STA 141A: Fundamentals of Statistical Data Science

Course Material Summary *Part 2*

University of California at Davis

Last Edit Date: 06/23/2022

Disclaimer and Term of Use:

1. We do not guarantee the accuracy and completeness of the summary content. Some of the course material may not be included, and some of the content in the summary may not be correct. You should use this file properly and legally. We are not responsible for any results from using this file.
2. Although most of the content in this summary is originally written by the creator, there is still some of the content that is adapted (derived) from the slides and codes from *Professor Emanuela Furfaro*. We use those as references and quotes in this file. Please contact us to delete this file if you think your rights have been violated.
3. This work is licensed under a Creative Commons Attribution 4.0 International License.

STA 141A Summary Part 2

I. Statistical Learning

Supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs.

With **unsupervised** statistical learning, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.

II. Linear Regression

1. Basics

Linear regression is a tool for predicting a quantitative response.

We assume $\boxed{\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}}$ or assume the following matrix:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}), \boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$$

Assumptions:

$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$

1) $E(\boldsymbol{\varepsilon}) = 0$

2) $Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$

*3) $\sigma^2 \boldsymbol{I} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$

The variance-covariance matrix tells the ε_i are independent and have constant variance σ^2. All covariance terms are zero, only diagonal is σ^2, which means they are uncorrelated. As a result, under normality, this indicating independence.

2. Least Square Estimation (OLS)

Residuals: $\boxed{\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X\beta}}$.

$\widehat{\boldsymbol{\beta}}: Q(\boldsymbol{\beta}) = \min_{\beta} (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta}) = \min_{\beta} (\boldsymbol{Y}^T \boldsymbol{Y} - 2\boldsymbol{\beta^T X^T Y} + \boldsymbol{\beta^T X^T X \beta})$

$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X^T Y} + \boldsymbol{2X^T X\beta} = 0 \Leftrightarrow \boldsymbol{X^T X \beta} = \boldsymbol{X^T Y} \Leftrightarrow \boxed{\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}}$

The minimizer is unique provided that the columns of $\boldsymbol{X}$ are linearly independent. It is not possible if the sample size n is smaller than the number of predictors $p + 1$. If the assumptions $E(\boldsymbol{\varepsilon}) = 0$ and $E(\boldsymbol{\varepsilon\varepsilon^T}) = \boldsymbol{\sigma^2 I}$ hold, then $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$.

The $\boldsymbol{MSE}$ $(\boldsymbol{s^2})$ is an unbiased estimator of σ^2, $\boxed{s^2 = \frac{(e)^2}{n-p-1}}$, where $n > p + 1$.

$\sqrt{MSE} = s$ is called the residual standard error (RSE).

3. Accuracy of Estimates

The standard error (SE) of an estimator of a parameter is the standard deviation of its sampling distribution. $Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X^T X})^{-1} \sigma^2$, $s(\widehat{\boldsymbol{\beta}}) = \sqrt{(\boldsymbol{X^T X})^{-1} \sigma^2}$.

4. Hypothesis Test

The aim of a linear model is often to help us determine whether there is a relationship between the response and the predictor.

Step 1: State null and alternative hypotheses

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$, H_1: At least one β_j is non-zero.

Step 2: Calculate F-statistic or t-statistic

$$\boxed{F^* = \frac{MSR}{MSE} = \frac{\frac{TSS-RSS}{p}}{\frac{RSS}{n-p-1}} \quad or \ \frac{(RSS_1 - RSS_2)/p}{RSS_2/(n-p-1)} \sim F(n-p-1)}$$

$$\boxed{t^* = \frac{\widehat{\beta} - c}{s(\widehat{\beta})} \sim t(n-p-1)}$$

$SSTO = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 + n\bar{Y}^2 = \boldsymbol{Y^T Y} - n\bar{Y}^2$

$SSE = \sum (Y_i - \hat{Y}_i)^2 = \boldsymbol{Y^T Y} - \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X^T Y}$

$SSR = \sum (\hat{Y}_i - \bar{Y})^2 = SSTO - SSE = \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X^T Y} - n\bar{Y}^2$

$MSE = \frac{SSE}{df(SSE)} = \frac{SSE}{n-p}$ (unbiased estimator of σ^2)

$MSR = \frac{SSR}{df(SSR)}$ $\quad MSTO = \frac{SSTO}{df(SSTO)}$

Step 3: Calculate $p - value$.

