STATISTICAL LEARNING

Modeling Problems

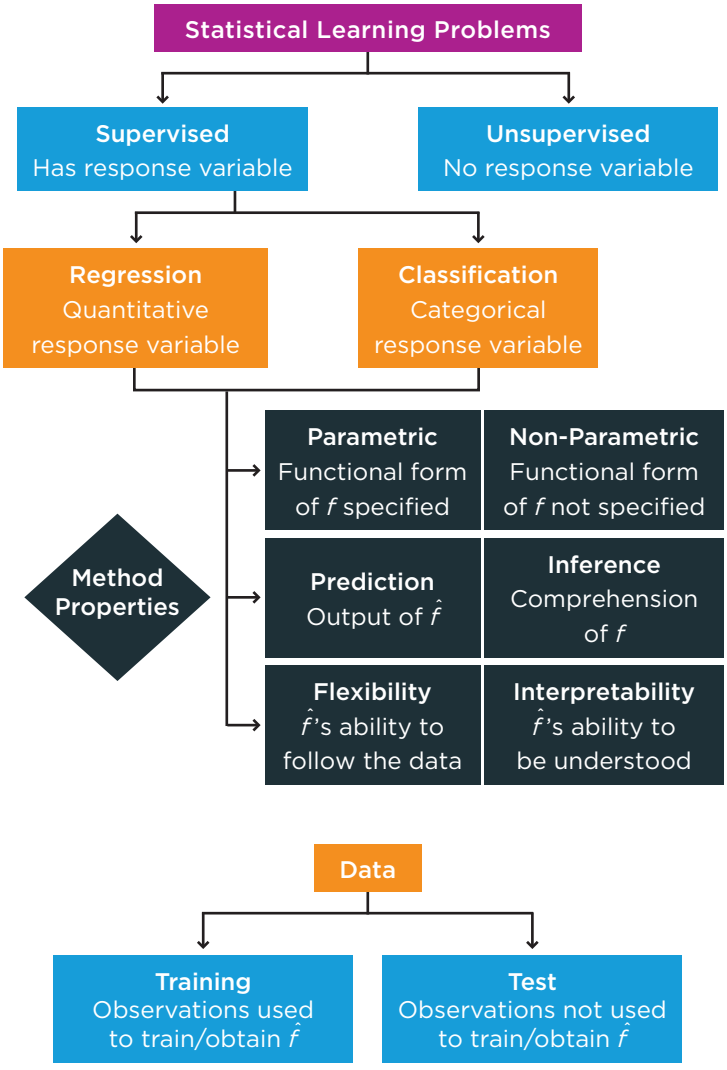
Types of Variables

Response	A variable of primary interest
Explanatory	A variable used to study the response variable
Count	A quantitative variable usually valid on non-negative integers
Continuous	A real-valued quantitative variable
Nominal	A categorical/qualitative variable having categories without a meaningful or logical order
Ordinal	A categorical/qualitative variable having categories with a meaningful or logical order

Notation

y, Y	Response variable
x, X	Explanatory variable
Subscript i	Index for observations
n	No. of observations
Subscript j	Index for variables except response
p	No. of variables except response
$\mathbf{A}^T$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$
ε	Error term
$\hat{y}, \hat{Y}, \hat{f}(x)$	Estimate/Estimator of $f(x)$

Contrasting Statistical Learning Elements



Regression Problems

$Y = f(x_1, \dots, x_p) + \varepsilon$ where $\mathrm{E}[\varepsilon] = 0$, so $\mathrm{E}[Y] = f(x_1, \dots, x_p)$

Test MSE $= \mathrm{E}\left[(Y - \hat{Y})^2\right]$,

which can be estimated using $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

For fixed inputs $x_1, \dots, x_p$, the test MSE is

$$\underbrace{\mathrm{Var}\left[\hat{f}(x_1, \dots, x_p)\right] + \left(\mathrm{Bias}\left[\hat{f}(x_1, \dots, x_p)\right]\right)^2}_{\text{reducible error}} + \underbrace{\mathrm{Var}[\varepsilon]}_{\text{irreducible error}}$$

Classification Problems

Test Error Rate $= \mathrm{E}\left[I(Y \neq \hat{Y})\right]$,

which can be estimated using $\frac{\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)}{n}$

Bayes Classifier:

$$f(x_1, \dots, x_p) = \arg\max_c \Pr(Y = c | X_1 = x_1, \dots, X_p = x_p)$$

Key Ideas

- The disadvantage to parametric methods is the danger of choosing a form for f that is not close to the truth.
- The disadvantage to non-parametric methods is the need for an abundance of observations.
- Flexibility and interpretability are typically at odds.
- As flexibility increases, the training MSE (or error rate) decreases, but the test MSE (or error rate) follows a u-shaped pattern.
- Low flexibility leads to a method with low variance and high bias; high flexibility leads to a method with high variance and low bias.

