Linear Regression

Elif Yılmaz

11.04.2021

SIMPLE LINEAR REGRESSION

It predicts a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X$$

SIMPLE LINEAR REGRESSION

It predicts a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X$$

In the equation, β_0 and β_1 are two unknown constants representing the intercept and slope terms in the linear model. Together, β_0 and β_1 are known as the model coefficients or parameters.

Example

Let's try to predict future sales using TV advertising by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ indicates a prediction of Y on the basis of X = x. Here, $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{y}$ denote the estimated value for an unknown parameter or coefficient, or the predicted value of the response, respectively.

Estimating the Coefficients

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ represent n observation pairs. Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well.

Estimating the Coefficients

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ represent n observation pairs. Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well. That is, so that

$$\hat{y}_i \approx \hat{\beta}_1 + \hat{\beta}_1 x_i \quad \textit{for } i = 1, ..., n.$$

In other words, we want to find an intercept and a slope such that the resulting line is as close as possible to our data points. There are a number of ways of measuring closeness.

Estimating the Coefficients

Here, we will use the most common approach involving minimizing the least squares criterion. Let $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith residual which is the difference between the ith observed response value and the ith response value that is predicted by our linear model.

Estimating the Coefficients

Here, we will use the most common approach involving minimizing the least squares criterion. Let $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith residual which is the difference between the ith observed response value and the ith response value that is predicted by our linear model. We define the residual sum of squares as

$$RSS = {e_1}^2 + {e_2}^2 + ... + {e_n}^2$$

or equivalently

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

Estimating the Coefficients

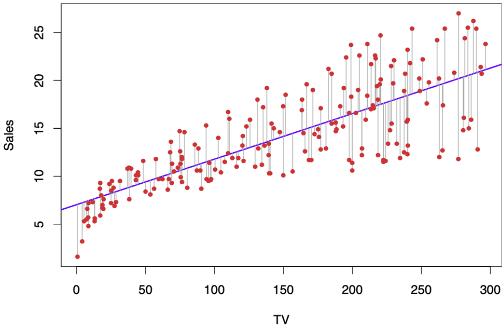
The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Therefore, we can show that the minimizers are

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

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means. That is, these minimizers defines *the least squares coefficient estimates* for simple linear regression.

Example



In the figure, for the advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the sum of squared errors.

Assessing the Accuracy of the Coefficient Estimates

We assume that the true relationship between X and Y takes the form $Y = f(X) + \varepsilon$ for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Assessing the Accuracy of the Coefficient Estimates

We assume that the true relationship between X and Y takes the form $Y = f(X) + \varepsilon$ for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The error term is a catch-all for what we miss with this simple model:

- The true relationship is probably not linear
- There may be other variables that cause variation in Y
- There may be measurement error.

Assessing the Accuracy of the Coefficient Estimates

We assume that the true relationship between X and Y takes the form $Y = f(X) + \varepsilon$ for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function, then we can write this relationship as

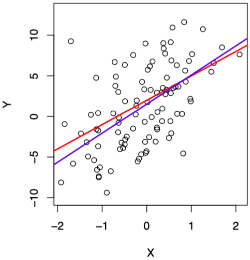
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The error term is a catch-all for what we miss with this simple model:

- The true relationship is probably not linear
- There may be other variables that cause variation in Y
- There may be measurement error.

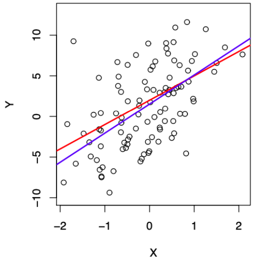
The model given by this equation defines the population regression line, which is the best linear approximation to the true relationship between X and Y.

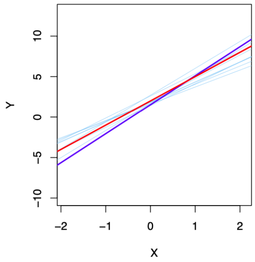
Example



The red line represents the true relationship which is known as the population regression line. The blue line is the least squares line.

Example





The red line represents the true relationship which is known as the population regression line. The blue line is the least squares line.

In light blue, 10 least squares lines are shown, each computed on the basis of a separate random set of observations. The least squares lines are quite close to the population regression line.

Assessing the Accuracy of the Coefficient Estimates

We should be careful about these points:

- We interested in population mean, let's say it μ. We do not know μ. However, we have some data points; that is, sample which has n observations. So, we can estimate μ by using these observations.

