BASIC STATISTICAL LEARNING

Prediction: Determining what the response will be for new sets of predictor values.

Inference: Understanding how the explanatory variables influence the response.

Parametric: Specifies a function, and statistics is used to fit the function's parameters

Non-parametric: Without specify a functional form for the relationship.

Supervised learning: Has a response variable that is influenced by explanatory variables.

Unsupervised learning: Relates the observations to each other or finds patterns in the observations.

Training MSE: $\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\hat{f}\left(\mathbf{x}_i\right)\right]^2$, the training MSE decreases monotonically when the model flexibility increases.

Test MSE: Average $\left[y_0-\hat{f}\left(\mathbf{x}_0\right)\right]^2$, the test MSE exhibits a U-shape, first declining as flexibility increases, then starting to increase again.

Bias: Measures the extent to which the expected value of the estimator differs from the true value

Variance: Measures how much the estimator varies with different random samples of data

Bias-variance trade-off: Minimize $\mathbf{E}_{y_0}\left[\left(y_0-\hat{f}\left(\mathbf{x}_0\right)\right)^2\right]=\operatorname{Var}\left[\hat{f}\left(\mathbf{x}_0\right)\right]+\left[\operatorname{Bias}\left(\hat{f}\left(\mathbf{x}_0\right)\right)\right]^2+\operatorname{Var}(\varepsilon)$

Increasing complexity/flexibility of a model may result in an increase in variance but a reduction in bias.

Overfitting: The model will replicate historical data very well but cannot predict future outcomes reliably.

Underfitting: The model will predict future outcomes reliably but cannot explain what is driving the result.

Sum of squares:

$$S_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}$$

$$S_{xx}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

Sample variance of x and y: $s_x^2=\frac{S_{xx}}{n-1}\quad s_y^2=\frac{S_{yy}}{n-1}$

Sample covariance: $cv_{xy}=\frac{S_{xy}}{n-1}$

Sample correlation between x and y: $r_{xy}=\frac{cv_{xy}}{s_xs_y}$

Linear Regression

SIMPLE LINEAR REGRESSION

Model equation: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$

β_0(intercept) and β_1(slope parameter) are regression coefficients, ε_i is the random error term.

Model assumptions: ε_i are independent with $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Fitted regression: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad \text{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Specialized formulas for $\widehat{\beta}_1$: $\widehat{\beta}_1 = r_{xy} \times \dfrac{s_y}{s_x}$

Properties:

$\mathrm{E}\left[\widehat{\beta}_0\right] = \beta_0$ $\quad \mathrm{Var}\left(\widehat{\beta}_0\right) = \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)$ $\quad \mathrm{SE}\left(\widehat{\beta}_0\right) = \sqrt{s^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}$ $\quad s^2$ is an unbiased estimator of σ^2

$\mathrm{E}\left[\widehat{\beta}_1\right] = \beta_1$ $\quad \mathrm{Var}\left(\widehat{\beta}_1\right) = \sigma^2\left(\dfrac{1}{S_{xx}}\right)$ $\quad \mathrm{SE}\left(\widehat{\beta}_1\right) = \sqrt{s^2\left(\dfrac{1}{S_{xx}}\right)}$ $\quad \mathrm{Cov}\left(\beta_0, \beta_1\right) = -\dfrac{\sigma^2}{S_{xx}}\bar{x}$

$\mathrm{E}\left[\widehat{y}_i\right] = \beta_0 + \beta_1 x_i$ $\quad \mathrm{Var}\left(\widehat{y}_i\right) = \sigma^2\left(\dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{S_{xx}}\right)$ $\quad \mathrm{SE}\left(\widehat{y}_i\right) = \sqrt{s^2\left(\dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{S_{xx}}\right)}$

Confidence interval: $\widehat{y}_* \pm t_{n-2,\alpha/2} \times \mathrm{SE}\left(\mathrm{E}[y_*] - \widehat{y}_*\right) = \widehat{y}_* \pm t_{n-2,\alpha/2}\sqrt{s^2\left[\dfrac{1}{n} + \dfrac{(x_* - \bar{x})^2}{S_{xx}}\right]}$

Prediction interval: $\widehat{y}_* \pm t_{n-2,\alpha/2} \times \mathrm{SE}\left(y_* - \widehat{y}_*\right) = \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_*\right) \pm t_{n-2,\alpha/2}\sqrt{s^2\left[1 + \dfrac{1}{n} + \dfrac{(x_* - \bar{x})^2}{S_{xx}}\right]}$

Residual(Raw residual)**:** $e_i = y_i - \widehat{y}_i =$ observed value $-$ fitted value

Sum-to-zero constraints on residuals: $\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} x_i e_i = 0$

GOODNESS OF FIT OF A MODEL

Partition sum of squares:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Total sum of squares}} = \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\text{Residual sum of squares}} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}_{\text{Regression sum of squares}} + \underbrace{2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y})}_{\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \bar{y}\right) = 0}$$

Total sum of squares(TSS)**:** Amount of variability inherent in the response prior to performing regression

Residual sum of squares(RSS)**:** Variation unexplained by the linear regression model

Mean square error(MSE) in SLR: $s^2 = \dfrac{\text{RSS}}{n-2} = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2}$ Note: s^2 is an unbiased estimator of $\sigma^2 = \text{Var}(y) = \text{Var}(\varepsilon)$.

Regression sum of squares(Reg SS): Variation explained by the linear regression model

Coefficient of determination: The proportion of the sum of squares explained by the regression.

$R^2 = \dfrac{\text{Reg SS}}{\text{TSS}} = 1 - \dfrac{\text{RSS}}{\text{TSS}}$, adding more variables to the model always increases R^2.

$R^2 = r^2$ (the square of the sample correlation coefficient)

Specialized formulas for Reg SS: Reg SS$= \hat{\beta}_1^2 S_{xx}$

MULTIPLE LINEAR REGRESSION

Model equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$

β_j is the regression coefficient attached to the j th predictor, for $j = 1, \ldots, p$

Model fitting: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{y}$ is the n x 1 response vector, $\mathbf{X}$ is the **design matrix**, $\boldsymbol{\beta}$ is the vector of $p+1$ regression coefficients

$$\overbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}^{\mathbf{y}} = \overbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}^{\mathbf{X}} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}^{\boldsymbol{\varepsilon}},$$

LSE of $\boldsymbol{\beta}$: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\mathrm{y}$

$$\mathbf{X'X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix},$$

$$(\mathbf{X'X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (n\bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix},$$

$$\mathbf{X'y} = \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix},$$

Fitted regression: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

ANOVA table:

Sum of Squares	degree of freedom	Mean Square
Reg SS	p	Reg SS/p
RSS	$n-(p+1)$	RSS/$[n-(p+1)]$=s^2
TSS	$n-1$	

Properties: $\mathrm{E}\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$ $\quad \mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ $\quad \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$

Predicted value: $\widehat{y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_{*1} + \cdots + \widehat{\beta}_p x_{*p}$

Confidence interval: $\widehat{y}_* \pm t_{n-p-1,\alpha/2} \times \mathrm{SE}\left(\mathrm{E}[y_*] - \widehat{y}_*\right) = \widehat{y}_* \pm t_{n-p-1,\alpha/2}\sqrt{s^2\left[\mathbf{x}_*'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_*\right]}$

