Machine Learning (ML)

Chapter 3:

Linear Regression and assessing the accuracy of Models

Saeed Saeedvand, Ph.D.

Outline

In this Chapter:

- ✓ Linear regression
- ✓ Standard Error Evaluation
- ✓ Confidence Intervals
- ✓ Hypothesis testing
- ✓ Model Overall Accuracy

Aim of this chapter:

- ✓ Understanding the undelaying concepts of a linear regression and its objectives. Then evaluating models accuracy from different perspectives.

What is the Linear regression?

- ✓ A statistical method that models the **relationship between two variables** (X and Y)
- ✓ We assuming there is a **linear relationship** between variables (for now).
- ✓ Find the best-fit line that describes this relationship.
- ✓ Can be used to make predictions about Y for a given value of X.

Linear regression is often considered a simple statistical technique, but it holds **significant conceptual importance** and it has practical value.

Linear regression

Questions of Relationship between data

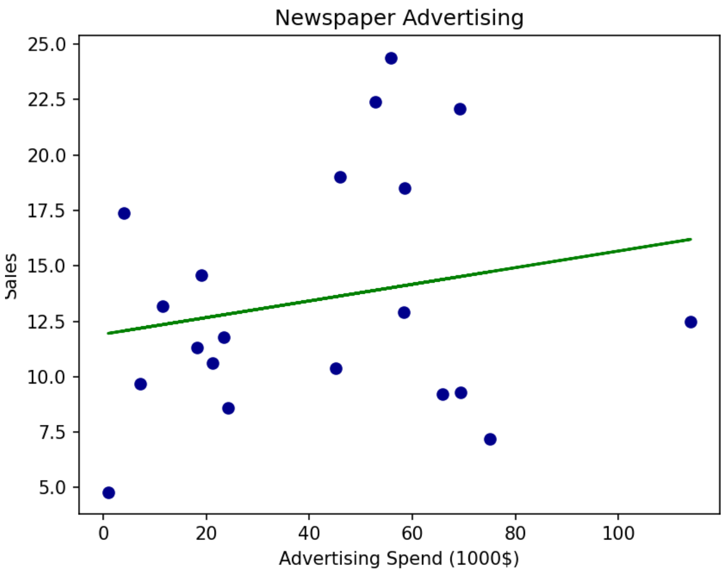
We may ask several questions when examining the relationship between for instance advertising and sales, such as:

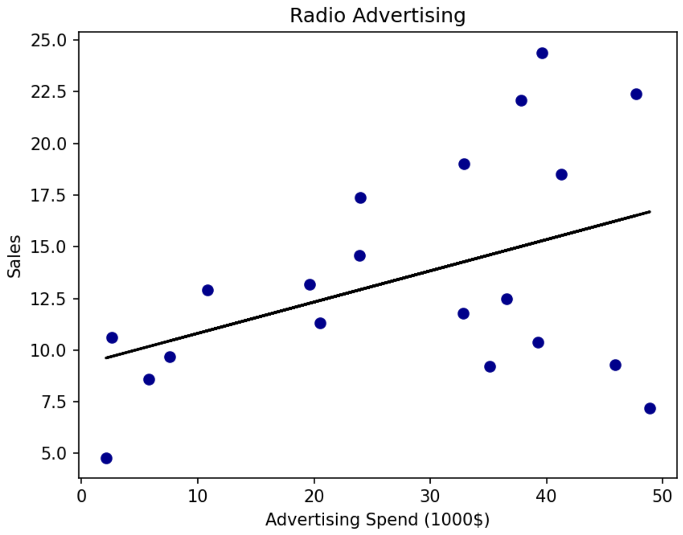
- Does a correlation **exist** between advertising budget and sales?
- How strong is the **correlation** between advertising budget and sales?
  - E.g. one variable increases if the other increases
- **Which media** channels have an impact on sales?
- **How precise** our sales predictions can be?
- Is the **relationship** between advertising and sales **linear**?
- Do the **advertising media** channels work better together or better individually?

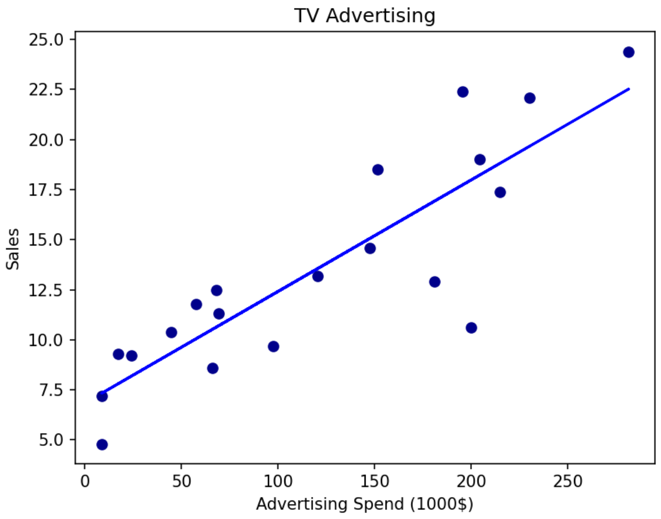
Linear regression

Example

- ✓ Sales if we do advertisements on TV, Radio and Newspaper.
- ✓ The lines are linear-regression fit to each.







$$\text{Sales} \approx f(\text{Newspaper}, \text{Radio}, \text{TV})$$

Linear regression- Definition

What is the Dependent variable:

✓ The dependent variable is the variable that we want to predict or explain using the independent variable(s).

