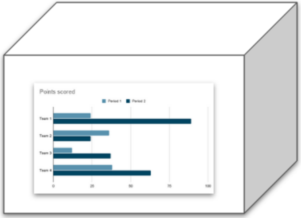
# Interpretable Machine Learning

# Linear Regression Model



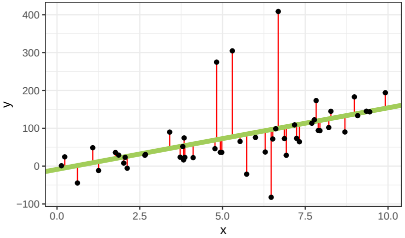

**Learning goals**

- Interpretation of main effects in LM
- Inclusion of high-order and interaction effects
- Regularization via LASSO

# LINEAR REGRESSION

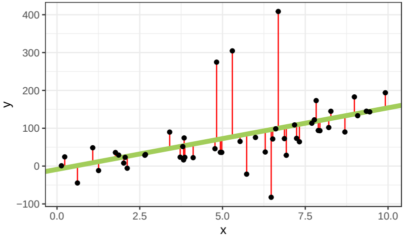
$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon$$

- $y$: target / output
- $\epsilon$: remaining error / residual (e.g., due to noise)
- $\theta_j$: weight of input feature $x_j$ (with intercept $\theta_0$)
  $\rightsquigarrow$ model consists of $p + 1$ weights



# LINEAR REGRESSION

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

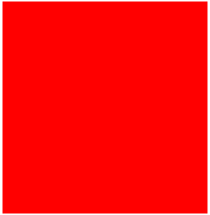



- $y$: target / output
- $\epsilon$: remaining error / residual (e.g., due to noise)
- $\theta_j$: weight of input feature $x_j$ (with intercept $\theta_0$)
  $\rightsquigarrow$ model consists of $p + 1$ weights

## Properties and assumptions

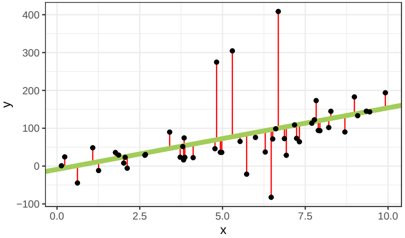
▸ Faraway (2002), Ch. 7 ▸ Checking assumptions in R & Python

- **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
- $\epsilon$: remaining error / residual (e.g., due to noise)
- $\theta_j$: weight of input feature $x_j$ (with intercept $\theta_0$)
  $\rightsquigarrow$ model consists of $p + 1$ weights

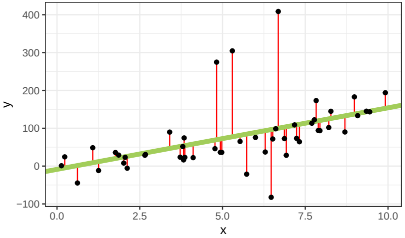
## Properties and assumptions

▸ Faraway (2002), Ch. 7 ▸ Checking assumptions in R & Python

- **Linear** relationship between features and target
- $\epsilon$ 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
- $\epsilon$: remaining error / residual (e.g., due to noise)
- $\theta_j$: weight of input feature $x_j$ (with intercept $\theta_0$)
  $\rightsquigarrow$ model consists of $p + 1$ weights

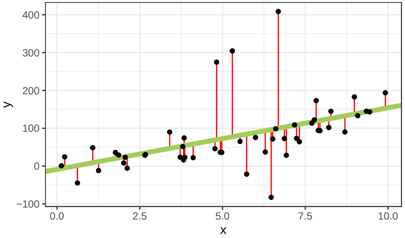
## Properties and assumptions

▸ Faraway (2002), Ch. 7 ▸ Checking assumptions in R & Python

- **Linear** relationship between features and target
- $\epsilon$ 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

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



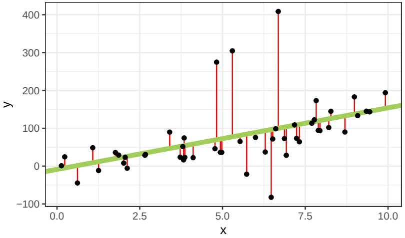

- $y$: target / output
- $\epsilon$: remaining error / residual (e.g., due to noise)
- $\theta_j$: weight of input feature $x_j$ (with intercept $\theta_0$)
  $\rightsquigarrow$ model consists of $p+1$ weights

