Advanced Data Analysis with Python

Cecilia Graiff

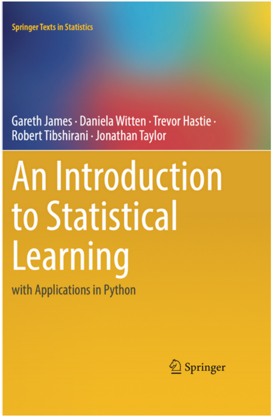
September 18, 2025

ALMAnaCH, Inria Paris - École d'Affaires Publiques, SciencesPo
cecilia.graiff@sciencespo.fr

Other recommended resources

- Think Python by Allen Downey offers electronic resources here that will allow you to directly code there
- Kaggle - Learn Python is a more interactive guide
- Invent your Own Computer Games with Python by Al Sweigart

Resources for this class



Springer Texts in Statistics

Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor

An Introduction to Statistical Learning

with Applications in Python

Springer

An Introduction to Statistical Learning

- This class follows **Chapter 3** of this book
- Refer to the book if you want to deepen today's topic
- It is a bit more mathematically intensive than this class!

Homework Correction

Statistical Modeling Foundations

What is a statistical model?

Definition (Statistical Model)

A statistical model is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population).

Source: Wikipedia

Example: Regression

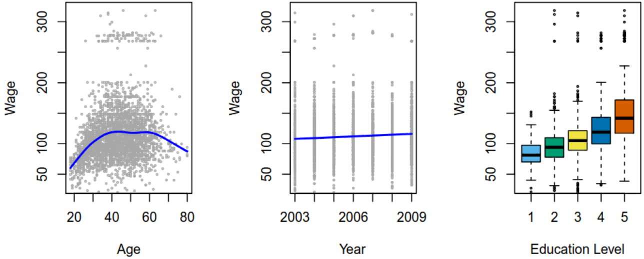


FIGURE 1.1. Wage *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: wage *as a function of* age. *On average,* wage *increases with* age *until about* 60 *years of age, at which point it begins to decline.* Center: wage *as a function of* year. *There is a slow but steady increase of approximately* $10,000 *in the average* wage *between* 2003 *and* 2009. Right: *Boxplots displaying* wage *as a function of* education, *with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* wage *increases with the level of education.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Classification