Assessing the Accuracy of the Coefficient Estimates

We should be careful about these points:

- We interested in population mean, let's say it μ. We do not know μ. However, we have some data points; that is, sample which has n observations. So, we can estimate μ by using these observations.
- We do not know true regression line and again we do not have β_0 and β_1. Similarly, we calculate these parameters by using $\hat{\beta}_0$ and $\hat{\beta}_1$ because we have sample and we can calculate these parameters. Therefore we try to estimate β_0 and β_1 in population regression line.

Assessing the Accuracy of the Coefficient Estimates

We should be careful about these points:

- We interested in population mean, let's say it μ. We do not know μ. However, we have some data points; that is, sample which has n observations. So, we can estimate μ by using these observations.
- We do not know true regression line and again we do not have β_0 and β_1. Similarly, we calculate these parameters by using $\hat{\beta}_0$ and $\hat{\beta}_1$ because we have sample and we can calculate these parameters. Therefore we try to estimate β_0 and β_1 in population regression line.

The analogy between linear regression and estimation of the mean of a random variable is based on the concept of bias. If we use the sample mean $\hat{\mu}$ to estimate μ, this estimate is unbiased.

Assessing the Accuracy of the Coefficient Estimates

How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
We can calculate standard error of $\hat{\mu}$, $SE(\hat{\mu})$.

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

where σ is the standard deviation of each of the realizations y_i of Y.

Assessing the Accuracy of the Coefficient Estimates

How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
We can calculate standard error of $\hat{\mu}$, $SE(\hat{\mu})$.

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

where σ is the standard deviation of each of the realizations y_i of Y.
Similarly, we can apply this to $\hat{\beta}_0$ and $\hat{\beta}_1$ to find how close them are to true values β_0 and β_1.

$$(SE(\hat{\beta}_0))^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad (SE(\hat{\beta}_1))^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\sigma^2 = Var(\varepsilon)$.

Assessing the Accuracy of the Coefficient Estimates

How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
We can calculate standard error of $\hat{\mu}$, $SE(\hat{\mu})$.

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

where σ is the standard deviation of each of the realizations y_i of Y.
Similarly, we can apply this to $\hat{\beta}_0$ and $\hat{\beta}_1$ to find how close them are to true values β_0 and β_1.

$$(SE(\hat{\beta}_0))^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad (SE(\hat{\beta}_1))^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\sigma^2 = Var(\varepsilon)$.
Note: In general, σ^2 is not known and we can estimate it from data by using residual standard error

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

Assessing the Accuracy of the Coefficient Estimates

Standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Assessing the Accuracy of the Coefficient Estimates

Standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

For linear regression, the 95% confidence interval for β_1 approximately takes the form $\bar{\beta}_1 \pm 2SE(\bar{\beta}_1)$. That is, there is approximately a 95% chance that the interval

$$[\bar{\beta}_1 - 2SE(\bar{\beta}_1), \bar{\beta}_1 + 2SE(\bar{\beta}_1)]$$

Similarly, for $\bar{\beta}_0$,

$$[\bar{\beta}_0 - 2SE(\bar{\beta}_0), \bar{\beta}_0 + 2SE(\bar{\beta}_0)]$$

Assessing the Accuracy of the Coefficient Estimates

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

$$H_0: \text{ There is no relationship between X and Y} \rightarrow H_0 : \beta_1 = 0$$

and the alternative hypothesis

$$H_A: \text{ There is some relationship between X and Y.} \rightarrow H_1 : \beta_1 \neq 0$$

Assessing the Accuracy of the Coefficient Estimates

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

$$H_0: \text{ There is no relationship between X and Y} \rightarrow H_0 : \beta_1 = 0$$

and the alternative hypothesis

$$H_A: \text{ There is some relationship between X and Y.} \rightarrow H_1 : \beta_1 \neq 0$$

Here, we can use t-statistic to measure the number of standard deviations that β_1 is away from 0.

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

Assessing the Accuracy of the Coefficient Estimates

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

$$H_0: \text{ There is no relationship between X and Y} \rightarrow H_0 : \beta_1 = 0$$

and the alternative hypothesis

$$H_A: \text{ There is some relationship between X and Y.} \rightarrow H_1 : \beta_1 \neq 0$$

Here, we can use t-statistic to measure the number of standard deviations that β_1 is away from 0.

