Linear Regression B

Yuan Yao

Department of Mathematics
Hong Kong University of Science and Technology

Chapter 3

Spring, 2018

Outline

Example: Advertising data

The data contains 200 observations.

- Sample size: $n = 200$.
- Number of parameters: $p = 3$.
- Sales as responses: y_i, $i = 1..., n$.
- Covariates:
  1. TV (bugdets): $x_{i1}, i = 1, ..., n$.
  2. Radio (budgets): $x_{i2}, i = 1, ..., n$.
  3. Newspaper (budgets): $x_{i3}, i = 1, ..., n$.

Advertising data

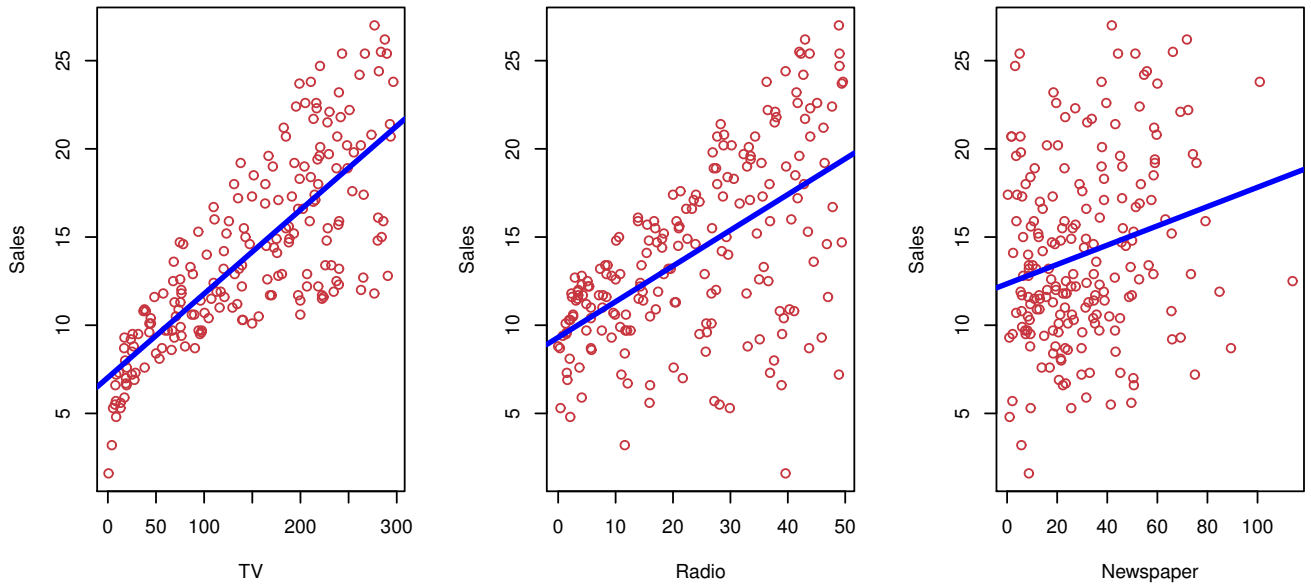

Figure 1: The **Advertising** data set. The plot displays *sales*, in thousands of units, as a function of *TV*, *radio*, and *newspaper* budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively.

Here are a few important questions that we might seek to address:

- 1. Is there a relationship between advertising budget and sales?
- 2. How strong is the relationship between advertising budget and sales?
- 3. Which media contribute to sales?
- 4. How accurately can we estimate the effect of each medium on sales?
- 5. How accurately can we predict future sales?
- 6. Is the relationship linear?
- 7. Is there synergy among the advertising media?

It turns out that linear regression can be used to answer each of these questions. We will first discuss all of these questions in a general context, and then return to them later.

Outline

Illustrating python linear regression

In [11]:

```
est = smf.ols('Sales ~ TV', advertising).fit()
est.summary().tables[1]
```

Out[11]:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

Figure 2: What are 't(-values)', 'P(-values)' and so on? Reference:

https://github.com/JWarmenhoven/ISLR-python/blob/master/Notebooks/Chapter%203.ipynb

Simple linear regression

- It is a very straightforward simple 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. Mathematically, we can write this linear relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

  e.g., sales $\approx \beta_0 + \beta_1 \times$ TV.
- Here β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, β_0 and β_1 are intercept slope known as the model coefficients or parameters.
- Once we have used our coefficient parameter training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising by computing

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2}$$

  where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. Here we use a hat symbol, $\hat{}$, to denote the **estimated value** for an unknown parameter or coefficient, or to denote the **predicted value** of the response.

Estimating the Coefficients

- The most common approach involves minimizing the **least squares** criterion. Alternative approaches will be discussed later.
- The least squares coefficient estimates are given as

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{3}$$

  where $\bar{y} \equiv \frac{1}{n}\sum_i^n y_i$ and $\bar{x} \equiv \sum_i^n x_i$.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i-th value of X. Then $e_i = y_i - \hat{y}_i$ represents the i-th residual– this is the difference between residual the i-th observed response value and the i-th response value that is predicted by our linear model. We define the residual sum of squares (RSS) as

$$\begin{aligned} RSS &= e_1^2 + e_2^2 + \cdots + e_n^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \end{aligned} \tag{4}$$

