Linear Regression 2

Linear algebra revisited 1

The identity matrix

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Linear algebra revisited 1

The identity matrix

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

An inverse of a $n \times n$ **square matrix** X is a matrix X^{-1} such that

$$XX^{-1} = X^{-1}X = I_n.$$

Linear algebra revisited 2

Two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are perpendicular if

$$\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{u}^T \boldsymbol{v} = 0$$

Linear Regression

- Quantitative response Y.

- Predictor variable $X_1, X_2, \ldots, X_p$.

Goal: Study a linear relationship between X_i's and Y:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon.$$

Linear Regression

- Quantitative response Y.

- Predictor variable $X_1, X_2, \ldots, X_p$.

Goal: Study a linear relationship between X_i's and Y:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon.$$

Example: We study the effects of TV, radio and newspaper advertising budgets on the sales of a product.

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

Data: $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
As in the simple case, we find the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ which give the prediction

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip},$$

Data: $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
As in the simple case, we find the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ which give the prediction

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip},$$

and we want to minimize the RSS

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

$$
\begin{aligned}
\hat{y}_1 &= \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12} + \ldots + \hat{\beta}_p x_{1p} \\
\hat{y}_2 &= \hat{\beta}_0 + \hat{\beta}_1 x_{21} + \hat{\beta}_2 x_{22} + \ldots + \hat{\beta}_p x_{2p} \\
\vdots &= \qquad\qquad\qquad \vdots \\
\hat{y}_n &= \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \hat{\beta}_2 x_{n2} + \ldots + \hat{\beta}_p x_{np}.
\end{aligned}
$$

$$
\begin{aligned}
\hat{y}_1 &= \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12} + \ldots + \hat{\beta}_p x_{1p} \\
\hat{y}_2 &= \hat{\beta}_0 + \hat{\beta}_1 x_{21} + \hat{\beta}_2 x_{22} + \ldots + \hat{\beta}_p x_{2p} \\
\vdots &= \qquad\qquad \vdots \\
\hat{y}_n &= \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \hat{\beta}_2 x_{n2} + \ldots + \hat{\beta}_p x_{np}.
\end{aligned}
$$

$$
\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}
$$

Equations in a matrix form

Let

$$\widehat{\boldsymbol{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T.$$

. Then the linear equations can be written as

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

$$\widehat{\boldsymbol{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{X}_1 + \hat{\beta}_2 \boldsymbol{X}_2 + \ldots + \hat{\beta}_p \boldsymbol{X}_p$$

$$\widehat{\boldsymbol{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{X}_1 + \hat{\beta}_2 \boldsymbol{X}_2 + \ldots + \hat{\beta}_p \boldsymbol{X}_p$$

Find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is perpendicular to $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.
In other words,

$$\boldsymbol{X}_i \cdot \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{X}_i^T \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = 0 \qquad i = 0, 1, \ldots, p.$$

Find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is perpendicular to $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.
In other words,

$$\boldsymbol{X}_i \cdot \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{X}_i^T \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = 0 \qquad i = 0, 1, \ldots, p.$$

$$\begin{pmatrix} \longleftarrow & \boldsymbol{X}_0 & \longrightarrow \\ \longleftarrow & \boldsymbol{X}_1 & \longrightarrow \\ & \vdots & \\ \longleftarrow & \boldsymbol{X}_p & \longrightarrow \end{pmatrix} \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\boldsymbol{X}^T \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{0}.$$

Find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is perpendicular to $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.
In other words,

$$\boldsymbol{X}_i \cdot \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{X}_i^T \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = 0 \qquad i = 0, 1, \ldots, p.$$

$$\begin{pmatrix} \longleftarrow & \boldsymbol{X}_0 & \longrightarrow \\ \longleftarrow & \boldsymbol{X}_1 & \longrightarrow \\ & \vdots & \\ \longleftarrow & \boldsymbol{X}_p & \longrightarrow \end{pmatrix} \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\boldsymbol{X}^T \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{0}.$$

OLS estimator $\hat{\beta}$

$$\boldsymbol{X}^T \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \right) = \mathbf{0}$$

