta_2, \frac{m_2\sigma^2}{m_2m_4 - m_3^2}\right).$$

Step 4. To study the properties of $\hat{\sigma}^2$, we can prove the **analysis of variance identity**:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_1 x_i - \beta_2 x_i^2)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 - \beta_1 x_i - \beta_2 x_i^2)^2.$$

Let the predicted value $\hat{Y}_i = \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$. We have:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_1 x_i - \beta_2 x_i^2)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \beta_1 x_i - \beta_2 x_i^2)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - \beta_1 x_i - \beta_2 x_i^2)^2 - \frac{2}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \beta_1 x_i - \beta_2 x_i^2).$$

We need to show that the third term on the right hand side is zero. Recall that $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy these equations:

$$\begin{cases} \hat{\beta}_1 \sum_{i=1}^n x_i^2 + \hat{\beta}_2 \sum_{i=1}^n x_i^3 = \sum_{i=1}^n x_i Y_i, \\ \hat{\beta}_1 \sum_{i=1}^n x_i^3 + \hat{\beta}_2 \sum_{i=1}^n x_i^4 = \sum_{i=1}^n x_i^2 Y_i. \end{cases}$$

We can rewrite these equations:

$$\begin{cases} \sum_{i=1}^n (Y_i - \hat{Y}_i) x_i = 0, \\ \sum_{i=1}^n (Y_i - \hat{Y}_i) x_i^2 = 0. \end{cases}$$

We find:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \beta_1 x_i - \beta_2 x_i^2) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i - \beta_1\sum_{i=1}^{n}(Y_i - \hat{Y}_i)x_i - \beta_2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)x_i^2$$

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2) = \hat{\beta}_1\sum_{i=1}^{n}(Y_i - \hat{Y}_i)x_i + \hat{\beta}_2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)x_i^2 = 0.$$

So we proved that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_1 x_i - \beta_2 x_i^2)^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\hat{\sigma}^2} + \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 - \beta_1 x_i - \beta_2 x_i^2)^2.$$

It follows that

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n}E\{(Y_i - \beta_1 x_i - \beta_2 x_i^2)^2\} - \frac{1}{n}\sum_{i=1}^{n}E\{(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 - \beta_1 x_i - \beta_2 x_i^2)^2\}.$$

We have:

$$E\{(Y_i - \beta_1 x_i - \beta_2 x_i^2)^2\} = E\{(Y_i - E(Y_i))^2\} = Var(Y_i) = \sigma^2,$$

$$E\{(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 - \beta_1 x_i - \beta_2 x_i^2)^2\} = E\{(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 - E(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2))^2\} = Var(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2).$$

Note that

$$Var(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2) = x_i^2 Var(\hat{\beta}_1) + x_i^4 Var(\hat{\beta}_2) + 2x_i^3 Cov(\hat{\beta}_1, \hat{\beta}_2) \tag{3}$$

because $\hat{\beta}_1$ and $\hat{\beta}_2$ are dependent. Using independence of $Y_1, \ldots, Y_n$, linearity property of covariance and (1), we find:

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = \frac{\sum_{i=1}^n (m_4 x_i - m_3 x_i^2)(m_2 x_i^2 - m_3 x_i) Cov(Y_i, Y_i)}{(m_2 m_4 - m_3^2)^2}.$$

since all terms $Cov(Y_i, Y_j) = 0$ for $i \neq j$. Further, $Cov(Y_i, Y_i) = Var(Y_i) = \sigma^2$ and

$$\sum_{i=1}^n (m_4 x_i - m_3 x_i^2)(m_2 x_i^2 - m_3 x_i) = m_4 m_2 \sum_{i=1}^n x_i^3 - m_4 m_3 \sum_{i=1}^n x_i^2 - m_3 m_2 \sum_{i=1}^n x_i^4 + m_3^2 \sum_{i=1}^n x_i^3$$

$$= m_2 m_3 m_4 - m_2 m_3 m_4 - m_2 m_3 m_4 + m_3^3 = -m_3(m_2 m_4 - m_3^2),$$

and hence

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{m_3 \sigma^2}{m_2 m_4 - m_3^2}.$$

