Assessing accuracy of linear regression

Chris Cornwell

Oct 9, 2025

Outline

Outline

Difference between parameters from “population” vs. from data

- Modeling relationship between independent variables and y with a linear model (with noise in the y-coordinate direction). In other words, the modeled relationship is that

$$y \approx b + \mathbf{x}^T \mathbf{w}.$$

for some parameters $\mathbf{w}$, b.

[1] There is some number M so that $f : \mathbb{R}^N \to \mathbb{R}^M$.

Difference between parameters from “population” vs. from data

- Modeling relationship between independent variables and y with a linear model (with noise in the y-coordinate direction). In other words, the modeled relationship is that

$$y \approx b + \mathbf{x}^T \mathbf{w}.$$

  for some parameters $\mathbf{w}$, b.
- Alternatively, we may have some function[1] of the variables $f(\mathbf{x})$ and a modeled relationship

$$y \approx b + f(\mathbf{x})^T \mathbf{w}.$$

[1] There is some number M so that $f : \mathbb{R}^N \to \mathbb{R}^M$.

Difference between parameters from "population" vs. from data

- Modeling relationship between independent variables and y with a linear model (with noise in the y-coordinate direction). In other words, the modeled relationship is that

$$y \approx b + \mathbf{x}^T \mathbf{w}.$$

  for some parameters $\mathbf{w}$, b.
- Alternatively, we may have some function[1] of the variables $f(\mathbf{x})$ and a modeled relationship

$$y \approx b + f(\mathbf{x})^T \mathbf{w}.$$

However, we find values for $\mathbf{w}$ and b by using observed data, *sampled* from a "population." Among the entire population, there is a best fit linear model having some parameters. However, from the observed data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_P, y_P)$ if our procedure determines best fit parameters $b^\star$ and $\mathbf{w}^\star$, these are not (necessarily) the parameters for the population linear model.

[1] There is some number M so that $f : \mathbb{R}^N \to \mathbb{R}^M$.

Example

Simulate noisy linear data: make 30 points, from a line with slope -1.6 and y-intercept 0.8; put noise into the y-coordinate with standard deviation $\sigma = 0.5$.

Example

Simulate noisy linear data: make 30 points, from a line with slope -1.6 and y-intercept 0.8; put noise into the y-coordinate with standard deviation $\sigma = 0.5$.

```
x = np.random.uniform(0, 2, size=30)

def simulate_data(x, std):
    return -1.6*x + 0.8 + np.random.normal(0, std, size=len(x))
y = simulate_data(x, 0.5)
```

Example

Simulate noisy linear data: make 30 points, from a line with slope -1.6 and y-intercept 0.8; put noise into the y-coordinate with standard deviation $\sigma = 0.5$.

```
x = np.random.uniform(0, 2, size=30)

def simulate_data(x, std):
    return -1.6*x + 0.8 + np.random.normal(0, std, size=len(x))
y = simulate_data(x, 0.5)
```

With this simulated data set, compute the slope $w^\star$ and intercept $b^\star$ from linear regression; then put $w^\star$, and $b^\star$, into a list of slopes, intercepts respectively. Iterate this 1000 times $\rightarrow$ a list of 1000 slopes, another with 1000 intercepts.

Example

Simulate noisy linear data: make 30 points, from a line with slope -1.6 and y-intercept 0.8; put noise into the y-coordinate with standard deviation $\sigma = 0.5$.

```python
x = np.random.uniform(0, 2, size=30)

def simulate_data(x, std):
    return -1.6*x + 0.8 + np.random.normal(0, std, size=len(x))
y = simulate_data(x, 0.5)
```

With this simulated data set, compute the slope $w^\star$ and intercept $b^\star$ from linear regression; then put $w^\star$, and $b^\star$, into a list of slopes, intercepts respectively. Iterate this 1000 times $\rightarrow$ a list of 1000 slopes, another with 1000 intercepts.

What is the mean of these slopes and intercepts that were found?

Sample statistic, relation to population statistic

This is fundamental to statistics.

- Say that a sample of 2000 people are selected from around the country and their height is measured. Mean of these 2000 heights: sample mean.

Sample statistic, relation to population statistic

This is fundamental to statistics.

- Say that a sample of 2000 people are selected from around the country and their height is measured. Mean of these 2000 heights: sample mean.
- Sample mean differs from the true mean height of the entire population of the country. (Not by much, perhaps.)
  - Weak Law of Large Numbers: if s random samples of 2000 people taken, and each sample mean calculated then, as $s \to \infty$, the average of those s sample means limits (in probability) to the population mean.

