Linear Regression
Ch3-ISLP; Ch2-IMLP

Zhongjian Lin
University of Georgia

October 12, 2024

Linear Regression

- Linear regression is a useful tool for predicting a quantitative response.
- It has been around for a long time and is the topic of innumerable textbooks. Though it may seem somewhat dull compared to some of the more modern statistical approaches is still a useful and widely used statistical learning method.
- It serves as a good jumping-off point for newer approaches: as we will see in later chapters, many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression.
- The importance of having a good understanding of linear regression before studying more complex learning methods cannot be overstated.

Simple Linear Regression

Simple linear regression lives up to its name: it is a very straightforward simple linear approach for predicting 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

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

or

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{1}$$

We usually assume ϵ is independent of X and with mean 0. We refer β_0 and β_1 as two unknown constants that represent the *intercept* and *slope* terms. They are known as the model *coefficients* or *parameters*
For example, in the Advertising case,

$$sales \approx \beta_0 + \beta_1 TV.$$

Estimating the Coefficients

Let $(x_i, y_i)_{i=1}^n$ represent n observation pairs, each of which consists of a measurement of X and a measurement of Y. 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 $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \cdots, n$.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith residual—this is the difference between the ith observed response value and its predicted value by the linear model.

We define the residual sum of squares (RSS) as

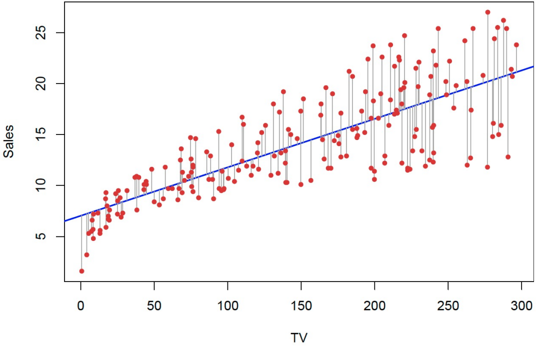
$$RSS = e_1^2 + \cdots + e_n^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{2}$$

The (ordinary) least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.

Estimation

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

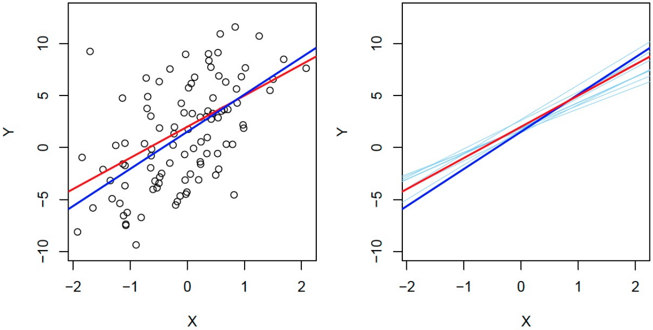


In the example of *Advertising*, $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$.

Accuracy of the Coefficient Estimates

Consider a simulation study with true population line

$$Y = 2 + 3X + \epsilon, \epsilon \sim N(0, \sigma^2).$$

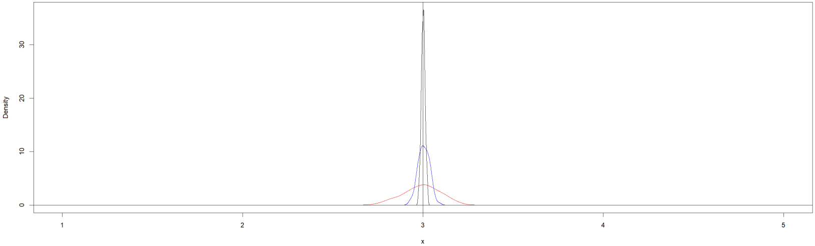


Accuracy of LS Estimation

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \xrightarrow{p} \beta_1 + 0 \tag{3}$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1}{n}\frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx O\left(\frac{1}{n}\right)$$

Notice in the formula that $SE(\hat{\beta}_1)$ is smaller when the x_i are more spread out; intuitively we have more leverage to estimate a slope when this is the case. Furthermore, $SE(\hat{\beta}_1)^2$ decreases proportionally as the sample size n increases.



Confidence Intervals

Standard errors can be used to compute confidence intervals. A 95 % confidence confidence interval is defined as a range of values such that with 95% interval 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

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1).$$

That is, there is approximately a 95 % chance that the interval

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

will contain the true value of β_1. We have similar result for β_0. In the case of the advertising data, the 95 % confidence interval for β_0 is $[6.130, 7.935]$ and the 95 % confidence interval for β_1 is $[0.042, 0.053]$.

