

Statistical Learning for Data Science

Lecture 12

唐晓颖

电子与电气工程系
南方科技大学

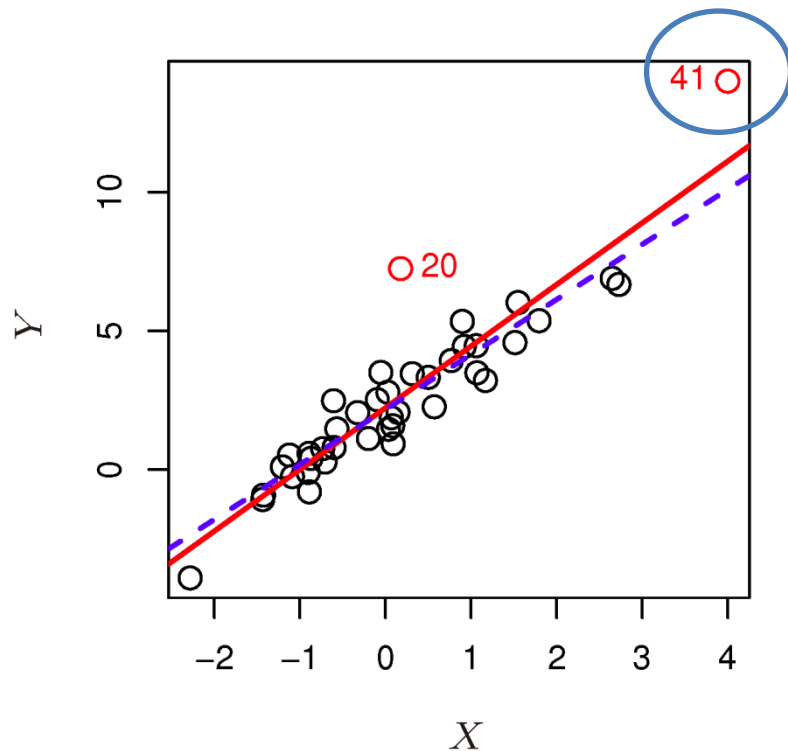
April 10, 2023

Other considerations in the regression model

- Potential problems of fitting a linear regression model

High leverage points

Observations with *high leverage* have an unusual value for x_i .



--red solid line: the least squares regression fit of all data

--blue dash line: the least squares regression fit after removal of the high leverage point

Removing the high leverage observation has a much more substantial impact on the least squares line than removing the outlier

Definition [\[edit\]](#)

In the [linear regression](#) model, the leverage score for the i -th data unit is defined as:

$$h_{ii} = [\mathbf{H}]_{ii},$$

the i -th diagonal element of the [projection matrix](#) $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, where \mathbf{X} is the [design matrix](#). The leverage score is also known as the observation self-sensitivity or self-influence,^[2] as shown by

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i},$$

where \hat{y}_i and y_i are the fitted and measured observation, respectively.

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \varepsilon_i$$

This model can be written in matrix terms as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & w_1 & x_1 \\ 1 & w_2 & x_2 \\ 1 & w_3 & x_3 \\ 1 & w_4 & x_4 \\ 1 & w_5 & x_5 \\ 1 & w_6 & x_6 \\ 1 & w_7 & x_7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

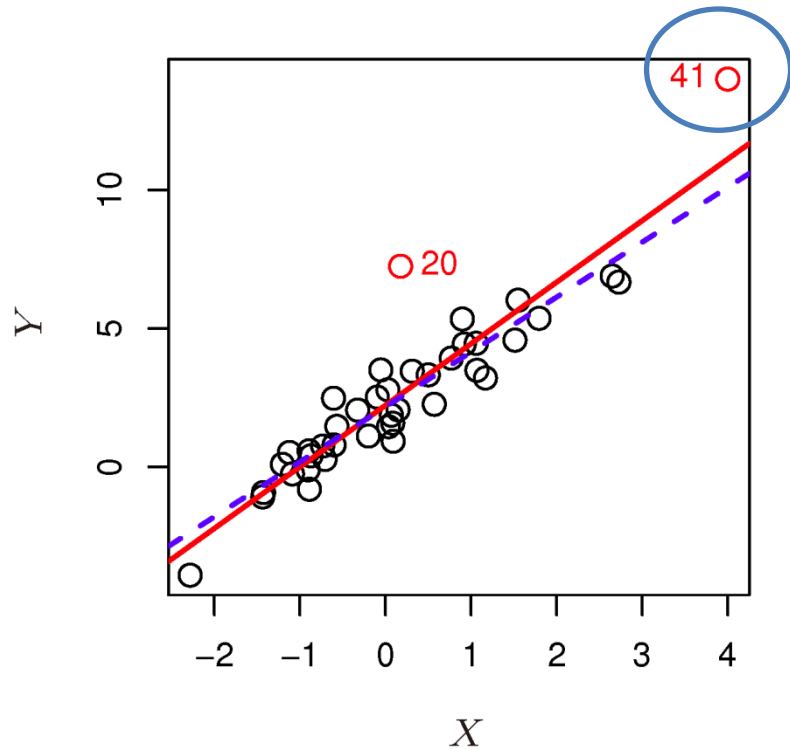
Here the 7×3 matrix on the right side is the design matrix.

Other considerations in the regression model

■ Potential problems of fitting a linear regression model

High leverage points

Observations with *high leverage* have an unusual value for x_i .



Note:

- High leverage observations tend to have a sizable impact on the estimated regression line.
- It is cause for concern if the least squares line is heavily affected by just a couple of observations, because any problem with these points may invalidate the entire fit.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

High leverage points

Observations with *high leverage* have an unusual value for x_i .

