Notes: Introduction to Statistical Learning

A. John Woodill

Revision 1 – October 2018

Contents

1 Introduction

The following notes are from Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. For updated document and Python code see https://github.com/johnwoo Intro-to-Stat-Learning. For online book content see www.statlearning.com

2 Statistical Learning

2.1 Overview

Suppose we observe a quantitative response Y with different predictors, $X_1, X_2, ..., X_p$. We assume there is some relationship between $Y = X_p$. A general form is,

$$Y = f(x) + \epsilon$$

where f is some systematic information that X_p provides about Y. ϵ is the random error term, which is **independent** of X_p and mean zero.

Statistical learning refers to approaches that estimate f where we can estimate to provide predictions or inference.

2.1.1 Predictions

Predictions assume a set of inputs X_p are available but outputs Y may not be. Predictions follow the form,

$$\hat{Y} = f(\hat{X})$$

The accuracy of $\hat{Y}$ by predicting $f(\hat{X})$ is not perfect, which introductions an error of two quantities:

Reducible Error: Error can be improved with appropriate modeling strategies.

Irreducible Error: variability of ϵ affects the accuracy of the prediction; thus, cannot reduce the error introduced by ϵ. The error may contain (1) unmeasured variables, (2) unmeasured variation.

Derive reducible and irreducible errors by simply differencing $Y - \hat{Y}$, then find mean squared.

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(\hat{X})]^2$$

$$E(Y - \hat{Y})^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}}$$

Thus, the irreducible error depends on the variation in the error. Statistical Learning focuses on improving (minimizing) the reducible error. Note that the irreducible error will always provide an upper bound on the accuracy of the prediction, which is almost always unknown in practice

2.1.2 Inference

Inference relates to understanding the relationship between X and Y, or how Y changes in response to X.

- Which predictors are associated with the response?

- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

2.1.3 Estimating $f()$

Our goal is to apply statistical learning method to train data to estimate and unknown function f. Methods include parametric and nonparametric methods.

Parametric: Methods that use distributional assumptions are called parametric methods, because we estimate the parameters of the distribution assumed for the data. In OLS, assumptions about function form are linear with fixed parameters. No matter how much data you have, there will always be fixed parameters.

Examples:

- Logistic Regression
- Linear Discriminant Analysis
- Perceptron
- Naive Bayes
- Simple Neural Networks

Advantage: simplifies estimating $f()$ because it is easier to estimate a set of parameters, β_0, β_1.

Disadvantage: model does not usually match $f()$.

Non-Parametric: Methods do not make explicit assumptions about the functional form of $f()$. Goal is to get as close to the data points as possible without being to rough or widdly.

Examples:

- k-Nearest Neighbors
- Decision Trees like CART and C4.5
- Support Vector Machines

Advantage: potential to accurately fit a wider range of possible shapes for $f()$.

Disadvantage: do not reduce the problem of estimating f to a small number of parameters, thus a large number of observations is required to accurately estimate f.

2.1.4 Prediction Accuracy versus Model Interpretability

Why would we ever choose to use a more restrictive method instead of a very flexible approach?

Restrictive models, such as linear models, are more interpretable. In contrast, flexible approaches, such as splines, can provide complicated estimates of $f()$ that may improve prediction accuracy.

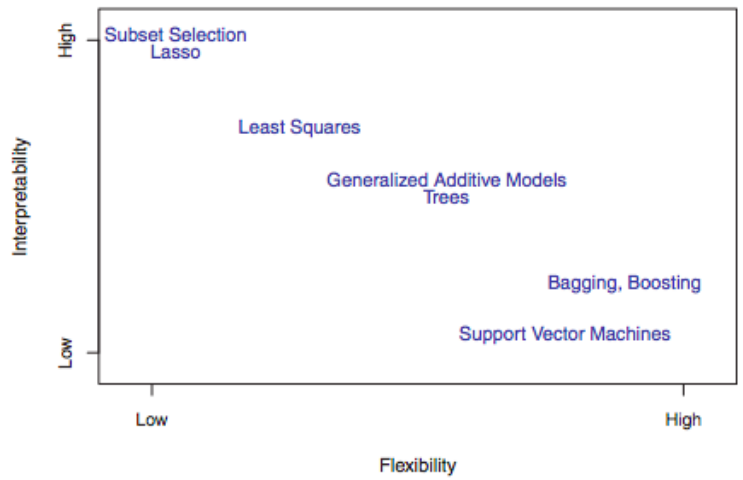

FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

The choice of modeling strategy comes down to the end goal: prediction or inference. Less flexible models are easy to interpret, thus prefered when the goal is inference. If prediction is prefered, then more flexible models may be prefered – although, more flexible models are not always prefered for prediction accuracy due to overfitting.

2.1.5 Supervised Versus Unsupervised Learning

Supervised: Each observation of the predictor measurements x_i there is an associated response measurement y_i. Goal is to fit a model that relates to the response predictors with an aim to accurately predict the response variable in the future.

Examples:

- Linear Regression
- Logistic Regression
- GAM
- Boosting
- Support Vector Machines

Unsupervised: Observations of the predictor x_i does not contain a response variable, y.

Examples:

- Cluster Analysis
- PCA

2.1.6 Regression versus Classification Problems

Variables are characterized by quantitative or qualitative (categorical). Quantitative values are numerical whereas qualitative variables take values in classes or categories.

- Regression Analysis: Uses quantitative variables
- Classification Analysis: Uses qualitative variables

We select statistical learning methods based on the response variable being quantitative or qualitative.
Note: distribution of predictors being qualitative or quantitative is less important.