**Properties and assumptions** ▸ Faraway (2002), Ch. 7 ▸ Checking assumptions in R & Python

- **Linear** relationship between features and target
- $\epsilon$ 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)
- Independence of features $x_j$ with error term $\epsilon$

# LINEAR REGRESSION

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

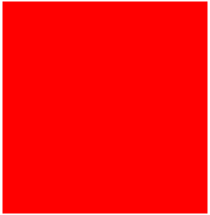



- $y$: target / output
- $\epsilon$: remaining error / residual (e.g., due to noise)
- $\theta_j$: weight of input feature $x_j$ (with intercept $\theta_0$)
  $\rightsquigarrow$ model consists of $p + 1$ weights

## Properties and assumptions

Faraway (2002), Ch. 7 | Checking assumptions in R & Python

- **Linear** relationship between features and target
- $\epsilon$ 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)
- Independence of features $x_j$ with error term $\epsilon$
- No or little multicollinearity (i.e., no strong feature correlations)

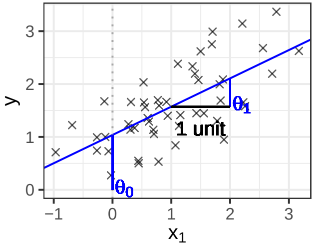


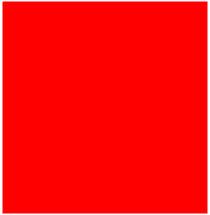
# LINEAR REGRESSION - INTERPRETATION

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

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

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



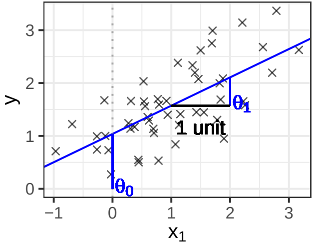



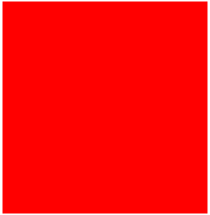
# LINEAR REGRESSION - INTERPRETATION

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

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

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



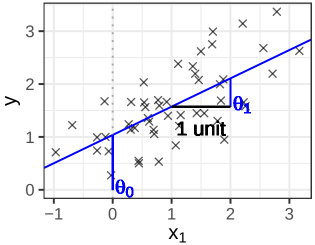



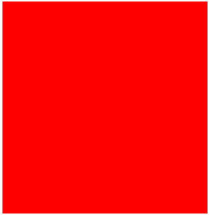
# LINEAR REGRESSION - INTERPRETATION

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

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

- **Numerical** $x_j$: Increasing $x_j$ by one unit changes outcome by $\theta_j$, ceteris paribus (c.p.) (*ceteris paribus* means "everything else held constant".)
- **Binary** $x_j$: Weight $\theta_j$ is active or not (multiplication with 1 or 0) where 0 is reference category
- **Categorical** $x_j$ **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 category is reference ($\hat{=}$ dummy encoding)
  $\rightsquigarrow$ Interpretation: Outcome changes by $\theta_{j,l}$ for category $l$ compared to reference cat., c.p.



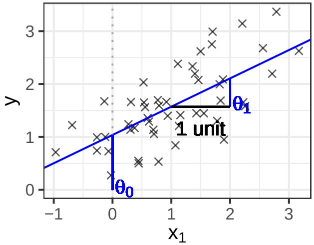



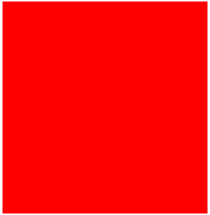
# LINEAR REGRESSION - INTERPRETATION

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

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

- **Numerical** $x_j$: Increasing $x_j$ by one unit changes outcome by $\theta_j$, ceteris paribus (c.p.) (*ceteris paribus* means "everything else held constant".)
- **Binary** $x_j$: Weight $\theta_j$ is active or not (multiplication with 1 or 0) where 0 is reference category
- **Categorical** $x_j$ **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 category is reference ($\hat{=}$ dummy encoding) $\rightsquigarrow$ Interpretation: Outcome changes by $\theta_{j,l}$ for category $l$ compared to reference cat., c.p.
- **Intercept** $\theta_0$: Expected outcome if all feature values are set to 0