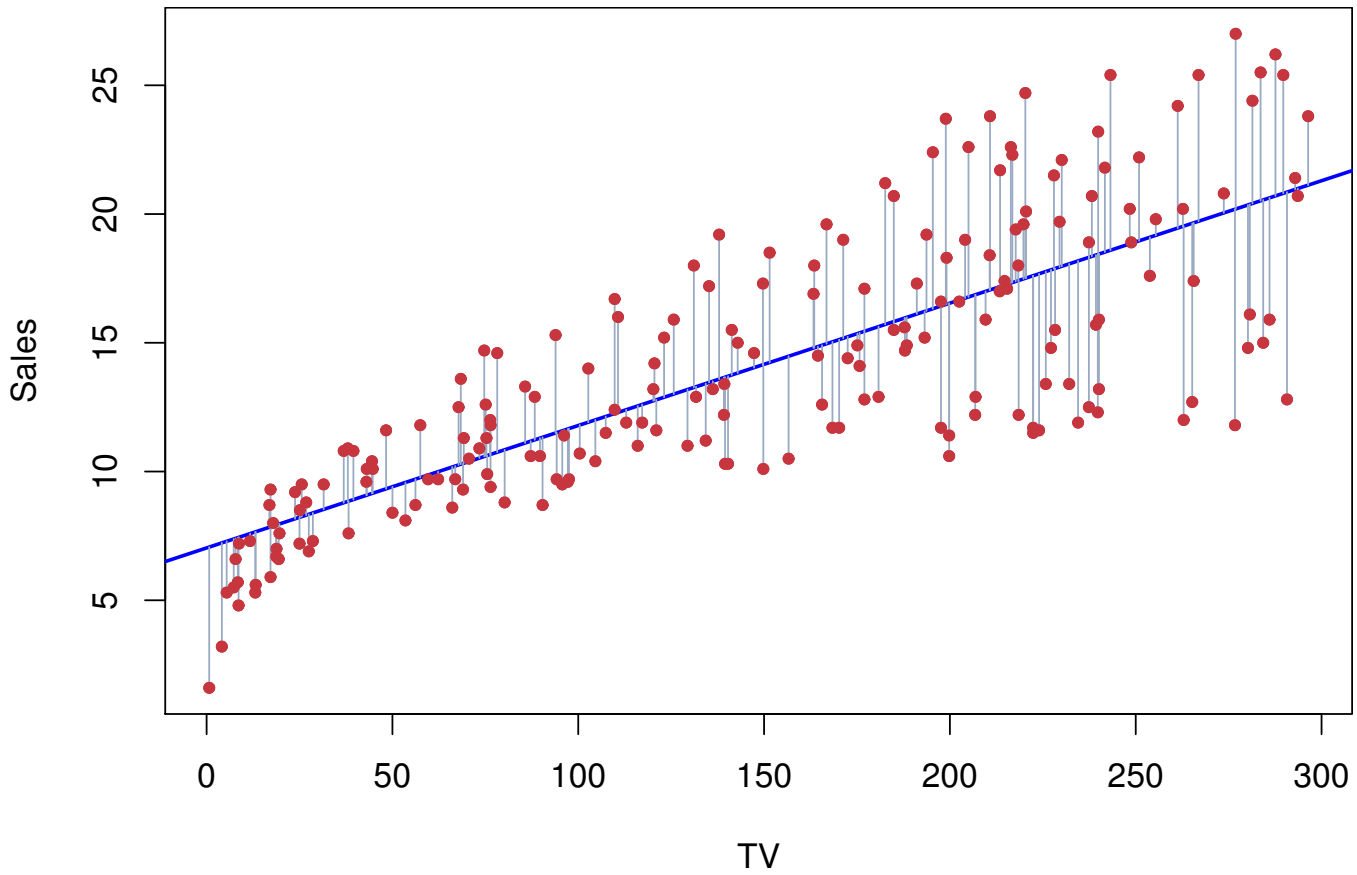

Figure 3: For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit ($\hat{\beta}_0 = 7.03$, $\hat{\beta}_1 = 0.0475$) is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

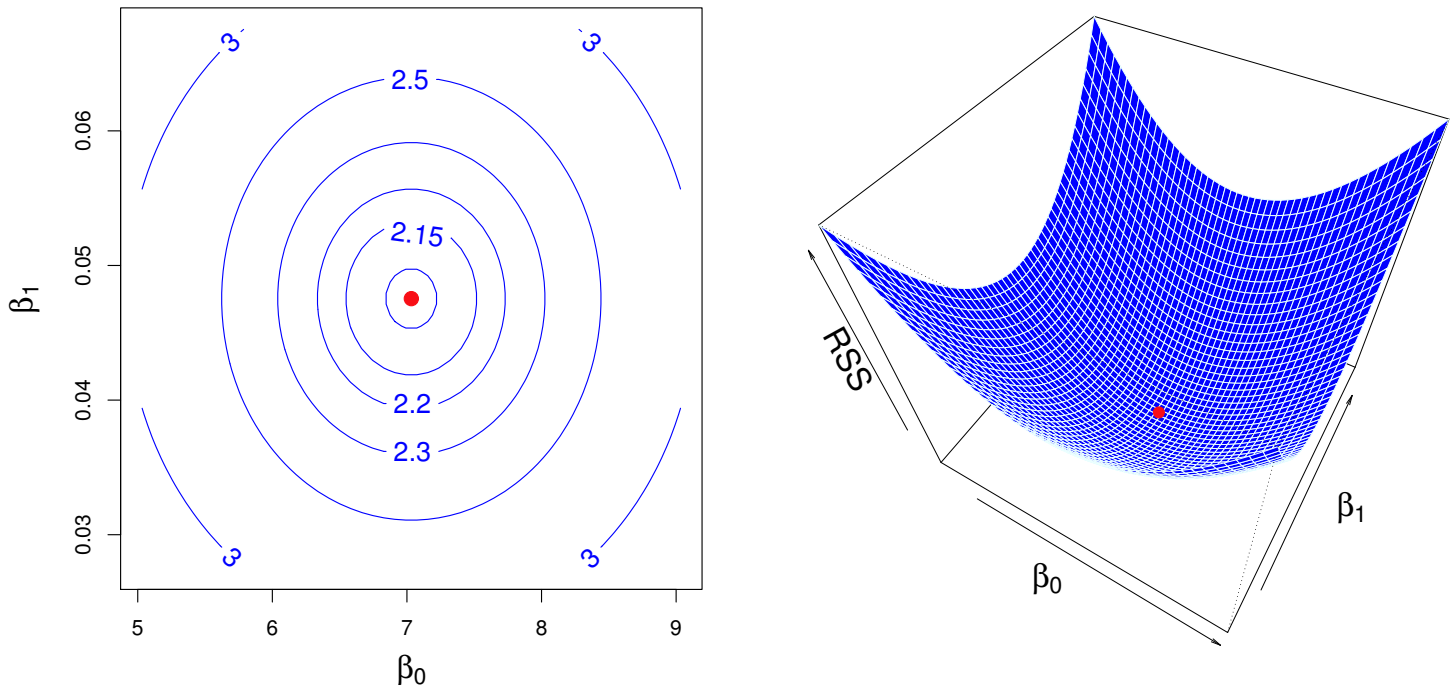

Figure 4: Contour and three-dimensional plots of the RSS on the Advertising data, using sales as the response and TV as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

Confidence Intervals of Estimated Parameters

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{s\sqrt{1/\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

where

$$s^2 = \mathrm{RSS}/(n-2)$$

is an unbiased estimator of the variance of the error.

