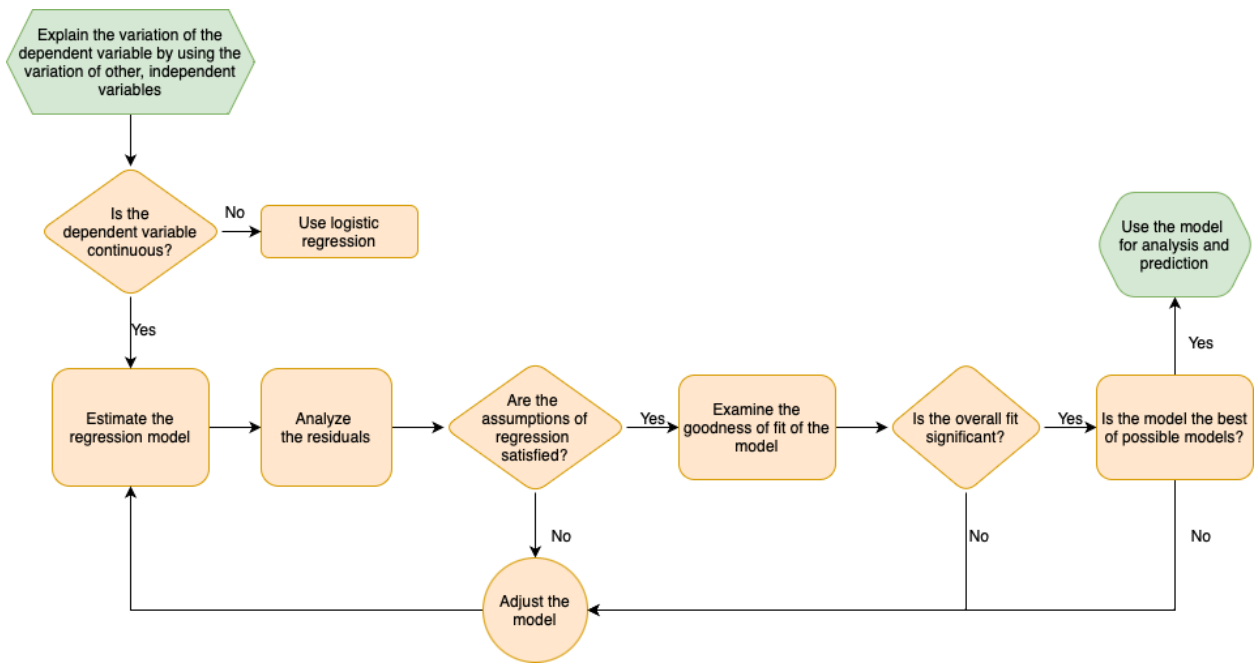
1. Quantitative Methods

▼ Learning Module 1: Basics of multiple regression and underlying assumptions

1. Uses of multiple linear regression
    - Multiple linear regression is used to model the linear relationship between one dependent variable and two or more independent variables.
    - In practice, multiple regressions are used to explain relationships between financial variables, to test existing theories, or to make forecasts.
2. The basics of multiple regression
    - The regression process covers several decisions the analyst must make, such as identifying the dependent and independent variables, selecting the appropriate regression model, testing if the assumptions behind linear regression are satisfied, examining goodness of fit, and making needed adjustments.
    - The regression process