2.2 Assessing Model Accuracy

2.2.1 Measuring Quality of Fit

To assess performance of statistical learning methods, we need to quantify the extent to which the predicted response value is close to the true value. The most commonly-used measure is mean-squared-error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Calculating training MSE (MSE on training data) doesn't tell us much about out-of-sample performance, which is prefered. Suppose (x_0, y_0) are previously unseen test observations. The test MSE is

$$Ave(y_0 - \hat{f}(x_0))^2)$$

It is important to note that minimizing the training data provides no guarantee that the method will also minimize the test data.

There is a trade-off between inflexibility versus flexible models. Degrees of freedom define the flexibility of a curve. A more restricted (smoother) curve has fewer degrees of freedom than a wiggly curve. As flexibility increases training MSE declines monotonically.

In the figure below, as the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE. As model flexibility increases, training MSE will decrease, but the test MSE may not.

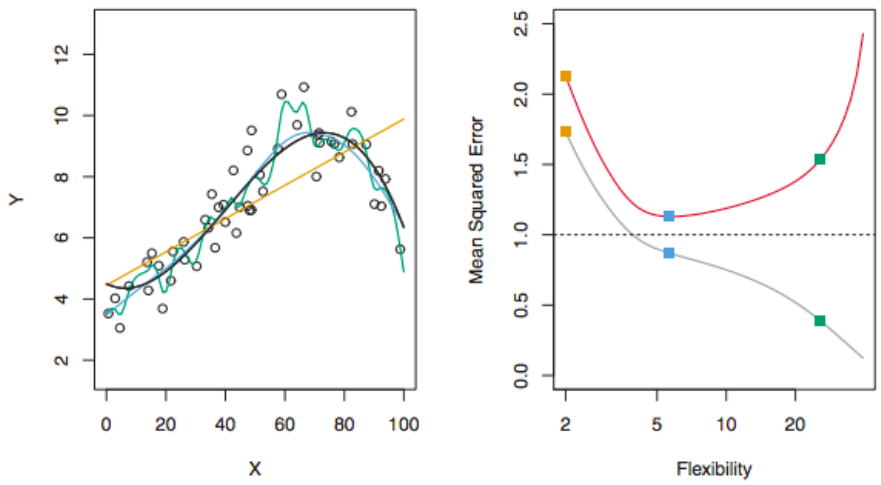

FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed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

Note: Overfitting occurs when a small training MSE but a large test MSE exists. This happens because the training model is find patterns in the data and not establishing the signal.

2.2.2 The Bias-Variance Trade-Off

The U-shaped observed in the test MSE curve is a result of two competing properties in statistical learning:

The test MSE, for a given value x_0, can be decomposed into the sum of three fundamental quantities: variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$, and the variance of the error terms ϵ,

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{Expected Test MSE}} = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

To minimize the expected test error, we need to select a stat. method that achieves a low variance and a low bias.

- Variance: amount by which $\hat{f}$ changes if estimated using different raining data.
- Bias:error that is introduced by approximating a real-life problem

Variance between training data sets shouldn't change $\hat{f}$ too much; however, methods that are more flexible have higher variance that will shift the MSE larger whereas restricted methods have low variance and will only cause small shifts.

In terms of bias, the inverse is true. Restricted methods do not identify the true response variable, which results in large bias; however, flexible methods are usually better at predicting the true response variable which provides less bias.

Bias-Variance Trade-Off

- Flexible Methods: Variance increases and bias will decrease MSE
- Restricted Methods: Variance decreases and bias will increases MSE

Note: The challenge lies in finding a method for which both the variance and the squared bias are low.

2.2.3 Classification Strategy

Model accuracy transfers over to classification problems. The most common approach is to quantify the accuracy of $\hat{f}$ using a training error rate, or the proportion of mistakes that are made to the training observations,

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

where $I(y_i \neq \hat{y}_i)$ is an indicator variable that equals 1 if $y_i \neq \hat{y}_i$, and zero if $y_i = \hat{y}_i$. If $I(y_i \neq \hat{y}_i) = 0$, then the observation was classified correctly. The test error is calculated as,

$$Ave(I(Y - 0 \neq \hat{y}_0))$$

The Bayes Classifier The error rate can be classified by assigning each observation to the most likely class, given its predictor values.The Bayes Classifier is,

$$Pr(Y = j|X = x_0)$$

or the probability that $Y = j$ given the observed predictor vector x_0. The Bayes Classifier establishes a Bayes decision boundary that falls on one side or the other of the classification.

Bayes error rate maximizes the probability of selecting,

$$1 - E(\max_j PR(Y = j|X))$$

and is analogous to the irreducible error.

K-Nearest Neighbors In theory, Bayes Classifier is the gold standard, but we don't always know the conditional distribution of Y given X;thus, we need to estimate the probability – K-Nearest Neighbor (KNN) is one method.

Given a positive integer K, and a test observation x_0, the KNN classifier identifies the K points in the training data that are closest to x_0, represented by N_0. Condition probabilities are estimated for class j as a fraction of N_0 whose response values equal j:

$$Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = J)$$

The choice of K has a drastic effect on the classifier obtained; small $K = 1$ provides a boundary that is overfly flexible and has a low bias but high variance. As K increases, method becomes less flexible and is closer to linear (high bias low variance). No strong relationship between test and train error rates. Flexible $K = 1$ have a low training rate (0), but test error will be high.

