Machine Learning

Linear regression

Souhaib Ben Taieb

March 24, 2020

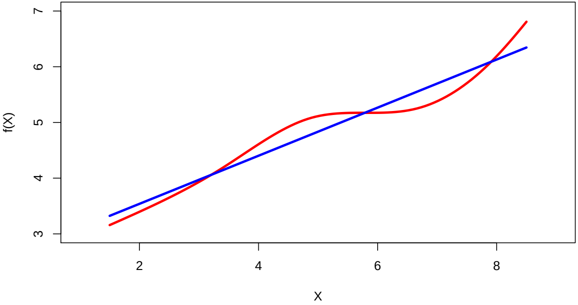
University of Mons

Table of contents

Table of contents

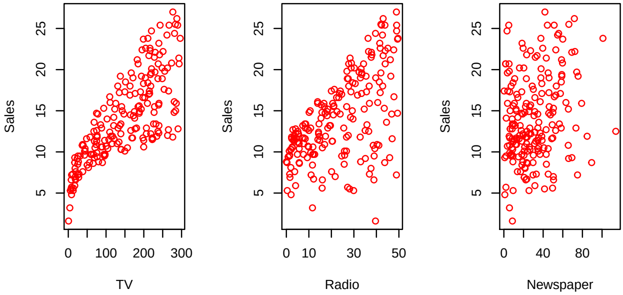
Linear regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.
- True regression functions are never linear!



- although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Advetising data



- Is there a relationship between advertising budget and sales? If so, how strong is it? Is the relationship linear?
- Which media contribute to sales? Is there synergy among the media?
- How accurately can we predict future sales?

Advertising data

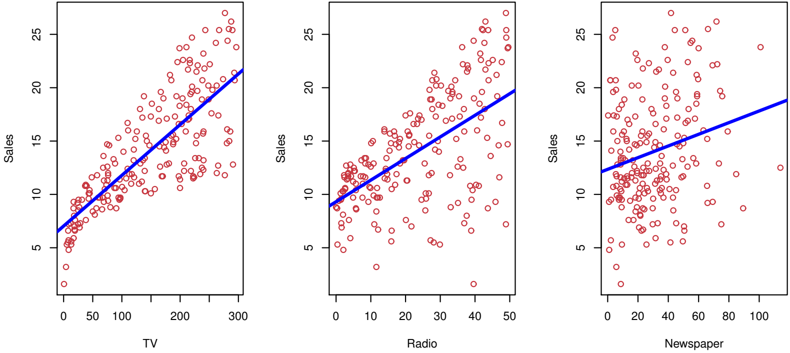


Table of contents

Simple linear regression - Optimal predictions

Let us consider simple linear regression, i.e. linear regression using a single input (or predictor) $x \in \mathbb{R}$. In other words, we consider the hypothesis set $\mathcal{H} = \{h : h(x) = \beta_0 + \beta_1 x; \beta_0, \beta_1 \in \mathbb{R}\}$ and a squared error loss function. What are the **optimal linear predictions**?, i.e. the predictions that minimize the expected out-of-sample squared errors.

In other words, we want to solve the following optimization problem:

$$\underset{h \in \mathcal{H}}{\text{Minimize}}\ E_{\text{out}}(h) \equiv \mathbb{E}_{x,y}[(y - h(x))^2].$$

Since $h(x) = \beta_0 + \beta_1 x$, where β_0 and β_1 completlety characterize h, we can rewrite the problem as

$$\underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\text{Minimize}}\ E_{\text{out}}(\beta_0, \beta_1) \equiv \mathbb{E}_{x,y}[(y - (\beta_0 + \beta_1 x))^2].$$