Descriptive Data Analysis

Numerical Summaries

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

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

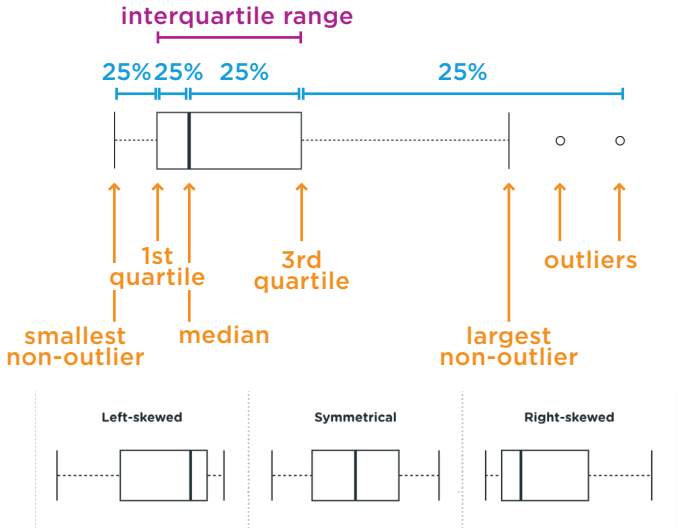
$$r_{x,y} = \frac{cov_{x,y}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad -1 \leq r_{x,y} \leq 1$$

Scatterplots

Plots values of two variables to investigate their relationship.

Box Plots

Captures a variable's distribution using its median, 1st and 3rd quartiles, and distribution tails.



qq Plots

Plots sample quantiles against theoretical quantiles to determine whether the sample and theoretical distributions have similar shapes.

LINEAR MODELS

Simple Linear Regression (SLR)

Special case of MLR where $p = 1$

Estimation

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

SLR Inferences

Standard Errors

$$se_{b_0} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$se_{b_1} = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$se_{\hat{y}} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$se_{\hat{y}_{n+1}} = \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

Multiple Linear Regression (MLR)

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Notation

β_j	The j^{th} regression coefficient
b_j	Estimate of β_j
σ^2	Variance of response / Irreducible error
MSE	Estimate of σ^2
$\mathbf{X}$	Design matrix
$\mathbf{H}$	Hat matrix
e	Residual
SST	Total sum of squares
SSR	Regression sum of squares
SSE	Error sum of squares

Assumptions

1. $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$
2. x_i's are non-random
3. $\text{E}[\varepsilon_i] = 0$
4. $\text{Var}[\varepsilon_i] = \sigma^2$
5. ε_i's are independent
6. ε_i's are normally distributed
7. The predictor x_j is not a linear combination of the other p predictors, for $j = 0, 1, \ldots, p$

Estimation – Ordinary Least Squares (OLS)

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_p x_p$$

$$\begin{bmatrix} b_0 \\ \vdots \\ b_p \end{bmatrix} = \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\text{MSE} = \text{SSE}/(n - p - 1)$$

$$\text{residual standard error} = \sqrt{\text{MSE}}$$

Other Numerical Results

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

$$e = y - \hat{y}$$

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{total variability}$$

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \text{explained}$$

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \text{unexplained}$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$R^2 = \text{SSR}/\text{SST}$$

$$R^2_{\text{adj.}} = 1 - \frac{\text{MSE}}{s_y^2} = 1 - (1 - R^2)\left(\frac{n - 1}{n - p - 1}\right)$$

Key Ideas

- R^2 is a poor measure for model comparison because it will increase simply by adding more predictors to a model.
- Polynomials do not change consistently by unit increases of its variable, i.e. no constant slope.
- Only $w - 1$ dummy variables are needed to represent w classes of a categorical predictor; one of the classes acts as a baseline.
- In effect, dummy variables define a distinct intercept for each class. Without the interaction between a dummy variable and a predictor, the dummy variable cannot additionally affect that predictor's regression coefficient.

MLR Inferences

Notation

$\hat{\beta}_j$	Estimator for β_j
$\hat{Y}$	Estimator for $\text{E}[Y]$
se	Estimated standard error
H_0	Null hypothesis
H_1	Alternative hypothesis
df	Degrees of freedom
$t_{1-q,\text{df}}$	q quantile of a t-distribution
α	Significance level
k	Confidence level
ndf	Numerator degrees of freedom
ddf	Denominator degrees of freedom
$F_{1-q,\text{ndf,ddf}}$	q quantile of an F-distribution
y_{n+1}	Response of new observation
Subscript r	Reduced model
Subscript f	Full model

Standard Errors

$$se_{b_j} = \sqrt{\widehat{\text{Var}}[\hat{\beta}_j]}$$

Variance-Covariance Matrix

$$\widehat{\text{Var}}[\hat{\boldsymbol{\beta}}] = \text{MSE}(\mathbf{X}^T\mathbf{X})^{-1} =$$

$$\begin{bmatrix} \widehat{\text{Var}}[\hat{\beta}_0] & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \cdots & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_p] \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \widehat{\text{Var}}[\hat{\beta}_1] & \cdots & \widehat{\text{Cov}}[\hat{\beta}_1, \hat{\beta}_p] \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_p] & \widehat{\text{Cov}}[\hat{\beta}_1, \hat{\beta}_p] & \cdots & \widehat{\text{Var}}[\hat{\beta}_p] \end{bmatrix}$$

t Tests

$$t \text{ statistic} = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

Test Type	*Rejection Region*
Two-tailed	$\lvert t \text{ statistic}\rvert \geq t_{\alpha/2,n-p-1}$
Left-tailed	$t \text{ statistic} \leq -t_{\alpha,n-p-1}$
Right-tailed	$t \text{ statistic} \geq t_{\alpha,n-p-1}$

