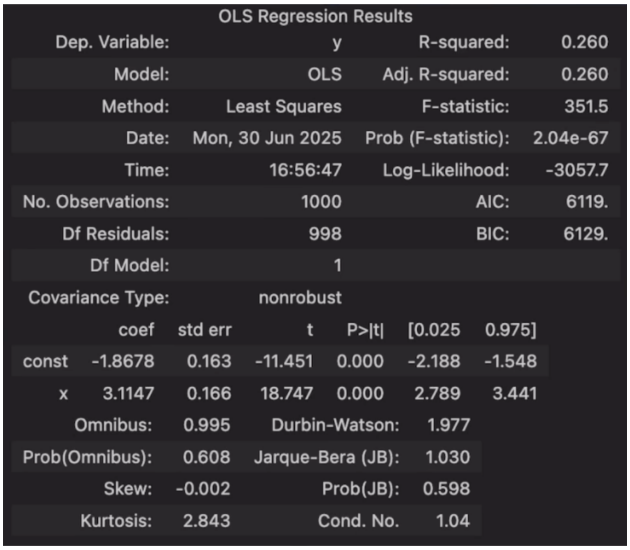
Chapter 1

Linear Regression Models

Linear regression models are among the simplest and most fundamental models in statistical machine learning, often serving as the starting point for learning more advanced techniques. These models have the lowest complexity and is widely used to analyze low signal-to-noise ratio datasets like low-frequency financial data in quantitative research, where *overfitting* is the main issue to be avoided[1]. Additionally, they are usually used as the baseline for benching more sophisticated models.

Linear regression models are implemented in many statistical packages. The table in the Fig 1.1 contains regression results generated by the "statsmodels" package. Our goal in this set of notes is reproducing this table by ourselves.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.260
Model:	OLS	Adj. R-squared:	0.260
Method:	Least Squares	F-statistic:	351.5
Date:	Mon, 30 Jun 2025	Prob (F-statistic):	2.04e-67
Time:	16:56:47	Log-Likelihood:	-3057.7
No. Observations:	1000	AIC:	6119.
Df Residuals:	998	BIC:	6129.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.8678	0.163	-11.451	0.000	-2.188	-1.548
x	3.1147	0.166	18.747	0.000	2.789	3.441

Omnibus:	0.995	Durbin-Watson:	1.977
Prob(Omnibus):	0.608	Jarque-Bera (JB):	1.030
Skew:	-0.002	Prob(JB):	0.598
Kurtosis:	2.843	Cond. No.	1.04

Figure 1.1: Result sample generated by the "statsmodels" package.

1.1 Introduction

Given *predictors* (*features*) $X^{(1)}, X^{(2)}, \ldots, X^{(p-1)}$, $p \geq 2$ and the corresponding *outcome* (*target*) Y, the goal of linear regression models is finding linear relationships between the outcomes and all predictors

$$Y = \beta_0 + \beta_1 X^{(1)} + \cdots + \beta_{p-1} X^{(p-1)} + \varepsilon, \tag{1.1}$$

[1] In the context of statistical machine learning, this is often refereed as bias-variance tradeoff.

where ε is a random noise. In practice, we usually have some observed data points $\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p-1)}; y_i\}_{i=1}^n$ that can be used to estimate the model, i.e., those unknown regression coefficients, namely $\beta_0, \beta_1, \ldots, \beta_{p-1}$, we denote the estimated coefficients as $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}$ correspondingly[2]. When a new data point (whose outcome is missing) is given, we can let the model process the predictors and thus generate the prediction of this missing outcome.

The joint distribution of $X^{(1)}, X^{(2)}, \ldots, X^{(p-1)}, Y$ is extremely complicated in real world, especially when p is large. Therefore, we focus on modeling the conditional statistical properties instead. To be specific, we assume that the conditional expectation of Y given $X^{(j)} = x^{(j)}$, $1 \le j \le p-1$ is linear, that is

$$\mathbb{E}\big[Y | X^{(j)} = x^{(j)}, 1 \le j \le p-1\big] = \beta_0 + \sum_{j=1}^{p-1} \beta_j x^{(j)}. \tag{1.2}$$

Which means that, for each data point, we have

$$\mathbb{E}\big[Y | X_i^{(j)} = x_i^{(j)}, 1 \le j \le p-1\big] = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_i^{(j)}. \tag{1.3}$$

We also assume that $\{Y_i\}_{i=1}^N$ are conditional independent.

For simplicity, we omit the conditional expectation notation in the following discussions. Nevertheless, it is important to keep in mind that all expectations are taken conditionally.