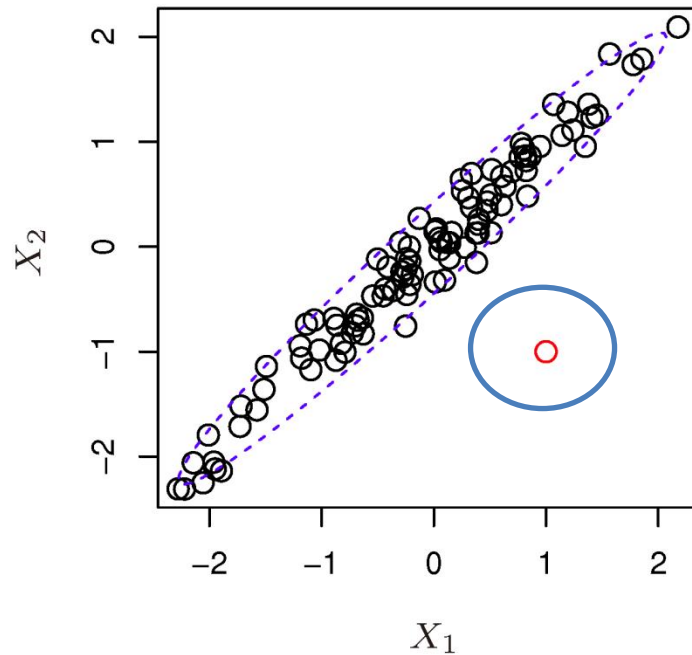
- In a simple linear regression, high leverage observations are fairly easy to identify.
- In a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictors.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

High leverage points

Observations with *high leverage* have an unusual value for x_i .



This problem is more pronounced in multiple linear regression settings with more than two predictors, because then there is no simple way to plot all dimensions of the data simultaneously.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

High leverage points

How to identify?

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,

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

Other considerations in the regression model

- Potential problems of fitting a linear regression model

High leverage points

For a simple linear regression,

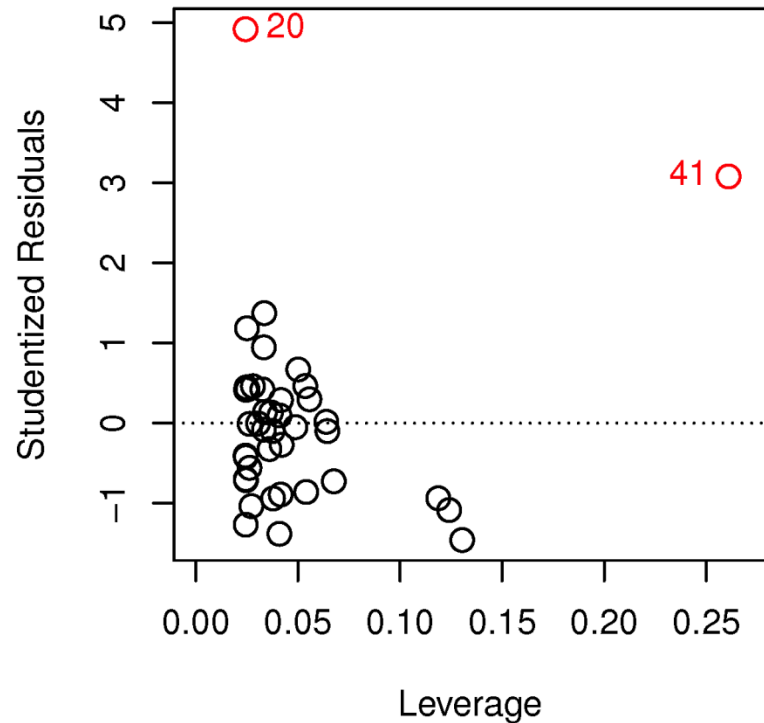
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^n (x_{i'} - \bar{x})^2}$$

- h_i increases with the distance of x_i from \bar{x}
- If a given observation has a leverage statistic that greatly exceeds the average, then we may suspect that the corresponding point has high leverage

Other considerations in the regression model

■ Potential problems of fitting a linear regression model

High leverage points



- Observation 41 stands out as having a very high leverage statistic as well as a high studentized residual. This is a particularly dangerous combination.
- This plot also reveals the reason that observation 20 had relatively little effect on the least squares fit: it has low leverage.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

High leverage points

How to identify?

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 multiple linear regression,

$$h_{ii} = [\mathbf{H}]_{ii}, \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$



The design matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

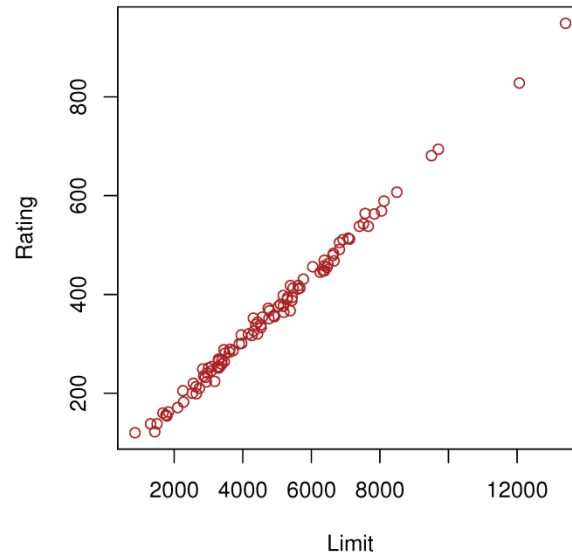
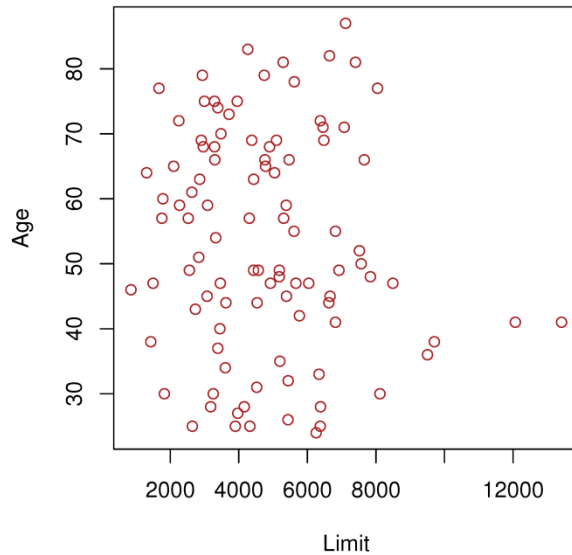
$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

Collinearity refers to the situation in which two or more predictors/variables are closely related to each other.