F Tests

$$F \text{ statistic} = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/p}{\text{SSE}/(n - p - 1)}$$

Reject H_0 if F statistic $\geq F_{\alpha,\text{ndf,ddf}}$

- $\text{ndf} = p$
- $\text{ddf} = n - p - 1$

Partial F Tests

F statistic $= \dfrac{(\text{SSE}_r - \text{SSE}_f)/(p_f - p_r)}{\text{SSE}_f/(n - p_f - 1)}$

Reject H_0 if F statistic $\geq F_{\alpha,\text{ndf},\text{ddf}}$

- ndf $= p_f - p_r$
- ddf $= n - p_f - 1$

For all hypothesis tests, reject H_0 if p-value $\leq \alpha$.

Confidence and Prediction Intervals

estimate $\pm$ (t quantile)(standard error)

Quantity	*Interval Expression*
β_j	$b_j \pm t_{(1-k)/2,n-p-1} \cdot se_{b_j}$
$\text{E}[Y]$	$\hat{y} \pm t_{(1-k)/2,n-p-1} \cdot se_{\hat{y}}$
y_{n+1}	$\hat{y}_{n+1} \pm t_{(1-k)/2,n-p-1} \cdot se_{\hat{y}_{n+1}}$

Linear Model Assumptions

Leverage

$h_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \dfrac{se_{\hat{y}_i}{}^2}{\text{MSE}}$

$h_i = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{\sum_{u=1}^{n}(x_u - \bar{x})^2}$ for SLR

- $1/n \leq h_i \leq 1$
- $\sum_{i=1}^{n} h_i = p + 1$

Cook's Distance

$D_i = \dfrac{\sum_{u=1}^{n}(\hat{y}_u - \hat{y}_{(i)u})^2}{\text{MSE}(p+1)}$

$= \dfrac{e_i^2 h_i}{\text{MSE}(p+1)(1-h_i)^2}$

Plots of Residuals

- e versus $\hat{y}$
  Residuals are well-behaved if
  - Points appear to be randomly scattered
  - Residuals seem to average to 0
  - Spread of residuals does not change
- e versus i
  Detects dependence of error terms
- qq plot of e

Variance Inflation Factor

$\text{VIF}_j = \dfrac{1}{1 - R_j^2} = \dfrac{s_{x_j}^2(n-1)}{\text{MSE}} se_{b_j}^2$

Tolerance is the reciprocal of VIF.

Key Ideas

- As realizations of a t-distribution, studentized residuals can help identify outliers.
- When residuals have a larger spread for larger predictions, one solution is to transform the response variable with a concave function.
- There is no universal approach to handling multicollinearity; it is even possible to accept it, such as when there is a suppressor variable. On the other hand, it can be eliminated by using a set of orthogonal predictors.

Model Selection

Notation

g	Total no. of predictors in consideration
p	No. of predictors for a specific model
MSE_g	MSE of the model that uses all g predictors
M_p	The "best" model with p predictors

Best Subset Selection

1. For $p = 0, 1, \ldots, g$, fit all $\binom{g}{p}$ models with p predictors. The model with the largest R^2 is M_p.
2. Choose the best model among $\text{M}_0, \ldots, \text{M}_g$ using a selection criterion of choice.

Forward Stepwise Selection

1. Fit all g simple linear regression models. The model with the largest R^2 is M_1.
2. For $p = 2, \ldots, g$, fit the models that add one of the remaining predictors to M_{p-1}. The model with the largest R^2 is M_p.
3. Choose the best model among $\text{M}_0, \ldots, \text{M}_g$ using a selection criterion of choice.

Backward Stepwise Selection

1. Fit the model with all g predictors, M_g.
2. For $p = g - 1, \ldots, 1$, fit the models that drop one of the predictors from M_{p+1}. The model with the largest R^2 is M_p.
3. Choose the best model among $\text{M}_0, \ldots, \text{M}_g$ using a selection criterion of choice.

Selection Criteria

- Mallows' C_p

  $C_p = \dfrac{\text{SSE} + 2p \cdot \text{MSE}_g}{n}$

  $C_p = \dfrac{\text{SSE}}{\text{MSE}_g} + 2p - n$
- Akaike information criterion

  $\text{AIC} = \dfrac{\text{SSE} + 2p \cdot \text{MSE}_g}{n \cdot \text{MSE}_g}$
- Bayesian information criterion

  $\text{BIC} = \dfrac{\text{SSE} + \ln n \cdot p \cdot \text{MSE}_g}{n \cdot \text{MSE}_g}$
- Adjusted R^2
- Cross-validation error

Validation Set

- Randomly splits all available observations into two groups: the training set and the validation set.
- Only the observations in the training set are used to attain the fitted model, and those in validation set are used to estimate the test MSE.

k-fold Cross-Validation

1. Randomly divide all available observations into k folds.
2. For $v = 1, \ldots, k$, obtain the v^{th} fit by training with all observations except those in the v^{th} fold.
3. For $v = 1, \ldots, k$, use $\hat{y}$ from the v^{th} fit to calculate a test MSE estimate with observations in the v^{th} fold.
4. To calculate CV error, average the k test MSE estimates in the previous step.