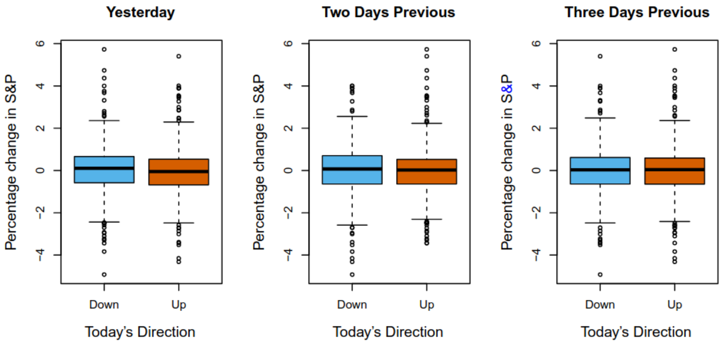


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Classification

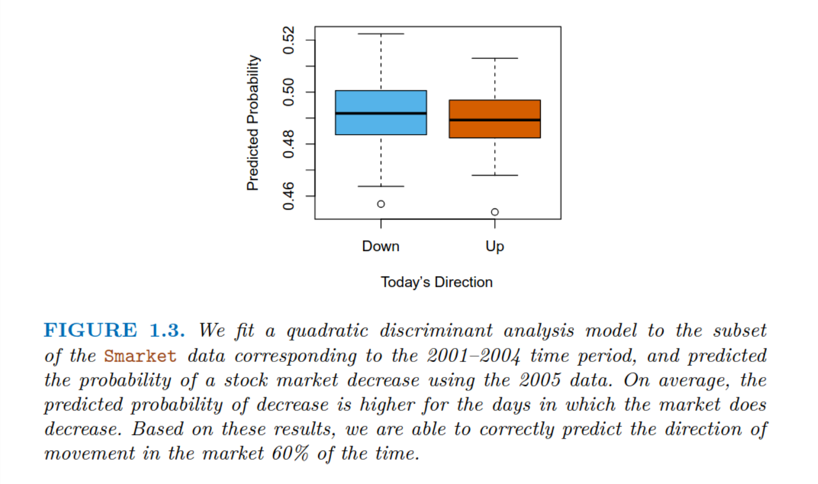


FIGURE 1.3. *We fit a quadratic discriminant analysis model to the subset of the* `Smarket` *data corresponding to the 2001–2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market 60% of the time.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Regression: Example

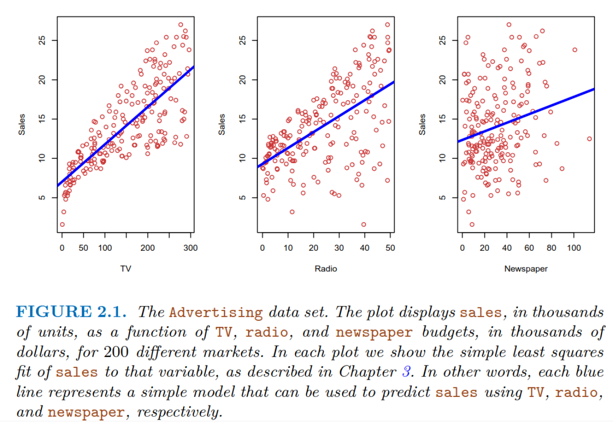


FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

Source: An Introduction to Statistical Learning with Application in Python

Regression: Research Questions

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media are associated with sales?
4. How large is the association between each medium and sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?

Source: An Introduction to Statistical Learning with Application in Python

Predictors and dependent variables

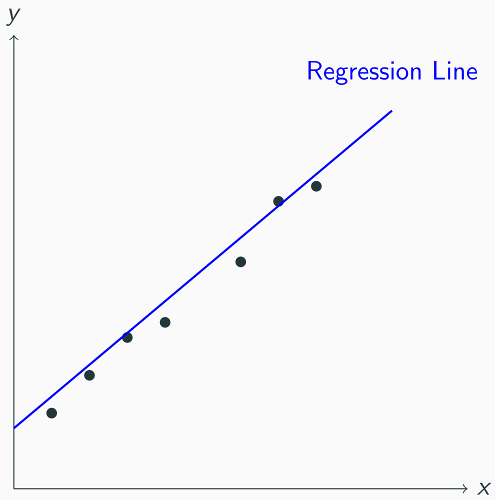
- **Dependent variable**: the output you are trying to predict.
- **Predictor(s)**: the variable(s) used to forecast the output.

What is Simple Linear Regression?

- **Linear Regression** predicts a quantitative response assuming a **linear relationship** with the predictor(s).
- Single linear regression predicts the quantitative output Y on the basis of a **single** predictor variable X.
- Assumes linear relationship:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Parameters or coefficients:
  - β_0: intercept (value of y when $x = 0$)
  - β_1: slope (how much y changes for 1 unit increase in x)
- ϵ: error term


y
Regression Line
x

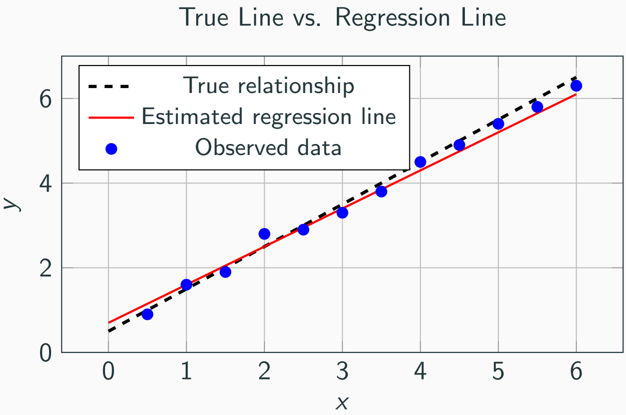


Figure 1:

Comparison of the true relationship (black dashed line) and the estimated regression line (red). The regression line approximates the true line but is not perfectly identical due to random noise in the data. Readapted from An introduction to Statistical Learning with Python.

