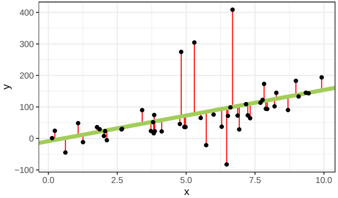
Interpretable Machine Learning

Interpretable Models 1
Linear Regression Model



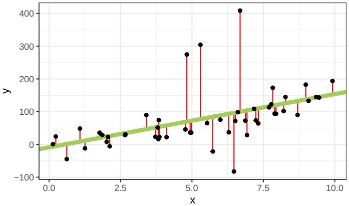
Learning goals

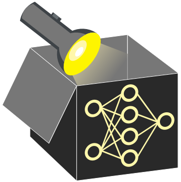
- LM basics and assumptions
- Interpretation of main effects in LM
- What are significant features?

LINEAR REGRESSION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

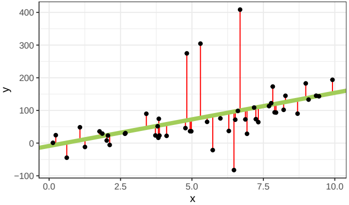
- y: target / output
- ϵ: remaining error / residual
- θ_j: weight of input feature x_j (intercept θ_0)
  $\rightsquigarrow$ model consists of $p + 1$ weights





LINEAR REGRESSION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$



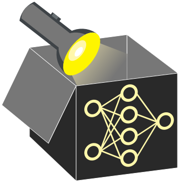
- y: target / output
- ϵ: remaining error / residual
- θ_j: weight of input feature x_j (intercept θ_0)
  $\rightsquigarrow$ model consists of $p + 1$ weights

Properties and assumptions

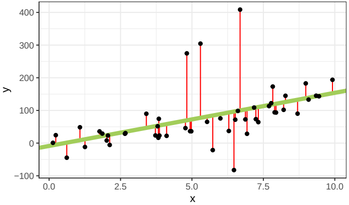
▶ "Faraway, Ch. 7" 2002

- **Linear** relationship between features and target

LINEAR REGRESSION



$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$



- y: target / output
- ϵ: remaining error / residual
- θ_j: weight of input feature x_j (intercept θ_0)
  $\rightsquigarrow$ model consists of $p + 1$ weights

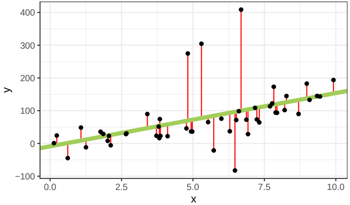
Properties and assumptions

▶ "Faraway, Ch. 7" 2002

- **Linear** relationship between features and target
- ϵ and $y|\mathbf{x}$ are **normally** distributed with **constant variance** (homoscedastic)
  $\rightsquigarrow \epsilon \sim N(0, \sigma^2) \;\Rightarrow\; (y|\mathbf{x}) \sim N(\mathbf{x}^\top\theta, \sigma^2)$
  $\rightsquigarrow$ if violated, inference-based metrics (e.g., p-values) are invalid

LINEAR REGRESSION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$



- y: target / output
- ϵ: remaining error / residual
- θ_j: weight of input feature x_j (intercept θ_0)
  $\rightsquigarrow$ model consists of $p + 1$ weights

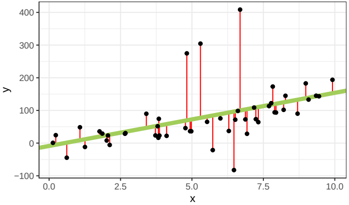
Properties and assumptions

▸ "Faraway, Ch. 7" 2002

- **Linear** relationship between features and target
- ϵ and $y|\mathbf{x}$ are **normally** distributed with **constant variance** (homoscedastic)
  $\rightsquigarrow \epsilon \sim N(0, \sigma^2) \;\Rightarrow\; (y|\mathbf{x}) \sim N(\mathbf{x}^\top\theta, \sigma^2)$
  $\rightsquigarrow$ if violated, inference-based metrics (e.g., p-values) are invalid
- Independence of observations (e.g., no repeated measurements)

LINEAR REGRESSION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$



- y: target / output
- ϵ: remaining error / residual
- θ_j: weight of input feature x_j (intercept θ_0)
  $\rightsquigarrow$ model consists of $p + 1$ weights

