

Lecture 01. Introduction to Statistical Learning (Chapters 1-3)

Ping Yu

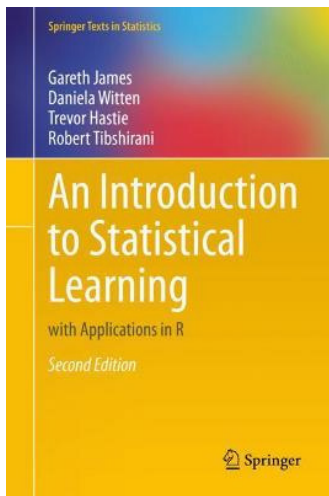
HKU Business School
The University of Hong Kong

Course Information

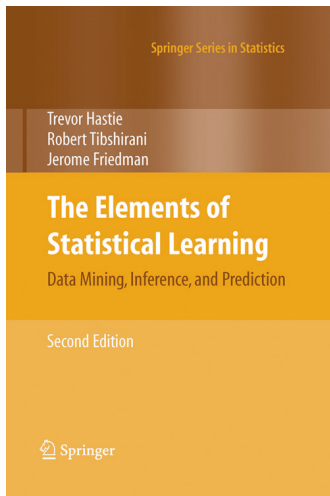
- **Instructor:** Yu, Ping
- **Email:** pingyu@hku.hk
- **Teaching Time:** 9:00-9:50, 10:00-10:50 and 11:00-11:50am + 12:00nn-12:30pm & 2:00-2:50, 3:00-3:50 and 4:00-4:50 + 5:00-5:30pm, Friday | Sunday
 - In both morning and afternoon, the first three sessions teach intuitions and details of all kinds of algorithms, and the fourth session teaches how to implement these algorithms by the software **R**.
- **Teaching Location:** Cyberport Classroom H | KKL101 | Virtual by Zoom (the lectures will be recorded and uploaded on moodle)
- **Office Hour:** 11:00-12:00am, Thursday, f2f at KKL1110 | Virtual by Zoom
 - I will **NOT** answer questions in email if the answer is long or is not easy to explain exactly by words. Please stop by during my office hour (please make an appointment beforehand) either physically or online (by the same zoom link as lectures).
- **Tutor:** Lily Wang
 - The tutor will teach how to implement the algorithms by the software.
 - **ANY** issues on administration (e.g., enrolment, moodle, time clash, software, etc.) and HWs (e.g., clarification of problems) should contact the tutor.

Our Textbook and Three References

- I will roughly follow the following textbook:
 1. [An Introduction to Statistical Learning with Applications in R \(ISLR hereafter\)](#), 2nd edition, by James, G., D. Witten, T. Hastie and R. Tibshirani, Springer, 2021.
 - For illustration of methods by empirical applications, we will concentrate on business and economic examples, and neglect biostatistical ones, e.g., gene expression data.
- A more advanced textbook is only occasionally mentioned:
 2. [The Elements of Statistical Learning \(ESL hereafter\)](#), 2nd edition, by Hastie, T., R. Tibshirani, and J. Friedman, Springer, 2009.
 - ISLR and ESL are freely available on moodle. [[figure here](#)]
- Two other useful reference books at the level of ISLR:
 3. [Data Science for Business \(DSB hereafter\)](#), by Provost, Foster and Tom Fawcett, O'Reilly, 2013.
 4. [Business Data Science \(BDS hereafter\)](#), by Matt Taddy, McGraw Hill, 2019.
 - The relevant chapters of DSB and BDS are available on moodle. [[figure here](#)]
- ISLR is the default textbook, so unless necessary, I will only indicate DSB, BDS and ESL (and suppress ISLR) when citing chapters or sections.



ISLR

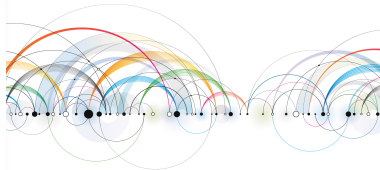


ESL

"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

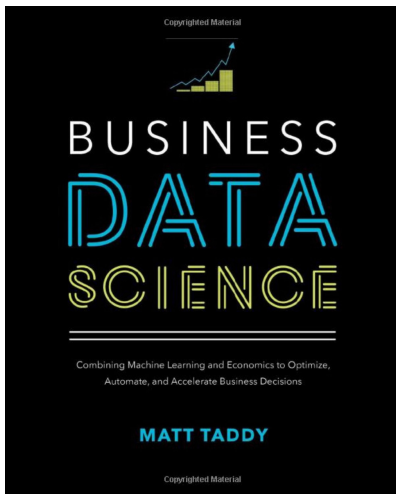
Data Science *for Business*

What You Need to Know
About Data Mining and
Data-Analytic Thinking



Foster Provost & Tom Fawcett

DSB



BDS

- We will use **R** and RStudio as the statistical software in this course; RStudio is an Integrated Development Environment (IDE) for **R**.





- Different from STATA or other softwares, both of them are free to download and install.
- **Website for R**: <https://www.r-project.org/>
- **Website for RStudio**: <https://www.rstudio.com/>
- **Wiki for R**: [https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))
- **Wiki for RStudio**: <https://en.wikipedia.org/wiki/RStudio>

Evaluation

- **Evaluation:** Four Assignments (20% each, posted already) and one Final Project (20%, not posted yet)
 - No in-class midterm or final.
- More Details on the **Assignments:**
 - Turn in your HW on moodle on the due day (usually before midnight, 11:59pm, of some Thursday).
 - Late HWs are not acceptable for whatever reasons. To avoid any risk, start your HW early (the HWs indicate clearly which problems can be solved after each lecture).
 - Each assignment contains both analytical and empirical problems; both types of problems are mainly from ISLR, and some analytical and empirical problems are written by the instructor.
 - You can form teams up to **five** members to solve assignments and the final project (each team need turn in only one copy of solution); the team members for each of the four assignments and the project need not be the same or from the same session, i.e., each student can join at most five teams with team members from both sessions.
 - All documents turned in must be typed (e.g., by LaTeX or Word).
 - For empirical problems, besides the final answers, **R** codes should also be submitted.

- More Details on the [Project](#):
 - Two datasets (\mathcal{D}_1 has a quantitative response and \mathcal{D}_2 has a qualitative response) are provided with some responses randomly deleted, and you can choose **one and only one** dataset for prediction.
 - You can use any techniques **in or out of** this course in your prediction.
 - For **each** dataset, the scores are **uniformly** distributed between 60 and 100.
 - If only one group solves, say \mathcal{D}_1 , then this group will automatically get a full score.
 - If two groups, then the group with better predictions gets a full score and the other group gets 60.
 - If three groups, then the scores would be 60, 80, 100.¹
 - **R** codes should be submitted; especially, the predictions should be submitted as a separate data file.

¹You will learn what "quantitative", "qualitative", "response", "better prediction" mean in the coming lectures.  

The Philosophy of This Course

- Emphasize concepts (i.e., ideas) and algorithm, rather than mathematics and asymptotic theories.
- The main mathematical tools employed are **linear algebra** and **optimization**.
- The main statistical tools employed are **least squares (LS)** and **maximum likelihood (ML)**.
 - No GMM since this is a statistical book (this is covered in recent econometric literature and may be incorporated in the future teaching).
 - LS can be interpreted as ML.
 - Bayesian methods are related to ML and will also be covered.
 - So ML is the focal point of our statistical tools.
- Two traditions in statistics: the **frequentist** approach and the **Bayesian** approach.
 - The former treats the parameter as fixed (i.e., there is only one true value) and the samples as random; LS and ML belong to the former.
 - The latter treats the parameter as random and the samples as fixed.