To fully determine the conditional distribution of Y, we also need to specify the conditional distribution of the random noise ε. In these notes, we use normal distributions with common variance σ^2 to model ε. Thus, the conditional distribution of Y_i is

$$Y \Big| \left\{ X_i^{(j)} = x_i^{(j)} \right\}_{j=1}^{p-1} \sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_i^{(j)}, \sigma^2 \right). \tag{1.4}$$

As previously explained, we may omit the conditional notation in Eq. (1.4) for simplicity. The expression can then be written as follows

$$Y \sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_i^{(j)}, \sigma^2 \right). \tag{1.5}$$

However, we emphasize that Eq. (1.5), along with all results derived from it, should be understood as conditional statements.

Additionally, we emphasize that the variance σ^2 of the error term ε is assumed to be a constant across all inputs $\{x_i^{(j)}\}_{j=1}^n$, a property known as *homoscedasticity*. In contrast, if the variance depends on the inputs $\{x_i^{(j)}\}_{j=1}^n$, the model is said to exhibit *heteroscedasticity*.

1.2 Simple Linear Regression

When $p = 2$, the linear regression model becomes $Y = \beta_0 + \beta_1 X + \varepsilon$, which is called *simple linear regression model*.

1.2.1 Fitting the Model

Theorem 1.2.1. The regression coefficients of a simple linear regression model are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{and } \hat{\beta}_0 = \bar{Y} - \beta_1 \bar{x}, \tag{1.6}$$

[2] They are random variables

where

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{1.7}$$

Notation Remark: We will use $\hat{\xi}$ as our estimated ξ, it can be a number (without randomness) or a random variable (with randomness) depending the context.

Proof. Our best estimates for β_0 and β_1 should minimize the error between the observed responses and the predicted values. One reasonable choice of the error is the mean square error (MSE)[3]

$$F(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 x_i)\right]^2,$$

and we have the following equation

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{\beta_0, \beta_1} F(\beta_0, \beta_1).$$

The minimum condition gives the equations for $\hat{\beta}_0, \hat{\beta}_1$

$$\begin{cases} \left.\dfrac{\partial F}{\partial \beta_0}\right|_{\hat{\beta}_0} = 0, & (1.8) \\ \left.\dfrac{\partial F}{\partial \beta_1}\right|_{\hat{\beta}_1} = 0. & (1.9) \end{cases}$$

Eq. (1.8) gives

$$\begin{aligned} \frac{-2}{n}\sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right] &= 0, \qquad (1.10) \\ \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i) &= \sum_{i=1}^{n} Y_i, \\ n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} Y_i, \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

We obtained the expression for $\hat{\beta}_0$ in the theorem. We proceed our calculation with Eq. (1.9), it gives

$$\begin{aligned} \frac{-2}{n}\sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right] x_i &= 0, \qquad (1.11) \\ \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i &= \sum_{i=1}^{n} Y_i x_i, \end{aligned}$$

[3]We briefly note that this is also referred to as the L_2 error. Other choices, such as the L_1 error, are also possible.

Inserting the result for $\hat{\beta}_0$, yields

$$\begin{aligned}
(\bar{Y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} Y_i x_i, \\
\hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i) &= \sum_{i=1}^{n} Y_i x_i - \bar{Y} \sum_{i=1}^{n} x_i, \\
\hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i) &= \sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} Y_i \bar{x}, \\
\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i &= \sum_{i=1}^{n} (x_i - \bar{x}) Y_i, \\
\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) &= \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}), \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.
\end{aligned}$$

Note that, in the calculation, we used the observation that

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x}) \xi_i &= \sum_{i=1}^{n} (x_i - \bar{x}) \xi_i - \bar{\xi} \sum_{i=1}^{n} (x_i - \bar{x}) \\
&= - \sum_{i=1}^{n} (x_i - \bar{x})(\bar{\xi} - \xi_i),
\end{aligned}$$

since $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$. This concludes our proof. □

Corollary 1.2.1. The regression line must pass through the point $(\bar{x}, \bar{Y})$.

Proof. This is a simple observation from the regression coefficient $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \Rightarrow \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. □

Suppose that the observed values of $\mathbf{Y} = \{Y_i\}_{i=1}^n$ are $\mathbf{y} = \{y_i\}_{i=1}^n$. With the numbers $\mathbf{y} = \{y_i\}_{i=1}^n$, we can calculate the *fitted values*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{1.12}$$

and residuals

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{1.13}$$

Notation Remark: We will use boldface symbols like $\mathbf{A}, \ldots, \mathbf{Z}, \boldsymbol{\alpha}, \ldots, \boldsymbol{\omega}$ to represent vectors and matrices.