Sample statistic, relation to population statistic

This is fundamental to statistics.

- Say that a sample of 2000 people are selected from around the country and their height is measured. Mean of these 2000 heights: sample mean.
- Sample mean differs from the true mean height of the entire population of the country. (Not by much, perhaps.)
  - Weak Law of Large Numbers: if s random samples of 2000 people taken, and each sample mean calculated then, as $s \to \infty$, the average of those s sample means limits (in probability) to the population mean.
- Analogous thing happens with data from linear relationship with noise – think of parameters $w^\star$ and $b^\star$ as sample statistics (like the sample mean).

Confidence intervals

How close do we suspect $w^\star$ and $b^\star$ to be to the *true* (population) slope and intercept?

[2]These formulae are just for single-variable linear regression.

Confidence intervals

How close do we suspect $w^\star$ and $b^\star$ 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: P.

[2] These formulae are just for single-variable linear regression.

Confidence intervals

How close do we suspect $w^\star$ and $b^\star$ 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: P.

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

$$(SE(w^\star))^2 = \frac{\sigma^2}{\sum_{i=1}^{P}(x_i - \bar{x})^2};$$

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

[2] These formulae are just for single-variable linear regression.

Confidence intervals

How close do we suspect $w^\star$ and $b^\star$ 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: P.

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

$$(SE(w^\star))^2 = \frac{\sigma^2}{\sum_{i=1}^{P}(x_i - \bar{x})^2};$$

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

Roughly, the Standard Error is the amount, on average, that $w^\star$ (resp. $b^\star$) differs from true slope w (resp. true intercept b).[2]

[2] These formulae are just for single-variable linear regression.

Confidence intervals

How close do we suspect $w^\star$ and $b^\star$ 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: P.

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

$$(SE(w^\star))^2 = \frac{\sigma^2}{\sum_{i=1}^{P}(x_i - \bar{x})^2};$$

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

Roughly, the Standard Error is the amount, on average, that $w^\star$ (resp. $b^\star$) differs from true slope w (resp. true intercept b).[2]

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

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

---

[2] These formulae are just for single-variable linear regression.

Confidence intervals, cont'd

How close do we suspect $w^\star$ and $b^\star$ to be to the *true* (population) slope and intercept?

Formulae:

$$(SE(w^\star))^2 = \frac{\sigma^2}{\sum_{i=1}^{P}(x_i - \bar{x})^2};$$

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

Estimate:

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

[3] 95% of the time, these intervals contain population w, b.

Confidence intervals, cont'd

How close do we suspect $w^\star$ and $b^\star$ to be to the *true* (population) slope and intercept?

Formulae:

$$\left(SE(w^\star)\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{P}(x_i - \bar{x})^2};$$

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

Estimate:

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

Can get (roughly) 95% confidence interval[3] with $\pm 2SE$:

$$\left(w^\star - 2SE(w^\star), w^\star + 2SE(w^\star)\right)$$

and

$$\left(b^\star - 2SE(b^\star), b^\star + 2SE(b^\star)\right).$$

[3] 95% of the time, these intervals contain population w, b.

Outline

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 "predicted" $\hat{y}_i = b^\star + \mathbf{x}_i^T \mathbf{w}^\star, 1 \leq i \leq P$ so that the points $(\mathbf{x}_1, \hat{y}_1), \ldots, (\mathbf{x}_P, \hat{y}_P)$ fit a line. Could use the Mean Squared Error as our measure.

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

Mean Squared Error

How to measure how well the data fits to regression line?
In linear regression, we found "predicted" $\hat{y}_i = b^\star + \mathbf{x}_i^T \mathbf{w}^\star, 1 \leq i \leq P$ so that the points $(\mathbf{x}_1, \hat{y}_1), \ldots, (\mathbf{x}_P, \hat{y}_P)$ fit a line. Could use the Mean Squared Error as our measure.

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

- For the same sample size, the larger the MSE is 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 "predicted" $\hat{y}_i = b^\star + \mathbf{x}_i^T \mathbf{w}^\star, 1 \leq i \leq P$ so that the points $(\mathbf{x}_1, \hat{y}_1), \ldots, (\mathbf{x}_P, \hat{y}_P)$ fit a line. Could use the Mean Squared Error as our measure.

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

- For the same sample size, the larger the MSE is the farther y_i is from $\hat{y}_i$, on average.