Using (2) and (3), we find:

$$Var(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2) = \frac{\sigma^2 (m_4 x_i^2 + m_2 x_i^4 - 2x_i^3 m_3)}{m_2 m_4 - m_3^2},$$

and

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^n Var(Y_i) - \frac{1}{n}\sum_{i=1}^n Var(\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2) = \sigma^2 - \frac{1}{n}\frac{\sigma^2 (\sum_{i=1}^n (m_4 x_i^2 + m_2 x_i^4 - 2x_i^3 m_3)}{m_2 m_4 - m_3^2}$$

$$= \sigma^2 - \frac{1}{n}\frac{\sigma^2 (m_4 m_2 + m_2 m_4 - 2m_3 m_3)}{m_2 m_4 - m_3^2} = \left(1 - \frac{2}{n}\right)\sigma^2.$$

We conclude that $\hat{\sigma}^2$ is biased and we can define an unbiased estimator of σ^2:

$$S^2 = \left(1 - \frac{2}{n}\right)^{-1} \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

It is possible to prove that $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$ and that

$$\frac{\sqrt{n(m_2 m_4 - m_3^2)}(\hat{\beta}_1 - \beta_1)}{S\sqrt{m_4}} \sim t(n-2), \quad \frac{\sqrt{n(m_2 m_4 - m_3^2)}(\hat{\beta}_2 - \beta_2)}{S\sqrt{m_2}} \sim t(n-2).$$

These three pivots can be used to construct the 95% confidence intervals for β_1, β_2 and σ^2.

We can see that lengthy calculations might be required to derive the properties of our estimators. Normality assumption in a linear regression model helps to obtain estimates in closed form. Things become more complicated if the distribution of errors, ϵ_i, is not normal. A model with non-normal errors is required if the normal qq-plot of $Y_1, \ldots, Y_n$ shows significant departures from a straight line. In particular, a distribution with heavier tails is required if the empirical quantiles of Y_i-s are larger than theoretical quantiles of the normal distribution.

Example 2: Consider the following model:

$$Y_i = \beta x_i + \epsilon_i, \quad \epsilon_1, \ldots, \epsilon_n \text{ are independent and } \epsilon_i \sim \text{Laplace}(\mu_i = x_i/\lambda),$$

where the density of the Laplace distribution with the scale parameter μ is

$$f(x;\mu) = \frac{1}{2\mu} \exp\left\{-\frac{|x|}{\mu}\right\}, \quad -\infty < x < \infty,\ \mu > 0.$$

There are two parameters in this model: β and λ. Again, x_i-s are some constants and Y_i-s are random variables. We can use the likelihood approach to estimate these parameters.

Step 1. We need to find the pdf of Y_i. Let $F(x;\mu_i)$ be the cdf of ϵ_i. We find the cdf of Y_i:

$$\Pr(Y_i < x) = \Pr(\beta x_i + \epsilon_i < x) = \Pr(\epsilon_i < x - \beta x_i) = F(x - \beta x_i; \mu_i).$$

The pdf of Y_i is

$$f_{Y_i}(x) = \frac{\partial \Pr(Y_i < x)}{\partial x} = f(x - \beta x_i; \mu_i) = \frac{1}{2\mu_i} \exp\left\{-\frac{|x - \beta x_i|}{\mu_i}\right\}.$$

The likelihood function is a product of marginal pdfs, $f_{Y_i}(Y_i)$, with $\mu_i = x_i/\lambda$:

$$L(\beta,\lambda) = \prod_{i=1}^{n} \frac{1}{2\mu_i} \exp\left\{-\frac{|Y_i - \beta x_i|}{\mu_i}\right\} = C\lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} \left|\frac{Y_i}{x_i} - \beta\right|\right\}$$

where $C = 2^{-n} \left(\prod_{i=1}^{n} x_i\right)^{-1}$. The log-likelihood function is

$$\ell(\beta,\lambda) = \ln C + n \ln \lambda - \lambda \sum_{i=1}^{n} \left|\frac{Y_i}{x_i} - \beta\right|.$$