From Abstract X and Y to Real Data

Substitute the example values:

$$Y \longrightarrow sales$$

$$X \longrightarrow TV$$

From Abstract X and Y to Real Data

Substitute the example values:

$$Y \longrightarrow \mathit{sales}$$

$$X \longrightarrow \mathit{TV}$$

Apply the regression formula:

$$\mathit{sales} = \beta_0 + \beta_1 \times \mathit{TV}$$

Problem

You cannot compute the above example without **knowing the values of the coefficients**!

Estimating the coefficients

- Let's consider a sequence of data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Estimating the coefficients

- Let's consider a sequence of data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- We want to find β_0 and β_1 so that:

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimating the coefficients

- Let's consider a sequence of data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- We want to find β_0 and β_1 so that:

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The most common approach is minimizing the **least squares criterion**.

Residual Sum of Squares (RSS) in Linear Regression

- The **residual** for each observation is

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Residual Sum of Squares (RSS) in Linear Regression

- The **residual** for each observation is

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- The **Residual Sum of Squares (RSS)** is:

$$RSS = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

- Equivalent to:

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Residual Sum of Squares (RSS) in Linear Regression

- The **residual** for each observation is

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- The **Residual Sum of Squares (RSS)** is:

$$RSS = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

- Equivalent to:

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

- To sum up:

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

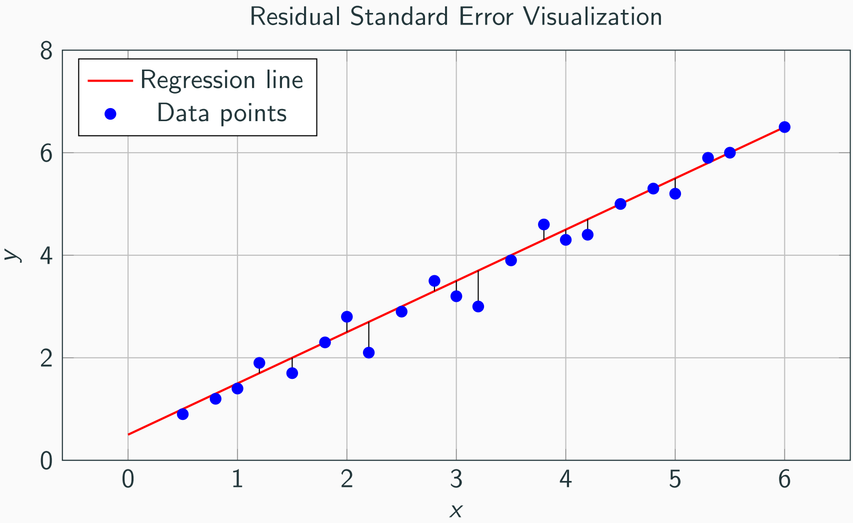


Figure 2: Residual standard error: vertical distances from observed data points to the fitted regression line. Readapted from Statistical Learning with Applications in Python.

It's not over yet!

Next step is **assessing the accuracy of the coefficients** that you estimated!

Assessing the accuracy of the coefficients

- Let's recall the formula:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- The error term ϵ captures what we missed with this model
- Therefore, **it is important to know how well the estimated coefficients are estimated**.

Mean Standard Error (SE)

How accurate is the sample mean $\bar{\mu}$ as an estimate of μ?

- **Standard Error** measures how much a sample statistic (like the mean) would vary if we repeated the experiment multiple times.
- This means that it tells us how accurately a sample statistic represents the population value.

Mean Standard Error (SE)

How accurate is the sample mean $\bar{\mu}$ as an estimate of μ?