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

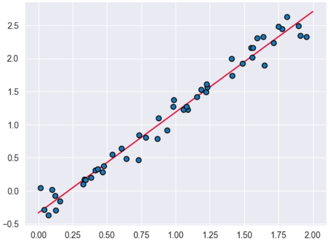
$$\mathrm{RSE} = \sqrt{\frac{1}{P-2} \sum_{i=1}^{P} (y_i - \hat{y}_i)^2}.$$

So $\mathrm{MSE} = \frac{P-2}{P} \mathrm{RSE}^2$.

Mean Squared Error, example

Recall, `'Example1.csv'` data. Its best fit line is

$$y = 1.520275x - 0.33458.$$

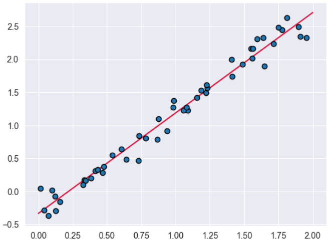


Mean Squared Error, example

Recall, `'Example1.csv'` data. Its best fit line is

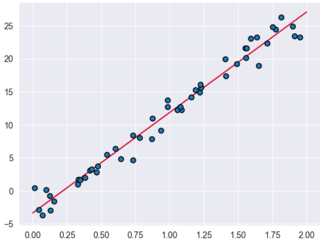
$$y = 1.520275x - 0.33458.$$

The MSE for this data and its predictions is ≈ 0.0197.
Does that mean that the linear model is a "good fit"?



Mean Squared Error, scaling

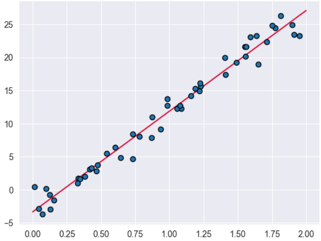
What about the following data and its best fit line? Here, the MSE is 1.9746.



Is it still a good fit?

Mean Squared Error, scaling

What about the following data and its best fit line? Here, the MSE is 1.9746.



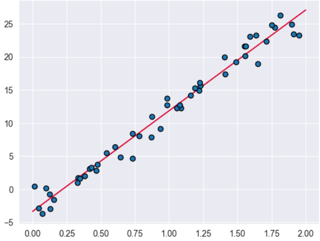
Is it still a good fit?

The data here is from `'Example1.csv'` again, except that the y-coordinates have been multiplied by 10. Its regression line is

$$y = 15.20275x - 3.3458.$$

Mean Squared Error, scaling

What about the following data and its best fit line? Here, the MSE is 1.9746.



Is it still a good fit?

The data here is from `'Example1.csv'` again, except that the y-coordinates have been multiplied by 10. Its regression line is

$$y = 15.20275x - 3.3458.$$

MSE is still a good measure to think about, but its size depends on scale of y-coordinates (equivalently, depends on units y is measured in).

R^2: Proportion of "variance explained"

Get a measure that is unchanged by scaling: first, set **total sum of squares** (TSS) to

$$\mathrm{TSS} = \sum_{i=1}^{P} (y_i - \bar{y})^2,$$

where $\bar{y} = \frac{1}{P} \sum_{i=1}^{P} y_i$.

R^2: Proportion of "variance explained"

Get a measure that is unchanged by scaling: first, set **total sum of squares** (TSS) to

$$\text{TSS} = \sum_{i=1}^{P} (y_i - \bar{y})^2,$$

where $\bar{y} = \frac{1}{P} \sum_{i=1}^{P} y_i$.

Then,

$$R^2 = \frac{\text{TSS} - PMSE}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{P} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{P} (y_i - \bar{y}_i)^2}.$$

R^2: Proportion of "variance explained"

Get a measure that is unchanged by scaling: first, set **total sum of squares** (TSS) to

$$\mathrm{TSS} = \sum_{i=1}^{P} (y_i - \bar{y})^2,$$

where $\bar{y} = \frac{1}{P} \sum_{i=1}^{P} y_i$.

Then,

$$R^2 = \frac{\mathrm{TSS} - PMSE}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{P} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{P} (y_i - \bar{y}_i)^2}.$$

- R^2 does *not* depend on the scale of the y-coordinates.

R^2: Proportion of "variance explained"

Get a measure that is unchanged by scaling: first, set **total sum of squares** (TSS) to

$$\text{TSS} = \sum_{i=1}^{P} (y_i - \bar{y})^2,$$

where $\bar{y} = \frac{1}{P} \sum_{i=1}^{P} y_i$.