Using R or calculator.

Step 4: When $p - value < \alpha$,

reject null hypothesis.

R code approach:

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

formula: an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

data: an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which lm is called.

5. Residual Standard Error (RSE)

$\boxed{RSE = \sqrt{MSE} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}}$ is the average amount that the response will deviate from the true regression line. If the predictions using the model are very close to the true outcome values, then the RSE will be small, and we can conclude the model fits the data very well, vice versa. RSE uses units of Y.

6. Coefficient of Determination

1) $\boxed{R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \ or\ 1 - \frac{RSE}{TSS}}$

2) $\boxed{R_{adj}{}^2 = \frac{MSR}{MSTO} = 1 - \frac{MSE}{MSTO} \ or\ 1 - \frac{\frac{RSE}{n-p-1}}{\frac{TSS}{n-1}}}$

3) Interpretation: We can say R^2% of the variability in Y can be explained by X.

4) R^2 is at least as large as $R_{adj}{}^2$. 5) $0 \le R^2 \le 1$ 6) $r^2 = [corr(X,Y)]^2$

7) R^2 is unit free (does not depend on X or Y)

8) $\boldsymbol{R^2 = 1}$ means perfect linear association between X and Y ($R^2 = 1 \Leftrightarrow SSE=0 \Leftrightarrow \sum (Y_i - \hat{Y}_i)^2 = 0 \Leftrightarrow Y_i = \hat{Y}_i$ for all i)

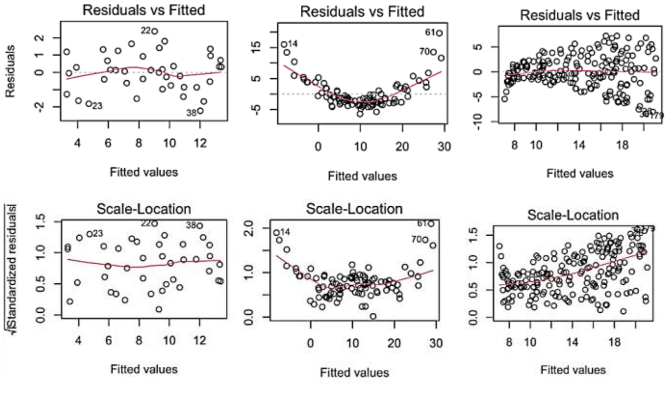
9) $\boldsymbol{R^2 = 0}$ means no linear association between X and Y. However, a nonlinear association may exist. ($R^2 = 0 \Leftrightarrow SSR=0 \Leftrightarrow \sum (\hat{Y}_i - \bar{Y})^2 = 0 \Leftrightarrow \hat{Y}_i = \bar{Y}$ for all i)

7. Non-linearity of The Response-predictor Relationships

The linear regression model assumes that there is a straight-line relationship between the predictors and the response.

If there are non-linear associations in the data, a simple approach is to use **non-linear transformations** of the predictors, such as $\log X$, $\sqrt{X}, X^2$, in the regression model.

Plot approach:



No apparent problems   No linearity   Heteroskedasticity

8. Correlation of Error Terms

The errors are assumed to be **uncorrelated**. If the errors are correlated, the estimated standard errors will tend to **underestimate** the true standard errors. The p-values associated with the model will be **lower** than they should be. Such correlations frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time.

Use Ljung–Box test or Durbin-Watson test. R code as following:

```
Box.test(model$residuals)
dwt(model)
```

When p-value is large or greater than α, conclude **no correlation**.