Hypothesis testing

- The most common hypothesis test involves testing the null hypothesis of

$$\mathcal{H}_0 : \beta_1 = 0, \tag{5}$$

  i.e., there is no relationship between X and Y.
- We compute a t-statistics

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}, \tag{6}$$

  which will have a t-distribution with $n - 2$ degrees of freedom.
- Consequently, it is a simple matter to compute the probability of observing any value equal to $|t|$ or larger, assuming $\beta_1 = 0$. We call this probability the *p-value*. We reject the null hypothesis if the p-value is small enough, e.g., <0.05 or <0.01.

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

Table 1: For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units.

Assessing the accuracy of the model

- The residual standard error (RSE):

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum^{n}(y_i - \hat{y}_i)^2}. \tag{7}$$

  It provides an absolute measure of lack of fit of the linear model.
- **R^2 statsitic** 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.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \tag{8}$$

  where $TSS = \sum(y_i - \bar{y})^2$ is the *total sum of squares*.
- Correlation between X and Y is define as

$$r = Cor(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}. \tag{9}$$

  In fact, it can be shown that $R^2 = r^2$ in linear regression.

Outline

Model

- Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors. In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{10}$$

  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, *holding all other predictors fixed*.
- In the advertising example, (18) becomes

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon. \tag{11}$$

Illustrating python linear regression

In [17]:

```
est = smf.ols('Sales ~ TV + Radio + Newspaper', advertising).fit()
est.summary()
```

Out[17]:

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Tue, 09 Jan 2018	**Prob (F-statistic):**	1.58e-96
Time:	23:14:15	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Estimating the regression coefficients

- Let $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p]$ be the least squares

$$\hat{\boldsymbol{\beta}} = \arg \min_{\tilde{\boldsymbol{\beta}}=[\tilde{\beta}_0,\tilde{\beta}_1,\ldots,\tilde{\beta}_p]} \sum_{i}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_p x_{ip})^2 \qquad (12)$$

- Note that we use $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ to denote the population parameter, the estimate and the optimization variable.
- The least square solution is

$$\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}. \qquad (13)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & x_{np} \end{bmatrix} \qquad (14)$$

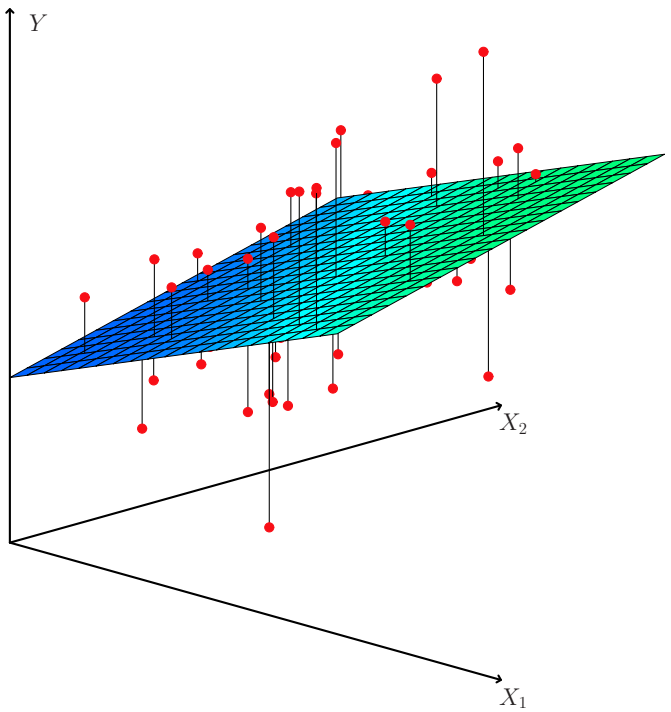

Figure 5: In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

Some important Questions

- 1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- 2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- 3. How well does the model fit the data?
- 4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

One. Is There a Relationship Between the Response and Predictors?

- In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{15}$$

  versus the alternative H_a: at least one β_j is non-zero.
- This hypothesis test is performed by computing the F-statistics,

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}, \tag{16}$$

  where $TSS = \sum_i (y_i - \bar{y})^2$ and $RSS = \sum_i (y_i - \hat{y}_i)^2$.
- Sometimes we want to test that a particular subset of q of the coefficients are zero. This corresponds to a null hypothesis

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0, \tag{17}$$

  Then the appropriate F-statistic is

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}, \tag{18}$$

Two: Deciding on Importance Variables

- If we conclude on the basis of that p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!
- We could look at the individual p-values as in Table 2, but as discussed, if p is large we are likely to make some false discoveries.

	Coefficient	Std.error	t-statistics	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 2: For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

- It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only related to a subset of the predictors.
- The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as **variable selection**.
- Ideally, we would like to perform variable selection by trying out a lot of different models, each containing a different subset of the predictors. For instance, 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 both X_1 and X_2.
- We can then select the best model out of all of the models that we have considered. How do we determine which model is best? Various statistics can be used to judge the quality of a model. These include Mallow's Cp, Akaike information criterion (AIC), Bayesian information criterion (BIC) and adjusted R^2.

