Chapter 3

Linear Regression

STAT303-2

Northwestern | WEINBERG COLLEGE OF ARTS & SCIENCES

Simple Linear Regression - Recap

Training

In the last SLR lecture, we derived the optimum parameters for:

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Last week's focus

In terms of the data at hand:

$$X^1, X^2, \ldots X^N$$

$$Y^1, Y^2, \ldots Y^N$$

and used the formula with optimum parameters for prediction.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\boxed{RSS = e_1^2 + e_2^2 + \cdots + e_N^2}$$

Find RMSE/MAE for test and training data

One addition: Residual Standard Error (RSE)

$$\boxed{RSE = \sqrt{\frac{RSS}{n-2}}}$$

- A common alternate stat. measure
- Used for assessing training data

Simple Linear Regression - Inference

Today, we will focus more on the estimated function itself

$$\hat{Y} = \hat{f}(X)$$

- Straightforward for SLR
- Only important parameter: $\hat{\beta}_1$
  - The estimated slope
  - Determines the magnitude and direction of the linear association between X and Y
  - Change in Y = Change in X * $\hat{\beta}_1$

Look for insights:

- The parameters themselves
- How sure we are that the real function, f, is close to $\hat{f}$, and how close — Uncertainty
- How much $\hat{f}$ explains the variation in the data → Coefficient of Determination (R^2)

Simple Linear Regression - Uncertainty

Remember that the formula that **we assumed** to be linear

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

is just an **estimation** of the underlying function f. → Still assumed to be $\beta_0 + \beta_1 X$

- f has no way to be known or analytically found
- f is only observed through the collected data

Question: How to actually assess how close the estimation is?

We discussed RMSE/MAE to assess the results using the data – Can we do anything a little more analytic about the function?

Find the confidence intervals (CIs)

Simple Linear Regression - Uncertainty

- Recall that the residual sum of squares error is:

$$RSS = \left(Y^1 - \hat{\beta}_0 - \hat{\beta}_1 X^1\right)^2 + \left(Y^2 - \hat{\beta}_0 - \hat{\beta}_1 X^2\right)^2 + \cdots + \left(Y^N - \hat{\beta}_0 - \hat{\beta}_1 X^N\right)^2, \text{ or}$$

$$RSS = e_1^2 + e_2^2 + \cdots + e_N^2$$

- The coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimated by minimizing RSS
- However, the observations $(X^1, Y^1), (X^2, Y^2), \ldots (X^N, Y^N)$ correspond to a particular sample of observations
- For a different sample, the values of $(X^1, Y^1), (X^2, Y^2), \ldots (X^N, Y^N)$ are likely to be different, which implies that $e_1^2, e_2^2, \ldots, e_n^2$ are likely to be different
- Thus, the uncertainty in the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is due to the uncertainties in the residuals $e_1^2, e_2^2, \ldots, e_n^2$

Simple Linear Regression - Uncertainty

- Assuming the variance of the residuals to be a constant, i.e.,

$$Var(\epsilon) = \sigma^2,$$

it can be shown that the uncertainties in the coefficients $\widehat{\beta}_0$ and $\hat{\beta}_1$ can be estimated as:

$$SE\left(\hat{\beta}_0\right)^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] \qquad SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Note that σ^2 is unknown.
- However, it can be estimated from the data as:

$$\sigma^2 = \sqrt{\frac{RSS}{N-2}}$$

Simple Linear Regression - Uncertainty

$$SE\left(\hat{\beta}_0\right)^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{X}^2}{\sum_{i=1}^{N}(X^i - \bar{X})^2}\right] \qquad SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(X^i - \bar{X})^2}$$

where $\sigma^2 = Var(\epsilon) = \sqrt{\frac{RSS}{n-2}}$

- The standard errors can be used to compute **the confidence intervals**. (CIs)
- 95% CI → A range of values that contains the true unknown value of the parameter with 95% probability.
- 95% CI of a variable $= \text{ variable} \pm\ 2 * SE(\text{variable})$
- For SLR:

$$\hat{\beta}_1 \pm\ 2 * SE(\hat{\beta}_1)$$

$$\hat{\beta}_0 \pm\ 2 * SE(\hat{\beta}_0)$$

- We are 95% sure that the real intercept (β_0) and slope (β_1) values of the underlying function, are within the ranges of:

$$\left.\begin{array}{l} [\hat{\beta}_0 - 2 * SE\left(\hat{\beta}_0\right), \hat{\beta}_0 +\ 2 * SE\left(\hat{\beta}_0\right)] \\ [\hat{\beta}_1 - 2 * SE\left(\hat{\beta}_1\right), \hat{\beta}_1 +\ 2 * SE\left(\hat{\beta}_1\right)] \end{array}\right\}$$