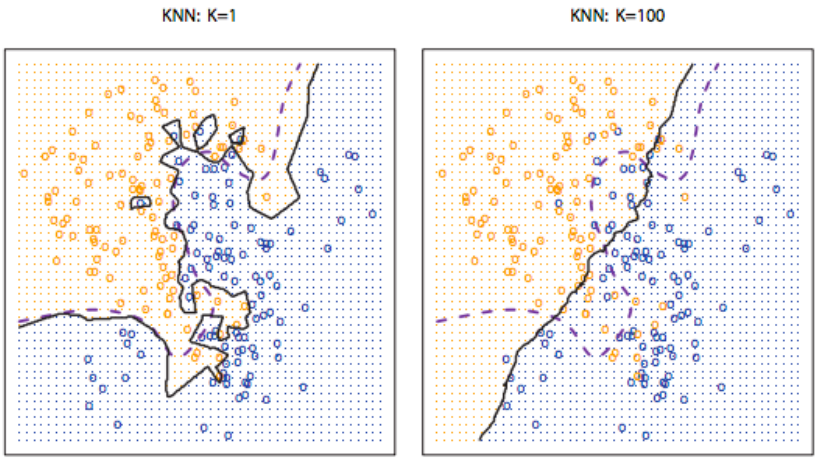

FIGURE 2.16. *A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple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

Note: In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method.

3 Linear Regression

3.1 Simple Linear Regression

Approach to predicting a quantitative response Y on the basis of a single predictor variable X, assuming an approximate linear relationship.

$$Y \approx \beta_0 + \beta_1 X$$

where β_0, β_1 are unknown constants that represent an intercept and slope, known as coefficients or parameters.

3.1.1 Estimating the Coefficients

Goal is to minimize the relationship between a linear line and the actual value, also known as residuals. Most approaches involve minimizing the least squares criterion.

Residual from linear regression is,

$$e_i = y_i - \hat{y}_i$$

where residual sum of squares (RSS) is,

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

and the minimization problem reduces to,

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{2}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \hat{\beta}\bar{x}$$

3.1.2 Assessing the Accuracy of the Coefficient Estimates

A sample mean is unbiased in the sense that on average the estimated sample equals population mean. By selecting multiply samples, calculating mean, and estimate mean of sample means, the mean should be close to the population mean, which produces unbiased mean. The regression mean provides a reasonable estimate of this sampling procedure.

To calculate how over-or-under the average estimate of the population mean is, we use the standard error,

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma}{n}$$

A regression line provides a reasonable estimate of the sample mean assuming the sample mean is randomly drawn multiple times and averaged.

Standard Error: tells us the average amount that is estimate $\hat{\mu}$ differs from the actual value of μ

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

where $\sigma^2 = Var(\epsilon)$ and we assume the errors are uncorrelated with common variance σ^2. Generally, we don't know σ, so we estimate from the residual standard error,

$$RSE = \sqrt{RSS/(n-2)}$$

Standard errors can then be used to calculate confidence intervals at a 95% confidence interval as,

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_2 + 2 \cdot SE(\hat{\beta}_1)]$$

SE can also be used to perform hypothesis tests on coefficients (stat. sign),

$$H_0 : \beta_1 (\text{There is no relationship between X and Y.})$$

$$H_a : \beta_1 \neq (\text{There is some relationship between X and Y.})$$

A t-stat measures the number of standard deviations that $\hat{\beta}_1$ is away from 0. Generally, a t-stat above 2 implied statistical significance.

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

p-value: a small p-value indicates that it is is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real associations.

In other words, a small p-value infers that there is an association between the predictor and the response, in which case we reject the null hypothesis.

3.1.3 Assessing the Accuracy of the Model

Model accuracy is typically assessed with residual standard error (RSE) and the R^2

Residual Standard Error RSE is an estimate of the standard deviation of ϵ, or the average amount that the response will deviate from the true regression line. RSE is thought of as a measure of lack of fit – low RSE indicates model fits data well, high RSE indicates poor fit.

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

R^2 Statistic R^2 takes the form of a proportion – the portion of the variance explained – and will be between 0 and 1

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

where $TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares. TSS measures the total variance in the response and is thought of as the amount of variability inherent in the response before the regression is performed. RSS measures the amount of variability that is left unexplained after performing the regression.

$R^2 = 0$ regression did not explain much of the variability in the response

$R^2 = 1$ indicates a large proportion of the variability in the response has been explained in the regression.

TSS uses mean of y_i whereas RSS uses residual differences.

An R-squared of 0.65 might mean that the model explains about 65% of the variation in our dependent variable.

Problems with R-squared (https://data.library.virginia.edu/is-r-squared-useless/)

- R-squared does not measure goodness of fit. It can be arbitrarily low when the model is completely correct. By making σ^2 large, we drive R-squared towards 0, even when every assumption of the simple linear regression model is correct in every particular.
- R-squared can be arbitrarily close to 1 when the model is totally wrong.
- R-squared says nothing about prediction error, even with σ^2 exactly the same, and no change in the coefficients. R-squared can be anywhere between 0 and 1 just by changing the range of X. We're better off using Mean Square Error (MSE) as a measure of prediction error.
- R-squared cannot be compared between a model with untransformed Y and one with transformed Y, or between different transformations of Y. R-squared can easily go down when the model assumptions are better fulfilled.

3.2 Multiple Linear Regression

Estimating separate simple linear regression models for each predictor is not entirely satisfactory: (1) unclear how single predictions affect or variables (2) individual regressions ignore the other regressors. A better solution is to provide individual slopes for each of the regressors,

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

β_j quantifies the association between that variable and the effect on a one unit increase in X_j, holding all other predictors fixed.

3.2.1 Estimating the Regression Coefficients

The parametric estimates are obtained using the same least squares minimization problem as before, although slightly more complicated. The differences is that covariates are adjusted based on correlations and will obtain different results than individual least squares estimates[1].

1. Is there a relationship between the response and predictors? In multiple variable linear regressions we need to consider whether there is a relationship between all variables. Thus, the null hypothesis is,

$$H_0 = \beta_1 = \beta_2 = ... = \beta_p = 0$$

An F-statistic establishes the hypothesis test,

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

where a value close to 1 establishes no relationship between the response and predictors where as an F-stat greater than 1 establishes the covariates represent a relationship to the response variable Y.