- Unfortunately, there are a total of 2^p models that contain subsets of p variables. This means that even for moderate p, trying out every possible subset of the predictors is infeasible. For instance, we saw that if $p = 2$, then there are $2^2 = 4$ models to consider. But if $p = 30$, then we must consider $2^{30} = 1,073,741,824$ models! This is not practical.
- There are three classical approaches for this task:
  - Forward selection
  - Backward selection
  - Mixed section
- Many other methods appeared in recent years, such as Lasso.

Three: Model Fit

- Recall that in simple regression, R^2 is the square of the correlation of the response and the variable. In multiple linear regression, it turns out that it equals $Cor(Y, \hat{Y})^2$, the square of the correlation between the response and the fitted linear model.
- In fact one property of the fitted linear model is that it maximizes this correlation among all possible linear models.
- An R^2 value close to 1 indicates that the model explains a large portion of the variance in the response variable.
- Take Advertising data as an example. The model that uses all three advertising media to predict sales has an R^2 of 0.8972. On the other hand, the model that uses only TV and radio to predict sales has an R^2 value of 0.89719 (see Table 2).
- Note that R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. This is due to the fact that adding another variable to the least squares equations must allow us to fit the training data (though not necessarily the testing data) more accurately.

Visualization for Model Checking

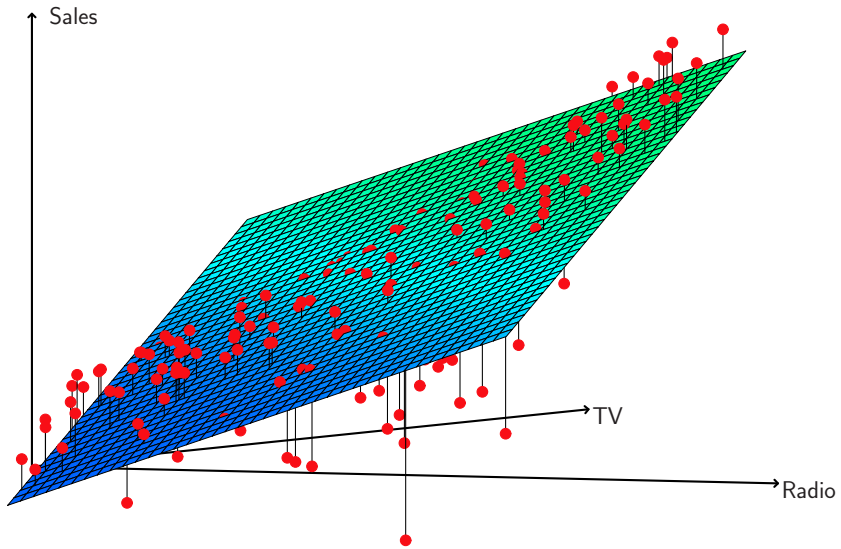

Figure 6: For the Advertising data, a linear regression fit to sales using TV and radio as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals (those visible above the surface), tend to lie along the 45-degree line, where TV and Radio budgets are split evenly. The negative residuals (most not visible), tend to lie away from this line, where budgets are more lopsided.

Four: Prediction interval

- To predict the actual response y, rather than its mean, we would use the same point estimator $\hat{\beta}^T \mathbf{x}$, but the accuracy is much decreased as more uncertainty in the randomness of the actual response from the error is involved.
- The confidence interval, often called prediction interval, for y is

$$\hat{\beta}^T \mathbf{x} \pm t_{n-p-1}(\alpha/2) s \sqrt{1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}.$$

Extensions of the Linear Model

For linear models, two of the most important assumptions state that the relationship between the predictors and response are *additive* and *linear*.

- The additive assumption means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors, i.e.,

$$Y = \sum_j f_j(X_j) + \epsilon.$$

  The "additive" assumption is also known as the "separable" condition.

- The linear assumption states that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j.
- For example, $Y = 2 + 3X_1 - 4sin(X_1) + X_2$ is additive but non-linear in X_1.

Removing the Additive Assumption

- Consider the standard linear regression model with two variables,

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

- One way of extending this model to allow for interaction effects is to include a third predictor, which is constructed by computing the product of X_1 and X_2. This results in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon. \tag{19}$$

- When $\beta_3 \neq 0$, the effect of changes in a predictor X_1 on the response Y depends on X_2, and vice versa.
- This is known as a *synergy* effect or an *interaction* effect.
- The parameters $\beta_1, \beta_2, \beta_3$ can be estimated using least squares.

Non-linear Relationships

- Consider the model with a quadratic term,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon.$$

- Although Y is non-linear in X, the model $f(X; \beta) = \beta_0 + \beta_1 X + \beta_2 X^2$ is **linear in parameters** $[\beta_0, \beta_1, \beta_2]$.

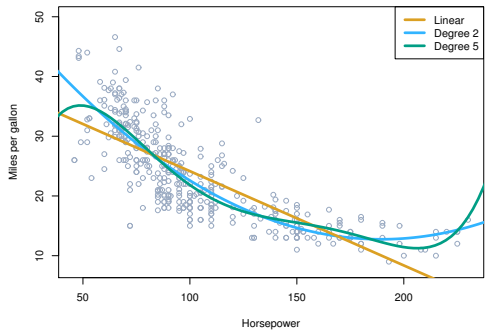

Figure 7: The Auto data set. For a number of cars, *mpg* and *horsepower* are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes horsepower^2 is shown as a blue curve. The linear regression fit for a model that includes all polynomials of horsepower up to fifth-degree is shown in green.

Potential problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

- 1. Non-linearity of the response-predictor relationships.
- 2. Correlation of error terms.
- 3. Non-constant variance of error terms.
- 4. Outliers.
- 5. High-leverage points.
- 6. Collinearity.

In practice, identifying and overcoming these problems is as much an art as a science. Here we provide only a brief summary of some key points.

1. Non-linearity of the Data

- **Residual plots** are a useful graphical tool for identifying non-linearity. We may plot the residuals versus the predicted (or fitted) values $\hat{y}_i$. We may add nonlinear terms such as $log(X)$, $\sqrt{X}$ and X^2 if necessary.