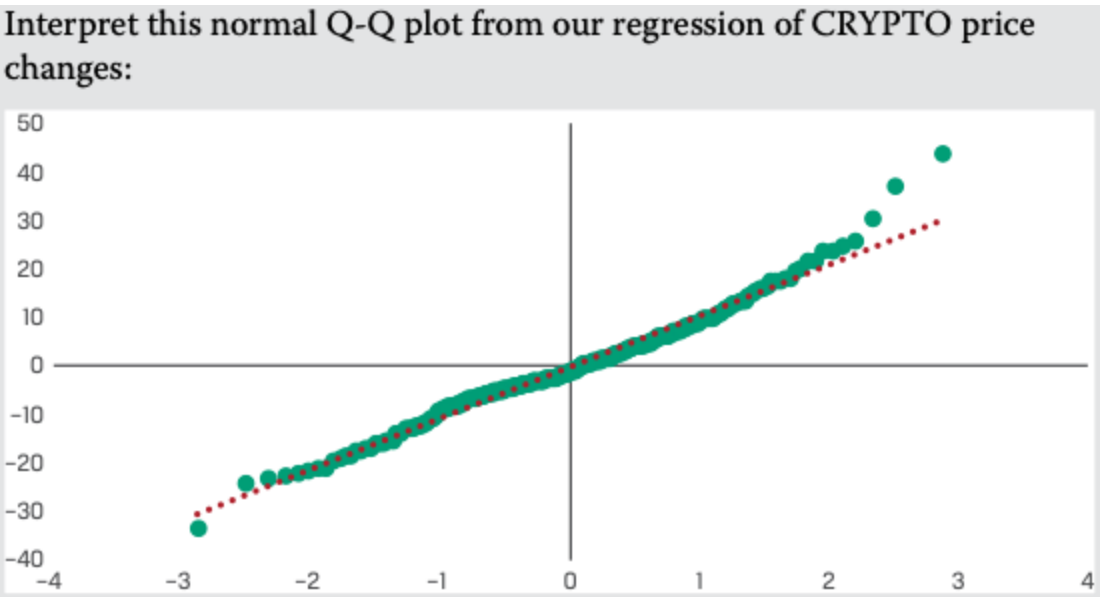
    - A multiple regression model is represented by the following equation:
      $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_kX_{ki} + \epsilon_i, i = 1, 2, 3, ..., n$
      where
      Y is the dependent variable, Xs are the independent variables from 1 to k, and the model is estimated using n observations
    - Coefficient b_0 is the model's "intercept" representing the expected value of Y if all independent variables are zero.
    - Parameters b_1 to b_k are the slope coefficients (or partial regression coefficients) for independent variables X_1 to X_k. Slope coefficient b_j describes the impact of independent variable X_j on Y, holding all the other independent variables constant.
3. Assumptions underlying multiple linear regression

Assumptions	Description	Diagnose
Linearity	The relationship between the dependent variable and the independent variables is linear.	Scatterplots of dependent versus and independent variables (pairs plot) — also identify extreme values and outliers

Assumptions	Description	Diagnose
Homoskedasticity	The variance of the regression residuals is the same for all observations.	Scatter plot of residuals against the dependent variable
Independence of errors	The observations are independent of one another. This implies the regression residuals are uncorrelated across observations.	Scatter plot of residuals against the dependent variable
Normality	The regression residuals are normally distributed.	Normal Q-Q plot
Independence of independent variables	1. Independent variables are not random. 2. There is no exact linear relation between two or more of the independent variables or combinations of the independent variables.	

- Fat-tailed Q-Q plot

Interpret this normal Q-Q plot from our regression of CRYPTO price changes:



Solution

Based on the plot, the residuals are not normally distributed, as indicated by the deviation of residuals from the diagonal evident past ±2 standard deviations, and several potential outliers are also apparent. This normal Q-Q plot suggests the distribution of residuals is "fat-tailed." Note that fat-tailed distributions of residuals are a commonly observed feature of financial data time series.

▼ Learning Module 2: Evaluating regression model fit and interpreting model results

0. Analysis of variance, ANOVA 方差分析

	degrees of freedom, df	sum of squares, SS	mean sum of squares, MS
Regression	k	SSR	MSR = SSR/k
Error	n-k-1	SSE	MSE = SSE/(n-k-1)
Total	n-1	SST	

1. Goodness of fit
   - 一元线性回归中 coefficient of determination $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
   - 多元线性回归中 adjusted $R^2 = 1 - [\frac{n-1}{n-k-1} \times (1 - R^2)]$
     - adjusted R^2 一定小于 R^2，甚至可能小于0
     - R^2 的含义是因变量变化被解释的比率，但是 adjusted R^2 并无此含义
     - R^2 和 adjusted R^2 都不能说明回归系数是否有显著性，也不能说明模型拟合度的显著性，需要通过方差分析和建设检验才能得出结论
   - AIC：用于比较因变量相同的各个模型的拟合优度 AIC = $n \times ln(\frac{SSE}{n}) + 2(k+1)$
   - BIC：BIC = $n \times ln(\frac{SSE}{n}) + ln(n) \times (k+1)$
     - 通常 BIC > AIC
     - 若更关注模型的预测能力，用 AIC；若关注模型的拟合优度，用 BIC
     - AIC 和 BIC 越小越好
2. Testing joint hypotheses for coefficients