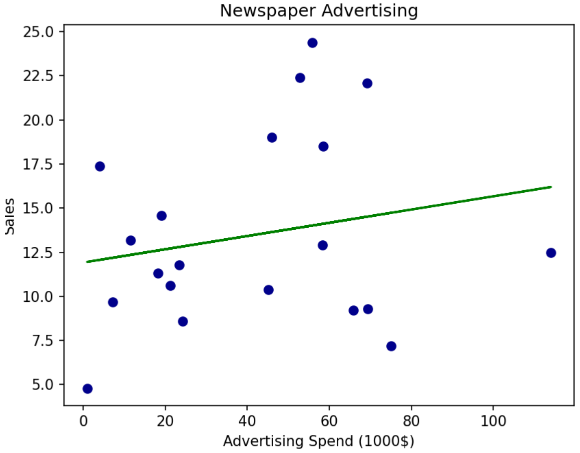
What is the Independent variable:

✓ A variable that we believe may have an effect on the dependent variable.

Example:

✓ The amount of money spent on advertising is the independent variable.

✓ Sales revenue is the dependent variable.

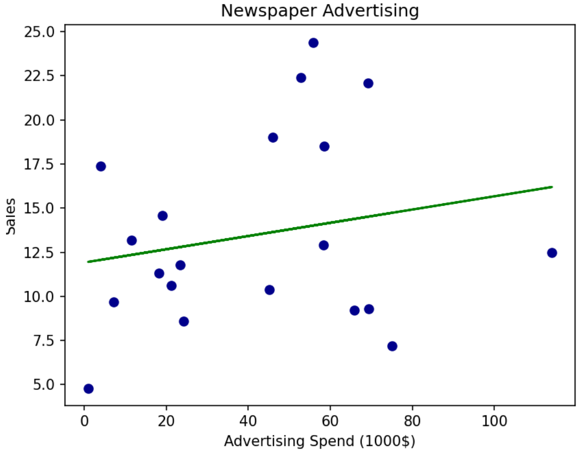


Note: When we include independent variables in a regression model, we want to model the relationship between the independent and dependent variables.

Linear regression

- ✓ If we assume a model as follows:

$$Y = \underset{\textit{Intercept}}{\boxed{\theta_0}} + \overset{\textit{slope}}{\boxed{\theta_1 X}} + \varepsilon$$



- ✓ We can call *intercept* and *slope* as *coefficients* or *parameters*.
- ✓ If want to write the prediction for that model we can write (ignoring error):

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 X$$

Prediction of Y for the $X = x$

Linear regression

- ✓ We can write the prediction for Y based on the $\boldsymbol{i^{th}}$ **value** of X.

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$$

What is the Residual?

- ✓ The difference between the actual observed value of a dependent variable and the predicted value of that variable (from a model).

$$\underbrace{e_i}_{\text{Residual}} = y_i - \hat{y}_i$$

Linear regression

Residual Sum of Squares (RSS)

- ✓ The difference between the actual observed value of a dependent variable and the predicted value of that variable from a model.

$$RSS = (y_1 - \overset{\hat{y}}{(\hat{\theta}_0 + \hat{\theta}_1 x_1)})^2 + (y_2 - (\hat{\theta}_0 + \hat{\theta}_1 x_2))^2 + \ldots + (y_n - (\hat{\theta}_0 + \hat{\theta}_1 x_n))^2$$

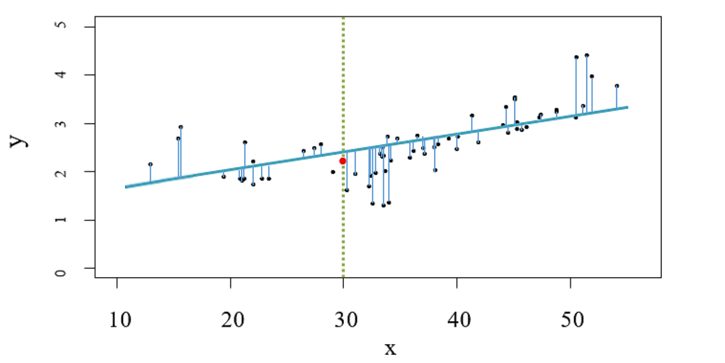
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

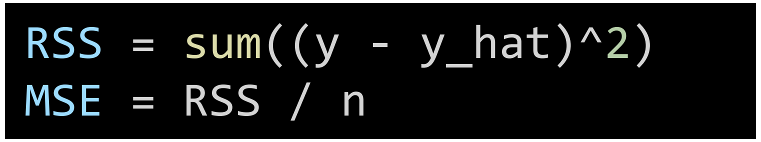
With minimizing the RSS, we can effectively find the "best" line (or curve) that fits the data (we need two parts).