- **Standard Error** measures how much a sample statistic (like the mean) would vary if we repeated the experiment multiple times.
- This means that it tells us how accurately a sample statistic represents the population value.
- Example: Let μ be the population mean, and $\bar{\mu}$ be mean of a sample:

$$SE(\bar{\mu})^2 = \frac{s^2}{n}$$

where

  - $s =$ sample standard deviation
  - $n =$ sample size

Mean Standard Error (SE)

How accurate is the sample mean $\bar{\mu}$ as an estimate of μ?

- **Standard Error** measures how much a sample statistic (like the mean) would vary if we repeated the experiment multiple times.
- This means that it tells us how accurately a sample statistic represents the population value.
- Example: Let μ be the population mean, and $\bar{\mu}$ be mean of a sample:

$$SE(\bar{\mu})^2 = \frac{s^2}{n}$$

  where

  - $s =$ sample standard deviation
  - $n =$ sample size

- **Larger samples = smaller SE = more precise estimates**

Standard Errors of the Coefficients

How close are $\bar{\beta}_0$ and $\bar{\beta}_1$ are to the true values $\bar{\beta}_0$ and $\bar{\beta}_1$?

Standard Errors of the Coefficients

How close are $\bar{\beta}_0$ and $\bar{\beta}_1$ are to the true values $\bar{\beta}_0$ and $\bar{\beta}_1$?

Formulas:

$$\mathrm{SE}(\beta_0)^2 = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)$$
$$\mathrm{SE}(\beta_1)^2 = \frac{s^2}{\sum(x_i - \bar{x})^2}$$

Where:

- $s^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$ is the residual variance ($\mathrm{Var}(\epsilon)$)
- x_i = independent variable values
- $\bar{x}$ = mean of x
- n = number of observations

Standard Errors of the Coefficients

How close are $\bar{\beta}_0$ and $\bar{\beta}_1$ are to the true values $\bar{\beta}_0$ and $\bar{\beta}_1$?

Formulas:

$$\mathrm{SE}(\beta_0)^2 = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$
$$\mathrm{SE}(\beta_1)^2 = \frac{s^2}{\sum (x_i - \bar{x})^2}$$

Where:

- $s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$ is the residual variance ($\mathrm{Var}(\epsilon)$)
- x_i = independent variable values
- $\bar{x}$ = mean of x
- n = number of observations

Keep in mind: Smaller standard error = more precise coefficient estimates

What is a Confidence Interval (CI)?

- A confidence interval gives a **range** of plausible values for a population parameter.

What is a Confidence Interval (CI)?

- A confidence interval gives a **range** of plausible values for a population parameter.
- Common choice: 95% confidence interval

> With 95% probability, the range will contain the true unknown value of the parameter.

What is a Confidence Interval (CI)?

- A confidence interval gives a **range** of plausible values for a population parameter.
- Common choice: 95% confidence interval

> With 95% probability, the range will contain the true unknown value of the parameter.

- Formula:

$$\text{CI} = \bar{x} \pm Z \cdot SE$$

where

- $\bar{x}$ = sample mean
- SE = standard error
- Z = critical value from standard normal distribution (1.96 for 95%)

Example of a 95% Confidence Interval

- Sample parameter value: 150
- Standard error: 20
- 95% CI:

$$150 \pm 1.96 \cdot 20 = 150 \pm 39.2$$

- Interval: [110.8, 189.2]

We are 95% confident that the true value lies within this range.

Example

Table 1: Linear Regression of Wage on Age: Key Statistics

Measure	**Value**	**Meaning**
Mean Standard Error (SEM)	50	Average variation of sample mean wage if we resampled
Standard deviation of b_0 (SE b_0)	300	Intercept (2000) varies ±300 due to sample randomness
Standard deviation of b_1 (SE b_1)	20	Slope (150) varies ±20; uncertainty in wage increase per age
Confidence Interval (95%)	[110, 190]	We are 95% confident true slope lies between 110 and 190

Hypothesis testing

Null Hypothesis

H_0: There is no relationship between X and Y.