We can also test a subset of covariates to determine whether a relationship exists.
If we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will incorrectly conclude that there is a relationship. However, the F-statistic does not suffer from this problem because it adjusts for the number of predictors.

2. Deciding on Important Variables 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. This association is referred to as variable selection.

To determine the best variable selections, we can utilize BIC, AIC, R^2, or even RMSE. However, the size of model selection grows exponentially, so costs increase substantially.

Three approaches exist to validate model selection:

- Forward selection: start with intercept with no predictors and add variables that minimize the RSS.
- Backward selection: start with all variables and remove variables that are the least stat sign.
- Mixed selection: combination of forward and backward selection. Start with no variables, add variables that provides the best fit, but only add variables below a certain threshold.

3. Model Fit Most common numerical measures of model fit are RSE and R^2. It is important to note that R^2 will always increase with additional variables, so care needs to be taken when utilize R^2 as a model fit discussion. Additional RSE can increase when variables are added.

[1] A regression of shark attacks and ice cream sales reveals a significant result that shark attacks increase with ice cream sale due to increases in temperatures; however, when temperatures are included in the analysis to adjust for correlations between increases in temperature and ice cream sales, the estimate becomes insignificant.

4. Predictions Once the model has been fit, predictions are relatively straightforward. However, uncertainty exists,

- Are the coefficient estimates of the true population? Inaccuracies related to the reducible error.
- Does the linear model provide accurate approximations? Model bias may bias results.
- Even if we know the true values, we cannot perfectly predict the response because of the random error. Therefore, irreducible errors always exist in linear approximations.

Note: Confidence intervals are used to quantify uncertainty around model estimates.

3.3 Other Considerations in the Regression Model

3.3.1 Qualitative Predictors

Predictors can be qualitative.

Predictors with two levels Create a dummy variable, D, for two possible numerical values, such as 0 or 1. The level that is associated with 1 can be interpreted as,

$$D(1) = \beta_0 + \beta_1 + \epsilon$$

$$D(0) = \beta_0 + \epsilon$$

It is also possible to code with 1 and -1. In this case, the interpretation of the coefficients change.

$$D(1) = \beta_0 + \beta_1 + \epsilon$$

$$D(-1) = \beta_0 - \beta_1 + \epsilon$$

Predictors with more than two levels With more than two levels, dummy variables need to be spread out for each factor (level). When including more than two factors, there will always be one fewer dummy variables. The level with no dummy variable is known as the baseline and includes the constant and errors.
Note: The baseline establishes the number for which coefficients are differenced or added. For example, a baseline (intercept) reports 500. β_1 reports a coefficient of -5. Therefore, the dummy variable representing β_1 has a value of 495.

3.3.2 Extensions of the Linear Model

Two of the most important assumptions in linear regressions is,

- **Additive**: the effect of changes in predictor X on the response Y is independent of the values of other predictors.
- **Linear**: change in response Y due to a one-unit change in X is constant, regardless of the value of X

Removing the Additive Assumption Additive assumption assumes no relationship between predictors, which may not always hold (temp and precipitation). In statistics, this is known as an interaction effect. We can relax the additive assumption by including an **interaction term**.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Example,

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned} \quad (3.33)$$

We can interpret β_3 as the increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa). The coefficients that result from fitting the model (3.33) are given in Table 3.9.

textbfNote: The **hierarchical principle** states that if we include an interaction in a model, we should also include the main effects, even if the p-values associate with their coefficients are not sign.

Quantitative and qualitative variables can be interacted to remove the additive assumption.

Non-linear relationship The relationship between the response and predictor may be non-linear. We can accommodate this relationship by using a polynomial regression.

A simple way to fit a polynomial is to use quadratic functional form of variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

Note that this is still a linear model!!!

3.3.3 Potential Problems

Most common problems when fitting a linear regression,

- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- Collinearity.

1. Non-linearity of the Data If a true linear relationship exists between response and predictors, then we can utilize the linear interpretation discussed. However, nonlinearities can throw off modeling aspects and interpretations.

Residual plots are useful for identify non-linearities.

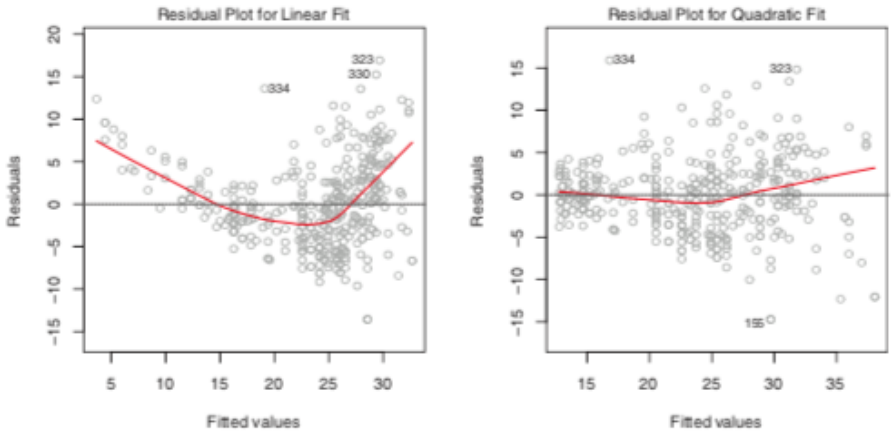

FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* `Auto` *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* `mpg` *on* `horsepower`*. A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* `mpg` *on* `horsepower` *and* $\texttt{horsepower}^2$*. There is little pattern in the residuals.*

Simple approaches to transform variables include log X, $sqrtX$, and X^2.

