Introduction to Statistical Learning and Application - Master of Applied Mathematics

Razan MHANNA

STATIFY team, Inria, University Grenoble Alpes
Team 5, Grenoble Institute of Neurosciences GIN

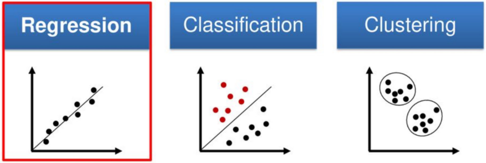
February 12, 2026

Contents

1. Statistical Learning
2. Simple Linear Regression
3. Model assessment
4. Multiple Linear Regression
5. Final notes

Course Information

- This is the complementary course of "Introduction to Statistical Learning and Applications" given to students from ENSIMAG and UGA by Professor Pedro Rodrigues.
- The classes will be given on Tuesdays from 11h30 to 13h at IM2AG, room F118.
- The materials used will be made available in this page https://github.com/ISLA-Grenoble/2026-complementary. The course reference book here.

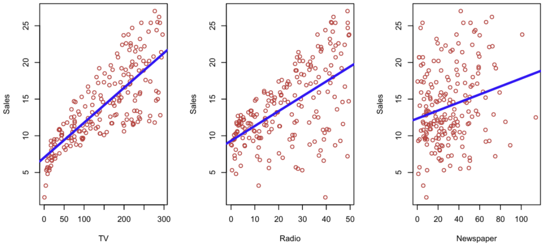


Current Section

1. Statistical Learning
2. Simple Linear Regression
3. Model assessment
4. Multiple Linear Regression
5. Final notes

Motivation for Statistical Learning

Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.



Statistical Learning Setup

- Goal: Understand the relationship between advertising and sales.
- Data: Sales in 200 markets, with advertising budgets for TV, radio, and newspaper.
- Objective: Develop an accurate model that can be used to predict sales on the basis of the three media budgets.
- Inputs(Predictors): Advertising budgets (TV, radio, newspaper), denoted as X_1, X_2, X_3.
- Quantative Response: Sales (denoted as Y).

$$Y = f(X) + \varepsilon$$

where is some fixed but unknown function of X1, ..., Xp, and ε is a random error term, which is independent of X and has mean zero.

In essence, statistical learning refers to a set of approaches for estimating f.

Reasons for Estimating *f*

Prediction:

- Inputs X are readily available, but output Y is difficult to obtain.
- We estimate f to predict Y using $\hat{Y} = \hat{f}(X)$.
- The accuracy of $\hat{(}Y)$ as a prediction for Y depends on two quantities: the reducible error and the irreducible error.
- Why is the irreducible error larger than zero? Due to unmeasurable factors that are useful in predicting Y, or simply the quantity ε may contain unmeasurable variations.

$$E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 + \text{Var}(\varepsilon)$$

- The focus of this course is on techniques for estimating f with the aim of minimizing the reducible error.
- The irreducible error sets an upper bound on prediction accuracy. This bound is unkown in pratice.

Inference

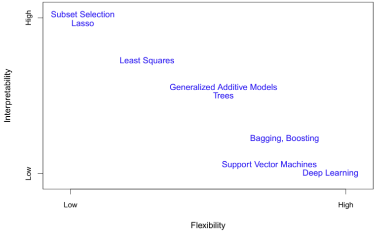
Inference: Understand the association between Y and $X_1, X_2, \ldots, X_p$. Unlike prediction, fcan't be trated as a blackbox; we need the exact form of $\hat{f}$.

- Key questions:
  1. Which predictors are associated with the response?
  2. What is the relationship between Y and each predictor? Some predictors may have a positive relationship with Y. Other predictors may have the opposite relationship.
  3. Can the relationship be summarized with a linear model, or is it more complex?

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate. For example:

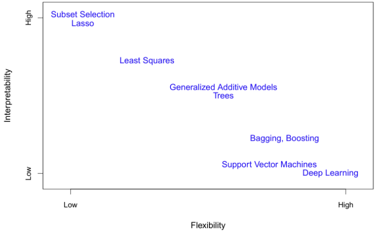
- *Linear Models: Easy to interpret.*
- *Non-linear Models: More accurate predictions but harder to interpret.*

The Trade-Off Between Prediction Accuracy and Model Interpretability



Why Use Restrictive Models? When inference is the goal, restrictive models are more interpretable. For instance, the linear model may be a good choice for inference since it will be quite easy to understand the relationship between Y and X1,X2,...,Xp. In contrast, very flexible approaches, such as the splines and boosting.

The Trade-Off Between Prediction Accuracy and Model Interpretability



Why Use Restrictive Models? When inference is the goal, restrictive models are more interpretable. For instance, the linear model may be a good choice for inference since it will be quite easy to understand the relationship between Y and X1,X2,...,Xp. In contrast, very flexible approaches, such as the splines and boosting.