Prediction interval: $\widehat{y}_* \pm t_{n-p-1,\alpha/2} \times \mathrm{SE}\left(y_* - \widehat{y}_*\right) = \widehat{y}_* \pm t_{n-p-1,\alpha/2}\sqrt{s^2\left[1 + \mathbf{x}_*'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_*\right]}$

T TESTS, F TESTS, AND GENERALIZED F TEST

Hypothesis tests: t-test

Test Null hypothesis H_0 : $\beta_j = d.$

t-statistic: $t\left(\widehat{\beta}_j\right) = \dfrac{\text{LSE} - \text{hypothesized value}}{\text{standard error of LSE}} = \dfrac{\widehat{\beta}_j - d}{\mathrm{SE}\left(\widehat{\beta}_j\right)}$

Alternative Hypothesis H_a	Reject H_0 in favor of H_a if...	p -value (t is the observed value of $t\left(\widehat{\beta}_j\right)$)
$\beta_j \neq d$	$\left\lvert t\left(\widehat{\beta}_j\right)\right\rvert > t_{n-p-1,\alpha/2}$	$\Pr\left(\lvert t_{n-p-1}\rvert > \lvert t\rvert\right) = 2\Pr\left(t_{n-p-1} > \lvert t\rvert\right)$
$\beta_j > d$	$t\left(\widehat{\beta}_j\right) > t_{n-p-1,\alpha}$	$\Pr\left(t_{n-p-1} > t\right)$
$\beta_j < d$	$t\left(\widehat{\beta}_j\right) < -t_{n-p-1,\alpha}$	$\Pr\left(t_{n-p-1} < t\right)$

CI for β_j: $\widehat{\beta}_j \pm t_{n-p-1,\alpha/2} \times \mathrm{SE}\left(\widehat{\beta}_j\right)$

F-tests: Test Null hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0.$

F-statistic: $F_{p,n-p-1} = \dfrac{\text{Reg SS}/p}{\text{RSS}/(n-p-1)} = \dfrac{R^2/p}{(1-R^2)/(n-p-1)}$

For the single linear regression model, the F statistic is the square of the t statistic for $\mathrm{H}_0 : \beta_1 = 0$

$$t\left(\widehat{\beta}_1\right)^2 = \frac{\widehat{\beta}_1^2}{s^2/S_{xx}} = \frac{\widehat{\beta}_1^2 S_{xx}}{s^2} = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)} = F_{1,n-2}$$

Generalized F-test:

full model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$

reduced model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-q} x_{p-q} + \varepsilon$ where $q \le p$ is a positive integer

Test Null hypothesis H_0 : $\beta_{p-q+1} = \cdots = \beta_p = 0$.

	Reduced model		Full model
RSS	RSS_r	$\ge$	RSS_f
Reg SS	Reg SS_r	$\le$	Reg SS_f
TSS	TSS	=	TSS

F-statistic: $F_{q,n-p-1} = \dfrac{(\text{RSS}_r - \text{RSS}_f)/q}{\text{RSS}_f/(n-p-1)} = \dfrac{(R_f^2 - R_r^2)/q}{(1-R_f^2)/(n-p-1)}$

MODEL CONSTRUCTION

Continuous predictors: $y = \beta_0 + \boxed{\beta_1 x} + \varepsilon$, A unit increase in x is expected to increase y by $\dfrac{\partial}{\partial x}\text{E}[y] = \beta_1$, while all other predictors are held constant.

Categorical predictors: k levels of variables with the **base level** variable

Special case: **binary(dummy)** variable, Each dummy variable corresponds to one possible value of the categorical variable.

It is equal to 1 if the variable is equal to that value, 0 otherwise

Polynomial regression: $y = \beta_0 + \beta_1 x + \boxed{\beta_2 x^2 + \cdots + \beta_m x^m} + \varepsilon$

$m = 2 \rightarrow$ quadratic regression, $m = 3 \rightarrow$ cubic polynomial.

Regression with interaction term: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2} + \varepsilon$ $x_1 x_2$ is called an **interaction term**.

Interactions between continuous and categorical predictors:

$$\text{E}[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \begin{cases} \beta_0 + \beta_1 x_1, & \text{if } x_2 = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1, & \text{if } x_2 = 1 \end{cases}. \quad x_2 \text{ is the base level}$$

Piecewise linear regression models:

Model 1: $\text{E}[y] = \beta_0 + \beta_1 x + \beta_2 [z(x-c) = \beta_0 + \beta_1 x + \beta_2 (x-c)_+ = \begin{cases} \beta_0 + \beta_1 x, & \text{if } x < c, \\ (\beta_0 - \beta_2 c) + (\beta_1 + \beta_2) x, & \text{if } x \ge c \end{cases}$

Model 2: $\text{E}[y] = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 [zx] = \begin{cases} \beta_0 + \beta_1 x, & \text{if } z = 0, \text{ or } x < c, \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x, & \text{if } z = 1, \text{ or } x \ge c \end{cases}$

PARTIAL CORRELATION COEFFICIENTS

Scatterplot matrix: Each entry of the matrix is a scatterplot for a pair of variables identified by the corresponding row and column labels

Drawback: The pairwise relationships are may be by other variables in the data
Each entry of the matrix can only depict the relationship between a pair of variables at one time

Construct a plot to remove the effects:

1. Regress y on $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$, i.e., all predictors except x_j.
   Let $e_{i1}\ (= y_i - \widehat{y}_i)$ for $i = 1, 2, \ldots, n$ be the residuals from this regression.
2. Regress x_j on $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$. Let $e_{i2}\ (= x_{ij} - \widehat{x}_{ij})$ for $i = 1, 2, \ldots, n$ be the residuals from this regression.
3. Construct a scatter plot of e_1 on e_2.

Partial correlation t-statistic: $$r\left(y, x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p\right) = \frac{t\left(\widehat{\beta}_j\right)}{\sqrt{t\left(\widehat{\beta}_j\right)^2 + (n-p-1)}},$$

RESIDUAL ANALYSIS

Hat matrix: $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ $\rightarrow$ $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.

Residual: $\widehat{\boldsymbol{e}} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$

Variance: $\mathrm{Var}\left(e_i\right) = \sigma^2\left(1 - h_{ii}\right)$ $\quad$ $\widehat{\mathrm{Var}}\left(e_i\right) = s^2\left(1 - h_{ii}\right)$

Standardized residuals: Comparable with one another unlike raw residuals, and share approximately the same variance.

$$e_i^{\mathrm{st}} = \frac{\widehat{e}_i}{\sqrt{s^2\left(1 - h_{ii}\right)}} = \frac{y_i - \widehat{y}_i}{\sqrt{s^2\left(1 - h_{ii}\right)}}, \quad i = 1, 2, \ldots, n$$

Studentized residual: Remove the effect of the i^{th} observation by deleting it from the data, it follows Student's t distribution with $n - (p+1)$ df

$$e_i^{\mathrm{stud}} = \frac{\widehat{e}_i}{\sqrt{s_{(i)}^2\left(1 - h_{ii}\right)}}, \quad i = 1, 2, \ldots, n$$

We need to validate model assumptions by checking the pattern of residuals.