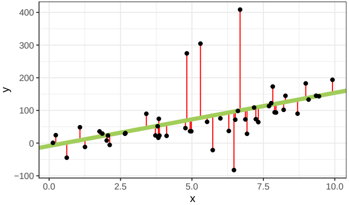
Properties and assumptions

▸ "Faraway, Ch. 7" 2002

- **Linear** relationship between features and target
- ϵ and $y|\mathbf{x}$ are **normally** distributed with **constant variance** (homoscedastic)
  $\rightsquigarrow \epsilon \sim N(0, \sigma^2) \;\Rightarrow\; (y|\mathbf{x}) \sim N(\mathbf{x}^\top \theta, \sigma^2)$
  $\rightsquigarrow$ if violated, inference-based metrics (e.g., p-values) are invalid
- Independence of observations (e.g., no repeated measurements)
- Features x_j independent from error term ϵ

LINEAR REGRESSION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$



- y: target / output
- ϵ: remaining error / residual
- θ_j: weight of input feature x_j (intercept θ_0)
  $\rightsquigarrow$ model consists of $p + 1$ weights

Properties and assumptions

▸ "Faraway, Ch. 7" 2002

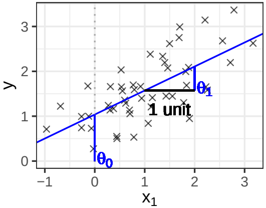
- **Linear** relationship between features and target
- ϵ and $y|\mathbf{x}$ are **normally** distributed with **constant variance** (homoscedastic)
  $\rightsquigarrow \epsilon \sim N(0, \sigma^2) \;\Rightarrow\; (y|\mathbf{x}) \sim N(\mathbf{x}^\top\theta, \sigma^2)$
  $\rightsquigarrow$ if violated, inference-based metrics (e.g., p-values) are invalid
- Independence of observations (e.g., no repeated measurements)
- Features x_j independent from error term ϵ
- No or little multicollinearity (i.e., no strong feature correlations)

LINEAR REGRESSION - INTERPRETATION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

Interpretation of weights (**feature effects**) depend on type of feature:

- **Numerical** x_j: Increasing x_j by one unit changes outcome by θ_j, ceteris paribus
  (*ceteris paribus* (c.p.) means "everything else held constant".)

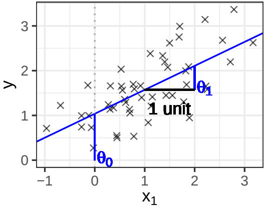


LINEAR REGRESSION - INTERPRETATION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

Interpretation of weights (**feature effects**) depend on type of feature:

- **Numerical** x_j: Increasing x_j by one unit changes outcome by θ_j, ceteris paribus
  (*ceteris paribus* (c.p.) means "everything else held constant".)
- **Binary** x_j: Weight θ_j is active or not (multiplication with 1 or 0)
  $\rightsquigarrow$ reference category $x_j = 0$



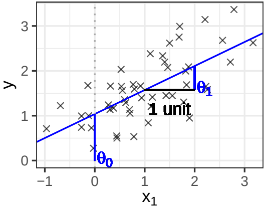
LINEAR REGRESSION - INTERPRETATION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

Interpretation of weights (**feature effects**) depend on type of feature:

- **Numerical** x_j: Increasing x_j by one unit changes outcome by θ_j, ceteris paribus
  (*ceteris paribus* (c.p.) means "everything else held constant".)
- **Binary** x_j: Weight θ_j is active or not (multiplication with 1 or 0)
  $\rightsquigarrow$ reference category $x_j = 0$
- **Categorical feature** x_j **with** L **categories**:
  - Create $L - 1$ one-hot-encoded features $x_{j,1}, \dots, x_{j,L-1}$ (each having its own weight)
  - Left out cat. is reference ($\hat{=}$ dummy encoding)

  $\rightsquigarrow$ Interpretation: Outcome changes by $\theta_{j,i}$ for category i compared to reference cat., c.p.