Hypothesis testing

Null Hypothesis

H_0: There is no relationship between X and Y.

Mathematically, this equals to:

$$\beta_1 = 0$$

Hypothesis testing

Null Hypothesis

H_0: There is no relationship between X and Y.

Mathematically, this equals to:

$$\beta_1 = 0$$

Alternative Hypothesis

H_a: There is a relationship between X and Y.

Hypothesis testing

Null Hypothesis

H_0: There is no relationship between X and Y.

Mathematically, this equals to:

$$\beta_1 = 0$$

Alternative Hypothesis

H_a: There is a relationship between X and Y.

Mathematically, this equals to:

$$\beta_1 \neq 0$$

T-test (for a coefficient)

- A **t-test** checks whether a coefficient is significantly different from 0.
- Translated: it checks **if a predictor actually has an effect**.

T-test (for a coefficient)

- A **t-test** checks whether a coefficient is significantly different from 0.
- Translated: it checks **if a predictor actually has an effect**.
- Test statistic:

$$t = \frac{\text{Estimated Coefficient} - 0}{\text{Standard Error}}$$

T-test (for a coefficient)

- A **t-test** checks whether a coefficient is significantly different from 0.
- Translated: it checks **if a predictor actually has an effect**.
- Test statistic:

$$t = \frac{\text{Estimated Coefficient} - 0}{\text{Standard Error}}$$

- How to interpret values:
  - Large $|t| \rightarrow$ coefficient far from 0 $\rightarrow$ likely real effect
  - Small $|t| \rightarrow$ coefficient close to 0 $\rightarrow$ effect might be random

P-value (for a coefficient)

- **P-value** $=$ probability of observing a t-statistic at least as extreme as the one obtained if H_0 were true.
- Translated: tells you how unusual the result would be if H_0 was true ($=$ no effect of x on y)

P-value (for a coefficient)

- **P-value** $=$ probability of observing a t-statistic at least as extreme as the one obtained if H_0 were true.
- Translated: tells you how unusual the result would be if H_0 was true ($=$ no effect of x on y)
- Interpretation:
  - Small p-value (e.g., < 0.05) $\rightarrow$ reject H_0, coefficient likely significant
  - Large p-value (e.g., > 0.05) $\rightarrow$ fail to reject H_0, no strong evidence

Example

Regression of wage on age: $wage = \beta_0 + \beta_1 \times age + \epsilon$ Random example values: $wage = 2000 + 150 \times age + \epsilon$

Table 2: Linear Regression of Wage on Age: Key Statistics

Measure	**Value**	**Meaning**
Mean Standard Error (SEM)	50	Average variation of sample mean wage if we resampled
Standard deviation of b_0 (SE b_0)	300	Intercept (2000) varies ±300 due to sample randomness
Standard deviation of b_1 (SE b_1)	20	Slope (150) varies ±20; uncertainty in wage increase per age
Confidence Interval (95%)	[110, 190]	We are 95% confident true slope lies between 110 and 190
T-test (for b_1)	7.5	Tests if age significantly affects wage; higher t → more evidence
P-value (for b_1)	0.0001	Very low → strong evidence that age significantly affects wage

These values are not real, they just deem as an example. We will see real values in the practical session.

We talked about how to evaluate the accuracy of the coefficients. What about the **accuracy of the model?**

- Coefficient accuracy: how well is the relationship between X and Y modeled?
- Model accuracy: how well does the model predict?

Assessing the accuracy of the model

The **error** of an observation is the deviation of the observed value from the true value of a quantity of interest (for example, a population mean).

Assessing the accuracy of the model

The **error** of an observation is the deviation of the observed value from the true value of a quantity of interest (for example, a population mean).

The **residual** is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean).