Credit card set

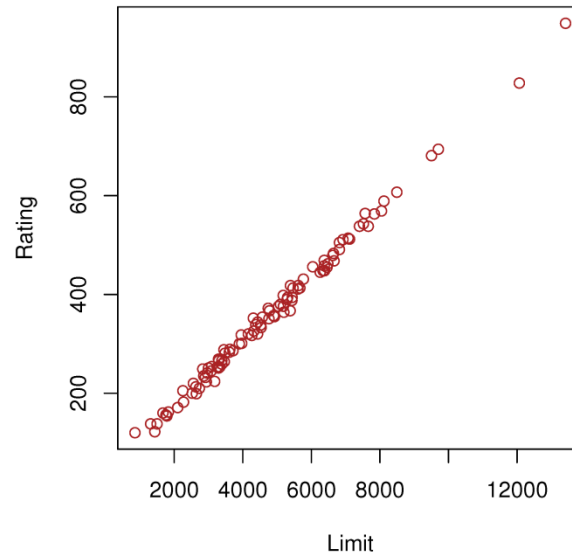
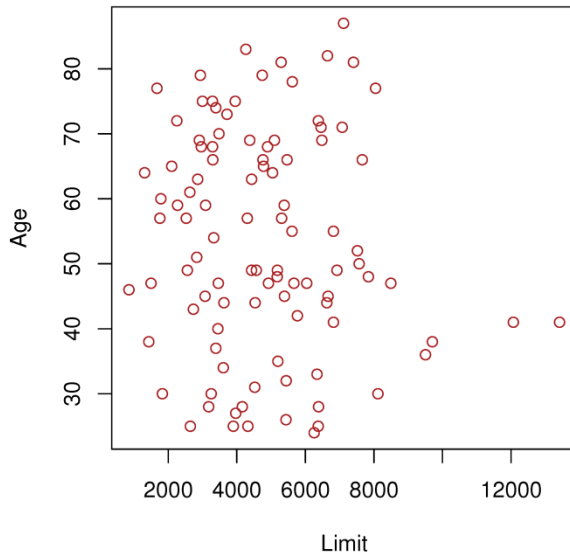
- The two predictors **limit** and **age** appear to have no obvious relationship.
- The predictors **limit** and **rating** are very highly correlated with each other, and we say that they are **collinear**.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

Collinearity refers to the situation in which two or more predictors/variables are closely related to each other.



Credit card set

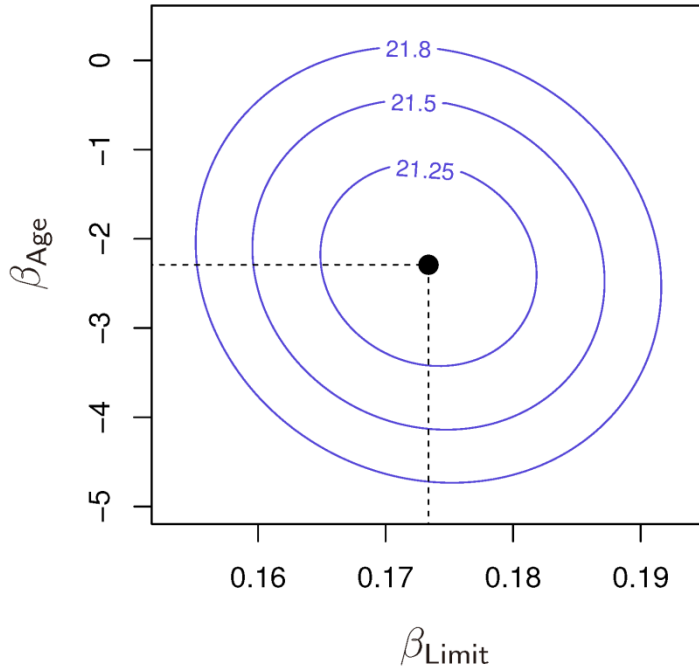
Since **limit** and **rating** tend to increase or decrease together, it can be difficult to determine how each one separately is associated with the response, **balance**.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

Collinearity refers to the situation in which two or more predictors variables are closely related to each other.



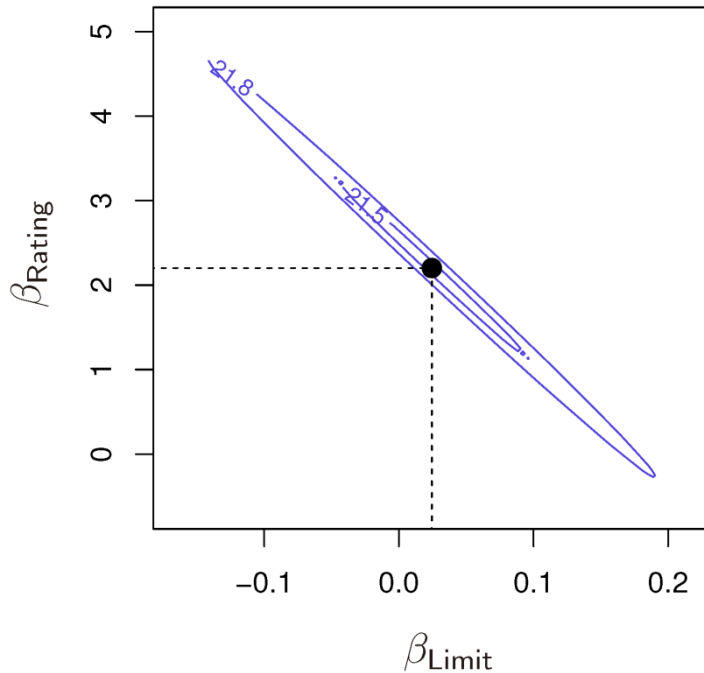
The axes for limit and age have been scaled so that the plot includes possible coefficient estimates that are up to four standard errors on each side of the least squares estimates. Thus the plot includes all plausible values for the coefficients. For example, we see that the true limit coefficient is almost certainly somewhere between 0.15 and 0.2

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to each other.