Variance-covariance of the estimators

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

Since σ^2 is unknown, we instead use its estimator

$$\text{RSE} = \sqrt{\frac{RSS}{n-p-1}}.$$

What we will use instead of $\text{Cov}(\hat{\boldsymbol{\beta}})$ is

$$C = \text{RSE}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Variance-covariance of the estimators

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

Since σ^2 is unknown, we instead use its estimator

$$\text{RSE} = \sqrt{\frac{RSS}{n-p-1}}.$$

Variance-covariance of the estimators

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

Since σ^2 is unknown, we instead use its estimator

$$\text{RSE} = \sqrt{\frac{RSS}{n-p-1}}.$$

What we will use instead of $\text{Cov}(\hat{\boldsymbol{\beta}})$ is

$$C = \text{RSE}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$C \approx \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \mathrm{Var}(\beta_0) & \mathrm{Cov}(\beta_0, \beta_1) & \dots & \mathrm{Cov}(\beta_0, \beta_p) \\ \mathrm{Cov}(\beta_1, \beta_0) & \mathrm{Var}(\beta_1) & \dots & \mathrm{Cov}(\beta_1, \beta_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\beta_p, \beta_0) & \mathrm{Cov}(\beta_p, \beta_1) & \dots & \mathrm{Cov}(\beta_p, \beta_p) \end{pmatrix}$$

Example

In the following regression:

$$\widehat{sales} = \hat{\beta}_0 + \hat{\beta}_1 \times TV + \hat{\beta}_2 \times radio + \hat{\beta}_3 \times newspaper,$$

We have $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (2.939, 0.046, 0.189, -0.001)$

Example

In the following regression:

$$\widehat{sales} = \hat{\beta}_0 + \hat{\beta}_1 \times TV + \hat{\beta}_2 \times radio + \hat{\beta}_3 \times newspaper,$$

We have $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (2.939, 0.046, 0.189, -0.001)$

$\text{RSE} = \sqrt{\text{RSS}/(n-3-1)} = 1.69$ and

$$C = \begin{pmatrix} 9.7 \times 10^{-2} & -2.7 \times 10^{-4} & -1.1 \times 10^{-3} & -6.0 \times 10^{-4} \\ -2.7 \times 10^{-4} & 1.9 \times 10^{-6} & -4.5 \times 10^{-7} & -3.3 \times 10^{-7} \\ -1.1 \times 10^{-3} & -4.5 \times 10^{-7} & 7.4 \times 10^{-5} & -1.8 \times 10^{-5} \\ -5.9 \times 10^{-4} & -3.3 \times 10^{-7} & -1.8 \times 10^{-5} & 3.4 \times 10^{-5} \end{pmatrix}$$

$\text{SE}(\hat{\beta}_3) = \sqrt{3.4 \times 10^{-5}} = 0.0059.$

Important questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all predictors help explaining Y, or only a subset of them?
3. How well does model fit the data?

Relationship between the response and the predictors

We use a hypothesis test:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_a : \text{at least one of } \beta_j\text{'s is non-zero.}$$

Relationship between the response and the predictors

We use a hypothesis test:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_a : \text{at least one of } \beta_j\text{'s is non-zero.}$$

The decision will be made after looking at the $F-$statistic:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}.$$

Recall that TSS $= \sum_{i=1}^{n}(y_i - \bar{y})^2$ and RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

How should we look at F– statistic?

One can show that

$$\mathbb{E}[\mathsf{RSS}/(n-p-1)] = \sigma^2$$

and provided that H_0 is true, we also have

$$\mathbb{E}[(\mathsf{TSS}-\mathsf{RSS})/p] = \sigma^2.$$

How should we look at F– statistic?

One can show that

$$\mathbb{E}[\text{RSS}/(n-p-1)] = \sigma^2$$

and provided that H_0 is true, we also have

$$\mathbb{E}[(\text{TSS} - \text{RSS})/p] = \sigma^2.$$

- If H_0 is true, then we expect $F-$statistic to be **very close to 1**.