The Trade-Off Between Prediction Accuracy and Model Interpretability

- Least Squares Regression: A simple and interpretable method that fits a linear relationship between input and output but may lack flexibility for complex patterns.
- Lasso: A linear regression variant that adds sparsity by shrinking some coefficients to zero, improving interpretability while maintaining predictive power.
- Generalized Additive Models (GAMs): Extend linear models by allowing non-linear relationships, increasing flexibility but making interpretation harder.
- Non-Linear Methods: Techniques like bagging, boosting, support vector machines, and neural networks offer high predictive accuracy but at the cost of interpretability and complex inference.

How Do We Estimate f?

Given n data points, the training data is used to estimate f by finding a function $\hat{f}$ such that $Y \approx \hat{f}(X)$. These methods are classified into parametric and non-parametric approaches.

Parametric methods: assume a specific functional form for f, simplifying the estimation process, then use training data to estimate the model's parameters:

- Assume a functional form, e.g., linear model

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Although efficient, parametric methods may suffer if the true function f deviates significantly from the assumed model, leading to poor estimates.

How Do We Estimate f?

Given n data points, the training data is used to estimate f by finding a function $\hat{f}$ such that $Y \approx \hat{f}(X)$. These methods are classified into parametric and non-parametric approaches.

Parametric methods: assume a specific functional form for f, simplifying the estimation process, then use training data to estimate the model's parameters:

- Assume a functional form, e.g., linear model

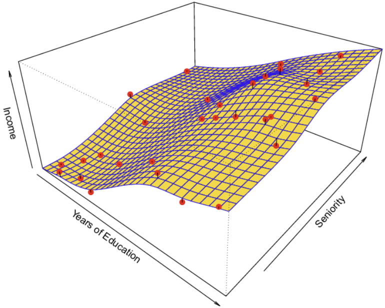
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Although efficient, parametric methods may suffer if the true function f deviates significantly from the assumed model, leading to poor estimates.

Non-Parametric Methods

Non-parametric methods do not assume a functional form for *f*, aiming for a flexible fit:

- These methods can more accurately capture complex relationships but require a larger number of observations.
- Example: Thin-plate splines fit the data closely. Overfitting!



Supervised vs Unsupervised Learning

- Supervised learning involves predictor measurements $x_1, \ldots, x_n$ with associated response measurements y_i.
  - Linear regression, logistic regression, Generalized Additive Models (GAM), boosting, support vector machines.
- In unsupervised learning, only predictors $x_1, \ldots, x_n$ are available, without corresponding responses y_i.
  - Clustering analysisâgrouping similar observations based on predictor variables.
- Semi-supervised learning aims to incorporate both labeled and unlabeled data. Some observations have both predictors and responses, while others only have predictors.
  - Scenario arises when predictors are cheap to collect, but responses are expensive.

Current Section

1. Statistical Learning
2. Simple Linear Regression
   - Model definition
   - Estimation of the parameters by least squares
   - Example
3. Model assessment
4. Multiple Linear Regression
5. Final notes

Outline

1. Statistical Learning

2. Simple Linear Regression
   - Model definition
   - Estimation of the parameters by least squares
   - Example

3. Model assessment

4. Multiple Linear Regression

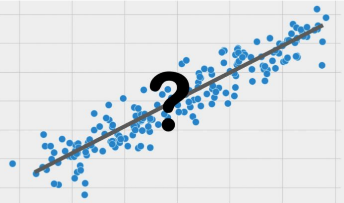
5. Final notes

Simple Linear Regression (SLR)

Idea

We model the relationship between an input variable x and an output variable y using a straight line, with the goal of prediction and explanation.

- We assume an approximately linear trend between x and y.
- We fit a line that minimizes the vertical errors between observed values y_i and predicted values $\hat{y}_i$.
- Once fitted, the line provides predictions of y for new values of x.



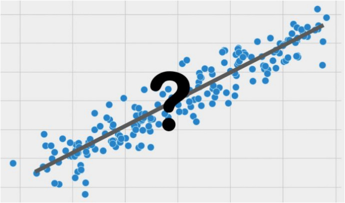
Example: scatterplot + fitted line

Simple Linear Regression (SLR)

Idea

We model the relationship between an input variable x and an output variable y using a straight line, with the goal of prediction and explanation.

- We assume an approximately linear trend between x and y.
- We fit a line that minimizes the vertical errors between observed values y_i and predicted values $\hat{y}_i$.
- Once fitted, the line provides predictions of y for new values of x.