1. Response is linear: Plot $\widehat{y}$ or $\widehat{\boldsymbol{e}}$ against each x.
2. Response is normal: Plot $\boldsymbol{e}^{\mathrm{st}}$ against $Z \sim N(0, 1)$.
3. Response has constant variance (homoscedasticity) Plot $\boldsymbol{e}^{\mathrm{st}}$ or $\widehat{\boldsymbol{e}}$ against $\widehat{y}$.
4. Observations are independent: Plot $\widehat{\boldsymbol{e}}$ in the order.

MODEL DIAGNOSTICS

Outliers: The observations with unusual values of the response variable y relative to the predicted values.

Potential outliers: Can be determined by using standardized residuals

Potential solutions:
1. Delete the outlier with care.
2. Retain the outlier in the model analysis, but make a note of its potential effects.
3. Create an indicator variable that is 1 only for outlier.

Influential points: Some observations may have a strong influence on the predicted value of y.

High leverage: Involves anomalous (unusual) values of predictors.
Note: high-leverage observation need not be influential

h_{ii} is the **leverage** of the i^{th} observation where h_{ii} **is between** $1/n$ **and 1**.

For a simple lienar regression: $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ $\quad \sum_{i=1}^{n} h_{ii} = p+1 \quad \rightarrow \quad \bar{h}_{ii} = (p+1)/n.$

Cook's Distance: An influence measure combines outliers and leverage.
$\widehat{y}_j$ is the fitted value of whole data set and $\widehat{y}_{j(i)}$ is the fitted value of data set with i^{th} observation removed.

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{(p+1)s^2} = D_i = \frac{1}{p+1} \times \underbrace{\left(e_i^{\text{st}}\right)^2}_{\text{outlier}} \times \underbrace{\left(\frac{h_{ii}}{1-h_{ii}}\right)}_{\text{leverage}}.$$

COLLINEARITY

Collinearity: One predictor is or is nearly a linear combination of the other predictors.
Standard error tend to overestimate the true standard errors.
Confidence and prediction intervals will be wider.
p-values for hypothesis tests will be lower.

Potential solutions:
1. Combine the collinear predictors into a single predictor
2. Delete one or more problematic predictors that seem to be causing collinearity.
3. Standardize each predictor by subtracting its average and dividing by its sample standard deviation.

Variance inflation factor: $\text{VIF}_j = \frac{1}{1-R_j^2}.$

Specialized formulas for SE $\left(\widehat{\beta}_j\right)$: $\text{d SE}\left(\widehat{\beta}_j\right) = \frac{s}{s_{x_j}}\sqrt{\frac{\text{VIF}_j}{n-1}}.$

Tolerance: The reciprocal of VIF, $1 - R_J^2$.

HETEROSCEDASTICITY

Heteroscedasticity: The amount of variability are changes throughout the plot.
Cause unreliable MSE, adjusted R^2, and F statistic.

Homoscedastic: The n observations are all subject to the same amount of variability from their expected values.

Detection by residual plots: Create residual plot (e) against fitted values $(\widehat{y})$ of the regression model.

Detection by Breusch–Pagan Test: $\mathrm{H}_a: \quad \mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2 + \mathbf{z}_i'\gamma, \quad i = 1, 2, \ldots, n.$

Breusch–Pagan testing procedure:

1. Fit the original MLR model, and obtain the raw residuals e_i's for $i = 1, \ldots, n$,
2. Calculate the squared standardized residuals $e_i^{*2} = e_i^2/s^2$,
3. Regress e^{*2} on z and calculate the resulting regression sum of squares,
4. Test statistic = Regression sum of squares / 2 follow a chi-square distribution with q degrees of freedom, where q is the dimension of γ.

Solutions to Heteroscedasticity

Weighted least squares: $\mathrm{Var}(\varepsilon_i) = \sigma^2/w_i$, with w_i varying by observation.

$$\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} = (\mathbf{X'WX})^{-1}\mathbf{X'Wy}, \text{ where } \mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}.$$

Specialized formulas for $\widehat{\beta}_1$: $\hat{\beta}_1 = \sum_{i=1}^{n} w_i y_i, \text{ where } w_i = \dfrac{x_i - \bar{x}}{S_{xx}} \text{ for } i = 1, 2, \ldots, n$

Transformations of the response variable: Apply a variance-stabilizing transformation to the response variable.
logarithmic transformations $\ln y$ and **square root transformations** $\sqrt{y}$.

RESAMPLING METHODS

Training set: The subset of the available observations used for fitting or "training" the given statistical learning method.

Validation set: The fitted statistical learning model is tested out by making predictions for observations in the validation set.

Out-of-sample validation: Randomly splitting the observations into the training and validation set, The model is fit to the training set, then the fitted model is used to predict the responses for the observations in the validation set.

MSE is calculate using the validation set MSE $= \dfrac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(y_i - \widehat{y}_i)^2$.

Draw back: Fewer observations in training leads to higher variance and overestimate the MSE.
The MSE is highly variable

Cross-validation: Select test data sets in a systematic non-random fashion, and then average the MSE results from all runs

Leave One Out Cross-Validation (LOOCV) $= \text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_{(i)}\right)^2 \frac{1}{n} = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$

$\sum_{i=1}^{n}\left(y_i - \widehat{y}_{(i)}\right)^2$ is called **predicted residual sum of squares (PRESS)** $= \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$

One observation as test data, the rest as training data. There is a total of n fits.

LOOCV minimizes bias but maximizes variance. MSE_i is calculated using the test data.

k-fold cross-validation $= \text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i$ LOOCV is a special case of k-fold CV when $k = n$

The data is randomly partitioned into k subsets, each subset as test data, the rest as training data. There is a total of k fits.

k-fold CV has higher bias but lower variance. k-fold CV has computational advantage over LOOCV.

MODEL COMPARISON

R^2 is not suitable for model comparison with different number of predictors since adding more predictors into a model will always improve R^2.

Model comparison statistic:

Adjusted R^2: $R_a^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} = 1 - \frac{n-1}{n-p-1}\left(1-R^2\right) = 1 - \frac{s^2}{s_y^2} \quad \rightarrow \quad R_a^2 \le R^2 \le 1$

s^2 is a model-dependent quantity, the model with the highest R_a^2 is equivalent to model one with the smallest MSE.

as p increases, R_a^2 decrease

Mallow's C_p: $C_p = \begin{cases} \frac{(\text{RSS})_p + 2ps^2}{n}, & \text{An Introduction to Statistical Learning} \\ \frac{(\text{RSS})_p}{s^2} - n + 2(p+1), & \text{Regression Modeling with Actuarial and Financial Applications} \end{cases}$

$(\text{RSS})_p$ is the residual sum of squares of the current model with p predictors.

the best model on the basis of C_p is the model with the lowest C_p.

Alternative AIC: $\text{AIC} = \frac{(\text{RSS})_p + 2ps^2}{ns^2}$

AIC is proportional to C_p. When used as model selection criteria, they give equivalent results.

Alternative BIC: $\text{BIC} = \frac{(\text{RSS})_p + \ln(n)ps^2}{ns^2}$

Models with the same number of predictors:

- The one with the lower SSE is better.
- The one with the higher R^2 is better.

Models with different number of predictors:

- The one with higher adjusted R^2 is better
- The one with lower Mallow's C_p is better
- The one with lower AIC/BIC is better.
- The one with lower cross-validation is better.