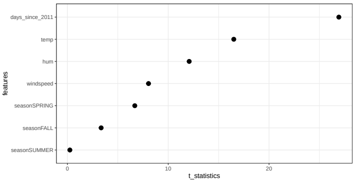
# LINEAR REGRESSION - INTERPRETATION

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

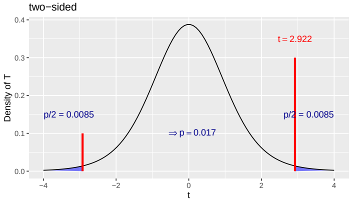
**Feature importance**:

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

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




- High values indicate important (i.e. significant) features
- **p-value**: probability of obtaining a test statistic that is more extreme (values that speak against $H_0$) as the test statistic computed from the sample, assuming $H_0$ is correct.
- The smaller the p-value, the less likely it is to obtain a more extreme test statistic.

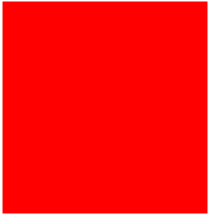


# EXAMPLE: LINEAR REGRESSION - MAIN EFFECTS

**Bike data**: predict number of rented bikes using 4 numeric and 1 categorical feature (season)

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

| | 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 |



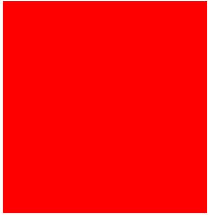
# EXAMPLE: LINEAR REGRESSION - MAIN EFFECTS

**Bike data**: predict number of rented bikes using 4 numeric and 1 categorical feature (season)

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

| | 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 $\hat{\theta}_0 = 3229.3$



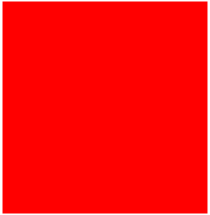
# EXAMPLE: LINEAR REGRESSION - MAIN EFFECTS

**Bike data**: predict number of rented bikes using 4 numeric and 1 categorical feature (season)

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

| | 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 $\hat{\theta}_0 = 3229.3$
- **Interpretation categorical**: Rentals in `SPRING` are by $\hat{\theta}_1 = 862$ higher than in `WINTER`, c.p.



# EXAMPLE: LINEAR REGRESSION - MAIN EFFECTS

**Bike data**: predict number of rented bikes using 4 numeric and 1 categorical feature (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 $\hat{\theta}_0 = 3229.3$
- **Interpretation categorical**: Rentals in `SPRING` are by $\hat{\theta}_1 = 862$ higher than in `WINTER`, c.p.
- **Interpretation numerical**: Rentals increase by $\hat{\theta}_4 = 120.5$ if `temp` increases by 1 °C, c.p.

# LINEAR REGRESSION - INTERACTION AND HIGH-ORDER EFFECTS

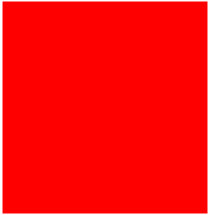
$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon$$

Equation above can be extended (polynomial regression) by including

- **high-order effects** which have their own weights
  $\rightsquigarrow$ e.g., quadratic effect: $\theta_{x_j^2} \cdot x_j^2$
- **interaction effects** as the product of multiple feat.
  $\rightsquigarrow$ e.g., 2-way interaction: $\theta_{x_i, x_j} \cdot x_i \cdot x_j$

**Bike Data**

| Method | $R^2$ | adj. $R^2$ |
|---|---|---|
| Simple LM | 0.85 | 0.84 |
| Higher-order | 0.87 | 0.87 |
| Interaction | 0.96 | 0.93 |



# LINEAR REGRESSION - INTERACTION AND HIGH-ORDER EFFECTS

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \epsilon$$

Equation above can be extended (polynomial regression) by including

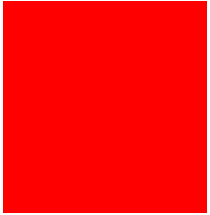
- **high-order effects** which have their own weights
  $\rightsquigarrow$ e.g., quadratic effect: $\theta_{x_j^2} \cdot x_j^2$
- **interaction effects** as the product of multiple feat.
  $\rightsquigarrow$ e.g., 2-way interaction: $\theta_{x_i, x_j} \cdot x_i \cdot x_j$