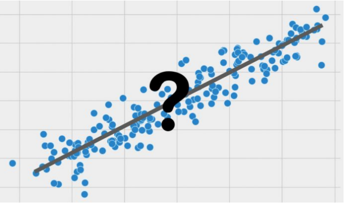
Example: scatterplot + fitted line

Simple Linear Regression (SLR)

Idea

We model the relationship between an input variable x and an output variable y using a straight line, with the goal of prediction and explanation.

- We assume an approximately linear trend between x and y.
- We fit a line that minimizes the vertical errors between observed values y_i and predicted values $\hat{y}_i$.
- Once fitted, the line provides predictions of y for new values of x.



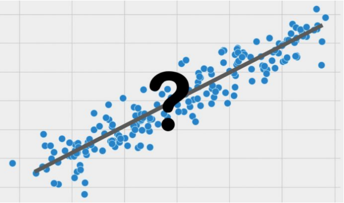
Example: scatterplot + fitted line

Simple Linear Regression (SLR)

Idea

We model the relationship between an input variable x and an output variable y using a straight line, with the goal of prediction and explanation.

- We assume an approximately linear trend between x and y.
- We fit a line that minimizes the vertical errors between observed values y_i and predicted values $\hat{y}_i$.
- Once fitted, the line provides predictions of y for new values of x.



Example: scatterplot + fitted line

SLR Model, Variables, and Interpretation

Model

For each observation $i = 1, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{and} \qquad \hat{y}_i = b_0 + b_1 x_i$$

Vocabulary

- x: predictor / explanatory / independent variable
- y: response / outcome / dependent variable
- ε_i: random error (noise, unmodeled effects)

Interpretation

- b_0: predicted value of y when $x = 0$
- b_1: expected change in y for a one-unit increase in x
- Residual: $e_i = y_i - \hat{y}_i$

SLR Model, Variables, and Interpretation

Model

For each observation $i = 1, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{and} \qquad \hat{y}_i = b_0 + b_1 x_i$$

Vocabulary

- x: predictor / explanatory / independent variable
- y: response / outcome / dependent variable
- ε_i: random error (noise, unmodeled effects)

Interpretation

- b_0: predicted value of y when $x = 0$
- b_1: expected change in y for a one-unit increase in x
- Residual: $e_i = y_i - \hat{y}_i$

SLR Model, Variables, and Interpretation

Model

For each observation $i = 1, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{and} \qquad \hat{y}_i = b_0 + b_1 x_i$$

Vocabulary

- x: predictor / explanatory / independent variable
- y: response / outcome / dependent variable
- ε_i: random error (noise, unmodeled effects)

Interpretation

- b_0: predicted value of y when $x = 0$
- b_1: expected change in y for a one-unit increase in x
- Residual: $e_i = y_i - \hat{y}_i$

SLR Model, Variables, and Interpretation

Model

For each observation $i = 1, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{and} \qquad \hat{y}_i = b_0 + b_1 x_i$$

Vocabulary

- x: predictor / explanatory / independent variable
- y: response / outcome / dependent variable
- ε_i: random error (noise, unmodeled effects)

Interpretation

- b_0: predicted value of y when $x = 0$
- b_1: expected change in y for a one-unit increase in x
- Residual: $e_i = y_i - \hat{y}_i$

Outline

1. Statistical Learning
2. Simple Linear Regression
   - Model definition
   - Estimation of the parameters by least squares
   - Example
3. Model assessment
4. Multiple Linear Regression
5. Final notes

Least Squares: Setup

Model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Prediction

$$\hat{y}_i = b_0 + b_1 x_i$$

Least Squares Criterion

Choose b_0, b_1 to minimize

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Normal Equations

Minimization

At the optimum:

$$\frac{\partial S}{\partial b_0} = 0, \qquad \frac{\partial S}{\partial b_1} = 0$$

Normal equations

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

Key consequence

$$b_0 = \bar{y} - b_1 \bar{x}$$

Normal Equations

Minimization

At the optimum:

$$\frac{\partial S}{\partial b_0} = 0, \qquad \frac{\partial S}{\partial b_1} = 0$$

Normal equations

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

Key consequence

$$b_0 = \bar{y} - b_1 \bar{x}$$

Normal Equations

Minimization

At the optimum:

$$\frac{\partial S}{\partial b_0} = 0, \qquad \frac{\partial S}{\partial b_1} = 0$$

Normal equations

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

Key consequence

$$b_0 = \bar{y} - b_1 \bar{x}$$

Least Squares Estimates

Slope

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

Intercept

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Interpretation

- $\hat{b}_0$: mean of Y when $X = 0$
- $\hat{b}_1$: mean change in Y per unit increase in X

Least Squares Estimates

Slope

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

Intercept

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Interpretation

- $\hat{b}_0$: mean of Y when $X = 0$
- $\hat{b}_1$: mean change in Y per unit increase in X

