Chp. 3: Linear Regression

CS 6823/MBS 6251

Outline

Regression

- Linear Regression
- Non-parametric approach (KNN Regression)
- Considerations in High Dimensions
- Lab 3

Advertising dataset

Describes sales (in thousands of units) as a function of advertising budgets (in thousands of dollars) for TV, radio, and newspaper for $n = 200$ different markets.

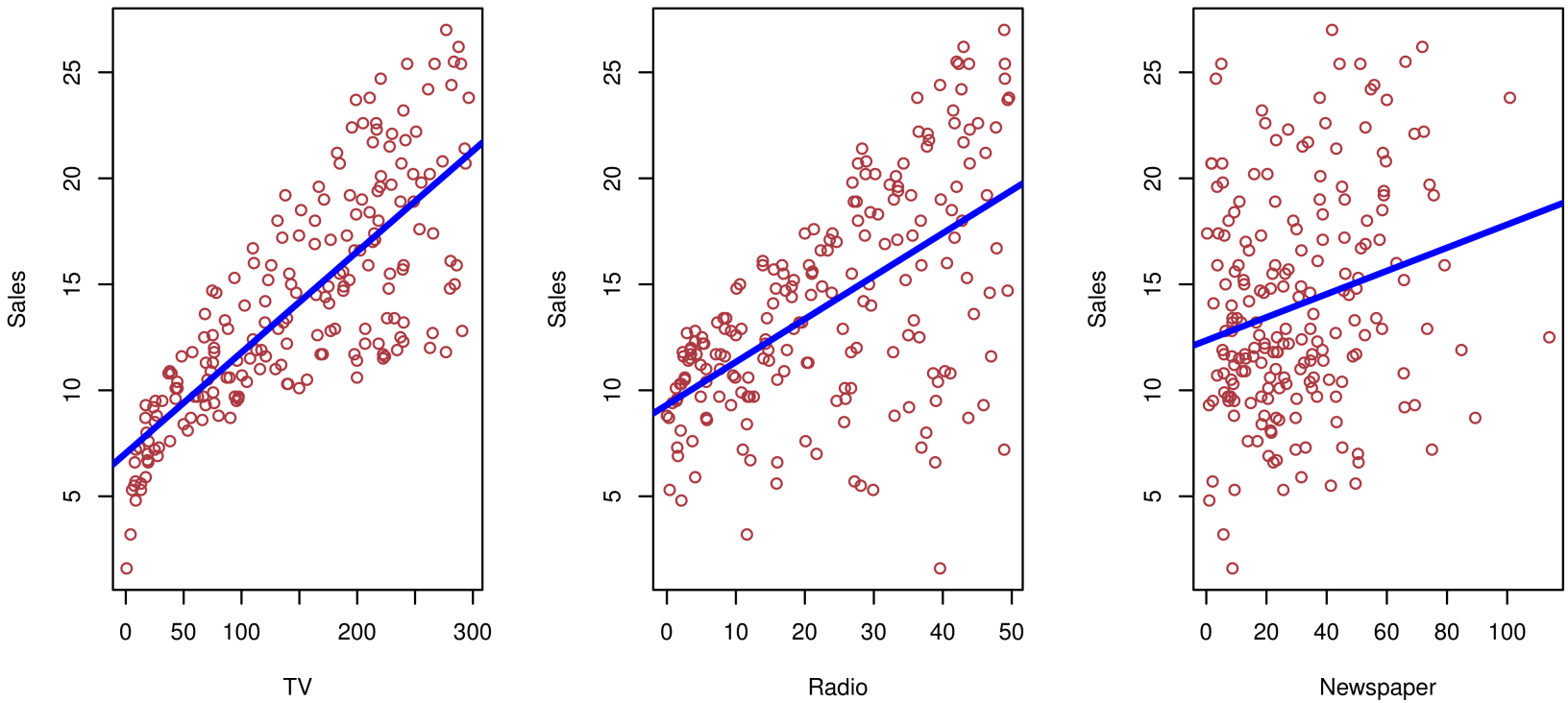

Figure 1: Fig 2.1

Questions you can answer with linear regression:

- Is there a relationship between predictor(s) and response?
- How strong is this relationship?
- Which predictors contribute to the response?
- How accurately can we estimate the effect of each predictor?
- How accurately can we predict the response?
- Is the relationship linear?
- Is there synergy among predictors?