**Bike Data**

| Method | $R^2$ | adj. $R^2$ |
|---|---|---|
| Simple LM | 0.85 | 0.84 |
| Higher-order | 0.87 | 0.87 |
| Interaction | 0.96 | 0.93 |

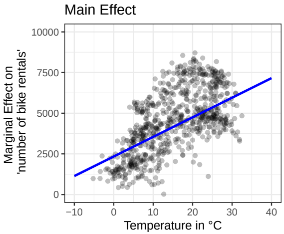
Implications of including high-order and interaction effects:

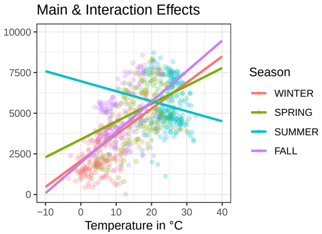
- Both make the model more flexible but also less interpretable
  $\rightsquigarrow$ More weights to interpret
- Both need to be specified manually (inconvenient and sometimes infeasible)
  $\rightsquigarrow$ Other ML models learn them often automatically
- Marginal effect of a feature cannot be interpreted by single weights anymore
  $\rightsquigarrow$ Feature $x_j$ occurs multiple times (with different weights) in equation



# EXAMPLE: LINEAR REGRESSION - INTERACTION EFFECT

**Example**: Interaction between `temp` and `season` will affect marginal effect of `temp`

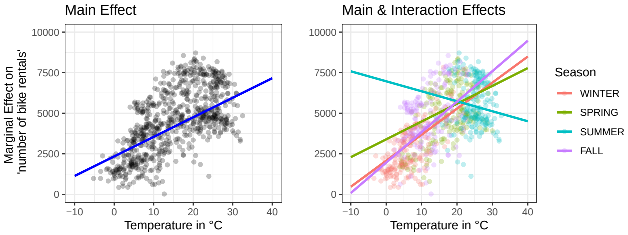



| | Weights |
|---|---|
| (Intercept) | 3453.9 |
| seasonSPRING | 1317.0 |
| seasonSUMMER | 4894.1 |
| seasonFALL | -114.2 |
| temp | 160.5 |
| hum | -37.6 |
| windspeed | -61.9 |
| days_since_2011 | 4.9 |
| seasonSPRING:temp | -50.7 |
| seasonSUMMER:temp | -222.0 |
| seasonFALL:temp | 27.2 |

# EXAMPLE: LINEAR REGRESSION - INTERACTION EFFECT

**Example**: Interaction between `temp` and `season` will affect marginal effect of `temp`



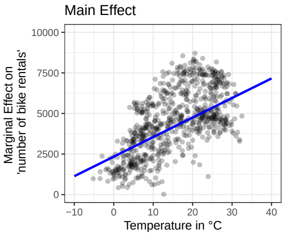

| | Weights |
|---|---|
| (Intercept) | 3453.9 |
| seasonSPRING | 1317.0 |
| seasonSUMMER | 4894.1 |
| seasonFALL | -114.2 |
| temp | 160.5 |
| hum | -37.6 |
| windspeed | -61.9 |
| days_since_2011 | 4.9 |
| seasonSPRING:temp | -50.7 |
| seasonSUMMER:temp | -222.0 |
| seasonFALL:temp | 27.2 |

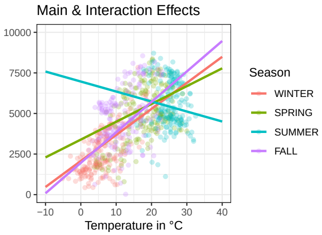
**Interpretation**: If `temp` increases by 1 °C, bike rentals

- increase by 160.5 in `WINTER` (reference)

# EXAMPLE: LINEAR REGRESSION - INTERACTION EFFECT

**Example**: Interaction between `temp` and `season` will affect marginal effect of `temp`



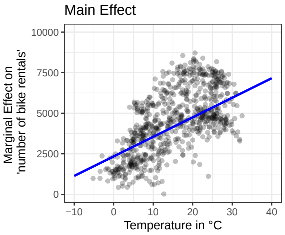

| | Weights |
|---:|---:|
| (Intercept) | 3453.9 |
| seasonSPRING | 1317.0 |
| seasonSUMMER | 4894.1 |
| seasonFALL | -114.2 |
| temp | 160.5 |
| hum | -37.6 |
| windspeed | -61.9 |
| days_since_2011 | 4.9 |
| seasonSPRING:temp | -50.7 |
| seasonSUMMER:temp | -222.0 |
| seasonFALL:temp | 27.2 |