Source: Wikipedia

Assessing the accuracy of the model

- **Residual Standard Error (RSE)**
- **Coefficient of Determination (R-Squared)**

Residual Standard Error

The **Residual Standard Error (RSE)** is the standard deviation of **observed values** from the predicted values.

Residual Standard Error

The **Residual Standard Error (RSE)** is the standard deviation of **observed values** from the predicted values.

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

Residual Standard Error

The **Residual Standard Error (RSE)** is the standard deviation of **observed values** from the predicted values.

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

- The **smaller** your RSE, the better your model.

R-squared

The **coefficient of determination**, denoted as r^2, is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

Source: Wikipedia

R-squared

The **coefficient of determination**, denoted as r^2, is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

Source: Wikipedia

$$1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}}$$

- $R^2 = 1$: the regression predictions perfectly fit the data
- $R^2 = 0$: the regression does not fit the data

Example

Table 3: Linear Regression of Wage on Age: Key Statistics

Measure	**Value**	**Meaning**
Mean Standard Error (SEM)	50	Average variation of sample mean wage
Standard deviation of b_0 (SE b_0)	300	Intercept (2000) varies ±300 due to sa
Standard deviation of b_1 (SE b_1)	20	Slope (150) varies ±20; uncertainty in
Confidence Interval (95%)	[110, 190]	We are 95% confident true slope lies be
T-test (for b_1)	7.5	Tests if age significantly affects wage; h
P-value (for b_1)	0.0001	Very low ⟶ strong evidence that age s
Residual Standard Error (RSS)	12,000	Difference between observed and predic
0.85	85% of wage variation explained by age; higher ⟶ better model fit	

What happens when we have more than one predictor?

What is multiple linear regression?

- Models the relationship between a dependent variable Y and multiple independent variables $X_1, X_2, \ldots, X_p$:

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

What is multiple linear regression?

- Models the relationship between a dependent variable Y and multiple independent variables $X_1, X_2, \ldots, X_p$:

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

  where:

  - Y = dependent variable (outcome)
  - $X_1, X_2, \ldots, X_p$ = independent variables (predictors)
  - β_0 = intercept
  - $\beta_1, \ldots, \beta_p$ = regression coefficients
  - ε = error term

What is multiple linear regression?

- Models the relationship between a dependent variable Y and multiple independent variables $X_1, X_2, \ldots, X_p$:

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

  where:

  - Y = dependent variable (outcome)
  - $X_1, X_2, \ldots, X_p$ = independent variables (predictors)
  - β_0 = intercept
  - $\beta_1, \ldots, \beta_p$ = regression coefficients
  - ε = error term

Remember

The coefficient β_i represents the change in Y for a one-unit increase in X_i *(remember single linear regression!)*, **holding other variables constant**.

From Abstract X and Y to Real Data

Substitute the example values:

$$Y \longrightarrow \textit{sales}$$

$$X_1 \longrightarrow \textit{TV}$$

$$X_2 \longrightarrow \textit{radio}$$

$$X_3 \longrightarrow \textit{newspaper}$$

From Abstract X and Y to Real Data

Substitute the example values:

$$Y \longrightarrow \mathit{sales}$$

$$X_1 \longrightarrow \mathit{TV}$$

$$X_2 \longrightarrow \mathit{radio}$$

$$X_3 \longrightarrow \mathit{newspaper}$$

Apply the regression formula:

$$\mathit{sales} = \beta_0 + \beta_1 \times \mathit{TV} + \beta_2 \times \mathit{radio} + \beta_3 \times \mathit{newspaper}$$

Frame Title

Remember

Correlation among predictors can be problematic: low correlation is desired.