Theorem 1.2.2. Residuals always satisfy

$$\begin{cases} \sum_{i=1}^{n} e_i = 0, \\ \sum_{i=1}^{n} x_i e_i = 0. \end{cases} \tag{1.14}$$

Proof. This is a simple corollary from Eq. (1.10) and (1.11). □

The simple linear regression model can be trained by maximum likelihood estimators (MLE) as well.

Theorem 1.2.3. Define $\mathbf{x} = \{x_i\}_{i=1}^n$, then the likelihood function $\mathcal{L}_n$ evaluated at $\mathbf{y}$ is

$$\mathcal{L}_n(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right]. \tag{1.15}$$

The MLE of the regression coefficients are given in Eq. (1.6), while the MLE of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n}S^2, \tag{1.16}$$

where S^2 is the sum of squared errors under the best estimation for β_0 and β_1, that is, $S^2 = F(\hat{\beta}_0, \hat{\beta}_1)$.
Remark: The estimated $\hat{\sigma}^2$ is biased, in fact,

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

We will show this later.

Proof. Suppose that we are given a sample y_i from Y_i, the possibility of observing this particular sample is, i.e., its' likelihood function $\mathcal{L}$ is

$$\mathcal{L}_i(y_i|x_i, \beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]. \tag{1.17}$$

Since $\{Y_i\}_{i=1}^N$ are conditional independent, if we are giving a set of samples $\mathbf{y} = \{y_i\}_{i=1}^N$, we can simply multiply the likelihood function of single data point to get the likelihood function for $\mathbf{y}$:

$$\mathcal{L}_n(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right].$$

When training the model, we aim to maximize the likelihood function $\mathcal{L}_n$. In other words, we seek the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize the likelihood of observing the given dataset. Mathematically, its optimization is equivalent to minimizing the MSE[4], $nF(\beta_0, \beta_1) = \sum_{i=1}^{n}\left[Y_i - (\beta_0 + \beta_1 x_i)\right]^2$. As a result, both approaches yield the same estimates for β_0 and β_1 in the linear regression model, with their expressions given in Eq. (1.6).

Additionally, the best estimate of σ^2 should maximize $\mathcal{L}_n$ (the probability) as well, we have

$$\begin{aligned}
\frac{\partial \mathcal{L}_n}{\partial \sigma^2} &= 0, \\
\frac{-n/2}{(2\pi)^{n/2}}\frac{1}{(\sigma^2)^{n/2+1}} \exp\left[-\frac{F(\beta_0,\beta_1|\mathbf{y})}{2\sigma^2}\right] + \frac{-1}{(2\pi)^{n/2}}\frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{F(\beta_0,\beta_1|\mathbf{y})}{2\sigma^2}\right]\left(\frac{F(\beta_0,\beta_1|\mathbf{y})}{2}\frac{1}{(\sigma^2)^2}\right) &= 0, \\
-n + \frac{F(\beta_0,\beta_1|\mathbf{y})}{\sigma^2} &= 0, \\
\sigma^2 &= \frac{F(\beta_0,\beta_1|\mathbf{y})}{n}.
\end{aligned}$$

Since this analysis holds for any dataset, we can remove the "conditioned on $\mathbf{y}$" in the equations, and get the best estimation for σ^2 as $\hat{\sigma}^2 = \frac{1}{n}F(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n}S^2$. □

Notation Remark: We will use the notation

$$s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}, \tag{1.18}$$

in the following discussions.

[4]In the following equation, y_i is replaced to Y_i, since the argument holds for any dataset. It is similar to the argument we make later in the text for $\hat{\sigma}^2$.

Distributions of the estimators

Theorem 1.2.4. The following claims hold for distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

1.

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}\right)\right). \tag{1.19}$$

2.

