Linear Regression 2

229351

Regression with Multiple Predictors

- Response Y.

- Predictors $X_1, X_2, \dots, X_n$.

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

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

where $\epsilon \sim N(0, \sigma^2)$, σ is unknown

Regression with Multiple Predictors

- Response Y.

- Predictors $X_1, X_2, \ldots, X_n$.

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

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

where $\epsilon \sim N(0, \sigma^2)$, σ is unknown

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

Data: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, where

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \qquad i = 1, \ldots, n$$

Example:

	TV **($\boldsymbol{X}_1$)**	**Radio** **($\boldsymbol{X}_2$)**	**Newspaper** **($\boldsymbol{X}_3$)**	**Sales** **($\boldsymbol{Y}$)**
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
⋮	⋮	⋮	⋮	⋮
200	232.1	8.6	8.7	13.4

As in the simple case, we find the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ which give the prediction

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

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

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

Equations in a matrix form

Let

$$
\begin{aligned}
\boldsymbol{Y} &= (y_1, y_2, \ldots, y_n)^T \\
\widehat{\boldsymbol{Y}} &= (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T \\
\boldsymbol{X} &= \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \\
\hat{\boldsymbol{\beta}} &= (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T.
\end{aligned}
$$

. Then the linear equations can be written as

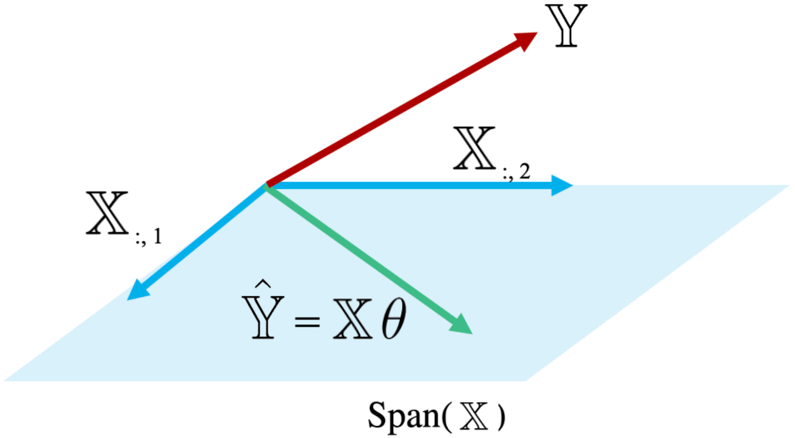
$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

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

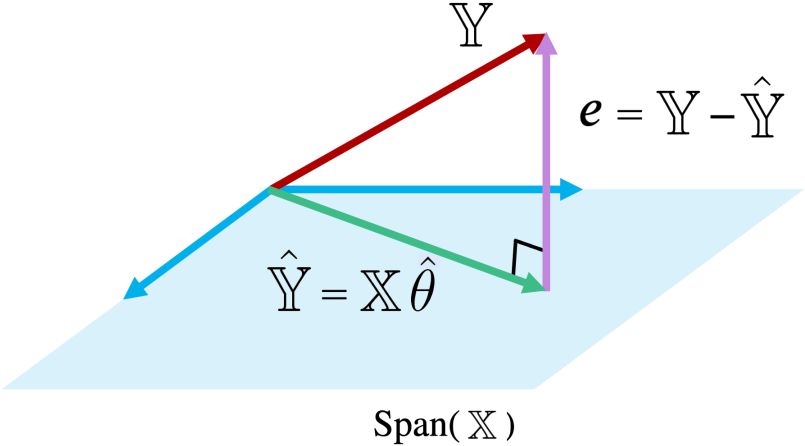
$$\widehat{\boldsymbol{Y}} = \mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_1\hat{\beta}_1 + \boldsymbol{X}_2\hat{\beta}_2 + \ldots + \boldsymbol{X}_p\hat{\beta}_p$$

In other words, $\widehat{Y}$ **must lie in the space spanned by** $\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p$

True Response vs Prediction



Finding the Best Prediction



Find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \perp \mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.

