Machine Learning I

Linear regression

Souhaib Ben Taieb

March 16, 2022

University of Mons

Table of contents

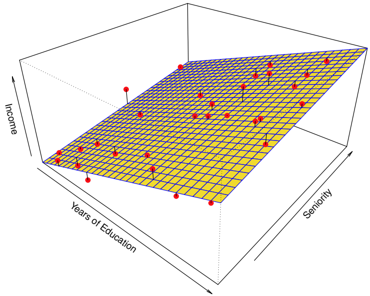
Table of contents

Linear regression

- Hypothesis set with affine (linear) functions. If $x \in \mathbb{R}^p$, we have

$$\mathcal{H}_{\text{lin}} = \{h(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p : \beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}\}$$

- The squared error loss function: $L(y, h(x)) = (y - h(x))^2$.
- Although true functions are *very rarely linear*, linear regression models are useful both conceptually and practically.



Simple linear regression

We will first consider linear regression with a single input $x \in \mathbb{R}$, i.e. $p = 1$, also called *simple* linear regression

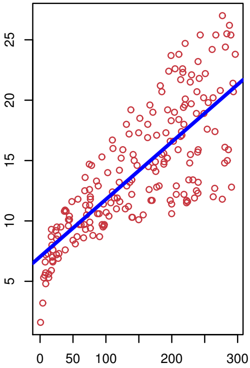


Table of contents

Optimal predictions

What are the **optimal predictions** in simple linear regression? In other words, we want to compute

$$g^* = \underset{h \in \mathcal{H}_{\text{lin}}}{\operatorname{argmin}} \, E_{\text{out}}(h) := \mathbb{E}_{x,y}[(y - h(x))^2],$$

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

Optimal predictions

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

$$E_{\text{out}}(h) = \mathbb{E}_{x,y}[(y - h(x))^2] = \mathbb{E}_{x,y}[(y - (\beta_0 + \beta_1 x))^2] = E_{\text{out}}(\beta_0, \beta_1)$$

The problem

$$g^* = \underset{h \in \mathcal{H}_{\text{lin}}}{\text{argmin}} \; E_{\text{out}}(h)$$

reduces to

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\text{argmin}} \; E_{\text{out}}(\beta_0, \beta_1),$$

where

$$g^*(x) = \beta_0^* + \beta_1^* x.$$

Optimal predictions

To compute (β_0^*, β_1^*), we need to set two partial derivatives to zero, which will give us two equations in two unknowns:

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

Table of contents

Parameter estimation

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, we can compute

$$g = \underset{h \in \mathcal{H}_{\text{lin}}}{\operatorname{argmin}} \; E_{\text{in}}(h) := \frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i))^2,$$

or, equivalently,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \; E_{\text{in}}(\beta_0, \beta_1) := \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

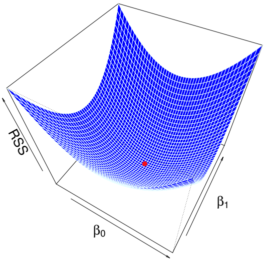
Geometry of least squares

If we let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $e_i = y_i - \hat{y}_i$ represent the ith residual, we define the residual sum of squares (RSS) as

$$\text{RSS} = \sum_{i=1}^{n} e_i^2.$$

Minimizing E_{in} is equivalent to minimize RSS since $E_{\text{in}} = \frac{\text{RSS}}{n}$.

This method is also know as the (ordinary) least squares (OLS).



Parameter estimation

The minimizing values $\hat{\beta}_1$ and $\hat{\beta}_0$ can be shown to be

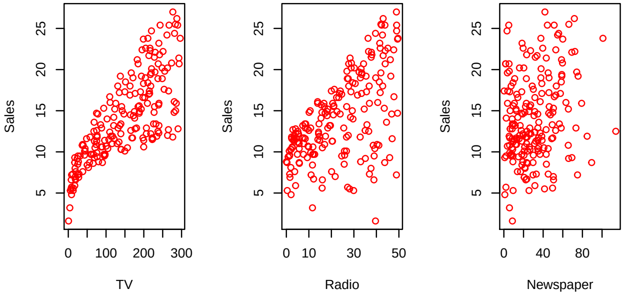
$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