Simple Linear Regression

Predict a quantitative response Y on the basis of a single predictor X.

$$\textit{sales} \approx \beta_0 + \beta_1 \times \textit{TV}$$

Use training data to produce estimates for the model coefficients, $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Estimating the Coefficients

Most common approach to produce coefficient estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is to minimize the *least squares* criterion.

Estimating the Coefficients

Consider the least squares fit for regression of *sales* onto *TV*:

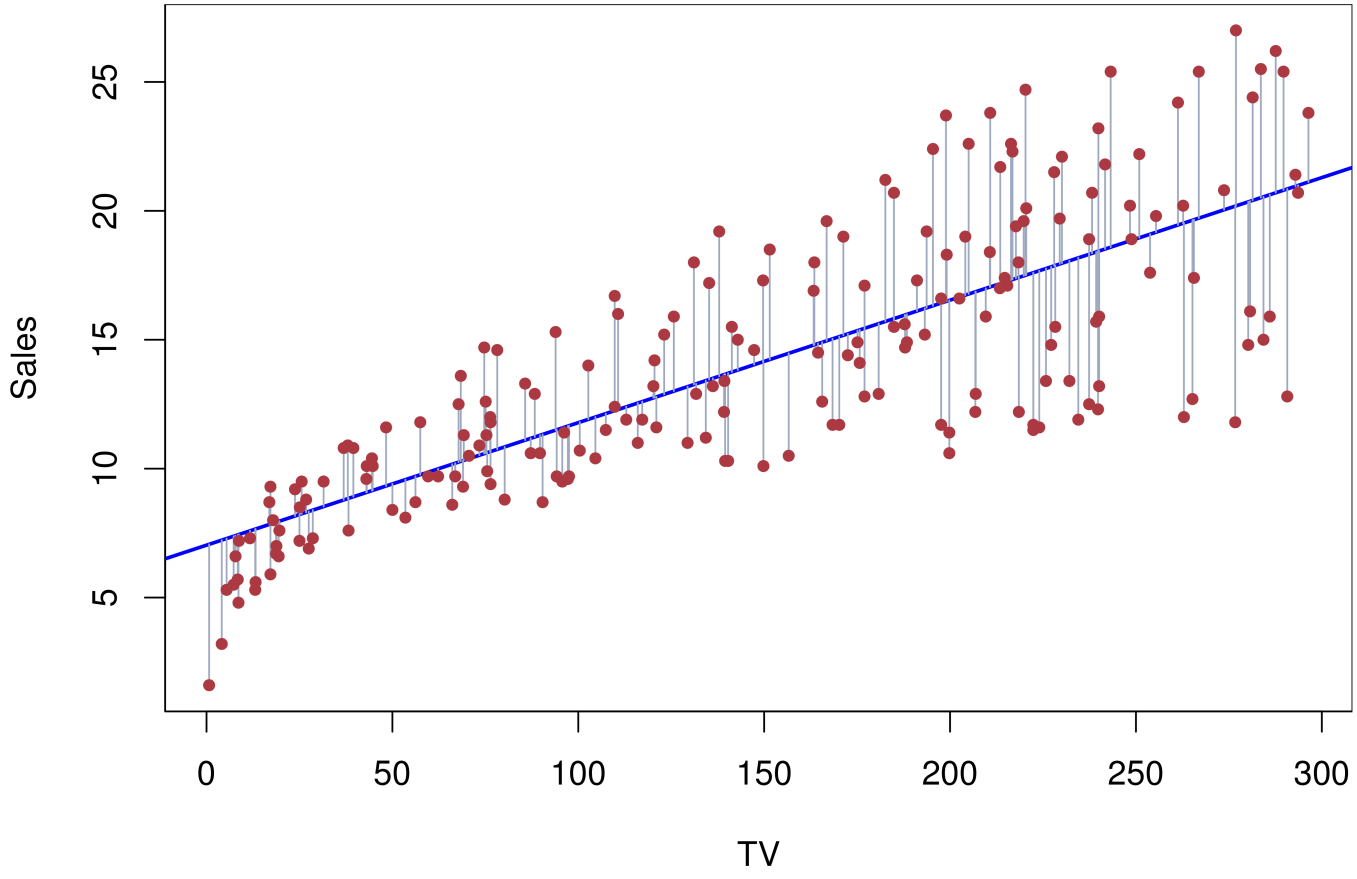

Figure 2: Fig 3.1

Fig. 3.1

What does each grey line represent?