95% CIs
for SLR

Simple Linear Regression - Uncertainty

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{X}^2}{\sum_{i=1}^{N}(X^i - \bar{X})^2}\right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(X^i - \bar{X})^2}$$

where $\sigma^2 = Var(\epsilon) = \sqrt{\frac{RSS}{n-2}}$

$$\left.\begin{matrix}[\hat{\beta}_0 - 2 * SE(\hat{\beta}_0), \hat{\beta}_0 + 2 * SE(\hat{\beta}_0)] \\ [\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]\end{matrix}\right\}$$

95% CIs for SLR

- The ols function from statsmodels calculates the CIs
- We still need to interpret them
  - Higher RSS – higher standard errors, wider CI, less certainty
  - Lower variation in X – higher standard errors , wider CI, less certainty
  - Higher variation in X – lower standard errors, tighter CI, more certainty

Simple Linear Regression - Uncertainty

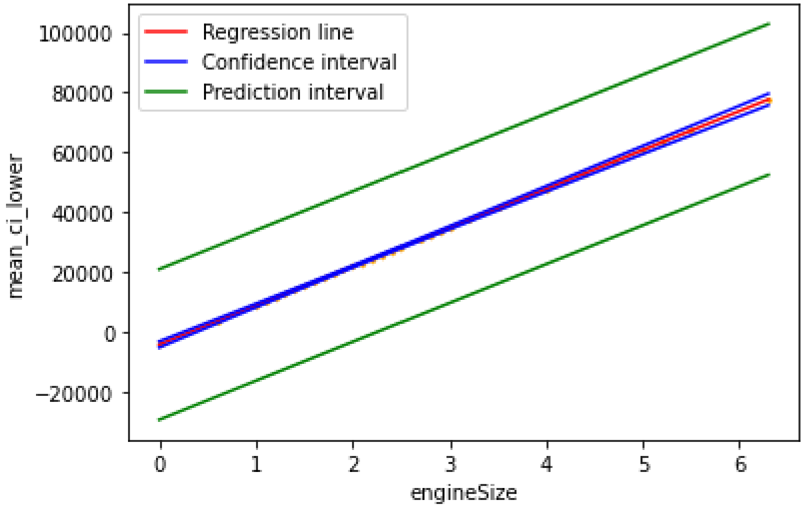
Very similar to the confidence intervals, there is also **prediction intervals.**

- A more conservative approach to uncertainty, that also takes into account the irreducible error in statistical modeling.
- Recall from Chapter 2 that:

$$E\left(Y-\hat{Y}\right)^2 = \underbrace{\left[f(X)-\hat{f}(X)\right]^2}_{\text{Reducible error}} + \underbrace{Var(\in)}_{\text{Irreducible error}}$$

- For this reason, the prediction interval of a Linear Regression model is always wider than its confidence interval.

Simple Linear Regression - Uncertainty



- In this course, you only need to know the prediction interval conceptually – as a more conservative approach that takes $Var(\in)$ into account.
- A model created by ols function in Python will easily return both prediction and confidence intervals

Simple Linear Regression - Uncertainty

Another important application for Standard Errors: **Hypothesis Testing**

- A purely statistical concept – covered in more detail in Statistics courses
- In Hypothesis testing, the starting point is a Null Hypothesis (**H_0**) and an Alternate Hypothesis (**H_A**)

$\mathbf{H_0}: \beta_1 = 0$ ⟶ X and Y are unrelated

$\mathbf{H_A}: \beta_1 \neq 0$ ⟶ X and Y are related

The whole point of Hypothesis testing:

- Using the data to calculate a probability
- That probability of the calculated nonzero $\hat{\beta}_1$ being due to random chance
- This probability is also called the **p-value**.

To find the probability, we first need to calculate a value called the t-statistic.