Least Squares Estimates

Slope

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

Intercept

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Interpretation

- $\hat{b}_0$: mean of Y when $X = 0$
- $\hat{b}_1$: mean change in Y per unit increase in X

Assumptions of Simple Linear Regression

The validity of estimation and inference relies on the following assumptions, summarized by the acronym **LINE**:

- *Linearity*: the conditional mean of Y is a linear function of X
$$\mathbb{E}[Y \mid X] = b_0 + b_1 X$$
- *Independence*: the error terms ε_i are independent
- *Normality*: the errors are normally distributed
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$
- *Equal variance (Homoscedasticity)*:
$$\mathrm{Var}(\varepsilon \mid X) = \sigma^2$$

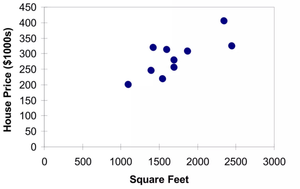
Outline

1. Statistical Learning
2. Simple Linear Regression
   - Model definition
   - Estimation of the parameters by least squares
   - Example
3. Model assessment
4. Multiple Linear Regression
5. Final notes

Example: House Price vs. Size (Data)

A real-estate agent studies the relationship between house selling price and size. Let Y be the house price (in \$1000s) and X the size (in ft^2).

Y (\$1000s)	X (ft^2)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



- Goal: fit a simple linear regression model.
- Interpret slope and make predictions.

Example: Fitted Model and Prediction

The fitted simple linear regression model is:

$$\widehat{Y} = 98.248 + 0.10977\,X$$

where Y is in \$1000s and X is in ft^2.

Interpretation

- Intercept (98.248): predicted price when $X = 0$ (not meaningful physically, but part of the model).
- Slope (0.10977): +0.10977 (\$1000s) per ft^2 $\approx$ \$110 per additional ft^2.

Prediction for $X = 2000$ ft^2

$$\widehat{Y}(2000) = 98.248 + 0.10977 \times 2000 \approx$$

Predicted price $\approx$ \$317,790.

- Units check: Y is in thousands of dollars.
- Report with appropriate rounding.

Manual Solution: Least Squares Estimates

We fit $Y = b_0 + b_1 X + \varepsilon$ by minimizing $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$.

Step 1: Compute sample means

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Step 2: Compute sums of squares

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Step 3: Closed-form estimators

$$b_1 = \frac{S_{xy}}{S_{xx}}, \qquad b_0 = \bar{y} - b_1 \bar{x}$$

Solution in Python: Two Approaches to Linear Regression

- **`scikit-learn` (machine-learning oriented)**
  - Designed primarily for prediction and model deployment
  - Emphasizes algorithmic efficiency and scalability
  - Regression is treated as a black-box estimator
  - Outputs: fitted coefficients and predictions
- **`statsmodels` (statistics-oriented)**
  - Designed for statistical modeling and inference
  - Explicitly follows the classical Ordinary Least Squares (OLS) framework
  - Provides full access to uncertainty quantification and hypothesis testing

Key message: the choice of library reflects whether the goal is prediction or statistical interpretation.

Python Solution `scikit-learn`

```
import numpy as np
from sklearn.linear_model import LinearRegression

X = np.array([1400, 1600, 1700, 1875, 1100,
              1550, 2350, 2450, 1425, 1700]).reshape(-1,
     1)
y = np.array([245, 312, 279, 308, 199,
              219, 405, 324, 319, 255])  # \ $1000s

model = LinearRegression().fit(X, y)
b0 = model.intercept_
b1 = model.coef_[0]
pred_2000 = model.predict([[2000]])[0]

print("b0 =", b0)
print("b1 =", b1)
print("Pred(2000) =", pred_2000, "($1000s)")
```

What Does `statsmodels` Provide?

Model-based inference

For the linear model

$$Y = b_0 + b_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

`statsmodels` estimates:

- Regression coefficients (b_0, b_1)
- Variance of the noise σ^2
- Standard errors of the estimators

Hypothesis testing and model quality

- t-tests for $H_0 : b_j = 0$ (significance of predictors)
- p-values and confidence intervals
- R^2: proportion of variance explained by the model

What Does `statsmodels` Provide?