Simple linear regression - Optimal predictions

$$
\begin{aligned}
E_{\text{out}}(\beta_0, \beta_1) &= \mathbb{E}[(y - (\beta_0 + \beta_1 x))^2] \\
&= \mathbb{E}[y^2] - 2\beta_0 \mathbb{E}[y] - 2\beta_1 \mathbb{E}[xy] + \mathbb{E}[(\beta_0 + \beta_1 x)^2] \\
&= \mathbb{E}[y^2] - 2\beta_0 \mathbb{E}[y] - 2\beta_1 (\text{Cov}(x, y) + \mathbb{E}[x]\mathbb{E}[y]) \\
&+ \mathbb{E}[(\beta_0 + \beta_1 x)^2] \\
&= \mathbb{E}[y^2] - 2\beta_0 \mathbb{E}[y] - 2\beta_1 (\text{Cov}(x, y) + \mathbb{E}[x]\mathbb{E}[y]) \\
&+ \beta_0^2 + \beta_1^2 \mathbb{E}[x^2] + 2\beta_0 \beta_1 \mathbb{E}[x] \\
&= \mathbb{E}[y^2] - 2\beta_0 \mathbb{E}[y] - 2\beta_1 \text{Cov}(x, y) - 2\beta_1 \mathbb{E}[x]\mathbb{E}[y] \\
&+ \beta_0^2 + \beta_1^2 \text{Var}(x) + \beta_1^2 (\mathbb{E}[x])^2 + 2\beta_0 \beta_1 \mathbb{E}[x]
\end{aligned}
$$

where we used the following identities:

$$
\text{Cov}(xy) = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y], \quad \text{Var}(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2.
$$

Simple linear regression - Optimal predictions

We minimize by setting derivatives to zero; we need to take two partial derivatives, which will give us two equations in two unknowns:

$$\frac{\partial E_{\text{out}}(\beta_0, \beta_1)}{\partial \beta_0} = 0 \qquad \iff \qquad \beta_0 = \mathbb{E}[y] - \beta_1 \mathbb{E}[x]$$
$$\frac{\partial E_{\text{out}}(\beta_0, \beta_1)}{\partial \beta_1} = 0 \qquad \qquad \beta_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

- We **did not** assume that the relationship between x and y really is linear.
- We **did not** assume anything about the marginal distributions of x and y, or about their joint distributions.

Table of contents

We saw that the optimal predictions are obtained using $\beta_0 = \mathbb{E}[y] - \beta_1 \mathbb{E}[x]$ and $\beta_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$. However, in practice, we **do not know** $p(x)$, $p(y)$ or $p(x,y)$ which are required to compute β_0 and β_1.

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} p(x,y)$, we could compute $\hat{\beta}_0$ and $\hat{\beta}_1$ by replacing the population quantities with their sample counterparts, which is called the "**plug-in principle**".

Another approach is to directly minimize the in-sample error by solving the following optimization problem:

$$\underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\text{Minimize}}\ E_{\text{in}}(\beta_0, \beta_1) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2,$$

which is also called the *least squares*.

Estimation of the parameters by least squares

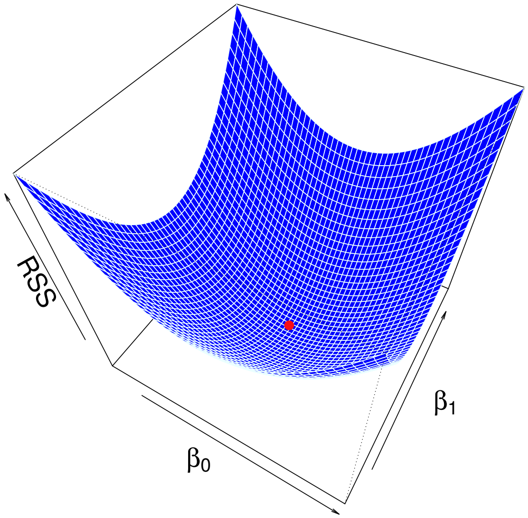
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

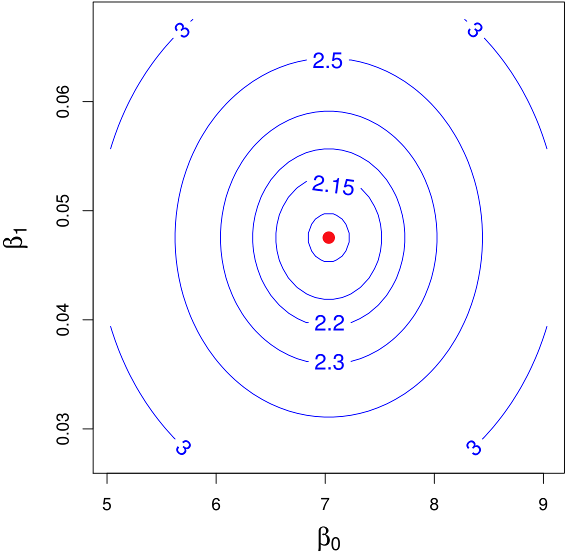
or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Simple linear regression - Geometry of least squares