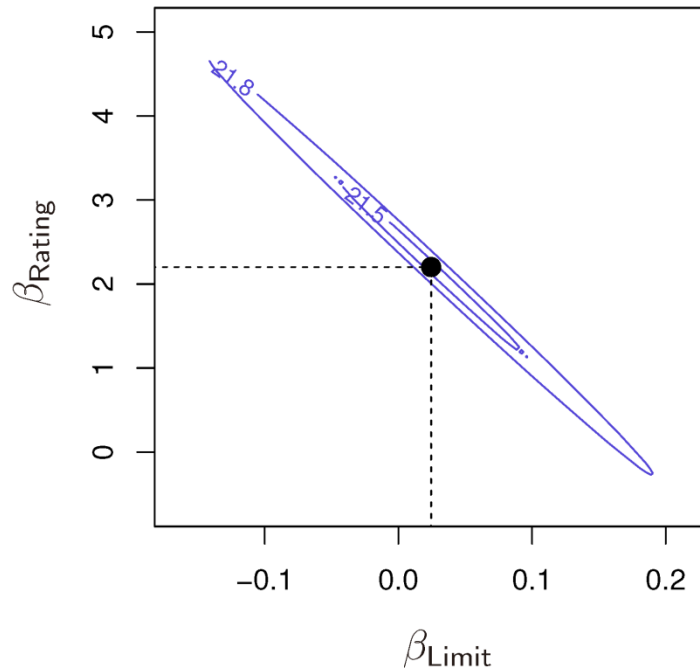
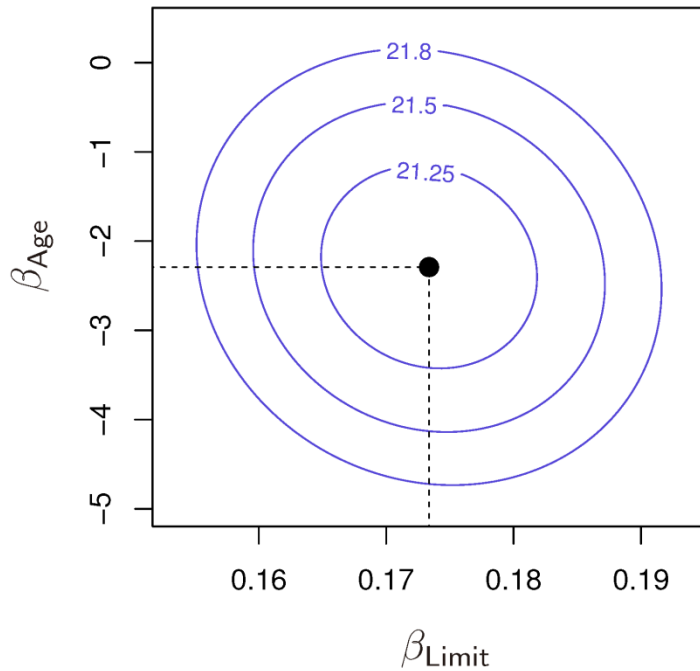
The contours run along a narrow valley; there is a broad range of values for the coefficient estimates that result in equal values for RSS. Hence a small change in the data could cause the pair of coefficient values that yield the smallest RSS – that is, the least squares estimates – to move anywhere along this valley. This results in a great deal of uncertainty in the coefficient estimates.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to each other.



[0.15,0.2]
Versus
[-0.2,0.2]

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

- Collinearity reduces the accuracy of the estimates of the regression coefficients.
- The standard error for each coefficient estimate grows.
- $t\text{-statistic} = \text{coefficient estimate} / \text{SE}$, a decline in the t-statistic
- We might fail to reject the null hypothesis $H_0 : \beta_j = 0$

Conclusion: the power of the hypothesis test – the probability of correctly detecting a non-zero coefficient – is reduced by collinearity

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

Observations:

1. SE for limit: 0.005 versus 0.064
2. p-value for limit: <0.0001 versus 0.7012

The importance of the limit variable has been masked due to the presence of collinearity

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The correlation matrix of the predictors

An element of the correlation matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF)

Not all collinearity problems can be detected by inspection of the correlation matrix: it is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation (*multicollinearity*).

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF)

As the name suggests, a VIF quantifies how much the variance is inflated. But what variance? Recall that we learned previously that the standard errors — and hence the variances — of the estimated coefficients are inflated when multicollinearity exists. So, the VIF for the estimated coefficient $\hat{\beta}_j$ — denoted VIF_j — is just the factor by which the variance is inflated.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF)

For the model in which x_j is the only predictor

$$y_i = \beta_0 + \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$$\text{var}(\hat{\beta}_j)_{\min} = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

Note that we add the subscript "min" in order to denote that it is the smallest the variance can be

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF)

For the full model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_j x_{ij} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2}$$

→ The R^2 -statistic obtained by regressing the j^{th} predictor on the remaining predictors

Other considerations in the regression model


- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF)

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2}$$

 The R^2 -statistic obtained by regressing the j^{th} predictor on the remaining predictors

The greater the linear dependence among the predictor x_j and the other predictors, the larger the R_j^2 value. And, as the above formula suggests, the larger the R_j^2 value, the larger the variance of $\hat{\beta}_j$.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF)

$$VIF_j = \frac{\text{var}(\hat{\beta}_j)}{\text{var}(\hat{\beta}_j)_{\min}} = \frac{\frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2}}{\frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}} = \frac{1}{1 - R_j^2}$$

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF) $VIF_j = \frac{1}{1 - R_j^2}$

- The smallest possible value for VIF is 1 (complete absence of collinearity).
- A VIF exists for *each of the p predictors* in a multiple regression model.
- VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to identify?

The variance inflation factor (VIF) $VIF_j = \frac{1}{1 - R_j^2}$

Credit card set

$VIF(\text{age}) = 1.01$, $VIF(\text{rating}) = 160.67$, $VIF(\text{limit}) = 160.59$



There is considerable collinearity in the data!

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Collinearity

How to fix?

- Drop one of the problematic variables from the regression
e.g. Drop the rating predictor in the credit card data
- Combine the collinear variables together into a single predictor
e.g. take the average of the **standardized** versions of limit and rating in order to create a new variable.

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

Advantages:

- Easy to fit (only need to estimate a small number of coefficients)
- The coefficients have simple interpretations
- Tests of statistical significance can be easily performed

Disadvantages:

- Make a strong assumption 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.