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- t is a random variable that belongs to a distribution, called t-distribution.
- You can calculate the t-statistic plugging in $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$
- You can then check (using Python) where the t-statistic falls in the distribution and find the **probability** of that value t-statistic or a higher/lower value occurring. (Again, using Python)

Simple Linear Regression - Uncertainty

To find the probability, we first need to calculate a value called the t-statistic.

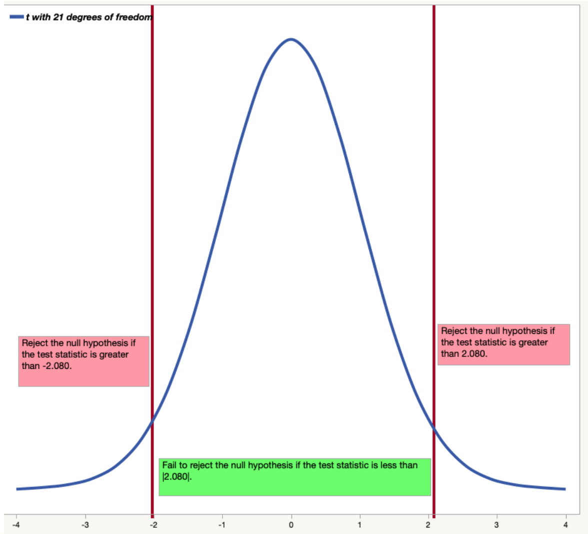
$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- t is a random variable that belongs to a distribution, called t-distribution.
- You can calculate the t-statistic plugging in $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$
- You can then check (using Python) where the t-statistic falls in the distribution and find the **probability** of that value of the t-statistic or a higher/lower value occurring. (again, using Python)

- This probability is also called the **p-value**.
  - If p-value is below 0.05 (5%) the Null Hypothesis (**$\mathbf{H_0}$**) is rejected.
  - That means the non-zero linear association between X and Y is **statistically significant**.

$\mathbf{H_0}: \beta_1 = 0$ ⟶ X and Y are unrelated.

$\mathbf{H_A}: \beta_1 \neq 0$ ⟶ X and Y are related.



In plain English, using $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$, we can figure out how sure we are that there is an actual linear relation between X and Y.

- $\hat{\beta}_1$ is still an estimation of that linear relation.
- How good is $\hat{\beta}_1$? That is determined by RMSE/MAE and another metric - R^2 (up next)

Simple Linear Regression - R^2

Question: How much does $\hat{f}$ explain the variation in the data?

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

- The responses (Y's) in the data has a certain variation. → Quantified by Var(Y)
- Using X's and the optimum parameters, ($\hat{\beta}_0$ and $\hat{\beta}_1$) how much of this variance can the model explain/capture?

$N * var(Y)$

- Let's start with the formula for the variation in Y, what we try to explain:

$$\underbrace{\sum_{i=1}^{N}(Y^i - \bar{Y})^2}_{} = \underbrace{\sum_{i=1}^{N}(Y^i - \hat{Y}^i)^2}_{} + \underbrace{\sum_{i=1}^{N}(\hat{Y}^i - \bar{Y})^2}_{}$$

- Total sum of squares (TSS)
- Total variation

- Residual sum of squares (RSS)
- Unexplained variation

- Explained sum of squares (RSS)
- Explained variation

Simple Linear Regression - R^2

Question: How much does $\hat{f}$ explain the variation in the data?

R^2 is the metric that quantifies this question by the ratio of the explained variation to the total variation.

$$R^2 = \frac{Explained\ variation}{Total\ variation} = \frac{Total\ variation - Unexplained\ variation}{Total\ variation} = 1 - \frac{Unexplained\ variation}{Total\ variation}$$

$$\sum_{i=1}^{N}(Y^i - \bar{Y})^2 = \sum_{i=1}^{N}(Y^i - \hat{Y}^i)^2 + \sum_{i=1}^{N}(\hat{Y}^i - \bar{Y})^2$$

- Total sum of squares (TSS)
- Total variation

- Residual sum of squares (RSS)
- Unexplained variation

- Explained sum of squares (RSS)
- Explained variation

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{N}[Y^i - \hat{Y}^i]^2}{\sum_{i=1}^{N}[Y^i - \bar{Y}]^2}}$$

Simple Linear Regression - R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{N}[Y^i - \hat{Y}^i]^2}{\sum_{i=1}^{N}[Y^i - \bar{Y}]^2} = 1 - \frac{RSS}{TSS}$$