Remainder (The Model’s Accuracy)

- ✓ Mean Square Error (MSE):

$$MSE_{Train} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \underbrace{\hat{f}(x_i)}_{\hat{y}_i}]^2, n = |train|$$





```
RSS = sum((y - y_hat)^2)
MSE = RSS / n
```

Linear regression **(Objective)**

Minimize the RSS:

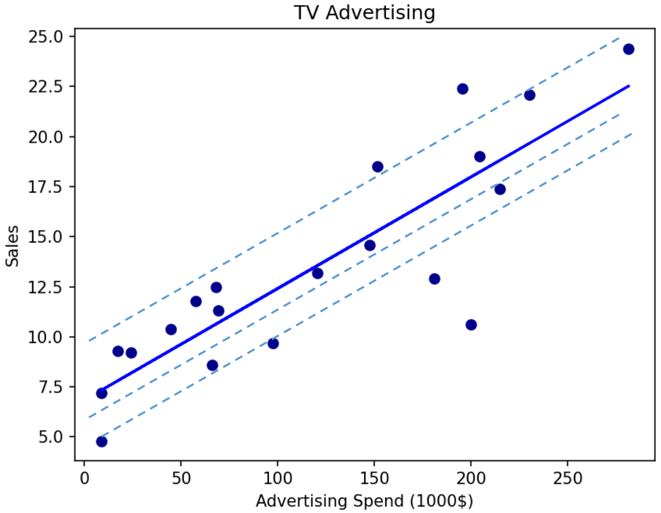
- ✓ The minimizing values can be shown (first part):

Intercept error

$$\hat{\theta}_0 = \boxed{\frac{1}{n}\sum_{i=1}^{n} y_i} - \hat{\theta}_1 \bar{x}$$

$\bar{y}$

n: Total number of observations in the dataset



Note: measurement of the linear regression line accuracy on the y-axis, (intercept).

Linear regression **(Objective)**

Minimize the RSS:

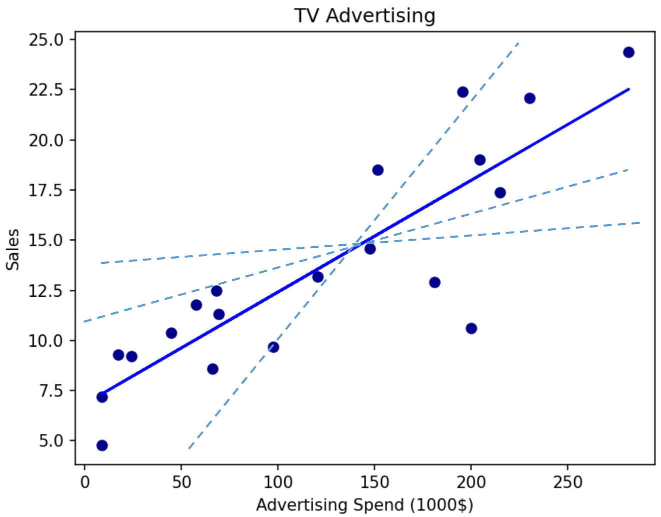
✓ The minimizing values can be shown (second part):

Slope error

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Estimating the value of the slope coefficient

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Linear regression (Objective)

Minimize the RSS:

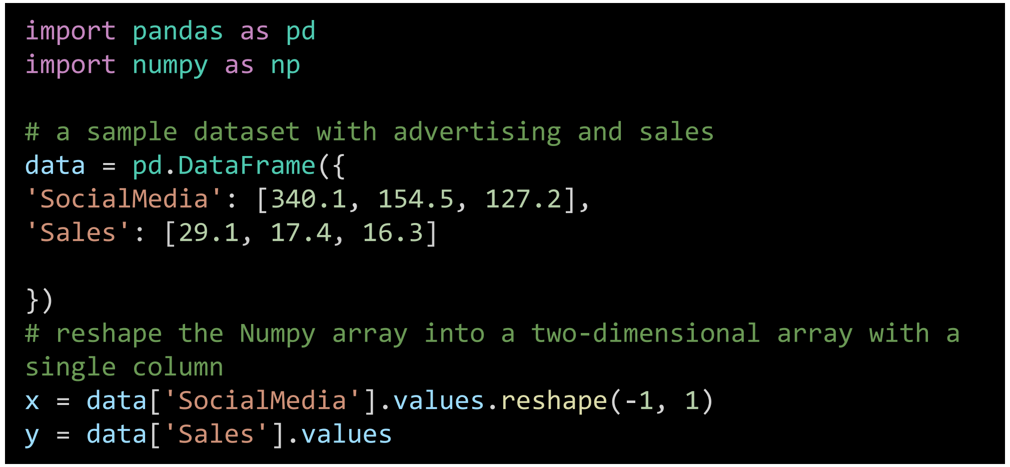
Slope error

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Intercept error

$$\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\theta}_1 \bar{x}$$

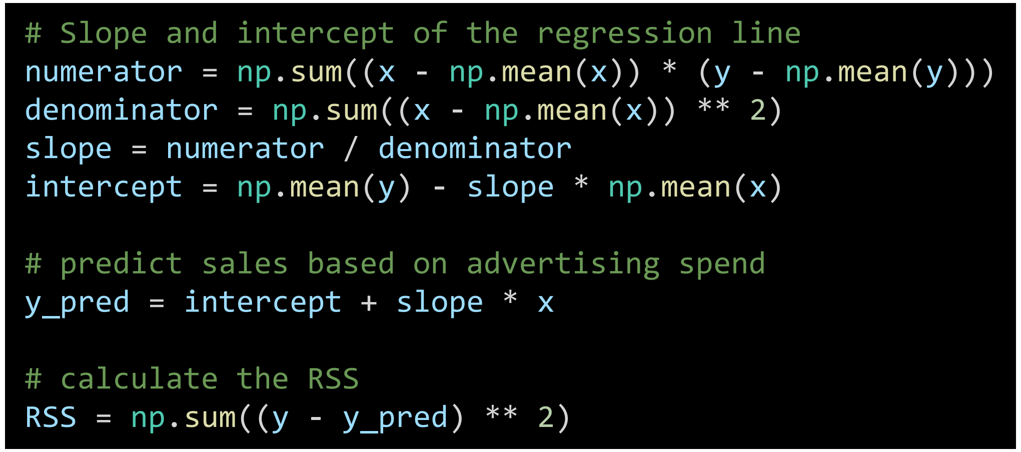
$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



```
import pandas as pd
import numpy as np

# a sample dataset with advertising and sales
data = pd.DataFrame({
'SocialMedia': [340.1, 154.5, 127.2],
'Sales': [29.1, 17.4, 16.3]

})
# reshape the Numpy array into a two-dimensional array with a
single column
x = data['SocialMedia'].values.reshape(-1, 1)
y = data['Sales'].values
```



```
# Slope and intercept of the regression line
numerator = np.sum((x - np.mean(x)) * (y - np.mean(y)))
denominator = np.sum((x - np.mean(x)) ** 2)
slope = numerator / denominator
intercept = np.mean(y) - slope * np.mean(x)

# predict sales based on advertising spend
y_pred = intercept + slope * x

# calculate the RSS
RSS = np.sum((y - y_pred) ** 2)
```

Linear regression (**Objective**)

Minimize the RSS:

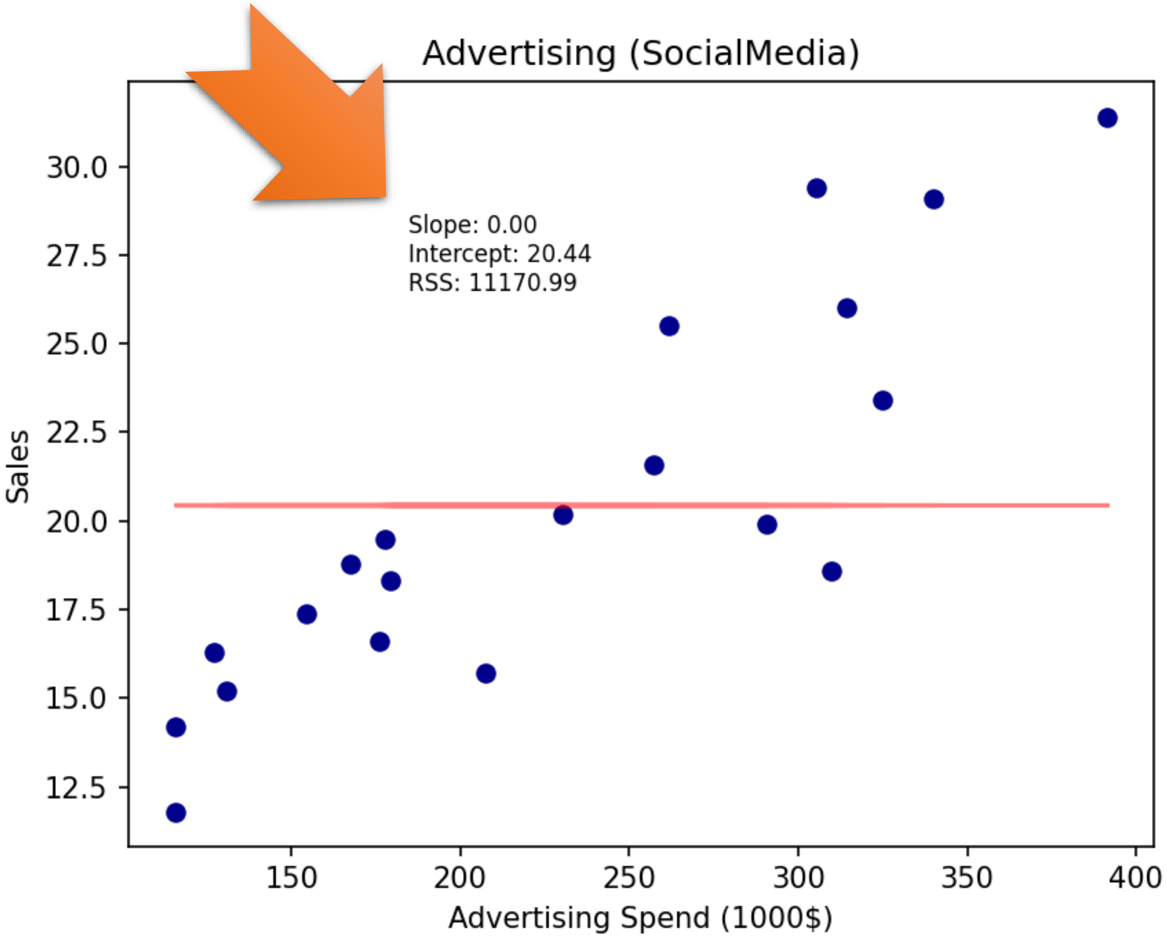
Slope error

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Intercept error

$$\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\theta}_1\bar{x}$$

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



Linear regression **(Objective)**

Minimize the RSS:

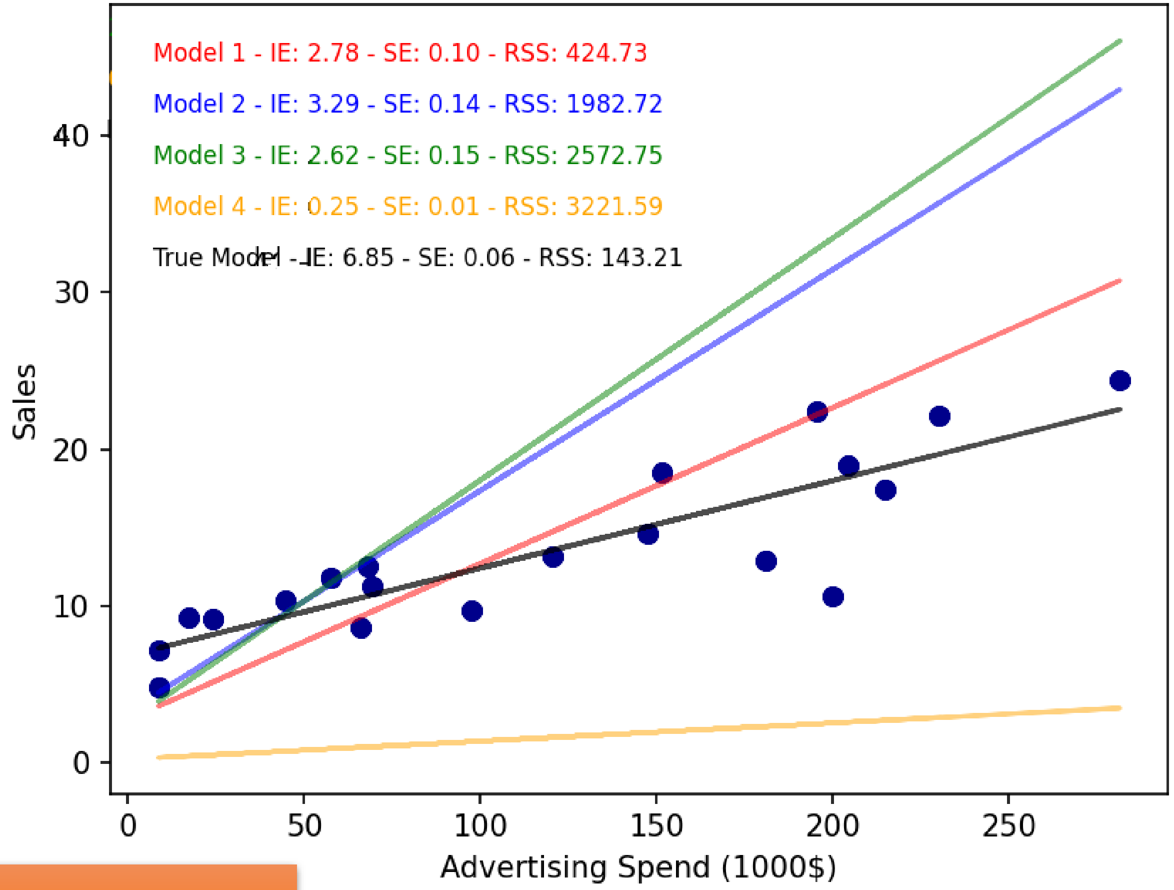
The goal of linear regression

To find the best values of:

✓ $\hat{\theta}_0$ **Intercept Error (IE)**

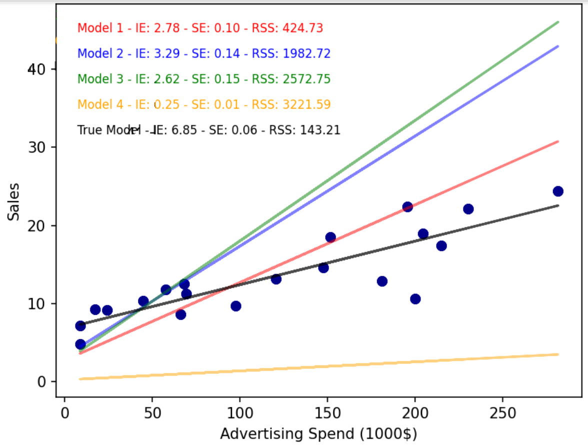
✓ $\hat{\theta}_1$ **Slope Error (SE)**

That minimize the RSS equation.