SUBSET SELECTION

Best subset selection: For each number of predictors, find the best candidate.
Then, select the best among these candidates.
There is a total of 2^k fitted models, best subset selection is rarely used when $k \geq 20$.
It is susceptible to overfitting and high variance of the regression coefficient estimates.

Forward stepwise selection: There is a total of $1 + \frac{k\,(k+1)}{2}$ fitted models.

1. Start with model having intercept only.
2. Create $p+1$ predictor models by fitting a model with the current p predictors plus one of the $k-p$ unused predictors.
3. Select the best $p+1$ predictor model based on RSS or R^2.
4. If $p+1 < k$, repeat steps $2-3$ with the $p+1$ parameter model.
5. Select the best model from the various models based on cross-validation or a statistic (Mallow's C_p, AIC, BIC, adjusted R^2).

Nested model: the predictors in the p-predictor model are always a subset of the predictors in the $(p+1)$-predictor model.

Backward stepwise selection: There is a total of $1 + \frac{k\,(k+1)}{2}$ fitted models.

1. Start with full model.
2. create $p-1$ predictor models by fitting a model removing one of the parameters from the current p predictors.
3. Select the best $p-1$ predictor model based on RSS or R^2.
4. If $p-1 > 1$, repeat steps $2-3$ with the $p-1$ parameter model.
5. Select the best model from the various models based on cross-validation or a statistic (Mallow's C_p, AIC, BIC, adjusted R^2).

Backward selection cannot be implemented in the high-dimensional setting with $n \leq k$

SHRINKAGE METHODS

Shrinkage methods: Reduce the variance of the coefficient estimators and improve the prediction accuracy of the model

Ridge regression and the lasso apply a specific penalty function to the sum of square differences.

Ridge regression: Minimize $\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})\right]^2+\lambda\sum_{j=1}^{p}\beta_j^2$ where $\lambda\geq 0$, λ is the **tuning parameter**

or

Minimize $\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})\right]^2$ subject to $\sum_{j=1}^{p}|\beta_j|\leq s$, s is the **budget parameter**.

Shrinkage penalty not applied to β_0 and **Standardization of predictors** $\tilde{x}_{ij}=\dfrac{x_{ij}}{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2/n}}$.

- When $\lambda=0$, the shrinkage penalty vanishes and ridge regression is identical to least squares regression.
- The larger the value of λ, the more of a price we have to pay for making the coefficient estimates non-zero.
- When $\lambda\to\infty$, the shrinkage penalty dominates and the coefficient estimates have no choice but to be all zero, and the model becomes the null model

Lasso regression: Minimize $\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})\right]^2+\lambda\sum_{j=1}^{p}|\beta_j|$

or

Minimize $\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})\right]^2$ subject to $\sum_{j=1}^{p}|\beta_j|\leq s$

- The lasso shrinking the coefficient estimates to be exactly zero when the tuning parameter λ is large enough. (Variable Selection)

Case	Model Complexity	Squared Bias	Variance	Training Error	Test Error
(**tuning parameter**) $\lambda\uparrow$	$\downarrow$ less complex	$\uparrow$	$\downarrow$	$\uparrow$	U-shape
(**budget parameter**) $s\uparrow$	$\uparrow$ more complex	$\downarrow$	$\uparrow$	$\downarrow$	U-shape

- When $s=0$ (equivalently, $\lambda=\infty$), all of the coefficient estimates are forced to be zero, resulting in perfect shrinkage.
- As s increases, so does the flexibility of the model
- As $s\to\infty$, there is essentially no constraint on the coefficient estimates, which is the least squares regression.

$\|\cdot\|_1(\ell_1\ norm)$: $\left\|\widehat{\boldsymbol{\beta}}_\lambda\right\|_1=|\beta_{1,\lambda}|+\cdots+|\beta_{p,\lambda}|$, $\|\cdot\|_2(\ell_2\ norm)$: $\left\|\widehat{\boldsymbol{\beta}}_\lambda\right\|_2=\sqrt{\beta_{1,\lambda}^2+\cdots+\beta_{p,\lambda}^2}$

Coefficient estimates: $\widehat{\beta}_i^R=\dfrac{y_i}{1+\lambda}$ and $\widehat{\beta}_i^L=\begin{cases}y_i-\lambda/2, & \text{if } y_i>\lambda/2\\ y_i+\lambda/2, & \text{if } y_i<-\lambda/2\\ 0, & \text{if } |y_i|\leq\lambda/2\end{cases}$

Subset selection and shrinkage methods can sidestep the curse of dimensionality (the test error tends to increase with the dimensionality)

Generalized Linear Model

GENERALIZED LINEAR MODELS

Linear exponential family of distributions: $f(y;\theta,\phi)=\exp\left[\dfrac{y\theta-b(\theta)}{\phi}+S(y,\phi)\right]$

- ϕ is the scale parameter.
- $b(\theta)$ and $S(y,\phi)$ are functions of θ and y,ϕ, respectively.
- y is the argument of the probability function.
- θ is the parameter of interest.

Distribution	θ	$b(\theta)$	ϕ
Binomial, Bin(n,π)(n is known)	$\ln[\pi/(1-\pi)]$	$n\ln\left(1+\mathrm{e}^{\theta}\right)$	1
Normal, N$\left(\mu,\sigma^2\right)$	μ	$\theta^2/2$	σ^2
Poisson (λ)	$\ln\lambda$	e^{θ}	1
Gamma, $\Gamma(\alpha,\gamma)$	$-\gamma/\alpha$	$-\ln(-\theta)$	$1/\alpha$
Inverse Gaussian, IG(μ,λ)	$-1/\left(2\mu^2\right)$	$-\sqrt{-2\theta}$	$1/\lambda$
Negative binomial, NB(r,p)(r is known)	$\ln(1-p)$	$-r\ln\left(1-\mathrm{e}^{\theta}\right)$	1

Mean and Variance: $\mu=\mathbb{E}[y]=b'(\theta)$ and $\mathrm{Var}(y)=\phi b''(\theta)$

Link functions: $g(\mu)=\mathbf{x}'\boldsymbol{\beta}=\beta_0+\beta_1x_1+\cdots+\beta_px_p$

Inverse of the link function: $\mu=g^{-1}(\eta)=g^{-1}\left(\beta_0+\beta_1x_1+\cdots+\beta_px_p\right)$ for $\eta=\mathbf{x}'\boldsymbol{\beta}$.

Canonical link function: Sets the systematic component equal to the parameter of interest; $\eta=\theta$.

Tweedie distributions: a member of the linear exponential family with $\mathrm{Var}(y)=\phi\mu^p$, $1<p<2$

Members of exponential family in canonical form:

Distribution	Canonical Link Function	Mathematical Form	Tweedie Distribution
Normal	Identity	$g(\mu)=\mu$	$b=0$
Poisson	Natural log	$g(\mu)=\ln\mu$	$b=1$
Gamma	Inverse	$g(\mu)=-1/\mu$	$b=2$
Inverse Gaussian	Squared inverse	$g(\mu)=-1/\mu^2$	$b=3$
Binomial	Logit	$g(\pi)=\ln[\pi/(1-\pi)]$	

ESTIMATION

Likelihood function: $L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + S(y_i, \phi_i)\right]$

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

Scores: The partial derivative of $l(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, and they are set equal to 0.