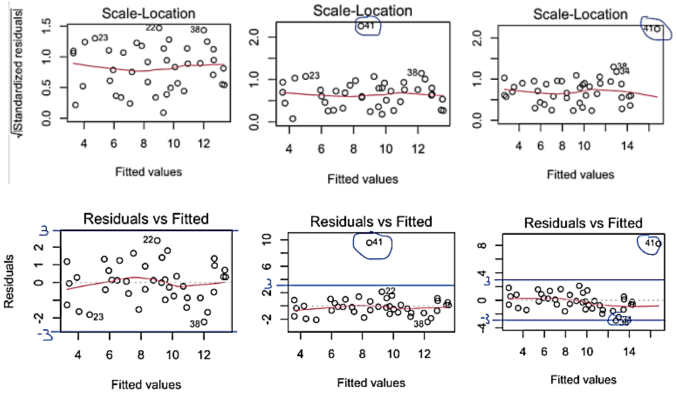
9. Non-constant Variance of Errors Terms

The errors are assumed to be **homoscedasticity**. The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption. One can identify non-constant variances in the errors, or heteroscedasticity, from the presence of a funnel shape in the residual plot. One possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$.

10. Outliers

An outlier is a point for which Y_i is **far from the value** predicted by the model. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection. Typically, an outlier that does not have an unusual predictor value have little effect on the least squares fit. Outliers might inflate the RSE. The RSE is used to compute all confidence intervals and p-values.

Plot approach:



No apparent problems   Outlier   Outlier

R code approach:

```
rstandard(model)[rstandard(model) < -3 |
rstandard(model) > 3]
```

11. High Leverage Points

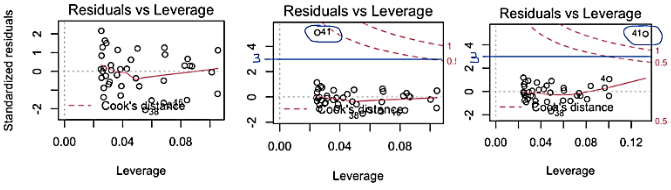
Outliers are observations for which the response Y_i is unusual given the predictor X_i. Observations with high leverage have an **unusual** value for X_i. Any problems with these points may invalidate the entire fit.

Leverage statistic for the i-th observation is defined as following:

$$\boxed{h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i'=1}^{n}(X_{i'} - \bar{X})^2}}.$$

The leverage statistic h_i is always between $\frac{1}{n}$ and 1. The average leverage for all the observations is always equal to $\frac{p+1}{n}$. If a given observation has a leverage statistic that greatly **exceeds** $\frac{p+1}{n}$, then we may suspect that the corresponding point has **high leverage**.

Plot approach:



No apparent problems   Outlier   Outlier and leverage and influential point

R code approach:

```
hatvalues(model)[hatvalues(model) > (p + 1) / n]
```

12. Collinearity

Collinearity refers to the situation in which two or more predictor variables are **closely related to each other**.

This results in a great deal of uncertainty in the coefficient estimates. Since collinearity **reduces the accuracy** of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to **grow**.

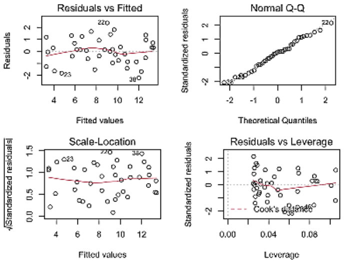
It is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation. We call this situation **multicollinearity**.

The variance inflation factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis. Formula is as following:

$\boxed{VIF(\hat{\beta}_j) = \frac{1}{1-R^2_{X_j|X_{-j}}}}$, where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X_{-j}}$ is **close to one**, then collinearity is **present**, and so the VIF will be **large**. VIF value that **exceeds 5 or 10** indicates a problematic amount of collinearity.

We can solve the collinearity problem by **(1)** dropping one of the problematic variables from the regression; **(2)** combining the colinear variables together into a single predictor.

13. Diagnostic (Residual) Plot



Residuals vs. Fitted

(1) Linearity; (2) Heteroskedasticity; (3) **Suspected** outliers and leverage points.