Comparison of Linear Regression with K-Nearest Neighbors

KNN regression is an example of a non-*parametric* approach

Definition: 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 N_0 . It then estimates $f(x_0)$ using the average of all the training responses in N_0 . In other words,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

Advantages:

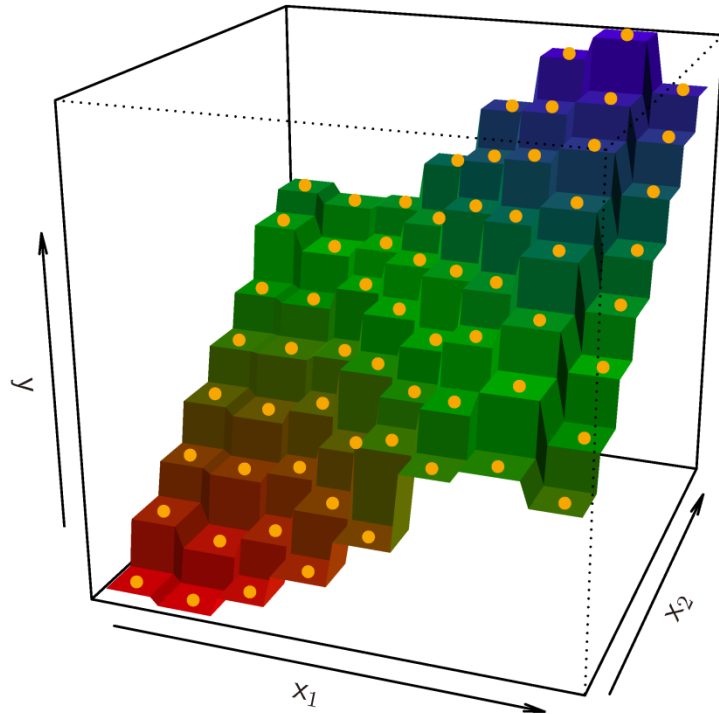
- Don't explicitly assume a parametric form for $f(X)$, and thereby provide an alternative and more flexible approach for performing regression.

Comparison of Linear Regression with K-Nearest Neighbors

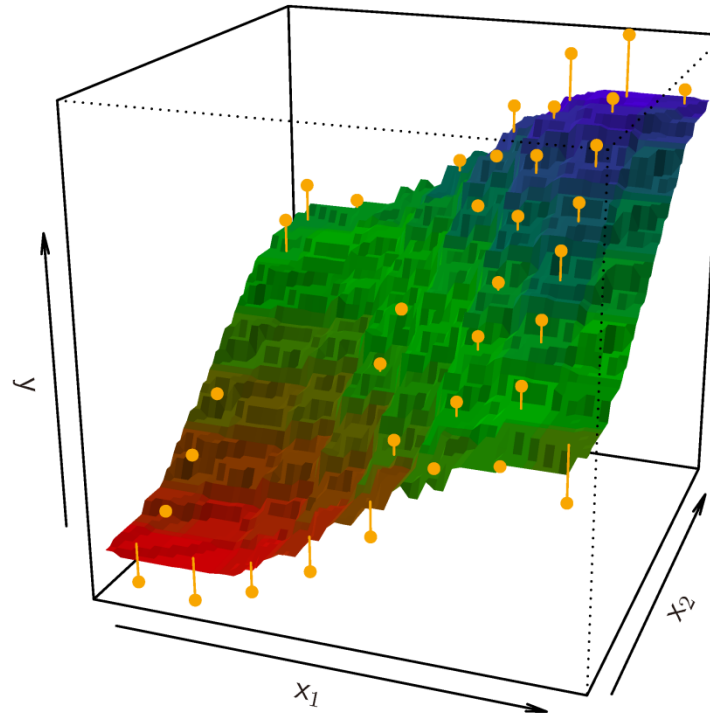
Linear regression is an example of a *parametric* approach

KNN regression

$K = 1$



$K = 9$



Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

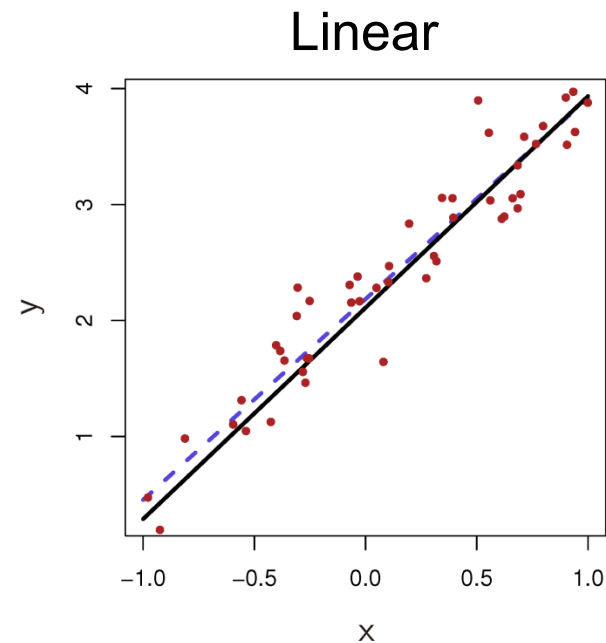
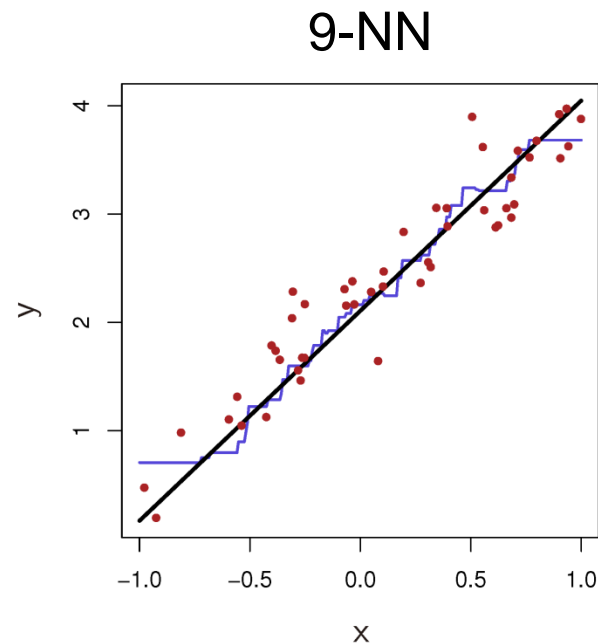
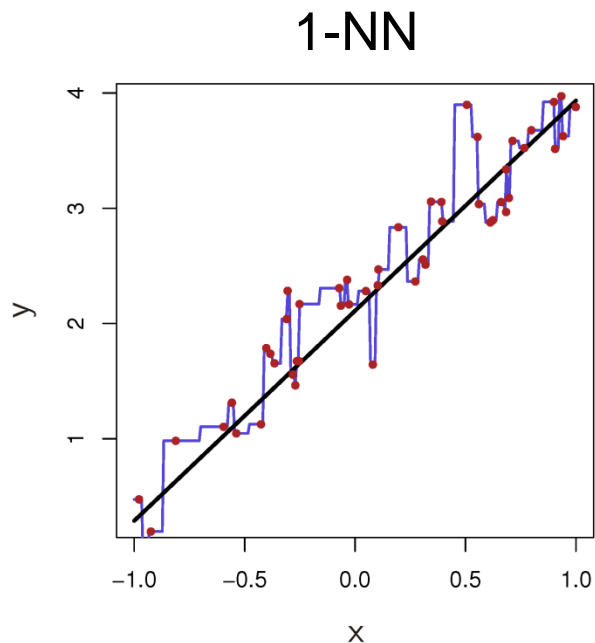
KNN regression

