Variations on theme of Linear Regression

Chris Cornwell

Feb 18, 2025

Outline

Outline

Working with multiple independent variables

Before now, we focused on so-called *simple* linear regression, where there is a single independent variable x from which we predict y-values. Recall the `'Advertising.csv'` data set.

- Before, looked at the `Sales` as a function of `TV` (advertising budget). The data has budgets for other media: `Radio` and `Newspaper`.

Working with multiple independent variables

Before now, we focused on so-called *simple* linear regression, where there is a single independent variable x from which we predict y-values. Recall the `'Advertising.csv'` data set.

- Before, looked at the `Sales` as a function of `TV` (advertising budget). The data has budgets for other media: `Radio` and `Newspaper`.
- Fitting `Sales` to each one with simple linear regression (one for `TV`, one for `Radio`, one for `Newspaper`) is not right.
  - Ignores that all are contributing together to `Sales`.
  - Doesn't give predictive ability that matches data.

Working with multiple independent variables

Rather than fit separate simple linear regressions, use a single model with multiple independent variables.

Working with multiple independent variables

Rather than fit separate simple linear regressions, use a single model with multiple independent variables.
If $x_1, x_2, \ldots, x_d$ are the variables, use the model

$$p_0 x_1 + p_1 x_2 + \ldots + p_{d-1} x_d + p_d + \varepsilon$$

where $p_i, i = 0, 1, \ldots, d$ are the coefficients to be fit from the data and ε is a random variable with expected value 0.

- Simple linear regression case, $d = 1$: p_0 is the slope, p_1 is intercept.
- Advertising data set: independent variables are `TV`, `Radio`, `Newspaper`

Working with multiple independent variables

Rather than fit separate simple linear regressions, use a single model with multiple independent variables.
If $x_1, x_2, \ldots, x_d$ are the variables, use the model

$$p_0 x_1 + p_1 x_2 + \ldots + p_{d-1} x_d + p_d + \varepsilon$$

where $p_i, i = 0, 1, \ldots, d$ are the coefficients to be fit from the data and ε is a random variable with expected value 0.

- Simple linear regression case, $d = 1$: p_0 is the slope, p_1 is intercept.
- Advertising data set: independent variables are `TV`, `Radio`, `Newspaper`

To find the coefficients, alter procedure a bit. Now, A has column for each variable and last column of ones: $A = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_d, \vec{1}]$.
Just as before, the coefficients $\mathbf{p} = (p_0, \ldots, p_d)$ are given by $(A^T A)^{-1}(A^T \mathbf{y})$.

Advertising example

Using x_1 for the `TV` budget, x_2 for `Radio`, and x_3 for `Newspaper`, multiple linear regression for the Advertising data set is approximately

$$\text{Sales} = 0.04576x_1 + 0.18853x_2 + -0.00104x_3 + 2.93889 + \varepsilon.$$

Contrast this with what you get if you do three separate linear regressions (below, with R^2 coefficient).

Independent var.	`TV`	`Radio`	`New`
LSR line	$0.04754x_1 + 7.03259$	$0.2025x_2 + 9.31164$	$0.05469x$
R^2	0.612	0.332	0

R^2 coefficient

On previous slide, saw that R^2 for predicting `Sales` from `TV` was significantly higher than predicting from either `Radio` or `Newspaper`.

R^2 coefficient

On previous slide, saw that R^2 for predicting `Sales` from `TV` was significantly higher than predicting from either `Radio` or `Newspaper`.

Can compute R^2 from multiple linear regression model, either using just two of the variables or using all three.

First, recall result from simple regression:

Independent var.	`TV`	`Radio`	`Newspaper`
R^2	0.612	0.332	0.052

R^2 coefficient

On previous slide, saw that R^2 for predicting `Sales` from `TV` was significantly higher than predicting from either `Radio` or `Newspaper`.

Can compute R^2 from multiple linear regression model, either using just two of the variables or using all three.

First, recall result from simple regression:

Independent var.	`TV`	`Radio`	`Newspaper`
R^2	0.612	0.332	0.052

Now, R^2 for all possible pairs of two:

Two vars.	`TV`, `Radio`	`TV`, `Newspaper`	`Radio`, `Newspaper`
R^2	0.89719	0.646	0.333

R^2 coefficient

On previous slide, saw that R^2 for predicting `Sales` from `TV` was significantly higher than predicting from either `Radio` or `Newspaper`.