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

We can also calculate p-value to decide the probability of observing any value equal to $|t|$ or larger. It is just one way to decide whether our null hypothesis is true or not.

Assessing the Accuracy of the Model

The quality of a linear regression fit is typically assessed using two related quantities:

- Residual standard error (RSE)

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

  It is the average amount that the response will deviate from the true regression line.

Assessing the Accuracy of the Model

The quality of a linear regression fit is typically assessed using two related quantities:

- Residual standard error (RSE)

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

  It is the average amount that the response will deviate from the true regression line.

- R^2 statistic

$$R^2 = \frac{TSS - RSS}{TSS}$$

  where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares. R^2 always takes on a value between 0 and 1, and is independent of the scale of Y. It is a measure of the linear relationship between X and Y.

Assessing the Accuracy of the Model

Correlation is also a measure of the linear relationship between X and Y.

$$Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Then, if we say $r = Cor(X, Y)$, we can show that $R^2 = r^2$ in the simple linear regression setting.

MULTIPLE LINEAR REGRESSION

The multiple linear regression model is in the form

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

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the average effect on Y of a one unit increase in X_j when we hold all other predictors fixed.

Estimating the Regression Coefficients

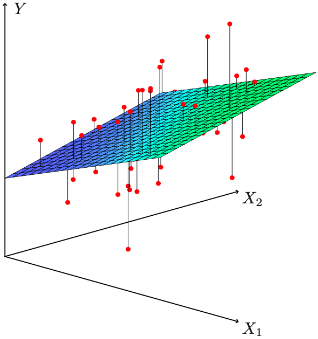
By using given estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$, we can make predictions as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i + ... + \hat{\beta}_p x_p$$

We can estimate the parameters by using the same least squares approach in simple linear regression. We choose $\beta_0, \beta_1, ..., \beta_p$ to minimize the sum of squared residuals

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y})^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - .... - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Example



The figure illustrates an example of the least squares fit with p = 2 predictors.

Is there a relationship between the response and predictors?

In the simple linear regression setting, to determine whether there is a relationship between the response and the predictor, we can simply check whether $\beta_1 = 0$ or not.

Is there a relationship between the response and predictors?

In the simple linear regression setting, to determine whether there is a relationship between the response and the predictor, we can simply check whether $\beta_1 = 0$ or not. Here, we can apply similar way:

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

$$H_a : \text{at least one } \beta_j \text{ is non-zero.}$$

Is there a relationship between the response and predictors?

In the simple linear regression setting, to determine whether there is a relationship between the response and the predictor, we can simply check whether $\beta_1 = 0$ or not. Here, we can apply similar way:

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

$$H_a : \text{at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Is there a relationship between the response and predictors?

If the linear model assumptions are correct; H_0 is true, we can show that

$$E\{RSS/(n-p-1)\} = \sigma^2$$

and

$$E\{(TSS-RSS)/p\} = \sigma^2$$

Is there a relationship between the response and predictors?

If the linear model assumptions are correct; H_0 is true, we can show that

$$E\{RSS/(n-p-1)\} = \sigma^2$$

and

$$E\{(TSS-RSS)/p\} = \sigma^2$$

Note: When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1.

Is there a relationship between the response and predictors?

If the linear model assumptions are correct; H_0 is true, we can show that

$$E\{RSS/(n-p-1)\} = \sigma^2$$

and

$$E\{(TSS-RSS)/p\} = \sigma^2$$

Note: When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1.
Note: If H_a is true, then $E\{(TSS-RSS)/p\} > \sigma^2$, so F should be greater than 1.

Is there a relationship between the response and predictors?

Sometimes we want to test that a particular subset of q of the coefficients are zero. This corresponds to a null hypothesis

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0$$

where for convenience we have put the variables chosen for omission at the end of the list. In this case we fit a second model that uses all the variables except those last q. Suppose that the residual sum of squares for that model is RSS_0. Then the appropriate F-statistic is

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

Deciding on Important Variables

If $p = 2$, then we can consider four models:

1. a model containing no variables
2. a model containing X_1 only
3. a model containing X_2 only
4. a model containing X_1 and X_2

Deciding on Important Variables

If p = 2, then we can consider four models:

1. a model containing no variables
2. a model containing X_1 only
3. a model containing X_2 only
4. a model containing X_1 and X_2

Therefore, we can select the best model out of all of the models. Various statistics can be used to judge the quality of a model. These include

- Mallow's C_p
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Adjusted R^2
- Plotting various model outputs, such as the residuals, in order to search for patterns.