How to do this minimization?

Linear regression **(Objective)**



Minimize the RSS:

This can be done using various optimization algorithms such as:

✓ **Normal Equations**

- Closed-form solution (exact solution)
- Good choice when the number of features is small (computationally expensive)

✓ **Matrix inversion**

- Closed-form solution
- Good choice when the number of features is small (computationally expensive)
- Can be numerically unstable (small change in the input of the problem leads to a large change in the output e.g. fining roots of a high-degree polynomial)

✓ **Gradient descent**

- Iterative optimization algorithm used to find the minimum of a function
- Commonly used in machine learning and other optimization problems

Standard Error

Standard Error Evaluation

How precise the estimates of those coefficients are?

- ✓ RSS tells us how well the model parameters (e.g. regression) fits the data.
- ✓ With Standard Error of the regression coefficients (intercept and slope) we can say how precise the estimates of those coefficients are.
- ✓ Both RSS and standard error of regression coefficients can assess model performance in gradient descent regression.
- ✓ Both measures are important for evaluating model quality and inferring independent or dependent variable relationships.

Standard Error

Standard Error (SE) Evaluation

- ✓ Standard Error (SE) is a measure of the precision or accuracy of an estimate.

$$SE(\hat{\theta}_1) = \frac{var(\varepsilon)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- ✓ Usually SE is estimated using the **RSS** and the **Degrees of Freedom (DoF)** of the model.

$$var(\varepsilon) = \text{sqrt(RSS / (n−2))}$$

Number of independent variables (slopes) plus one (for the intercept)

Number of samples in dataset (DoF)

Standard Error

Standard Error (SE) Evaluation

✓ Standard error equation for the intercept term in a simple linear regression model:

$$SE(\hat{\theta}_0) = var(\varepsilon)\sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]}$$

- SE of intercept **estimates** variability of intercept term in population of all possible samples.
- Provides **information on** expected variation of estimated intercept from sample to sample.

Standard Error

Sampling:

✓ Randomly selecting a subset of observations from the available dataset (also known as train and test set:

- In some cases, a **75/25 split** might be appropriate.
- In others a **90/10 split** or even **50/50 split** might be more appropriate.

✓The **train subset** is then used to **estimate the regression parameters**, such as the intercept and slope.

✓ Important Note: for standard error and confidence intervals (next slide) calculations we usually use **only** the **entire sample of data**.

Standard Error Evaluation

Confidence Intervals

✓Expressing the uncertainty in an estimate by providing a range of plausible values for the true population parameter.

✓**Constructed** based on the standard error of the estimate.

✓Reflects the variability of the estimate across repeated samples.

Standard Error Evaluation

Confidence Intervals

- ✓ We use Confidence Interval with the **slope parameter $\hat{\boldsymbol{\theta}}_1$** in linear regression.

We can say 95% confidence interval

- ✓ If we do repeat our sampling (create new train set) and then create a new regression model and perform analysis many times:
  - We expect the all the slope parameter to fall within this range in **95% of the samples**:

$$range = [\hat{\theta}_1 - 2.SE(\hat{\theta}_1), \hat{\theta}_1 + 2.SE(\hat{\theta}_1)]$$

$$= \hat{\theta}_1 \pm 2.SE(\hat{\theta}_1)$$

What if does not fall in range?

Standard Error Evaluation

Confidence Intervals

$$range = [\hat{\theta}_1 - 2.SE(\hat{\theta}_1), \hat{\theta}_1 + 2.SE(\hat{\theta}_1)]$$

$$= \hat{\theta}_1 \pm 2.SE(\hat{\theta}_1)$$

What if does not fall in range?

✓ It means that maybe one or all of following problems existing:

- Model may be is not good fit for the data (**estimate of the slope parameter)!**
- Sampling approach was not good!
- There are other factors affecting the relationship between the variables that we did not account for in our model!

Testing Hypothesis or validity of a claim

Hypothesis testing

- ✓ A statistical method to evaluate if a hypothesis about a population parameter is true (based on a sample of data)

- ✓ There are many different Hypothesis we can test, for instance:
  - E.g. the correlation between two variables is positive.
  - E.g. the mean height of a population is 180 cm.
  - E.g. the proportion of males in a population is 60%.
  - …

Testing Hypothesis or validity of a claim

Hypothesis testing

✓ Can be used to test various hypotheses:

- **Relationship** between input and output variables in a regression model.
- **Relationships** between different input variables.

✓ The most common Hypothesis test is formulating a null hypothesis and an alternative hypothesis.

✓ We can use Standard error for Hypothesis testing.

Relationship between multiple independent variables

Null hypothesis

H_0

✓ Null hypothesis assumes there is no relationship between X and Y:

$$H_0 \colon \hat{\theta}_1 = 0$$

H_A

✓ There is some relationship between X and Y (called alternative hypothesis):

$$H_A \colon \hat{\theta}_1 \neq 0$$

Proof

1. If $\hat{\theta}_1 = 0$, then the model form $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 X$ can be simplified to Y = $\hat{\theta}_0$
2. Then Y = $\hat{\theta}_0$ indicates that there is no relation between X and Y.