Model-based inference

For the linear model

$$Y = b_0 + b_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

`statsmodels` estimates:

- Regression coefficients (b_0, b_1)
- Variance of the noise σ^2
- Standard errors of the estimators

Hypothesis testing and model quality

- t-tests for $H_0 : b_j = 0$ (significance of predictors)
- p-values and confidence intervals
- R^2: proportion of variance explained by the model

Py solution using `statsmodels`

```
import statsmodels.api as sm

X = np.array([...])  # house sizes
y = np.array([...])  # prices in $1000s
```

Important: adding the intercept

`statsmodels` does not add the intercept automatically.

```
X_design = sm.add_constant(X)
```

This creates the design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

```
ols = sm.OLS(y, X_design).fit() # OLS = Ordinary Least S
```

Py solution using `statsmodels`

```
import statsmodels.api as sm

X = np.array([...])  # house sizes
y = np.array([...])  # prices in $1000s
```

Important: adding the intercept

`statsmodels` does <u>not</u> add the intercept automatically.

```
X_design = sm.add_constant(X)
```

This creates the design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

```
ols = sm.OLS(y, X_design).fit() # OLS = Ordinary Least S
```

Py solution using `statsmodels`

```
import statsmodels.api as sm

X = np.array([...])  # house sizes
y = np.array([...])  # prices in $1000s
```

Important: adding the intercept

`statsmodels` does not add the intercept automatically.

```
X_design = sm.add_constant(X)
```

This creates the design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

```
ols = sm.OLS(y, X_design).fit() # OLS = Ordinary Least S
```

Py solution using `statsmodels`

The command: `ols.summary()` produces a complete statistical report.

Key quantities (focus on these)

- **coef**: estimated parameters (b_0, b_1)
- **std err**: standard error of each estimate
- **t**: t-statistic for $H_0 : b_j = 0$
- **P**$>|t|$: corresponding p-value
- R^2: proportion of variance explained

Py solution using `statsmodels`

From the output, we obtain:

$$\widehat{Y} = 98.248 + 0.10977\,X$$

Intercept (b_0)

- Estimated price when $X = 0$
- Not meaningful physically
- Needed to correctly position the regression line

Slope (b_1)

- Average increase in price per additional ft^2
- 0.10977 thousand dollars $\approx$ \$110
- Indicates a positive association between size and price

Py solution using `statsmodels`

From the output, we obtain:

$$\widehat{Y} = 98.248 + 0.10977\,X$$

Intercept (b_0)

- Estimated price when $X = 0$
- Not meaningful physically
- Needed to correctly position the regression line

Slope (b_1)

- Average increase in price per additional ft^2
- 0.10977 thousand dollars $\approx$ \$110
- Indicates a positive association between size and price

Py solution using `statsmodels`

To predict the price for a house of 2000 ft^2:

```
pred_2000 = ols.predict([1, 2000])[0]
```

Why `[1, 2000]`?

Because the model expects:

$$(1, X) \quad \text{(intercept + predictor)}$$

$$\widehat{Y}(2000) \approx 317.79$$

- Unit: \$1000s
- Final prediction: $\approx$ \$317,790

Py solution using `statsmodels`

To predict the price for a house of 2000 ft^2:

```
pred_2000 = ols.predict([1, 2000])[0]
```

Why `[1, 2000]`?

Because the model expects:

$$(1, X) \quad \text{(intercept + predictor)}$$

$$\widehat{Y}(2000) \approx 317.79$$

- Unit: \$1000s
- Final prediction: $\approx$ \$317,790

Py solution using `statsmodels`

To predict the price for a house of 2000 ft^2:

```
pred_2000 = ols.predict([1, 2000])[0]
```

Why `[1, 2000]`?

Because the model expects:

$$(1, X) \quad \text{(intercept + predictor)}$$

$$\widehat{Y}(2000) \approx 317.79$$

- Unit: $1000s
- Final prediction: $\approx$ $317,790

Current Section

1. Statistical Learning
2. Simple Linear Regression
3. Model assessment
4. Multiple Linear Regression
5. Final notes

Measure of Variation: SST, SSR, SSE

Sum of squares decomposition $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$

- Total Sum of Squares (SST / TSS):

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

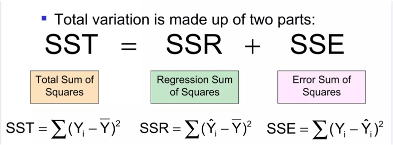
- Regression Sum of Squares (SSR):

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

- Error Sum of Squares (SSE / RSS):

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

Visual Interpretation of Variation Decomposition



- Total variation is made up of two parts:

$$SST \quad = \quad SSR \quad + \quad SSE$$

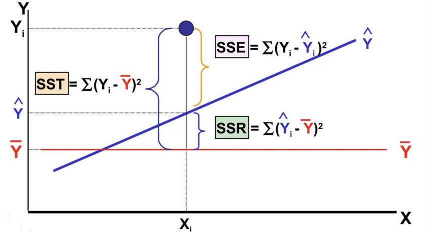
Total Sum of Squares

Regression Sum of Squares

Error Sum of Squares

$SST = \sum (Y_i - \bar{Y})^2$ $SSR = \sum (\hat{Y}_i - \bar{Y})^2$ $SSE = \sum (Y_i - \hat{Y}_i)^2$

Illustration of total variation and fitted values.