The Content of This Course

- I will teach this course as a complement of Econ6001 or Econ6005,² so all contents in these two courses (esp. OLS) are assumed to be known and will not be repeated (or repeated only briefly).
- I classify the **machine learning (ML)**³ problems into three groups: **regression**, **classification** and **clustering**.
- Mathematically, the former two are estimating the conditional mean and the last is estimating the joint density.
- Less and less information is provided in the three problems.
 - The **response** Y is available in the former two, but not available in the last, which is why the name "**supervised learning**" and "**unsupervised learning**".⁴
 - Y in regression is **quantitative** and in classification is **qualitative** (e.g., binary or categorical), where "quantitative" means the magnitude of variable conveys usual information, but "qualitative" is the converse (e.g., we can assign values to a binary variable as 0/1 or $-1/1$).

²Econ6001 and Econ6005 emphasize LS and GMM and this course emphasizes ML, and these three are the most important approaches in econometrics.

³ML stands for either maximum likelihood or machine learning, depending on the context.

⁴In the former two, our target is to build a prediction model, or a **learner**, while in the latter, we try to understand how the data are organized or clustered.

Course Outline

- Lecture 01: Introduction to Statistical Learning (Chapters 1-3)
- Lecture 02: Classification (Chapter 4)
- Lecture 03: Clustering (2nd half of Chapter 12)
- Lecture 04: Model Selection and Regularization (Chapter 6 and 1st half of Chap 5)
- Lecture 05: Principal Components Analysis (1st half of Chapter 12)
- Lecture 06*: Text As Data (Chapter 8 of BDS and Chapter 10 of DSB)
[ECON6087: Textual Analysis for Economists]
- Lecture 07: Moving Beyond Linearity (Chapter 7)
- Lecture 08: Tree-Based Methods (Chapter 8 and 2nd half of Chapter 5)
- Lecture 09: Support Vector Machines (Chapters 9)
- Lecture 10: Deep Learning (Chapter 10)

- Lectures 4, 8 and 10 may take more than one half day (i.e., 3 hours) to cover.

- Slides indexed by (*): may be covered in the lecture, but not related to the assignments.
- Slides indexed by (**) and Appendices: not covered in the lecture, only for after-class reading.
- **Suggestion:** Preview the slides before attending the lecture and concentrate in class on the points about which you found confused.

A Brief History of Statistical Learning

(Chapter 1)

Statistics in the News

- **How IBM Built Watson, Its 'Jeopardy'-Playing Supercomputer, 2011:**

▶ Watson in Jeopardy

- **Learning from its mistakes:** According to [David Ferrucci](#) (PI of Watson DeepQA technology for IBM Research), "Watson's software is wired for more than handling natural language processing".

- "It's **machine learning** allows the computer to become smarter as it tries to answer questions — and to learn as it gets them right or wrong."

- **FiveThirtyEight⁵ – Nate Silver's Political Calculus:** correctly predicted Obama and Biden in 50 and 48 out of 50 states; mistakenly predicted Trump, but less mistaken than other analysts. Something new:

▶ How (un)popular is Joe Biden?

▶ What Redistricting Looks Like In Every State?

- **For Today's Graduate, Just One Word: Statistics, 2009:**

▶ Statistics is the Future

- "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." — [Hal Varian](#), chief economist at Google.

⁵538 is the number of electors in the United States electoral college.

A Brief History of Statistical Learning

- Statistical learning refers to a set of tools for **making sense of/understanding complex datasets**.
- 1795: linear regression
- 1936: linear discriminant analysis (LDA)
- 1940s: logistic regression
- early 1970s: ridge regression and generalized linear models (GLM)
- until end of 1970s: all about linear methods due to computational constraints
- mid 1980s: classification and regression trees (CART), followed shortly by generalized additive models (GAM)
- late 1980s: neural networks
- 1990s: lasso, support vector machines (SVM), random forests (RF), boosting, BART
- new millennium: powerful and relatively user-friendly softwares, such as **R** and Python.

Our Plan

- **Traditional** toolkit of statistical learning:
 - Linear Regression.
 - Ridge Regression.
 - Logistic Regression.
 - Linear (or Quadratic) Discriminant Analysis.
 - K -Nearest Neighbors
 - Principal Components Analysis (PCA)
- **Modern** toolkit of statistical learning:
 - Lasso
 - Bagging
 - Random Forests
 - Boosting
 - BART
 - Support Vector Machines
 - Deep Learning
 - Matrix Completion
 - Clustering

What Is Statistical Learning?

(Section 2.1)

[Example] Advertising

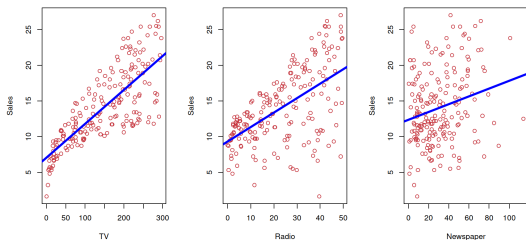


FIGURE 2.1. *The Advertising data set. The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively.*

- Can we predict sales using these three variables?
- Perhaps we can do better using a model

$$\text{sales} \approx f(\text{TV}, \text{radio}, \text{newspaper}).$$

Notation

- We will roughly follow the notations of ISLR.
- Here **sales** is a **response**, or **output variable**, or **target**, or **dependent variable** that we wish to predict. We generically refer to the response as Y .
- **TV** is a **feature**, or **input variable**, or **predictor**, or **independent variable**, or just a **variable**; we name it X_1 .
- Likewise name **radio** as X_2 , and so on.
- We can refer to the input vector (by default, a column) collectively as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = (X_1, X_2, X_3)^T, \text{ where } ^T \text{ is the transpose.}$$

- Now we write our model as

$$Y = f(X) + \varepsilon,$$

where ε is a random **error term**, independent of X and $E[\varepsilon] = 0$, capturing measurement errors and other discrepancies (e.g., missing variables), and f represent the **systematic** information that X provides about Y .

- In essence, statistical learning refers to a set of approaches for estimating f .

Why Estimate f ?

- **Prediction:** With a good f we can make predictions of Y at new points $X = x$.
 - $\hat{Y} = \hat{f}(X)$, where the form of \hat{f} does not matter, as long as it yields accurate predictions of Y , so can be treated as a **black box**.
- **Inference:** (i) We can understand which components of $X = (X_1, \dots, X_p)^T$ are important in explaining Y , and which are irrelevant, e.g., **seniority** (or experience) and **years of education** have a big impact on **income**, but **marital status** typically does not. (ii) Depending on the complexity of f , we may be able to understand how each component X_j of X affects Y . (iii) Is a linear \hat{f} enough to model the relationship between Y and X ?
 - The form of \hat{f} is important for this purpose.