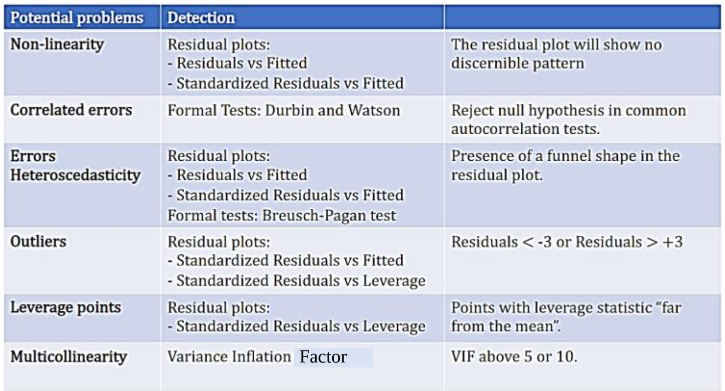
Standardized Residuals vs. Fitted

Outliers and **suspected** leverage points.

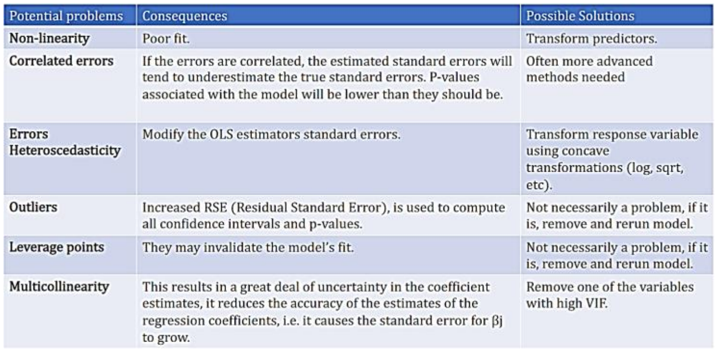
Standardized Residuals vs. Leverage

Outliers and leverage points.

14. Linear regression: detecting potential problems



Potential problems	Detection	
Non-linearity	Residual plots: - Residuals vs Fitted - Standardized Residuals vs Fitted	The residual plot will show no discernible pattern
Correlated errors	Formal Tests: Durbin and Watson	Reject null hypothesis in common autocorrelation tests.
Errors Heteroscedasticity	Residual plots: - Residuals vs Fitted - Standardized Residuals vs Fitted Formal tests: Breusch-Pagan test	Presence of a funnel shape in the residual plot.
Outliers	Residual plots: - Standardized Residuals vs Fitted - Standardized Residuals vs Leverage	Residuals < -3 or Residuals > +3
Leverage points	Residual plots: - Standardized Residuals vs Leverage	Points with leverage statistic "far from the mean".
Multicollinearity	Variance Inflation Factor	VIF above 5 or 10.



Potential problems	Consequences	Possible Solutions
Non-linearity	Poor fit.	Transform predictors.
Correlated errors	If the errors are correlated, the estimated standard errors will tend to underestimate the true standard errors. P-values associated with the model will be lower than they should be.	Often more advanced methods needed
Errors Heteroscedasticity	Modify the OLS estimators standard errors.	Transform response variable using concave transformations (log, sqrt, etc).
Outliers	Increased RSE (Residual Standard Error), is used to compute all confidence intervals and p-values.	Not necessarily a problem, if it is, remove and rerun model.
Leverage points	They may invalidate the model's fit.	Not necessarily a problem, if it is, remove and rerun model.
Multicollinearity	This results in a great deal of uncertainty in the coefficient estimates, it reduces the accuracy of the estimates of the regression coefficients, i.e. it causes the standard error for βj to grow.	Remove one of the variables with high VIF.

III. Logistic Regression

1. Basics

We can use logistic regression when the variable to be predicted is qualitative (categorical).

We will first consider a situation where the response variable is binary, which means $Y \in \{1,0\}$. Rather than modeling the response Y directly, we will try to model the probability $p(X) = \Pr(Y = 1|X)$.

In logistic regression, we use the logistic function: $f(x) = \frac{e^x}{1+e^x}$, where $x \in \mathbb{R}$.

Then we have $\boxed{p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$ and this guarantees that predicted probability is in $(0, 1)$ and hence we will obtain a sensible prediction.