Prediction:

1. Calculate the estimated value of the linear predictor $g(\widehat{\mu}) = \mathbf{x}'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p$.
2. Invert g to obtain the fitted mean $\widehat{\mu} = g^{-1}\left(\mathbf{x}'\widehat{\boldsymbol{\beta}}\right)$.

MEASURES OF FIT

Scaled deviance: A goodness-of-fit measure of how much the fitted GLM departs from the saturated model $D^* = 2(l_{\text{SAT}} - l)$

Saturated model: The fitted values exactly equal the observed values, $\widehat{\mu}_i = y_i$ for all $i = 1, \ldots, n$ under the saturated model.

Deviance The scaled deviance multiplied by the scale parameter ϕ: $D = \phi D^*$

Derive and calculate the (scaled) deviance:

step 1: Write a generic expression for the loglikelihood function in terms of the unknown means $\mu_1, \ldots, \mu_n$.

step 2: Make the substitutions $\mu_i \to y_i$ for the saturated model and $\mu_i \to \widehat{\mu}_i$ for the fitted model to obtain l_{SAT} and l, respectively.

step 3: Determine the scaled deviance $D^* = 2(l_{\text{SAT}} - l)$ in terms of the y_i's and $\widehat{\mu}_i$'s.

step 4: Multiply the scaled deviance by the scale parameter ϕ to get the deviance.

Max-scaled R^2 : $R^2_{ms} = \frac{R^2}{R^2_{\text{SAT}}} = \frac{R^2}{1 - [\exp(l_0/n)]^2}$ where R^2 can be defined as $R^2 = 1 - \left[\frac{\exp(l_0/n)}{\exp(l/n)}\right]^2$

l_0 is the maximized loglikelihood of the null model

***Pseudo-R^2*:** $pseudo\text{-}R^2 = \frac{l - l_0}{l_{\text{SAT}} - l_0}$

Three major forms of residuals in generalized linear model

Pearson residuals: $r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\text{Var}}(y_i)}}$, $q_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$ (Poisson) $q_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i(1 - \widehat{\mu}_i)}}$ (Bernoulli)

Anscombe residuals: $\frac{h(y_i) - \mathbf{E}[h(y_i)]}{\sqrt{\text{Var}(h(y_i))}}$

where h is a transformation that makes $h(y_i)$ approximately normally distributed.

Deviance residuals: $d_i = \text{sign}(y_i - \widehat{y}_i)\sqrt{2\left(\ln f(y_i; \theta_{i,SAT}) - \ln f\left(y_i; \widehat{\theta}_i\right)\right)}$ where $\sum_{i=1}^{n} d_i^2 = D^*$.

Poisson regression: $d_i = \text{sign}\,(y_i - \hat{y}_i)\sqrt{2\left(y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right)}$.

Bernoulli regressio: $d_i = \text{sign}\,(y_i - \hat{y}_i)\sqrt{2\left(y_i \ln\frac{y_i}{\hat{y}_i} + (1-y_i)\ln\frac{1-y_i}{1-\hat{y}_i}\right)}$.

LIKELIHOOD RATIO TEST, AIC, AND BIC

Likelihood ratio test: Testing a null hypothesis H_0 involving r constraints.

$\text{LRT} = 2\,(l_1 - l_0)$ have an approximate chi-square distribution under H_0 with r degrees of freedom.

For a minimal model: use $\mu_i = \bar{y}$ to obtain the log-likelihood function

$\text{LRT} = \Delta D^* = D_0^* - D_1^*$ for $D_0^* = 2\,(l_{\text{SAT}} - l_0)$ and $D_1^* = 2\,(l_{\text{SAT}} - l_1)$.

Akaike Information Criterion: $\text{AIC} = -2l + 2 \times (p+1)$

the smaller the AIC, the better the model

Bayesian Information Criterion: $\text{BIC} = -2l + (p+1)\ln n$

the smaller the BIC, the better the model

GLM:CATEGORICAL RESPONSE VARIABLES

Binomial Response: There are two possible outcomes, True or False.

Link	Distribution Function
Probit	$g(\pi) = \Phi^{-1}(\pi)$
Complementary log-log	$g(\pi) = \ln[-\ln(1-\pi)]$
Logit	$g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$

Logistic regression: $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad \rightarrow \quad \pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}} = \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}}$

Odds: $\pi/(1-\pi)$ the ratio of the probability of occurrence of the event to the probability of non-occurrence

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

Nominal regression: The generalized logit model with one category as the baseline category

$\ln\frac{\pi_j}{\pi_c} = \mathbf{x}'\boldsymbol{\beta}_j = \beta_{0j} + \beta_{1j}x_1 + \cdots + \beta_{pj}x_p, \quad j = 1, 2, \ldots, c.$

$$\pi_c = \frac{1}{1 + \sum_{k=1}^{c-1}\exp\left(\mathbf{x}'\boldsymbol{\beta}_k\right)} \qquad \widehat{\pi}_j = \frac{\exp\left(\mathbf{x}'\widehat{\boldsymbol{\beta}}_j\right)}{1 + \sum_{k=1}^{c-1}\exp\left(\mathbf{x}'\widehat{\boldsymbol{\beta}}_k\right)}, \quad j = 1, 2, \ldots, c.$$

Ordinal regression: Cumulative logit model and Proportional odds model

Cumulative logit model: $\ln \dfrac{\tau_j}{1-\tau_j} = \ln \dfrac{\pi_1 + \cdots + \pi_j}{\underbrace{\pi_{j+1} + \cdots + \pi_c}_{\text{complement of } \pi_1 + \cdots + \pi_j}}$

Proportional odds model: $\ln \dfrac{\tau_j}{1-\tau_j} = \beta_{0j} + \underbrace{\beta_1 x_1 + \cdots + \beta_p x_p}_{\text{does not depend on } j}$

GLM:COUNT RESPONSE VARIABLES

Poisson response: $\ln \mu_i = \underset{\text{offset}}{\ln E_i} + \mathbf{x}_i'\boldsymbol{\beta} = \ln E_i + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

Likelihood function: $L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \dfrac{\mathrm{e}^{-\mu_i}\mu_i^{y_i}}{y_i!} \propto \prod_{i=1}^{n} \mathrm{e}^{-\mu_i}\mu_i^{y_i}$

Loglikelihood function: $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \ln \mu_i - \mu_i) + \text{constants free of } \boldsymbol{\beta}$

Scores: $\sum_{i=1}^{n} (y_i - \widehat{\mu}_i)\,\mathbf{x}_i = \mathbf{0}$

Information matrix: $\mathbf{I}(\boldsymbol{\beta}) = -\mathbb{E}\left[\dfrac{\partial}{\partial \boldsymbol{\beta}'} \sum_{i=1}^{n} (y_i - \mu_i)\,\mathbf{x}_i\right] = \sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'$

Pearson chi-square goodness-of-fit statistic

Pearson chi-square for Poisson response: $Q = \sum_{i=1}^{n} \dfrac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$

Pearson chi-square for Binomial response: $Q = \sum_{j=0}^{m} \dfrac{(n_j - n\widehat{p}_j)^2}{n\widehat{p}_j}$

