Variations on theme of Linear Regression

Chris Cornwell

Feb 18, 2025

Outline

Outline

Multiple variables

Polynomial fitting

Working with multiple independent variables

Before now, we focused on *simple* linear regression, with a single independent variable x used to predict values of $\hat{y}$.

Recall the `'Advertising.csv'` data set.

Working with multiple independent variables

Before now, we focused on *simple* linear regression, with a single independent variable x used to predict values of $\hat{y}$.
Recall the `'Advertising.csv'` data set.

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

Working with multiple independent variables

Before now, we focused on *simple* linear regression, with a single independent variable x used to predict values of $\hat{y}$.

Recall the `'Advertising.csv'` data set.

- Before, looked at the `Sales` (y) as a function of `TV` (advertising budget, x). In data set, 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 inadequate.

Working with multiple independent variables

Before now, we focused on *simple* linear regression, with a single independent variable x used to predict values of $\hat{y}$.

Recall the `'Advertising.csv'` data set.

- Before, looked at the `Sales` (y) as a function of `TV` (advertising budget, x). In data set, 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 inadequate.
  - 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 more than one independent variable – **multiple linear regression**.

Working with multiple independent variables

Rather than fit separate simple linear regressions, use a single model with more than one independent variable – **multiple linear regression**.

If $x_0, x_1, \ldots, x_{d-1}$ are the variables, use the model

$$y = p_0 x_0 + p_1 x_1 + \ldots + p_{d-1} x_{d-1} + p_d + \varepsilon$$

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

Working with multiple independent variables

Rather than fit separate simple linear regressions, use a single model with more than one independent variable – **multiple linear regression**.
If $x_0, x_1, \ldots, x_{d-1}$ are the variables, use the model

$$y = p_0 x_0 + p_1 x_1 + \ldots + p_{d-1} x_{d-1} + p_d + \varepsilon$$

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

- Simple linear regression case, $d = 1$: p_0 is the slope, p_1 is intercept.

Working with multiple independent variables

Rather than fit separate simple linear regressions, use a single model with more than one independent variable – **multiple linear regression**.
If $x_0, x_1, \dots, x_{d-1}$ are the variables, use the model

$$y = p_0 x_0 + p_1 x_1 + \ldots + p_{d-1} x_{d-1} + p_d + \varepsilon$$

where $p_i, i = 0, 1, \dots, d$ are coefficients to be fit from the data; ε is 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`; $d = 3$.

Working with multiple independent variables

To find the coefficients, alter procedure a bit.
Matrix A is size $n \times (d+1)$ and has column for each variable (and a column of ones). That is, treating each $\vec{x}_i$ as a column vector (with one entry for each data point),

$$A = \left[\vec{x}_0, \quad \vec{x}_1, \quad \ldots, \quad \vec{x}_{d-1}, \quad \vec{1}\right].$$

[1]If a "real world" data set with $n \geq d+1$, almost surely.

Working with multiple independent variables

To find the coefficients, alter procedure a bit.
Matrix A is size $n \times (d+1)$ and has column for each variable (and a column of ones). That is, treating each $\vec{x}_i$ as a column vector (with one entry for each data point),

$$A = \begin{bmatrix} \vec{x}_0, & \vec{x}_1, & \ldots, & \vec{x}_{d-1}, & \vec{1} \end{bmatrix}.$$

Just as before, the coefficients $\mathbf{p} = (\hat{p}_0, \ldots, \hat{p}_d)$ are given by $(A^T A)^{-1}(A^T \mathbf{y})$, provided that $A^T A$ is invertible.

[1] If a "real world" data set with $n \geq d+1$, almost surely.