A. the ith residual

B. $y_i - \widehat{y}_i$

C. error of the ith observation

D. squared error of the ith observation

E. residual sum of squares

Estimating the Coefficients

The residual sum of squares (RSS) is given by:

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

where $e_i = y_i - \widehat{y}_i$ represents the ith residual and $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$.

Estimating the Coefficients

The *least squares* approach chooses $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize the residual sum of squares (RSS).

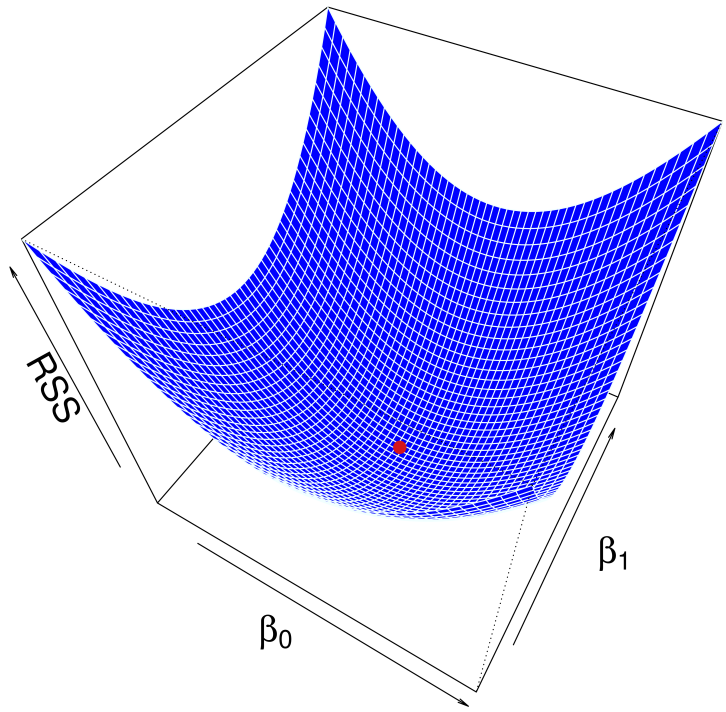

Figure 3: Fig 3.2

Fig. 3.2

Accuracy of the Model

How well does the model fit the data?

One measure of the lack of fit of model to the data is the **Residual Standard Error**.

- Estimate of the standard deviation of ϵ in $Y = f(X) + \epsilon$.
- Average amount that the response will deviate from the true regression line.
- Given by $\text{RSE} = \sqrt{\text{RSS}/(n-2)}$.

Accuracy of the Coefficient Estimates

The Residual Standard Error (σ) determines the standard error (SE) of $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\mathrm{SE}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Accuracy of the Coefficient Estimates

We can also compute *t-statistics* & *p-values* to test the null hypothesis that there is no relationship between X and Y, $H_0 : \beta_1 = 0$, against the alternative $H_a : \beta_1 \neq 0$.

t measures the number of standard deviations that $\widehat{\beta}_1$ is from 0:

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

p is the probability of observing any number equal to $|t|$ or larger, assuming $\beta_1 = 0$.

Accuracy of the Model

RSE is measured in units of Y and can be difficult to interpret.

R^2, the proportion of variability in Y explained by X, is not measured in units of Y

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

where the total sum of squares, $TSS = \sum(y_i - \bar{y})^2$

Multiple Linear Regression

Predict a response **on the basis of multiple predictors.**

$$\textit{sales} \approx \beta_0 + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio} + \beta_3 \times \textit{newspaper} + \epsilon$$

β_j is the average effect on Y of a one unit increase in X_j, *holding all other predictors fixed.*

Multiple Linear Regression

What do the vertical black lines represent?