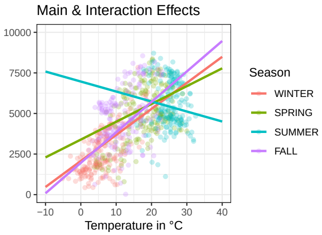
**Interpretation**: If `temp` increases by 1 °C, bike rentals

- increase by 160.5 in `WINTER` (reference)
- increase by 109.8 (= 160.5 - 50.7) in `SPRING`

# EXAMPLE: LINEAR REGRESSION - INTERACTION EFFECT

**Example**: Interaction between `temp` and `season` will affect marginal effect of `temp`



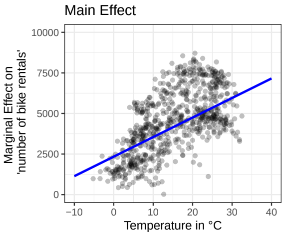

| | Weights |
|---|---|
| (Intercept) | 3453.9 |
| seasonSPRING | 1317.0 |
| seasonSUMMER | 4894.1 |
| seasonFALL | -114.2 |
| temp | 160.5 |
| hum | -37.6 |
| windspeed | -61.9 |
| days_since_2011 | 4.9 |
| seasonSPRING:temp | -50.7 |
| seasonSUMMER:temp | -222.0 |
| seasonFALL:temp | 27.2 |

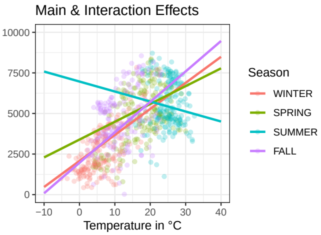
**Interpretation**: If `temp` increases by 1 °C, bike rentals

- increase by 160.5 in `WINTER` (reference)
- increase by 109.8 (= 160.5 - 50.7) in `SPRING`
- decrease by -61.5 (= 160.5 - 222) in `SUMMER`

# EXAMPLE: LINEAR REGRESSION - INTERACTION EFFECT

**Example**: Interaction between `temp` and `season` will affect marginal effect of `temp`



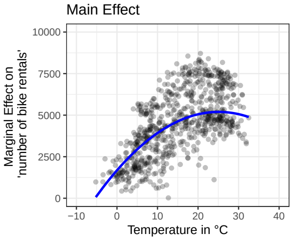

| | Weights |
|---|---|
| (Intercept) | 3453.9 |
| seasonSPRING | 1317.0 |
| seasonSUMMER | 4894.1 |
| seasonFALL | -114.2 |
| temp | 160.5 |
| hum | -37.6 |
| windspeed | -61.9 |
| days_since_2011 | 4.9 |
| seasonSPRING:temp | -50.7 |
| seasonSUMMER:temp | -222.0 |
| seasonFALL:temp | 27.2 |

**Interpretation**: If `temp` increases by 1 °C, bike rentals

- increase by 160.5 in `WINTER` (reference)
- increase by 109.8 (= 160.5 - 50.7) in `SPRING`
- decrease by -61.5 (= 160.5 - 222) in `SUMMER`
- increase by 187.7 (= 160.5 + 27.2) in `FALL`

# EXAMPLE: LINEAR REGRESSION - QUADRATIC EFFECT

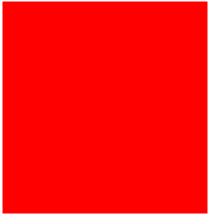
**Example**: Adding quadratic effect for `temp`




| | Weights |
|---|---|
| (Intercept) | 3094.1 |
| seasonSPRING | 619.2 |
| seasonSUMMER | 284.6 |
| seasonFALL | 123.1 |
| hum | -36.4 |
| windspeed | -65.7 |
| days_since_2011 | 4.7 |
| temp | 280.2 |
| $\text{temp}^2$ | -5.6 |

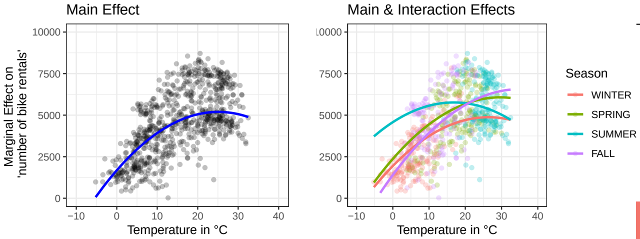
**Interpretation**: Not linear anymore!