Working with multiple independent variables

To find the coefficients, alter procedure a bit.
Matrix A is size $n \times (d+1)$ and has column for each variable (and a column of ones). That is, treating each $\vec{x}_i$ as a column vector (with one entry for each data point),

$$A = \begin{bmatrix} \vec{x}_0, & \vec{x}_1, & \dots, & \vec{x}_{d-1}, & \vec{1} \end{bmatrix}.$$

Just as before, the coefficients $\mathbf{p} = (\hat{p}_0, \dots, \hat{p}_d)$ are given by $(A^T A)^{-1}(A^T \mathbf{y})$, provided that $A^T A$ is invertible.
The matrix $A^T A$ is invertible if A has rank $d+1$ (when $\{\vec{x}_0, \vec{x}_1, \dots, \vec{x}_{d-1}, \vec{1}\}$ a linearly indpt. set). [1]

[1] If a "real world" data set with $n \geq d+1$, almost surely.

Working with multiple independent variables

To find the coefficients, alter procedure a bit.
Matrix A is size $n \times (d+1)$ and has column for each variable (and a column of ones). That is, treating each $\vec{x}_i$ as a column vector (with one entry for each data point),

$$A = \begin{bmatrix} \vec{x}_0, & \vec{x}_1, & \ldots, & \vec{x}_{d-1}, & \vec{1} \end{bmatrix}.$$

Just as before, the coefficients $\mathbf{p} = (\hat{p}_0, \ldots, \hat{p}_d)$ are given by $(A^T A)^{-1}(A^T \mathbf{y})$, provided that $A^T A$ is invertible.
The matrix $A^T A$ is invertible if A has rank $d+1$ (when $\{\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_{d-1}, \vec{1}\}$ a linearly indpt. set). [1]

- Larger $d \longrightarrow$ more likely $A^T A$ is poorly conditioned (potential issues from numerically computing its inverse).

[1] If a "real world" data set with $n \geq d+1$, almost surely.

Advertising example

x_0 for `TV` budget; x_1 for `Radio` budget; x_2 for `Newspaper` budget.

Advertising example

x_0 for TV budget; x_1 for Radio budget; x_2 for Newspaper budget.
Writing y for Sales, multiple linear regression model for Advertising data is approximately

$$\hat{y} = 0.0458x_0 + 0.1885x_1 - 0.001x_2 + 2.9389.$$

Advertising example

x_0 for `TV` budget; x_1 for `Radio` budget; x_2 for `Newspaper` budget.
Writing y for `Sales`, multiple linear regression model for Advertising data is approximately

$$\hat{y} = 0.0458x_0 + 0.1885x_1 - 0.001x_2 + 2.9389.$$

Interpretation: given fixed budget for radio and newspaper ads, increasing TV ad budget by $1000 will increase sales by around 46 units (in each market, on average).

Advertising example

x_0 for TV budget; x_1 for Radio budget; x_2 for Newspaper budget.
Writing y for Sales, multiple linear regression model for Advertising data is approximately

$$\hat{y} = 0.0458x_0 + 0.1885x_1 - 0.001x_2 + 2.9389.$$

Interpretation: given fixed budget for radio and newspaper ads, increasing TV ad budget by \$1000 will increase sales by around 46 units (in each market, on average).
Contrast with result of three separate linear regressions, below.

Variable	TV	Radio	Newspaper
LSR line	$0.0475x_0 + 7.0326$	$0.2025x_1 + 9.3116$	$0.0547x_2 + 12.3514$
R^2	0.612	0.332	0.052

R^2

Previously: R^2 for predicting `Sales` from `TV` significantly higher than from either `Radio` or `Newspaper`.

R^2

Previously: R^2 for predicting `Sales` from `TV` significantly higher than from either `Radio` or `Newspaper`.

Can get R^2 from regression model, either using 2 of the variables, or using all 3.

First, recall result from simple regression:

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

R^2

Previously: R^2 for predicting `Sales` from `TV` significantly higher than from either `Radio` or `Newspaper`.

Can get R^2 from regression model, either using 2 of the variables, or using all 3.