Then,

$$R^2 = \frac{\text{TSS} - PMSE}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{P} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{P} (y_i - \bar{y}_i)^2}.$$

- R^2 does *not* depend on the scale of the y-coordinates.
- Any data set, have $0 \leq R^2 \leq 1$ (provided R^2 is defined; i.e., we do not have $y_1, y_2, \ldots, y_n$ all the same).
  - Can you prove this?

R^2: Proportion of "variance explained"

Get a measure that is unchanged by scaling: first, set **total sum of squares** (TSS) to

$$\mathrm{TSS} = \sum_{i=1}^{P} (y_i - \bar{y})^2,$$

where $\bar{y} = \frac{1}{P}\sum_{i=1}^{P} y_i$.
Then,

$$R^2 = \frac{\mathrm{TSS} - PMSE}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{P}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{P}(y_i - \bar{y}_i)^2}.$$

- R^2 does *not* depend on the scale of the y-coordinates.
- Any data set, have $0 \leq R^2 \leq 1$ (provided R^2 is defined; i.e., we do not have $y_1, y_2, \ldots, y_n$ all the same).
  - Can you prove this?
- Checking that R^2 is "close" to 1 is often done to indicate that a linear model is a very good one.

Outline

Linear regression, significance of variables

Checking significance of variables, an example: Recall the `'Advertising.csv'` data set, found on the class GitHub site, in folder DataSets.

Linear regression, significance of variables

Checking significance of variables, an example: Recall the `'Advertising.csv'` data set, found on the class GitHub site, in folder DataSets.

- Model `Sales` (y) as a function of advertising budgets in `TV` (x_1), `Radio` (x_2) and `Newspaper` (x_3).

Linear regression, significance of variables

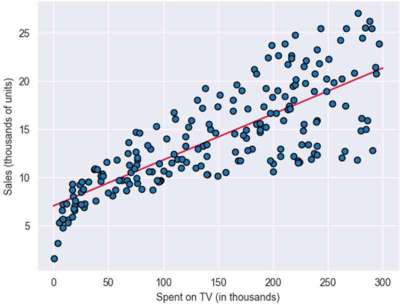
Checking significance of variables, an example: Recall the `'Advertising.csv'` data set, found on the class GitHub site, in folder DataSets.

- Model `Sales` (y) as a function of advertising budgets in `TV` (x_1), `Radio` (x_2) and `Newspaper` (x_3).

Linear regression, significance of variables

Checking significance of variables, an example: Recall the `'Advertising.csv'` data set, found on the class GitHub site, in folder DataSets.

- Model `Sales` (y) as a function of advertising budgets in `TV` (x_1), `Radio` (x_2) and `Newspaper` (x_3).
- Regression with just one of the variables ignores that all are contributing to `Sales` and doesn't predict y very well.



Working with multiple independent variables

Rather than separate single-variable linear regressions, use multivariable linear regression. (We did this before with this example.)

[4] When a "real world" data set with $P \geq N + 1$, almost surely.

Working with multiple independent variables

Rather than separate single-variable linear regressions, use multivariable linear regression. (We did this before with this example.)

With $\mathbf{x} = (x_1, x_2, x_3)$ as the variables, use the model

$$y \approx b + \mathbf{x}^T \mathbf{w} = b + w_1 x_1 + w_2 x_2 + w_3 x_3 .$$

[4]When a "real world" data set with $P \geq N + 1$, almost surely.

Working with multiple independent variables

Rather than separate single-variable linear regressions, use multivariable linear regression. (We did this before with this example.) With $\mathbf{x} = (x_1, x_2, x_3)$ as the variables, use the model

$$y \approx b + \mathbf{x}^T\mathbf{w} = b + w_1x_1 + w_2x_2 + w_3x_3.$$

As we discussed in Lecture 5, set A to be $P \times (N+1)$ matrix with a column of ones, and a column for each independent variable (in this example, $N = 3$). So,

$$A = \begin{bmatrix} \vec{1}, & \mathbf{x}_1, & \mathbf{x}_2, & \mathbf{x}_3 \end{bmatrix}.$$

$(b^\star, w_1^\star, w_2^\star, w_3^\star)$ is the solution to normal equations: $(A^TA)^{-1}(A^T\mathbf{y})$. *General note*: the matrix A^TA is invertible when A has rank $N+1$ (when $\vec{1}, \mathbf{x}_1, \ldots, \mathbf{x}_N$ are linearly independent).[4]

[4] When a "real world" data set with $P \geq N+1$, almost surely.