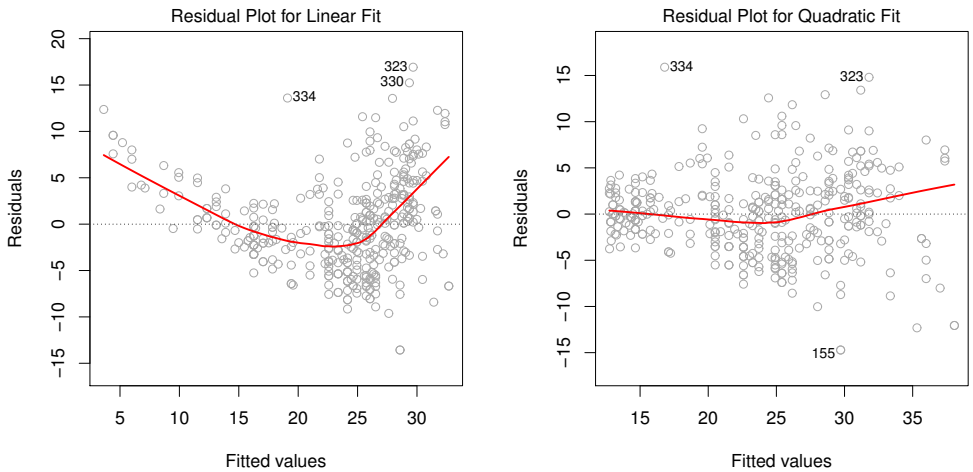

Figure 8: Plots of residuals versus predicted (or fitted) values for the Auto data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear regression of *mpg* on *horsepower*. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of *mpg* on *horsepower* and *horsepower*^{2}. There is little pattern in the residuals.

2. Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \ldots, \epsilon_n$, are uncorrelated. The standard errors that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrelated error terms.
- If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be. For example, a 95% confidence interval may in reality have a much lower probability than 0.95 of containing the true value of the parameter.
- In addition, p-values associated with the model will be lower than they should be; this could cause us to erroneously conclude that a parameter is statistically significant.
- In short, if the error terms are correlated, we may have an unwarranted sense of confidence in our model.
- As an extreme example, suppose we accidentally doubled our data, leading to observations and error terms identical in pairs.

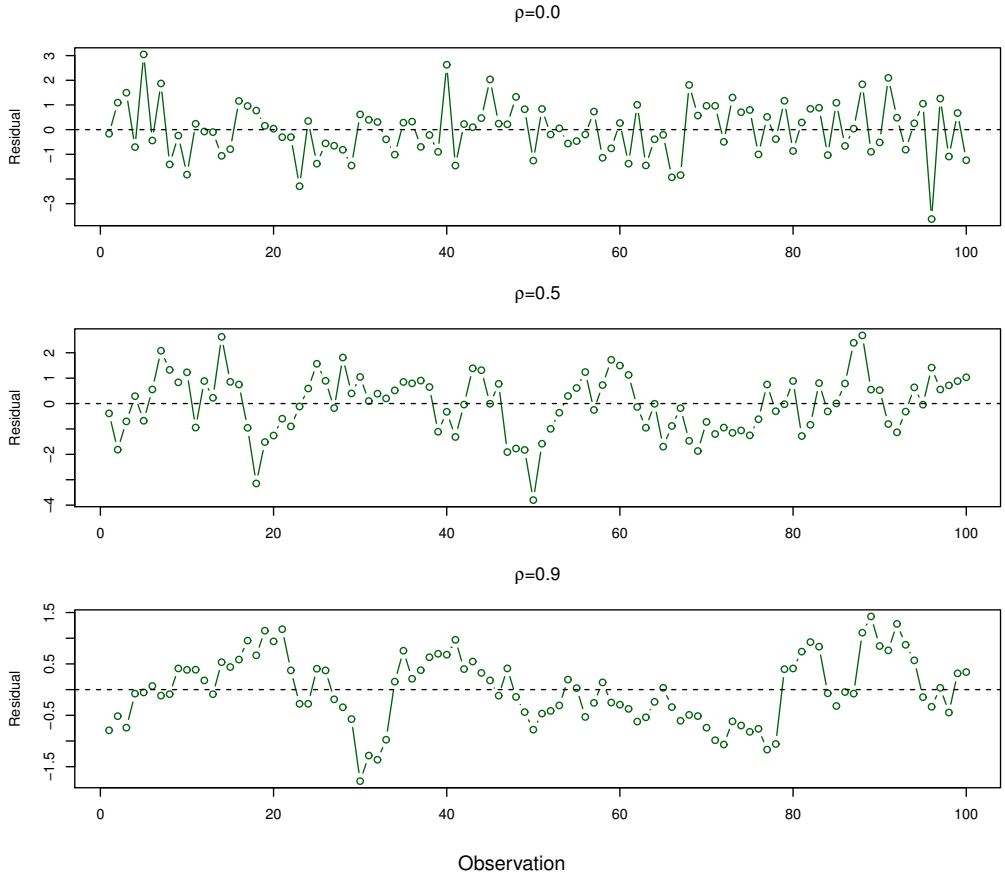

Figure 9: Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

3. Non-constant Variance of Error Terms

- An important assumption of the linear regression model is $Var(\epsilon_i) = \sigma^2$. Non-constant variances in the errors, known as **heteroscedasticity**, can be seen from the presence of a *funnel shape* in the residual plot.