Reducible and Irreducible Errors

- In prediction, assume both \hat{f} and X are fixed; then

$$\begin{aligned} E \left[(Y - \hat{Y})^2 \right] &= E \left[\left(f(X) + \varepsilon - \hat{f}(X) \right)^2 \right] \\ &= \underbrace{E \left[\left(f(X) - \hat{f}(X) \right)^2 \right]}_{\text{Reducible}} + \underbrace{\text{Var}[\varepsilon]}_{\text{Irreducible}}, \end{aligned}$$

where the randomness is only from ε .

- ε is not predictable from X since it is independent of X , so is irreducible.

- The irreducible error provides an upper bound on the accuracy of prediction, and this bound is usually unknown in practice.

- What is the optimal $f(X)$ in the sense of minimizing $E \left[(Y - g(X))^2 \right]$ among all functions g ?
- From Econ6001 or Econ6005, we know that the ideal $f(x) = E[Y|X=x]$, the **regression function**.⁶
- So one main target of machine learning is to find \hat{f} , aiming to minimize the reducible error $\left[f(X) - \hat{f}(X) \right]^2$ with $f(X) = E[Y|X]$, i.e., estimate $E[Y|X]$.

⁶We use **capital normal font** to denote random variables/vectors, and **lower case normal font** to denote their realizations.

How Do We Estimate f ?

- First, suppose we have n observations

$$\{(x_1, y_1), \dots, (x_n, y_n)\},$$

where $x_i = (x_{i1}, \dots, x_{ip})^T$.

- These observations are called the **training data** because they are used to train, or teach, our method how to estimate f .
 - **Training sample/data** and **training/learning** in machine learning are equivalent to the statistical terms "observed data" and "estimating".
- To estimate $E[Y|X = x]$, we need many data points at $X = x$, but typically we have few if any data points with $X = x$ exactly.

K-Nearest Neighbors Estimator

- Relax the definition and let

$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x)) := \frac{1}{K} \sum_{i \in \mathcal{N}(x)} y_i,$$

where $\mathcal{N}(x)$ is some **neighborhood** of x , e.g., including K (say, $10\%n$) nearest neighbors results in the **K-nearest neighbors (KNN)** estimator. [[figure here](#)]

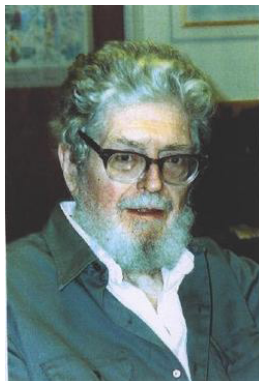
- Implicitly, we assume $f(x)$ is smooth since the data points near x can provide information on $f(x)$.

- Nearest neighbor averaging can be pretty good for small p — i.e., $p \leq 4$ and large-ish \mathcal{N} .
- We will discuss smoother versions, such as kernel and spline smoothing later in [Lecture 7](#); these three are all **nonparametric** methods.

History of the KNN Estimator



Evelyn Fix
(1904-1965, UC-Berkeley)



Joseph L. Hodges Jr.
(1922-2000, UC-Berkeley)

The Curse of Dimensionality

- Nearest neighbor methods can be lousy when p is large.
- **Reason:** Nearest neighbors (e.g., 10% neighborhood) tend to be far away in high dimensions (i.e., not neighbors anymore), called the **curse of dimensionality**.

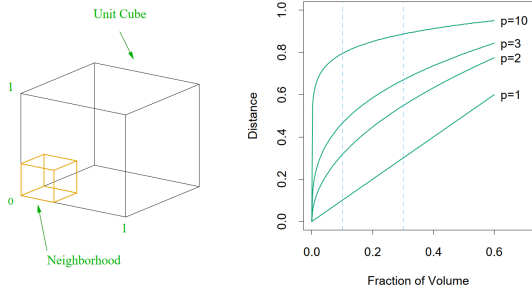


FIGURE The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p . In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.

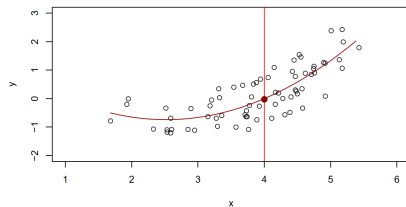
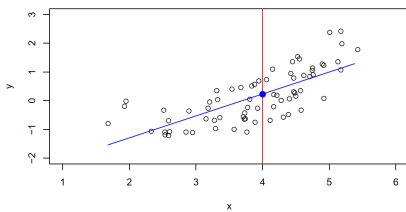
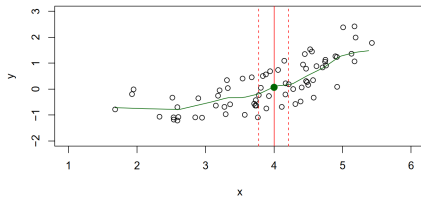
Parametric and Structured Models

- The **linear** model is an important example of a **parametric** model:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p =: \beta_0 + \beta^T X,$$

where $\beta = (\beta_1, \dots, \beta_p)^T$.

- A linear model is specified in terms of $p + 1$ parameters, $\beta_0, \beta_1, \dots, \beta_p$.
- We estimate the parameters by fitting the model to training data, e.g., by least squares.
- Although it is **almost never correct**, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.
- The next slide shows the difference in applying KNN, a linear model, and a quadratic model on a simulated dataset with $p = 1$, where $f_L(X) = \beta_0 + \beta_1 X$ in the linear model, $f_Q(X) = \beta_0 + \beta_1 X + \beta_2 X^2$ in the quadratic model, and both the linear and quadratic models are fitted by least squares.
 - Order of fitting: KNN \succ quadratic \succ linear.



[Example] Income: $p = 2$

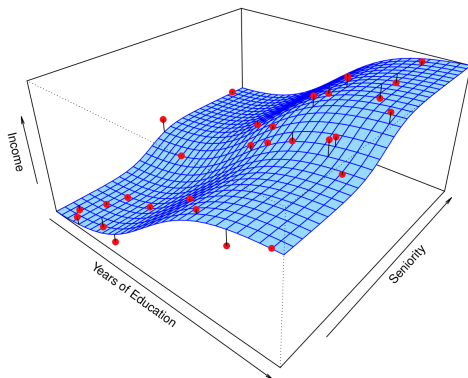


FIGURE 2.3. The plot displays **income** as a function of **years of education** and **seniority** in the **Income** data set. The blue surface represents the true underlying relationship between **income** and **years of education** and **seniority**, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

Linear Model

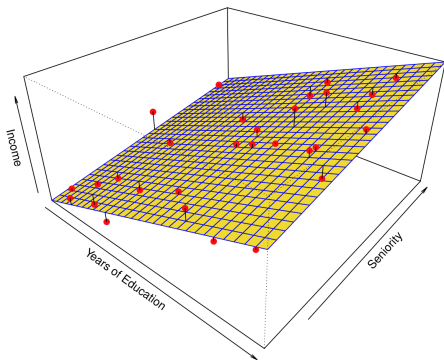


FIGURE 2.4. A linear model fit by least squares to the **Income** data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.

- The linear model does not capture the curvature of $f(X)$, but captures the positive relationship between **years of education** and **income** and slightly less positive relationship between **seniority** and **income**.