Deciding on Important Variables

There are a total of 2^p models that contain subsets of p variables. When p is large, there are three classical approaches to choose a smaller set of models to consider:

- Forward selection

Deciding on Important Variables

There are a total of 2^p models that contain subsets of p variables. When p is large, there are three classical approaches to choose a smaller set of models to consider:

- Forward selection
- Backward selection

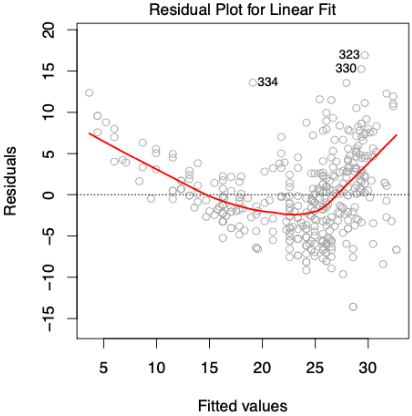
Deciding on Important Variables

There are a total of 2^p models that contain subsets of p variables. When p is large, there are three classical approaches to choose a smaller set of models to consider:

- Forward selection
- Backward selection
- Mixed selection

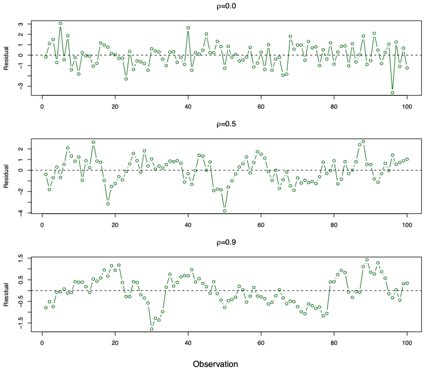
Potential Problems in a Linear Model

- Non-linearity of the response-predictor relationships



Potential Problems in a Linear Model

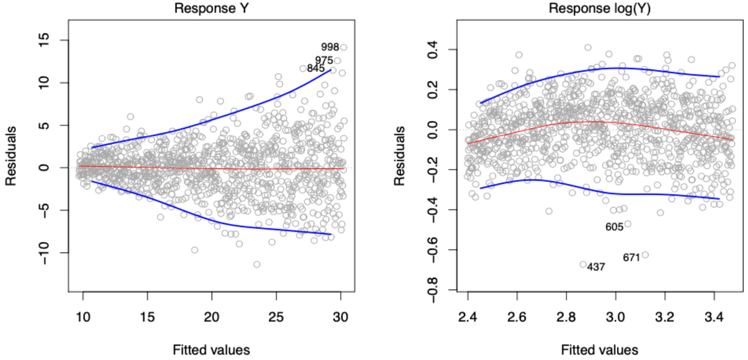
- Correlation of error terms



In the figure, we see plots of residuals from simulated time series data sets generated with differing levels of correlation p between error terms for adjacent time points.

Potential Problems in a Linear Model

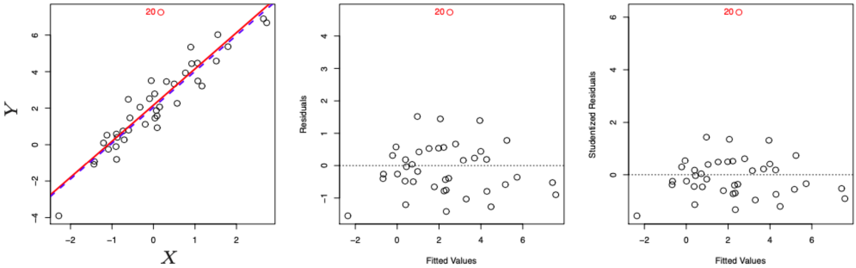
- Non-constant variance of error terms



In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals.