Estimating the coefficients

- **Minimizing least squares** as in simple linear regression

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

where:

- y_i = actual value of the i-th observation
- $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ = predicted value
- n = number of observations

Estimating the coefficients

- **Minimizing least squares** as in simple linear regression

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

where:

  - y_i = actual value of the i-th observation
  - $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ = predicted value
  - n = number of observations

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \times x_{i1} - \hat{\beta}_2 \times x_{i2} - \ldots - \hat{\beta}_n \times x_{in} \right)$$

Some assumptions of Linear Regression

Linearity

- **Linear relationship** between the dependent and independent variable(s)

Linearity

- **Linear relationship** between the dependent and independent variable(s)

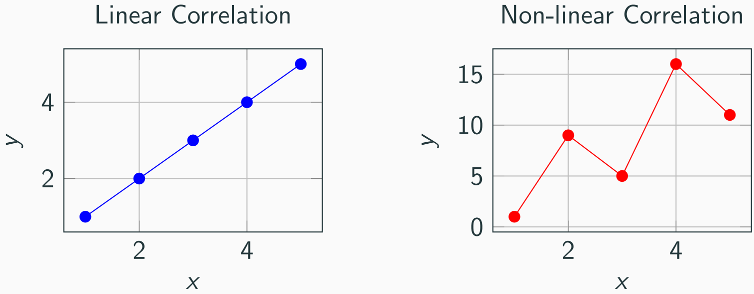


Figure 3: Left: Linear correlation. Right: Non-linear correlation.

Linearity

- **Linear relationship** between the dependent and independent variable(s)

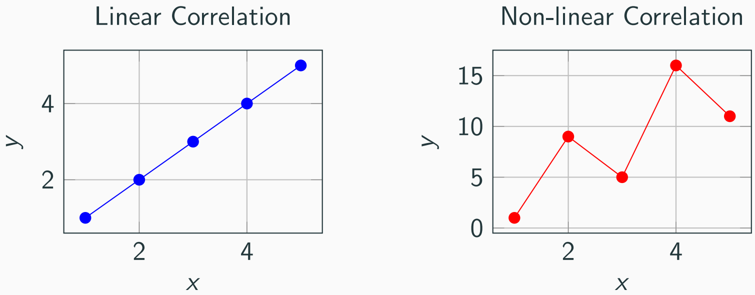


Figure 3: Left: Linear correlation. Right: Non-linear correlation.

- **How to check**: With a scatterplot.

No multicollinearity

- In multiple linear regression, independent variables must be **not too highly correlated with each other**.

No multicollinearity

- In multiple linear regression, independent variables must be **not too highly correlated with each other**.
- **How to check**: Correlation matrix.

No multicollinearity

- In multiple linear regression, independent variables must be **not too highly correlated with each other**.
- **How to check**: Correlation matrix.

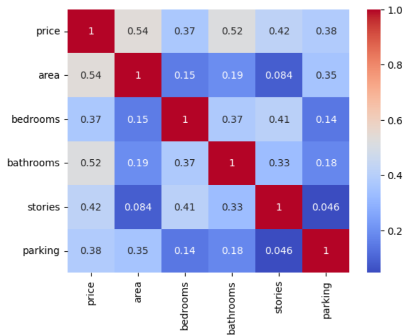


Figure 4: From last week's lab: a heatmap representing correlation values.

Homoscedasticy

- **Homoscedasticy** means that the variance of the errors is roughly the same across all values of the predictor(s) in your regression model.

Homoscedasticy

- **Homoscedasticy** means that the variance of the errors is roughly the same across all values of the predictor(s) in your regression model.
- **How to check** (intuitively, without formal tests): Plot the residuals and look for (approximately) even variance.

Other limitations

- Linear regression involves other limitations and assumptions, **not treated in this course due to its more limited scope**.

How can you elaborate a meaningful data analysis pipeline that involves linear regression?

Adapt to your own research question

- Elaborate a **public policy research question** related to your own studies and knowledge
- This way, **errors in your model will still be meaningful.**
- Example:
  - I evaluate **wage** in relation to **age**: this is meaningful!
  - I evaluate **wage** in relation to **height**: I might detect some patterns (for example, due to biases in my samples), but this research question makes no sense!
- This seems obvious, yes; but I am sure you can think of several field-specific examples, and of correlation that people who do not study public policies do not know!