Nonparametric Model

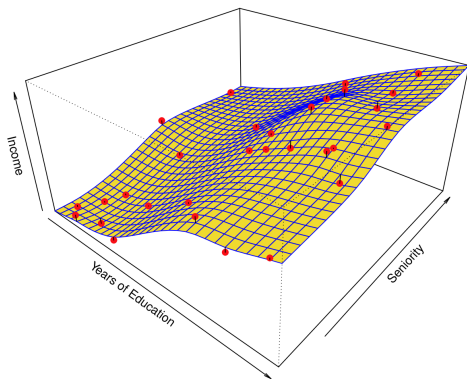


FIGURE 2.5. A smooth thin-plate spline fit to the **Income** data from Figure 2.3 is shown in yellow; the observations are displayed in red.

- By choosing the correct amount of smoothness, the nonparametric estimator \hat{f} is very close to the true f .

Overfitting

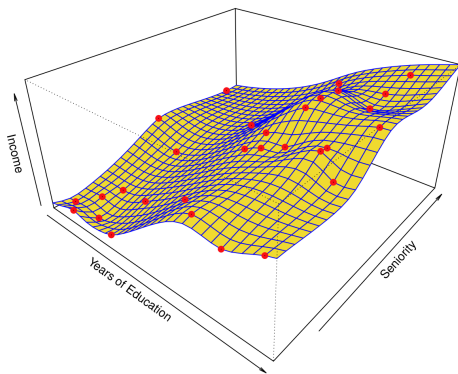


FIGURE 2.6. A rough thin-plate spline fit to the **Income** data from Figure 2.3. This fit makes zero errors on the training data.

- Overfitting means the fitted model follows the errors, or noise, too closely, which is not desirable when the fitted model is applied to **test data**.⁷

⁷Test data are independent of training data but following the same probability distribution. < > >> >>> >>>>

The Trade-Off Between Prediction Accuracy and Model Interpretability

- Linear models are easy to interpret (which is good for inference); thin-plate splines are not.
- On the other hand, more flexible models tend to predict well, although not always so.

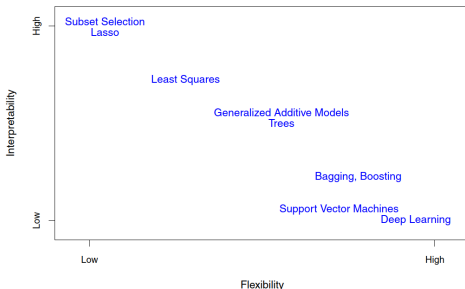


FIGURE 2.7. A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

Regression versus Classification Problems

- In all examples above, Y is **quantitative**, i.e., takes numerical values. We refer to the associated problems as **regression** problems.
- When Y is **qualitative** (aka **categorical**), i.e., takes K , $K \geq 2$, unordered values such as the brand of product purchased, we refer to the associated problems as **classification** problems.
 - Other popular classification problems: spam filtering of our mailbox, post code reading, face recognizing, disease diagnosis, potential customer detecting, etc.
 - **Logistic regression** in [Lecture 2](#) is termed as a regression, but is a classification method.
- Some methods can be used in either regression or classification, e.g., KNN, random forests, boosting, BART, deep learning, etc.
- The types of predictors are less important for most methods in this course.

Supervised versus Unsupervised Learning

- When there are y_i 's to monitor or supervise, we develop **supervised learning** methods, e.g., the regression and classification problems.
- When only x_i 's are available, we develop **unsupervised learning** methods.
- Except **cluster analysis** or **clustering** (Lecture 3) and **principal components analysis** (Lecture 5), all other lectures are devoted to supervised learning methods.
- Yet another machine learning method that we will not touch is the so-called **reinforcement learning**, which focuses on finding a balance between **exploration** (of uncharted territory) and **exploitation** (of current knowledge) as time passing.
 - The power of reinforcement learning was exaggerated by the success of AlphaGo.



AlphaGo

▶ Alphago Beats Lee Sedol

Philosophy

Supervised Learning

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.
- Statistical learning is a fundamental ingredient in the training of a modern **data scientist**.

Unsupervised Learning

- Objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Difficult to know how well your are doing.
- Different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of **Artificial Intelligence (AI)**.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap — both fields focus on supervised and unsupervised problems:
 - Machine learning has a greater emphasis on **large scale** applications and **prediction accuracy** based on algorithms.
 - Statistical learning emphasizes **models** and their interpretability, and **precision** and **uncertainty**.
 - There are a lot of pitfalls and caveats in machine learning without thinking statistically!
- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.
- Machine learning has the upper hand in **Marketing!**
- Along the following spectrum, you move from heavy focus on measuring and inferring the truth to a more pragmatic "useful is true" pattern discovery approach:

Econometrics \longrightarrow Statistics \longrightarrow Data Mining/Big Data/Data Science
 \longrightarrow Machine Learning and AI

Assessing Model Accuracy

(Section 2.2)

Measuring the Quality of Fit

- **Why** so many different machine learning methods, rather than a single best method?
- For different datasets, the best methods are different!
- **How** to choose the best method given a dataset? We first introduce some relevant concepts here, and explain how to apply these concepts in future lectures.
- Suppose we fit a model $\hat{f}(x)$ to some training data $\mathbf{Tr} = \{x_i, y_i\}_{i=1}^N$, and we wish to see how well it performs.
- We could compute the average squared prediction error over \mathbf{Tr} – the **mean squared error (MSE)**:

$$\text{MSE}_{\mathbf{Tr}} = \text{Ave}_{i \in \mathbf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2 := \frac{1}{N} \sum_{i \in \mathbf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2.$$

- This may be biased toward more overfit models.

- Instead we should, if possible, compute it using fresh test data $\mathbf{Te} = \{x_i, y_i\}_{i=1}^M$:

$$\text{MSE}_{\mathbf{Te}} = \text{Ave}_{i \in \mathbf{Te}} \left[y_i - \hat{f}(x_i) \right]^2 = \frac{1}{M} \sum_{i \in \mathbf{Te}} \left[y_i - \hat{f}(x_i) \right]^2.$$

Training MSE versus Test MSE

- **Figures 2.9–2.11**: training MSE tends to be quite a bit smaller than test MSE, and a low training MSE by no means guarantee a low test MSE.
 - Training MSE always decreases as flexibility increases, but test MSE exhibits a *U*-shape.
- In **Figures 2.10**, the test MSE of linear regression is close to the minimum of test MSE curve because the true f is close to linear.
- In **Figures 2.11**, there is a rapid decrease in both curves before the test MSE starts to increase slowly because the true f is highly nonlinear.