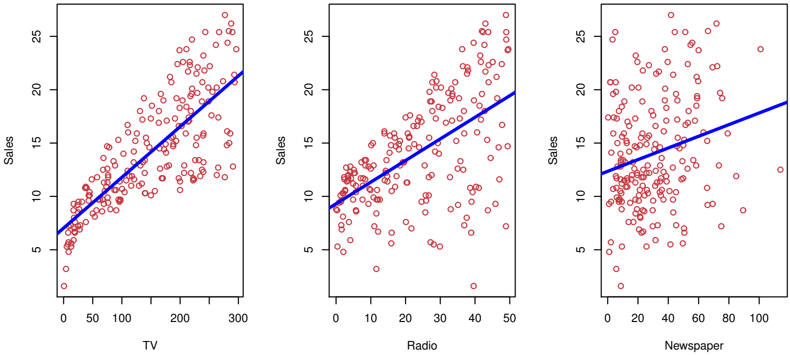
In other words, we obtain

$$g_{\mathcal{D}}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Advetising data



Advertising data



Plug-in principle

We saw that the optimal linear predictions are obtained using

$$\beta_1^* = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)},$$
$$\beta_0^* = \mathbb{E}[y] - \beta_1 \mathbb{E}[x].$$

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} p_{x,y}$, if we replace the population quantities with their sample counterparts, we obtain

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

This is known as the "**plug-in principle**".

Bias and variance in simple linear regression

Let us assume the data generating process 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 $\mathrm{Var}(\varepsilon|x) = \sigma^2$.

Then we can show that

$$\mathbb{E}[\hat{\beta}_1] = \beta_1^* \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

and

$$\mathbb{E}[\hat{\beta}_0] = \beta_0^* \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right],$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Table of contents

Simple linear regression and MLE

Let us assume the following linear data generating process for the data:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where $\beta_0, \beta_1 \in \mathbb{R}$ and $\varepsilon | x \sim \mathcal{N}(0, \sigma^2)$. This implies that

$$y | x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2) = p_{y|x}(y | x; \boldsymbol{\theta}).$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$.

Simple linear regression and MLE

The **(conditional) likelihood function** is given by

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma) &= \mathcal{L}(\beta_0, \beta_1, \sigma; \mathcal{D}) \\
&= p(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p_{y|x}(y_i | x_i; \beta_0, \beta_1, \sigma) \\
&\propto \sigma^{-n} exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}.
\end{aligned}$$

The **(conditional) log-likelihood** is given by

$$\log \mathcal{L}(\beta_0, \beta_1, \sigma) \propto -n log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Simple linear regression and MLE

To find the MLE of β_0 and β_1, we **maximize** the conditional log-likelihood

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

which is equivalent to **minimize**

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

In other words, if we assume a linear model with a normally distributed error term, **(ordinary) least squares is equivalent to MLE**.

Table of contents

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

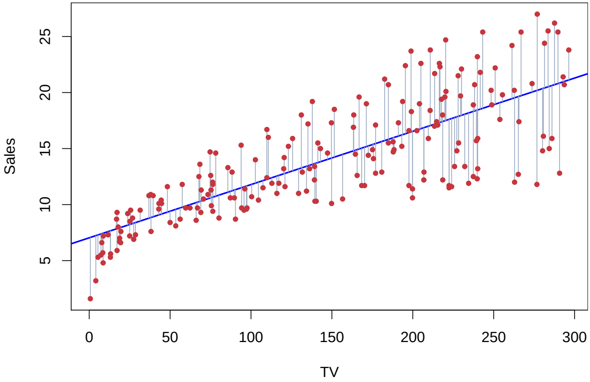
$$H_A : \beta_1 \neq 0,$$

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*.

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.

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

Residual standard error is

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

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

R-squared or the fraction of variance explained is

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

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

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

For the advertising data, RSE is 3.26 and R^2 is 0.612.