Relationship between multiple independent variables

How to check Null hypothesis?

- ✓ Simply checking $\hat{\theta}_1 = 0$ for the null hypothesis cannot be sufficient to say no relationship between X and Y, in all cases (it is not valid).
- ✓ **The reason:**
  - ➢ $\hat{\theta}_1$ depends on the sample data, and there may be some variability in the estimated value due to sampling error (not good sampling).

What is the solution then?

- ➢ **Using t-statistics:** Calculate the t-statistic using the estimated coefficient and its standard error.
- ➢ **Using confidence interval:** (similar to what we studied but for hypothesis testing).

Relationship between multiple independent variables

t-statistics

- ✓ t-statistic is a way of quantifying the strength of evidence against the null hypothesis (in favor of the alternative hypothesis).
- ✓ To test the null hypothesis we compute the t-statistic as follows:

$$\boxed{|t|} = \frac{\hat{\theta}_1 - 0}{SE(\hat{\theta}_1)}$$

This is how we present t-statistic usually

Relationship between multiple independent variables

t-statistics

$$|t| = \frac{\hat{\theta}_1 - 0}{SE(\hat{\theta}_1)}$$

✓ A larger t-statistic indicates stronger evidence against the null hypothesis.

- **If accept null hypothesis:** there is not enough evidence to suggest that some relationship between X and Y exists (between the independent variable(s) and the dependent variable)

✓ A smaller t-statistic suggests weaker evidence against the null hypothesis.

- **Reject the null hypothesis**: accept the alternative hypothesis (relationship exists).

Note: t-statistic alone still does not provide complete information about the significance of the coefficient.

We need to calculate the corresponding p-value to determine the level of significance.

Relationship between multiple independent variables

p-value

- ✓ We need to **calculate Probability** of observing any value equal to t-statistics or larger.
- ✓ **Two types** of tests (one-tailed or two-tailed) can be chosen based on the input data, depending on whether we want to **check** for a one-directional or two-directional relationship. The corresponding null and alternative hypotheses can then be formulated afterwards.
- ✓ The **p-value** is calculated using the t-statistic and the DoF of the samples.
- ✓ A p-value of 0.05 (or 5%) is often used as a threshold for statistical significance.
- ✓ If the p-value is less than or equal to 0.05, then the null hypothesis is rejected in favor of the alternative hypothesis. (sign of there is relationship).

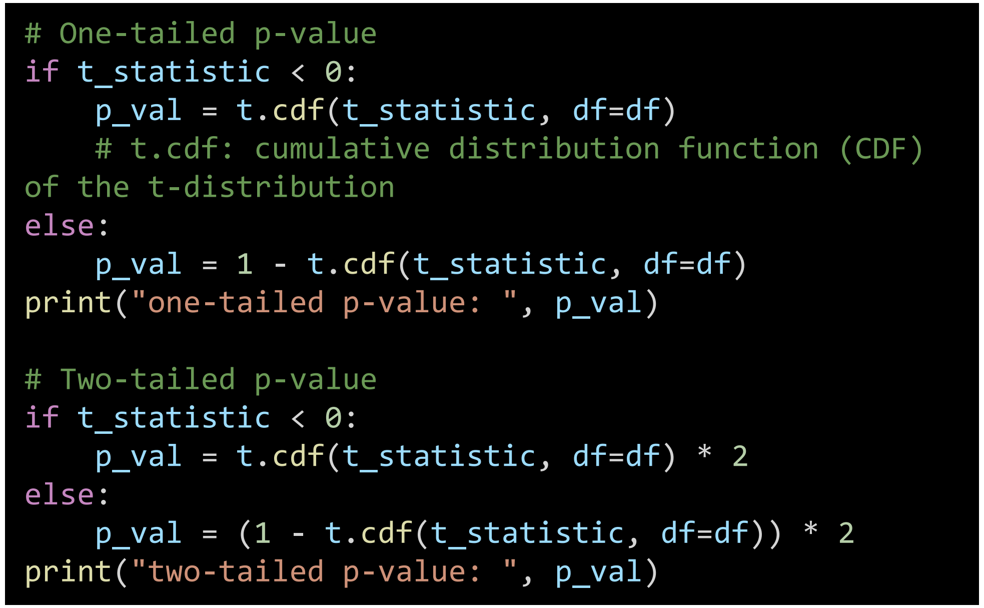
In general p-value indicates the **strength of evidence** against the null hypothesis provided by the sample data

Relationship between multiple independent variables

p-value

How to calculate?