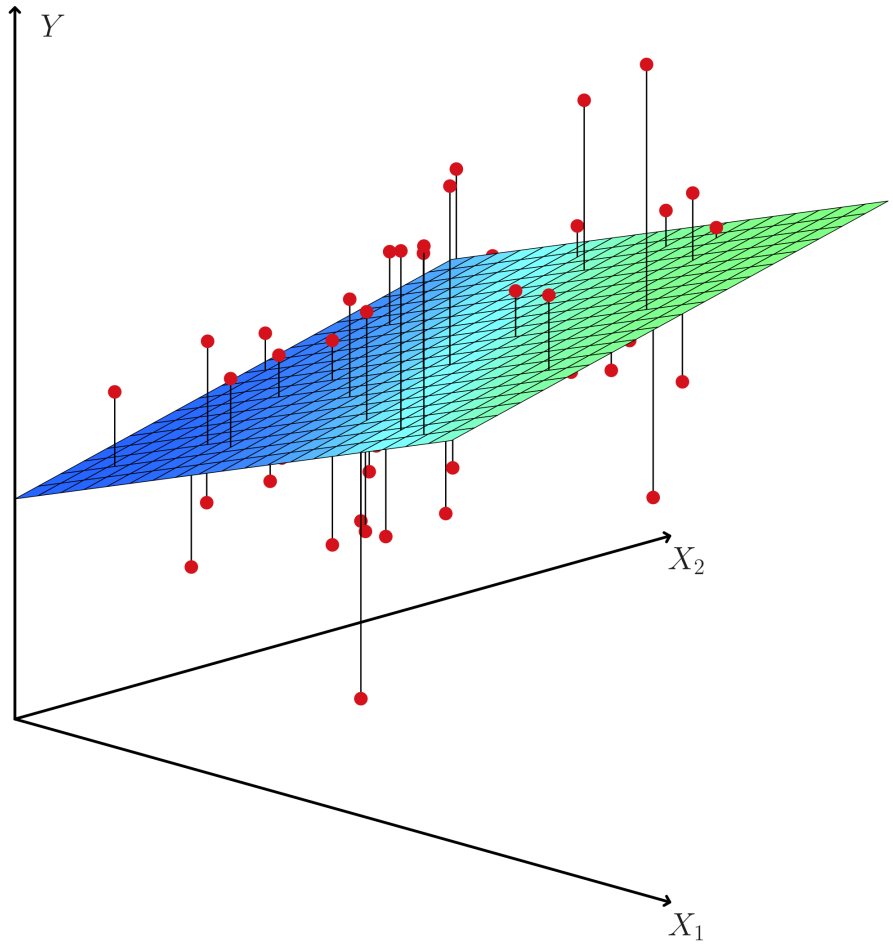

Figure 4: Fig 3.4

Multiple Linear Regression

F-statistic to test $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ vs. the alternative, at least one β_j is non-zero.

Unlike multiple individual tests for each β_j, the F-statistic adjusts for the number of predictors.

Considerations of Regression Models

- Qualitative predictors (code as 'dummy' variables), non-additive relationships b/w predictors and response (interaction terms), non-linear relationships (polynomial terms, log transformations)
- Potential issues such as non-linearity, correlation of error terms, non-constant variance of error terms, outliers, high leverage points, collinearity.
- Plots of residuals, correlation scatterplots of the predictors, and calculation of other statistics (leverage, studentized residuals, variance inflation factor) can help diagnose these issues.

Parametric vs. non-parametric regression

Which is true for *non-parametric* approaches?

A. Do not assume a specific form of $f(X)$.

B. Are more flexible than parametric approaches.

C. Are better than parametric approaches.

Non-parametric regression

One example of a non-parametric approach is *K-nearest neighbors regression*.

KNN:

1. Identifies the K training observations that are nearest to x_0, represented by $\mathcal{N}_0$.
2. Estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$

Non-parametric regression

Consider a dataset with one predictor, where the true relationship between the predictor and the response is linear.

KNN with K = 1 (left) and K = 9 (right):