Simple linear regression - Geometry of least squares



Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

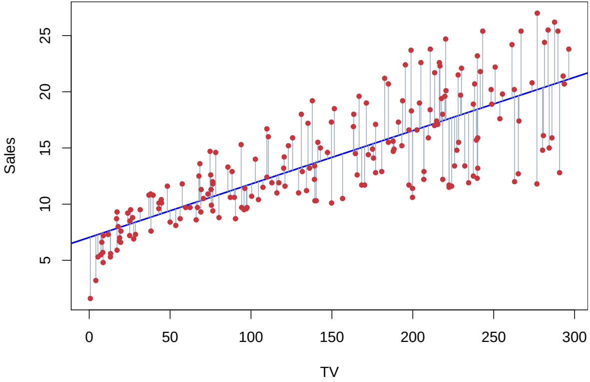
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

Example: advertising data



The least squares fit for the regression of `sales` onto `TV`.
In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Table of contents

Simple linear regression - Least Squares and MLE

In a linear model, if the errors are normally distributed, (ordinary) least squares is equivalent to Maximum Likelihood Estimation (MLE).

Suppose that $z_1, z_2, \ldots, z_n \overset{i.i.d.}{\sim} p(z; \theta)$ where $z_i = (y_i, x_i)$ and p_θ denotes either the pmf or pdf. We will also write $p(z; \theta)$ in place of $p_\theta(z)$.

The **likelihood function** is defined by

$$L(\theta) \equiv L(\theta; z_1, z_2, \ldots, z_n) = \Pi_{i=1}^{n} p_\theta(z_i) = \Pi_{i=1}^{n} p_\theta(y_i, x_i).$$

The **log-likelihood function** is

$$l(\theta) \equiv l(\theta; z_1, z_2, \ldots, z_n) = \log\, L(\theta).$$

The **maximum likelihood estimator**, or mle – denoted by $\hat{\theta}$ – is the value of θ that maximizes $L(\theta)$. Note that $\hat{\theta}$ also maximizes $l(\theta)$. We write

$$\hat{\theta} = \text{argmax}\ L(\theta) = \text{argmax}\ l(\theta).$$

Simple linear regression - Least Squares and MLE

The linear model is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $E[\varepsilon_i|x] = 0$ and $\text{Var}(\varepsilon_i|x) = \sigma^2$.

Let us assume $\varepsilon_i|x_i \sim \mathcal{N}(0, \sigma^2)$, which implies that

$$y_i|x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) = p_{Y|X}(y_i|x_i;\theta).$$

where $\theta = (\beta_0, \beta_1)$.

The **likelihood function** is

$$\begin{aligned} L(\theta) = \Pi_{i=1}^n p(y_i, x_i;\theta) &= \Pi_{i=1}^n p_X(x_i) p_{Y|X}(y_i|x_i;\theta) \\ &= \underbrace{\Pi_{i=1}^n p_X(x_i)}_{\mathcal{L}_1} \underbrace{\Pi_{i=1}^n p_{Y|X}(y_i|x_i;\theta)}_{\mathcal{L}_2} \end{aligned}$$

Simple linear regression - Least Squares and MLE

The term $\mathcal{L}_1$ does not involve the parameters β_0 and β_1. We shall focus on the second term $\mathcal{L}_2$ which is called the **conditional likelihood**, given by

$$\begin{aligned}\mathcal{L}_2 = \mathcal{L}(\beta_0, \beta_1, \sigma) = \Pi_{i=1}^{n} f_{Y|X}(y_i|x_i) \\ \propto \sigma^{-n} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\}\end{aligned}$$

The **conditional log-likelihood** is

$$l(\beta_0, \beta_1, \sigma) \propto -nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

To find the MLE of (β_0, β_1), we **maximize** $l(\beta_0, \beta_1, \sigma)$, which is equivalent to **minimize** $\text{RSS} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$.