Table of contents

Multiple linear regression

In multiple linear regression, we consider a multivariate input $x \in \mathbb{R}^p$ where $p > 1$. In the advertising example, we would consider

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

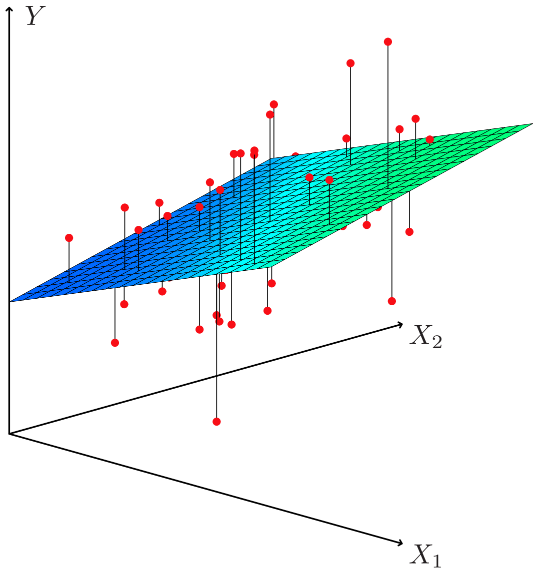

Y
X_2
X_1

Table of contents

Parameter estimation - Matrix notation

A dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^p$ can be represented, in matrix notation, as

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \quad \text{with} \quad \hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T.$$

Parameter estimation - Matrix notation

The residual sum of squares (RSS) can be written as

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

Assuming $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, we have

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

Note: $(\boldsymbol{X}^T\boldsymbol{X})$ is not always invertible, e.g. in high dimensions ($p > n$) or when some input variables are highly correlated.

Table of contents

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.

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

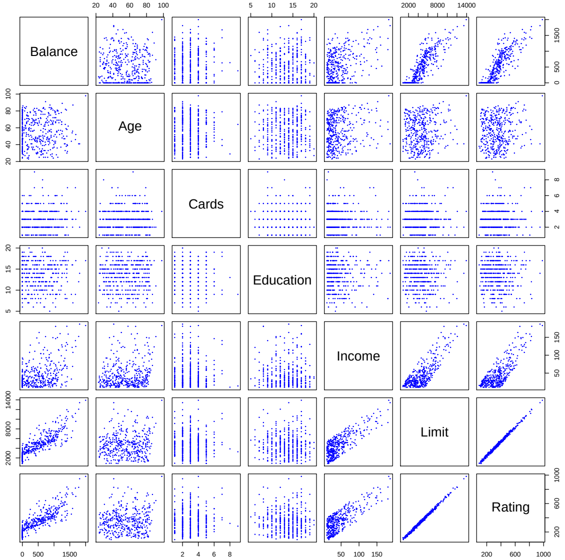
Other Considerations in the Regression Model

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

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

- 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

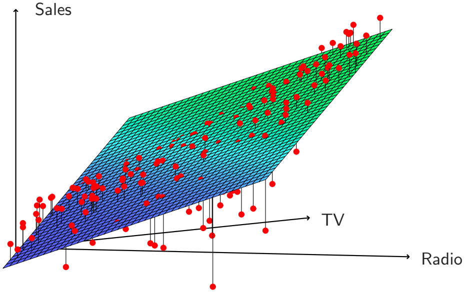
$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

  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`$\times$`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.

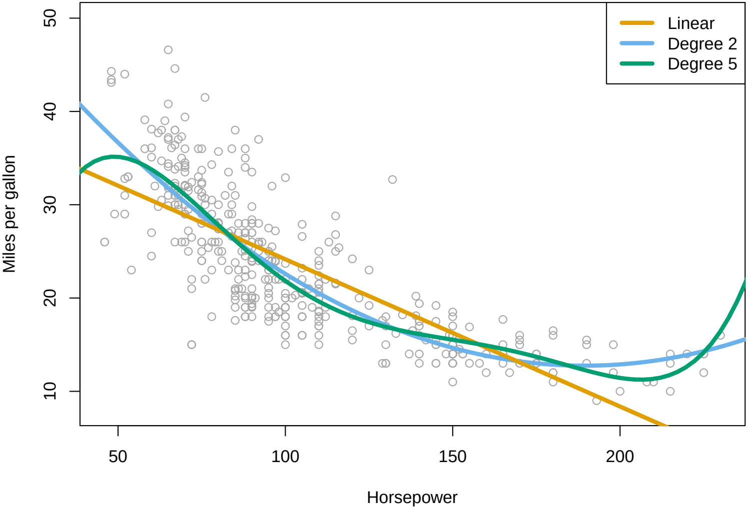
Interpretation — continued

- 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.

