Introduction to Data Science

- Introduction to Machine Learning: Linear Regression -

Mauricio Molina

Keio University, Faculty of Economics

June 18, 2025

Contents

What is Machine Learning?

Definition

A subfield of AI where systems learn patterns from data to make decisions or predictions without explicit programming.

- **Traditional Programming**: Rules + Data $\rightarrow$ Output
- **Machine Learning**: Data + Output $\rightarrow$ Rules (Model)
- **Key Goal**: Generalize from examples to unseen data.

Supervised Learning

Supervised learning trains a model using labeled data where each input has a known correct output. The model learns by comparing its predictions with these correct answers and improves over time. It is used for both classification and regression problems.

Key Features

- Uses labeled data (input-output pairs).
- Goal: Learn a mapping from inputs to outputs.
- Two subtypes: **Classification** (discrete labels) and **Regression** (continuous labels).

Examples

- Email spam detection (Classification).
- House price prediction (Regression).
- Algorithms: Linear Regression, SVM, Random Forest.

- **Classification:** A shows a dataset of a shopping store that is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age, and salary. Input: Gender, Age, Salary Output: Purchased i.e. 0 or 1; 1 means yes the customer will purchase and 0 means that the customer won't purchase it.
- **Regression:** B shows a Meteorological dataset that serves the purpose of predicting wind speed based on different parameters. Input: Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction Output: Wind Speed

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Figure: Example of Labeled Data

Source: `https://www.geeksforgeeks.org/supervised-machine-learning/`

Unsupervised Learning

Unlike supervised learning, where the data is labeled with a specific category or outcome, unsupervised learning algorithms are tasked with finding patterns and relationships within the data without any prior knowledge of the data's meaning.

Key Features

- Uses unlabeled data (no predefined outputs).
- Goal: Discover hidden patterns or groupings.
- Common tasks: Clustering, Dimensionality Reduction.

Examples

- Customer segmentation (Clustering with K-Means).
- Image compression (Dimensionality Reduction with PCA).
- Algorithms: DBSCAN, Apriori, t-SNE.

This Data-set is Mall data that contains information about its clients that subscribe to them. Once subscribed they are provided a membership card and the mall has complete information about the customer and his/her every purchase. Now using this data and unsupervised learning techniques, the mall can easily group clients based on the parameters we are feeding in.

CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35

Figure A

Figure: Example of Labeled Data

Source: `https://www.geeksforgeeks.org/unsupervised-learning/`

The input to the unsupervised learning models is as follows:

- **Unstructured data:** May contain noisy(meaningless) data, missing values, or unknown data
- **Unlabeled data:** Data only contains a value for input parameters, there is no targeted value(output). It is easy to collect as compared to the labeled one in the Supervised approach.

Reinforcement Learning

Reinforcement Learning (RL) trains an agent to make decisions by interacting with an environment. Instead of being told the correct answers, agent learns by trial and error method and gets rewards for good actions and penalties for bad ones. Over time it develops a strategy to maximize rewards and achieve goals. This approach is good for problems having sequential decision making such as robotics, gaming and autonomous systems.

Key Features

- Learns via trial-and-error with rewards/penalties.
- Agent interacts with an environment to maximize cumulative reward.
- No explicit data: Focus on **sequential decision-making**.

Examples

- Game-playing AI (AlphaGo, Chess engines).
- Autonomous vehicle navigation.
- Algorithms: Q-Learning, Deep Q-Networks (DQN).

Reinforcement Learning revolves around the idea that an agent (the learner or decision-maker) interacts with an environment to achieve a goal. The agent performs actions and receives feedback to optimize its decision-making over time.

- **Agent**: The decision-maker that performs actions.
- **Environment**: The world or system in which the agent operates.
- **State**: The situation or condition the agent is currently in.
- **Action**: The possible moves or decisions the agent can make.
- **Reward**: The feedback or result from the environment based on the agent's action.

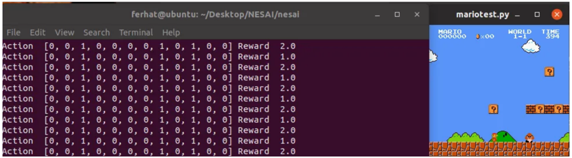


Figure: Super Mario Bros. Bot with Reinforcement Learning

Source: https://medium.datadriveninvestor.com/super-mario-bros-reinforcement-learning-77d6615a805e
see also: https://www.youtube.com/watch?v=rD9JVm3_ke0

Summary

- **Supervised**: Labeled data $\rightarrow$ Predict outcomes.
- **Unsupervised**: Unlabeled data $\rightarrow$ Find structure.
- **Reinforcement**: Agent-environment interaction $\rightarrow$ Maximize rewards.

Key Takeaway

Choose the type based on the problem and data availability!