- If H_a is true, then $\mathbb{E}[(\text{TSS} - \text{RSS})/p]$ and so we expect F to be **greater than 1**.

Sales data

$$\widehat{sales} = \hat{\beta}_0 + \hat{\beta}_1 \times TV + \hat{\beta}_2 \times radio + \hat{\beta}_3 \times newspaper$$

- The F-value is 570 with its corresponding p-value $= 1.58 \times 10^{-96}$.
- We are certain that **at least** one of the advertising media must be related to the sales.

Relationship between the response and a subset of the predictors

Suppose we want to make the same test for **a subset** of q predictors:

$$H_0 : \beta_{i+1} = \beta_{i+2} = \ldots = \beta_{i+q} = 0$$

H_a : at least one of these β_j's is non-zero.

Relationship between the response and a subset of the predictors

Suppose we want to make the same test for **a subset** of q predictors:

$$H_0 : \beta_{i+1} = \beta_{i+2} = \ldots = \beta_{i+q} = 0$$

$$H_a : \text{at least one of these } \beta_j\text{'s is non-zero.}$$

The decision will be made after looking at the $F-$statistic:

$$F = \frac{(\text{RSS}_{-q} - \text{RSS})/q}{\text{RSS}/(n-p-1)},$$

where RSS_{-q} is the residual sum of squares of the model **without those q predictors**.

Relationship between the response and a single predictor

The hypothesis test is

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

Relationship between the response and a single predictor

The hypothesis test is

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

The decision will be made after looking at the $t-$statistic:

$$t = \frac{\hat{\beta}_j - 0}{\text{SE}(\hat{\beta}_j)}.$$

Here, $\text{SE}(\hat{\beta}_j)$ is the square root of entry (j, j) of C, which is an estimate of the covariance matrix of the coefficients.

Example

	Coefficient	SE	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

For example, t-statistic of $\hat{\beta}_3$ (newspaper) is

$$t = \frac{-0.0001}{0.0059} = -0.18$$

Example

However, newspaper strongly affects sales in the simple linear regression.

	Coefficient	SE	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.071	3.30	< 0.0001

Example

However, newspaper strongly affects sales in the simple linear regression.

	Coefficient	SE	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.071	3.30	< 0.0001

This is because of the correlation between newspaper and radio

	TV	radio	newspaper	sales
TV	1.000	0.055	0.057	0.78
radio		1.000	0.35	0.58
newspaper			1.000	0.23
sales				1.000

F-statistic vs t-statistic

Why do we prefer F-statistic over t-statistic when testing $\beta_0 = \beta_1 = \ldots, \beta_p = 0$?

- Calculating F is easier than t, especially for a high p.

- For large p, even $\beta_0 = \beta_1 = \ldots, \beta_p = 0$ is true, there is a small chance that the p-value of some β_j is low enough that we reject $\beta_j = 0$. The F-statistic does not suffer from this issue since it is calculated only once.

Variable selection

- Forward selection:
    - Start with 0 variable. In each step: add a variable that results in the lowest RSS.
    - Stop when RSS barely improves by adding any of the remaining variables.
    - For example, if adding any of the remaining variables reduces the RSS by less that 0.0001, then we will stop here.

Variable selection

- Forward selection:
  - Start with 0 variable. In each step: add a variable that results in the lowest RSS.
  - Stop when RSS barely improves by adding any of the remaining variables.
  - For example, if adding any of the remaining variables reduces the RSS by less that 0.0001, then we will stop here.

- Backward selection:
  - Start with all variables. In each step: remove a variable with the largest p-value.
  - Stop when all p-values are below some threshold e.g. 0.001.

Model evaluation

- Residual standard error (RSE):

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n - p - 1}}$$

- R^2 measures the variance of Y that is explained by the model:

$$R^2 = \left[\text{Cor}(Y, \widehat{Y})\right]^2$$

Example

Predictos	RSE	R2
TV	3.26	0.612
TV + radio	1.68	0.897
TV + radio + newspaper	1.69	0.897

In both metrics, we can conclude that

- Adding `radio` helps significantly improve the model.
- There is no point in adding `newspaper` to the model.