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{s_x^2}\right). \tag{1.20}$$

3.

$$\mathbb{C}\mathrm{ov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{s_x^2}. \tag{1.21}$$

Proof. Let's start with $\hat{\beta}_1$ in Eq. (1.6), we have

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n Y_i(x_i - \bar{x})}{s_x^2}, \\
&= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \varepsilon_i)(x_i - \bar{x})}{s_x^2}, \\
&= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i)(x_i - \bar{x})}{s_x^2} + \sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x^2}\varepsilon_i, \\
&\sim \beta_1 \frac{\sum_{i=1}^n x_i(x_i - \bar{x})}{s_x^2} + \sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x^2}\mathcal{N}(0, \sigma^2), \\
&= \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{s_x^2} + \mathcal{N}\left(0, \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(s_x^2)^2}\sigma^2\right), \quad (\varepsilon_i \text{ are independent}) \\
&= \beta_1 + \mathcal{N}\left(0, \frac{\sigma^2}{s_x^2}\right), \\
&= \mathcal{N}\left(\beta_1, \frac{\sigma^2}{s_x^2}\right).
\end{aligned}$$

We obtain the result in Eq. (1.20). Note that, in the calculation, we repeatedly use the property that $\sum_i c(x_i - \bar{x}) = 0$, where c is a constant.

Next, let's calculate the distribution of $\hat{\beta}_0$, starting from Eq. (1.6),

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x}, \\
&= \bar{Y} - \frac{\sum_{i=1}^n Y_i (x_i - \bar{x})}{s_x^2} \bar{x}, \\
&= \frac{\sum_{i=1}^n \left[Y_i \frac{s_x^2}{n} - Y_i \bar{x}(x_i - \bar{x}) \right]}{s_x^2}, \\
&= \frac{1}{s_x^2} \sum_{i=1}^n \left[\frac{s_x^2}{n} - \bar{x}(x_i - \bar{x}) \right] (\beta_0 + \beta_1 x_i + \varepsilon_i), \\
&= \frac{1}{s_x^2} \left\{ \sum_{i=1}^n \left[\frac{s_x^2}{n} - \bar{x}(x_i - \bar{x}) \right] (\beta_0 + \beta_1 x_i) + \sum_{i=1}^n \left[\frac{s_x^2}{n} - \bar{x}(x_i - \bar{x}) \right] \varepsilon_i \right\}, \\
&\sim \frac{1}{s_x^2} \left\{ \beta_0 s_x^2 + \beta_1 \bar{x} s_x^2 - \beta_1 \bar{x} \sum_{i=1}^n (x_i - \bar{x})(x_i) + \sum_{i=1}^n \left[\frac{s_x^2}{n} - \bar{x}(x_i - \bar{x}) \right] \mathcal{N}(0, \sigma^2) \right\}, \\
&= \beta_0 + \sum_{i=1}^n \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_x^2} \right] \mathcal{N}(0, \sigma^2), \\
&= \beta_0 + \mathcal{N}\left(0, \sigma^2 \sum_{i=1}^n \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_x^2} \right]^2 \right), \quad (\varepsilon_i \text{ are independent}) \\
&= \mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right) \right),
\end{aligned}$$

where the sum of the variances in the final step is straightforward to compute,

$$\begin{aligned}
\sum_{i=1}^n \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_x^2} \right]^2 &= \sum_{i=1}^n \left[\frac{1}{n^2} + \frac{\bar{x}^2 (x_i - \bar{x})^2}{(s_x^2)^2} - \frac{2}{n} \frac{\bar{x}(x_i - \bar{x})}{s_x^2} \right], \\
&= \frac{1}{n} + \frac{\bar{x}^2}{s_x^2}.
\end{aligned}$$

We have to emphasize that we can not use the distribution of $\hat{\beta}_1$ we just calculated, since it is correlated with Y_i and we don't know the covariance yet.

To complete the proof of the theorem, it remains to show the covariance, for which we have

$$\begin{aligned}
\mathbb{C}\text{ov}(\hat{\beta}_0, \hat{\beta}_1) &= \mathbb{C}\text{ov}(\bar{Y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1), \\
&= \mathbb{C}\text{ov}(\bar{Y}, \hat{\beta}_1) - \bar{x}\, \mathbb{V}\text{ar}(\hat{\beta}_1), \\
&= \mathbb{C}\text{ov}\left(\sum_{i=1}^n \frac{1}{n} Y_i, \frac{\sum_{j=1}^n Y_j (x_j - \bar{x})}{s_x^2} \right) - \bar{x}\, \mathbb{V}\text{ar}(\hat{\beta}_1), \\
&= \sum_{i=1}^n \sum_{j=1}^n \frac{1}{n} \frac{(x_j - \bar{x})}{s_x^2} \mathbb{C}\text{ov}(Y_i, Y_j) - \bar{x} \frac{\sigma^2}{s_x^2}, \\
&= \sum_{i=1}^n \sum_{j=1}^n \frac{1}{n} \frac{(x_j - \bar{x})}{s_x^2} \sigma^2 \delta_{ij} - \frac{\bar{x}\sigma^2}{s_x^2}, \\
&= -\frac{\bar{x}\sigma^2}{s_x^2}.
\end{aligned}$$

