LINEAR REGRESSION

Part II

Outline

- Multiple linear regression

Advertising data

- How can we extend our analysis of the advertising data in order to accommodate radio and newspaper?

	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

`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

Drawbacks of fitting separate models

- It is difficult to make predictions of sales given TV, radio, and newspaper
- Each of the three regressions ignores the other two media in forming coefficient estimates

Multiple linear regression

- The model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

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

- Prediction formula

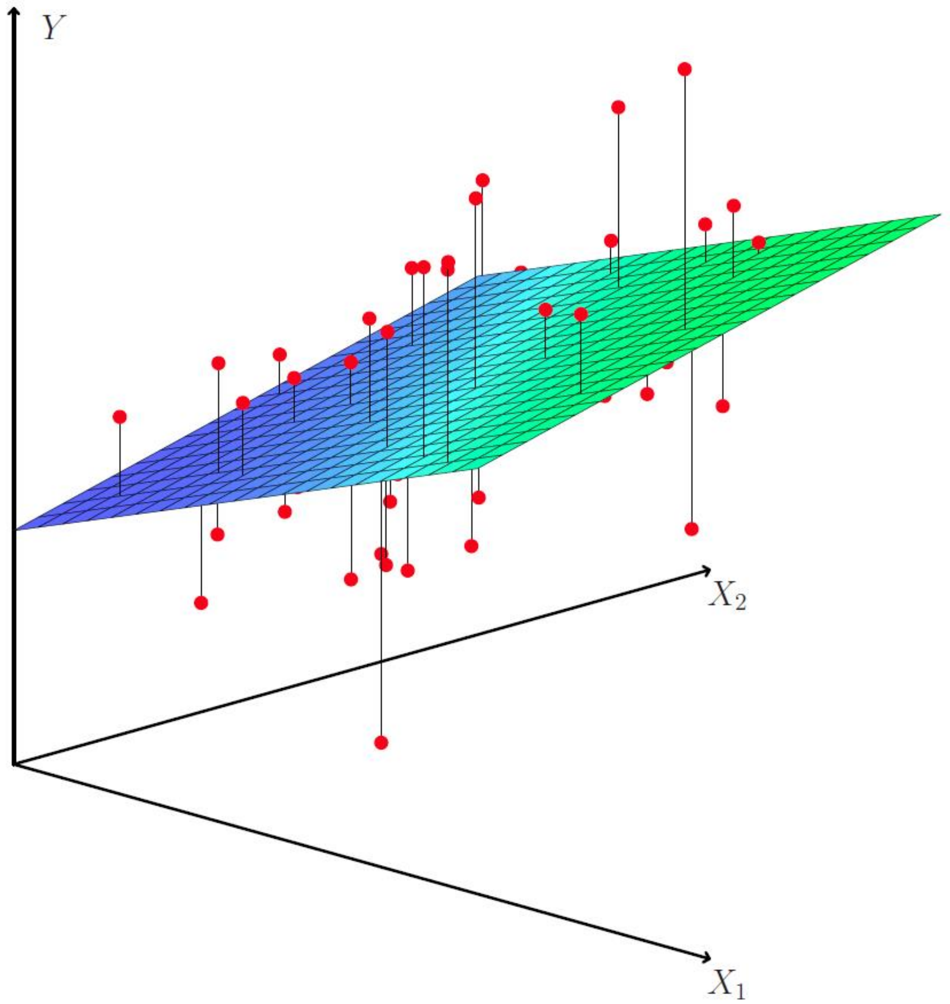
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

  - E.g., $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$

Least squares coefficient estimation

- Let $e_i = y_i - \hat{y}_i$
- We choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimize the sum of squared residuals

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$



An illustrative example with $p = 2$ predictors

	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.*

	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

Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

	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				1.0000

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

Important questions

- *Is at least one of the predictors useful in predicting the response?*
- *How well does the model fit the data?*
- *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*
- *Do all the predictors help to explain Y, or is only a subset of the predictors useful?*

Is there a relationship?

