# Interpretable Maching Learning

## Chapter 03 Interpretable Linear Regression Model

Hao ZHAN

haozhan1993@gmail.com

2020.5

# Table of Contents

# Table of Contents

# 1.Assumptions of linear regression model

A linear regression model predicts the target as a weighted sum of the feature inputs.
Linear models can be used to model the dependence of a regression target y on some features x. The learned relationships are linear and can be written for a single instance i as follows:

$$y = \beta_0 + \beta_0 x_1 + \beta_0 x_2 + ... + \beta_p x_p + \epsilon \tag{1}$$

The ordinary least squares method is usually used to find the weights that minimize the squared differences between the actual and the estimated outcomes (Hastie, Tibshirani, and Friedman, 2009):

$$\hat{\boldsymbol{\beta}} = \arg \min_{\beta_0,...,\beta_p} \sum_{i=1}^{n} \left( y^{(i)} - \left( \beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2 \tag{2}$$

# 1.Assumptions of linear regression model

Whether the model is the "correct" model depends on whether the relationships in the data meet certain assumptions, which are linearity, normality, homoscedasticity, independence, fixed features, and absence of multicollinearity.

## (1) Linearity

The linear regression model forces the prediction to be a linear combination of features, which is both its greatest strength and its greatest limitation.

## (2) Normality

It is assumed that the target outcome given the features follows a normal distribution.

# 1.Assumptions of linear regression model

## (3) Homoscedasticity (constant variance)

The variance of the error terms is assumed to be constant over the entire feature space..

## (4) Independence

It is assumed that each instance is independent of any other instance.

## (5) Fixed features

The input features are considered "fixed".

## (6) Absence of multicollinearity

People do not want strongly correlated features, because this messes up the estimation of the weights.

# Table of Contents

# 2.Interpretation

The interpretation of a weight in the linear regression model depends on the type of the correspond- ing feature.

## (1) Numerical feature

Increasing the numerical feature by one unit changes the estimated outcome by its weight.

## (2) Binary feature

A feature that takes one of two possible values for each instance.

## (3) Categorical feature with multiple categories

A feature with a fixed number of possible values.

## (4) Intercept $\beta_0$

The intercept is the feature weight for the "constant feature", which is always 1 for all instances.

# 2.Interpretation

The interpretation:

## (1) Numerical feature

Anincrease of feature $x_k$ by one unit increases the prediction for $y$ by $\beta_k$ units when all other feature values remain fixed.

## (2) Binary feature

Changing feature $x_k$ from the reference category to the other category increases the prediction for $y$ by $\beta_k$ when all other features remain fixed.

# 2.Interpretation

The interpretation:

## (3) Interpretation of model: R-squared

R- squared tells you how much of the total variance of your target outcome is explained by the model.

$$R^2 = 1 - SSE/SST \tag{3}$$

SSE is the squared sum of the error terms:

$$SSE = \sum_{i=1}^{n} \left( y^{(i)} - \hat{y}^{(i)} \right)^2 \tag{4}$$

SST is the squared sum of the data variance:

$$SST = \sum_{i=1}^{n} \left( y^{(i)} - \bar{y} \right)^2 \tag{5}$$

The interpretation:

## (3) Interpretation of model: R-squared

There is a catch, because R-squared increases with the number of features in the model, even if they do not contain any information about the target value at all.
Therefore, it is better to use the adjusted R-squared, which accounts for the number of features used in the model.

$$\bar{R}^2 = R^2 - \left(1 - R^2\right) \frac{p}{n - p - 1} \tag{6}$$

The interpretation:

## (4) Interpretation of features: t-statistic

The importance of a feature in a linear regression model can be measured by the absolute value of its t-statistic.

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)} \tag{7}$$

*$SE()$ means standard error of the estimate.

The importance of a feature increases with increasing weight.
The more variance the estimated weight has, the less important the feature is. This also makes sense.