$\rightsquigarrow$ `temp` depends on two weights:
$280.2 \cdot x_{temp} - 5.6 \cdot x_{temp}^2$



# EXAMPLE: LINEAR REGRESSION - QUADRATIC EFFECT

**Example**: Adding quadratic effect for `temp` (left) and an interaction with `season` (right)




| | Weights |
|---:|---:|
| (Intercept) | 3802.1 |
| seasonSPRING | -1345.1 |
| seasonSUMMER | -6006.3 |
| seasonFALL | -681.4 |
| hum | -38.9 |
| windspeed | -64.1 |
| days_since_2011 | 4.8 |
| temp | 39.1 |
| $\text{temp}^2$ | 8.6 |
| seasonSPRING:temp | 407.4 |
| seasonSPRING:$\text{temp}^2$ | -18.7 |
| seasonSUMMER:temp | 801.1 |
| seasonSUMMER:$\text{temp}^2$ | -27.2 |
| seasonFALL:temp | 217.4 |
| seasonFALL:$\text{temp}^2$ | -11.3 |

**Interpretation**: Not linear anymore!

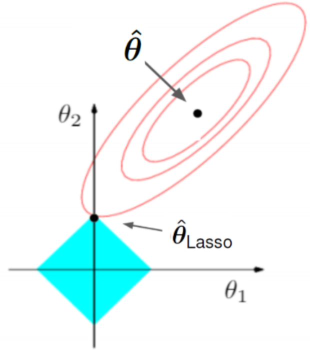
- ⇝ `temp` depends on multiple weights due to `season`:
  - ⇝ `WINTER`: $39.1 \cdot x_{temp} + 8.6 \cdot x_{temp}^2$
  - ⇝ `SPRING`: $(39.1+407.4) \cdot x_{temp} + (8.6-18.7) \cdot x_{temp}^2$
  - ⇝ `SUMMER`: $(39.1+801.1) \cdot x_{temp} + (8.6-27.2) \cdot x_{temp}^2$
  - ⇝ `FALL`: $(39.1+217.4) \cdot x_{temp} + (8.6-11.3) \cdot x_{temp}^2$

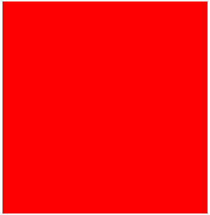
# REGULARIZATION VIA LASSO

Tibshirani (1996)

- LASSO adds an $L_1$-norm penalization term ($\lambda||\theta||_1$)
  - ⇝ Shrinks some feature weights to zero (feature selection)
  - ⇝ Sparser models (fewer features): more interpretable
- Penalization parameter $\lambda$ must be chosen (e.g., by CV)

$$min_\theta \left( \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \mathbf{x}^{(i)\top}\theta)^2}_{\text{Least square estimate for LM}} + \lambda||\theta||_1 \right)$$



# REGULARIZATION VIA LASSO ▸ Tibshirani (1996)

**Example** (interpretation of weights analogous to LM):

- LASSO with main effects and interaction `temp` with `season`
- $\lambda$ is chosen such that 6 features are selected (not zero)
- For categorical features, LASSO shrinks weights of single categories separately (due to dummy encoding)

  $\rightsquigarrow$ No feature selection of whole categorical features

  $\rightsquigarrow$ Solution: group LASSO ▸ Yuan and Lin (2006)

| | Weights |
|---|---|
| (Intercept) | 3135.2 |
| seasonSPRING | 767.4 |
| seasonSUMMER | 0.0 |
| seasonFALL | 0.0 |
| temp | 116.7 |
| hum | -28.9 |
| windspeed | -50.5 |
| days_since_2011 | 4.8 |
| seasonSPRING:temp | 0.0 |
| seasonSUMMER:temp | 0.0 |
| seasonFALL:temp | 30.2 |