- Testing

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

versus

H_a: at least one β_j is non-zero

- F-statistic

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

- On one hand,

$$\mathrm{E}\left(\frac{\mathrm{RSS}}{n - p - 1}\right) = \sigma^2$$

- On the other hand, under H_0,

$$\mathrm{E}\left(\frac{\mathrm{TSS} - \mathrm{RSS}}{p}\right) = \sigma^2$$

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

TABLE 3.6. *More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the* `Advertising` *data. Other information about this model was displayed in Table 3.4.*

- Testing whether a subset of the coefficients are zero, e.g.,

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

- F-statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n - p - 1)}$$

	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.*

- Given these individual p-values for each variable, why do we need to look at the overall F-statistic?

Lack of fit

- Residual standard error

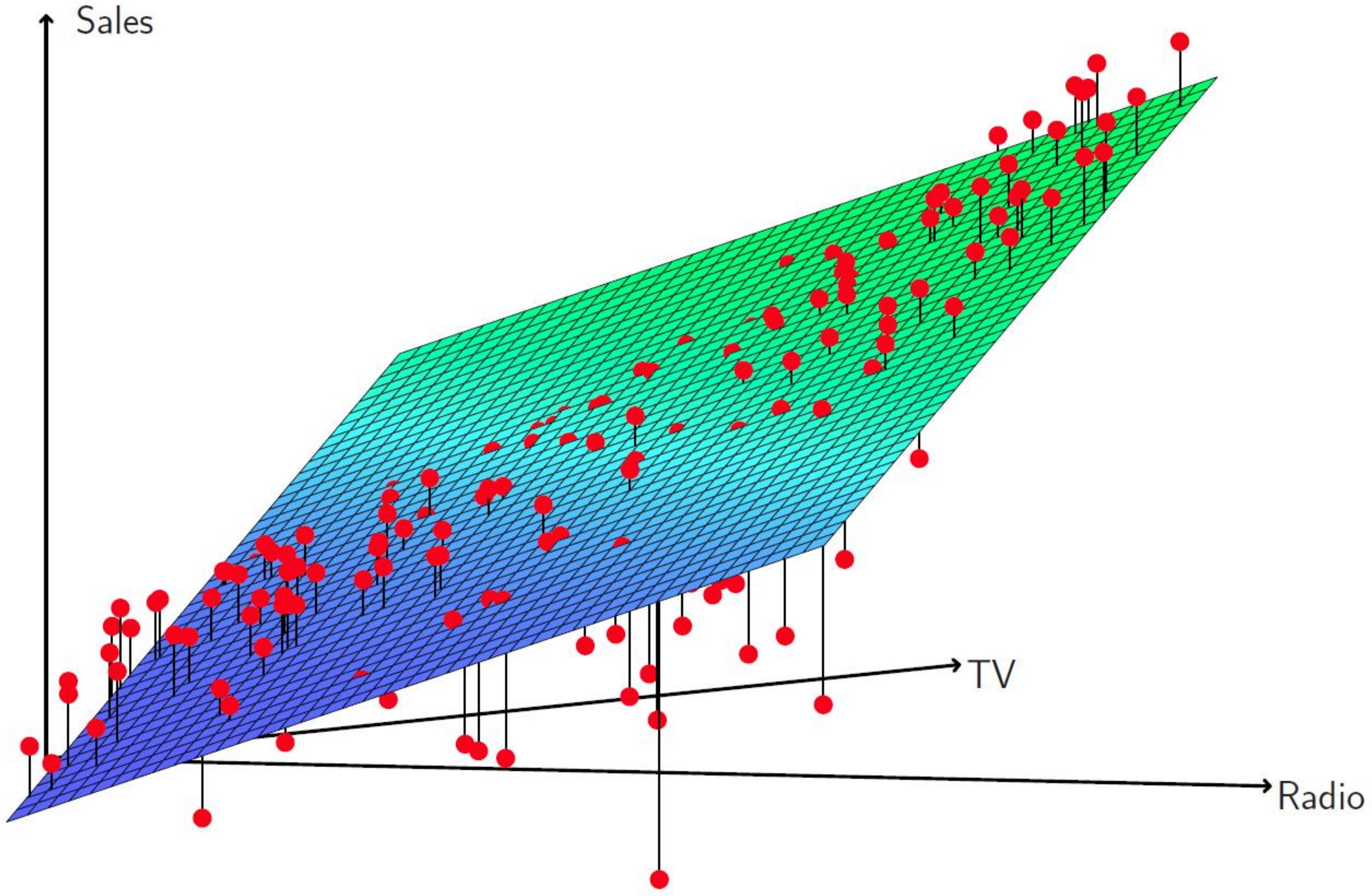
$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-p-1}}$$