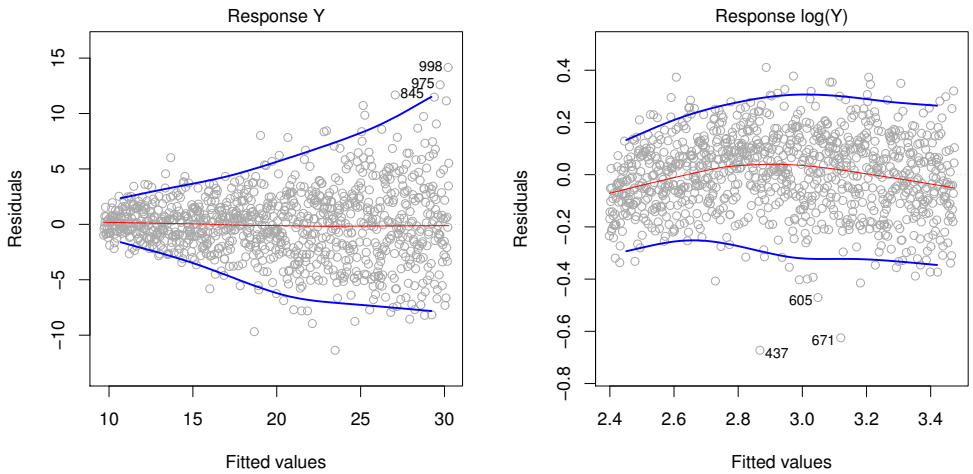

Figure 10: Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates **heteroscedasticity**. Right: The predictor has been **log-transformed**, and there is now no evidence of heteroscedasticity.

- When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.
- **Weighted least squares**: Sometimes we have a good idea of the variance of each response. For example, the ith response could be an average of n_i raw observations. If each of these raw observations is uncorrelated with variance σ^2, then their average has variance $\sigma_i^2 = \sigma^2/n_i$. In this case a simple remedy is to fit our variances-i.e. $w_i = n_i$ in this case.

4. Outliers

- An *outlier* is a point for which y_i is far from the value predicted by the outlier model. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection.

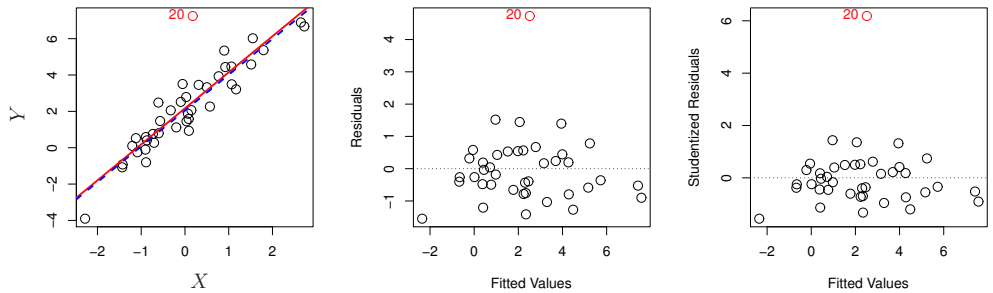

Figure 11: Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized residual of 6; typically we expect values between -3 and 3. The studentized residual can be computed by dividing each residual e_i by its estimated standard studentized residual error. Many other methods in **robust** linear regression are available to handle outliers.

5. High leverage Points

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i. In contrast, observations with high leverage have an unusual value for x_i.

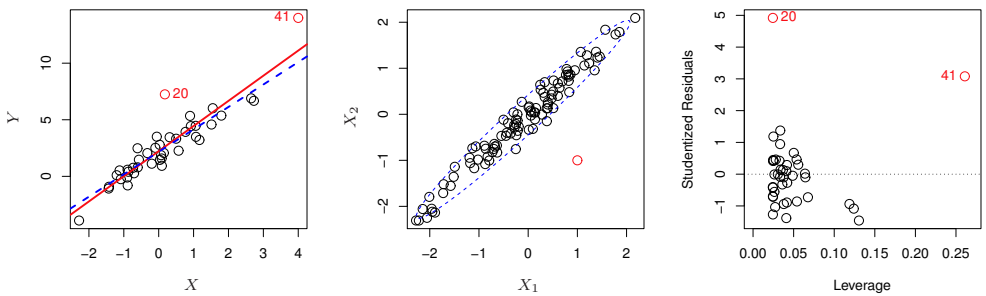

Figure 12: Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage. Right: Observation 41 has a high leverage and a high residual.

6. Collinearity: Credit Data

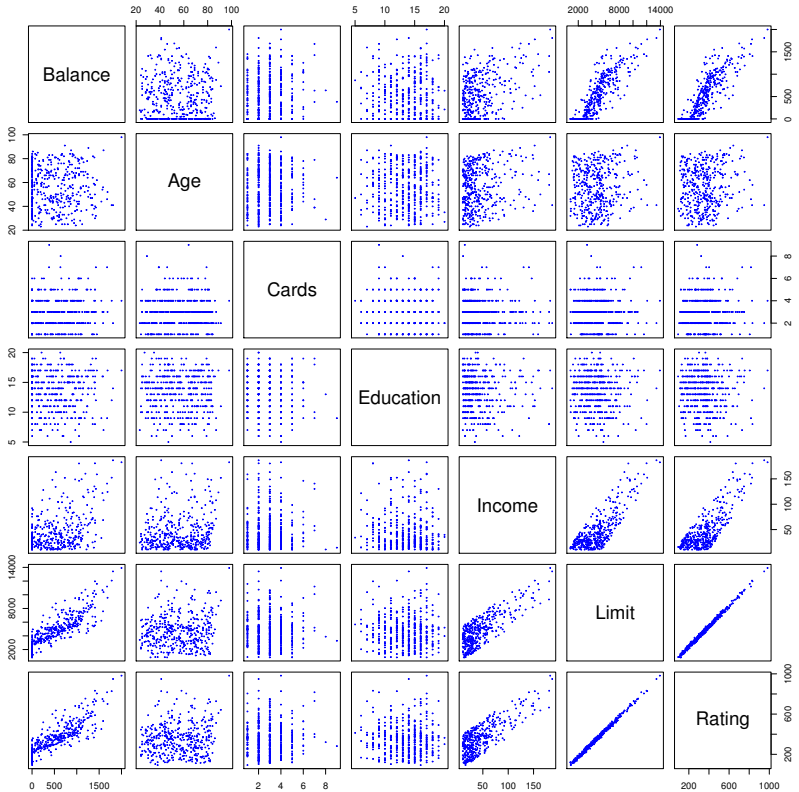

Figure 13: The *Credit* data set contains information about *balance*, *age*, *cards*, *education*, *income*, *limit*,and *rating* for a number of potential customers.