Y
Y_i
$SSE = \Sigma(Y_i - \hat{Y}_i)^2$
$\hat{Y}$
$SST = \Sigma(Y_i - \bar{Y})^2$
$\hat{Y}$
$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$
$\bar{Y}$
$\bar{Y}$
X_i
X

SST decomposes into explained (SSR) and unexplained (SSE) parts.

Coefficient of Determination: R^2 (or r^2)

Definition

The coefficient of determination is the proportion of total variation in Y explained by the linear regression on X:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

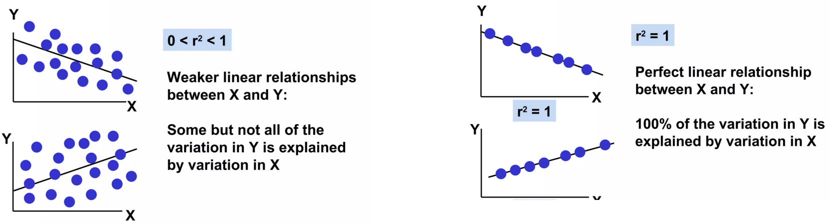
Interpretation

- $R^2 \in [0, 1]$
- $R^2 \approx 1$: the model explains most variability in Y
- $R^2 \approx 0$: the model explains little variability in Y

Remark

In **simple** linear regression (one predictor with intercept), $R^2 = \mathrm{Cor}(X, Y)^2$

Examples of Typical R^2 Values



$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Interpretation of a “good” R^2 is application-dependent (physics vs biology vs social sciences).

Standard Error: Measuring Accuracy

Warm-up: sample mean

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

More observations ⇒ smaller uncertainty.

Regression: uncertainty of coefficients

We also want to quantify how far $(\hat{\beta}_0, \hat{\beta}_1)$ can be from (β_0, β_1). This is captured by:

$$\mathrm{SE}(\hat{\beta}_0), \qquad \mathrm{SE}(\hat{\beta}_1)$$

Interpretation: SE is the typical sampling variability of the estimator over repeated samples.

Standard Errors and Residual Standard Error

Assuming $\mathrm{Var}(\varepsilon) = \sigma^2$ and uncorrelated errors:

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

Key implication

Larger spread in X (larger $\sum(x_i - \bar{x})^2$) $\Rightarrow$ more precise slope estimate.

Estimating σ from data (Residual Standard Error)

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

Deriving $\mathrm{SE}(\hat{\beta}_1)$ and $\mathrm{SE}(\hat{\beta}_0)$ (SLR)

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. Let $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.

Step 1: write OLS estimators as linear combinations of the errors

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} \Rightarrow \hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} \Rightarrow \hat{\beta}_0 - \beta_0 = \bar{\varepsilon} - \bar{x}(\hat{\beta}_1 - \beta_1)$$

where $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i$.

Step 2: variance of $\hat{\beta}_1$ (hence $\mathrm{SE}(\hat{\beta}_1)$)

Using $\mathrm{Var}(\sum a_i \varepsilon_i) = \sum a_i^2 \sigma^2$ (uncorrelated errors):

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})\varepsilon_i\right) = \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

Deriving $\mathrm{SE}(\hat{\beta}_1)$ and $\mathrm{SE}(\hat{\beta}_0)$ (SLR)

Step 3: Variance of $\hat{\beta}_0$ (hence $\mathrm{SE}(\hat{\beta}_0)$)

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{\varepsilon}) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\,\mathrm{Cov}(\bar{\varepsilon}, \hat{\beta}_1 - \beta_1)$$

$$\mathrm{Var}(\bar{\varepsilon}) = \frac{\sigma^2}{n}, \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\mathrm{Cov}(\bar{\varepsilon}, \hat{\beta}_1 - \beta_1) = \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i, \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right)$$

$$= \frac{1}{nS_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\mathrm{Var}(\varepsilon_i) = 0 \quad \left(because \sum_{i=1}^{n}(x_i - \bar{x}) = 0\right)$$

$$\Rightarrow \quad \mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Confidence Intervals for β_0 and β_1

A 95% confidence interval provides a range of plausible values for the parameter.

Approximate 95% CI (large n)

$$\beta_1 \in \hat{\beta}_1 \pm 2\,\mathrm{SE}(\hat{\beta}_1), \qquad \beta_0 \in \hat{\beta}_0 \pm 2\,\mathrm{SE}(\hat{\beta}_0)$$

More precise 95% CI (exact under normal errors)

Replace “2” by the t quantile with $n-2$ degrees of freedom:

$$\beta_1 \in \hat{\beta}_1 \pm t_{n-2,0.975}\,\mathrm{SE}(\hat{\beta}_1), \qquad \beta_0 \in \hat{\beta}_0 \pm t_{n-2,0.975}\,\mathrm{SE}(\hat{\beta}_0)$$

Why 0.975? For a 95% CI, $\alpha = 0.05$ and we split it into two tails: $0.975 = 1-\alpha/2$.