Table of contents

Simple linear regression - Bias and variance

We would like to compute the **bias and variance** of $\hat{\beta}_0$ and $\hat{\beta}_1$. To do so, let us assume the data generating process (DGP) is given by

$$y = \beta_0^* + \beta_1^* x + \varepsilon, \tag{1}$$

where ε is a random noise term with $\mathbb{E}[\varepsilon|x] = 0$ and $\mathrm{Var}(\varepsilon|x) = \sigma^2$.

In other words, we are assuming the relationship between x and y is linear[1]. Note that we did not specify the distribution of x (it is arbitrary, possibly even non-random).

For a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where the data points are sampled i.i.d. from (1), the model says that

$$y_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i,$$

where $\mathbb{E}[\varepsilon_i|x] = 0$, $\mathrm{Var}(\varepsilon_i|x) = 0$, and $\mathrm{Cov}(\varepsilon_i \varepsilon_j|x) = 0$ for $i \neq j$.

[1] To be really pedantic, it is an affine rather than a linear function.

Simple linear regression - Bias and variance

Recall that if $y = f(x) + \varepsilon$, the bias and variance of $\hat{f}$ at a new x_0 are given by

$$\text{Bias}(\hat{f}(x_0)) = \mathbb{E}[\hat{f}(x_0)] - f(x_0) \;\&\; \text{Var}(\hat{f}(x_0)) = \mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right]$$

In simple linear regression, we have $f(x) = \beta_0^* + \beta_1^* x$ and $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. Therefore, the bias and variance terms are given by

$$\begin{aligned}\mathbb{E}[\hat{f}(x_0)] - f(x_0) &= \left((\mathbb{E}[\hat{\beta}_0] - \beta_0^*) + (\mathbb{E}[\hat{\beta}_1] - \beta_1^*)x_0\right) \\ &= \text{Bias}(\hat{\beta}_0) + \text{Bias}(\hat{\beta}_1)x_0\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right] &= \mathbb{E}\left[\left((\hat{\beta}_0 - \mathbb{E}[\hat{\beta}_0]) + (\hat{\beta}_1 - \mathbb{E}[\hat{\beta}_1])x_0\right)^2\right] \\ &= \text{Var}(\hat{\beta}_0) + \text{Var}(\hat{\beta}_1)x_0^2 + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)x_0 \\ &= \text{Var}(\hat{\beta}_0) + \text{Var}(\hat{\beta}_1)x_0^2 - 2\bar{x}\text{Var}(\hat{\beta}_1)x_0\end{aligned}$$

Simple linear regression - Bias and variance of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1^* + \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$$

where we used the fact that $y_i = \beta_0^* + \beta_1^* x + \varepsilon_i$ and $\bar{x}\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \bar{x}\varepsilon_i$ with $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i$.

In the following, we are going to assume that the x_i are fixed (non-random).

$$\mathbb{E}[\hat{\beta}_1] = \beta_1^* + \mathbb{E}\left[\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \beta_1^* + \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[\varepsilon_i]}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1^*$$

$$\begin{aligned}\text{Bias}(\hat{\beta}_1) &= \mathbb{E}[\hat{\beta}_1] - \beta_1^* \\ &= 0\end{aligned}$$

Simple linear regression - Bias and variance of $\hat{\beta}_1$

$$
\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \text{Var}\left(\beta_1^* + \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(\varepsilon_i)}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}
$$

Simple linear regression - Bias and variance of $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\bar{y} - \hat{\beta}_1 \bar{x}] \\
&= \beta_0^* + \beta_1^* \bar{x} - \mathbb{E}[\hat{\beta}_1]\bar{x} \\
&= \beta_0^* + \beta_1^* \bar{x} - \beta_1^* \bar{x} \\
&= \beta_0^*
\end{aligned}$$

$$\begin{aligned}
\text{Bias}(\hat{\beta}_0) &= \mathbb{E}[\hat{\beta}_0] - \beta_0^* \\
&= 0
\end{aligned}$$