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

Topics not covered

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

Table of contents

In-sample and out-of-sample errors

Let us compare the **expected** in-sample and out-of-sample **MSE** in a specific scenario. We assume that the training data is given by

$$\{(x_i, y_i)\}_{i=1}^n \text{ with } y_i = f(x_i) + \varepsilon_i,$$

and the test data is given by

$$\{(x_i, y_i^{'})\}_{i=1}^n \text{ with } y_i^{'} = f(x_i) + \varepsilon_i^{'},$$

where x_i are fixed (not random) and ε_i and $\varepsilon_i^{'}$ are independent but identically distributed random noise variables.

In other words, the training and test data share the **same** input variables x_i but have **different** random noise terms. This scenario is a particular case (simpler to analyze) of the more general scenario where the x_i in the training and test data can be different.

In-sample and out-of-sample errors

We compute $\hat{y}_i = g(x_i)$ using the training data $\{(x_i, y_i)\}_{i=1}^n$. We want to compare

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2\right] \text{ and } \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{y}_i]^2\right].$$

Note that

- y_i and $\hat{y}_i$ are dependent since $\hat{y}_i$ depends on $\{(x_i, y_i)\}_{i=1}^n$, and hence on y_i too.
- y_i' and $\hat{y}_i$ are independent since $\hat{y}_i$ depends on ε_i (through y_i) which is independent of ε_i'.

In-sample and out-of-sample errors

$$\mathbb{E}\left[(y_i - \hat{y}_i)^2\right] = \mathrm{Var}(y_i - \hat{y}_i) + (\mathbb{E}[y_i - \hat{y}_i])^2$$
$$= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{y}_i) - 2\mathrm{Cov}(y_i, \hat{y}_i) + (\mathbb{E}[y_i] - \mathbb{E}[\hat{y}_i])^2$$

$$\mathbb{E}\left[(y_i^{'} - \hat{y}_i)^2\right] = \mathrm{Var}(y_i^{'} - \hat{y}_i) + (\mathbb{E}[y_i^{'} - \hat{y}_i])^2$$
$$= \mathrm{Var}(y_i^{'}) + \mathrm{Var}(\hat{y}_i) - 2\mathrm{Cov}(y_i^{'}, \hat{y}_i) + (\mathbb{E}[y_i^{'}] - \mathbb{E}[\hat{y}_i])^2$$
$$= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{y}_i) + (\mathbb{E}[y_i] - \mathbb{E}[\hat{y}_i])^2$$
$$= \mathbb{E}\left[(y_i - \hat{y}_i)^2\right] + 2\mathrm{Cov}(y_i, \hat{y}_i)$$

since

- y_i is independent of $y_i^{'}$ but has the same distribution: $\mathbb{E}[y_i] = \mathbb{E}[y_i^{'}]$ and $\mathrm{Var}(y_i) = \mathrm{Var}(y_i^{'})$.
- $\mathrm{Cov}(y_i^{'}, \hat{y}_i) = 0$.

In-sample and out-of-sample errors

In summary, we have

$$\mathbb{E}\left[(y_i^{'} - \hat{y}_i)^2\right] = \mathbb{E}\left[(y_i - \hat{y}_i)^2\right] + 2\text{Cov}(y_i, \hat{y}_i),$$

which implies

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i^{'} - \hat{y}_i]^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2\right] + \frac{2}{n}\sum_{i=1}^{n}\text{Cov}(y_i, \hat{y}_i)$$

The expected out-of-sample error can be approximated as

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i^{'} - \hat{y}_i]^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2 + \frac{2}{n}\sum_{i=1}^{n}\text{Cov}(y_i, \hat{y}_i)$$

The last term in the RHS. of the previous expression is called the **optimism**, which is the amount by which the training error systematically under-estimates the expected test error.

Optimism in linear models

If we assume the data generating process is linear and if we use the least square estimator, we can show that

$$\frac{2}{n}\sum_{i=1}^{n}\text{Cov}(y_i, \hat{y}_i) = \frac{2}{n}\sigma^2(p+1),$$

In other words, we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{y}_i]^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2 + \frac{2}{n}\sigma^2(p+1)$$

Notice that the optimism:

- Grows with σ^2
- Shrinks with n
- Grows with p

Table of contents

How to estimate the out-of-sample error?

$$E_{\text{out}}(h) = E_{\text{in}}(h) + \underbrace{[E_{\text{out}}(h) - E_{\text{in}}(h)]}_{\textbf{overfit penalty}}, \quad h \in \mathcal{H}.$$

1. **Directly estimate it** using a large designated test set.
2. **Directly estimate it** using resampling methods.
3. **Estimate the overfiit penalty/optimism** and **add it to the in-sample (training) error**.