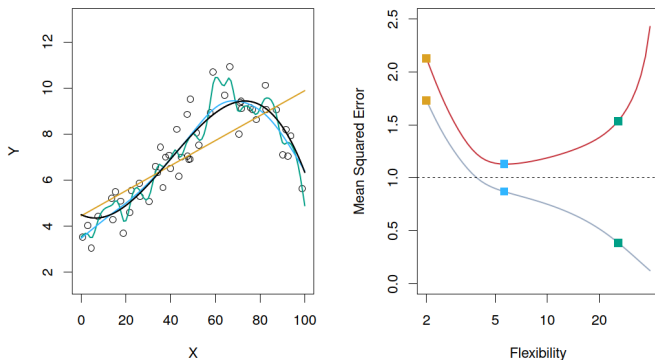


FIGURE 2.9. Left: Data simulated from f , shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

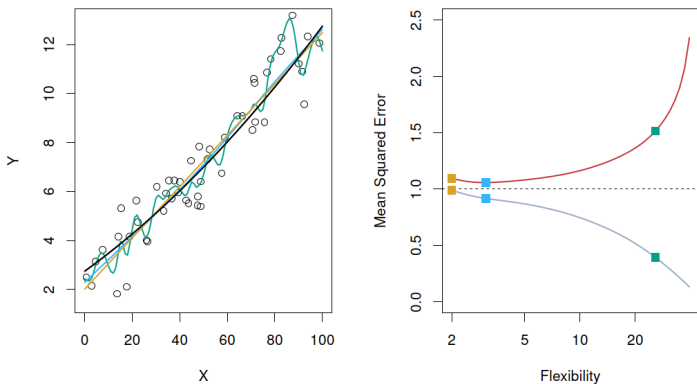


FIGURE 2.10. Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

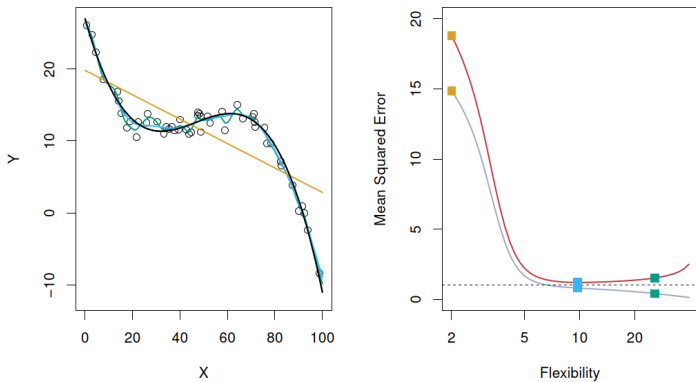


FIGURE 2.11. Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.

The Bias-Variance Trade-Off

- The expected test MSE, for a given value x_0 , is

$$E \left[\left(Y - \hat{f}(X) \right)^2 \mid X = x_0 \right] = \text{Var}[\varepsilon] + \left[\text{Bias} \left(\hat{f}(x_0) \right) \right]^2 + \text{Var} \left(\hat{f}(x_0) \right)$$

$$= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}.$$

- The overall expected test MSE can be computed by averaging $E[(Y - \hat{f}(x_0))^2 \mid X = x_0]$ over all possible values of x_0 in the test set.
- In general, more flexible statistical methods have higher variance.
 - The green curve in [Figures 2.9](#) follows the observations closely, so changing any data point would change \hat{f} considerably. The orange least squares line is the opposite case.
- On the other hand, more flexible statistical methods have lower bias.
- The optimal flexibility is determined by the trade-off between squared bias and variance. [see [Figures 2.12](#)]
- It is easy to find a low bias method (e.g., passing through all training data points) or a low variance method (e.g., a horizontal line), but the wisdom of statistical learning is to balance these two factors.

Trade-Off Between Squared Bias and Variance

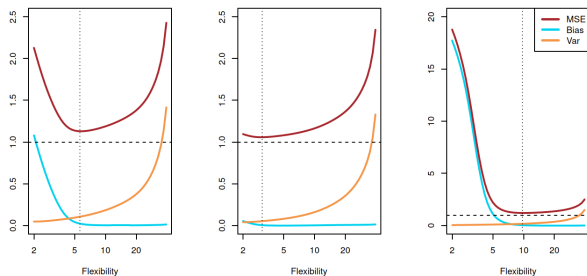


FIGURE 2.12. Squared bias (blue curve), variance (orange curve), $\text{Var}(\epsilon)$ (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

- The optimal flexibility varies a lot among the three datasets, depending on f (given the distribution of X and ϵ).
- In practice, f is **unknown**; we would use **cross validation** to estimate the test MSE in [Lecture 4](#).

The Classification Setting

- When Y is qualitative, we often use the training **error rate**, the fraction of incorrect classification, to quantify the accuracy of \hat{f} :

$$\text{Err}_{\text{Tr}} = \text{Ave}_{i \in \text{Tr}} I(y_i \neq \hat{f}(x_i)) := \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i), \quad (1)$$

where $I(y_i \neq \hat{y}_i)$ is an **indicator variable**, equal to 1 if $y_i \neq \hat{y}_i$ and 0 otherwise.

- The test error rate applies (1) to test data.
- A good classifier minimizes the test error.

The Bayes Classifier

- It can be shown that [Exercise for $K = 2$]

$$\arg \min_{g(x_0)} E[I(Y \neq g(X)) | X = x_0] = \arg \max_{k \in \{1, \dots, K\}} \Pr(Y = k | X = x_0),$$

i.e., assigns each test observation to the most likely class, given its predictor values, where $\Pr(Y = k | X = x_0)$, $k = 1, \dots, K$, are the **conditional class probabilities** at x_0 .

- This classifier is called the **Bayes classifier**, which is parallel to $E[Y | X = x_0]$ in the regression case.
- When $K = 2$, the Bayes classifier reduces to predict 1 at x_0 if $\Pr(Y = 1 | X = x_0) > 0.5$ and 2 otherwise.
 - The Bayes classifier's prediction is determined by the **Bayes decision boundary**, i.e., the line where $\Pr(Y = 1 | X = x_0) = 0.5$. [see [Figure 2.13](#)]
- The test error rate of the Bayes classifier is called the **Bayes error rate**, which is equal to $1 - \max_k \Pr(Y = k | X = x_0)$ at $X = x_0$.
 - The overall Bayes error rate is given by $1 - E[\max_k \Pr(Y = k | X)]$.
- The Bayes error rate is analogous to the irreducible error in the regression setting.

Bayes Decision Boundary

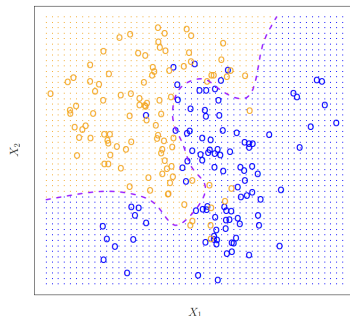


FIGURE 2.13. A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

- The (overall) Bayes error rate is $0.133 > 0$ since the classes overlap in the true population such that $\max_k \Pr(Y = k | X = x_0) < 1$ for some x_0 .

KNN

- To apply the Bayes classifier at x_0 , we need to estimate

$$\Pr(Y = k|X = x_0) = E[I(Y = k)|X = x_0].$$

- Since $E[I(Y = k)|X = x_0]$ is a conditional mean, we can estimate it by KNN:

$$\hat{\Pr}(Y = k|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}(x_0)} I(y_i = k),$$

and classify the test observation x_0 to $\arg \max_k \hat{\Pr}(Y = k|X = x_0)$, where K is the number of nearest neighbors, not the number of classes.

- KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary.
 - The counterpart of linear regression, logistic regression, will be discussed in [Lecture 2](#).
 - As a nonparametric method, $\hat{\Pr}(Y = k|X = x_0)$ will also suffer from the curse of dimensionality; however, the impact of p on $\arg \max_k \hat{\Pr}(Y = k|X = x_0)$ is less than on $\hat{\Pr}(Y = k|X = x_0)$.
- See [Figure 2.14](#) for $K = 3$, and [Figure 2.15](#) for $K = 10$.