NOTE: In general, 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. The variance is due to the fact that the prediction in a given region is entirely dependent on just one or several observations. In contrast, large values for K provide a smoother and less variable fit; the prediction in a region is an average of multiple points, and so change one observation has a smaller effect. However, the smoothing may cause bias by masking some of the structure in $f(X)$.

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

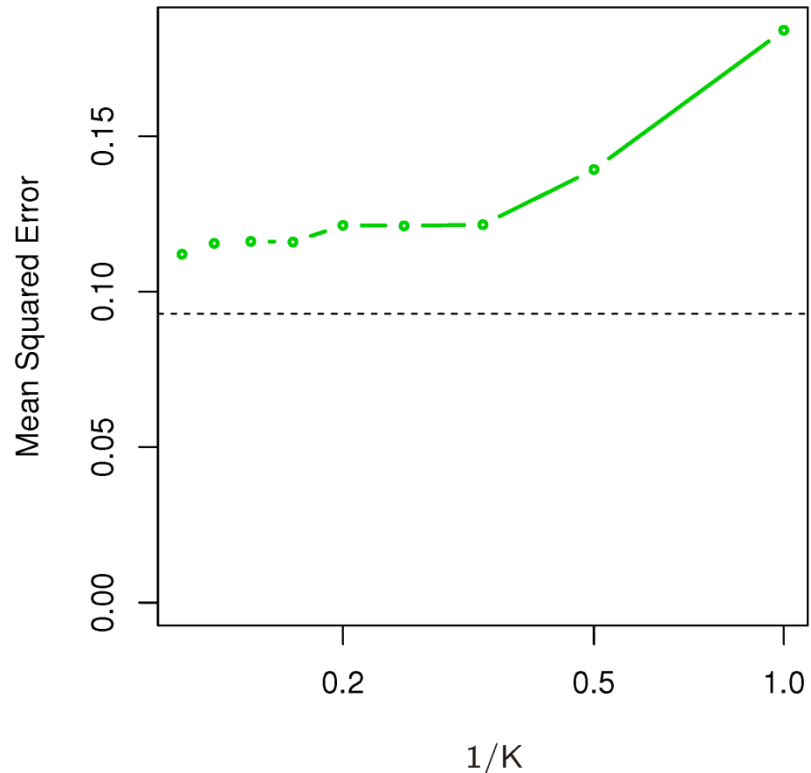
NOTE: the parametric approach (such as the least squares linear regression) will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of f .



Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

NOTE: the parametric approach (such as the least squares linear regression) will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of f .



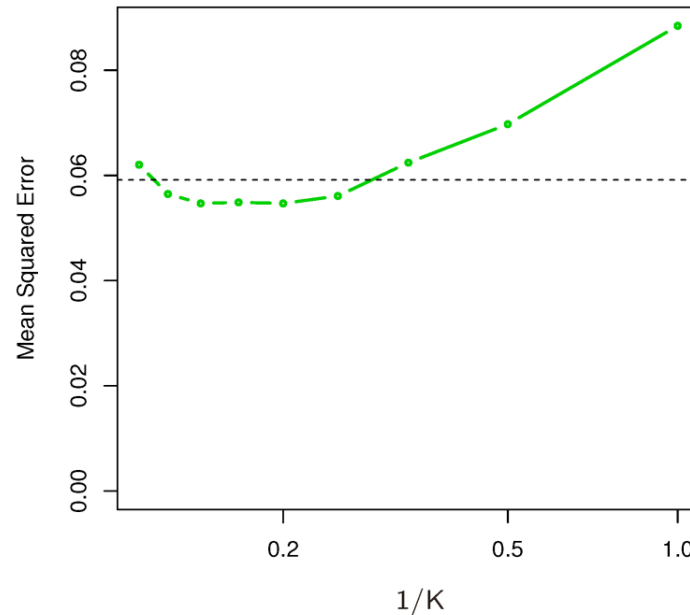
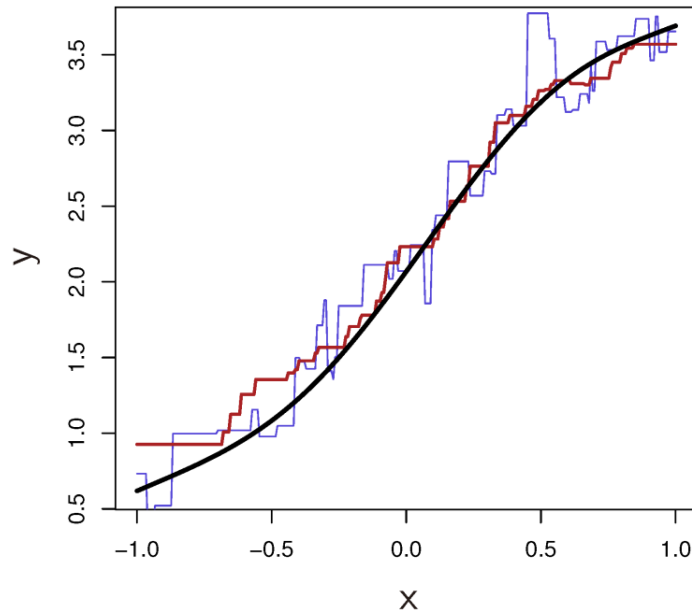
Black: the test MSE for linear regression

Green: the test MSE for KNN regression

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

NOTE: the parametric approach (such as the least squares linear regression) will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of f .