Step 2. We need to maximize the log-likelihood function with respect to unknown parameters. Unfortunately, $\ell(\beta,\lambda)$ is not a differentiable function of β so we cannot take derivative with respect to β. To maximize $\ell(\beta,\lambda)$ with respect to β, we need to minimize $\sum_{i=1}^{n} \left|\frac{Y_i}{x_i} - \beta\right|$. We know that $\hat{\beta} = \text{median}(Y_i/x_i)$ (check question 1b from week 8 workshop).

The log-likelihood is a differentiable function of λ so

$$\frac{\partial \ell(\beta,\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} \left|\frac{Y_i}{x_i} - \hat{\beta}\right| = 0 \quad \Rightarrow \quad \hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} \left|\frac{Y_i}{x_i} - \hat{\beta}\right|\right)^{-1}.$$

Step 3. What are the properties of $\hat{\beta}$ and $\hat{\lambda}$? For simplicity we assume that $n = 2k+1$ and let $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ are variables $Z_i = Y_i/x_i$ sorted in an increasing order. Then $\hat{\beta} = Z_{(k+1)}$ and $\hat{\lambda} = \left(\frac{1}{n}\sum_{j=1}^{k}(Z_{(k+1+j)} - Z_{(j)})\right)^{-1}$. Note that

$$\Pr(Z_i < z) = \Pr(Y_i/x_i < z) = \Pr(Y_i < z x_i) = \Pr(\beta x_i + \epsilon_i < z x_i)$$

$$= \Pr(\epsilon_i < (z - \beta) x_i) = F((z - \beta) x_i; \mu_i),$$

and hence the density of Z_i is

$$f_Z(z) = \frac{\partial \Pr(Z_i < z)}{\partial z} = x_i f\left((z-\beta)x_i; \mu_i = \frac{x_i}{\lambda}\right) = \frac{\lambda}{2}\exp\{-\lambda|z-\beta|\}, \quad -\infty < z < \infty.$$

This is a Laplace distribution Laplace(β, λ) with the shift parameter β. $\hat{\beta} = Z_{(k+1)}$ is therefore the $(k+1)$-st order statistic and its pdf is (check material from Module 7.5)

$$f_{\hat{\beta}}(z) = \frac{(2k+1)!}{(k!)^2}\{F_Z(z)(1-F_Z(z))\}^k f_Z(z),$$

where F_Z is the cdf of $Z \sim$ Laplace(β, λ). Clearly, this is not a normal distribution and it can be shown that

$$E(\hat{\beta}) = \frac{(2k+1)!}{(k!)^2}\int_{-\infty}^{\infty} z\{F_Z(z)(1-F_Z(z))\}^k f_Z(z)dz = \beta,$$

so that $\hat{\beta}$ is an unbiased estimator of β. Interestingly, $\tilde{\beta} = \hat{\beta} - \beta$ is a pivot if λ is known because the pdf of $\tilde{\beta}$ is

$$f_{\tilde{\beta}}(z) = f_{\hat{\beta}}(z+\beta) = \frac{(2k+1)!}{(k!)^2}\left\{F\left(z; \mu = \frac{1}{\lambda}\right)\left[1 - F\left(z; \mu = \frac{1}{\lambda}\right)\right]\right\}^k f\left(z; \mu = \frac{1}{\lambda}\right),$$

where $F(z;\mu)$ and $f(z;\mu)$ is the cdf and pdf of the Laplace distribution with the scale parameter μ (and shift parameter is zero). Quantiles of $\tilde{\beta}$ can be used to construct the 95% confidence interval for β. It is also possible to check that $\tilde{\lambda} = \hat{\lambda}/\lambda$ is a pivot and can be used to construct confidence intervals for λ. The distribution of $\tilde{\lambda}$ is very complex and is not shown here.