Interpretation

If we repeatedly sample data and compute the interval the same way, about 95% of these intervals contain the true parameter.

Testing Association: $H_0 : \beta_1 = 0$

Test statistic

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

Under H_0 (and normal errors), t follows a t-distribution with $n-2$ degrees of freedom:

$$t \sim t_{n-2}.$$

Decision rule at 5% (two-sided)

$$\text{Reject } H_0 \quad \Longleftrightarrow \quad |t| > t_{n-2,0.975}$$

Equivalently: reject H_0 if the p-value < 0.05.

Testing Association: $H_0 : \beta_1 = 0$

p-value (interpretation)

The p-value is the probability (under H_0) of observing a value at least as extreme as $|t|$:

$$p\text{-value} = \mathbb{P}\big(|T_{n-2}| \geq |t_{\mathrm{obs}}|\big).$$

- small p-value ⇒ evidence against H_0
- typical thresholds: 0.05 or 0.01 (context-dependent)

Assessing Model Fit Summary

Residual Standard Error (absolute scale)

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

Interpreted as the typical prediction error in the units of Y.

R^2 (scale-free proportion of variance explained)

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

- $R^2 \in [0,1]$
- close to 1: large explained variability (good linear fit)
- close to 0: weak linear explanatory power

Adjusted (Corrected) R^2

Motivation

The usual R^2 always increases when adding predictors, even if they are irrelevant. Adjusted R^2 penalizes model complexity.

Definition

Adjusted R^2 measures the proportion of variance explained by the model after accounting for the number of predictors.

Formula

$$R^2_{\text{adj}} = 1 - (1 - R^2)\,\frac{n-1}{n-p-1}$$

- n: number of observations
- p: number of predictors (excluding intercept)

Current Section

1. Statistical Learning
2. Simple Linear Regression
3. Model assessment
4. Multiple Linear Regression
5. Final notes

Multiple Linear Regression: Objective

Goal

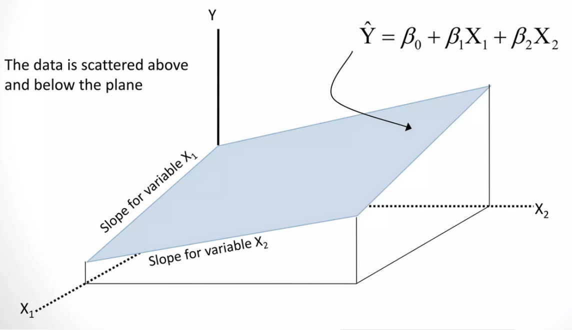
Given *n* observations and *p* predictors, we aim to model the conditional mean of Y:

$$\mathbb{E}(Y \mid X_1,\ldots,X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_p X_p + \varepsilon$$

- Each predictor contributes **additively** to the response.
- The effect of one predictor is measured while keeping others fixed.
- When $p = 1$, we recover simple linear regression.

Multiple Linear regression


Y
$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
The data is scattered above and below the plane
Slope for variable X_1
Slope for variable X_2
X_1
X_2

Estimation by Least Squares

Least Squares Criterion

We estimate $\hat{\beta}$ by minimizing the residual sum of squares:

$$\mathrm{RSS}(\beta) = \|Y - X\beta\|_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Closed-form Solution

If $X^T X$ is invertible, the solution is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Geometric Interpretation

- The vector $Y \in \mathbb{R}^n$ lives in an *n*-dimensional space.
- The columns of X span a subspace of $\mathbb{R}^n$.
- $\hat{Y} = X\hat{\beta}$ is the **orthogonal projection** of Y onto $\mathrm{col}(X)$.

Orthogonality Condition

The residual vector satisfies:

$$X^T(Y - X\hat{\beta}) = 0$$

Interpretation of Regression Coefficients

Meaning of β_j

β_j represents the expected change in Y when X_j increases by one unit, all other predictors being held constant.

- Interpretation is conditional, not marginal.
- Coefficients depend on the scale of predictors.
- Changing units changes the numerical value of β_j.

Model Assumptions

The multiple linear regression model relies on:

1. Linearity: $Y = X\beta + \varepsilon$
2. Independence of observations
3. Zero-mean errors: $\mathbb{E}(\varepsilon \mid X) = 0$
4. Homoscedasticity: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
5. No perfect multicollinearity among predictors

Important

Violation of these assumptions affects inference (standard errors, tests), not necessarily prediction.

Multicollinearity

Definition

Multicollinearity occurs when predictors are highly correlated.

- $X^T X$ becomes nearly singular.
- Coefficient estimates become unstable.
- Standard errors increase.