✓ It is important to note that statistical software packages, such as **Python** or R, can perform these calculations **automatically**.



```
# One-tailed p-value
if t_statistic < 0:
    p_val = t.cdf(t_statistic, df=df)
    # t.cdf: cumulative distribution function (CDF)
of the t-distribution
else:
    p_val = 1 - t.cdf(t_statistic, df=df)
print("one-tailed p-value: ", p_val)

# Two-tailed p-value
if t_statistic < 0:
    p_val = t.cdf(t_statistic, df=df) * 2
else:
    p_val = (1 - t.cdf(t_statistic, df=df)) * 2
print("two-tailed p-value: ", p_val)
```

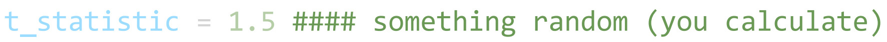
The steps to calculate the p-value from a t-statistic:

1. Determine the degrees of freedom (df) for the t-distribution: df = n - 1.
2. Look up the t-distribution table to find the probability associated with the t-statistic at the given degrees of freedom and level of significance.
3. For a one-tailed test: compare the calculated probability to the level of significance (α) for the test. If the calculated probability is less than α, reject the null hypothesis. If the calculated probability is greater than α, fail to reject the null hypothesis.
4. For a two-tailed test: multiply the calculated probability by 2 to obtain the p-value. Then, compare the p-value to the level of significance (α) for the test. If the p-value is less than α, reject the null hypothesis. If the p-value is greater than α, fail to reject the null hypothesis.

Refer to book: "SticiGui: Statistical Tools for Internet and Classroom Instruction with a Graphical User Interface (GUI)“, **available online**. (chapter 7)

Python Example

Practice: In the code calculate t-statistics.



```
t_statistic = 1.5 #### something random (you calculate)
```

Model Overall Accuracy

What is the Motivation for Overall Accuracy?

- ✓ We need to show models’ overall accuracy to assess total accuracy.
- ✓ Showing different accuracy in different ways at the same time provides a more complete picture of how well the model is performing.
- ✓ **Each method** may capture different aspects of the model's performance.
- ✓ **Comparing** the results from **different methods** can help to identify potential issues or areas **for improvement** in the model.

Model Overall Accuracy (Metrics)

Different metrics

- ✓MSE (Mean Squared Error)
- ✓RSE (Residual Standard Error)
- ✓R-squared (R^2)
- ✓TSS (Total Sum of Squares)
- ✓F-statistic
- ✓…

Assignment

Find publications (since 2023) that used these metrics to show their results first and then interpret exactly the relevant papers' results. (you may need to find more than one paper).
https://scholar.google.com/

Model Overall Accuracy

Residual Standard Error (RSE)

- ✓ It is standard deviation of the residuals (errors) in a regression model.
- ✓ A measure of how well a regression model fits the data

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

Number of samples in dataset (DoF)

residual sum-of-squares
(we already have seen it)

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

- ✓ A smaller RSE indicates a **better fit** of the model to the data.

Model Overall Accuracy

Total Sum of Squares (TSS)

- ✓ Represents the difference between the observed values of Y and their mean.
- ✓ Measure of **amount of variation** (spread of samples on Y) that exists in the dependent variable Y in the entire dataset.

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

Mean

- ✓ A larger value of TSS is generally desirable, it is not the only factor that determines the **accuracy or quality** of a regression model.

Model Overall Accuracy

R-squared (R^2)

- ✓ Statistical metric that is used to evaluate the goodness of fit of a regression model.
- ✓ How much of the variation in Y is explained by the variation in X that is included in the model.

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

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

Mean (labeled data)

- ✓ R-squared value ranges between 0 to 1 with higher values indicating a better fit.
- ✓ It indicates that the model explains all of the variance (does not necessarily mean overfit if 1).
- ✓ Does **not necessarily mean** that the **model is a good fit** and it is important to use other metrics beside it.

Model Overall Accuracy

F-statistic

- ✓ A statistical measure that is used to **test the overall significance** of a model in hypothesis testing.

$$\text{F-statistic} = \frac{RSS/k}{ESS/(n-k-1)}$$

k is Number of independent variables

n is the sample size

Explained Sum of Squares

$$\boxed{ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}$$

With F-statistics we can answer is at least one of the independent variables $X_1, X_2, \ldots, X_p$ useful in predicting the output or not?

Model Overall Accuracy

F-statistic

$$\text{F-statistic} = \frac{RSS/k}{ESS/(n-k-1)}$$

✓ If the calculated F-statistic is larger than the **critical value** (a threshold), we reject the null hypothesis.

✓ In addition a larger F-statistic indicates that the independent variables have a **stronger effect** on the dependent variable and that the model is a better fit to the data.

Assignment

How to determine critical value for F-statistic?
(formulate similar to slides)

Python Example

Model Overall Accuracy (Metrics)

Different metrics

- ✓MSE (Mean Squared Error) ↓ Smaller MSE indicates a **better fit**
- ✓RSE (Residual Standard Error) ↓ Smaller RSE indicates a **better fit**
- ✓R-squared (R^2) ↑ Higher values indicates a better fit [between 0 to 1]
- ✓TSS (Total Sum of Squares) ↑ Higher value of TSS is generally desirable
- ✓F-statistic ↑ Larger F-statistic indicates that the independent variables have a **stronger effect** dependent variable and better fit
- ✓…

Summery

- ✓ We discussed why Relationship between data is important
- ✓ We saw how Linear regression objective can be defined
- ✓ We answered how to check Null hypothesis.
- ✓ We defined four different important model’s overall accuracy