Hypothesis Testing

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

$$H_0\text{: There is no relationship between } X \text{ and } Y\text{, i.e., } \beta_1 = 0,$$

versus the alternative hypothesis

$$H_a\text{: There is some relationship between } X \text{ and } Y\text{, i.e., } \beta_1 \neq 0,$$

In practice, we compute a t-statistic, *t-statistic* given by

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

which measures the number of standard deviations that $\hat{\beta}_1$ is away from 0. We *reject the null hypothesis* if $|t|$ is large (compared to some thresholds).

Advertising vs Sales

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

Accuracy of the Model

Once we have rejected the null hypothesis in favor of the alternative hypothesis, it is natural to want to quantify the extent to which the model fits the data. The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic.

The residual standard error (RSE) is an estimate of the standard deviation of ϵ. Roughly speaking, it is the average amount that the response will deviate from the true regression line. It is computed using the formula

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

The RSE is considered a measure of the lack of fit of the model to the data.

Accuracy of the Model

The RSE provides an absolute measure of lack of fit of the model to the data. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE. The R^2 statistic provides an alternative measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1, and is independent of the scale of Y.

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

TSS measures the total variance in the response Y, and can be squares thought of as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression. Hence, $TSS - RSS$ measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using X.

Example: Marketing Strategy

Recall the Advertising case with sales as a function of advertising budgets for TV, radio, and newspaper media. Suppose that in our role as statistical consultants we are asked to suggest, a marketing plan for next year that will result in high product sales. What information would be useful in order to provide such a recommendation? Here are a few important questions that we might seek to address:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media are associated with sales?
- How large is the association between each medium and sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Simple Linear Regression vs Multiple Linear Regression

Drawbacks of SLR and motivation of MLR

- Only one input included? Correlation between the input and the error, i.e., $Cov(X, \epsilon) \neq 0$.
$$\hat{\beta} \nrightarrow \beta$$
- Cannot conduct ceteris paribus analysis, i.e., explicitly hold fixed other factors that otherwise would be in ϵ.
- Only linear relationship between Y and X.

In the advertising example, we have

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon$$

In general, multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon. \tag{4}$$

Ceteris Paribus Analysis

The notion of ceteris paribus—that is, holding all other (relevant) factors fixed—is at the crux of establishing a causal relationship. Simply finding that two variables are correlated is rarely enough to conclude that a change in one variable causes a change in another. After all, rarely can we run a controlled experiment that allows a simple correlation analysis to uncover causality. Instead, we can use linear regression to effectively hold other factors fixed.

$$\triangle Y = \beta_1 \triangle X_1 + \cdots + \triangle X_p + \triangle \epsilon$$

The ceteris paribus effects of X_1, holding $X_2, \cdots, X_p$ constant, i.e., $\triangle X_2 = \cdots = \triangle X_p = 0$, is then summarized by β_1 in

$$\triangle Y = \beta_1 \triangle X_1 + 0 + 0 \Rightarrow \beta_1 = \frac{\triangle Y}{\triangle X_1}$$

Estimation

We follow similar least square strategy as that in SLR to find estimation of MLR.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

We define the RSS as

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p)^2$$

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x'\beta)^2 \tag{5}$$

Advertising

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1

Does it make sense for the multiple regression to suggest no relationship between sales and newspaper while the simple linear regression implies the opposite? In fact it does. This slightly counterintuitive result is very common in many real life situations. Consider an absurd example to illustrate the point. Running a regression of shark attacks versus ice cream sales for data collected at a given beach community over a period of time would show a positive relationship, similar to that seen between sales and newspaper.

Multiple Linear Regression

We usually are interested in answering a few important questions.

1. Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting the response?
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful? *Forward selection, Backward selection, Mixed selection*
3. How well does the model fit the data? RSE and R^2.
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p.$$
We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.

Qualitative Predictor

In practice, predictor is not necessarily quantitative; often some predictors are qualitative.

Example: Card Balance

Consider an example of the analysis of individual *balance*, with quantitative predictors: *age*, *cards*, *education*, *income*, *limit*, *rating*; and qualitative predictors: *own* (house ownership), *student* (student status), *status* (marital status), and *region* (East, West or South).

Example: Wage Equation

Consider the example of *wage*, with quantitative predictors: *education*, *experience*; and qualitative predictors: *female* (gender), *status* (marital status).

Dummy Variable

We use *Dummy Variable* to incorporate qualitative predictors.

$$x_i(own_i) = \begin{cases} 1, & \text{if ith person owns a house} \\ 0, & \text{if ith person does not own a house} \end{cases}$$

For a predictor with more than two categories, e.g., *region*, we can use dummy variables as well.

$$east_i = \begin{cases} 1, & \text{if ith person is from the East} \\ 0, & \text{if ith person does not live in the east} \end{cases}$$

$$west_i = \begin{cases} 1, & \text{if ith person si from the West} \\ 0, & \text{if ith person does not live in the west} \end{cases}$$

$$south_i = \begin{cases} 1, & \text{if ith person is from the South} \\ 0, & \text{if ith person does not live in the south} \end{cases}$$

Interpretation

Consider the simple linear regression of balance on own.

$$balance_i = \beta_0 + \beta_1 own_i + \epsilon = \begin{cases} \beta_0 + \beta_1 + \epsilon, & \text{if ith person owns a house} \\ \beta_0 + \epsilon, & \text{if ith person does not own a house} \end{cases}$$

Now β_0 can be interpreted as the average credit card balance among those who do not own, $\beta_0 + \beta_1$ as the average credit card balance among those who do own their house, and β_1 as the average difference in credit card balance between owners and non-owners.