Overdispersion: The variance of the response variable exceed its mean

Quasi-likelihood: $\mathrm{Var}(y) = \phi v(\mu) = \phi\mu \quad \rightarrow \quad \widehat{\phi} = \dfrac{1}{n-(p+1)} \sum_{i=1}^{n} \dfrac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$

Negative binomial: $\mu = r(1-p)/p$ and $\mathrm{Var}(Y) = \phi\mu$ where $\phi = 1/p$

Zero-inflated models:

$$y_i \begin{cases} = 0, & \text{with probability } \pi_i \\ \sim g_i(\cdot), & \text{with probability } 1-\pi_i \end{cases}$$

$$\text{and } \Pr(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)\,g_i(0), & \text{for } j = 0, \\ (1-\pi_i)\,g_i(j), & \text{for } j = 1, 2, \ldots. \end{cases}$$

$\mathbf{E}[y_i] = (1-\pi_i)\,\mu_i$ and $\mathrm{Var}(y_i) = (1-\pi_i)\,\mu_i + \mu_i^2 \pi_i (1-\pi_i) > \mathbf{E}[y_i]$

$\rightarrow$ handle overdispersion

Hurdle models:

$$y_i \begin{cases} = 0, & \text{with probability } \pi_i \\ \sim \dfrac{g_i(\cdot)}{1-g_i(0)}, & \text{with probability } 1-\pi_i \end{cases}$$

$$\text{and } \Pr(y_i = j) = \begin{cases} \pi_i, & \text{for } j = 0, \\ k_i g_i(j), & \text{for } j = 1, 2, \ldots, \end{cases} \quad \text{with } k_i = (1-\pi_i)\,/\,(1-g_i(0))$$

$$\mathbf{E}[y_i] = k\mu_i \text{ and } \text{Var}(y_i) = \mathbf{E}[y_i] + k(1-k)\mu_i^2$$

$\rightarrow$ handle overdispersion and underdispersion

Heterogeneity models: Poisson with mean $\exp(\alpha_i + \mathbf{x}_i'\boldsymbol{\beta})$.

$$\mathbf{E}[y_i] = \mathbf{E}[\mathbf{E}[y_i \mid \alpha_i]] = \mathbf{E}\left[e^{\alpha_i + \mathbf{x}_i'\beta}\right] = \mu_i \mathbf{E}[e^{\alpha_i}] = \mu_i$$
and $\text{Var}(y_i) = \mu_i + \mu_i^2 \text{Var}(e^{\alpha_i})$ $\rightarrow$ handle overdispersion

Latent class models: Handle both under-dispersion and over-dispersion

Time Series

TREND AND SEASONAL

Deterministic trends: Non-random phenomena and can be predicted with a high degree of certainty

Stochastic trends: Inexplicable and unpredictable in direction and are caused by random variation

Let $\boldsymbol{T_t}$ be the **trend**(Deterministic trend/Stochastic trend), $\boldsymbol{S_t}$ be the **seasonal variation**, and $\boldsymbol{\varepsilon_t}$ be the **Random term**.

Additive model: $y_t = T_t + S_t + \varepsilon_t$

Multiplicative model: $y_t = T_t \times S_t + \varepsilon_t$

Drawbacks: Ignore other sources of information other than the time series itself.

Give the most weight to observations at the beginning and at the end of series.

STATIONARITY AND AUTOCORRELATIONS

Stationary: Properties like mean and variance do not depend on the time.

Weakly stationary: Stationary in mean and variance, the covariance between y_{t-k} and y_t depends on the two time points $t+k$ and t only through the time lag $|k|$.

Lag-k autocorrelation

Ture autocorrelation:
$$\rho_k = \text{Corr}(y_t, y_{t-k}) = \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t)\,\text{Var}(y_{t-k})}} = \frac{\text{Cov}(y_t, y_{t-k})}{\sigma_y^2}$$

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

r_0 is always equal to 1:
$$r_0 = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = 1$$

Correlogram:

Axes: The x-axis of the correlogram gives the lag k and the y-axis represents the sample autocorrelation.

Critical bands: the approximate 95% critical bands for the hypothesis that the population autocorrelations are equal to zero, the two critical bands are horizontal lines positioned at $= \pm 2/\sqrt{T}$

WHITE NOISE AND RANDOM WALKS

A **white noise** time series c_t is a stationary time series: $c_t \overset{iid}{\sim} N\left(\mu_c,\ \sigma_c^2\right)$

Mean: $\mathrm{E}\left[c_t\right] = \mu_c$

Variance: $\mathrm{Var}\left(c_t\right) = \sigma_c^2$

Point forecast: $\widehat{y}_{T+1} = \bar{y}$

Interval forecast: $\bar{y} \pm t_{T-1,\alpha/2}\sqrt{s_y^2(1+\frac{1}{T})}$

Correlogram: The ACF is close to 0 for $k \geq 1$.

A **random walk** is a nonstationary time series: $y_t = y_{t-1} + c_t = y_0 + (c_1 + \cdots + c_t)$

The difference of a random walk is a white noise

Mean: $\mathrm{E}\left[y_t\right] = y_0 + t\mu_c$

Variance: $\mathrm{Var}\left(y_t\right) = t\sigma_c^2$

Point forecast: $\widehat{y}_{T+1} = y_T + l\bar{c}$

Interval forecast: $(y_T + l\bar{c}) \pm 2\sqrt{s_c^2 \times l}$

Correlogram: The ACF will slowly decrease from 1 to 0.

SMOOTHING

Moving Averages

Moving average of length k at time t: $\widehat{s}_t = \dfrac{y_t + y_{t-1} + \cdots + y_{t-k+1}}{k} = s_{t-1} + \dfrac{y_t - y_{t-k}}{k}$

Double moving average smoothing: $\widehat{s}_t^{(2)} = \dfrac{\widehat{s}_t + \widehat{s}_{t-1} + \cdots + \widehat{s}_{t-k+1}}{k}$

Point forecast: $\widehat{y}_{T+l} = \widehat{y}_T + b_{1,T} \times l$ for all $l \geq 0$ and $b_{1,T} = \dfrac{2}{k-1}\left(\widehat{s}_T - \widehat{s}_T^{(2)}\right)$.

Exponential Smoothing

Simple exponential smoothing

$$\widehat{s}_t = \frac{y_t + wy_{t-1} + w^2 y_{t-2} + \cdots + w^{t-1}y_1 + w^t y_0}{1/(1-w)}, \quad t = 0, 1, \ldots$$

$$\widehat{s}_t = \widehat{s}_{t-1} + (1-w)\left(y_t - \widehat{s}_{t-1}\right) = (1-w)y_t + w\widehat{s}_{t-1}.$$

Double exponential smoothing

$$\widehat{s}_t^{(2)} = \frac{s_t + ws_{t-1} + w^2 s_{t-2} + \cdots + w^{t-1}s_1 + w^t s_0}{1/(1-w)} = (1-w)s_t + w\widehat{s}_{t-1}^{(2)}.$$

Point forecast:

$$\widehat{y}_{T+l} = b_{0,T} + b_{1,T} \times l \text{ for all } l \geq 0,$$

$$b_{0,T} = 2\widehat{s}_T - \widehat{s}_T^{(2)}, \text{ and } b_{1,T} = \frac{1-w}{w}\left(\widehat{s}_T - \widehat{s}_T^{(2)}\right)$$

SS one-step prediction error:

$$\text{SS}(w) := \sum_{t=1}^{T}\left(y_t - \widehat{s}_{t-1}\right)^2$$

AUTOREGRESSIVE MODEL

First-order AR model or AR(1): $Y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ The model is stationary if $|\beta_1| < 1$

if $\beta_1 = 0$, then the model reduces to a white noise process.

if $\beta_1 = 1$, then the model is a random walk.