R code approach:

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)
```

formula: an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

family: a description of the error distribution and link function to be used in the model. For glm this can be a character string naming a family function, a family function or the result of a call to a family function. For glm.fit only the third option is supported. (See family for details of family functions.)

data: an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which glm is called.

2. Odds

We have that $\boxed{\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}}$. The quantity $\frac{p(X)}{1-p(X)}$ is called the odds and can take on any value in $(0, \infty)$.

3. Log-odds (logit)

By taking the logarithm of both sides, we arrive at $\boxed{\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X}$. The left-hand side is called the log-odds or logit.

The amount that $p(X)$ changes due to a one-unit change in X will depend on the current value of X.

If β_1 is positive, then increasing X will be associated with increasing $p(X)$, and if β_1 is negative then increasing X will be associated with decreasing $p(X)$.

4. Estimation of parameters

The preferred method to estimate the parameters of the logistic regression is the method of maximum likelihood (MLE).

The maximum likelihood estimators of the logistic regression are the maximizers of the likelihood function $l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0}(1 - p(x_i))$.

A closed form solution does not exist, an iterative numerical optimization method (for example, gradient descent or Newton's method) is needed to find the estimates of the parameters.

5. Deviance

The deviance residuals: $\boxed{d_i = \sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right]}}$, with same sign as the raw residual $y_i - \hat{p}_i$.

The residual deviance is defined by $\sum_{i=1}^{n} d_i^2$. The null deviance is calculated for the model with no slope.

6. Akaike Information Criterion (AIC)

Akaike information criterion (AIC) estimates the relative amount of information lost by a given model: the **less** information a model loses, the **higher** the quality of that model.

Useful when comparing several models. Given a set of candidate models for the data, the preferred model is the one with the **minimum AIC value**.

7. Multiple Logistic Regression

Instead of having just one predictor, we might try to use p predictors to model the log odds: $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.

Then we have that $\boxed{p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}}$. We use the method of maximum likelihood to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$.

8. Logistic Regression as a Binary Classifier

Logistic regression allows to estimate the probability of belonging to a specific class (category, label) given a number of predictors.

We can use this probability to predict the class (category) of a unit, i.e. to actually classify the unit in a class.

Therefore, we can use logistic regression as a classifier, in particular as a binary classified since Y is a **binary** variable.

In order to do that we set a threshold, for instance 0.5, and we classify the unit as **'1'** if the estimated probability is **above 0.5**.

Regression is the problem of predicting a continuous outcome such as price, salary, BMI, QI, etc.

Classification refers to the problem of predicting / classifying a qualitative variable values such as Male or Female, True or False, Spam or Not Spam.

9. Errors in Binary Classifiers

It can incorrectly assign an individual of **category 1 to category 0 (False Negative).** It can incorrectly assign an individual of **category 0 to category 1 (False Positive)**.

True / Predicted	0	1	
0	True Negative	**False Positive**	N
1	**False Negative**	True Positive	P
	N^*	T^*	Total

Accuracy = (True Positive + True Negative) / Total

- The **accuracy** is simply the proportion of correctly classified units.

Error Rate = (False Positive + False Negative) / Total

- The **error rate** is simply the proportion of incorrectly classified units.

False Positive Rate = False Positive / (False Positive + True Negative)

- The **false positive rate** is the proportion of incorrectly classified units among those with true label 0 (i.e. the proportion of negative units classified as positive).

False Negative Rate = False Negative / (False Negative + True Positive)

- The **false negative rate** is the proportion of misclassified units among those with true label 1 (i.e. the proportion of positive units classified as negative).

Sensitivity = 1 - False Negative Rate

The probability of positive units classified as positive.

Specificity = 1 - False Positive Rate

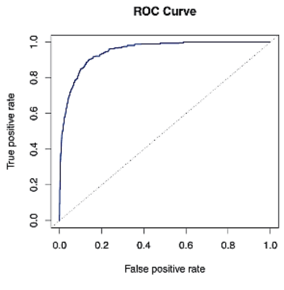