Simple linear regression - Bias and variance of $\hat{\beta}_0$

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1\bar{x}) \\ &= \mathrm{Var}(\bar{y}) + (\bar{x})^2\mathrm{Var}(\hat{\beta}_1) - 2\mathrm{Cov}(\bar{y}, \hat{\beta}_1\bar{x}) \\ &= \mathrm{Var}(\bar{y}) + (\bar{x})^2\mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\mathrm{Cov}(\bar{y}, \hat{\beta}_1)\end{aligned}$$

where

$$\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n},$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

Simple linear regression - Bias and variance of $\hat{\beta}_0$

$$
\begin{aligned}
\mathrm{Cov}(\bar{y}, \hat{\beta}_1) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \frac{\sum_{j=1}^{n}(x_j - \bar{x})y_j}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\
&= \frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\mathrm{Cov}\left(\sum_{i=1}^{n} y_i, \sum_{j=1}^{n}(x_j - \bar{x})y_j\right) \\
&= \frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})\sum_{j=1}^{n}\mathrm{Cov}(y_i, y_j) \\
&= \frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})\sigma^2 \\
&= 0 \quad \left(\text{since } \sum_{i=1}^{n}(x_j - \bar{x}) = 0\right).
\end{aligned}
$$

$$
\implies \mathrm{Var}(\hat{\beta}_0) = \tfrac{\sigma^2}{n} + \tfrac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
$$

Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$

  where $\sigma^2 = \mathrm{Var}(\epsilon)$

- These standard errors can be used to compute *confidence intervals.* A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Confidence intervals — continued

That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$

Hypothesis testing

- Standard errors can also be used to perform *hypothesis tests* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

Hypothesis testing

- Standard errors can also be used to perform *hypothesis tests* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of

  H_0 : There is no relationship between X and Y

  versus the *alternative hypothesis*

  H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis testing — continued

- To test the null hypothesis, we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the *p-value*.

Results for the advertising data

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error*

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where the *residual sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error*

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where the *residual sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- *R-squared* or fraction of variance explained is

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

  where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

R squared

$$\mathrm{R}^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $\hat{y}_i = y_i \implies \mathrm{R}^2 = 1$
- $\hat{y}_i = \bar{y} \implies \mathrm{R}^2 = 0$

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612

Table of contents

Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- We interpret β_j as the *average* effect on Y of a one unit increase in X_j, *holding all other predictors fixed*. In the advertising example, the model becomes

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated — a *balanced design*:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as *"a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed"*, are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous — when X_j changes, everything else changes.
- *Claims of causality* should be avoided for observational data.

The woes of (interpreting) regression coefficients

"Data Analysis and Regression" Mosteller and Tukey 1977

- a regression coefficient β_j estimates the expected change in Y per unit change in X_j, *with all other predictors held fixed*. But predictors usually change together!

The woes of (interpreting) regression coefficients

"Data Analysis and Regression" Mosteller and Tukey 1977

- a regression coefficient β_j estimates the expected change in Y per unit change in X_j, *with all other predictors held fixed*. But predictors usually change together!
- Example: Y total amount of change in your pocket; $X_1 = \#$ of coins; $X_2 = \#$ of pennies, nickels and dimes. By itself, regression coefficient of Y on X_2 will be > 0. But how about with X_1 in model?
- $Y =$ number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $\hat{Y} = b_0 + .50W - .10H$. How do we interpret $\hat{\beta}_2 < 0$?

Two quotes by famous Statisticians

"Essentially, all models are wrong, but some are useful"

George Box

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"

Fred Mosteller and John Tukey, paraphrasing George Box

Estimation and Prediction for Multiple Regression

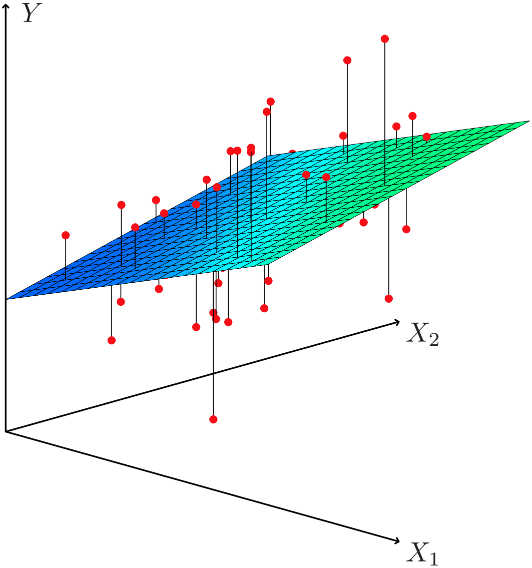
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2.
\end{aligned}$$

This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.