- 单个回归系数：t 检验
- 联合假设检验（join hypothesis test）：F 检验
  - unrestricted model：$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i} + \epsilon_i$
  - restricted model：例如 $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \epsilon_i$
  - 联合检验：H_0：$b_3 = b_4 = 0$, H_a：b_3 和 b_4 中至少有一个不等于0（限制条件q=2）

$$F = \frac{(SSE_{restricted} - SSE_{unrestricted})/q}{SSE_{unrestricted}/(n-k-1)}$$

  - 假设检验步骤
    1. State the hypothesis
    2. Identify the appropriate test statistic
    3. Specify the level of significance
    4. State the decision rule
    5. Calculate the test statistic
    6. Make a decision

▼ Learning Module 3: Model misspecification

1. Model specification errors 模型设定错误
   - 模型设定 model specification 是指确定回归方程中选取的变量以及变量的函数形式
   - 模型设定一般遵循以下5条原则
     1. 模型应基于基本的经济理论
     2. 模型应精简，要选取关键变量
     3. 模型应通过样本外 (out of sample) 数据检测，以判断模型是否可以被推广使用
     4. 模型中变量的函数形式应符合变量数据的实际特征
     5. 模型应符合回归假设
2. misspecified functional form

Failures in regression functional form	Explanation	Consequence
Omitted variables 遗漏变量	遗漏了一个或多个重要变量	可能造成异方差和序列相关
Inappropriate form of variables 错误的变量形式	忽略了非线性关系	可能造成异方差
Inappropriate scaling of variables 未使用缩放的数据	变量可能需要通过transform再放进模型中	可能造成异方差和多重共线性
Inappropriate pooling of data 错误融合来自不同样本的数据	把不同样本集放到一起回归	可能造成异方差和序列相关

3. Violations of regression assumptions

	Heteroskedasticity	Serial correlation	Multicolinearity
Description	unconditional 异方差： 残差的方差不恒定，但与自变量不相关 conditional 异方差：残差的方差不恒定，且残差的方差与自变量相关	正序列相关：前一个残差大于 0，后一个残差大于 0 的概率较大。	两个或更多自变量之间高度线性相关
Consequences	可能会造成标准误偏小，容易犯一类错误	- 正序列相关 - 一类错误 - 负序列相关 - 二类错误 - 如果模型自变量中不存在因变量的滞后性，啧序列相关不影响系数估计的一致性；否则会导致系数估计无效。	计算的标准误偏大，容易犯二类错误

	Heteroskedasticity	Serial correlation	Multicolinearity
Testing	- 散点图 - Breusch-Pagan(BP)检验：BP = $n \times R^2_{res}$, 将残差的平方与自变量做回归，单尾检验，拒绝域在右尾	- DW 检验（一阶序列相关） - BG 检验（p阶序列相关）~ $F_{n-p-k-1,p}$	$VIF_j = \frac{1}{1-R^2_j}$ 其中 R^2_j 是将第 j 个自变量作为因变量，与其他k-1 个自变量做线性回归。 VIF > 5 可能存在多重共线，>10 严重多重共线
Correcting	- robust standard errors - heteroskedasticity-consistent standard errors - White-corrected standard errors	- serial-correlation consistent standard errors - serial correlation and heteroskedasticity adjusted standard errors - Newey-West standard errors - Robust standard errors	- 去掉一个或多个共线性的自变量 - 以替代变量来代替一个共线性的自变量 - 增加样本容量 n

▼ Learning Module 4: Extensions of multiple regression

1. Influence analysis 影响力分析
   - 强影响点（Influential observation）
     - 高杠杆点（high-leverage point）：指自变量为极值
     - 异常值（outlier）：因变量为极值
   - 检测方法总结