So far we have concluded our proof. □

Problem 1.2.1. Write a program to verify these claims.
Problem 1.2.2. Let $H = c_0\hat{\beta}_0 + c_1\hat{\beta}_1$ be a linear combination of $\hat{\beta}_0$ and $\hat{\beta}_1$ show that the variance of H is given by

$$\mathbb{V}\mathrm{ar}(H) = \sigma^2 \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2} \right].$$

Proof. We have

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(H) &= \mathbb{V}\mathrm{ar}(c_0\hat{\beta}_0 + c_1\hat{\beta}_1), \\
&= c_0^2\,\mathbb{V}\mathrm{ar}(\hat{\beta}_0) + c_1^2\,\mathbb{V}\mathrm{ar}(\hat{\beta}_1) + 2c_0c_1\,\mathbb{C}\mathrm{ov}(\hat{\beta}_0, \hat{\beta}_1), \\
&= c_0^2\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right) + c_1^2 \frac{\sigma^2}{s_x^2} - 2c_0c_1 \frac{\bar{x}\sigma^2}{s_x^2}, \\
&= \sigma^2 \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2} \right].
\end{aligned}$$

□

Now suppose that a linear regression model has been trained through data points $\{x_i, y_i\}_{i=1}^n$. When a test data point x is coming, we can the use the model to predict the outcome Y for x.

By the setting of simple linear regression models, Y is a random variable following the normal distribution with mean $\beta_0 + \beta_1 x$ and variance σ^2. It is natural to use $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ as the prediction of Y, which will surely introduce errors. We can measure this error through the mean squared error.
Theorem 1.2.5. The MSE of Y is given by

$$\mathbb{E}\Big[\big(Y - \hat{Y}\big)^2\Big] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_x^2} \right]. \tag{1.22}$$

Remark: The MSE gets larger as x moves away from the observed training data. Is this correct intuitively?

Proof. The expectation we want to show can be written as

$$\begin{aligned}
\mathbb{E}\Big[\big(Y - \hat{Y}\big)^2\Big] &= \mathbb{E}\big[Y^2 + \hat{Y}^2 - 2Y\hat{Y}\big], \\
&= \mathbb{E}\big[Y^2\big] + \mathbb{E}\big[\hat{Y}^2\big] - 2\mathbb{E}\big[Y\hat{Y}\big].
\end{aligned}$$

For each term on the RHS, we have

$$\begin{aligned}
\mathbb{E}\big[Y^2\big] &= \mathbb{V}\mathrm{ar}(Y) + \mathbb{E}[Y]^2, \\
&= \sigma^2 + (\beta_0 + \beta_1 x)^2, \\
\mathbb{E}\big[\hat{Y}^2\big] &= \mathbb{V}\mathrm{ar}(\hat{Y}) + \mathbb{E}[\hat{Y}]^2, \\
&= \mathbb{V}\mathrm{ar}(\hat{\beta}_0 + \hat{\beta}_1 x) + (\beta_0 + \beta_1 x)^2, \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x} - x)^2}{s_x^2} \right) + (\beta_0 + \beta_1 x)^2, \\
\mathbb{E}\big[Y\hat{Y}\big] &= \mathbb{E}\big[Y(\hat{\beta}_0 + \hat{\beta}_1 x)\big], \\
&= \mathbb{E}\big[Y\hat{\beta}_0\big] + x\mathbb{E}\big[Y\hat{\beta}_1\big], \\
&= \mathbb{E}[Y]\mathbb{E}\big[\hat{\beta}_0\big] + x\mathbb{E}[Y]\mathbb{E}\big[\hat{\beta}_1\big], \quad (Y \text{ and } \hat{\beta}_{0,1} \text{ are independent by definition}) \\
&= (\beta_0 + \beta_1 x)^2.
\end{aligned}$$

Put them together, we find

$$\begin{aligned}\mathbb{E}\Big[\big(Y-\hat{Y}\big)^2\Big] &= \sigma^2 + (\beta_0+\beta_1 x)^2 + \sigma^2\left(\frac{1}{n}+\frac{(\bar{x}-x)^2}{s_x^2}\right) + (\beta_0+\beta_1 x)^2 - 2(\beta_0+\beta_1 x)^2,\\ &= \sigma^2\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{s_x^2}\right].\end{aligned}$$

That concludes our calculation. □

In order to obtain the joint distribution of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$, we need the following lemmas.