Types of Machine Learning: Key Differences

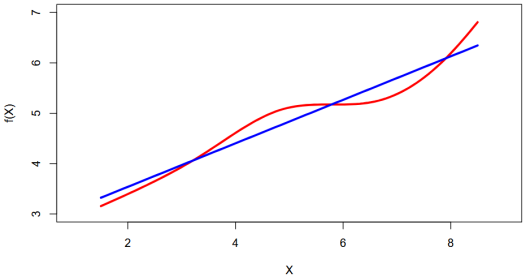
Characteristic	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Data Type	Labeled (input-output pairs)	Unlabeled (only inputs)	No initial data; learns via interaction
Goal	Predict outputs for new inputs	Discover hidden patterns or groupings	Maximize cumulative reward through actions
Feedback	Explicit (correct answers provided)	No feedback	Delayed feedback (rewards/penalties)
Examples	• Spam detection • House price prediction	• Customer segmentation • Anomaly detection	• Game-playing AI • Robotics control
Algorithms	• Linear Regression • SVM, Random Forest	• K-Means • PCA	• Q-Learning • Deep Q-Networks (DQN)
Data Structure	Input-Output pairs	Only input data	States, Actions, Rewards

Key Takeaway

Supervised: Needs labels. **Unsupervised**: Finds structure. **Reinforcement**: Learns by trial-and-error.

Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.
- True regression functions are never linear!



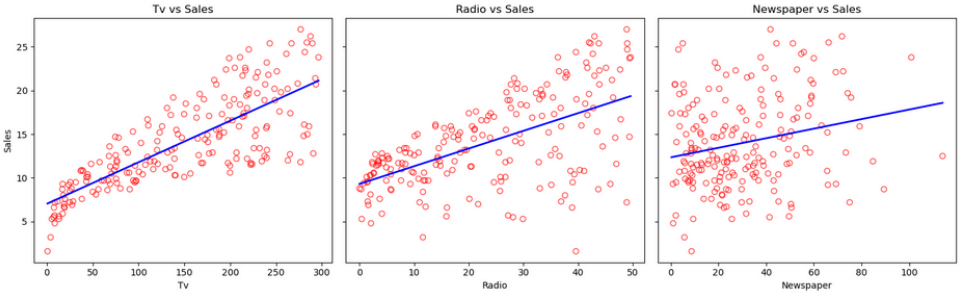
- although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Consider the advertising data shown on the next slide.
Questions we might ask:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Advertising data

Figure: Sales vs Each Predictor with OLS Fit



Simple linear regression using a single predictor X

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The hat symbol denotes an estimated value.

Estimation of the parameters by ordinary least squares (OLS)

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i th value of X. Then $e_i = y_i - \hat{y}_i$ represents the i th residual
- We define the residual sum of squares (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

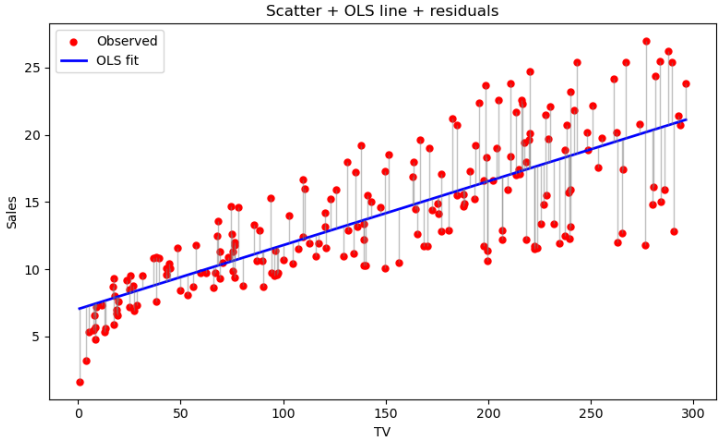
$$\text{RSS} = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \ldots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2.$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$ are the sample means.

OLS for Sales vs. TV



The least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

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

  where $\sigma^2 = \text{Var}(\epsilon)$

- These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_1 \pm 2 \cdot \text{SE}\left(\hat{\beta}_1\right)$$

That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \text{SE}\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2 \cdot \text{SE}\left(\hat{\beta}_1\right)\right]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)
For the advertising data, the 95% confidence interval for β_1 is [0.042, 0.053]

Output of sm.OLS()

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.612
Model:	OLS	Adj. R-squared:	0.610
Method:	Least Squares	F-statistic:	312.1
Date:	Mon, 26 May 2025	Prob (F-statistic):	1.47e-42
Time:	14:00:15	Log-Likelihood:	-519.05
No. Observations:	200	AIC:	1042.
Df Residuals:	198	BIC:	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

Omnibus:	0.531	Durbin-Watson:	1.935
Prob(Omnibus):	0.767	Jarque-Bera (JB):	0.669
Skew:	-0.089	Prob(JB):	0.716
Kurtosis:	2.779	Cond. No.	338.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

p-values
const 1.406300e-35
TV 1.467390e-42

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of
  H_0 : There is no relationship between X and Y versus the alternative hypothesis
  H_A : There is some relationship between X and Y.
- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

- To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}\left(\hat{\beta}_1\right)}$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value.