# Table of Contents

# 3.Example of interpretation

We use the linear regression model to predict the number of rented bikes.

| | Weight | SE | \|t\| |
|---|---|---|---|
| (Intercept) | 2399.4 | 238.3 | 10.1 |
| seasonSUMMER | 899.3 | 122.3 | 7.4 |
| seasonFALL | 138.2 | 161.7 | 0.9 |
| seasonWINTER | 425.6 | 110.8 | 3.8 |
| holidayHOLIDAY | -686.1 | 203.3 | 3.4 |
| workingdayWORKING DAY | 124.9 | 73.3 | 1.7 |
| weathersitMISTY | -379.4 | 87.6 | 4.3 |
| weathersitRAIN/SNOW/STORM | -1901.5 | 223.6 | 8.5 |
| temp | 110.7 | 7.0 | 15.7 |
| hum | -17.4 | 3.2 | 5.5 |
| windspeed | -42.5 | 6.9 | 6.2 |
| $days_since_2011$ | 4.9 | 0.2 | 28.5 |

# 3.Example of interpretation

## Weight Plot

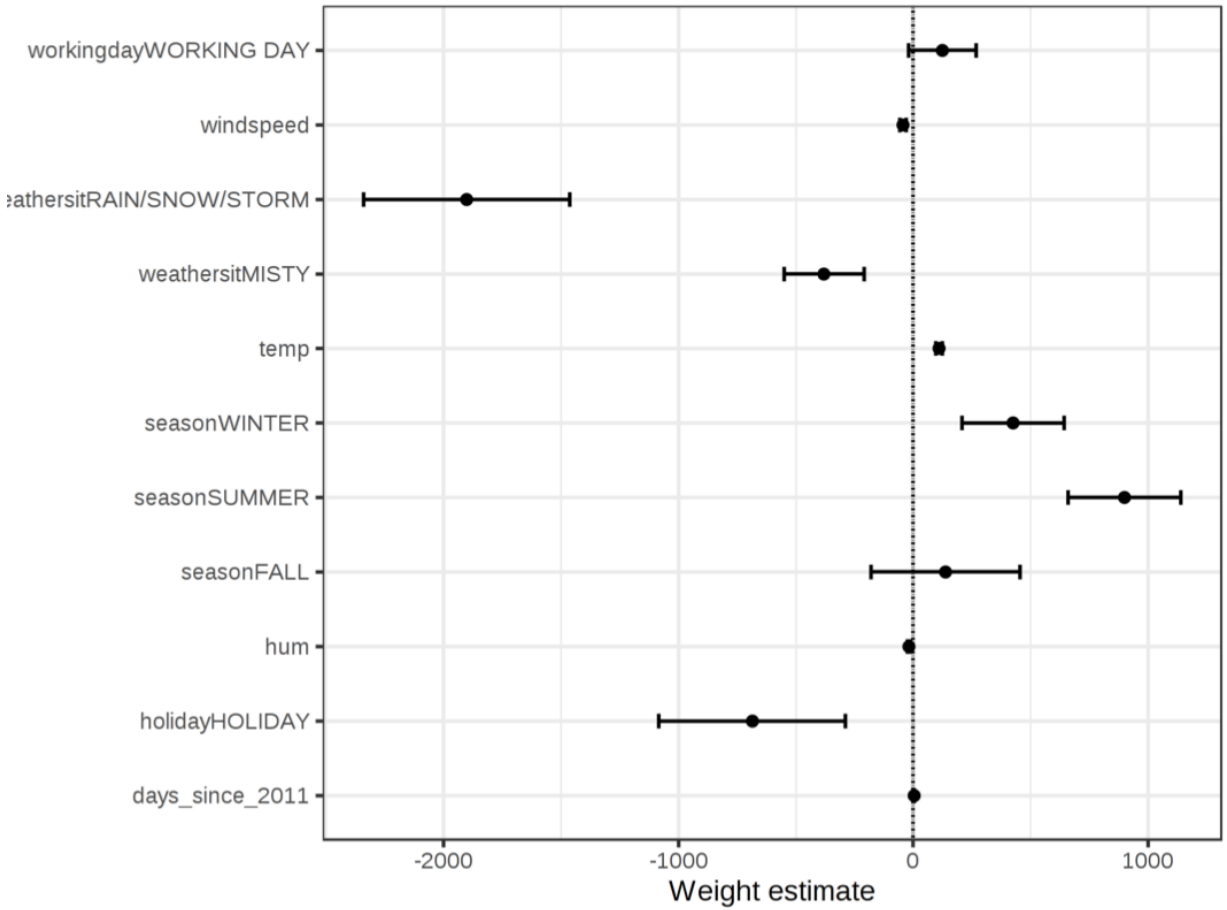
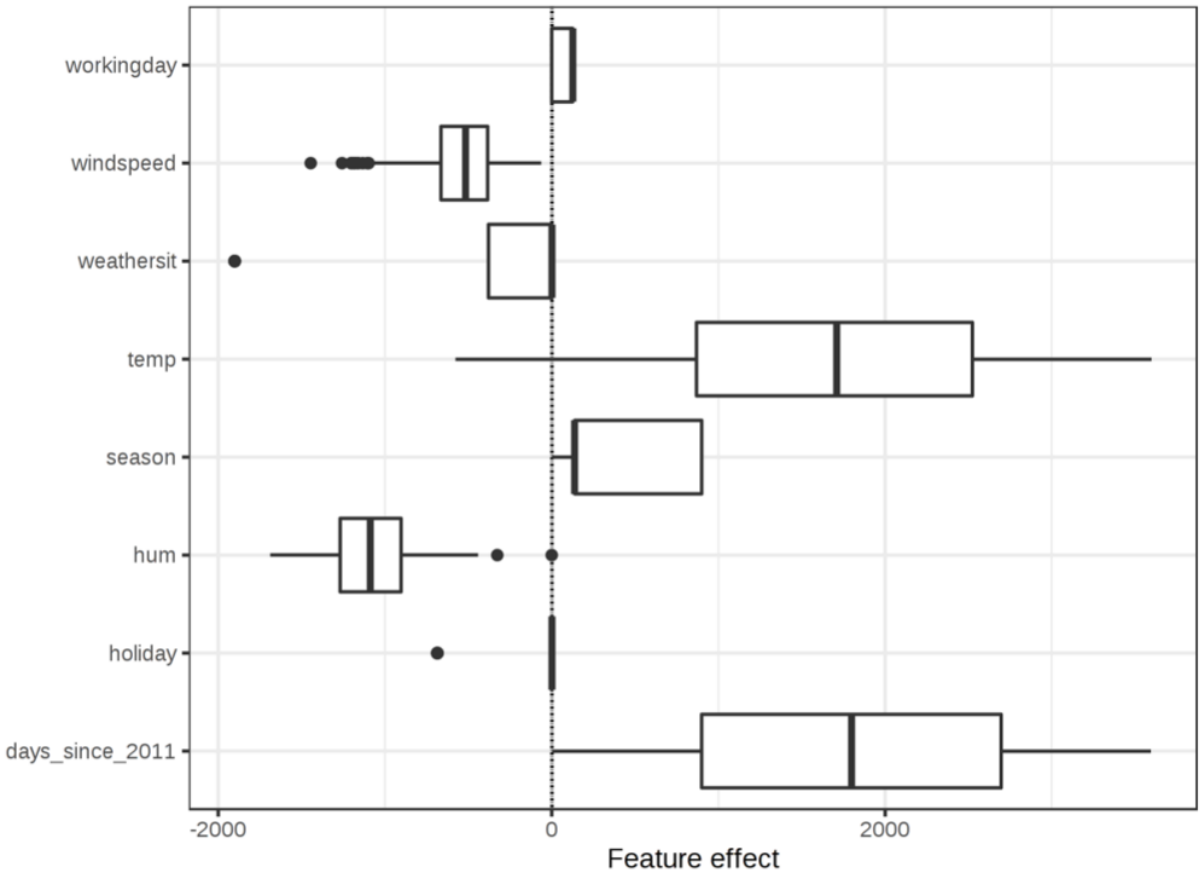