- R^2 statistic

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

Exercises

- Show that $R^2 = Corr^2(Y, \hat{Y})$, and that the least squares fit maximizes this correlation among all possible linear models
- Show that R^2 always increase when more variables are added to the model



For the Advertising data, a linear regression fit to sales using TV and radio as predictors

Predictions

- Population regression plane

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- Least squares plane

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

- The trade-off

$$E\{y_0 - \hat{f}(x_0)\}^2 = Var\{\hat{f}(x_0)\} + [Bias\{\hat{f}(x_0)\}]^2 + Var(\epsilon)$$
$$= Variance(x_0) \ + Bias^2(x_0) + Irreducible\ Error$$

- How close will $\hat{Y}$ be to $f(X)$?
  - Confidence intervals
- How much will Y vary from $\hat{Y}$?
  - Prediction intervals

Important predictors

- Variable selection
  - The task of determining which predictors are associated with the response
- How do we determine which model is best?
  - A total of 2^p models that contain subsets of p variables

Linear regression versus KNN

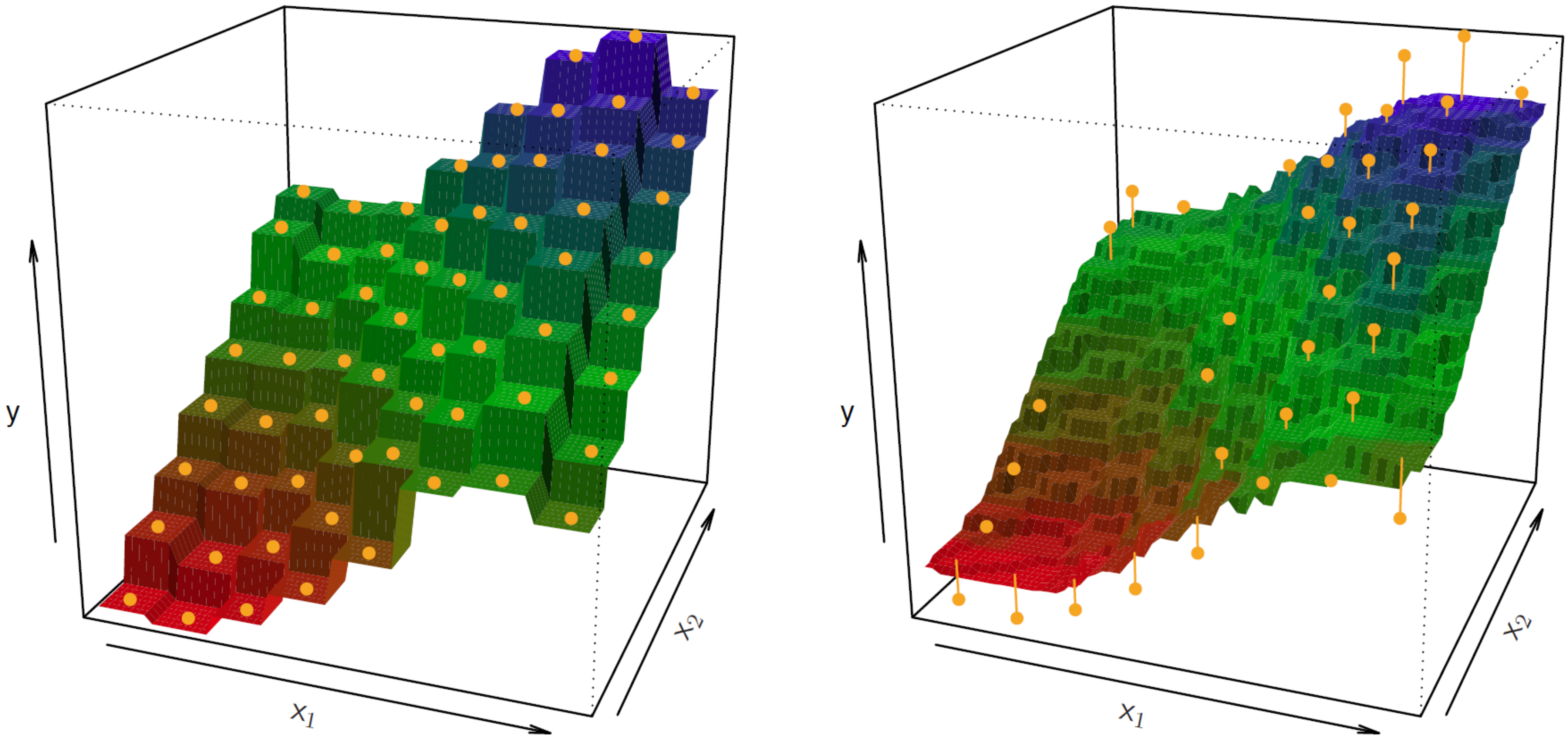
K-nearest neighbors regression

① Specify a positive integer K

② Identify the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$

③ Estimate $f(x_0)$ by the average of the responses in $\mathcal{N}_0$

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

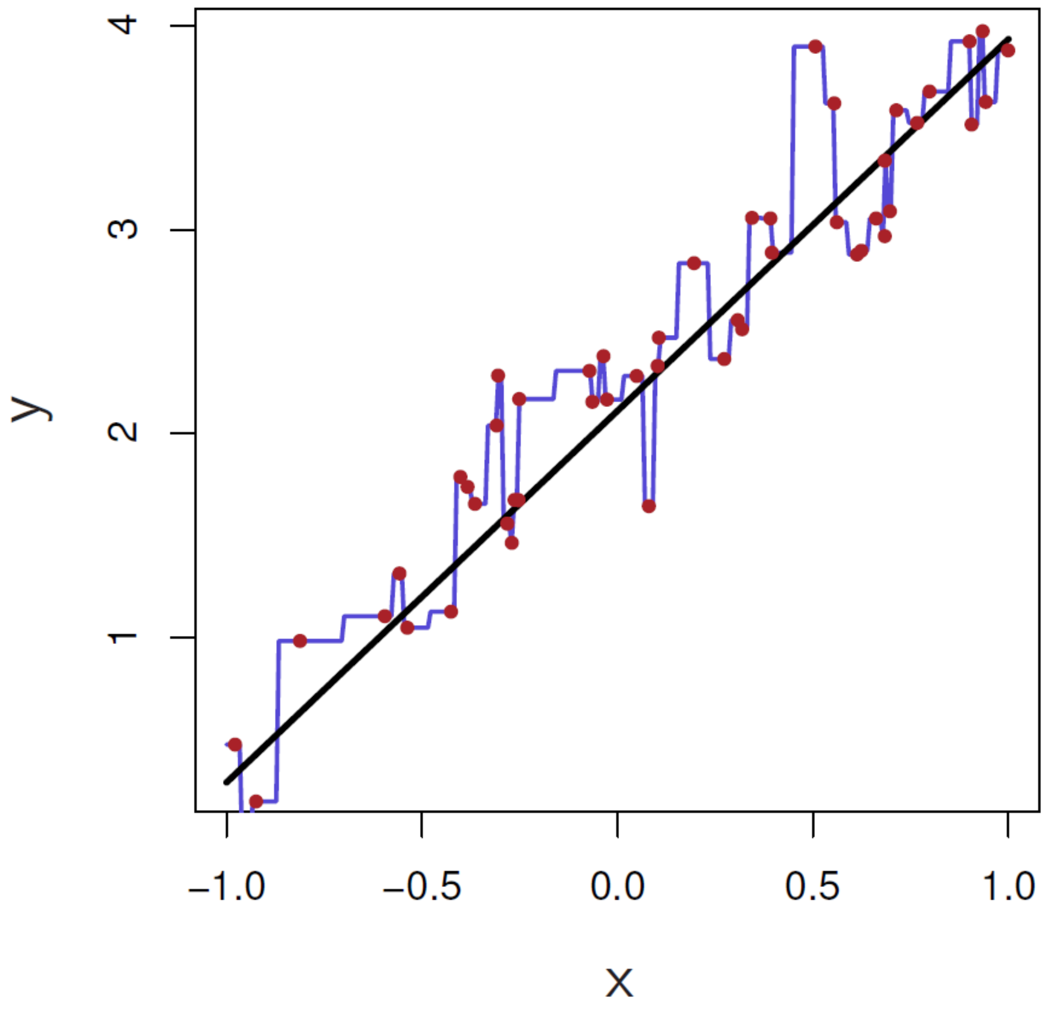


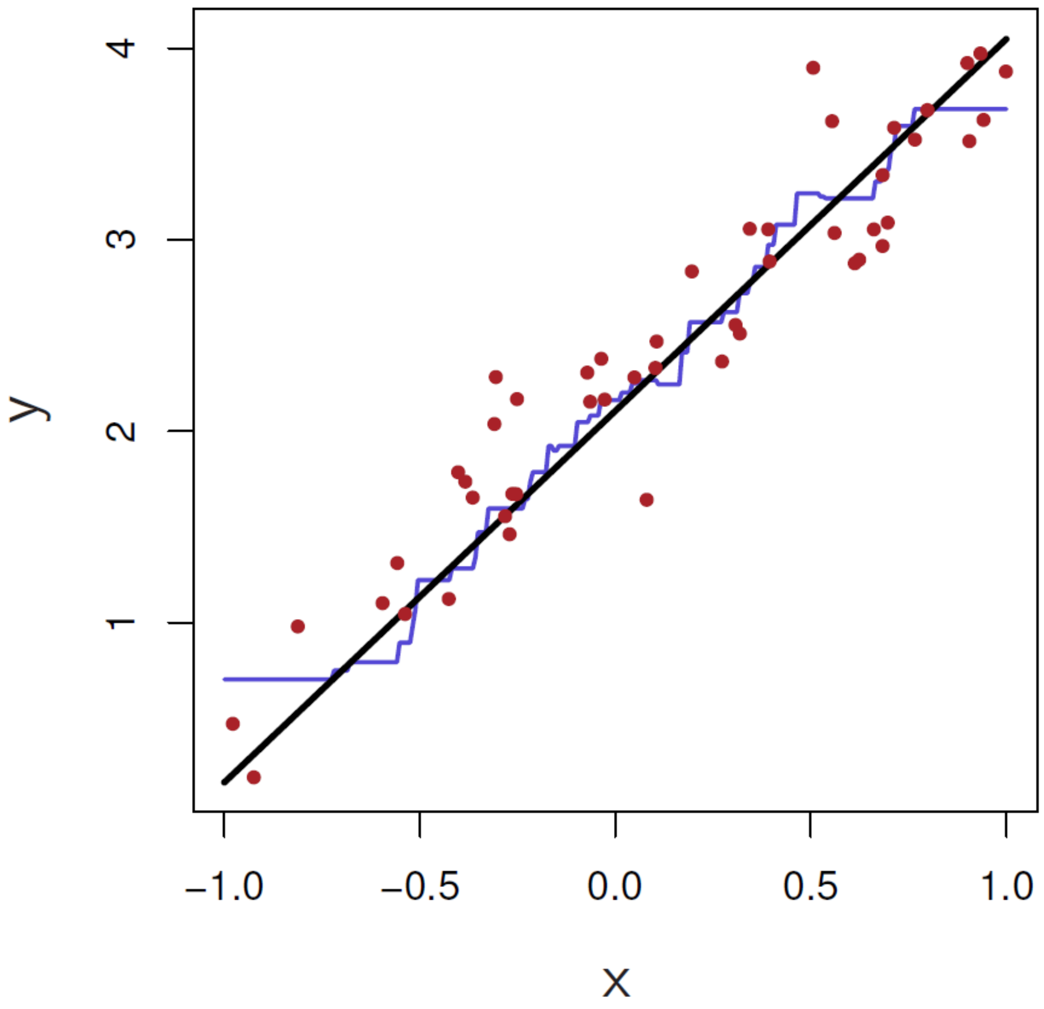
Two KNN fits on a two-dimensional data set with 64 observations.
Left: $K = 1$. Right: $K = 9$