6. Collinearity

- Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another.

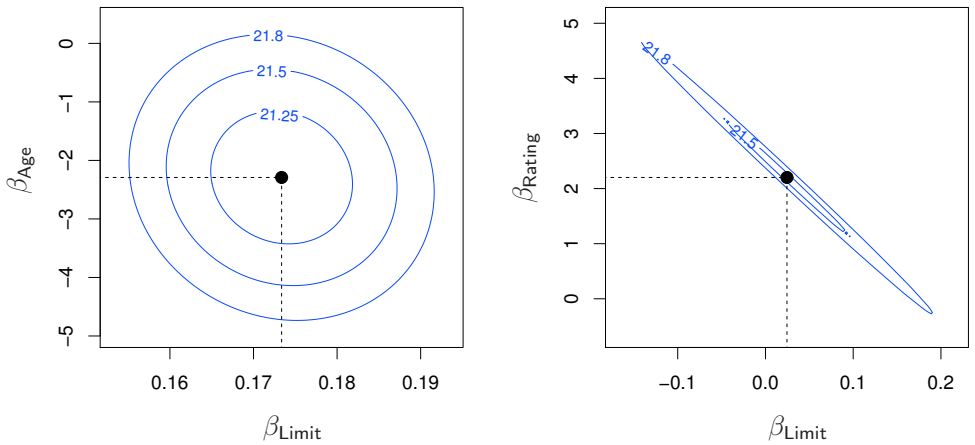

Figure 14: Contour plots for the RSS values as a function of the parameters β for various regressions involving the *Credit* data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS. Left: A contour plot of RSS for the regression of balance onto *age* and *limit*.The minimum value is well defined. Right: A contour plot of RSS for the regression of balance onto *rating* and *limit*. Because of the collinearity, there are many pairs $(\beta_{Limit}, \beta_{Rating})$ with a similar value for RSS.

- Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow. Recall that the t-statistic for each predictor is calculated by dividing by its standard error. Consequently, collinearity results in a decline in the t-statistic.
- As a result, in the presence of collinearity, we may fail to reject $\mathcal{H}_0$: $\beta_j = 0$. This means that the power of the hypothesis test–the probability of correctly power detecting a non-zero coefficient–is reduced by 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

Table 3: The results for two multiple regression models involving the *Credit* data set are shown. Model 1 is a regression of *balance* on *age* and *limit*, and Model 2 a regression of *balance* on *rating* and *limit*. The standard error of β_{limit} increases 12-fold in the second regression, due to collinearity.

- A simple way to detect collinearity is to look at the correlation matrix of the predictors. An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data.
- Unfortunately, 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. We call this situation **multicollinearity**.
- Instead of inspecting the correlation matrix, a better way to assess multicollinearity is to compute the *variance inflation factor* (VIF).

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}, \tag{20}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

Outline

Linear Regression v.s. K-Nearest Neighbors

- Linear regression is an example of a parametric approach because it assumes a linear functional form for $f(X)$.
- In contrast, 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 one of the simplest and best-known non-parametric methods, *K-nearest neighbors 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 $\mathcal{N}_0$. In other words,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i. \tag{21}$$

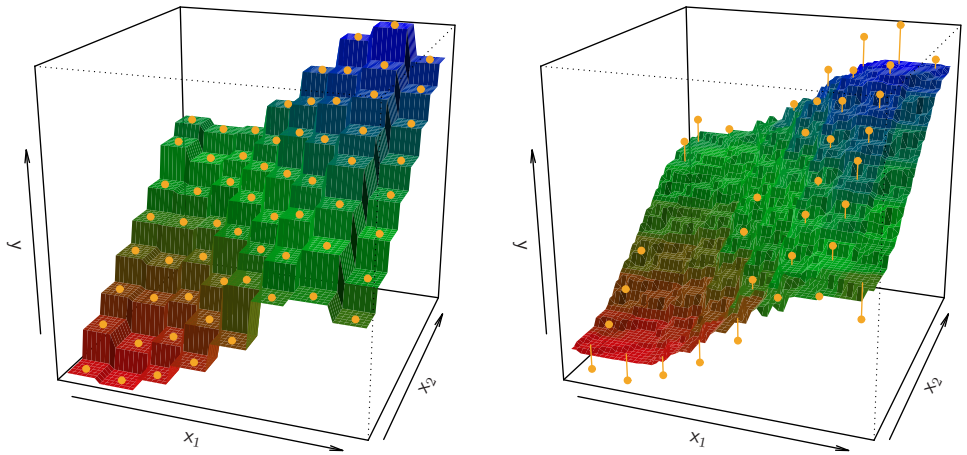

Figure 15: Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots). Left: $K = 1$ results in a rough step function fit. Right: $K = 9$ produces a much smoother fit.

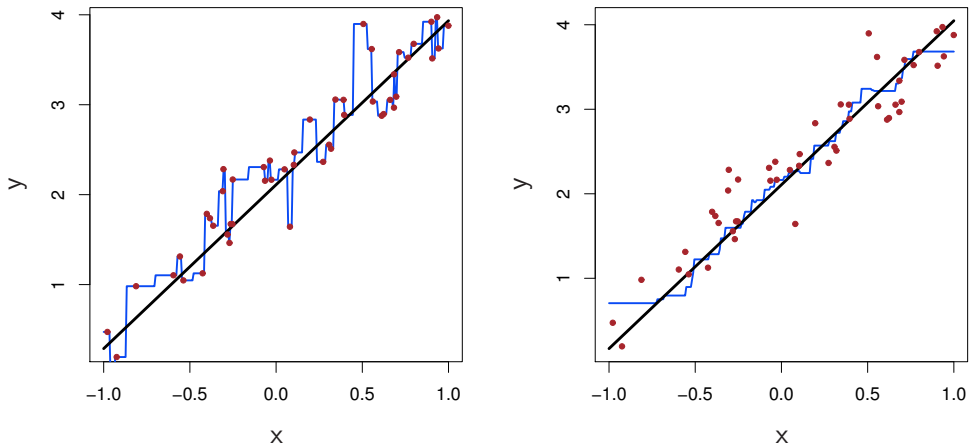

Figure 16: Plots of $\hat{f}(X)$ using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line. Left: The blue curve corresponds to $K = 1$ and interpolates (i.e. passes directly through) the training data. Right: The blue curve corresponds to $K = 9$,and represents a smoother fit.