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
own	19.73	46.05	0.429	0.6690

$$\begin{aligned} balance_i &= \beta_0 + \beta_1 south_i + \beta_2 west_i + \epsilon \\ &= \begin{cases} \beta_0 + \beta_1 + \epsilon, & \text{if ith person is from the South} \\ \beta_0 + \beta_2 + \epsilon, & \text{if ith person is from the West} \\ \beta_0 + \epsilon, & \text{if ith person is from the East} \end{cases} \end{aligned}$$

The level with no dummy variable-East in this example-is known as the *baseline* or *benchmark*.

Extension: Interaction Effects

Example: Advertising

We concluded that both *TV* and *radio* seem to be associated with sales.

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \epsilon. \tag{6}$$

The conclusion is based on the assumption that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media. The average increase in sales associated with a one-unit increase in *TV* is always β_1, regardless of the amount spent on *radio*. However, this simple model may be incorrect. Suppose that spending money on *radio* advertising actually increases the effectiveness of *TV* advertising, so that the slope term for TV should increase as radio increases. Spending half on *radio* and half on *TV* may increase sales more than allocating the entire amount to either TV or to radio. In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Extension: Interaction Effects

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$

Define $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$ and it captures the partial effect of X_1 on Y. Since $\tilde{\beta}_1$ is now a function of X_2, the association between X_1 and Y is no longer constant: a change in the value of X_2 will change the association between X_1 and Y. Similarly, we can define $\tilde{\beta}_2 = \beta_2 + \beta_3 X_1$ and the association between X_2 and Y depends on the value of X_1.

Example: Production

For example, suppose that we are interested in studying the productivity of a factory. We wish to predict the number of *units* produced on the number of production *lines* and the number of *workers*. It seems likely that the effect of increasing the number of production lines will depend on the number of workers. This suggests that it would be appropriate to include an interaction term to predict units.

$$units = \beta_0 + \beta_1 lines + \beta_2 workers + \beta_3 lines \cdot workers + \epsilon.$$

$$units \approx 1.2 + 3.4 lines + 0.22 workers + 1.4 lines \cdot workers$$

Testing the Interaction Effects

We could simply test the existence of interaction effects by hypothesis testing.

$$H_0 : \beta_3 = 0 \text{ v.s. } H_a : \beta_3 \neq 0.$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	< 0.0001
TV×radio	0.0011	0.000	20.73	< 0.0001

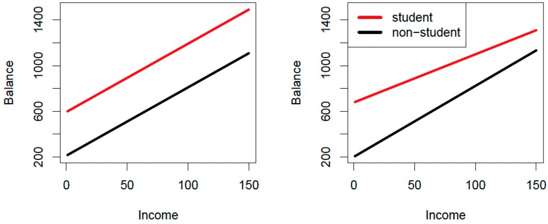
An increase in TV advertising of \$1,000 is associated with increased sales of $\hat{\hat{\beta}}_1 = 0.0191 + 0.0011 radio = 19.1 + 1.1 radio$ units. An increase in radio advertising of \$1,000 will be associated with an increase in sales of $\hat{\hat{\beta}}_2 = 0.0289 + 0.0011 TV = 28.9 + 1.1 TV$ units.

Interaction Effects with Qualitative Predictor

Example: Balance

The combination of interaction effects and qualitative predictor is very attractive for group comparison

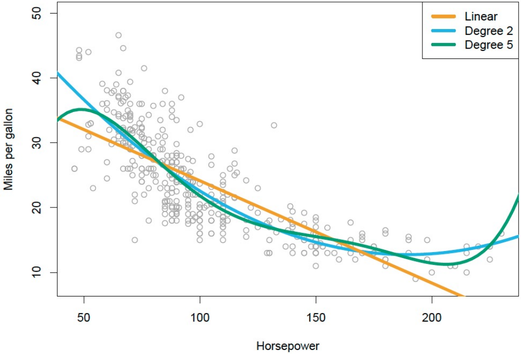
$$
\begin{aligned}
balance &\approx \beta_0 + \beta_1 income + \gamma_0 student + \gamma_1 income \cdot student, \\
&\approx \begin{cases} \beta_0 + \beta_1 income, & \text{if not student} \\ (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1) income, & \text{if student} \end{cases}
\end{aligned}
$$



Non-Linear Relationship

In some cases, the true relationship between the response and the predictors may be nonlinear. Here we present a very simple way to directly extend the linear model to accommodate non-linear relationships, using *polynomial regression*.

$$mpg \approx g(horsepower).$$



Comparison of Linear Regression with K-Nearest Neighbors

Parametric methods have several advantages.