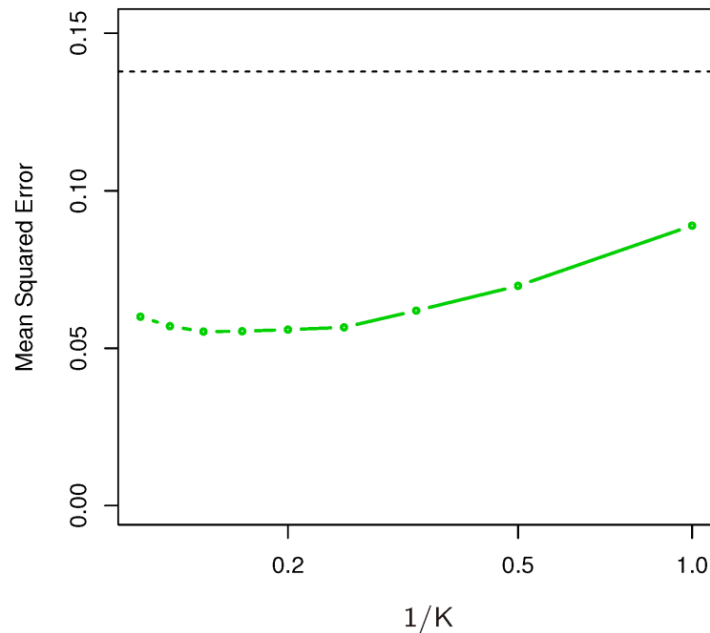
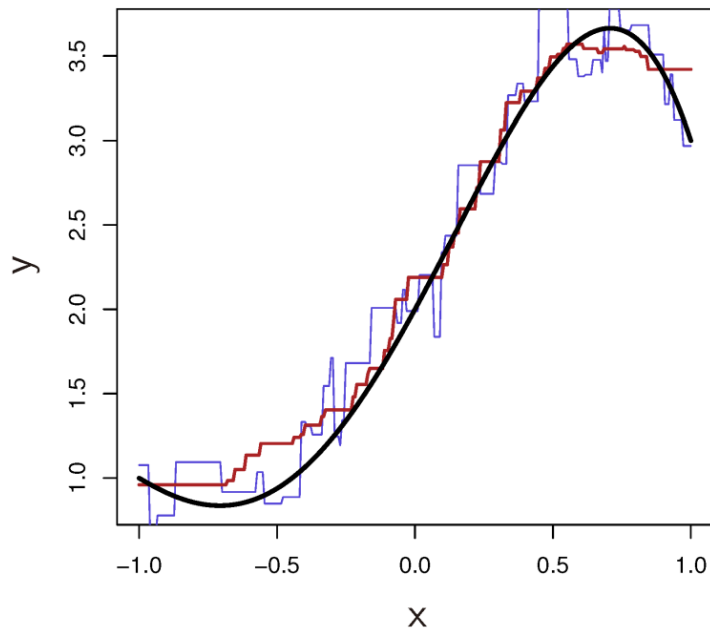


The test MSE for linear regression is still superior to that of KNN for low values of K . However, for $K \geq 4$, KNN outperforms linear regression.

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

NOTE: the parametric approach (such as the least squares linear regression) will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of f .



KNN substantially outperformed linear regression for all values of K .

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

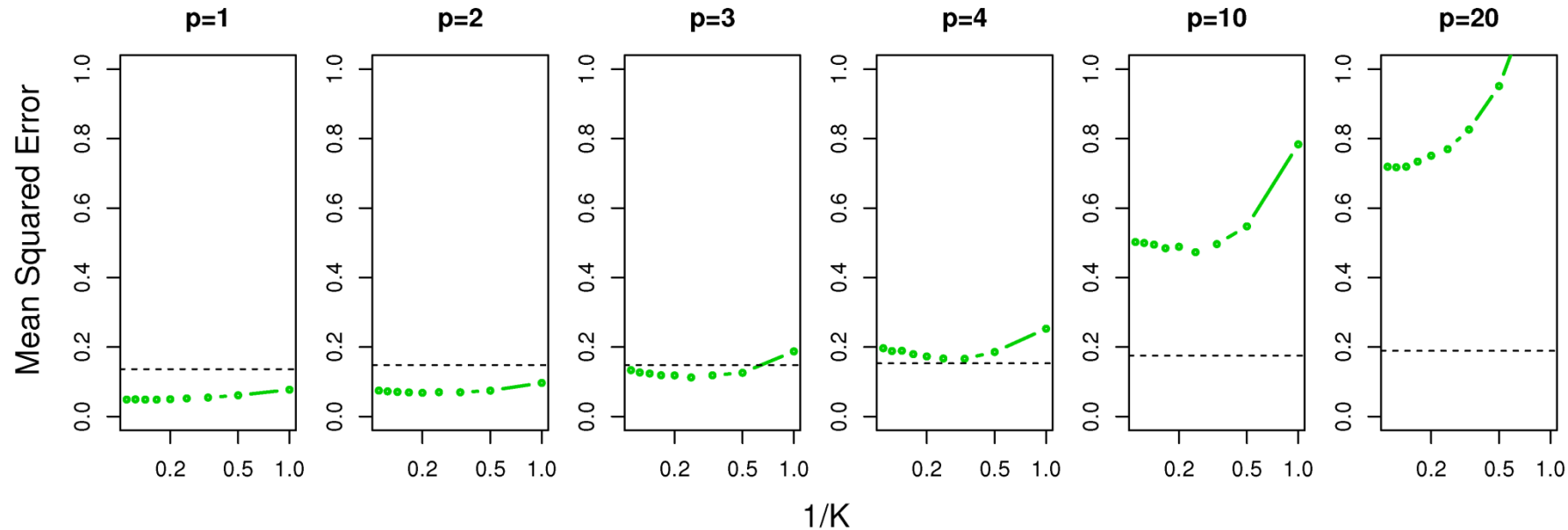
- In a real life situation in which the true relationship is unknown, one might draw the conclusion that KNN is better than linear regression because KNN may provide superior results over linear regression if the true relationship is nonlinear. However, KNN may provide substantially better results if the true relationship is nonlinear.
- In reality, even when the true relationship is highly non-linear, KNN may still provide inferior results to linear regression, especially in higher dimensions ($p > 1$).

Not completely correct!