Leave-one-out Cross-Validation (LOOCV)

- Calculate LOOCV error as a special case of k-fold cross-validation where $k = n$.
- For MLR:

  $\text{LOOCV Error} = \dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{y_i - \hat{y}_i}{1 - h_i}\right)^2$

Key Ideas on Cross-Validation

- The validation set approach has unstable results and will tend to overestimate the test MSE. The two other approaches mitigate these issues.
- With respect to bias, LOOCV < k-fold CV < Validation Set.
- With respect to variance, LOOCV > k-fold CV > Validation Set.

Other Regression Approaches

Standardizing Variables

- A centered variable is the result of subtracting the sample mean from a variable.
- A scaled variable is the result of dividing a variable by its sample standard deviation.
- A standardized variable is the result of first centering a variable, then scaling it.

Ridge Regression

Coefficients are estimated by minimizing the SSE while constrained by $\sum_{j=1}^{p} b_j^2 \leq a$ or equivalently, by minimizing the expression SSE $+ \lambda \sum_{j=1}^{p} b_j^2$.

Lasso Regression

Coefficients are estimated by minimizing the SSE while constrained by $\sum_{j=1}^{p} |b_j| \leq a$ or equivalently, by minimizing the expression SSE $+ \lambda \sum_{j=1}^{p} |b_j|$.

Key Ideas on Ridge and Lasso

- $x_1, \ldots, x_p$ are scaled predictors.
- λ is inversely related to flexibility.
- With a finite λ, none of the ridge estimates will equal 0, but the lasso estimates could equal 0.

Weighted Least Squares

- $\text{Var}[\varepsilon_i] = \sigma^2 / w_i$
- Equivalent to running OLS with $\sqrt{w}y$ as the response and $\sqrt{w}\mathbf{x}$ as the predictors, hence minimizing $\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$.
- $\mathbf{b} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$ where $\mathbf{W}$ is the diagonal matrix of the weights.

Partial Least Squares

- The first partial least squares direction, z_1, is a linear combination of standardized predictors $x_1, \ldots, x_p$, with coefficients based on the relation between x_j and y.
- Every subsequent partial least squares direction is calculated iteratively as a linear combination of "updated predictors" which are the residuals of fits with the "previous predictors" explained by the previous direction.
- The directions $z_1, \ldots, z_g$ are used as predictors in a multiple linear regression. The number of directions, g, is a measure of flexibility.

k-Nearest Neighbors (KNN)

1. Identify the "center of the neighborhood", i.e. the location of an observation with inputs $x_1, \ldots, x_p$.
2. Starting from the "center of the neighborhood", identify the k nearest training observations.
3. For classification, $\hat{y}$ is the most frequent category among the k observations; for regression, $\hat{y}$ is the average of the response among the k observations.

k is inversely related to flexibility.

NON-LINEAR MODELS

Generalized Linear Models

Notation

θ, ϕ	Linear exponential family parameters
$\text{E}[Y]$, μ	Mean response
$h(\mu)$	Link function
$\mathbf{b}$	Maximum likelihood estimate of $\boldsymbol{\beta}$
$l(\mathbf{b})$	Maximized log-likelihood
l_0	Maximized log-likelihood for null model
l_{sat}	Maximized log-likelihood for saturated model
e	Residual
$\mathbf{I}$	Information matrix
$\chi^2_{1-q,\text{df}}$	q quantile of a chi-square distribution
D^*	Scaled deviance
D	Deviance statistic

Linear Exponential Family

$$\text{Prob. fn. of } Y = \exp\left[\frac{y\theta - b(\theta)}{\phi} + a(y, \phi)\right]$$

$\text{E}[Y] = b'(\theta)$

$\text{Var}[Y] = \phi \cdot b''(\theta)$

Model Framework

- $h(\mu) = \mathbf{x}^T\boldsymbol{\beta}$
- ϕ_i is either a known constant regardless of i, or ϕ / w_i, where w_i is a predetermined weight.
- Canonical link is the link function where $h(\mu) = b'^{-1}(\mu)$.

Parameter Estimation

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + a(y_i, \phi_i)\right]$$

where $\theta_i = b'^{-1}\left[h^{-1}\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right]$

The score equations are the partial derivatives of $l(\boldsymbol{\beta})$ with respect to each β_j all set equal to 0. The solution to the score equations is $\mathbf{b}$. Then, $\hat{\mu} = h^{-1}(\mathbf{x}^T\mathbf{b})$.

Numerical Results

$D^* = 2[l_{\text{sat}} - l(\mathbf{b})]$

$D = \phi^* D^*$ where $\phi^* = \phi_i$ or ϕ

$$R^2_{\text{ms}} = \frac{1 - \exp\{2[l_0 - l(\mathbf{b})]/n\}}{1 - \exp\{2l_0/n\}}$$

$$R^2_{\text{pse.}} = \frac{l(\mathbf{b}) - l_0}{l_{\text{sat}} - l_0}$$

$\text{AIC}^* = -2 \cdot l(\mathbf{b}) + 2 \cdot (p + 1)$

$\text{BIC}^* = -2 \cdot l(\mathbf{b}) + \ln n \cdot (p + 1)$

*Assumes only $\boldsymbol{\beta}$ need to be estimated. If estimating ϕ is required, replace $p + 1$ with $p + 2$.