Lemma 1.2.1. Random variables (P_1, P_2) follow *bivariate normal distribution*. Define (Q_1, Q_2) as linear combinations of constant 1, P_1 and P_2

$$\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{1.23}$$

where $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0$. Then (Q_1, Q_2) also follow bivariate normal distribution.

Proof. We can write Eq. (1.23) in matrix form for simplicity, i,e.,

$$\mathbf{Q} = \mathbf{A}\mathbf{P} + \mathbf{b}.$$

Let $\boldsymbol{\lambda} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$, we have

$$\begin{aligned}\boldsymbol{\lambda}^\mathsf{T}\mathbf{Q} &= \boldsymbol{\lambda}^\mathsf{T}(\mathbf{A}\mathbf{P}+\mathbf{b}),\\ &= (\boldsymbol{\lambda}^\mathsf{T}\mathbf{A})\mathbf{P} + \boldsymbol{\lambda}^\mathsf{T}\mathbf{b}.\end{aligned}$$

When $\det \mathbf{A} \neq 0$, we have $\boldsymbol{\lambda}^\mathsf{T}\mathbf{A} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$. Since $\mathbf{P}$ is bivariate normal, it follows by definition that $(\boldsymbol{\lambda}^\mathsf{T}\mathbf{A})\mathbf{P}$ is normal distributed. Therefore $\boldsymbol{\lambda}^\mathsf{T}\mathbf{Q} = (\boldsymbol{\lambda}^\mathsf{T}\mathbf{A})\mathbf{P} + \boldsymbol{\lambda}^\mathsf{T}\mathbf{b}$ is also normally distributed. Hence, $\mathbf{Q}$ follows a bivariate normal distribution. □

Lemma 1.2.2. Suppose that $\mathbf{P} = \{P_i\}_{i=1}^n$ are i.i.d. and each has distribution $\mathcal{N}(0,1)$. Let $\mathbf{A} \in \mathrm{O}(n)$ be an orthogonal matrix and $\mathbf{Q} = \mathbf{A}\mathbf{P}$, then $\mathbf{Q} = \{Q_i\}_{i=1}^n$ are i.i.d. and each has distribution $\mathcal{N}(0,1)$ as well. additionally, we have $\sum_{i=1}^n P_i^2 = \sum_{i=1}^n Q_i^2$.

Proof. The joint probability density function (PDF) of $\mathbf{P}$ is

$$f_{\mathbf{P}}(\mathbf{p}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{p_i^2}{2}\right) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{\sum_{i=1}^n p_i^2}{2}\right) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{\mathbf{p}^\mathsf{T}\mathbf{p}}{2}\right).$$

Since the variable transformation should conserve the probability in an infinitesimal volume in $\mathbb{R}^n$, we have

$$\begin{aligned} g_{\mathbf{Q}}(\mathbf{q})\mathrm{d}\mathbf{q} &= f_{\mathbf{P}}(\mathbf{p})\mathrm{d}\mathbf{p},\\ g_{\mathbf{Q}}(\mathbf{q}) &= f_{\mathbf{P}}(\mathbf{p})\left|\frac{\partial \mathbf{p}}{\partial \mathbf{q}}\right|,\\ &= f_{\mathbf{P}}(\mathbf{A}^{-1}\mathbf{q})|\det\big(A^{-1}\big)|.\end{aligned}$$

Noting that $\mathbf{A}^{-1} = \mathbf{A}^\mathsf{T}$, the PDF of $\mathbf{Q}$ is

$$\begin{aligned} g_{\mathbf{Q}}(\mathbf{q}) &= f_{\mathbf{P}}(\mathbf{A}^\mathsf{T}\mathbf{q}),\\ &= \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{\mathbf{q}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{q}}{2}\right),\\ &= \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{\mathbf{q}^\mathsf{T}\mathbf{q}}{2}\right).\end{aligned}$$

Therefore, we find that $\mathbf{Q} = \{Q_i\}_{i=1}^n$ are i.i.d. and each has distribution $\mathcal{N}(0,1)$.

Note that the squared sum property follows directly from the fact that $\mathbf{A} \in \mathrm{O}(n)$,

$$\sum_{i=1}^n Q_i^2 = \mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{P}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{P} = \mathbf{P}^\mathsf{T}\mathbf{P} = \sum_{i=1}^n P_i^2.$$

□

Back to our discussion in linear regression, we have the following theorem.