2. Correlation of Error Terms An important assumption is that the error terms are uncorrelated. Moreover, standard errors are calculated assuming uncorrelated error terms, thus may underestimate the true standard errors.

Correlation of errors terms may exists in time series data (serial correlation). One way to check for correlation in error terms is to plot residuals versus time series.

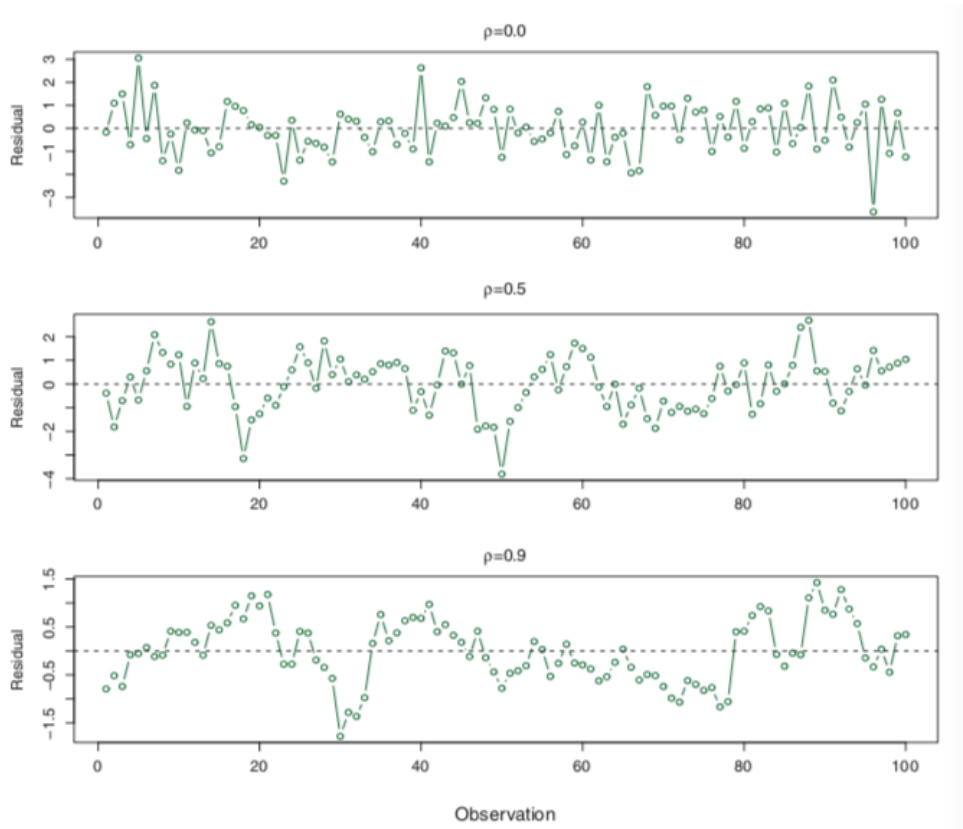

FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Correlation of error terms can exist outside of time series if groups (states or family members) are included in the variables.

Non-constant Variance of Error Terms Another important assumption is that the error terms have a constant variance, $Var(\epsilon) = \sigma^2$. Non-constant error terms exist with heteroscedastic data.

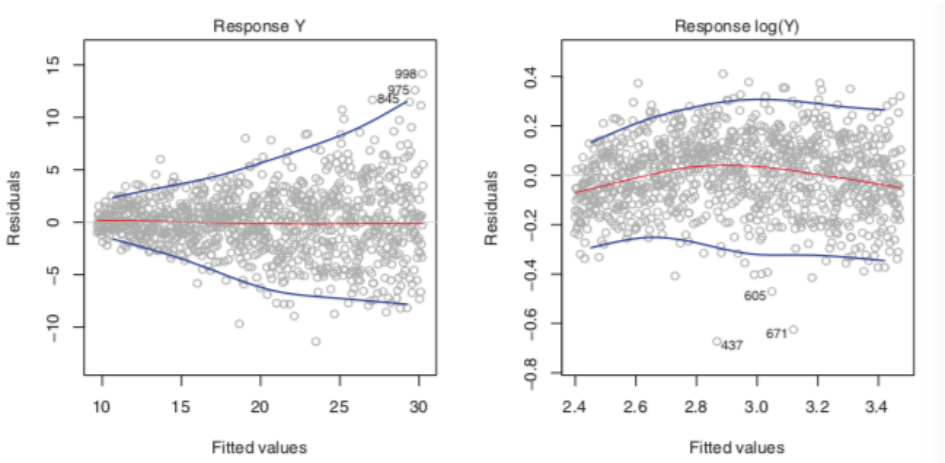

FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity.*

Ways to deal with heteroskedasticity is to log the response variable. Another option is to fit a weighted least squares.

Outliers Outlier is a point far beyond the value predicted by the model. An outlier may or may not affect a predictors slope and may also affect the RSE, which can affect confidence intervals and p-values, and can also affect the R^2.

Residual plots can be used to identify outliers or standardized residual (divide residuals by standard error.) plots,

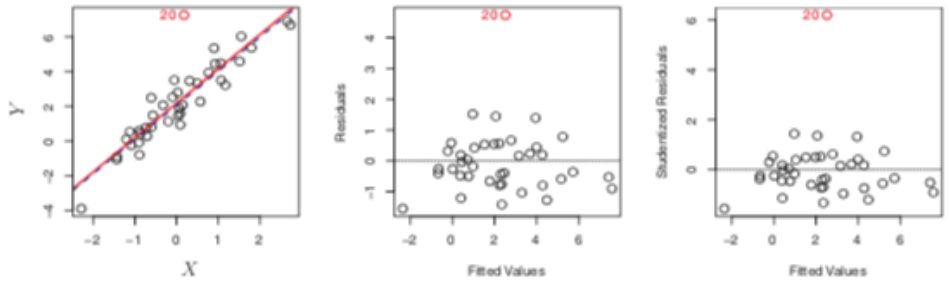

FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between −3 and 3.*

5. High Leverage Points High leverage points have an unusual value for x_i. These observations can heavily affect the least squares line.