- They are often easy to fit.
- The coefficients have simple interpretations
- Tests of statistical significance can be easily performed.

But parametric methods do have a disadvantage: by construction, they make strong assumptions about the form of $f(X)$. If the specified functional form is far from the truth, and prediction accuracy is our goal, then the parametric method will perform poorly. Non-parametric methods do not explicitly assume a parametric form for $f(X)$, and thereby provide an alternative and more flexible approach for performing regression. Here we consider K-nearest neighbors regression (KNN regression).

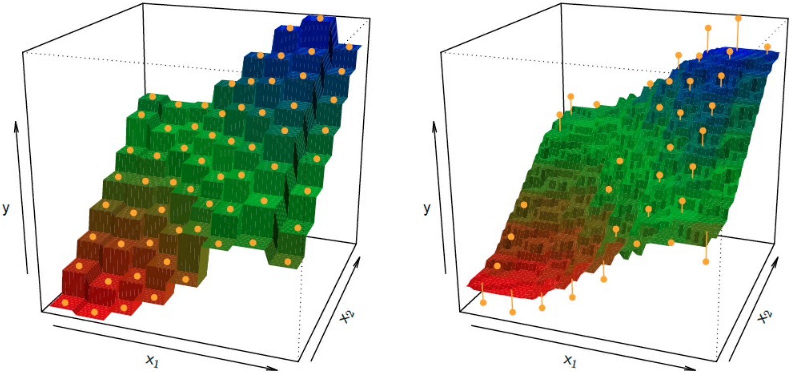
KNN regression

Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$.

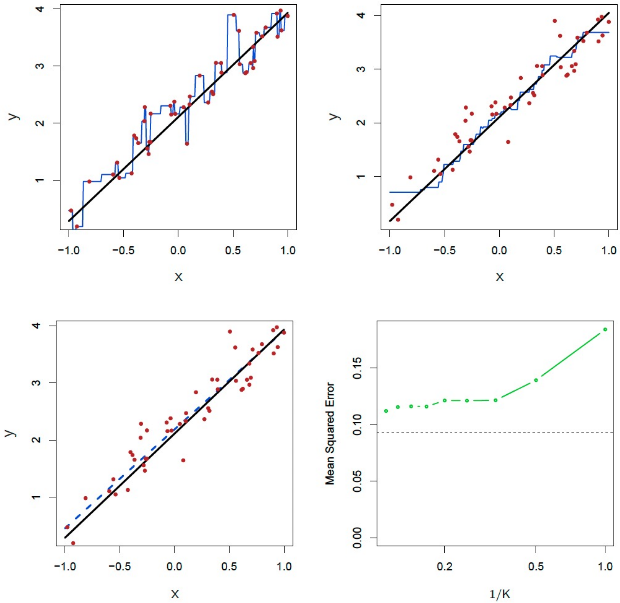
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

KNN regression

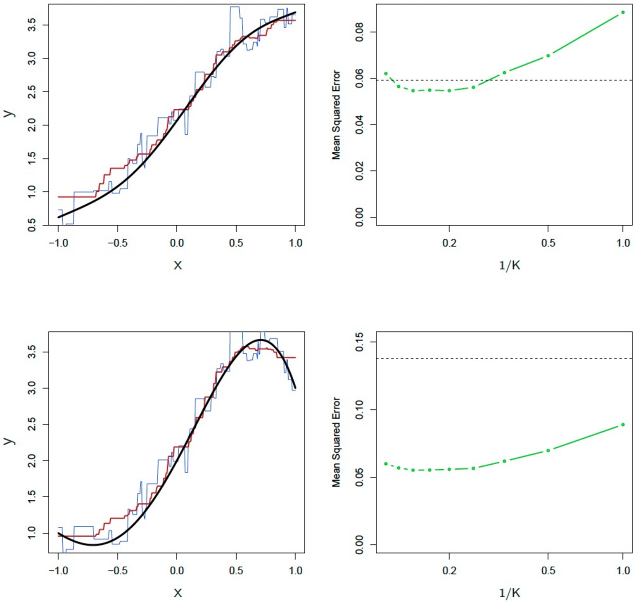
The optimal value for K will depend on the bias-variance tradeoff. A small value for K provides the most flexible fit, which will have low bias but high variance.



Parametric vs Non-parametric



Parametric vs Non-parametric


y
x
Mean Squared Error
1/K
y
x
Mean Squared Error
1/K