Working with multiple independent variables

Rather than separate single-variable linear regressions, use multivariable linear regression. (We did this before with this example.) With $\mathbf{x} = (x_1, x_2, x_3)$ as the variables, use the model

$$y \approx b + \mathbf{x}^T\mathbf{w} = b + w_1x_1 + w_2x_2 + w_3x_3.$$

As we discussed in Lecture 5, set A to be $P \times (N+1)$ matrix with a column of ones, and a column for each independent variable (in this example, $N = 3$). So,

$$A = \begin{bmatrix} \vec{1}, & \mathbf{x}_1, & \mathbf{x}_2, & \mathbf{x}_3 \end{bmatrix}.$$

$(b^\star, w_1^\star, w_2^\star, w_3^\star)$ is the solution to normal equations: $(A^TA)^{-1}(A^T\mathbf{y})$. *General note*: the matrix A^TA is invertible when A has rank $N+1$ (when $\vec{1}, \mathbf{x}_1, \ldots, \mathbf{x}_N$ are linearly independent).[4]

- Larger $N \to$ more likely there are numerical issues computing inverse of A^TA.

[4] When a "real world" data set with $P \geq N+1$, almost surely.

Back to the example

x_1 for `TV` budget; x_2 for `Radio` budget; x_3 for `Newspaper` budget.

Back to the example

x_1 for TV budget; x_2 for Radio budget; x_3 for Newspaper budget.
Multiple linear regression model for Advertising data is approximately

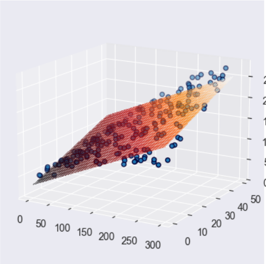
$$\hat{y} = 2.9389 + 0.0458x_1 + 0.1885x_2 - 0.001x_3.$$

Back to the example

x_1 for `TV` budget; x_2 for `Radio` budget; x_3 for `Newspaper` budget.
Multiple linear regression model for Advertising data is approximately

$$\hat{y} = 2.9389 + 0.0458x_1 + 0.1885x_2 - 0.001x_3.$$

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).



R^2

Here we consider the R^2 value from different choices for how many variables are used: one variable, two of the variables, or all 3 variables.

R^2

Here we consider the R^2 value from different choices for how many variables are used: one variable, two of the variables, or all 3 variables. First, the result from single-variable regression:

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

R^2

Here we consider the R^2 value from different choices for how many variables are used: one variable, two of the variables, or all 3 variables. First, the result from single-variable 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

Here we consider the R^2 value from different choices for how many variables are used: one variable, two of the variables, or all 3 variables. First, the result from single-variable 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 a change is ‘not significant’?

[5] Recall, on average, SE is how far $w_i^{\star}$ is from population coeff. w_i.

How small a change is '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 ($w_i^\star$ 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 $\rightarrow$ so, not significant.

[5] Recall, on average, SE is how far $w_i^\star$ is from population coeff. w_i.

How small a change is ‘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 ($w_i^\star$ 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 $\rightarrow$ so, not significant.

The formula for $\mathrm{SE}(w_i)$ ...

[5] Recall, on average, SE is how far $w_i^\star$ is from population coeff. w_i.

How small a change is ‘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 ($w_i^\star$ 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 $\to$ so, not significant.

The formula for $\mathrm{SE}(w_i)$ ...
Alternative: if sample $w_i^\star$ varies a lot (relative to its size) compared to coeff’s of the other var’s, that variable is not significant.

---

[5]Recall, on average, SE is how far $w_i^\star$ is from population coeff. w_i.

How small a change is ‘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 ($w_i^\star$ 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 $\to$ so, not significant.

The formula for $\mathrm{SE}\left(w_i\right)$ ...
Alternative: if sample $w_i^\star$ varies a lot (relative to its 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 $SE\left(w_i^\star\right)$ is large *relative to size of* $w_i^\star$. [5]

[5] Recall, on average, SE is how far $w_i^\star$ is from population coeff. w_i.

Intuitive estimate of significance (Bootstrapping)

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

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

[7]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 (Bootstrapping)

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

1. Take around 100 random subsamples[6] of data (or, resamplings with replacement); compute coefficient $w_i^\star$ for those. Standard deviation of those sample coefficients $\approx SE\left(w_i^\star\right)$.