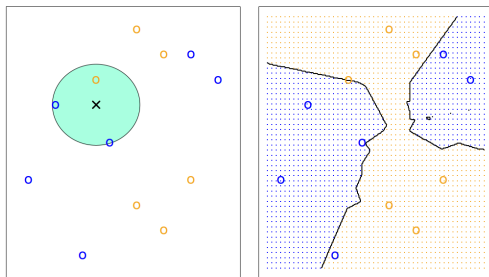


FIGURE 2.14. *The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue. Right: The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.*

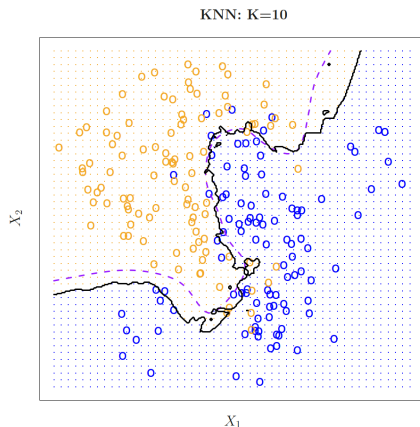


FIGURE 2.15. *The black curve indicates the KNN decision boundary on the data from Figure 2.13, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.*

- The KNN decision boundary is close to the Bayes decision boundary; the test error rate of KNN is also close the Bayes error rate.

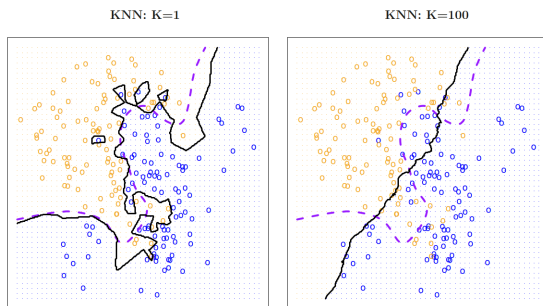
Choice of K 

FIGURE 2.16. A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.

- Low K : low bias and high variance;
- High K : high bias and low variance.

Training Error Rate versus Test Error Rate

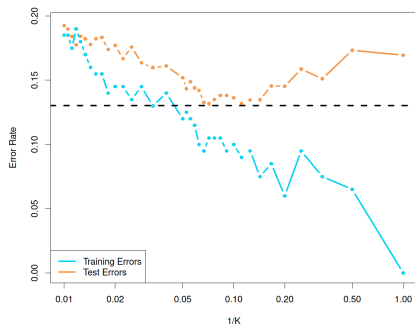


FIGURE 2.17. The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$ on the log scale) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.

- Training error rate does not match the test error rate, e.g., when $1/K = 1$, the former is zero but latter is high.
- As in the regression setting, the test error exhibits a U -shape; similarly, K can be chosen by cross validation.

Simple Linear Regression

(Section 3.1)

History of Least Squares

- The least-squares method is usually credited to Gauss (1809), but it was first published as an appendix to Legendre (1805) which is on the paths of comets. Nevertheless, Gauss claimed that he had been using the method since 1795 at the age of 18.



Carl F. Gauss
(1777-1855, Göttingen)



A.-M. Legendre
(1752-1833, École Normale)

Simple Linear Regression Using a Single Predictor X

- We assume a model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 and β_1 are two unknown constants that represent the **intercept** and **slope**, also known as **coefficients** or **parameters**, and ε is the error term.

- In the **Advertising** example, let Y be **sales** and X be **TV**.
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where \hat{y} indicates a prediction of Y on the basis of $X = x$.

- The hat symbol denotes an estimated value.

Estimating The Coefficients

- 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**.
- We define the **residual sum of squares (RSS)** as

$$\text{RSS} = \sum_{i=1}^n e_i^2$$

or equivalently,

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.
- The minimizing values can be shown to be

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

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

where $\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

[Example] Advertising

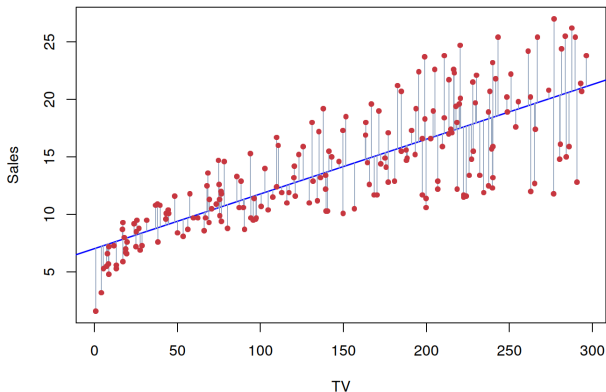


FIGURE 3.1. For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit captures the essence of the relationship, although it overestimates the trend in the left of the plot.

Assessing the Accuracy of the Coefficient Estimates

- The **standard error** of an estimator reflects how it varies under repeated sampling.
- We have

$$\text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right],$$

where $\sigma^2 = \text{Var}[\varepsilon]$ can be estimated by the **residual standard error**, $\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}$.

- These standard errors can be used to compute **confidence intervals** (CI).
- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- It has the form

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

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

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

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample).

- Similarly for the CI of β_0 .

Hypothesis Testing

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

H_0 : There is no relationship between X and Y

versus the **alternative hypothesis**

H_1 : There is some relationship between X and Y .

- Mathematically, this corresponds to testing

$$H_0: \beta_1 = 0$$

versus

$$H_1: \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and X is not associated with Y .

Continued

- To test the null hypothesis, we compute a ***t*-statistic**, given by

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

- This will have a *t*-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the ***p*-value**.
- Typical *p*-value cutoffs for rejecting the null are 5% or 1%, corresponding to *t*-statistics around 2 and 2.57 when $n \geq 30$.

Assessing the Overall Accuracy of the Model

- Residual Standard Error:

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

- R^2 or fraction of variance explained is

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

where $\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$ is the **total sum of squares**.

- RSE is an **absolute** measure of lack of fit, while R^2 is a **relative** measure, always in $[0, 1]$ and independent of the scale of Y .
- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y :

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

[Example] Advertising

- The 95% CIs for β_1 and β_0 are [0.042, 0.053] and [6.130, 7.935], respectively.

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

TABLE 3.1. 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. (Recall that the **sales** variable is in thousands of units, and the **TV** variable is in thousands of dollars.)

Quantity	Value
Residual standard error	3.26
R^2	0.612
<i>F</i> -statistic	312.1

TABLE 3.2. For the **Advertising** data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

Multiple Linear Regression

(Section 3.2)

Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

- 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, the model becomes

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

Interpreting Regression Coefficients

- The ideal scenario is when the predictors are uncorrelated — a **balanced design**:
 - Each coefficient can be estimated and tested separately.
 - Interpretations such as “a unit change in X_j is associated with a β_j change in Y , while all the other variables stay fixed”, are possible.
- Correlations amongst predictors cause problems:
 - The variance of all coefficients tends to increase, sometimes dramatically.
 - Interpretations become hazardous — when X_j changes, everything else changes.
- **Claims of causality** should be avoided for observational data.