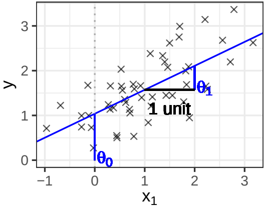
LINEAR REGRESSION - INTERPRETATION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

Interpretation of weights (**feature effects**) depend on type of feature:

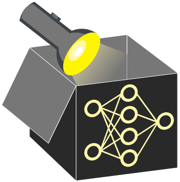
- **Numerical** x_j: Increasing x_j by one unit changes outcome by θ_j, ceteris paribus
  (*ceteris paribus* (c.p.) means "everything else held constant".)
- **Binary** x_j: Weight θ_j is active or not (multiplication with 1 or 0)
  $\rightsquigarrow$ reference category $x_j = 0$
- **Categorical feature** x_j **with** L **categories**:
  - Create $L - 1$ one-hot-encoded features $x_{j,1}, \ldots, x_{j,L-1}$ (each having its own weight)
  - Left out cat. is reference ($\hat{=}$ dummy encoding)

  $\rightsquigarrow$ Interpretation: Outcome changes by $\theta_{j,i}$ for category i compared to reference cat., c.p.



- **Intercept** θ_0: Expected outcome if all feature values are set to 0

LINEAR REGRESSION - INTERPRETATION

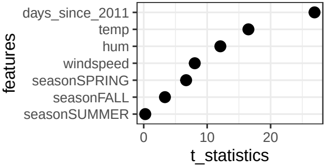


$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

Feature importance:

- Absolute **t-statistic** value: $\hat{\theta}_j$ scaled with standard error ($SE(\hat{\theta}_j) \hat{=}$ reliability of estimate)

$$|t_{\hat{\theta}_j}| = \left| \frac{\hat{\theta}_j}{SE(\hat{\theta}_j)} \right|$$



- High t-values ⇒ important (significant) feat.

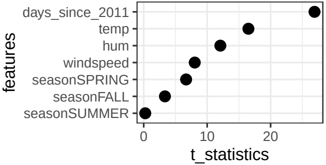
LINEAR REGRESSION - INTERPRETATION

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

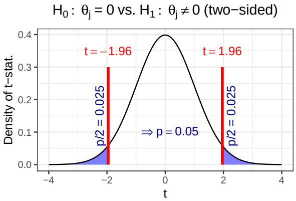
Feature importance:

- Absolute **t-statistic** value: $\hat{\theta}_j$ scaled with standard error ($SE(\hat{\theta}_j) \hat{=}$ reliability of estimate)

$$|t_{\hat{\theta}_j}| = \left| \frac{\hat{\theta}_j}{SE(\hat{\theta}_j)} \right|$$



- High t-values $\Rightarrow$ important (significant) feat.
- **p-value**: probability of obtaining a more extreme test statistic assuming H_0 is correct (here: $\theta_j = 0$, i.e., feat. j not significant) $\rightsquigarrow$ High $|t| \Rightarrow$ small p-val. (speak against H_0)



EXAMPLE: LIN. REGRESSION - MAIN EFFECTS

Bike data: predict no. of rented bikes using 4 numeric, 1 cat. feat. (season)

$$
\begin{aligned}
\hat{y} = & \hat{\theta}_0 + \hat{\theta}_1 \mathbb{1}_{x_{season}=SPRING} + \\
& \hat{\theta}_2 \mathbb{1}_{x_{season}=SUMMER} + \\
& \hat{\theta}_3 \mathbb{1}_{x_{season}=FALL} + \hat{\theta}_4 x_{temp} + \\
& \hat{\theta}_5 x_{hum} + \hat{\theta}_6 x_{windspeed} + \\
& \hat{\theta}_7 x_{days\_since\_2011}
\end{aligned}
$$

	Weights	SE	t-stat.	p-val.
(Intercept)	3229.3	220.6	14.6	0.00
seasonSPRING	862.0	129.0	6.7	0.00
seasonSUMMER	41.6	170.2	0.2	0.81
seasonFALL	390.1	116.6	3.3	0.00
temp	120.5	7.3	16.5	0.00
hum	-31.1	2.6	-12.1	0.00
windspeed	-56.9	7.1	-8.0	0.00
days_since_2011	4.9	0.2	26.9	0.00

EXAMPLE: LIN. REGRESSION - MAIN EFFECTS

Bike data: predict no. of rented bikes using 4 numeric, 1 cat. feat. (season)

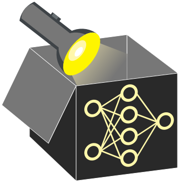
$$\begin{aligned}\hat{y} =\, & \hat{\theta}_0 + \hat{\theta}_1 \mathbb{1}_{x_{season}=SPRING} + \\ & \hat{\theta}_2 \mathbb{1}_{x_{season}=SUMMER} + \\ & \hat{\theta}_3 \mathbb{1}_{x_{season}=FALL} + \hat{\theta}_4 x_{temp} + \\ & \hat{\theta}_5 x_{hum} + \hat{\theta}_6 x_{windspeed} + \\ & \hat{\theta}_7 x_{days\_since\_2011}\end{aligned}$$