名称	影响来源	检测指标	计算方法	检测方法
高杠杆点	自变量	杠杆率 h_{ii}	度量某个自变量的第 i 个观测值与其 n 个观测值均值的距离	$h_{ii} > 3(\frac{k+1}{n})$，潜在的高杠杆点
异常值	因变量	学生化残差 t^*_i	1. 用全部样本建模，得到残差标准差s_{e^*}，然后依次剔除第i个样本重新建模 2. $\epsilon^*_i = Y_i - \hat{Y}_{i^*}$ 3. $t_{i^*} = \epsilon^*_i / s_{e^*} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \sim t_{n-k-2}$	$\lvert t^*_i \rvert > t$关键值，潜在的异常值， > 3 则认定为异常值
强影响点	自变量和因变量	Cook's distance D_i	$D_i = \frac{\epsilon^2_i}{k \times MSE} \times \frac{h_{ii}}{(1-h_{ii})^2}$	$D_i > \sqrt{k/n}$，很可能为强影响点 > 1， 很可能 > 0.5，可能

2. 虚拟变量（Dummy variables）
   - intercept dummy：$Y = b_0 + d_0 D + b_1 X + \epsilon$
   - slope dummy：$Y = b_0 + b_1 X + d_1 DX + \epsilon$
3. 定性因变量的多元线性回归 - logistic regression
   - odds = P/(1-P)
   - log odds (or logit) = ln(P/(1-P))
   - $ln(\frac{P}{1-P}) = b_0 + b_1 X_1 + \ldots + \epsilon$
   - $P = \frac{1}{1+exp[-(b_0+b_1 X_1+\ldots+\epsilon)]}$
   - 最大似然估计 MLE 进行回归系数估计
   - 似然比检验 LR test 检验拟合优度

▼ Learning Module 5: Time-series analysis

1. Trend models 趋势模型
   - Linear trend model: $y_t = b_0 + b_1 t + \epsilon_t$
   - Log-linear trend model (exponential trend): $y_t = e^{b_0 + b_1 t}$
     - 增长率为常数，$y_{t+1}/y_t - 1 = e^{b_1} - 1$

- 如果选取的趋势模型能很好地模拟时间序列，那么应当由残差序列不相关。可用DW检验。

2. Autoregressive model 自回归模型
   - 定义
     - AR(1)：$y_t = b_0 + b_1 y_{t-1} + \epsilon_t$
     - AR(p)：$y_t = b_0 + b_1 y_{t-1} + \cdots + b_p y_{t-p} + \epsilon_t$
   - 协方差平稳 covariance stationary
     - 如果时间序列不平稳，那么有关回归方程的系数估计是有偏的，统计推断非有效
     - 对一组时间序列数据，第一步就是判断是否平稳
     - 协方差平稳定义：
       - 均值平稳 $E(y_t) = \mu$
       - 方差平稳 $Var(y_t) = \sigma^2 < \infty$
       - 结构平稳（周期性）$Cov(y_t, y_{t-\tau}) = \gamma(\tau)$
   - AR 模型序列相关性检验
     - kth-order autocorrelation$(\rho_k) = \frac{cov(x_t, x_{t-k})}{\sigma_x^2}$
     - 步骤
       - 构建并估计AR（1）模型
       - 计算模型残差之间的自相关系数
       - 检验残差的各阶自相关系数是否显著不为0
   - 均值复归 Mean reversion
     - 具有均值复归特性的时间序列处于均值水平时，对下一期的预测仍然应当是均值
     - 均值复归水平 $y_t = \frac{b_0}{1-b_1}$
   - 模型预测
     - 如何选择预测模型：根据预测误差进行判断
     - in-sample forecast errors：回归标准差SEE
     - out-of-sample forecast errors：均方误RMSE，越小越好
     - 回归系数不稳定性：选取不同时间段的历史数据可能会得到不同的模型或回归系数
3. Random walk 随机游走 — $b_1 = 1$ 的AR (1)
   - 定义：$y_t = y_{t-1} + \epsilon_t$，$E(\epsilon_t) = 0, E(\epsilon_t^2) = \sigma^2, E(\epsilon_t \epsilon_s) = 0 (t \neq s)$
   - 均值不复归，方差不有限，不满足协方差平稳的条件
     - 处理方法：一阶差分 first-differencing $y_t' = \Delta y_t = y_t - y_{t-1}$
     - 差分后序列平稳
   - 含漂移项的随机游走：$y_t = b_0 + y_{t-1} + \epsilon, b_0 \neq 0$,....
4. Unit Root Test 非平稳的单位根检验
   - 如果一个时间序列有单位根，则序列非平稳。
   - 单位根检验基本思想：如果有 $|b_1| \geq 1$，则时间序列不平稳
   - Dickey-Fuller 检验
     - $\Delta y_t = b_0 + b_1' y_{t-1} + \epsilon_t$
     - $H_0 : b_1' = 0$（非平稳，具有单位根）$H_a : b_1' < 0$
5. Moving-average 移动平均时序模型
   - n-period moving average = $(y_t + y_{t-1} + \cdots + y_{t-(n-1)})/n$
   - MA (q) 定义：$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$, $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = \sigma^2$, $cov(\epsilon_t, \epsilon_s) = 0$ for $t \neq s$
   - S&P BSE 100 指数更适合用 MA 模型（相比于 AR 模型）