Model properties:

$$\text{E}[y_t] = \frac{\beta_0}{1-\beta_1} \qquad \text{Var}(y_t) = \frac{\sigma_\varepsilon^2}{1-\beta_1^2} \qquad \rho_k = \frac{\text{Cov}(y_t, y_{t-k})}{\text{Var}(y_t)} = \beta_1^k$$

Parameter Estimation:

$$\widehat{\beta}_1 = \frac{\sum_{t=2}^{T}\left(y_{t-1} - \bar{y}_{1,T-1}\right)\left(y_t - \bar{y}_{2,T}\right)}{\sum_{t=2}^{T}\left(y_{t-1} - \bar{y}_{1,T-1}\right)^2} \text{ where } \bar{y}_{1,T-1} = \sum_{t=1}^{T-1} y_t/(T-1)$$

and $\bar{y}_{2,T} = \sum_{t=2}^{T} y_t/(T-1)$

$$\widehat{\beta}_0 = \bar{y}_{2,T} - \widehat{\beta}_1 \bar{y}_{1,T-1}$$

Residuals: $e_t = \text{observed - fitted} = y_T - (\widehat{\beta}_0 + \widehat{\beta}_1 y_{t-1})$

Variance of the error term ε: $s^2 = \frac{1}{T-3}\sum_{t=2}^{T}(e_t - \bar{e})^2$

Point forecast: $\widehat{y}_{T+k} = \beta_0 + \beta_1 \widehat{y}_{T+k-1} = \mu + \beta_1^k(y_T - \mu), \text{ where } \mu = \frac{\beta_0}{1-\beta_1}$

Interval forecast: $\widehat{y}_{T+k} \pm t_{T-3,\alpha/2}\sqrt{s^2 \sum_{j=0}^{k-1} \widehat{\beta}_1^{2j}}$

UNIT ROOT TEST

Dickey–Fuller test:

$$y_t - y_{t-1} = \underbrace{(\phi - 1)y_{t-1}}_{\text{random walk}} + \underbrace{\beta_0 + \beta_1 t}_{\text{linear trend}} + c_t,$$

If $\phi = 1$ (and $\beta_1 = 0$), then the model retrieves a random walk model $y_t = y_{t-1} + \beta_0 + \varepsilon_t$.

If $\phi < 1$ (and $\beta_1 = 0$), then the model becomes an AR(1) model

If $\phi = 0$, then the model reduces to the linear trend model $y_t = \beta_0 + \beta_1 t + \varepsilon_t$

SEASONAL MODELS

Fixed seasonal effects models: $y_t = \beta_0 + S_t + \varepsilon_t = \beta_0 + \sum_{j=1}^{m} [\beta_{1j} \sin(2\pi f_j t) + \beta_{2j} \cos(2\pi f_j t)] + \varepsilon_t$ where $f_j = \frac{2\pi j}{\text{SB}}$

SB: seasonal base, the number of terms in the time series per year, or per whatever seasonal cycle is of interest

Trigonometric functions: $g(t) = a\sin(2\pi ft + b) = \beta_1 \sin(2\pi ft) + \beta_2 \cos(2\pi ft)$

where $\beta_1 = a\,cos\,b$ and $\beta_2 = a\,sin\,b$, $f = \frac{1}{\text{SB}}$

Autoregressive seasonal models: $y_t = \beta_0 + \beta_1 y_{t-SB} + \beta_2 y_{t-2SB} + \cdots + \beta_P y_{t-P\cdot SB}$

Seasonal exponential smoothing

$\widehat{y}_{T+l} = \widehat{\beta}_{0,T} + \widehat{\beta}_{1,T} \times l$

$b_{0,t} = (1 - w_1)\, y_t + w_1\, (b_{0,t-1} + b_{1,t-1})$

$b_{1,t} = (1 - w_2)\, (b_{0,t} - b_{0,t-1}) + w_2 b_{1,t-1}$

FORECAST EVALUATION AND FORECASTING VOLATILITY

To evaluate a forecast, we use out-of-sample validation techniques.

Mean error statistic	Mean absolute error	Mean percentage error
$\text{ME} = \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} e_t$	$\text{ME} = \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \lvert e_t \rvert$	$\text{MPE} = \frac{100}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \frac{e_t}{y_t}$

Mean absolute percentage error	Mean square error
$\text{MAPE} = \frac{100}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \left\lvert \frac{e_t}{y_t} \right\rvert$	$\text{MSE} = \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} e_t^2$

ARCH(p): $\sigma_t^2 = w + \gamma_1 \varepsilon_{t-1}^2 + \cdots + \gamma_p \varepsilon_{t-p}^2 = \sigma_t^2 = w + \sum_{j=1}^{p} \gamma_j \varepsilon_{t-j}^2$

where $w > 0$ and $\gamma_1, \ldots, \gamma_p$ are non-negative

unconditional variance: $\text{Var}(\varepsilon_t) = \dfrac{w}{1 - \sum_{j=1}^{p} \gamma_j}$

Forcast: $\widehat{\sigma}_{T+k}^2 = w + \gamma_1 \widehat{\sigma}_{T+k-1}^2 + \cdots + \gamma_{k-1} \sigma_{T+1}^2 + \gamma_k \varepsilon_T^2 + \cdots + \gamma_p \varepsilon_{T+k-p}^2$

GARCH(p): $\sigma_t^2 = w + \gamma_1 \varepsilon_{t-1}^2 + \cdots + \gamma_p \varepsilon_{t-p}^2 + \delta_1 \sigma_{t-1}^2 + \cdots + \delta_q \sigma_{t-q}^2$

where all the parameters are non-negative with $\sum_{i=1}^{p} \gamma_i + \sum_{j=1}^{q} \delta_j < 1$

unconditional variance: $\text{Var}(\varepsilon_t) = \dfrac{w}{1 - \left(\sum_{i=1}^{p} \gamma_i + \sum_{j=1}^{q} \delta_j \right)}$

Forcast: $\widehat{\sigma}_{T+k}^2 = w + (\gamma_1 + \delta_1)\, \widehat{\sigma}_{T+k-1}^2$

K-NEAREST NEIGHBORS

Classification error rate: Minimize $\frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i \neq \hat{y}_i\}}$ where $1_{\{y_i \neq \hat{y}_i\}} = 1$, if $y_i \neq \hat{y}_i$, else 0.

Bayes classifier: Assigns each case to its most likely class, given the explanatory variables.

Bayes error rate: $1 - \mathrm{E}\left[\max_j \Pr(Y = j \mid \mathrm{X})\right]$

KNN classifier: Identify the K nearest points in the training data whose X values are closest to X_0, where K is a pre-specified.

$\Pr\left(Y = j \mid \mathrm{X} = \mathrm{x}_0\right) \approx \frac{1}{K} \times \sum 1_{\{y_i = j\}}$

As K increases, bias increase and variance decrease.