Key idea

Good prediction can coexist with poor coefficient interpretability.

t-test for Individual Coefficients

Hypothesis Test

For each coefficient β_j, we test:

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0$$

Test Statistic

$$t_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$$

Under H_0, t_j follows a Student distribution with $n-p-1$ degrees of freedom.

Decision Rule

If the p-value $< \alpha$ (e.g. $\alpha = 0.05$), we reject H_0.

Global Significance of the Model

F-test

The F-statistic tests:

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

Interpretation

- Small p-value $\Rightarrow$ at least one predictor is significant
- Tests whether the model explains more variance than an intercept-only model

Current Section

1. Statistical Learning
2. Simple Linear Regression
3. Model assessment
4. Multiple Linear Regression
5. Final notes

MLR with a Categorical Predictor

A company wants to predict employees' salaries (in thousands of dollars) using:

- Years of Experience (X_1): numerical predictor.
- Education Level (X_2): categorical predictor with two levels:
  - Bachelors degree (reference category): $X_2 = 0$
  - Masters degree: $X_2 = 1$

Dummy (Indicator) Variable Encoding

The categorical variable Education Level is transformed into a numerical **dummy variable**:

$$X_2 = \begin{cases} 0 & \text{Bachelors degree} \\ 1 & \text{Masters degree} \end{cases}$$

MLR with a Categorical Predictor

Years of Experience (X_1)	**Education Level (X_2)**	**Salary (Y)**
1	Bachelor (0)	30
2	Bachelor (0)	35
3	Master (1)	50
4	Bachelor (0)	50
5	Master (1)	60
6	Master (1)	65

Regression Model with a Categorical Predictor

We consider the multiple linear regression model:

$$Y = a + b_1 X_1 + b_2 X_2$$

- a: intercept (expected salary for a Bachelorâs degree with $X_1 = 0$).
- b_1: effect of one additional year of experience.
- b_2: salary difference between a Master s and a Bachelor s degree, holding experience constant.

Interpretation of b_2

- If $b_2 > 0$: having a Master degree increases salary compared to a Bachelor one.
- If $b_2 < 0$: having a Master degree decreases salary compared to a Bachelor.

Step 1: Encoding the Categorical Predictor

We model salary (in thousands) using:

- X_1: Years of experience (numerical)
- X_2: Degree type (categorical)
  - Bachelor (reference): $X_2 = 0$
  - Master: $X_2 = 1$

The regression model is:

$$Y = a + b_1 X_1 + b_2 X_2 + \varepsilon$$

Observed data

X_1	X_2	Y
1	0	30
2	0	35
3	1	50

Step 2: Design Matrix and Least Squares

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 1 \\ 1 & 4 & 0 \\ 1 & 5 & 1 \\ 1 & 6 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 30 \\ 35 \\ 50 \\ 50 \\ 60 \\ 65 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{a} \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = (X^T X)^{-1} X^T Y$$

Interpretation of the solution

- $\hat{a}$: baseline salary for a Bachelor with 0 years of experience
- $\hat{b}_1$: average increase in salary per year of experience

Intuition

Interpretation of the solution

- $\hat{b}_2$: average salary difference between Master and Bachelor, holding experience fixed

- The slope b_1 is estimated using **all observations**
- The categorical coefficient b_2 measures a **vertical shift** between the two groups

Key intuition

The model fits two **parallel lines**:

- one for Bachelor ($X_2 = 0$)
- one for Master ($X_2 = 1$)

The distance between the lines is b_2.

Effect of Scaling a Predictor

Suppose we rescale experience:

$$X_1^\star = 2X_1$$

The model becomes:

$$Y = a + b_1^\star X_1^\star + b_2 X_2 \quad \text{with} \quad b_1^\star = \frac{b_1}{2}$$

What changes?

- The coefficient of the scaled predictor changes
- The intercept and other coefficients stay the same
- Predictions $\hat{Y}$ are unchanged

Effect of Adding a Constant (Centering)

Suppose we center experience:

$$X_1^{\star} = X_1 - c$$

The model becomes:

$$Y = a^{\star} + b_1 X_1^{\star} + b_2 X_2 \quad \text{with} \quad a^{\star} = a + b_1 c$$

What changes?

- Slopes (b_1, b_2) remain unchanged
- Only the intercept changes

Linear regression issues

- Sensitivity to outliers: a few extreme observations can strongly affect the estimated coefficients.
- Multicollinearity: highly correlated predictors lead to
  - large variance of coefficient estimates,
  - unstable and unreliable interpretation.
- Overfitting: when the number of predictors is large compared to the sample size, the model may fit noise rather than the underlying signal.
- Interpretability issues: with many predictors, it becomes difficult to identify which variables truly influence the response.