Figure: Weights are displayed as points and the .95 confidence intervals as lines.

# 3.Example of interpretation

## Effect Plot

$$\text{effect}_j^{(i)} = w_j x_j^{(i)} \tag{8}$$



# 3.Example of interpretation

## Explain Individual Predictions

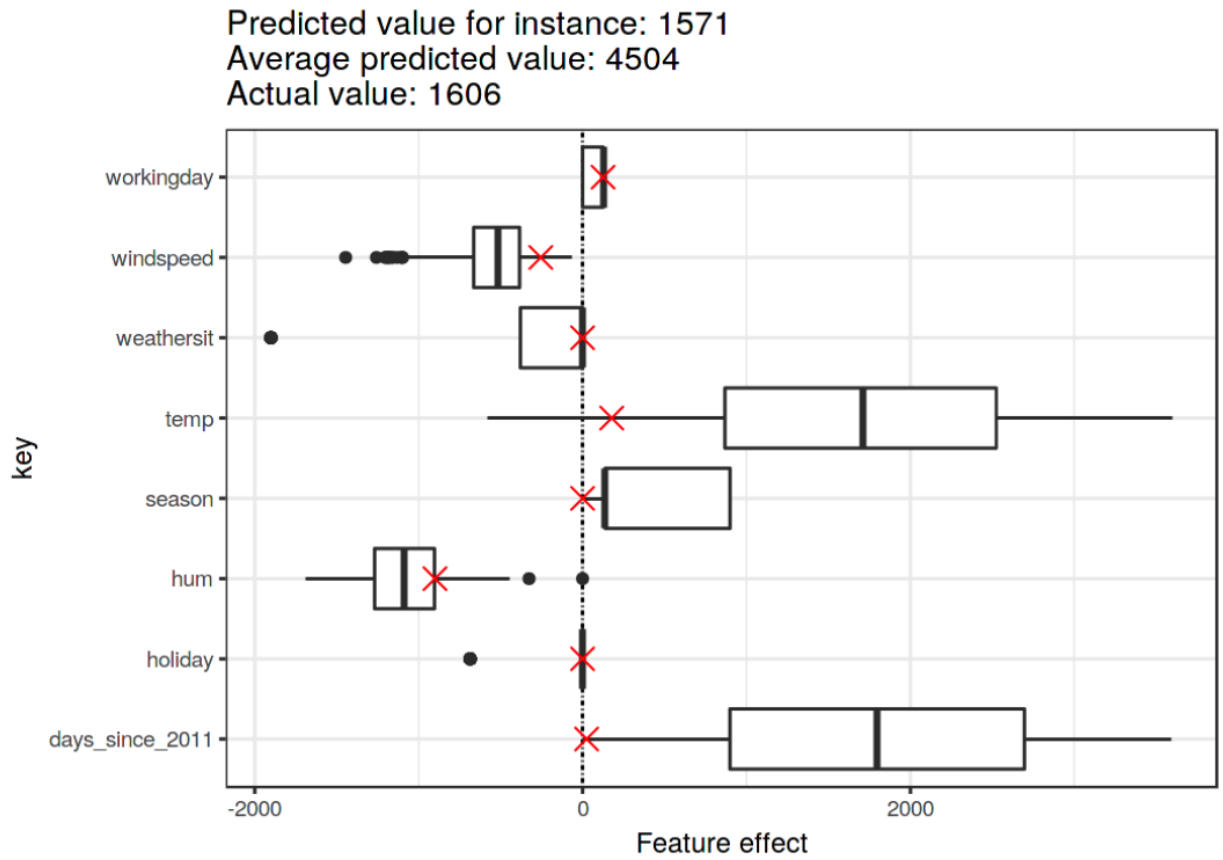



Figure: The effect plot for one instance shows the effect distribution and highlighting the effects of the instance of interest.

# Table of Contents

# 4.Sparse Linear Models

In reality we might not have just a handful of features, but hundreds or thousands.
You might even find yourself in a situation where there are more features than instances, and you cannot fit a standard linear model at all.

## (1) Lasso

Lasso is an automatic and convenient way to introduce sparsity into the linear regression model.

$$\min_{\beta} \left( \frac{1}{n} \sum_{i=1}^{n} \left( y^{(i)} - x_i^T \beta \right)^2 \right) \tag{9}$$

$$\min_{\beta} \left( \frac{1}{n} \sum_{i=1}^{n} \left( y^{(i)} - x_i^T \beta \right)^2 + \lambda \|\beta\|_1 \right) \tag{10}$$