Leave-one-out cross-validation with linear models

Let $\hat{y}_{[i]}$ be the predicted value obtained when the model is estimated with the ith observation deleted. If $e_{[i]} = y_i - \hat{y}_{[i]}$, then the leave-one-out cross-validaiton error is given by

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n} e_{[i]}^2,$$

It turns out that for linear models, we do not actually have to estimate the model n times, once for each omitted case.

Let $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$ with $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. If the diagonal values of $\boldsymbol{H}$ are denoted by $h_1, \dots, h_n$, then we have

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

where $e_i = y_i - \hat{y}_i$.

Training error adjustment

- These techniques adjust the training error for the "**model size**", and can be used to select among a set of models with **different numbers of variables**.
- One advantage of resampling methods compared to these methods is the fact that they can be used in a wider range of model selection tasks, even in cases where it is **hard to pinpoint the "model size"**.

The problem with Residual Sum of Squares and R^2

Recall that the **Residual Sum of Squares** (or RSS) is given by

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

Minimizing RSS will always choose the model with the most predictors.

The problem with Residual Sum of Squares and R^2

Recall that the **Residual Sum of Squares** (or RSS) is given by

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Minimizing RSS will always choose the model with the most predictors.

Recall that the R^2 **statistic** is given by

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

The R^2 gives the proportion of variance explained, and is independent of the scale of y. However ...

- R^2 does not allow for "degrees of freedom".
- Adding *any* variable tends to increase the value of R^2, even if that variable is irrelevant.

Estimated residual variance and adjusted R^2

Insead of minimizing RSS, we can minimize the **estimated residual variance**, given by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1},$$

where $p =$ no. predictors.

Minimizing $\hat{\sigma}^2$ works quite well for choosing predictors (but better methods to follow).

Also, instead of R^2, we can use the **adjusted** R^2, defined by

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)},$$

which pays a price for the inclusion of unnecessary variables.

Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.

Estimated residual variance and adjusted R^2

Minimizing $\hat{\sigma}^2$, what does that translate to? We have

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p-1} = \mathrm{MSE}\frac{n}{n-p-1} = \mathrm{MSE}\frac{1}{1-(p+1)/n}.$$

Using the binomial theorem which gives $(1-x)^{-1} = 1+x+x^2+\ldots$, and truncating the series at first order[1], we obtain

$$\hat{\sigma}^2 \approx \mathrm{MSE}\left(1+\frac{p+1}{n}\right) = \mathrm{MSE} + \mathrm{MSE}\frac{p+1}{n}.$$

Even for the right model (where MSE is a consistent estimator of σ^2), the penalty is half as big as what it should be, i.e.

$$\mathrm{MSE} + 2 \times \sigma^2\frac{(p+1)}{n}.$$

$\implies$ $\bar{R}^2$ is better than R^2 but it is still not going to work very well.

[1] For a fixed p, the approximation becomes exact as $n \to \infty$.

Mallow's C_p

The Mallows C_p statistic is given by

$$C_p = \frac{1}{n}(RSS + 2(p+1)\hat{\sigma}^2),$$

where p is the number of predictors in the model.

It essentially substitutes an estimator of σ^2 in the expression of the optimism for linear models. C_p penalizes more heavily than $\bar{R}^2$.

Akaike's Information Criterion

$$\text{AIC} = -2\log(\mathcal{L}) + 2(p+1)$$

where $\mathcal{L}$ is the likelihood and p is the number of predictors.

- AIC is defined for a large class of models fit by **maximum likelihood**. It is also called a **penalized likelihood** approach.
- In the case of the **linear model with Gaussian errors**, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.
- AIC is **asymptotically** equivalent to leave-one-out cross-validation.
- *Minimizing* the AIC gives the best model for **prediction** (not inference).