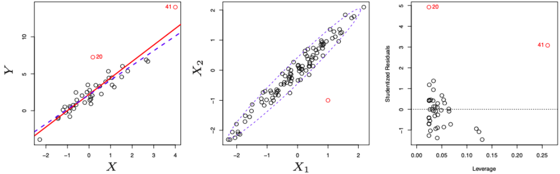
Potential Problems in a Linear Model

- Outliers



Potential Problems in a Linear Model

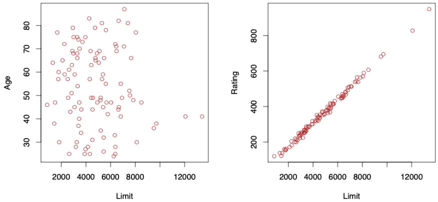
- High-leverage points



In order to quantify an observation's leverage, we compute the leverage statistic. A large value of this statistic indicates an observation with high leverage. For a simple linear regression,

$$h_i = \frac{1}{n} + \frac{(x_i \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} \bar{x})^2}$$

- Collinearity



A better way to assess multi-collinearity is to compute the variance inflation factor (VIF). The VIF for each variable can be computed using the formula

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is a parametric approach. Parametric methods have several advantages such that they need estimate small number of coefficients (or parameters). They have also some disadvantages; for example, if we assume a linear relationship between X and Y but the true relationship is far from linear, then the resulting model will provide a poor fit to the data.

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is a parametric approach. Parametric methods have several advantages such that they need estimate small number of coefficients (or parameters). They have also some disadvantages; for example, if we assume a linear relationship between X and Y but the true relationship is far from linear, then the resulting model will provide a poor fit to the data.

Non-parametric methods provide an alternative and more flexible approach to perform regression. Therefore, we can talk about K-nearest neighbors regression (KNN regression).

Comparison of Linear Regression with K-Nearest Neighbors

The KNN regression method is closely related to the KNN classifier.

- We have a value for K and a prediction point x_0.

Comparison of Linear Regression with K-Nearest Neighbors

The KNN regression method is closely related to the KNN classifier.

- We have a value for K and a prediction point x_o.
- KNN regression identifies K observations which are the closest points to x_o. Then, let represent these points by $\mathcal{N}_0$.

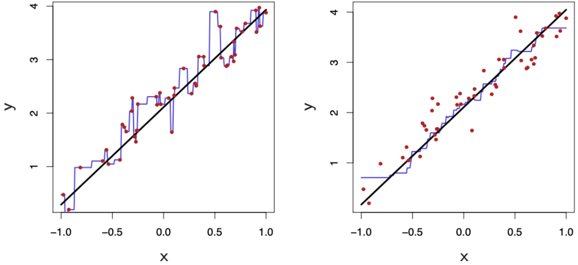
Comparison of Linear Regression with K-Nearest Neighbors

The KNN regression method is closely related to the KNN classifier.

- We have a value for K and a prediction point x_o.
- KNN regression identifies K observations which are the closest points to x_o. Then, let represent these points by $\mathcal{N}_0$.
- Therefore, we can estimate $f(x_o)$ using the average of all the training responses in $\mathcal{N}_0$. That is,

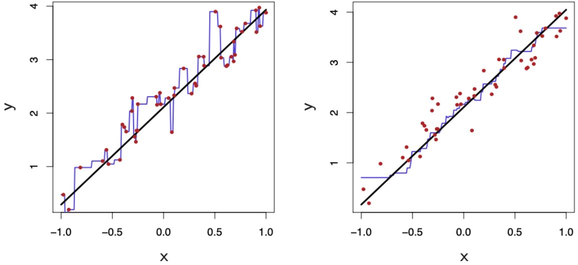
$$\hat{f}(x_o) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} x_i$$

Example



In the figure, we see plots of f(X) using KNN regression on a one-dimensional data set. Left: K = 1, Right: K = 9.

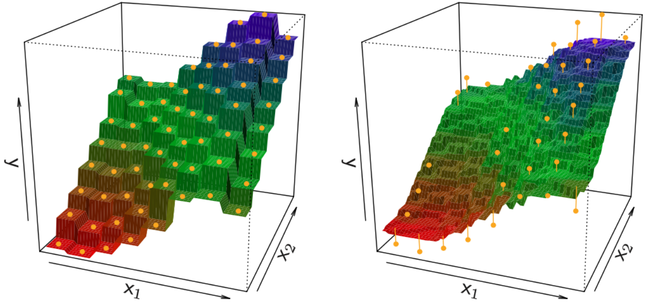
Example



In the figure, we see plots of f(X) using KNN regression on a one-dimensional data set. Left: K = 1, Right: K = 9.

Note: In general, the optimal value for K will depend on the bias-variance tradeoff.

Example



As an another example, we can observe plots of f(X) using KNN regression on a two-dimensional data set.

REFERENCES

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*, 112(18). New York: springer.
- https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/degrees-of-freedom/
- https://stats.stackexchange.com/questions/204238/why-divide-rss-by-n-2-to-get-rse