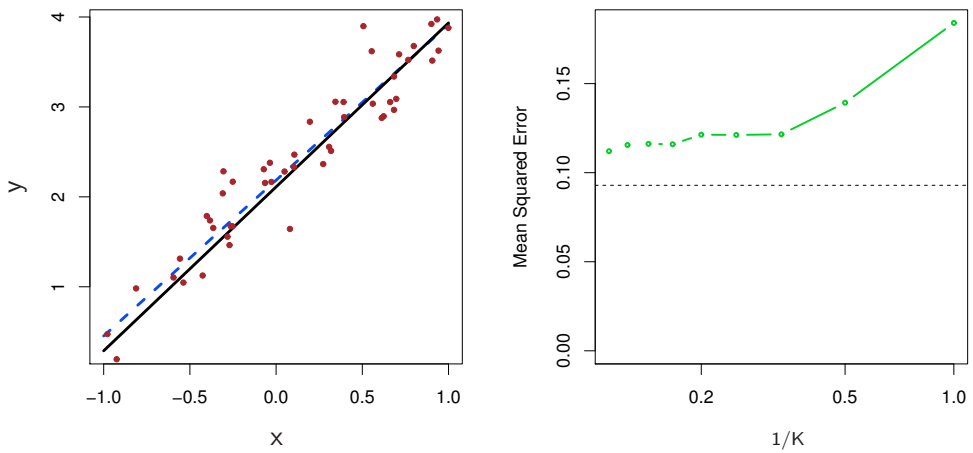

Figure 17: The same data set shown in Figure 16 is investigated further. Left: The blue dashed line is the least squares fit to the data. Since $f(X)$ is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of $f(X)$. Right: The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of $1/K$ (on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since $f(X)$ is in fact linear. For KNN regression, the best results occur with a very large value of K, corresponding to a small value of $1/K$.

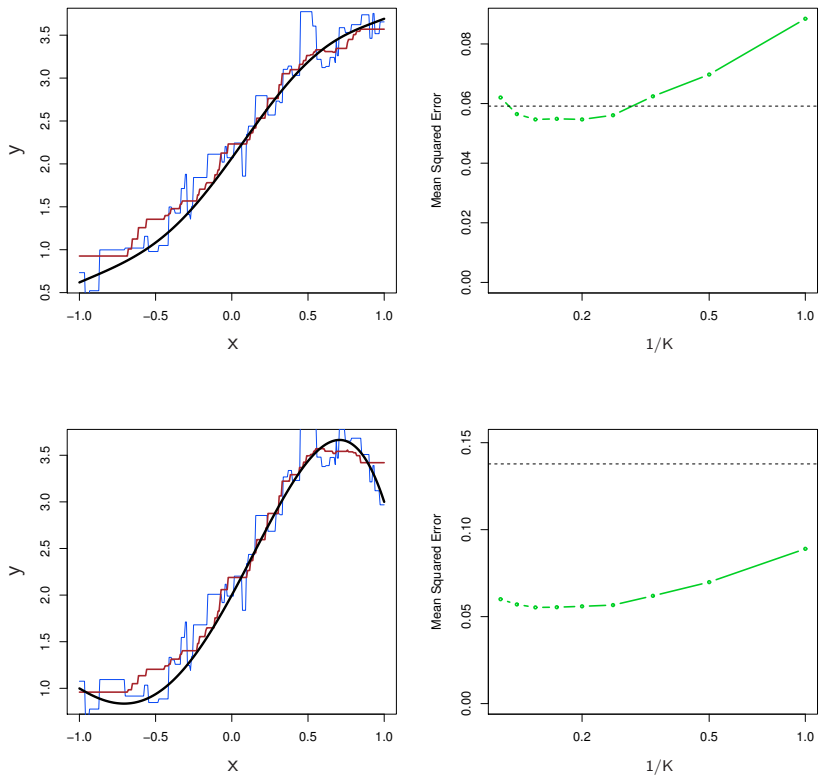

Figure 18: Top Left: In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with $K = 1$ (blue) and $K = 9$ (red) are displayed. Top Right: For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of $1/K$ (green) are displayed. Bottom Left and Bottom Right: As in the top panel, but with a strongly non-linear relationship between X and Y.

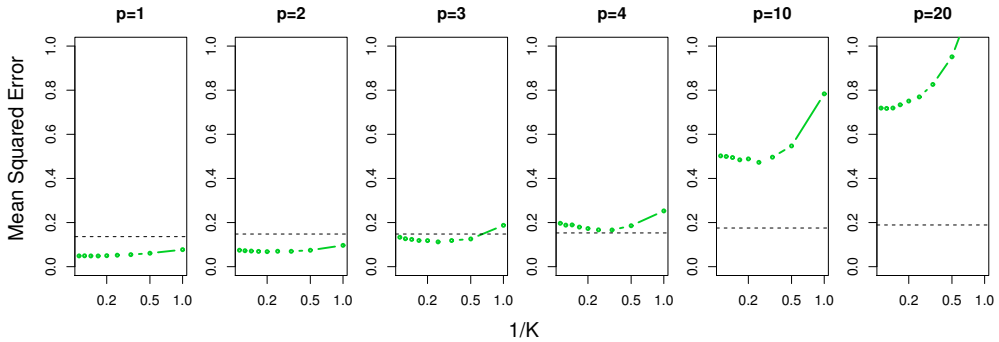

Figure 19: Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non-linear in the first variable, as in the lower panel in Figure 18, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.

Homework

- ISLR (Print 7) Chapter 3: 1; 2; 5; 8.