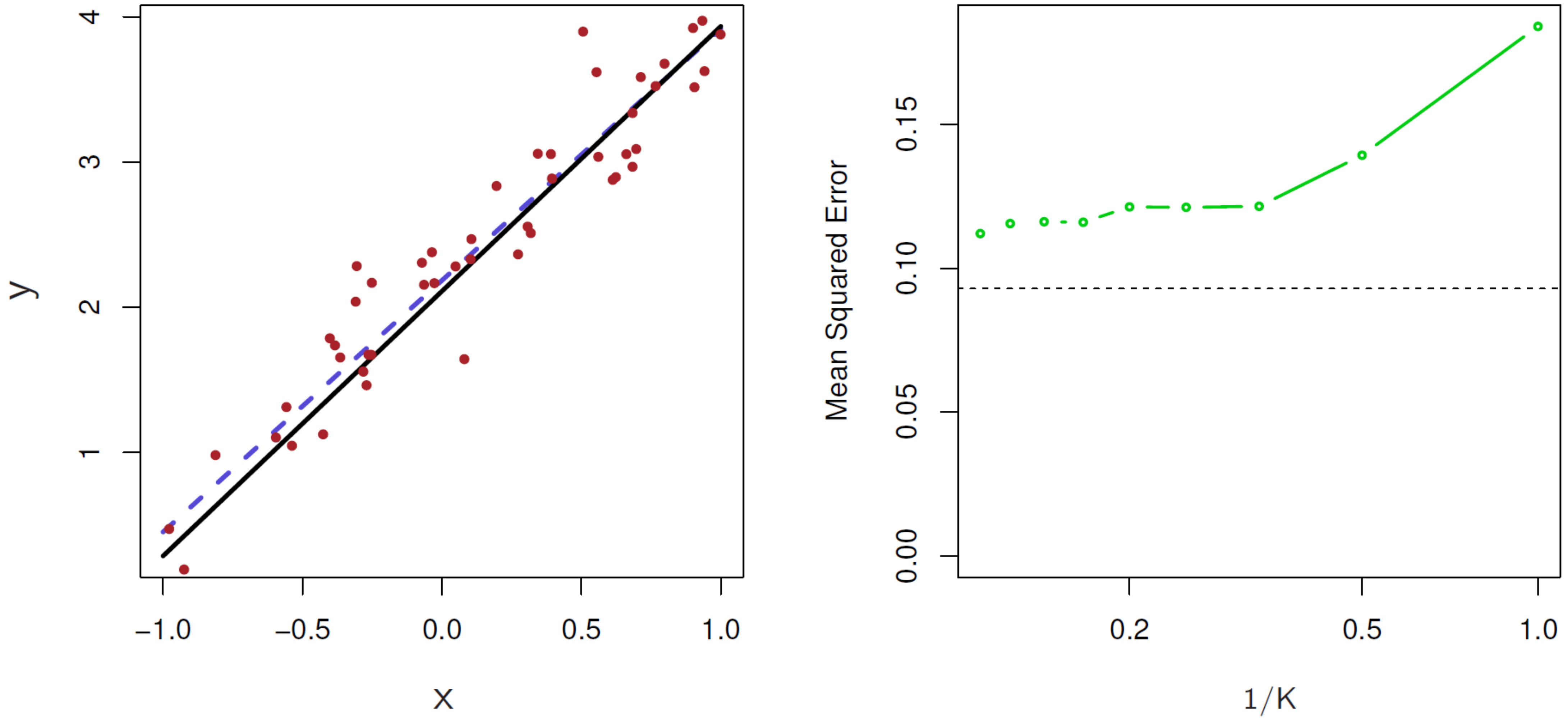
Comparison of linear regression with KNN