These can be identified similar to outliers or through a leverage statistic, calculated as,

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

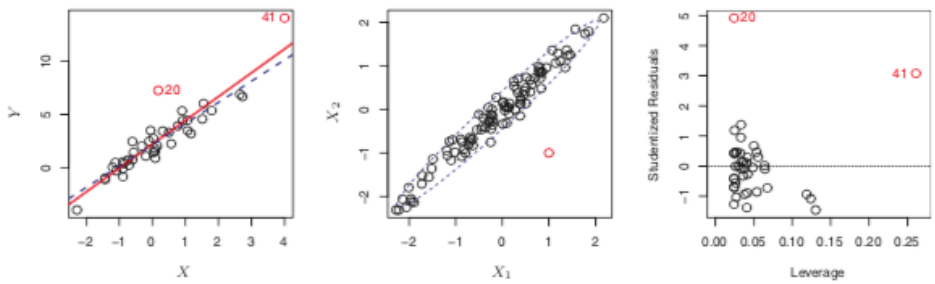

FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

6. Collinearity Collinearity occurs when two or more predictor variables are closely related to one another. High correlation relates to variables being collinear.

Collinearity introduces problems because the effects cannot be parsed out which can produce uncertainty around the coefficient estimates. Other problems exist, such as reduction in accuracy of coefficients causes standard errors to grow (due to calculation of t-stat and coefficient).

Ways to deal to collinearity include looking at correlation matrix of the predictors. A better way to assess multicollinearity (correlation of more than two variables) is to use variance inflation factor (VIF). The smallest possible value for VIF is 1, which indicates complete absence of collinearity. A VIF exceeds 5 or 10 indicates a problem.

VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ on its own.

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

To deal with collinearity, two solutions exist: (1) drop the problematic variables; (2) combine the collinear variables into a single predictor.

3.4 Comparison of Linear Regression with K-Nearest Neighbors

K-nearest neighbors regression (KNN regression) is one of the most well-known non-parametric regressions. KNN regressions first identify the K training observations that are closest to x_0. Then estimates $f(x_0)$ using the average of all the training responses in N_0,

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

Small K results in step-function that is most flexible of data while larger values smooth the plane and less flexible. The optimal value of K depends on the bias-variance tradeoff.

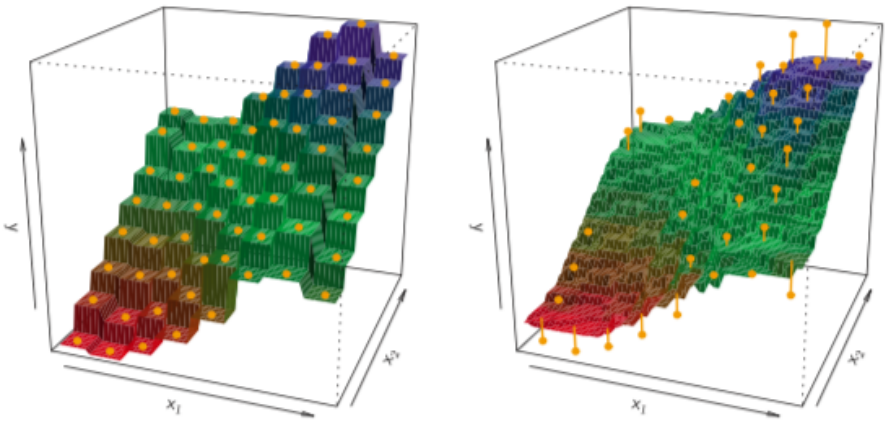

FIGURE 3.16. *Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots).* Left: $K = 1$ *results in a rough step function fit.* Right: $K = 9$ *produces a much smoother fit.*

Note: the parametric approach will outperform the non-parametric approach if the parametric form that has been selected is close to the true form of f.

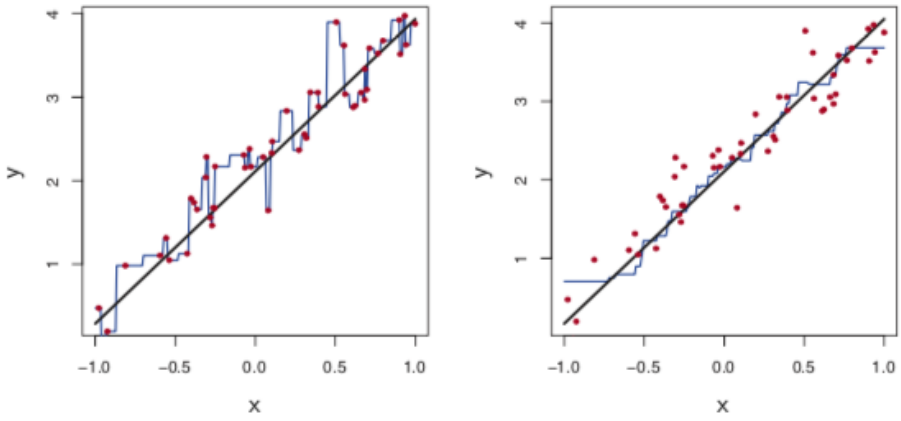

FIGURE 3.17. *Plots of $\hat{f}(X)$ using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line.* Left: *The blue curve corresponds to $K = 1$ and interpolates (i.e. passes directly through) the training data.* Right: *The blue curve corresponds to $K = 9$, and represents a smoother fit.*

Note: Generally, KNN regressions will outperform linear regressions with low number of variables. As the number of variables increase, KNN predictive power degrades (problem of dimensionality). As a general rule, parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor.