- TSS can be calculated before the training/prediction.
- RSS is calculated after the prediction – we need the $\hat{Y}$'s.

statsmodels table

- So, after a model and trained and the results are returned,
    - RMSE/MAE – the numeric assessment of the prediction accuracy → Prediction Task
    - R^2 – the numeric assessment of how well the function is fit → Inference Task
        - A ratio – between 0 and 1
        - **Also, R^2 is the square of Pearson correlation coefficient.**

Simple Linear Regression - R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left[Y^i - \hat{Y}^i\right]^2}{\sum_{i=1}^{N}\left[Y^i - \bar{Y}\right]^2} = 1 - \frac{RSS}{TSS}$$

- TSS can be calculated before the training/prediction.
- RSS is calculated after the prediction – we need the $\hat{Y}$’s.

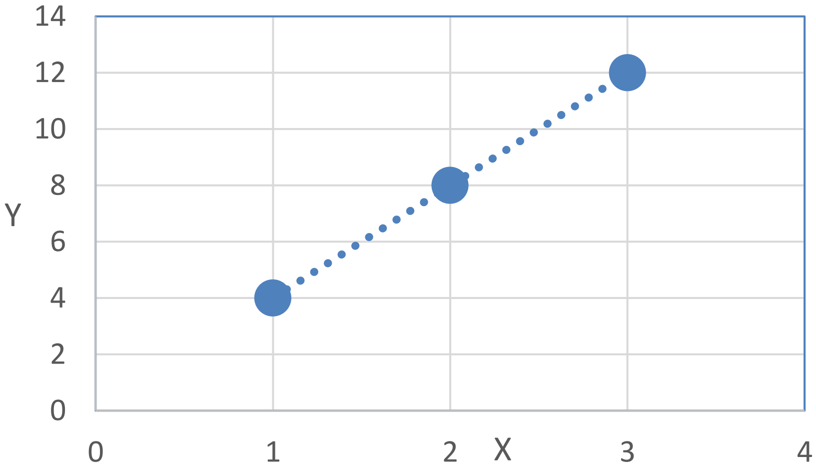
- So, after a model and trained and the results are returned,
  - RMSE/MAE – the numeric assessment of the prediction accuracy
  - R^2 – the numeric assessment of how well the function is fit
    - A ratio – between 0 and 1
    - **Also, R^2 is the square of Pearson correlation coefficient.**

- Ideal scenario: RMSE/MAE as low as possible and R^2 as close to 1 as possible.
- Note that RMSE/MAE is a continuous number
  - Its assessment is subjective
  - It can be too high, tolerable or good, depending on the prediction task.
- On the other hand, R^2 is a ratio – a more objective metric
  - A low RMSE/MAE can still end up with a relatively low R^2 if TSS is low

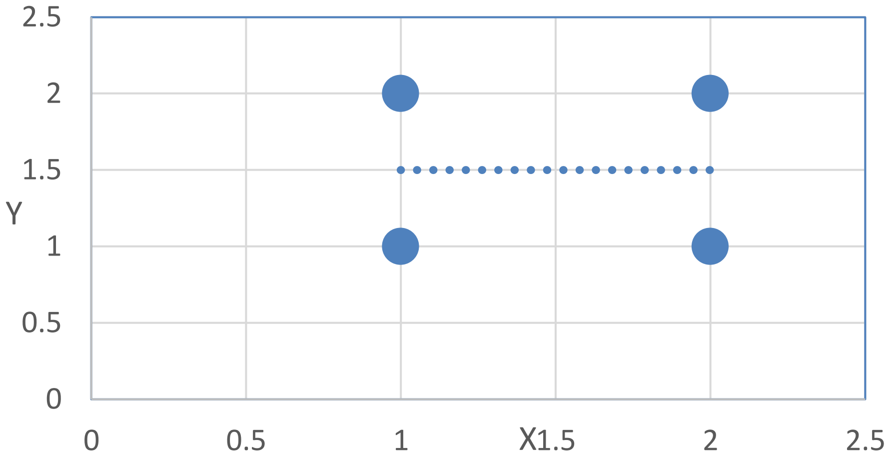
- If you are after accurate predictions: RMSE/MAE
- If you are after a good statistical explanation between X and Y: R^2
- Usually, a low RMSE/MAE and a high R^2 go hand-in-hand but not always