Residuals

Raw Residual

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

Pearson Residual

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}[Y_i]}} = \frac{y_i - h^{-1}\left(\mathbf{x}_i^T\mathbf{b}\right)}{\sqrt{\hat{\phi}_i \cdot b''\left(\hat{\theta}_i\right)}}$$

where $\hat{\theta}_i = b'^{-1}\left[h^{-1}\left(\mathbf{x}_i^T\mathbf{b}\right)\right]$. The Pearson chi-square statistic is $\sum_{i=1}^{n} e_i^2$.

Deviance Residual

$e_i = \pm\sqrt{D_i^*}$ whose sign follows the i^{th} raw residual

Anscombe Residual

$$e_i = \frac{t(y_i) - \widehat{\text{E}}[t(Y_i)]}{\sqrt{\widehat{\text{Var}}[t(Y_i)]}}$$

Inference

- Maximum likelihood estimators $\widehat{\boldsymbol{\beta}}$ asymptotically have a multivariate normal distribution with mean $\boldsymbol{\beta}$ and asymptotic variance-covariance matrix $\mathbf{I}^{-1}$.
- To address overdispersion, change the variance to $\text{Var}[Y_i] = \delta \cdot \phi_i \cdot b''(\theta_i)$ and estimate δ as the Pearson chi-square statistic divided by $n - p - 1$.

Likelihood Ratio Tests

χ^2 statistic $= 2\left[l(\mathbf{b}_f) - l(\mathbf{b}_r)\right]$

Reject H_0 if χ^2 statistic $\geq \chi^2_{\alpha, p_f - p_r}$

Goodness-of-Fit Tests

Y follows a distribution of choice with g free parameters, whose domain is split into w mutually exclusive intervals.

$$\chi^2 \text{ statistic} = \sum_{c=1}^{w} \frac{(n_c - nq_c)^2}{nq_c}$$

Reject H_0 if χ^2 statistic $\geq \chi^2_{\alpha, w-g-1}$

Tweedie Distribution

$\text{E}[Y] = \mu, \qquad \text{Var}[Y] = \phi \cdot \mu^d$

Distribution	d
Normal	0
Poisson	1
Gamma	2
Tweedie	$(1, 2)$
Inverse Gaussian	3

Logistic and Probit Regression

- The odds of an event are the ratio of the probability that the event will occur to the probability that the event will not occur.
- The odds ratio is the ratio of the odds of an event with the presence of a characteristic to the odds of the same event without the presence of that characteristic.

Binary Response

Function Name	$h(\mu)$
Logit	$\ln\left(\frac{\mu}{1-\mu}\right)$
Probit	$\Phi^{-1}(\mu)$
Complementary log-log	$\ln(-\ln(1-\mu))$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln \mu_i + (1 - y_i) \ln(1 - \mu_i)]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i (y_i - \mu_i) \frac{\mu_i'}{\mu_i(1 - \mu_i)} = \mathbf{0}$$

$$D = 2 \sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (1 - y_i) \ln\left(\frac{1 - y_i}{1 - \hat{\mu}_i}\right) \right]$$

$$\text{Pearson residual, } e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 - \hat{\mu}_i)}}$$

$$\text{Pearson chi-square statistic} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1 - \hat{\mu}_i)}$$

Nominal Response – Generalized Logit

Let $\pi_{i,c}$ be the probability that the i^{th} observation is classified as category c. k is the reference category.

$$\ln\left(\frac{\pi_{i,c}}{\pi_{i,k}}\right) = \mathbf{x}_i^T \boldsymbol{\beta}_c$$

$$\pi_{i,c} = \begin{cases} \dfrac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_c)}{1 + \sum_{m \neq k} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_m)}, & c \neq k \\ \dfrac{1}{1 + \sum_{m \neq k} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_m)}, & c = k \end{cases}$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{c=1}^{w} I(y_i = c) \ln \pi_{i,c}$$

Ordinal Response – Proportional Odds Cumulative

$h(\Pi_c) = \alpha_c + \mathbf{x}_i^T \boldsymbol{\beta}$ where

- $\Pi_c = \pi_1 + \cdots + \pi_c$
- $\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

Poisson Count Regression

$\ln \mu = \mathbf{x}^T \boldsymbol{\beta}$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln \mu_i - \mu_i - \ln(y_i!)\ ]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i (y_i - \mu_i) = \mathbf{0}$$

$$\mathbf{I} = \sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i^T$$

$$D = 2 \sum_{i=1}^{n} \left\{ y_i \left[\ln\left(\frac{y_i}{\hat{\mu}_i}\right) - 1 \right] + \hat{\mu}_i \right\}$$

$$\text{Pearson residual, } e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

$$\text{Pearson chi-square statistic} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Poisson Regression with Exposures Model

$\ln \mu = \ln w + \mathbf{x}^T \boldsymbol{\beta}$

Alternative Count Models

These models can incorporate a Poisson distribution while letting the mean of the response differ from the variance of the response:

Models	*Mean < Variance*	*Mean > Variance*
Negative binomial	Yes	No
Zero-inflated	Yes	No
Hurdle	Yes	Yes
Heterogeneity	Yes	No

TIME SERIES

Trend Models

Notation

Subscript t	Index for observations
T_t	Trends in time
S_t	Seasonal trends
ε_t	Random patterns
$\hat{y}_{n+l}$	l-step ahead forecast
se	Estimated standard error
$t_{1-q,\text{df}}$	q quantile of a t-distribution
n_1	Training sample size
n_2	Test sample size

Trends

Additive: $Y_t = T_t + S_t + \varepsilon_t$

Multiplicative: $Y_t = T_t \times S_t + \varepsilon_t$

Stationarity

Stationarity describes how something does not vary with respect to time. Control charts can be used to identify stationarity.

White Noise

$\hat{y}_{n+l} = \bar{y}$

$se_{\hat{y}_{n+l}} = s_y\sqrt{1 + 1/n}$

$100k\%$ prediction interval for y_{n+l} is

$\hat{y}_{n+l} \pm t_{(1-k)/2,n-1} \cdot se_{\hat{y}_{n+l}}$

Random Walk

$w_t = y_t - y_{t-1}$

$\hat{y}_{n+l} = y_n + l\bar{w}$

$se_{\hat{y}_{n+l}} = s_w\sqrt{l}$

Approximate 95% prediction interval for y_{n+l} is $\hat{y}_{n+l} \pm 2 \cdot se_{\hat{y}_{n+l}}$

Model Comparison

$$\text{ME} = \frac{1}{n_2}\sum_{t=n_1+1}^{n} e_t$$

$$\text{MPE} = 100 \cdot \frac{1}{n_2}\sum_{t=n_1+1}^{n} \frac{e_t}{y_t}$$

$$\text{MSE} = \frac{1}{n_2}\sum_{t=n_1+1}^{n} e_t^2$$

$$\text{MAE} = \frac{1}{n_2}\sum_{t=n_1+1}^{n} |e_t|$$

$$\text{MAPE} = 100 \cdot \frac{1}{n_2}\sum_{t=n_1+1}^{n} \left|\frac{e_t}{y_t}\right|$$

Autoregressive Models

Notation

ρ_k	Lag k autocorrelation
r_k	Lag k sample autocorrelation
σ^2	Variance of white noise
s^2	Estimate of σ^2
b_0	Estimate of β_0
b_1	Estimate of β_1
$\bar{y}_-$	Sample mean of first $n-1$ observations
$\bar{y}_+$	Sample mean of last $n-1$ observations

Autocorrelation

$$r_k = \frac{\sum_{t=k+1}^{n}(y_{t-k} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

To test $H_0: \rho_k = 0$ against $H_1: \rho_k \neq 0$

- $se_{r_k} = 1/\sqrt{n}$
- test statistic $= r_k/se_{r_k}$

AR(1) Model

$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$

Assumptions

1. $\text{E}[\varepsilon_t] = 0$
2. $\text{Var}[\varepsilon_t] = \sigma^2$
3. $\text{Cov}[\varepsilon_{t+k}, Y_t] = 0$ for $k > 0$

- If $\beta_1 = 0$, Y_t follows a white noise process.
- If $\beta_1 = 1$, Y_t follows a random walk process.
- If $-1 < \beta_1 < 1$, Y_t is stationary.

Properties of Stationary AR(1) Model

$$\text{E}[Y_t] = \frac{\beta_0}{1-\beta_1}$$

$$\text{Var}[Y_t] = \frac{\sigma^2}{1-\beta_1^2}$$

$\rho_k = \beta_1^k$

Estimation

$$b_1 = \frac{\sum_{t=2}^{n}(y_{t-1} - \bar{y}_-)(y_t - \bar{y}_+)}{\sum_{t=2}^{n}(y_{t-1} - \bar{y}_-)^2}$$

$b_0 = \bar{y}_+ - b_1\bar{y}_-$

$$s^2 = \frac{\sum_{t=2}^{n} e_t^2}{n-3}$$

$$\widehat{\text{Var}}[Y_t] = \frac{s^2}{1-b_1^2}$$

Smoothing and Predictions

$\hat{y}_t = b_0 + b_1 y_{t-1}, \quad 2 \le t \le n$

$$\hat{y}_{n+l} = \begin{cases} b_0 + b_1 y_{n+l-1}, & l = 1 \\ b_0 + b_1 \hat{y}_{n+l-1}, & l > 1 \end{cases}$$

$$se_{\hat{y}_{n+l}} = s\sqrt{1 + b_1^2 + b_1^4 + \cdots + b_1^{2(l-1)}}$$

$100k\%$ prediction interval for y_{n+l} is

$\hat{y}_{n+l} \pm t_{(1-k)/2,n-3} \cdot se_{\hat{y}_{n+l}}$

Other Time Series Models

Notation

k	Moving average length
w	Smoothing parameter
g	Seasonal base
d	No. of trigonometric functions

Smoothing with Moving Averages

$$\hat{s}_t = \frac{y_t + y_{t-1} + \cdots + y_{t-k+1}}{k}$$

$$\hat{s}_t = \hat{s}_{t-1} + \frac{y_t - y_{t-k}}{k}, \quad k = 1, 2, \ldots$$

Exponential Smoothing

$\hat{s}_t = (1-w)(y_t + wy_{t-1} + \cdots + w^t y_0)$

$\hat{s}_t = (1-w)y_t + w\hat{s}_{t-1}, \quad 0 \le w < 1$

Key Ideas for Smoothing

- It is only appropriate for time series data without a linear trend.
- It is related to weighted least squares.
- A double smoothing procedure can be used to forecast time series data with a linear trend.
- Holt-Winter double exponential smoothing is a generalization of the double exponential smoothing.