Y
X_2
X_1

Results for advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Table of contents

Is at least one predictor useful?

For the first question, we can use the F-statistic

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Which variables are important?

- The most direct approach is called *all subsets* or *best subsets* regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models!
  Instead we need an automated approach that searches through a subset of them. We will discuss better approaches for model selection with linear models (see Section on Model Selection).

Table of contents

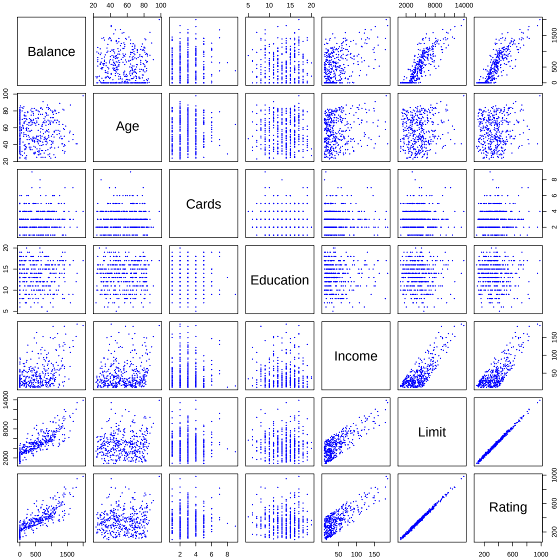
Other Considerations in the Regression Model

Qualitative Predictors

- Some predictors are not *quantitative* but are *qualitative*, taking a discrete set of values.
- These are also called *categorical* predictors or *factor variables*.
- See for example the scatterplot matrix of the credit card data in the next slide.

  In addition to the 7 quantitative variables shown, there are four qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American (AA) or Asian).

Credit Card Data



Qualitative Predictors — continued

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable **(dummy variable)**

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male (baseline).} \end{cases}$$

Intrepretation?

Credit card data — continued

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables. For example, for the `ethnicity` variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

  and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

ethnicity = {Asian, Caucasian, African American}

Qualitative predictors with more than two levels — continued.

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA} \end{cases}$$

(baseline).

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — African American in this example — is known as the *baseline*.

—> K-1 variables for K levels

Results for ethnicity

	Coefficient	Std. Error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

Table of contents

Extensions of the Linear Model

Removing the additive assumption: *interactions* and *nonlinearity*

Interactions:

- In our previous analysis of the `Advertising` data, we assumed that the effect on `sales` of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

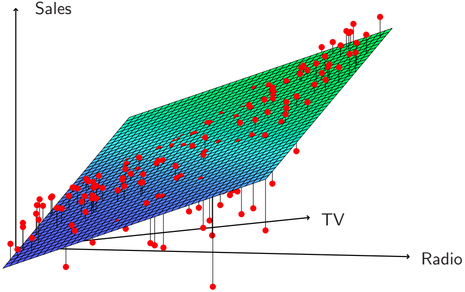
$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

  states that the average effect on `sales` of a one-unit increase in `TV` is always β_1, regardless of the amount spent on `radio`.

Interactions — continued

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of $\$100,000$, spending half on `radio` and half on `TV` may increase `sales` more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

Interaction in the Advertising data?



When levels of either TV or radio are low, then the true sales are lower than predicted by the linear model.
But when advertising is split between the two media, then the model tends to underestimate sales.

Modelling interactions — Advertising data

Model takes the form

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Interpretation

- The results in this table suggests that interactions are important.
- The p-value for the interaction term `TV`$\times$`radio` is extremely low, indicating that there is strong evidence for $H_A : \beta_3 \neq 0$.
- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts `sales` using `TV` and `radio` without an interaction term.