KNN regression: Use the average of the response values of the K nearest observations.

$\hat{f}\left(\mathbf{x}_0\right) = \frac{1}{K} \times \sum y_i$

DECISION TREE

Decision tree struction

Terminal nodes (leaves): The nodes that are not further split.

Internal nodes: The nodes in the tree that are split into two branches.

Branches: The lines connecting different nodes.

Recursive binary splitting

Regression trees: Minimize $\sum_{m=1}^{|T|} \sum_{i:\mathbf{x}_i \in R_m} \left(y_i - \widehat{y}_{R_m}\right)^2$.

Classification trees: Minimize $\sum_{m=1}^{|T|} \frac{n_m}{n} E_m = \frac{\#\text{ classification errors of whole tree}}{n}$.

More splits increase the flexibility of the model, which will result in overfitting

Cost complexity pruning

Regression trees: Minimize $\sum_{m=1}^{|T|} \sum_{i:\mathbf{x}_i \in R_m} \left(y_i - \widehat{y}_{R_m}\right)^2 + \alpha|\mathrm{T}|$.

Classification trees: Minimize $\sum_{m=1}^{|T|} \frac{n_m}{n} E_m + \alpha|\mathrm{T}|$.

node purity

Classification error rate: $\mathrm{E}_m = 1 - \max_{1 \leq k \leq K} \widehat{p}_{mk}$

Gini index: $G_m = \sum_{k=1}^{K} \widehat{p}_{mk}\left(1 - \widehat{p}_{mk}\right)$

Cross-entropy: $D_m = -\sum_{k=1}^{K} \widehat{p}_{mk} \ln \widehat{p}_{mk}$

Deviance: $D = -2\sum_m \sum_k n_{mk} \ln \widehat{p}_{mk}$

Residual mean deviance = Deviance/(n-|T|)

Pros and cons

Advantages	Disadvantages
easy to explain and understand	not robust
Closer to the way human decisions are made	low predictive accuracy compare to GLM-based models
Easier to handle categorical prediction	
Suitable for graphical representation	

Bagging: It is based on the **bootstrap**, generating additional samples by repeatedly sampling from the existing observations with replacement $\widehat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{*b}(x)$

increase B will not cause overfitting

Out-of-bag: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, for $\widehat{y}_i = \begin{cases}\text{average of } \widehat{f}^{*b}(\mathbf{x}_i)\\ \text{majority vote of } \widehat{f}^{*b}(\mathbf{x}_i)\end{cases}$

Random forest: Similar to bagging but use only m predictors instead of the full set of p predictors to decorrelate the bootstrapped tree.

Typically, $m \approx \sqrt{p}$.

Boosting:

1. Set $\widehat{f}(\mathbf{x}) = 0$ and $r_i = y_i$ for all $i = 1, \ldots, n$
2. (a) Fit a tree $\widehat{f}^b$ with d splits
   (b) Update $\widehat{f}$ by adding a shrunken version of the new tree $\widehat{f}(x) \leftarrow \widehat{f}(x) + \lambda f^b(x)$
   (c) Update the residuals by$r_i \leftarrow r_i - \lambda\widehat{f}^b(\mathbf{x}_i)$
3. $\widehat{f}(\mathbf{x}) = \sum_{b=1}^{B}\lambda\widehat{f}^b(\mathbf{x})$

Three parameters: Number of trees B, B too large will overfit the model, it is selected using cross-validation.

Number of splits d, d is the depth of interaction. if $d = 1$,, the boosted model is additive.

Shrinkage parameter λ, $0 < \lambda < 1$.

PRINCIPAL COMPONENTS ANALYSIS

Principal components: $Z_m = \phi_{1m}X_1 + \phi_{2m}X_2 + \cdots + \phi_{pm}X_p = \sum_{j=1}^{p}\phi_{jm}X_j$

Loading vector: $\phi_m = (\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm})'$

Scores: $z_{im} = \phi_{1m}x_{i1} + \phi_{2m}x_{i2} + \cdots + \phi_{pm}x_{ip} = \sum_{j=1}^{p}\phi_{jm}x_{ij}, \quad i = 1, \ldots, n$

Maximal variance: Maximize $\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2$ subject to $\sum_{j=1}^{p}\phi_{j1}^2 = 1$

Higher-order PCs: It is orthogonal (uncorrelated) with the first principal component.
The variance of the second principal component is always smaller than that of the first principal component.

Biplot: Plotting the first two PCs against each other allows us to visualize the data in a two-dimensional scatterplot.

Approximation formula: $x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}$

The higher the value of M(the number of principal components to use), the more accurate the approximation.

Uniqueness: Principal component loading vectors $\boldsymbol{\phi}_m$'s are unique up to a sign flip.

Proportion of variance explained

Total variance: $\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$

Variance of mth PC: $\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$

PVE: $\text{PVE}_m = \frac{\text{variance of } m\text{th PC}}{\text{total variance}} = \frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2}$

$\text{PVE}_1 \geq \text{PVE}_2 \geq \cdots \geq \text{PVE}_M$

Principal Components Regression

Dimension reduction: reducing the dimension of the model from p to m with linear transformation.

Feature selection: PCR is not a feature selection method, each principal component is defined in terms of all of the original p features, none of which can be removed in general.

Partial least squares a supervised dimension reduction method where the response Y plays no role in determining the principal components.

PLS directions: The first PLS direction $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$, set ϕ_{j1} to the estimated slope coefficient $\left(\widehat{\beta}_1\right)$ of the simple linear regression of Y on X_j, for $j = 1, \ldots, p$.

CLUSTER ANALYSIS

Cluster analysis: Group the observations into a small number of homogeneous clusters, groups of observations that are similar to each other.

K-Means Clustering: Group the observations in a data set into K disjoint clusters in which the observations are relatively homogeneous.

Within-cluster variation: $W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$ where $|C_k|$ is the number of observations in C_k.

Algorithm:

1. Split the observations arbitrarily into k clusters.
2. Determine the *centroid* $\bar{x}_k$ of the k th cluster, which is a vector of the p feature means: $\bar{\mathbf{x}}_k = (\bar{x}_{k1}, \bar{x}_{k2}, \ldots, \bar{x}_{kp})$ where $\bar{x}_{kj} = \sum_{i \in C_k} x_{ij} / |C_k|$
3. Assign each observation to the closest cluster in terms of Euclidean distance
4. Repeat steps 2–3 until cluster assignments do not change

The objective is minimize $\sum_{k=1}^{K} W(C_k)$ where $W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \times \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$

Hierarchical clustering: Consists of a series of fusions of observations results in bigger clusters containing smaller clusters containing smaller clusters.

linkage: the dissimilarity between two groups of observations

Dissimilarity measurement: Euclidean distance and Correlation-based distance

Euclidean distance: The square root of the sum of square differences between coordinates

Correlation-based distance: The correlation between the set of features of two observations and focuses on the shapes of the observations

Complete linkage The maximal dissimilarity between observations in the two clusters

Single linkage The minimal dissimilarity between observations in the two clusters

Average linkage The average of all dissimilarity between observations in the two clusters

Centroid linkage The dissimilarity between the two cluster centroids, it may result in inversions