The Woes of (Interpreting) Regression Coefficients

- “Data Analysis and Regression”, Mosteller and Tukey, 1977.
 - A regression coefficient β_j estimates the expected change in Y per unit change in X_j , with all other predictors held fixed. But predictors usually change together!
 - Example: Y = total amount of change in your pocket; X_1 = # of coins; X_2 = # of pennies, nickels and dimes. By itself, regression coefficient of Y on X_2 will be > 0 . But how about with X_1 in model?
 - Y = number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $\hat{Y} = b_0 + .50W - .10H$. How do we interpret $\hat{\beta}_2 < 0$?
- Two Quotes by Famous Statisticians:
 - “Essentially, all models are wrong, but some are useful” – George Box
 - “The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively” – Fred Mosteller and John Tukey, paraphrasing George Box.

Estimation and Prediction for Multiple Linear Regression

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, we can make predictions using the formula

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

- We estimate $\beta_0, \beta_1, \dots, \beta_p$ as the values that minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_p x_{ip} \right)^2. \end{aligned}$$

- This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

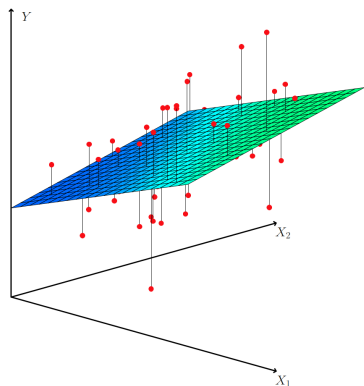
Illustration: Fitting Multiple Linear Regression with $p = 2$ 

FIGURE 3.4. 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.

[Example] 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

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.

- Different from the simple linear regression, the coefficient of newspaper is not significant now.

	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.

- newspaper is a surrogate for radio; it gets "credit" in the simple linear regression for the association between radio and sales.

Some Important Questions

- 1 Is at least one of the predictors X_1, X_2, \dots, 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?

Is There a Relationship Between the Response and Predictors?

- For the first question, we can use the **F-statistic**:

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

under $H_0: \beta_1 = \dots = \beta_p = 0$.

- $E\left[\frac{RSS}{n-p-1}\right] = \sigma^2$ as long as the linear model is correct, while $E\left[\frac{TSS-RSS}{p}\right] = \sigma^2$ under H_0 and $> \sigma^2$ under H_1 : at least one β_j is non-zero, so we reject H_0 for a large F .

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.

Deciding on the Important Variables, Model Fit and Prediction Uncertainty

- If there is at least one variable useful, which one or ones?
- Checking individual p -values does not work well especially when p is large because we are likely to make some false discoveries even if no variables are useful.
- In [Lecture 4](#), we will discuss this **variable selection** problem in details.
- $R^2 = \text{Cor}(Y, \hat{Y})^2$ now.
- $\text{RSE} = \sqrt{\frac{1}{n-p-1} \text{RSS}} =: \hat{\sigma}$ now.
- Prediction uncertainty still includes three parts: Irreducible Error + Bias² + Variance. Suppose we ignore the model bias; then compared with the CI of $f(X)$, the **prediction interval** of Y would be wider since it includes also the irreducible error component.

Other Considerations in the Regression Model

(Section 3.2)

Qualitative Predictors

- Some predictors are not **quantitative** but are **qualitative**, taking a discrete set of values.
- These are also called **categorical** predictors or **factor** variables.
- See for example the scatterplot matrix of the **Credit** card data in the next slide.
- In addition to the 7 quantitative variables shown, there are 4 qualitative variables: **own** (house ownership), **student** (student status), **status** (marital status), and **region** (East, West or South).

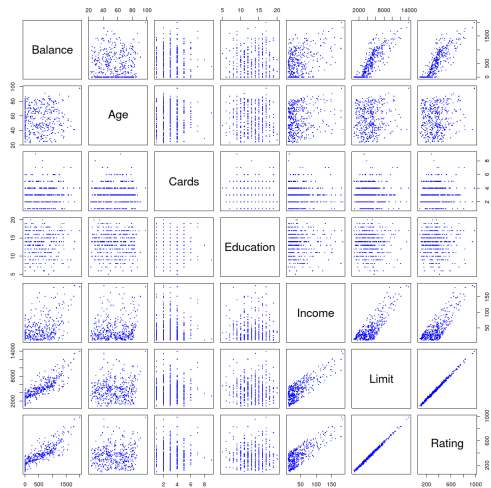


FIGURE 3.6. The Credit data set contains information about balance, age, cards, education, income, limit, and rating for a number of potential customers.

Predictors with Only Two Levels

- Suppose that we wish to investigate differences in credit card balance between those who own a house and those who don't, ignoring the other variables.
- We create a new **dummy variable**

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person does not own a house.} \end{cases} \quad (3.26)$$

- **Resulting Model:**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 + \varepsilon_i, & \text{if } i\text{th person does not own a house.} \end{cases} \quad (3.27)$$

Result for the Own Model

	Coefficient	Std. error	<i>t</i> -statistic	<i>p</i> -value
Intercept	509.80	33.13	15.389	< 0.0001
own[Yes]	19.73	46.05	0.429	0.6690

TABLE 3.7. *Least squares coefficient estimates associated with the regression of balance onto own in the Credit data set. The linear model is given in (3.27). That is, ownership is encoded as a dummy variable, as in (3.26).*

- What is the meaning of $\hat{\beta}_0$, $\hat{\beta}_0 + \hat{\beta}_1$ and $\hat{\beta}_1$?
- The *p*-value for the dummy variable is large. What does that mean?
- The final prediction does not depend on (i) which category is coded as 1, (ii) a 0/1 coding or a -1/1 coding is employed, which only affect the interpretation of coefficients.

Qualitative Predictors with More than Two Levels

- With more than two levels, we create additional dummy variables. For example, for the **region** variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1, & \text{if } i\text{th person is from the South} \\ 0, & \text{if } i\text{th person is not from the South,} \end{cases} \quad (3.28)$$

and the second could be

$$x_{i2} = \begin{cases} 1, & \text{if } i\text{th person is from the West} \\ 0, & \text{if } i\text{th person is not from the West.} \end{cases} \quad (3.29)$$

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \varepsilon_i, & \text{if } i\text{th person is from the West} \\ \beta_0 + \varepsilon_i, & \text{if } i\text{th person is from the East.} \end{cases} \quad (3.30)$$

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — East in this example — is known as the **baseline**.

Result for the Region Model

	Coefficient	Std. error	<i>t</i> -statistic	<i>p</i> -value
Intercept	531.00	46.32	11.464	< 0.0001
region[South]	-18.69	65.02	-0.287	0.7740
region[West]	-12.50	56.68	-0.221	0.8260

TABLE 3.8. Least squares coefficient estimates associated with the regression of **balance** onto **region** in the **Credit** data set. The linear model is given in (3.30). That is, *region* is encoded via two dummy variables (3.28) and (3.29).

- **What** is the meaning of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$?
- The *p*-values for the two dummy variables are large. **What** does that mean?
- Although the coefficients and their *p*-values depend on the choice of dummy variable coding, the final prediction does not.
- We can use the *F*-test to check whether there is any relationship between **balance** and **region**, $H_0: \beta_1 = \beta_2 = 0$, which does not depend on coding.
 - The *p*-value in this example is 0.96.

Extensions of the Linear Model: Interaction

- In our previous analysis of the **Advertising** data, we assumed that the effect on **sales** of increasing one advertising medium is independent of the amount spent on the other media.
- But 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.
- In this situation, given a fixed budget of \$100,000, spending half on **radio** and half on **TV** may increase sales more than allocating the entire amount to either **TV** or to **radio** . [see [Figure 3.5](#)]
- In marketing, this is known as a **synergy** effect, and in statistics it is referred to as an **interaction** effect.