Interpretation — continued

- This means that $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the variability in `sales` that remains after fitting the additive model has been explained by the interaction term.
- The coefficient estimates in the table suggest that an increase in TV advertising of $\$1,000$ is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, `TV` and `radio`) do not.
- The *hierarchy principle*:

  If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

Hierarchy — continued

- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

Interactions between qualitative and quantitative variables

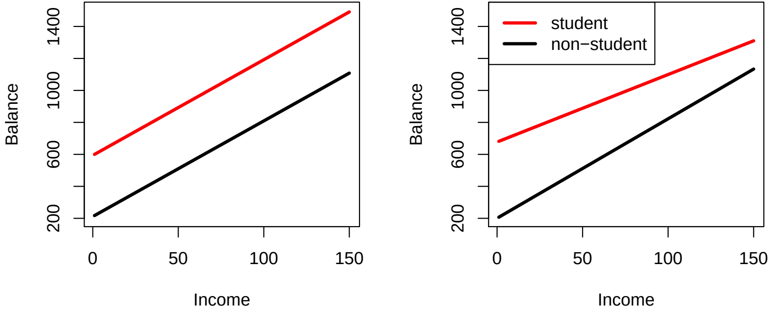
Consider the `Credit` data set, and suppose that we wish to predict `balance` using `income` (quantitative) and `student` (qualitative).

Without an interaction term, the model takes the form

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}
$$

With interactions, it takes the form

$$
\begin{aligned}
\texttt{balance}_i \quad &\approx \quad \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\
&= \quad \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}
\end{aligned}
$$

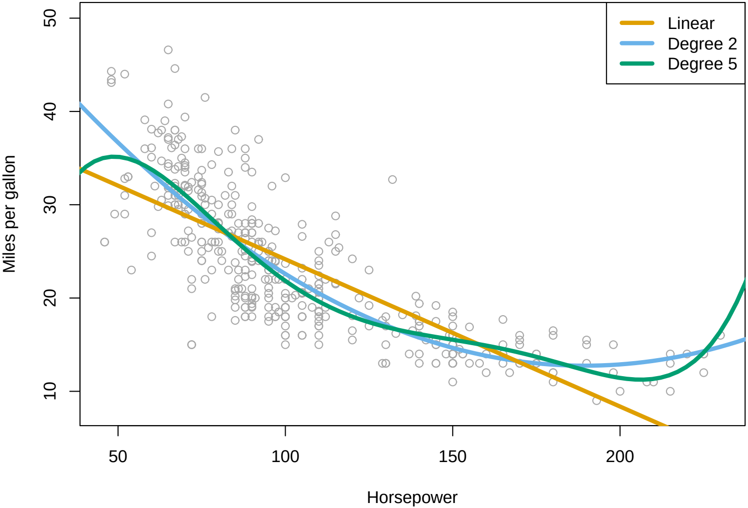


Credit data; Left: no interaction between income and student.
Right: with an interaction term between income and student.

Table of contents

Non-linear effects of predictors

polynomial regression on `Auto` data



The figure suggests that

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

What we did not cover

Outliers
Non-constant variance of error terms
High leverage points
Collinearity

See text Section 3.33

Table of contents

Matrix Notation - Linear Model

In matrix notation, the following linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, i = 1, \ldots, n,$$

where $\mathbb{E}[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$ can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

and $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an n dimensional identity matrix.

Matrix Notation - Ordinary Least Squares

In matrix notation, the residual sum of squares can be written as

$$\text{RSS} = \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta).$$

The ordinary least squares (OLS) solution is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

Note: $(\boldsymbol{X}^T\boldsymbol{X})$ is not always invertible, e.g. in high dimensions ($p > n$) or when the inputs are highly correlated. Another example is the dummy variable trap, where K instead of $K - 1$ dummy variables are used for a categorical varaibles with K levels.

Matrix Notation - Maximum Likelihood Estimation

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n),$$

where $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a n-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

So the likelihood is

$$L = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\right)$$

which is maximized when $(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$ is minimized.

So **MLE** $\equiv$ **OLS**.