Seasonal Time Series Models

Fixed Seasonal Effects – Trigonometric Functions

$$S_t = \sum_{i=1}^{d}\left[\beta_{1,i}\sin(f_i t) + \beta_{2,i}\cos(f_i t)\right]$$

- $f_i = 2\pi i/g$
- $d \le g/2$

Seasonal Autoregressive Models, SAR(p)

$Y_t = \beta_0 + \beta_1 Y_{t-g} + \cdots + \beta_p Y_{t-pg} + \varepsilon_t$

Holt-Winter Seasonal Additive Model

$Y_t = \beta_0 + \beta_1 t + S_t + \varepsilon_t$

- $S_t = S_{t-g}$
- $\sum_{t=1}^{g} S_t = 0$

Unit Root Test

- A unit root test is used to test whether a time series is stationary or not.
- A time series is not stationary if it possesses a unit root.
- The Dickey-Fuller test and augmented Dickey-Fuller test are two examples of unit root tests.

Volatility Models

ARCH(p) Model

$\sigma_t^2 = \theta + \gamma_1\varepsilon_{t-1}^2 + \cdots + \gamma_p\varepsilon_{t-p}^2$

GARCH(p, q) Model

$$\sigma_t^2 = \theta + \gamma_1\varepsilon_{t-1}^2 + \cdots + \gamma_p\varepsilon_{t-p}^2 + \delta_1\sigma_{t-1}^2 + \cdots + \delta_q\sigma_{t-q}^2$$

$$\text{Var}[\varepsilon_t] = \frac{\theta}{1 - \sum_{j=1}^{p}\gamma_j - \sum_{j=1}^{q}\delta_j}$$

Assumptions

- $\theta > 0$
- $\gamma_j \ge 0$
- $\delta_j \ge 0$
- $\sum_{j=1}^{p}\gamma_j + \sum_{j=1}^{q}\delta_j < 1$

DECISION TREES

Regression and Classification Trees

Notation

R	Region of predictor space
n_m	No. of observations in node m
$n_{m,c}$	No. of category c observations in node m
I	Impurity
E	Classification error rate
G	Gini index
D	Cross entropy
T	Subtree
$\|T\|$	No. of terminal nodes in T
λ	Tuning parameter

Algorithm

1. Construct a large tree with g terminal nodes using recursive binary splitting.
2. Obtain a sequence of best subtrees, as a function of λ, using cost complexity pruning.
3. Choose λ by applying k-fold cross validation. Select the λ that results in the lowest cross-validation error.
4. The best subtree is the subtree created in step 2 with the selected λ value.

Recursive Binary Splitting

Regression:

$$\text{Minimize} \sum_{m=1}^{g} \sum_{i:\mathbf{x}_i \in R_m} \left(y_i - \bar{y}_{R_m}\right)^2$$

Classification:

$$\text{Minimize} \frac{1}{n} \sum_{m=1}^{g} n_m \cdot I_m$$

More Under Classification:

$\hat{p}_{m,c} = n_{m,c}/n_m$

$E_m = 1 - \max_c \hat{p}_{m,c}$

$G_m = \sum_{c=1}^{w} \hat{p}_{m,c}\left(1 - \hat{p}_{m,c}\right)$

$D_m = -\sum_{c=1}^{w} \hat{p}_{m,c} \ln \hat{p}_{m,c}$

$\text{deviance} = -2 \sum_{m=1}^{g} \sum_{c=1}^{w} n_{m,c} \ln \hat{p}_{m,c}$

$$\text{residual mean deviance} = \frac{\text{deviance}}{n - g}$$

Cost Complexity Pruning

Regression:

$$\text{Minimize} \sum_{m=1}^{|T|} \sum_{i:\mathbf{x}_i \in R_m} \left(y_i - \bar{y}_{R_m}\right)^2 + \lambda |T|$$

Classification:

$$\text{Minimize} \frac{1}{n} \sum_{m=1}^{|T|} n_m \cdot I_m + \lambda |T|$$

Key Ideas

- Terminal nodes or leaves represent the partitions of the predictor space.
- Internal nodes are points along the tree where splits occur.
- Terminal nodes do not have child nodes, but internal nodes do.
- Branches are lines that connect any two nodes.
- A decision tree with only one internal node is called a stump.

Advantages of Trees

- Easy to interpret and explain
- Can be presented visually
- Manage categorical variables without the need of dummy variables
- Mimic human decision-making

Disadvantages of Trees

- Not robust
- Do not have the same degree of predictive accuracy as other statistical methods

Multiple Trees

Bagging

1. Create b bootstrap samples from the original training dataset.
2. Construct a decision tree for each bootstrap sample using recursive binary splitting.
3. Predict the response of a new observation by averaging the predictions (regression trees) or by using the most frequent category (classification trees) across all b trees.