Simple Linear Regression - R^2

Visualization with two toy models: Extreme cases



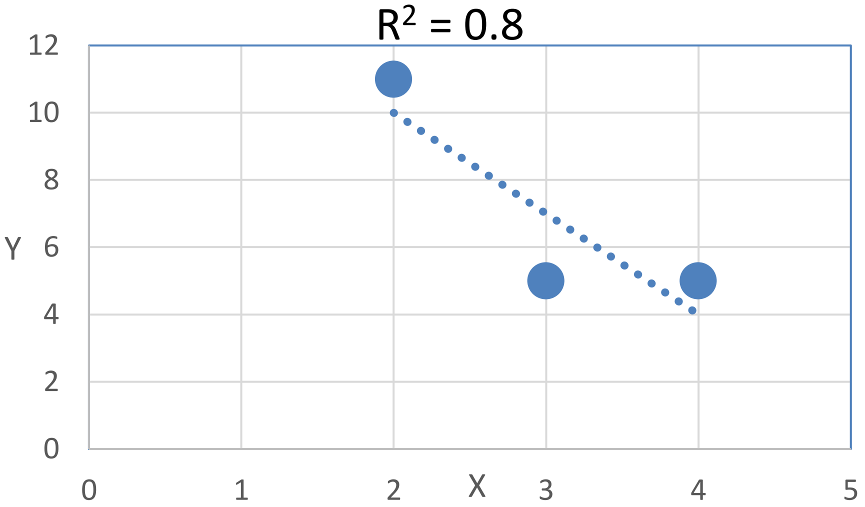
$R^2 = 1$

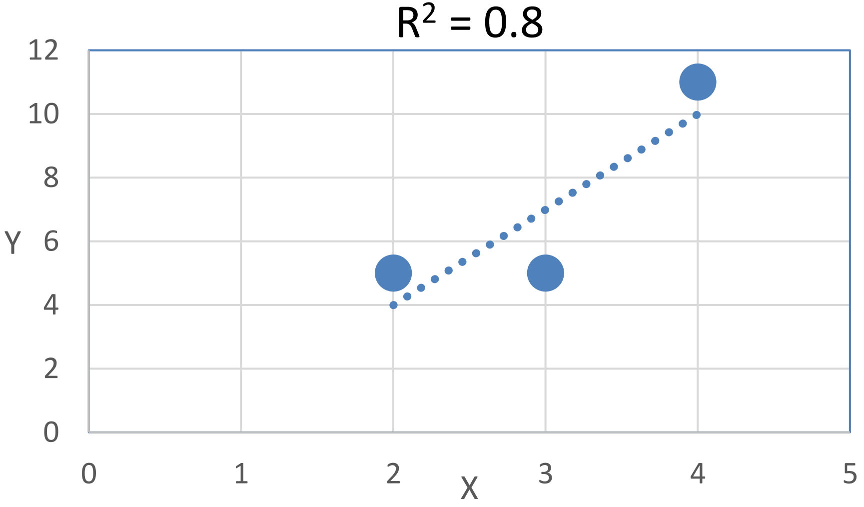


$R^2 = 0$

Simple Linear Regression - R^2

Two more toy models: Same R^2 with different slope





To wrap up:

A Simple Linear Regression (SLR) model: $Y = f(X) = \beta_0 + \beta_1 X$

One-line Python implementations:

```
ols_object = smf.ols(formula = 'price~engineSize', data = train)
```

Training:

- Find $\hat{\beta}_0$ and $\hat{\beta}_1$(using training data and the equations)
- Calculate **RSE for training data**

```
model = ols_object.fit()
```

```
np.sqrt(model.mse_resid)
```

Prediction:

- Using $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, find the predictions for test data
- Find training and **test RMSE/MAE**
- Compare test RMSE/MAE with training RMSE/MAE/MSE for **overfitting/underfitting**

```
pred_price = model.predict(testf)
```

```
np.sqrt(((testp.price - pred_price)**2).mean())
```

Inference:

- Find **95% confidence intervals** for $\hat{\beta}_0$ and $\hat{\beta}_1$
- Find the **prediction intervals**
- Find the t-statistic and p-value for **statistical significance**
- Find the **R^2 value.**

```
model.summary()
```

```
intervals = model.get_prediction(testf)
```

```
intervals.summary_frame(alpha=0.05)
```