- Case 1: $p = 1$ and $f(x)$ is linear
- Case 2: $p = 1$ and $f(x)$ is slightly non-linear
- Case 3: $p = 1$ and $f(x)$ is strongly non-linear
- Case 4: $p = \{1,2,3,4,10,20\}$ and $f(x)$ is strongly non-linear

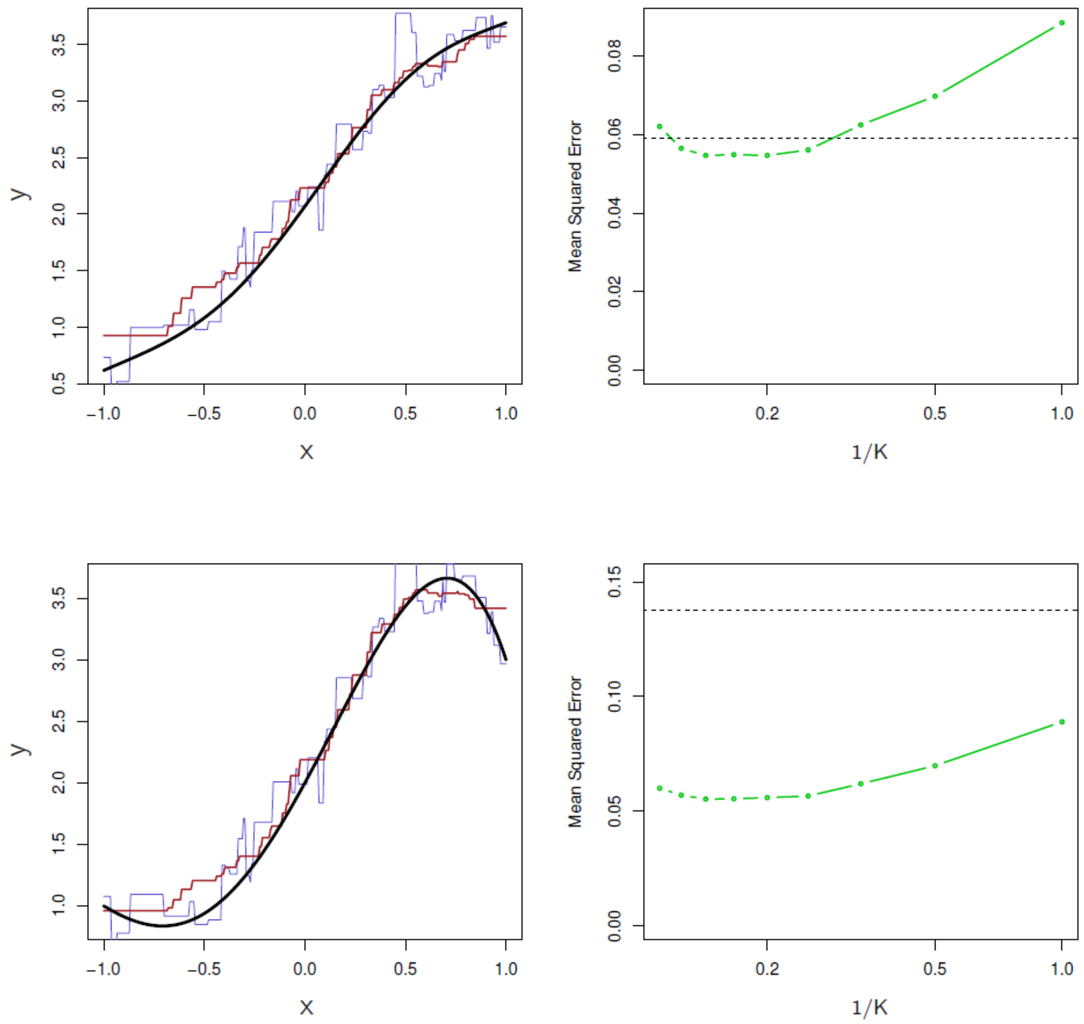
$p = 1$ and $f(x)$ is linear






y
1
2
3
4
−1.0
−0.5
0.0
0.5
1.0
x
Mean Squared Error
0.00
0.05
0.10
0.15
0.2
0.5
1.0
1/K

$f(x)$ is slightly or strongly non-linear



p varies and $f(x)$ strongly non-linear