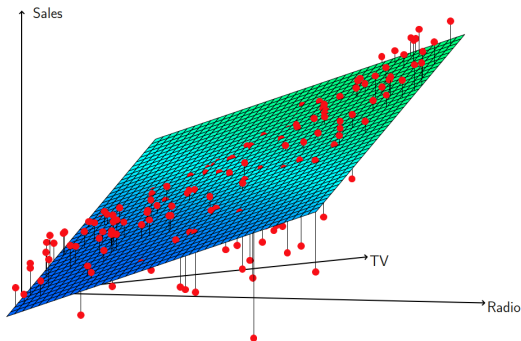


FIGURE 3.5. For the Advertising data, a linear regression fit to sales using TV and radio as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals (those visible above the surface), tend to lie along the 45-degree 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.

Interaction Term

- Adding an **interaction term** in the linear model, we have

$$\begin{aligned} \text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 (\text{radio} \times \text{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon, \end{aligned}$$

so the slope of **TV** depends on **radio**.

	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	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

TABLE 3.9. For the **Advertising** data, least squares coefficient estimates associated with the regression of **sales** onto **TV** and **radio**, with an interaction term

- The *p*-value for the interaction term is small, so the new model is superior to the model that contains only **main effects**.

Hierarchical Principle

- Sometimes it is the case that an interaction term has a very small p -value, but the associated main effects (in this case, **TV** and **radio**) do not.
- The **hierarchy principle**:
If we include an interaction in a model, we should also include the main effects, even if the p -values associated with their coefficients are not significant.
- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, if the main effect of **TV** does not exist, i.e., $\beta_1 = 0$, then

$$\frac{\partial \text{sales}}{\partial \text{TV}} = \beta_3 \times \text{radio},$$

i.e., **TV** does not affect **sales** solely but only through **radio**!

- **radio**×**TV** is correlated with **TV**, so it is hard to imagine to leave **TV** out but keep **radio**×**TV**.

Interactions between Qualitative and Quantitative Variables

- Consider the **Credit** data set, and suppose that we wish to predict balance using **income** (quantitative) and **student** (qualitative).
- Without an interaction term, the model takes the form [see the left panel of **Figure 3.7**]

$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2, & \text{if } i\text{th person is a student} \\ 0, & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2, & \text{if } i\text{th person is a student} \\ \beta_0, & \text{if } i\text{th person is not a student.} \end{cases} \end{aligned} \quad (3.34)$$

- With interactions, it takes the form [see the right panel of **Figure 3.7**]

$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i, & \text{if student} \\ 0, & \text{if not student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i, & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i, & \text{if not student.} \end{cases} \end{aligned} \quad (3.35)$$

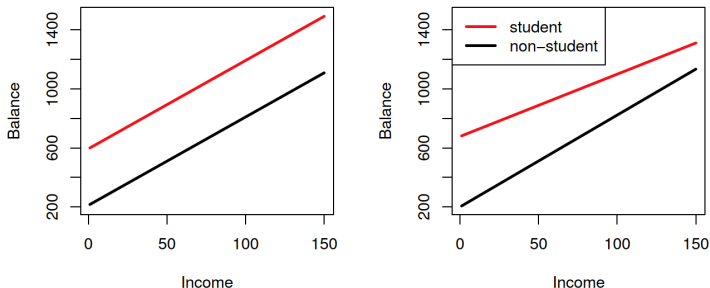


FIGURE 3.7. For the **Credit** data, the least squares lines are shown for prediction of **balance** from **income** for students and non-students. Left: The model (3.34) was fit. There is no interaction between **income** and **student**. Right: The model (3.35) was fit. There is an interaction term between **income** and **student**.

- $\hat{\beta}_2 > 0$, and $\hat{\beta}_3 < 0$.

Extensions of the Linear Model: Nonlinear Relationships

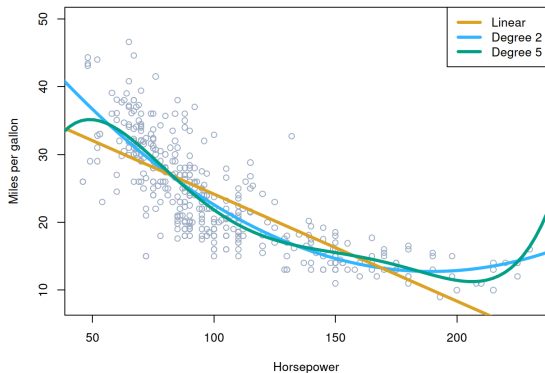


FIGURE 3.8. The `Auto` data set. For a number of cars, `mpg` and `horsepower` are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes `horsepower`² is shown as a blue curve. The linear regression fit for a model that includes all polynomials of `horsepower` up to fifth-degree is shown in green.

Polynomial Regression

- Figure 3.8 suggests that

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$$

may provide a better fit.

	Coefficient	Std. error	<i>t</i> -statistic	<i>p</i> -value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower ²	0.0012	0.0001	10.1	< 0.0001

TABLE 3.10. For the **Auto** data set, least squares coefficient estimates associated with the regression of **mpg** onto **horsepower** and **horsepower²**.

- The *p*-value for **horsepower²** is small.
- Are higher-order polynomials necessary? wiggly...
- Polynomial regression** and other nonlinear extensions of the linear model will be discussed in **Lecture 7**.

Potential Problems

- ① Nonlinearity of the data: **residual plot**, $e_i := y_i - \hat{y}_i$ against \hat{y}_i , as a detection tool.
- ② Correlation of error terms: in time series, plot e_i against time to detect tracking.
- ③ Non-constant variance of error terms: funnel shape in the residual plot; apply **weighted least squares**.
- ④ High leverage points: unusual x_i ; typically has great effects on $\hat{\beta}$; use **leverage statistic** h_i to detect high leverage points.
- ⑤ Outliers: unusual y_i ; typically has no great effect on $\hat{\beta}$ but has great effects on its inference; plot **studentized residuals** $\frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$ against \hat{y}_i to detect outliers ($|abs| > 3$).
- ⑥ Collinearity: use the **variance inflation factor (VIF)** to detect **multicollinearity** (> 5 or 10); either drop one of the collinear variables or average the standardized versions of collinear variables.

Generalizations of the Linear Model

- In much of the rest of this course, we discuss methods that expand the scope of linear models and how they are fit:
 - **Classification problems**: logistic regression, support vector machines
 - **Non-linearity**: kernel smoothing, splines and generalized additive models; nearest neighbor methods.
 - **Interactions**: Tree-based methods, bagging, random forests, boosting, BART and deep learning (these also capture non-linearities)
 - **Regularized fitting**: Ridge regression and lasso

Lab1: Introduction to R

(Section 2.3)

- Basic Commands
- Graphics
- Indexing Data
- Loading Data
- Additional Graphical and Numerical Summaries

Lab2: Linear Regression

(Section 3.6)

- Libraries
- Simple Linear Regression
- Multiple Linear Regression
- Interaction Terms
- Non-linear Transformations of the Predictors
- Qualitative Predictors
- Writing Functions