Properties

- Increasing b does not cause overfitting.
- Bagging reduces variance.
- Out-of-bag error is a valid estimate of test error.

Random Forests

1. Create b bootstrap samples from the original training dataset.
2. Construct a decision tree for each bootstrap sample using recursive binary splitting. At each split, a random subset of k variables are considered.
3. Predict the response of a new observation by averaging the predictions (regression trees) or by using the most frequent category (classification trees) across all b trees.

Properties

- Bagging is a special case of random forests.
- Increasing b does not cause overfitting.
- Decreasing k reduces the correlation between predictions.

Boosting

Let z_1 be the actual response variable, y.

1. For $k = 1, 2, \ldots, b$:
   - Use recursive binary splitting to fit a tree with d splits to the data with z_k as the response.
   - Update z_k by subtracting $\lambda \cdot \hat{f}_k(\mathbf{x})$, i.e. let $z_{k+1} = z_k - \lambda \cdot \hat{f}_k(\mathbf{x})$.
2. Calculate the boosted model prediction as $\hat{f}(\mathbf{x}) = \sum_{k=1}^{b} \lambda \cdot \hat{f}_k(\mathbf{x})$.

Properties

- Increasing b can cause overfitting.
- Boosting reduces bias.
- d controls complexity of the boosted model.
- λ controls the rate at which boosting learns.

UNSUPERVISED LEARNING

Principal Components Analysis

Notation

z, Z	Principal component (score)
Subscript m	Index for principal components
ϕ	Principal component loading
x, X	Centered explanatory variable

Principal Components

$$z_m = \sum_{j=1}^{p} \phi_{j,m} x_j, \qquad z_{i,m} = \sum_{j=1}^{p} \phi_{j,m} x_{i,j}$$

- $\sum_{j=1}^{p} \phi_{j,m}^2 = 1$
- $\sum_{j=1}^{p} \phi_{j,m} \cdot \phi_{j,u} = 0, \ m \neq u$

Proportion of Variance Explained (PVE)

$$\sum_{j=1}^{p} s_{x_j}^2 = \sum_{j=1}^{p} \frac{1}{n-1} \sum_{i=1}^{n} x_{i,j}^2$$

$$s_{z_m}^2 = \frac{1}{n-1} \sum_{i=1}^{n} z_{i,m}^2$$

$$\text{PVE} = \frac{s_{z_m}^2}{\sum_{j=1}^{p} s_{x_j}^2}$$

Key Ideas

- The variance explained by each subsequent principal component is always less than the variance explained by the previous principal component.
- All principal components are uncorrelated with one another.
- A dataset has $\min(n-1, p)$ distinct principal components.
- The first k principal component scores and loadings approximate the original dataset, $x_{i,j} \approx \sum_{m=1}^{k} z_{i,m} \phi_{j,m}$.

Principal Components Regression

$Y = \theta_0 + \theta_1 z_1 + \cdots + \theta_k z_k + \varepsilon$

- If $k = p$, then $\beta_j = \sum_{m=1}^{k} \theta_m \phi_{j,m}$.

Cluster Analysis

Notation

C	Cluster containing indices
$W(C)$	Within-cluster variation of cluster
$\lvert C \rvert$	No. of observations in cluster

$$\text{Euclidean Distance} = \sqrt{\sum_{j=1}^{p} \left(x_{i,j} - x_{m,j}\right)^2}$$

k-Means Clustering

1. Randomly assign a cluster to each observation. This serves as the initial cluster assignments.
2. Calculate the centroid of each cluster.
3. For each observation, identify the closest centroid and reassign to that cluster.
4. Repeat steps 2 and 3 until the cluster assignments stop changing.

$$W(C_u) = \frac{1}{|C_u|} \sum_{i,m \in C_u} \sum_{j=1}^{p} \left(x_{i,j} - x_{m,j}\right)^2$$

$$= 2 \sum_{i \in C_u} \sum_{j=1}^{p} \left(x_{i,j} - \bar{x}_{u,j}\right)^2$$

Hierarchical Clustering

1. Select the dissimilarity measure and linkage to be used. Treat each observation as its own cluster.
2. For $k = n, n-1, \ldots, 2$:
   - Compute the inter-cluster dissimilarity between all k clusters.
   - Examine all $\binom{k}{2}$ pairwise dissimilarities. The two clusters with the lowest inter-cluster dissimilarity are fused. The dissimilarity indicates the height in the dendrogram at which these two clusters join.

Linkage	*Inter-cluster dissimilarity =*
Complete	The largest dissimilarity
Single	The smallest dissimilarity
Average	The arithmetic mean
Centroid	The dissimilarity between the cluster centroids

Key Ideas

- For k-means clustering, the algorithm needs to be repeated for each k.
- For hierarchical clustering, the algorithm only needs to be performed once for any number of clusters.
- The result of clustering depends on many parameters, such as:
  - Choice of k in k-means clustering
  - Choice of number of clusters, linkage, and dissimilarity measure in hierarchical clustering
  - Choice to standardize variables