Theorem 1.2.6. Let $\mathbf{A} \in \mathrm{O}(n)$ be an orthogonal matrix and $\mathbf{Z} = \mathbf{A}\mathbf{Y}$, where $\mathbf{Y} = \{Y_i\}_{i=1}^n$ are independent normal distributions with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\mathbf{I}$. Then $\mathbf{Z} = \{Z_i\}_{i=1}^n$ are independent normal distributions with mean $\boldsymbol{\mu}' = \mathbf{A}\boldsymbol{\mu}$ and covariance matrix $\sigma^2\mathbf{I}$.

Proof. Let $\mathbf{R} = \frac{1}{\sigma}(\mathbf{Y} - \boldsymbol{\mu})$, then $\mathbf{R}$ follows i.i.d. standard normal distribution. From Lemma 1.2.2, we know that $\mathbf{S} = \mathbf{A}\mathbf{R}$ follows i.i.d. standard normal distribution. On the other hand, we have

$$\mathbf{Z} = \mathbf{A}\mathbf{Y} = \mathbf{A}(\sigma\mathbf{R} + \boldsymbol{\mu}) = \sigma\mathbf{S} + \mathbf{A}\boldsymbol{\mu}.$$

Thus, it is clear that $\mathbf{Z} = \{Z_i\}_{i=1}^n$ satisfies nulti-normal distributions $\mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \sigma^2\mathbf{I})$. □

Theorem 1.2.7. For a simple linear regression model, we have

1. $(\hat{\beta}_0, \hat{\beta}_1)$ follows bivariate normal distribution

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \sigma^2\begin{pmatrix}\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} & -\frac{\bar{x}}{s_x^2}\\ -\frac{\bar{x}}{s_x^2} & \frac{1}{s_x^2}\end{pmatrix},\right) \tag{1.24}$$

   and the marginal distributions are given in Eq. (1.19) and (1.20).

2. $\hat{\sigma}^2$ is independent of $(\hat{\beta}_0, \hat{\beta}_1)$ and

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2). \tag{1.25}$$

Proof. 1. Following theorem 1.2.4, we only need to show that $(\hat{\beta}_0, \hat{\beta}_1)$ follows bivariate normal distribution to finish the proof[5].

Let's construct the following $\mathbf{A} \in \mathrm{O}(n)$,

$$\mathbf{A} = \begin{pmatrix}\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}}\\ \frac{x_1-\bar{x}}{s_x} & \frac{x_2-\bar{x}}{s_x} & \cdots & \frac{x_n-\bar{x}}{s_x}\\ \vdots & \vdots & \vdots & \vdots\end{pmatrix}, \tag{1.26}$$

and denote the first row as $\mathbf{a}_1$, the second row as $\mathbf{a}_2$[6]. It is simple to verify that $\mathbf{a}_1$ and $\mathbf{a}_2$ are orthogonal and normalized to 1:

$$\begin{aligned}\|\mathbf{a}_1\|_2^2 &= \sum_{i=1}^n \frac{1}{n} = 1,\\ \|\mathbf{a}_2\|_2^2 &= \sum_{i=1}^n \frac{(x_i-\bar{x})^2}{s_x^2} = 1,\\ \mathbf{a}_1^\mathsf{T}\mathbf{a}_2 &= \sum_{i=1}^n \frac{x_i-\bar{x}}{\sqrt{n}s_x} = 0.\end{aligned}$$

[5]Note that marginally normal does not imply jointly normal!

[6]We define $\mathbf{a}_{1,2}$ as column vectors to make the notation clear, that a vector $\mathbf{v}$ is always a column vector.

The exact form of the rest rows are unimportant to our proof, and we simply note that they can be constructed from the Gram–Schmidt process.

Let $\mathbf{Z} = \mathbf{AY}$, from theorem 1.2.6, we know that $\mathbf{Z} = \{Z_i\}_{i=1}^n$ are independent normal distributions with the same variance σ^2. In particular, Z_1 and Z_2 are independent normal distributions with the same variance σ^2. Additionally, we have

$$
\begin{aligned}
Z_1 &= \sum_{i=1}^{n} a_{1i} Y_i = \sqrt{n}\bar{Y} = \sqrt{n}(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}), \\
Z_2 &= \sum_{i=1}^{n} a_{2i} Y_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) Y_i}{s_x} = s_x \hat{\beta}_1.
\end{aligned}
$$

Solving $\hat{\beta}_0$ and $\hat{\beta}_1$ from it, we have

$$
\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & -\frac{\bar{x}}{s_x} \\ 0 & \frac{1}{s_x} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \tag{1.27}
$$

in the matrix from. Clearly

$$
\begin{vmatrix} \frac{1}{\sqrt{n}} & -\frac{\bar{x}}{s_x} \\ 0 & \frac{1}{s_x} \end{vmatrix} = \frac{1}{\sqrt{n} s_x} \neq 0.
$$

From lemma 1.2.1, we conclude that $(\hat{\beta}_0, \hat{\beta}_1)$ follows bivariate normal distribution. Additionally, during the proof, we observe that Z_1, Z_2 have variance σ^2. By the property of normal distributions, we can get the covariance matrix immediately from Eq. (1.27).