The probability of negative units classified as negative.

10. ROC curve

The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the (ROC) curve (AUC).

An ideal ROC curve will hug the top left corner, so the larger area under the (ROC) curve the better the classifier.

We expect a classifier that performs no better than change to have an AUC of 0.5.



III. Cross Validation

1. Basic

Cross-validation is a class of methods that estimate the test error rate by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

2. Validation set approach

We estimate the test error associated with fitting a particular model on a set of observations.

The training sample is split into two parts: a training set and a validation set or hold-out set.

The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.

The resulting validation set error rate–typically assessed using MSE in the case of a quantitative response–provides an estimate of the test error rate.

The validation set approach has two potential **drawbacks**: (1) the validation estimate of the test error rate can be highly variable; (2) only a relatively small subset of the observations are used to fit the model.

3. Leave-one-out cross-validation (LOOCV)

For $i = 1, \dots, n$, fit the model using the training set $\{(x_1, x_2), \dots, (x_n, y_n)\}$; compute $\hat{y}_i$ using the fitted model; $MSE_i = (y_i - \hat{y}_i)^2$. The LOOCV estimate for the test MSE is the average of these n test error estimates: $\boxed{CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i}$.

The following formula holds for the least squares linear or polynomial regression $\boxed{CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2}$, where $\hat{y}_i$ is the i^{th} fitted value from the original least squares fit, and h_i is the leverage between $\frac{1}{n}$ and 1.

Advantages: (1) far less bias; (2) performing LOOCV multiple times will always yield the same result.

Disadvantages: expensive to implement.

4. k-fold cross validation

k-fold cross-validation works in the following way: (1) split the set of observations into k groups (or folds); (2) treat the first fold as a validation set and the remaining folds as the training set; (3) compute MSE_1 on the observations in the held-out fold; (4) repeat the k-fold CV estimate using the formula $CV_{(k)} = \frac{1}{k}\sum_{i=1}^{n} MSE_i$.

Notice: LOOCV is a special case of k-fold CV in which k is set to equal n.

IV. K Nearest Neighbor (k-NN)

1. Basics

Suppose that we want to classify an observation x. We first identify the k points in the training data that are closest to x (the set is denoted by N_0). Then estimate the conditional probability for class j as the fraction of points in N_0 whose response values equal to j.

$\boxed{P(Y = j | X = x) = \frac{1}{k}\sum_{i \in N_o} I(y_i = j)}$, where $j = 1, 2, \dots, k$. Assign the test observation x to the class with the largest probability.

Notice, the performance of the algorithm depends strongly on the parameter k. We can use cross-validation to select the value of the parameter k.

2. Steps to do k-NN with a training set and a testing set

Step 1: Calculate the Euclidean distance between $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$.

$\boxed{||x - y||_2 = \sqrt{(x_1 - y_1)^2 + \dots + (x_d - y_d)^2}}$. If one of the features has a broad range of values, the distance will be governed by this feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Sometimes we may need to **normalize** x. The min-max normalization: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$, where x is the original value and $x' \in [0,1]$ is the normalized value.

Sometimes we may also need to **standardize** x. $x' = \frac{x - \bar{x}}{S}$, where x is the original value, $\bar{x}$ is the sample mean and S is the standard deviation of the feature vector.

Step 2: Set up the value of k.

Step 3: Find the k smallest distances from the training data to the point we choose in testing data.

Step 4: Calculate the conditional probability: $P(Y = j | X = x) = \frac{1}{k}\sum_{i \in N_o} I(y_i = j)$.

Step 5: Generate a new categorical variable, and calculate confusion matrix.

V. Linear Discriminant Analysis (LDA)

1. Basics

Logistic regression involves modeling $P(Y = 1 | X = x)$, we model the conditional distribution of the response Y, given the predictor(s) X.

We now model the distribution of the predictors X **separately** in each of the response classes (Y), and then use Baye's theorem to flip these around into estimates for $P(Y = k | X = x)$.

The model will be very similar to the logistic regression when the distributions are assumed to be normal.

2. Why use LDA

If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is more stable than the logistic regression model.

Linear discriminant analysis is popular when we have more than two response classes.