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

Adding noise predictors that are not associated with the response

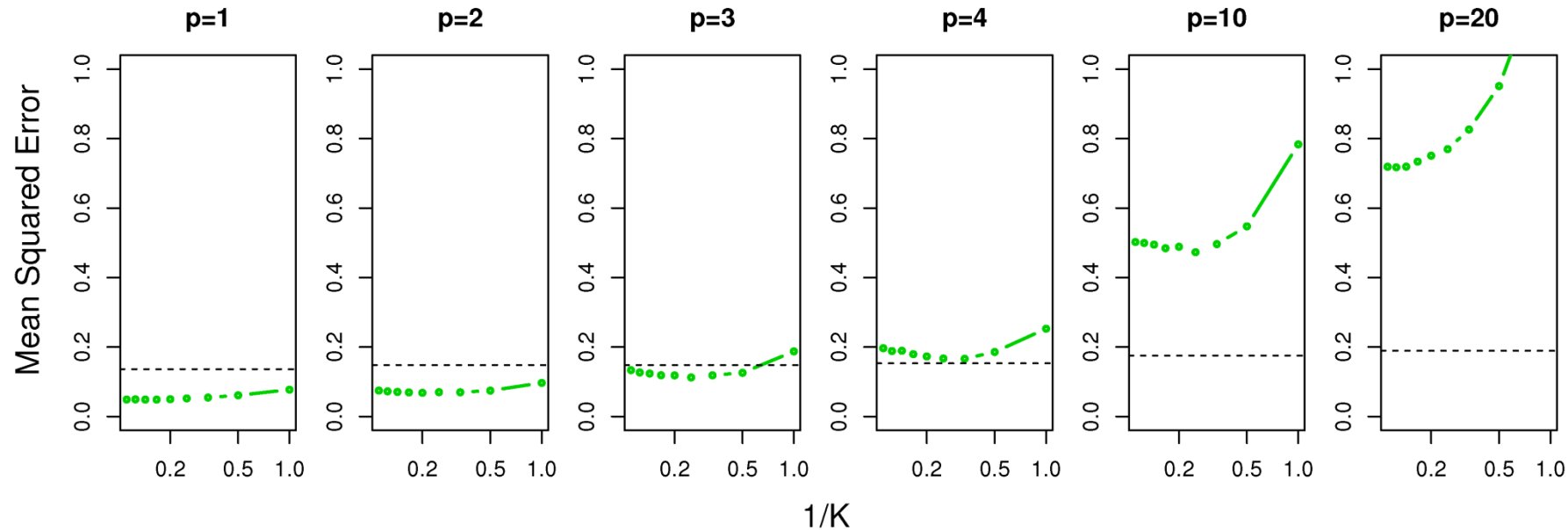


- When $p=1$ or $p=2$, KNN outperforms linear regression
- When $p=3$, the results are mixed
- When $p \geq 4$, linear regression is superior to KNN

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

Adding noise predictors that are not associated with the response

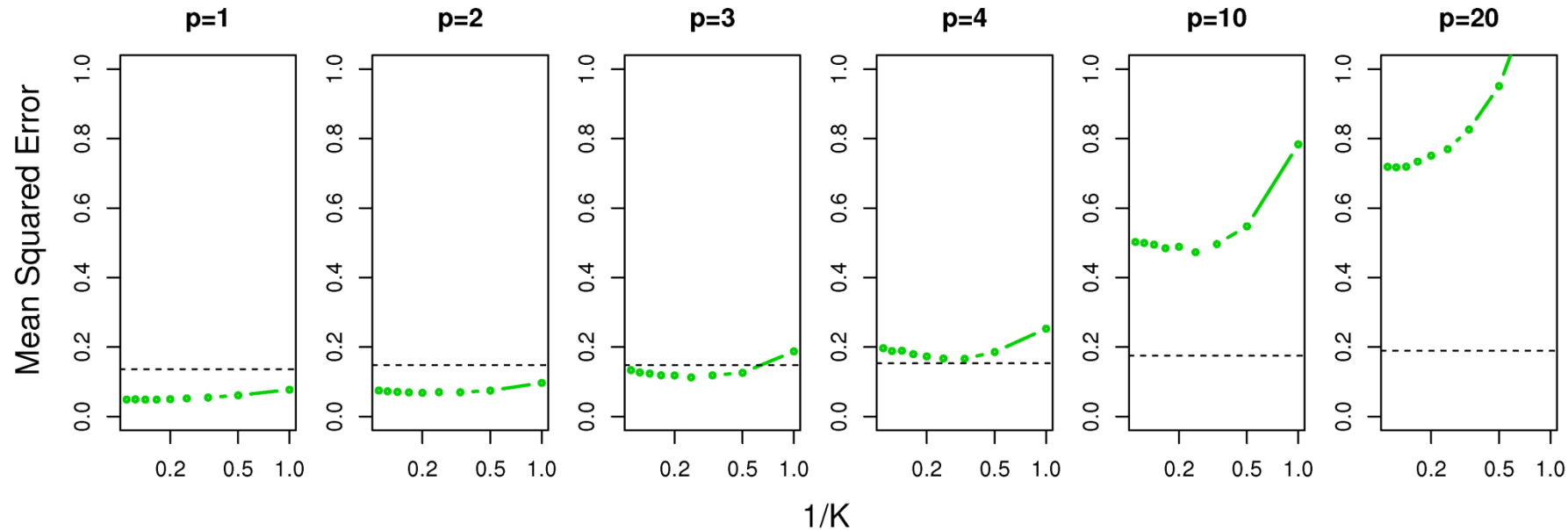


- The increase in dimension has only caused a small deterioration in the linear regression test set MSE, but it has caused more than a ten-fold increase in the MSE for KNN. This decrease in performance as the dimension increases is a common problem for KNN.

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

Adding noise predictors that are not associated with the response



- $n=100$, when $p=1$, this provides enough information to accurately estimate $f(X)$. However, spreading 100 observations over $p=20$ dimensions results in a phenomenon in which a given observation has no nearby neighbors – “curse of dimensionality”.

Comparison of Linear Regression with K-Nearest Neighbors

Linear regression is an example of a *parametric* approach

Suggestions

- As a general rule, parametric method will tend to outperform non-parametric approaches when there is a small number of observations per predictor.
- Even in problems in which the dimension is small, we might prefer linear regression to KNN from an interpretability standpoint.