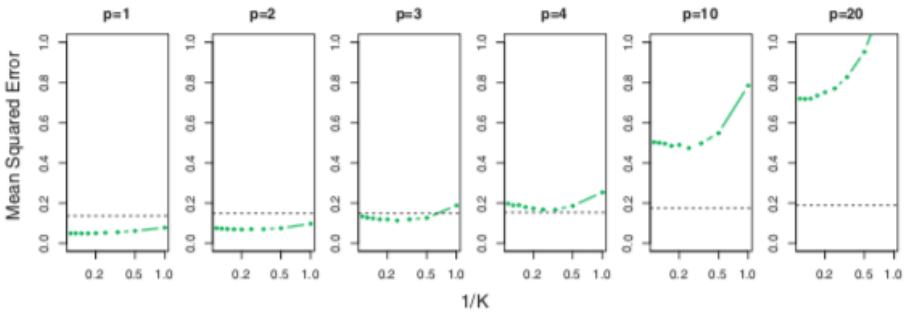

FIGURE 3.20. *Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non–linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.*

3.5 Classification

Classification problems invole dealing with categorical (qualitative) variables. These problems, generally, predict the probability of each category, so they behave like a regression problem.

Three most common classification problems: **logistic regression, linear discriminant analysis, and KNN**.

Why Not Linear Regression?: No natural way to convert a qualitative response variable with more than two levels into a quantitative response that is rady for linear regression, e.g. can't convert 1, 2, 3 to 1 to 3.

3.5.1 Logistic Regression

Logistic regression models the probability that Y belongs to a particular category, $p(X) = Pr(Y = 1|X)$. The general form is from a linear regression is,

$$p(X) = \beta_0 + \beta_1 X$$

However, the linear regression form includes a balance between negative values and positive values, which does not apply to probabilities.

The logistic function form is,

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1}}{1 + e^{\beta_0+\beta_1 X_1}}$$

To fit the model between zero and one, we use a maximum likelihood method (see next section). Solving for right-hand side, $\beta_0 + \beta_1 X$ equals,

$$\underbrace{log(\frac{p(X)}{1-p(X)})}_{\text{Log-odds or logit}} = \beta_0 + \beta_1 X$$

Thus, the logistic regression model has a logit that is linear in X. However, in a logistic regression, a one-unit increase in X changes the log odds by β_1.

3.5.2 Estimating the Regression Coefficients

General intuition behind maximum likelihood is to estimate β_0 and β_1 such that the predicted probability $\tilde{p}(x_i)$ of default for each individual from the logistic regression, corresponds as closely as possible to the individuals observed default status. In other words, we find coefficients that yields a number close to one for all individuals who defaulted and number close to zero for all individuals who did not. Formally, the likelihood function is,

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

β_0, β_1 are chosen to maximize the likelihood function.

Many aspects of logistic regression are similar to linear regression: measure accuracy of coefficients with standard errors, t-stats, null hypothesis testing. The intercept is generally not of interest and is used to fit probabilities to the proportion of ones in the data.

3.5.3 Making Prediction

Predictions are made from the simple logistic model,

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

3.5.4 Multiple Logistic Regression

Using multiple variables follows a similar approach to simple logistic regression,

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

As in the linear regression setting, the results obtained using one predictor may be quite different from those obtained using multiple predictors, especially when there is correlation among the predictors. In general, the phenomenon is known as confounding.

Logistic Regression for $>$ 2 Response Classes Multiple-class logistic regressions are available, but discriminant analysis is popular for multiple-class classification.

3.5.5 Linear Discriminant Analysis

Linear Discriminant Analysis involves modeling the distribution of the predictors X separately in each of the response classes, and then use Bayes theorem to flip those around into estimate for $Pr(Y = k|X = x)$.

Why chooce LDA?

- When classes are well separated in logistic regressions, the results are unstable. LDA does not suffer from this.
- If X predictors are approx normal and n is small, LDA are more stable
- Popular with two repsonse class

3.5.6 Using Bayes' Theorem for Classification

The Bayes' Theorem states,

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where π_k represents the overall, or prior, probability that a given obseration is associated with the kth category of the response variable Y. $f_k(x)$ denotes the density function of X for an observation that comes from the kth class.

Generally, estimating π_k is easy if we hae a random sample of Ys from the population (compute fraction of the training observationsthat belong to the kth class). However, estimating $f_k(x)$ is more challenging unless a form of density is assumed. The Bayes' classifier provides the lowest error rate, so if we can estimate $f_k(x)$ then we can get a way to classify Bayes.

3.5.7 Linear Discriminant Analysis for p=1

With only one predictor, assuming a normal or Gaussian, it is simlpy to estimate the normal density.

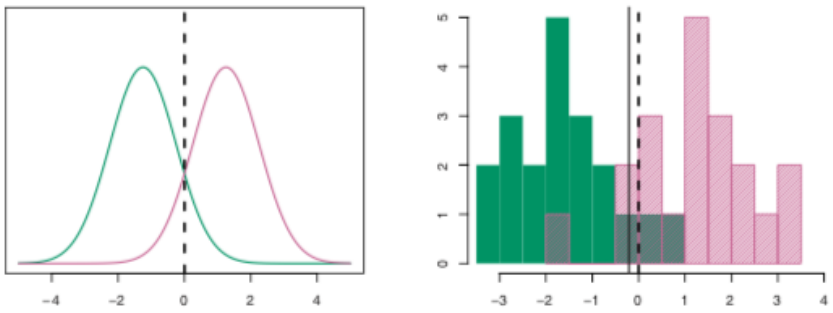

FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still have to estimate the parameters $\mu_1, ..., \mu_K, \pi_1, ..., \pi_K, and \sigma^2$. The linear discriminant analysis (LDA) method approximates the Bayes classifier by plugging estimates for $\pi_k, \mu_k, and \sigma^2$ as follows,

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

where n is total number of training observations, n_k is the number of training obs in the kth class. The estimate for μ_k is simply the average of all the training observations from the kth class, while

$\sigma\hat{}2$ can be seen as a weighted average of the sample variances for each of the K classes. We can estimate $\hat{\pi}_k$ as,

$$\hat{\pi}_k = n_k/n$$

The LDA clasifier is,

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k)$$

Note: the word linear in LDA comes from the fact the the discriminant function $\hat{\delta}_k(x)$ are linear functions of x.

