Variations on theme of Linear Regression

Chris Cornwell

Feb 18, 2025

Outline

Outline

Multiple variables

Measuring how well LSR line fits

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.

`TV`	`Radio`	`Newspaper`
$0.04754x_1 + 7.03259$	$0.2025x_2 + 9.31164$	$0.05469x_3 + 12.3514$

Confidence intervals

How close do we suspect $\hat{m}$ and $\hat{b}$ to be to the “true” (population) slope and intercept?

Confidence intervals

How close do we suspect $\hat{m}$ and $\hat{b}$ to be to the "true" (population) slope and intercept?

Standard error (SE): Suppose that for our error term ε, we have $\mathrm{Var}(\varepsilon) = \sigma^2$. Sample size: n.

Confidence intervals

How close do we suspect $\hat{m}$ and $\hat{b}$ to be to the “true” (population) slope and intercept?

Standard error (SE): Suppose that for our error term ε, we have $\mathrm{Var}(\varepsilon) = \sigma^2$. Sample size: n.

Using $\bar{x}$ for the average of $x_1, \ldots, x_n$,

$$SE(\hat{m})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$

$$SE(\hat{b})^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

Confidence intervals

How close do we suspect $\hat{m}$ and $\hat{b}$ to be to the "true" (population) slope and intercept?

Standard error (SE): Suppose that for our error term ε, we have $\mathrm{Var}(\varepsilon) = \sigma^2$. Sample size: n.

Using $\bar{x}$ for the average of $x_1, \ldots, x_n$,

$$SE(\hat{m})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$

$$SE(\hat{b})^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

Roughly, these are the amount, on average, that $\hat{m}$ (resp. $\hat{b}$) differs from true slope m (resp. true intercept b).

Confidence intervals

How close do we suspect $\hat{m}$ and $\hat{b}$ to be to the "true" (population) slope and intercept?

Standard error (SE): Suppose that for our error term ε, we have $\mathrm{Var}(\varepsilon) = \sigma^2$. Sample size: n.

Using $\bar{x}$ for the average of $x_1, \dots, x_n$,

$$SE(\hat{m})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$

$$SE(\hat{b})^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

Roughly, these are the amount, on average, that $\hat{m}$ (resp. $\hat{b}$) differs from true slope m (resp. true intercept b).

σ is unknown, but can estimate it with **residual standard error**:

$$\hat{\sigma}^2 = RSE^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

Confidence intervals

How close do we suspect $\hat{m}$ and $\hat{b}$ to be to the "true" (population) slope and intercept?

Formulae:

$$SE(\hat{m})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$

$$SE(\hat{b})^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

Estimate:

$$\sigma^2 \approx RSE^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

---

[1] 95% of the time, these intervals contain m, b.

Confidence intervals

How close do we suspect $\hat{m}$ and $\hat{b}$ to be to the "true" (population) slope and intercept?

Formulae:

$$SE(\hat{m})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$

$$SE(\hat{b})^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Estimate:

$$\sigma^2 \approx RSE^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

Can get (roughly) 95% confidence interval[1] with $\pm 2SE$:

$$(\hat{m} - 2SE(\hat{m}), \hat{m} + 2SE(\hat{m}))$$

and

$$(\hat{b} - 2SE(\hat{b}), \hat{b} + 2SE(\hat{b})).$$

[1] 95% of the time, these intervals contain m, b.

Outline

Multiple variables

Measuring how well LSR line fits

Mean Squared Error

How to measure how well the data fits to regression line?

Mean Squared Error

How to measure how well the data fits to regression line?

In linear regression, we found $\hat{y}_i, 1 \leq i \leq n$ so that the points $(x_1, \hat{y}_1), \ldots, (x_n, \hat{y}_n)$ fit exactly to a line. Could use average of $(y_i - \hat{y}_i)^2$ as our measure.

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

Mean Squared Error

How to measure how well the data fits to regression line?
In linear regression, we found $\hat{y}_i, 1 \leq i \leq n$ so that the points $(x_1, \hat{y}_1), \ldots, (x_n, \hat{y}_n)$ fit exactly to a line. Could use average of $(y_i - \hat{y}_i)^2$ as our measure.

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

- Called the mean squared error, MSE, of the LSR line.
- Larger MSE (for same sample size), the farther y_i is from $\hat{y}_i$, on average.

Mean Squared Error

How to measure how well the data fits to regression line?
In linear regression, we found $\hat{y}_i, 1 \leq i \leq n$ so that the points $(x_1, \hat{y}_1), \ldots, (x_n, \hat{y}_n)$ fit exactly to a line. Could use average of $(y_i - \hat{y}_i)^2$ as our measure.

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

- Called the mean squared error, MSE, of the LSR line.
- Larger MSE (for same sample size), the farther y_i is from $\hat{y}_i$, on average.

Closely related to RSE (residual standard error). Recall,

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

So $\text{MSE} = \frac{n-2}{n}\text{RSE}^2$.