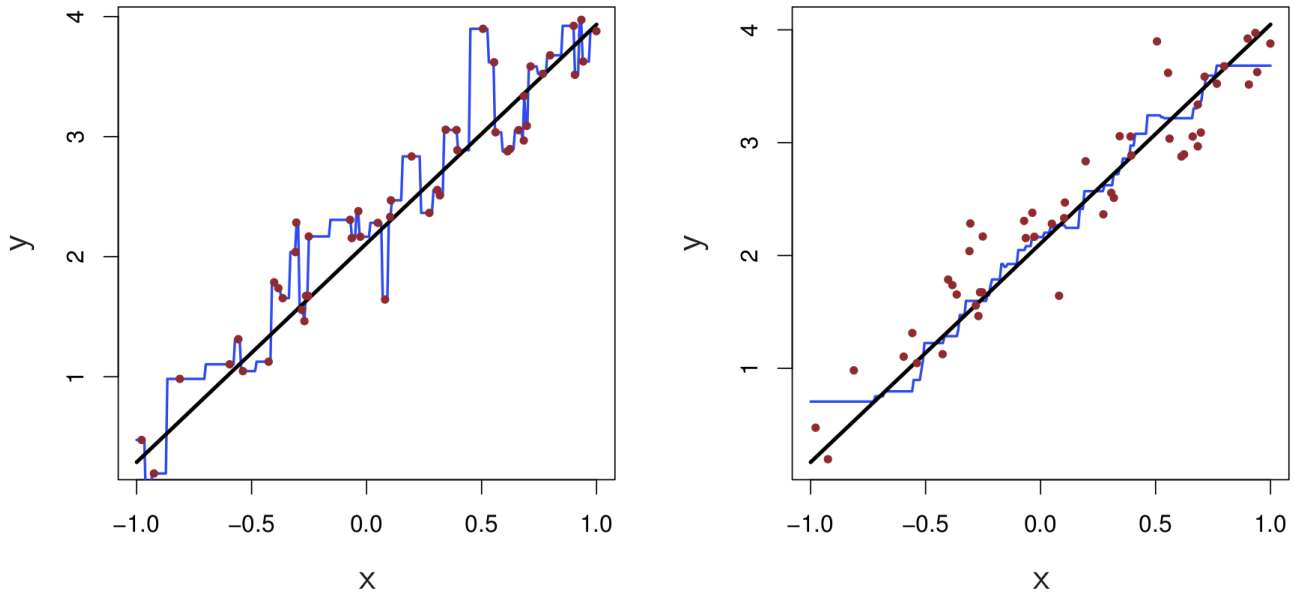

Figure 5: Fig 3.17

Fig. 3.17

Non-parametric regression

Even with large K, linear regression outperforms KNN when true relationship is linear. Cost in variance is not offset by a reduction in bias.

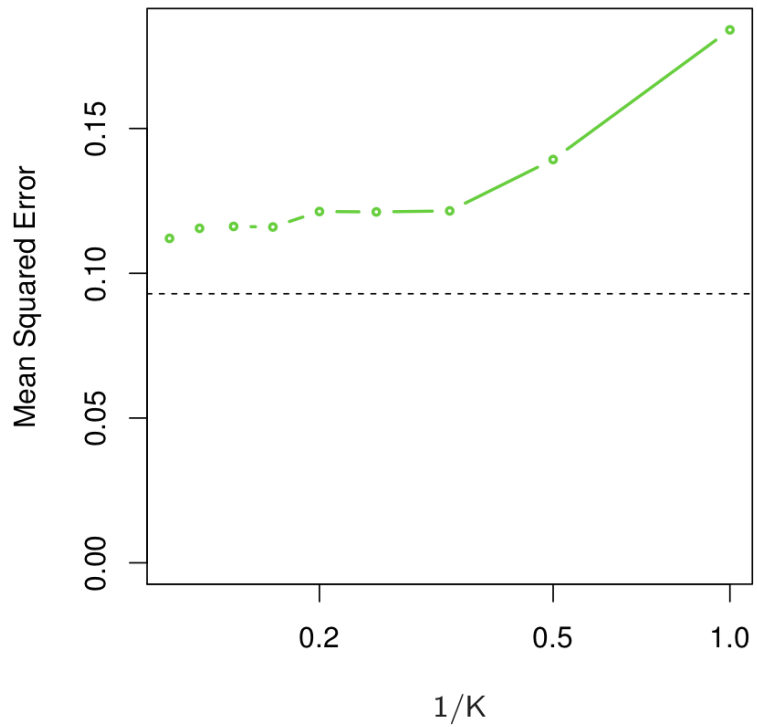

Figure 6: Fig 3.18

Fig. 3.18. Dashed line: least squares test set MSE; Green line: MSE for KNN as a function of K.