To reiterate, the LDA classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance σ^2, and plugging estimates for these parameters into the Bayes classifier.

3.5.8 Linear Discriminant Analysis for p > 1

In the case of $p > 1$ predictors, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \sum)$, where μ_k is a class-specific mean vector, and $\sum$ is a covariance matrix that is common to all K classes. The multivariation density function is plugged into LDA.

Problem: binary classifiers, such as LDA, can make two types of errors: (1) can incorrectly assign an individual who defaults to the no default category, or (2) it can incorrectly assign an individual who does not default to the default category. The solution to this is a confusion matrix,

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

TABLE 4.4. *A confusion matrix compares the LDA predictions to the true default statuses for the 10,000 training observations in the* `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.*

which describes the number predicted correctly versus not to compare strength of the model.

While the Bayes' Classifier will provide lowest error rate, it doesn't always do a good job predicting because of the threshold for the posterior probability, default 50%. If concerned about incorrect predictions, it's best to lower the threshold, but lowering too much will cause increases in prediction error. Deciding on the threshold is dependent on domain knowledge.

The ROC cuvrve (receiver operating characteristics) is used to simultaneously display the two types of errors for all thresholds. The performance is based on the area under the curve (AUC) of the ROC. ROC curves are useful for comparing different classifiers.

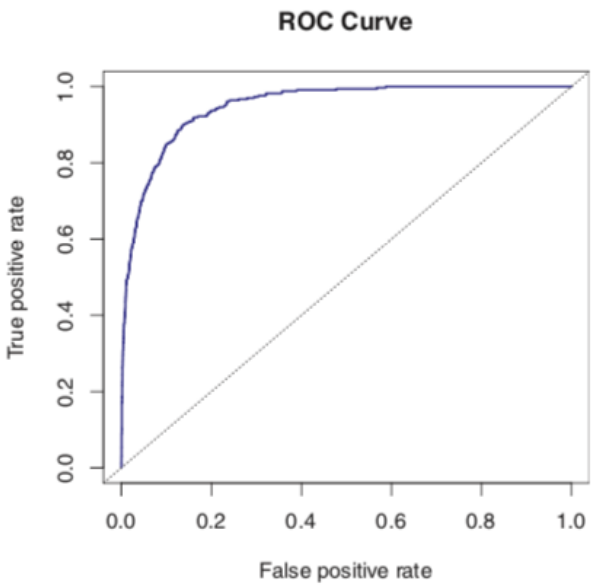

FIGURE 4.8. *A ROC curve for the LDA classifier on the* `Default` *data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value. The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate. The dotted line represents the "no information" classifier; this is what we would expect if student status and credit card balance are not associated with probability of default.*

3.5.9 Quadratic Discriminant Analysis (QDA)

Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to per- form prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix. Further, QDA assumes x is quandratic as opposed to linear.

Generally, LDA is more flexible with a lower variance, so model performance is improved over QDA. However, with larger training sets, QDA may perform better, so the variance is not of concern.

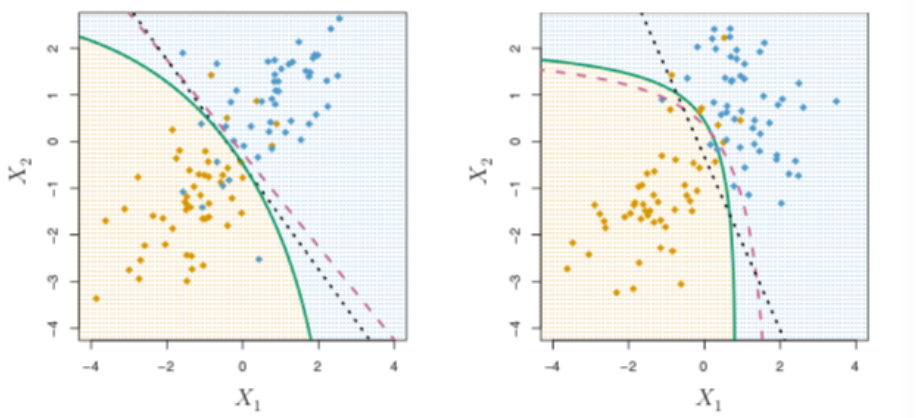

FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

3.6 Comparison of Classification Methods

LDA and logistic regressions are closely connected and differ only in their fitting procedure: logistic regression is estimating using maximum likelihood whereas LDA is estimated using mean and variance from a normal distribution.

LDA assumes normal distributions and common variances, so is an improvement over logistic; however, without those assumptions, logistic regressions will do better.

When the decision barrier is highly non-linear, KNN will perform better because of its non-parametric approach; however, KNN doesn't provide a table of cofficients to compare significance of variables.

QDA serves as a compromise between KNN, LDA, and logistic regressions.

In summary, when the true decision boundaries are linear, then the LDA and logistic regression approaches will tend to perform well. When the boundaries are moderately non-linear, QDA may give better results. Finally, for much more complicated decision boundaries, a non-parametric approach such as KNN can be superior.