Can compute R^2 from multiple linear regression model, either using just two of the variables or using all three.

First, recall result from simple regression:

Independent var.	`TV`	`Radio`	`Newspaper`
R^2	0.612	0.332	0.052

Now, R^2 for all possible pairs of two:

Two vars.	`TV`, `Radio`	`TV`, `Newspaper`	`Radio`, `Newspaper`
R^2	0.89719	0.646	0.333

The value of R^2 with all three predictor (independent) variables is: 0.89721.

Outline

Using powers “like” other variables

Often, a linear model does not seem like a good fit for our data. What about trying to fit the data to a polynomial?
i.e., consider the model

$$p_0 x^d + p_1 x^{d-1} + \ldots + p_{d-1} x + p_d + \varepsilon$$

for some degree d, and find the coefficients which give best fit polynomial.
For the procedure, use essentially the same idea for the matrix A, but using powers of your variable x (or, variables) instead of using different independent variables. Given data with x-coordinates $x_1, x_2, \ldots, x_m$, the matrix A is known as a **Vandermonde matrix**.

---

[1] Removed rows that contained schools receiving fewer than 2500 applications.

Using powers "like" other variables

Often, a linear model does not seem like a good fit for our data. What about trying to fit the data to a polynomial?
i.e., consider the model

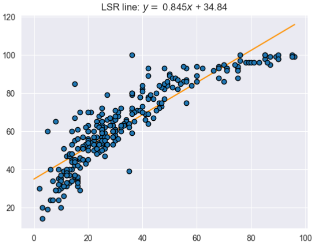
$$p_0x^d + p_1x^{d-1} + \ldots + p_{d-1}x + p_d + \varepsilon$$

for some degree d, and find the coefficients which give best fit polynomial.
For the procedure, use essentially the same idea for the matrix A, but using powers of your variable x (or, variables) instead of using different independent variables. Given data with x-coordinates $x_1, x_2, \ldots, x_m$, the matrix A is known as a **Vandermonde matrix**.
Example: Taking the `'College.csv'` data set from the DataSets folder. Two of the columns are `'Top10perc'` and `'Top25perc'`. For the schools in the data set, these columns give the percentage of the entering class that were in the top 10% (resp. 25%) of their graduating high school class.[1]

[1]Removed rows that contained schools receiving fewer than 2500 applications.

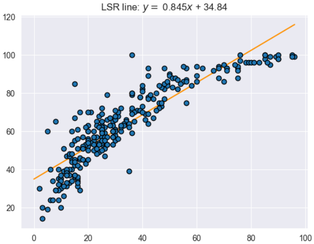
Example

Here is the data set with a least squares line. The value of R^2 is 0.791.

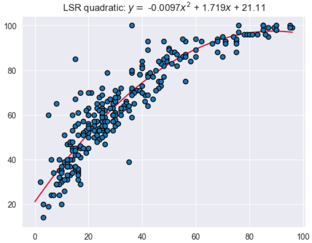


Example

Here is the data set with a least squares line. The value of R^2 is 0.791.



Next, we have the data set with a least squares quadratic polynomial fit. The R^2 value is 0.854.



Value of R^2 as polynomial degree increases

What will happen to the value of R^2 if we increase the degree of the polynomial that we fit to the data?

- Note: Suppose that $n > d$. A Vandermonde matrix for x-values $x_1, x_2, \ldots, x_n$, which has $d+1$ columns (so, highest power is x_i^d), will have rank $d+1$ if and only if there are $d+1$ of the x_i that are distinct.

  If $x_1, x_2, \ldots, x_{d+1}$ are pairwise distinct, say, then the determinant of the $(d+1)\times(d+1)$ submatrix for their corresponding rows is

$$\prod_{1\le i<j\le d+1}(x_j - x_i).$$

A_0: the Vandermonde matrix used in procedure to fit a polynomial of degree d; set A_1 to be one used for polynomial of degree $d+1$.[2] From Note, as long as enough of the x_i are distinct, $\mathrm{rank}(A_1) = \mathrm{rank}(A_0) + 1$. This means: $\mathrm{Col}(A_0)$ is proper subspace of $\mathrm{Col}(A_1)$. So, using it makes $|y - \hat{y}|^2$ smaller. Since $\sum(y-\bar{y})^2$ is unchanged, makes R^2 closer to 1.

[2]So, A_1 has all the columns of A_0, and one additional column.