# 4.Sparse Linear Models

Why lasso work?
I

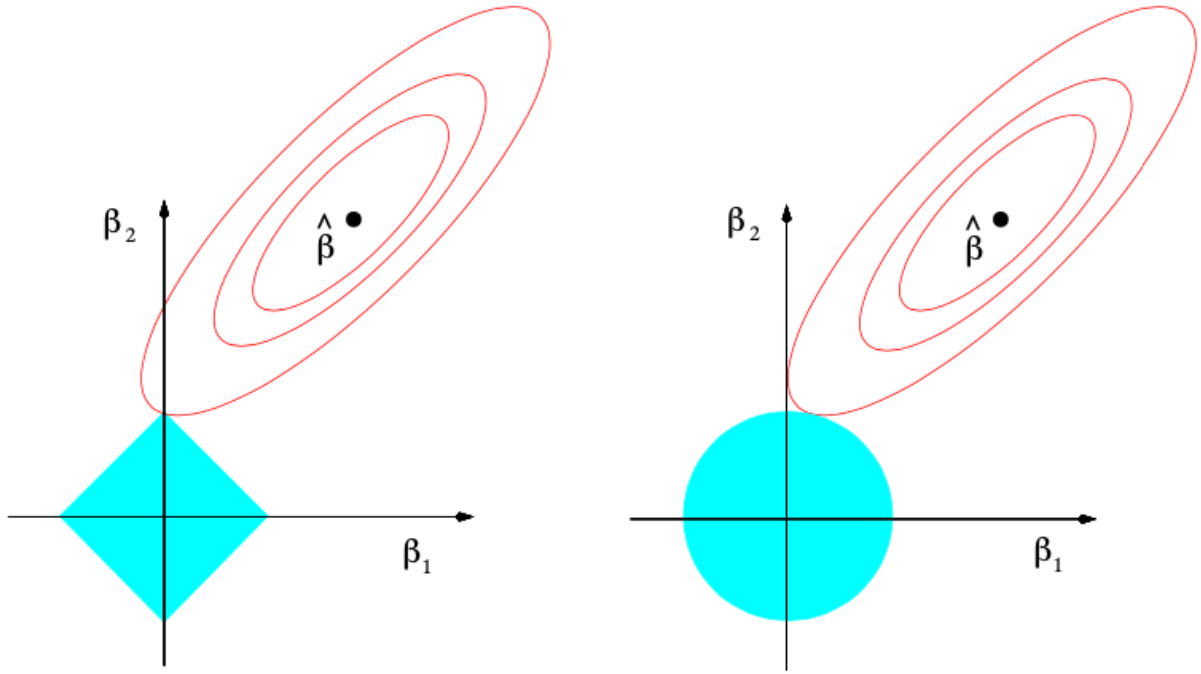



Figure: Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions.

**Example with Lasso**

2 features

| | Weight |
|---|---|
| seasonSPRING | 0.00 |
| seasonSUMMER | 0.00 |
| seasonFALL | 0.00 |
| seasonWINTER | 0.00 |
| holidayHOLIDAY | 0.00 |
| workingdayWORKING DAY | 0.00 |
| weathersitMISTY | 0.00 |
| weathersitRAIN/SNOW/STORM | 0.00 |
| temp | 52.33 |
| hum | 0.00 |
| windspeed | 0.00 |
| days_since_2011 | 2.15 |

**Example with Lasso**

5 features

| | Weight |
|---|---|
| seasonSPRING | -389.99 |
| seasonSUMMER | 0.00 |
| seasonFALL | 0.00 |
| seasonWINTER | 0.00 |
| holidayHOLIDAY | 0.00 |
| workingdayWORKING DAY | 0.00 |
| weathersitMISTY | 0.00 |
| weathersitRAIN/SNOW/STORM | -862.27 |
| temp | 85.58 |
| hum | -3.04 |
| windspeed | 0.00 |
| $days_since2011$ | 3.82 |

# 4.Sparse Linear Models

## Pre-processing methods

**Manually selected features:** You can always use expert knowledge to select or discard some features. The big drawback is that it cannot be automated and need be an expert.

**Univariate selection:** An example is the correlation coefficient. You only consider features that exceed a certain threshold of correlation between the feature and the target.

## Step-wise methods

**Forward selection:** Fit the linear model with one feature.

**Backward selection:** Similar to forward selection.

# Table of Contents

# 5.Advantages and disadvantages

## Advantages

- weighted sum makes model transparent
- high level of collective experience and expertise
- guarantee to find optimal weights

## Disadvantages

- nonlinearity or interaction has to be hand-crafted
- not that good regarding predictive performance
- can be unintuitive

# Refrence

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001.

Thank you for your time!