---

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

[7] 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 (Bootstrapping)

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

1. Take around 100 random subsamples[6] of data (or, resamplings with replacement); compute coefficient $w_i^\star$ for those. Standard deviation of those sample coefficients $\approx SE\left(w_i^\star\right)$.
2. Use regression coeff. from whole data set, $\approx w_i$. 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.

[7]

---

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

[7]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

Powers of x in place of multiple variables

If a linear model does not seem a good fit for our data, can try fitting to a polynomial. This is just like doing linear regression on transformed data, **except** with more than one function transformation: using x, and x^2, and x^3, etc.

Powers of x in place of multiple variables

If a linear model does not seem a good fit for our data, can try fitting to a polynomial. This is just like doing linear regression on transformed data, **except** with more than one function transformation: using x, and x^2, and x^3, etc.

As a regression model, we use

$$y \approx w_d x^d + w_{d-1} x^{d-1} + \ldots + w_1 x + b$$

for some degree d, then find the coefficients for a best fit polynomial.

Powers of x in place of multiple variables

If a linear model does not seem a good fit for our data, can try fitting to a polynomial. This is just like doing linear regression on transformed data, **except** with more than one function transformation: using x, and x^2, and x^3, etc.

As a regression model, we use

$$y \approx w_d x^d + w_{d-1} x^{d-1} + \ldots + w_1 x + b$$

for some degree d, then find the coefficients for a best fit polynomial.

Use essentially the same idea for the matrix A, but put powers of x in each column (new features $= x^p$), instead of different independent variables. Given data $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{P}$, the matrix A is known as a **Vandermonde matrix**.

Powers of x in place of multiple variables

If a linear model does not seem a good fit for our data, can try fitting to a polynomial. This is just like doing linear regression on transformed data, **except** with more than one function transformation: using x, and x^2, and x^3, etc.

As a regression model, we use

$$y \approx w_d x^d + w_{d-1} x^{d-1} + \ldots + w_1 x + b$$

for some degree d, then find the coefficients for a best fit polynomial.

Use essentially the same idea for the matrix A, but put powers of x in each column (new features $= x^p$), instead of different independent variables. Given data $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{P}$, the matrix A is known as a **Vandermonde matrix**.

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

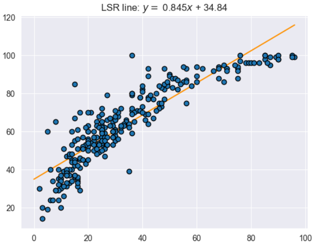
Example

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.[8]

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

Example

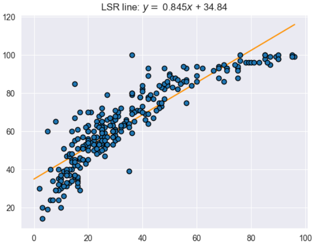
Taking the `'College.csv'` data set from the DataSets folder. Two of the columns are `'Top10perc'` and `'Top25perc'`. For the schools in the data set, these columns give the percentage of the entering class that were in the top 10% (resp. 25%) of their graduating high school class.[8] Here is the data set with a least squares line. The value of R^2 is 0.791.



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

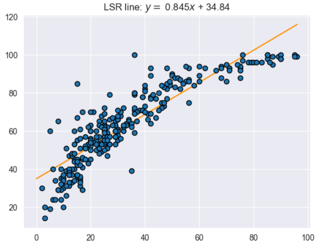
Example

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

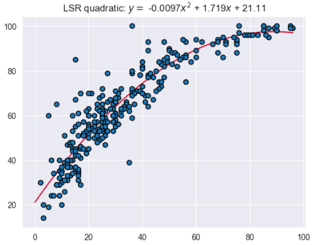


Example

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?

[9] 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 $P > d$. A Vandermonde matrix for x-values $x_1, x_2, \ldots, x_P$, 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.

---

[9] 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 $P > d$. A Vandermonde matrix for x-values $x_1, x_2, \ldots, x_P$, 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

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

[9]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 $P > d$. A Vandermonde matrix for x-values $x_1, x_2, \ldots, x_P$, 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

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

set A_0: the Vandermonde matrix used to fit polynomial of degree d; set A_1: the one used for polynomial of degree $d+1$. [9]

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

[9] 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 $P > d$. A Vandermonde matrix for x-values $x_1, x_2, \ldots, x_P$, 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

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

set A_0: the Vandermonde matrix used to fit polynomial of degree d; set A_1: the one used for polynomial of degree $d+1$. [9]

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

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

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