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

Previously: R^2 for predicting `Sales` from `TV` significantly higher than from either `Radio` or `Newspaper`.
Can get R^2 from regression model, either using 2 of the variables, or using all 3.
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. What conclusion can we draw?

How small to decide not significant?

[2] Recall, *SE* how far $\hat{p}_i$ is from population coeff. p_i, on average.

How small to decide not significant?

Hypothesis testing: choose a p-value threshold (often < 0.05 or < 0.01). The p-value corresponds to some t-statistic – use regression coefficient ($\hat{p}_i$ for x_i) and standard error.

- In example, if using simple linear regression on `Newspaper`, would get the variable is significant. However, using multiple regression with `TV`, `Radio`, and `Newspaper`, get very large p-value $\longrightarrow$ so, not significant.

[2] Recall, SE how far $\hat{p}_i$ is from population coeff. p_i, on average.

How small to decide not significant?

Hypothesis testing: choose a p-value threshold (often < 0.05 or < 0.01). The p-value corresponds to some t-statistic – use regression coefficient ($\hat{p}_i$ for x_i) and standard error.

- In example, if using simple linear regression on `Newspaper`, would get the variable is significant. However, using multiple regression with `TV`, `Radio`, and `Newspaper`, get very large p-value $\longrightarrow$ so, not significant.

Alternatively: if sample regression coeff. varies a lot (relative to size) compared to coeff.s of the other var's, that variable is not significant.

---

[2] Recall, SE how far $\hat{p}_i$ is from population coeff. p_i, on average.

How small to decide not significant?

Hypothesis testing: choose a p-value threshold (often < 0.05 or < 0.01). The p-value corresponds to some t-statistic – use regression coefficient ($\hat{p}_i$ for x_i) and standard error.

- In example, if using simple linear regression on `Newspaper`, would get the variable is significant. However, using multiple regression with `TV`, `Radio`, and `Newspaper`, get very large p-value $\longrightarrow$ so, not significant.

Alternatively: if sample regression coeff. varies a lot (relative to size) compared to coeff.s of the other var's, that variable is not significant.

- p-value large when t-statistic is small, which is when SE is large *relative to size of* $\hat{p}_i$. [2]

[2] Recall, SE how far $\hat{p}_i$ is from population coeff. p_i, on average.

Intuitive estimate of significance

Checking whether fluctuation of regression coefficient for an independent variable, relative to coeff.'s size, is large.

[3]*Some evidence in literature (Goodhue-Lewis, 2012) that not much precision is to be gained with more than 100 samples, for bootstrapping standard errors.

[4]This is an example of a bootstrapping procedure: the whole sample is used as a proxy for the population and the subsamples, or resamplings, are simulating samples from the population.

Intuitive estimate of significance

Checking whether fluctuation of regression coefficient for an independent variable, relative to coeff.'s size, is large.

1. Take around 100 random subsamples[3] of data (or, resamplings with replacement); compute $\hat{p}_i$ for those. Standard deviation of them $\approx$ SE.

---

[3]*Some evidence in literature (Goodhue-Lewis, 2012) that not much precision is to be gained with more than 100 samples, for bootstrapping standard errors.

[4]This is an example of a bootstrapping procedure: the whole sample is used as a proxy for the population and the subsamples, or resamplings, are simulating samples from the population.

Intuitive estimate of significance

Checking whether fluctuation of regression coefficient for an independent variable, relative to coeff.'s size, is large.

1. Take around 100 random subsamples[3] of data (or, resamplings with replacement); compute $\hat{p}_i$ for those. Standard deviation of them $\approx$ SE.
2. Use regression coeff. from whole data set, $\approx p_i$. If standard dev. found in 1., divided by this coeff., is larger than about 0.5, variable is not significant.
   - Since we are *estimating some things* here, don't use as a hard cutoff. Getting 0.48, versus 0.59, would perhaps both be *weakly* significant. However, if larger than 1.5, say, definitely not significant.