Find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \perp \mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.

In other words, find $\hat{\boldsymbol{\beta}}$ such that

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

Find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \perp \mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.

In other words, find $\hat{\boldsymbol{\beta}}$ such that

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

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

Find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \perp \mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.

In other words, find $\hat{\boldsymbol{\beta}}$ such that

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

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

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

Now we can solve for $\hat{\boldsymbol{\beta}}$!

Solving for $\hat{\boldsymbol{\beta}}$

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

Inference for β (not $\hat{\beta}$)

Covariance matrix of the estimators

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \text{Var}(\hat{\boldsymbol{\beta}}_0) & \text{Cov}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1) & \dots & \dots \\ \text{Cov}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1) & \text{Var}(\hat{\boldsymbol{\beta}}_1) & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

Covariance matrix of the estimators

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \text{Var}(\hat{\boldsymbol{\beta}}_0) & \text{Cov}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1) & \dots & \dots \\ \text{Cov}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1) & \text{Var}(\hat{\boldsymbol{\beta}}_1) & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$
$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Since σ is unknown, we use RSE to estimate σ:

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

Covariance matrix of the estimators

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \text{Var}(\hat{\boldsymbol{\beta}}_0) & \text{Cov}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1) & \dots & \dots \\ \text{Cov}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1) & \text{Var}(\hat{\boldsymbol{\beta}}_1) & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$
$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Since σ is unknown, we use RSE to estimate σ:

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

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

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \text{RSE}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

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

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

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

SE$(\hat{\beta}_3) =.$

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

Suppose we want to test between two hypotheses:

$$H_0 : \text{None of } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \text{ is related to } \boldsymbol{Y}$$
$$H_a : \text{at least one of } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \text{ is related to } \boldsymbol{Y}$$

Relationship between the response and the predictors

Suppose we want to test between two hypotheses:

$$H_0 : \text{None of } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \text{ is related to } \boldsymbol{Y}$$
$$H_a : \text{at least one of } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \text{ is related to } \boldsymbol{Y}$$

This is the same as

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

Relationship between the response and the predictors

Hypothesis test:

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

The decision will based on the following $F-$statistic:

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

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

How should we look at F– statistic?

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

Provided that H_0 is true,

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

How should we look at F– statistic?

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

Provided that H_0 is true,

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

And it can be shown that

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

How should we look at F– statistic?

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

Provided that H_0 is true,

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

And it can be shown that

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

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

How should we look at F– statistic?

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

Provided that H_0 is true,

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

And it can be shown that

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

- 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 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 $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$, 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

Higher values of newspaper $\rightarrow$ higher values of radio, which is the one that affects the sales.

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 one time is easier than calculating t for $\beta_0, \beta_1, \ldots, \beta_p$.

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 one time is easier than calculating t for $\beta_0, \beta_1, \ldots, \beta_p$.

- If we perform the t-test at significance level $\alpha = 0.05$ for $\beta_0 = \beta_1 = \ldots, \beta_p = 0$ the probability of us being wrong is:

Variable selection

Forward selection:

1. Start with 0 variable. In each step: add a variable that results in the lowest RSS.

Variable selection

Forward selection:

1. Start with 0 variable. In each step: add a variable that results in the lowest RSS.
2. Stop when RSS barely improves by adding any of the remaining variables.

Variable selection

Forward selection:

1. Start with 0 variable. In each step: add a variable that results in the lowest RSS.
2. Stop when RSS barely improves by adding any of the remaining variables.
3. 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:

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

Backward selection:

Variable selection

Forward selection:

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

Backward selection:

1. Start with all variables. In each step: remove a variable with the largest p-value.

Variable selection

Forward selection:

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

Backward selection:

1. Start with all variables. In each step: remove a variable with the largest p-value.
2. 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	R^2
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.