Results for the advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Assessing the Overall Accuracy of the Model

- We compute the Residual Standard Error

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

where the residual sum-of-squares is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- R-squared or fraction of variance explained is

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.

- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y :

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

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon.$$

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated - a *balanced design*:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as *"a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed"*, are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous - when X_j changes, everything else changes.
- *Claims of causality* should be avoided for observational data.

Estimation and Prediction for Multiple Regression

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$, we can make predictions using the formula

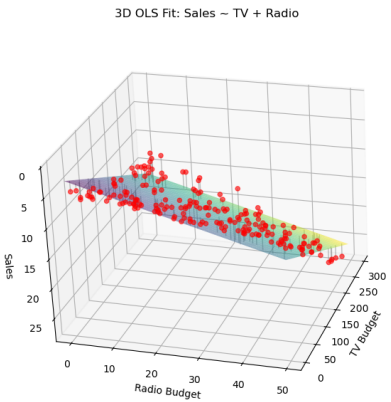
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2 . \end{aligned}$$

This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

Output of sm.OLS()



Output of sm.OLS()

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Mon, 26 May 2025   Prob (F-statistic):           1.58e-96
Time:                        14:59:15   Log-Likelihood:                -386.18
No. Observations:                 200   AIC:                             780.4
Df Residuals:                     196   BIC:                             793.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.9389      0.312      9.422      0.000      2.324      3.554
TV             0.0458      0.001     32.809      0.000      0.043      0.049
radio          0.1885      0.009     21.893      0.000      0.172      0.206
newspaper     -0.0010      0.006     -0.177      0.860     -0.013      0.011
==============================================================================
Omnibus:                       60.414   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.241
Skew:                          -1.327   Prob(JB):                     1.44e-33
Kurtosis:                       6.332   Cond. No.                         454.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie
d.
const        1.267295e-17
TV           1.509960e-81
radio        1.505339e-54
newspaper    8.599151e-01
dtype: float64
```

Results for advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Does it make sense for multiple regression to suggest no relationship between sales and newspaper while the simple linear regression implies the opposite?

Some important questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is There a Relationship Between the Response and Predictors?

Recall that in the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$. In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. As in the simple linear regression setting, we use a hypothesis test to answer this question. We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

Is at least one predictor useful?

For the first question, we can use the F-statistic

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Deciding on the important variables

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models!
  Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

Forward selection

- Begin with the null model - a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value - that is, the variable that is the least statistically significant.
- The new ($p-1$)-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p -value defined by some significance threshold.

Model selection - continued

- There is a systematic criteria for choosing an "optimal" member in the path of models produced by forward or backward stepwise selection.
- These include Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2 and Cross-validation (CV).

Model Fit

- Recall that in simple regression, R^2 is the square of the correlation of the response and the variable. In multiple linear regression, it turns out that it equals $\text{Cor}(Y, \hat{Y})^2$, the square of the correlation between the response and the fitted linear model; in fact one property of the fitted linear model is that it maximizes this correlation among all possible linear models.
- R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. This is due to the fact that adding another variable always results in a decrease in the residual sum of squares on the training data. In general RSE is defined as

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}$$

Assignment 10

Answer all the following questions in the Jupyter Notebook format. Show your Python code (when requested) and a short explanation for every result. Upload the completed `.ipynb` to K-LMS by next Tuesday at midnight.

Q1. Load the Student Perfomance dataset. Drop any null or `NaN` entries.

1. For each predictor in

   $$\{\texttt{absences}, \texttt{health}, \texttt{freetime}, \texttt{studytime}, \texttt{traveltime}\},$$

   fit *separate* simple linear regression with response `G1`.
2. For each model:
   - Report the estimated coefficient, its standard error, t-statistic, p-value and R^2.
   - State whether the coefficient is statistically significant at the 5% level.
3. Based on these results, summarize in a few sentences how each predictor relates to first-period grade `G1`.

Q2. Using the same (cleaned) dataset,

1. Fit a *multiple* linear regression with

$$\texttt{G1} \sim \texttt{absences} + \texttt{health} + \texttt{freetime} + \texttt{studytime} + \texttt{traveltime}.$$

2. Interpret each estimated coefficient (sign and magnitude) and state whether it is significant.
3. Compute and display the correlation matrix for the five predictors.
4. Answer: Are all coefficients significant in this full model?
5. Propose and fit a reduced model by removing one or more non-significant predictors.
   - Compare the reduced model's R^2 (and adjusted R^2) to the full model.
   - Which model better explains the variance in `G1`? Justify your choice.