# Assessing accuracy of the LSR line

Chris Cornwell

Feb 13, 2025

# Outline

Assuming the data does have a linear relationship

Measuring how well LSR line fits

# Outline

Assuming the data does have a linear relationship

Measuring how well LSR line fits

# Underlying assumption

- Modeled points in a plane as being from a line, but with noise in the $y$-coordinate direction. In other words, we assumed an underlying relationship

$$y = mx + b + \varepsilon$$

  for some $m$ and $b$, and a random variable $\varepsilon$[1] that has expected value 0. Alternatively, among the "entire population" there is an LSR line $mx + b$.
- Assumption: $\varepsilon$ is independent of $x$.

When we have a data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, from the population, our procedure determines an LSR line $\hat{m}x + \hat{b}$. However, $\hat{m}$ and $\hat{b}$ are not the slope and intercept for the population curve $m$ and $b$.

---

[1] $\varepsilon$ is called the error term.

# Example

Simulate noisy linear data: make 30 points, using a standard deviation $\sigma = 0.5$. We'll use slope $-1.6$ and intercept $0.8$.

```
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)
```

In groups, compute slope and intercept of the LSR line for a size 30 simulated data set; store $\hat{m}$ and $\hat{b}$ (in two lists). Iterate this 1000 times $\longrightarrow$ a list of 1000 slopes and intercepts.

What is the mean of the slopes and of the intercepts?

# 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, as $s \longrightarrow \infty$, mean of the sample means limits (in probability) to population mean.
- Analogous thing happens with data from linear relationship with noise – think of parameters $\hat{m}$ and $\hat{b}$ as sample statistics (like sample mean).

# 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 $\varepsilon$, 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 $b$).

$\sigma$ 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:

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

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

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

and

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

[2] 95% of the time, these intervals contain $m$, $b$.

# Outline

Assuming the data does have a linear relationship

Measuring how well LSR line fits

## 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,

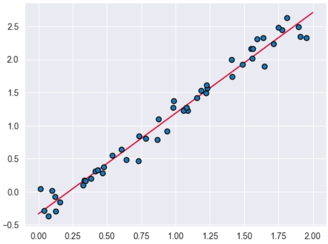
$$\text{RSE} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.$$

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

# Mean Squared Error, example

Recall, `'Example1.csv'` data. Its LSR line is

$$y = 1.520275x - 0.33458.$$

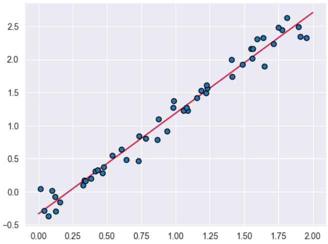


# Mean Squared Error, example

Recall, `'Example1.csv'` data. Its LSR line is

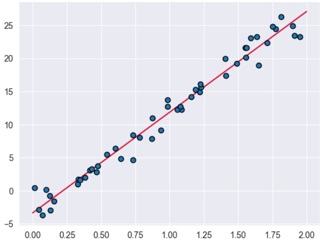
$$y = 1.520275x - 0.33458.$$

The MSE for this data and its LSR line is $\approx 0.0197$.
Does that mean that the linear model is a “good fit”?



# Mean Squared Error, scaling

What about the following data and its LSR 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 LSR 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

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

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

Then,

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

- $R^2$ 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?