	Weights	SE	t-stat.	p-val.
(Intercept)	3229.3	220.6	14.6	0.00
seasonSPRING	862.0	129.0	6.7	0.00
seasonSUMMER	41.6	170.2	0.2	0.81
seasonFALL	390.1	116.6	3.3	0.00
temp	120.5	7.3	16.5	0.00
hum	-31.1	2.6	-12.1	0.00
windspeed	-56.9	7.1	-8.0	0.00
days_since_2011	4.9	0.2	26.9	0.00

Interpretation:

- **Intercept**: If all feature values are 0 (and season is `WINTER` $\hat{=}$ reference cat.), the expected number of bike rentals is $\hat{\theta}_0 = 3229.3$

EXAMPLE: LIN. REGRESSION - MAIN EFFECTS



Bike data: predict no. of rented bikes using 4 numeric, 1 cat. feat. (season)

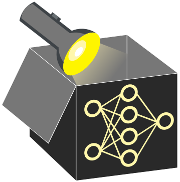
$$\begin{aligned}\hat{y} =& \hat{\theta}_0 + \hat{\theta}_1 \mathbb{1}_{x_{season}=SPRING} + \\ & \hat{\theta}_2 \mathbb{1}_{x_{season}=SUMMER} + \\ & \hat{\theta}_3 \mathbb{1}_{x_{season}=FALL} + \hat{\theta}_4 x_{temp} + \\ & \hat{\theta}_5 x_{hum} + \hat{\theta}_6 x_{windspeed} + \\ & \hat{\theta}_7 x_{days\_since\_2011}\end{aligned}$$

	Weights	SE	t-stat.	p-val.
(Intercept)	3229.3	220.6	14.6	0.00
seasonSPRING	862.0	129.0	6.7	0.00
seasonSUMMER	41.6	170.2	0.2	0.81
seasonFALL	390.1	116.6	3.3	0.00
temp	120.5	7.3	16.5	0.00
hum	-31.1	2.6	-12.1	0.00
windspeed	-56.9	7.1	-8.0	0.00
days_since_2011	4.9	0.2	26.9	0.00

Interpretation:

- **Intercept**: If all feature values are 0 (and season is `WINTER` $\hat{=}$ reference cat.), the expected number of bike rentals is $\hat{\theta}_0 = 3229.3$
- **Categ.**: Rentals in `SPRING` are by $\hat{\theta}_1 = 862$ higher than in `WINTER`, c.p.

EXAMPLE: LIN. REGRESSION - MAIN EFFECTS



Bike data: predict no. of rented bikes using 4 numeric, 1 cat. feat. (season)

$$\begin{aligned}\hat{y} =& \hat{\theta}_0 + \hat{\theta}_1 \mathbb{1}_{x_{season}=SPRING} + \\ & \hat{\theta}_2 \mathbb{1}_{x_{season}=SUMMER} + \\ & \hat{\theta}_3 \mathbb{1}_{x_{season}=FALL} + \hat{\theta}_4 x_{temp} + \\ & \hat{\theta}_5 x_{hum} + \hat{\theta}_6 x_{windspeed} + \\ & \hat{\theta}_7 x_{days\_since\_2011}\end{aligned}$$

	Weights	SE	t-stat.	p-val.
(Intercept)	3229.3	220.6	14.6	0.00
seasonSPRING	862.0	129.0	6.7	0.00
seasonSUMMER	41.6	170.2	0.2	0.81
seasonFALL	390.1	116.6	3.3	0.00
temp	120.5	7.3	16.5	0.00
hum	-31.1	2.6	-12.1	0.00
windspeed	-56.9	7.1	-8.0	0.00
days_since_2011	4.9	0.2	26.9	0.00

Interpretation:

- **Intercept**: If all feature values are 0 (and season is `WINTER` $\hat{=}$ reference cat.), the expected number of bike rentals is $\hat{\theta}_0 = 3229.3$
- **Categ.**: Rentals in `SPRING` are by $\hat{\theta}_1 = 862$ higher than in `WINTER`, c.p.
- **Numerical**: Rentals increase by $\hat{\theta}_4 = 120.5$ if `temp` increases by 1 °C, c.p.