2. Recall that our estimation for $\hat{\sigma}^2$ is

$$
\hat{\sigma}^2 = \frac{1}{n} S^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.
$$

The approach is to make connection with the $\mathbf{A}$ matrix we defined in Eq. (1.26). By Eq. (1.27), we

have

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \left(\frac{1}{\sqrt{n}}Z_1 - \frac{\bar{x}}{s_x}Z_2\right) - \frac{1}{s_x}Z_2 x_i\right]^2, \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \frac{1}{\sqrt{n}}Z_1 - \frac{x_i - \bar{x}}{s_x}Z_2\right]^2, \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - a_{1i}Z_1 - a_{2i}Z_2\right]^2, \\
&= \frac{1}{n}\left\|\mathbf{Y} - \mathbf{a}_1 Z_1 - \mathbf{a}_2 Z_2\right\|_2^2, \\
&= \frac{1}{n}\left(\mathbf{Y}^\mathsf{T}\mathbf{Y} + \mathbf{a}_1^\mathsf{T}\mathbf{a}_1 Z_1^2 + \mathbf{a}_2^\mathsf{T}\mathbf{a}_2 Z_2^2 - 2\mathbf{Y}^\mathsf{T}\mathbf{a}_1 Z_1 - 2\mathbf{Y}^\mathsf{T}\mathbf{a}_2 Z_2 + 2\mathbf{a}_1^\mathsf{T}\mathbf{a}_2 Z_1 Z_2\right), \\
&= \frac{1}{n}\left(\mathbf{Y}^\mathsf{T}\mathbf{Y} + Z_1^2 + Z_2^2 - 2(\mathbf{a}_1^\mathsf{T}\mathbf{Y})^\mathsf{T} Z_1 - 2(\mathbf{a}_2^\mathsf{T}\mathbf{Y})^\mathsf{T} Z_2\right), \quad (\mathbf{A} \in \mathrm{O}(n)) \\
&= \frac{1}{n}\left(\mathbf{Y}^\mathsf{T}\mathbf{Y} - Z_1^2 - Z_2^2\right), \quad (\mathbf{Z} = \mathbf{A}\mathbf{Y}) \\
&= \frac{1}{n}\left(\mathbf{Z}^\mathsf{T}\mathbf{Z} - Z_1^2 - Z_2^2\right), \quad \text{(lemma 1.2.2)} \\
&= \frac{1}{n}\sum_{i=3}^{n} Z_i^2,
\end{aligned}
$$

Next, for the purpose of normalization, we calculate the mean of Z_i,

$$
\begin{aligned}
\mathbb{E}[Z_i] &= \mathbb{E}[\mathbf{a}_i^\mathsf{T}\mathbf{Y}], \\
&= \mathbb{E}[\mathbf{a}_i^\mathsf{T}(\beta_0 + \beta_1\mathbf{x} + \boldsymbol{\varepsilon})], \\
&= \mathbb{E}[\mathbf{a}_i^\mathsf{T}(\hat{\beta}_0 + \hat{\beta}_1\mathbf{x})], \\
&= \mathbb{E}[\mathbf{a}_i^\mathsf{T}(\mathbf{a}_1 Z_1 + \mathbf{a}_2 Z_2)], \\
&= 0, \qquad \text{when } i \geq 3.
\end{aligned}
$$

Therefore, we have

$$
\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{i=3}^{n}\left(\frac{Z_i}{\sigma}\right)^2 \sim \chi^2(n-2).
$$

That concludes our proof.

□

Summary

In this section, we introduced two methods to train an simple linear regression model: (1) by minimizing the mean squared error, and (2) by maximizing the likelihood function. We show that their results are the same. Additionally, we derived the joint distribution of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$.

1.2.2 Statistical Tests

1.3 Multiple Linear Regression

1.3.1 Fitting the Model

1.3.2 Statistical Tests

1.4 Multicolinearity

1.5 R-Squared and Adjusted R-Squared

1.6 AIC and BIC

1.7 Other Details