Non-parametric regression

Extension to multiple predictors:

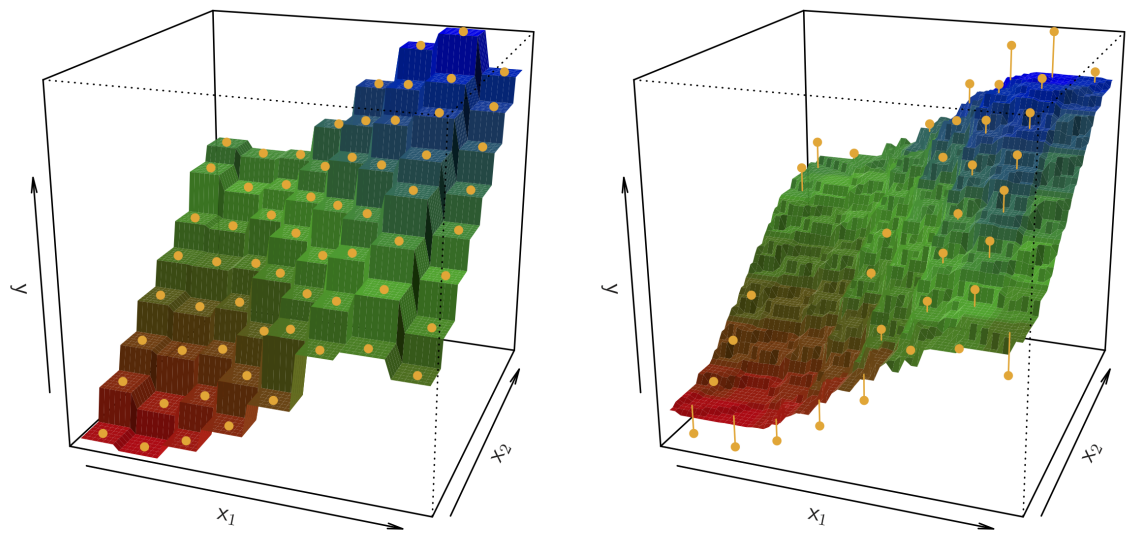

Figure 7: Fig 3.16

KNN with K = 1 (left) and K = 9 (right) in 2 dimensions

Non-parametric regression

- Parametric methods will generally outperform non-parametric methods when there is a small number of observations per predictor.
- In higher dimensions, a given observation may have no nearest neighbors (*curse of dimensionality*)

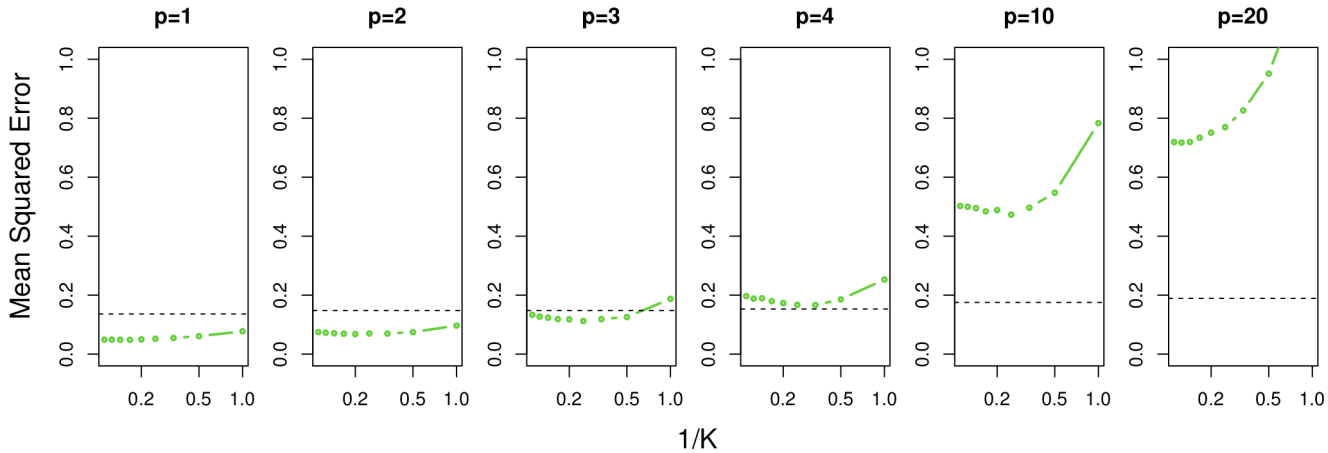

Figure 8: Fig 3.20

Fig. 3.20: True function is non-linear in first variable and does not depend on additional variables.

High-Dimensional Data

In high dimensions:

- Multicollinearity problem is extreme; any variable can be written as a linear combiation of all the other variables, and we can never know which variables are truly related to the outcome.
- Test error usually increases increases with the number of predictors, unless additional features truly are associated with the response
- Reducing flexibility of models (e.g. via regularization, Chp. 6) plays an important role