[4]

[3]*Some evidence in literature (Goodhue-Lewis, 2012) that not much precision is to be gained with more than 100 samples, for bootstrapping standard errors.

[4]This is an example of a bootstrapping procedure: the whole sample is used as a proxy for the population and the subsamples, or resamplings, are simulating samples from the population.

Outline

Multiple variables

Polynomial fitting

Powers of x in place of multiple 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?

[5]Could do mix, multivariate regression and powers.

Powers of x in place of multiple 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

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

for some degree d, and find the coefficients which give best fit polynomial.

[5]Could do mix, multivariate regression and powers.

Powers of x in place of multiple 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

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

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 single variable x instead of using different independent variables[5]. Given data with x-coordinates $x_1, x_2, \ldots, x_n$, the matrix A is known as a **Vandermonde matrix**.

[5] Could do mix, multivariate regression and powers.

Powers of x in place of multiple 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

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

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 single variable x instead of using different independent variables[5]. Given data with x-coordinates $x_1, x_2, \ldots, x_n$, the matrix A is known as a **Vandermonde matrix**.

$$A = \begin{bmatrix} x_1^d & \ldots & x_1^2 & x_1 & 1 \\ x_2^d & \ldots & x_2^2 & x_2 & 1 \\ \vdots & & \vdots & \vdots & \\ x_n^d & \ldots & x_n^2 & x_n & 1 \end{bmatrix}$$

[5]Could do mix, multivariate regression and powers.

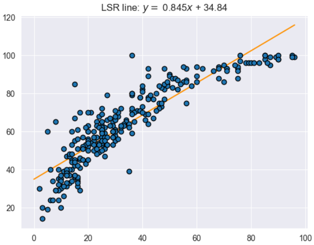
Example

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.[6]

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

Example

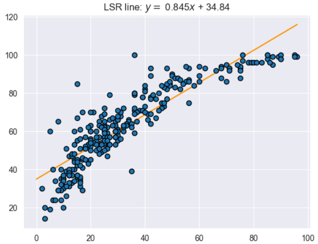
Taking the `'College.csv'` data set from the DataSets folder. Two of the columns are `'Top10perc'` and `'Top25perc'`. For the schools in the data set, these columns give the percentage of the entering class that were in the top 10% (resp. 25%) of their graduating high school class.[6] Here is the data set with a least squares line. The value of R^2 is 0.791.



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

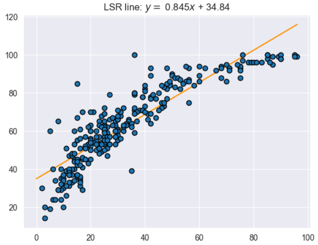
Example

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

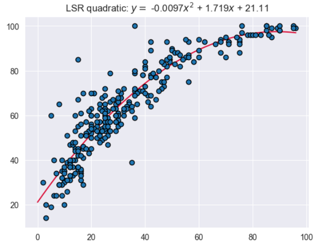


Example

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



Next, 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?

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

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.

---

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

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

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

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

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

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

set A_0: the Vandermonde matrix used to fit polynomial of degree d; set A_1: the one used for polynomial of degree $d+1$. [7]

From Note, as long as enough of the x_i are distinct,

$\mathrm{rank}(A_1) = \mathrm{rank}(A_0) + 1$.

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

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

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

set A_0: the Vandermonde matrix used to fit polynomial of degree d; set A_1: the one used for polynomial of degree $d+1$. [7]

From Note, as long as enough of the x_i are distinct, $\operatorname{rank}(A_1) = \operatorname{rank}(A_0) + 1$.

Meaning: $\mathrm{Col}(A_0)$ is proper subspace of $\mathrm{Col}(A_1)$. So, using A_1 makes $|y - \hat{y}|^2$ smaller. Since $\sum(y - \bar{y})^2$ is unchanged, makes R^2 closer to 1.

---

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