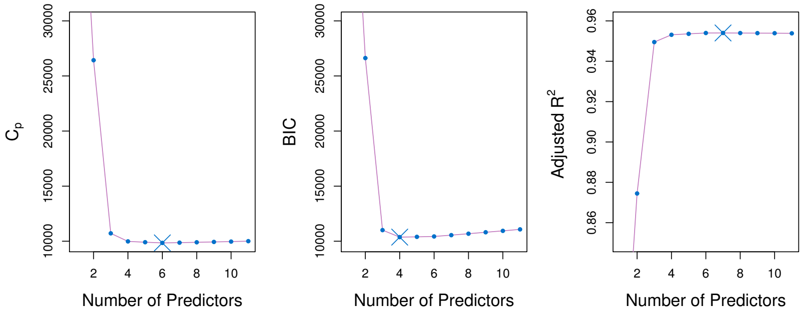
Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(\mathcal{L}) + (p+1)\log(n)$$

where $\mathcal{L}$ is the likelihood and p is the number of predictors.

- BIC penalizes more **heavily** than AIC
- Since $log(n) > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of **smaller models** than C_p/AIC.
- Also called SBIC and SC.
- BIC is **asymptotically** equivalent to leave-v-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

Credit data example



Credit data example

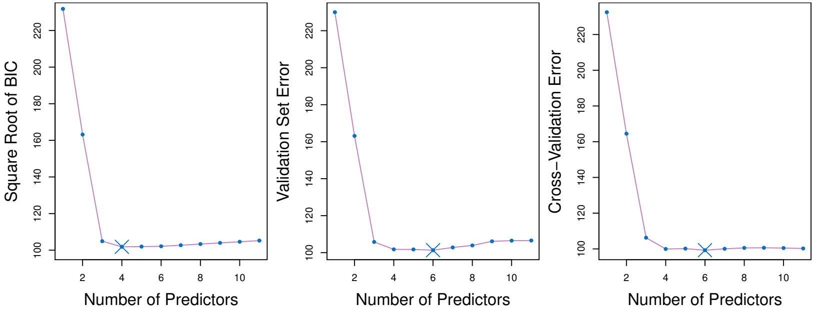


Table of contents

Variable selection

- When performing model selection, in addition to the selection of the best hyper-parameters, we often need to select the **best subset of input variables**.
- In fact, by removing irrelevant variables, we can obtain a model that provide **better predictions** and is more **easily interpreted**.
- If there are a limited number of predictors, we can study all possible models. Otherwise we need a **search strategy** to explore some potential models.
- Although we will present selection strategies for least squares regression, the same ideas apply to **other types of models**.
- The same problem arises for **hyperparameter optimization**. There are multiple search strategies: grid search, random search, evolutionary optimization, etc.

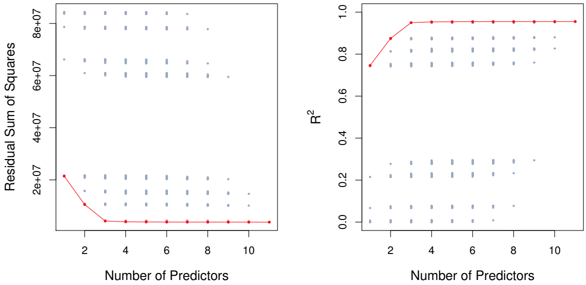
Subset Selection

Best subset and stepwise model selection procedures

Best Subset Selection

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Example- Credit data set



For each possible model containing a subset of the ten predictors in the `Credit` *data set, the RSS and R^2 are displayed. The red frontier tracks the* best *model for a given number of predictors, according to RSS and R^2. Though the data set contains only ten predictors, the x-axis ranges from 1 to 11, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables*

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p. *Why not?*
- Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to *overfitting* and high variance of the coefficient estimates.
- For both of these reasons, *stepwise* methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest *additional* improvement to the fit is added to the model.

In Detail

Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p - 1$:
   2.1 Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   2.2 Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

More on Forward Stepwise Selection

- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors. *Why not? Give an example.*

Credit data example

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income student, limit	rating, income, student, limit

The first four selected models for best subset selection and forward stepwise selection on the Credit data set. The first three models are identical but the fourth models differ.

Backward Stepwise Selection

- Like forward stepwise selection, *backward stepwise selection* provides an efficient alternative to best subset selection.
- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Backward Stepwise Selection: details

Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   2.2 Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

More on Backward Stepwise Selection

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection
- Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the *best* model containing a subset of the p predictors.
- Backward selection requires that the *number of samples n is larger than the number of variables p* (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.