- 特征：MA (q) 模型的前 q-autocorrelations 显著不等于0，而后突然变成0. 而 AR 模型的自相关系数是逐渐减小的。
- ARMA 模型 以此类推

6. 季节性因素
   - 存在季节性特征时，AR(1) 残差项会序列相关，需要将滞后 p 阶的时间序列也加入模型（假设考察对象为时间间隔为 p 的数据）
7. ARCH model 自回归条件异方差模型
   - ARCH(p) 定义：$\epsilon_t^2 = a_0 + a_1\epsilon_{t-1}^2 + \cdots + a_p\epsilon_{t-p}^2 + u_t$
8. 多个时间序列的回归
   - Cointegrated 协整

平稳性检验	结论与处理方法
两个时间序列均平稳	直接回归即可
一个平稳，一个非平稳	不能回归
两个时间序列均非平稳 回归后残差项非平稳	不存在协整
两个时间序列均非平稳 回归后残差项平稳	存在协整

   - Cointegrated 协整 — 多个均存在单位根的时间序列之间是否存在协整关系的判断：
     1. 多个时间序列进行回归
     2. 用 Dickey-Fuller 检验残差项是否为平稳序列
        a. 如果无法拒绝原假设（即残差项存在单位根），那么不存在协整
        b. 如果拒绝原假设（即残差项不存在单位根），那么存在协整
9. 时序预测分析步骤

   The following is a step-by-step guide to building a model to predict a time series.

   1. Understand the investment problem you have, and make an initial choice of model. One alternative is a regression model that predicts the future behavior of a variable based on hypothesized causal relationships with other variables. Another is a time-series model that attempts to predict the future behavior of a variable based on the past behavior of the same variable.
   2. If you have decided to use a time-series model, compile the time series and plot it to see whether it looks covariance stationary. The plot might show important deviations from covariance stationarity, including the following:
      a. a linear trend,
      b. an exponential trend,
      c. seasonality, or
      d. a significant shift in the time series during the sample period (for example, a change in mean or variance).
   3. If you find no significant seasonality or shift in the time series, then perhaps either a linear trend or an exponential trend will be sufficient to model the time series. In that case, take the following steps:
      a. Determine whether a linear or exponential trend seems most reasonable (usually by plotting the series).
      b. Estimate the trend.
      c. Compute the residuals.
      d. Use the Durbin–Watson statistic to determine whether the residuals have significant serial correlation. If you find no significant serial correlation in the residuals, then the trend model is sufficient to capture the dynamics of the time series and you can use that model for forecasting.
   4. If you find significant serial correlation in the residuals from the trend model, use a more complex model, such as an autoregressive model. First, however, reexamine whether the time series is

covariance stationary. The following is a list of violations of stationarity, along with potential methods to adjust the time series to make it covariance stationary:

   a. If the time series has a linear trend, first-difference the time series.
   b. If the time series has an exponential trend, take the natural log of the time series and then first-difference it.
   c. If the time series shifts significantly during the sample period, estimate different time-series models before and after the shift.
   d. If the time series has significant seasonality, include seasonal lags (discussed in Step 7)

5. After you have successfully transformed a raw time series into a covariance-stationary time series, you can usually model the transformed series with a short autoregression. To decide which autoregressive model to use, take the following steps:
   a. Estimate an AR(1) model.
   b. Test to see whether the residuals from this model have significant serial correlation.
   c. If you find no significant serial correlation in the residuals, you can use the AR(1) model to forecast.
6. If you find significant serial correlation in the residuals, use an AR(2) model and test for significant serial correlation of the residuals of the AR(2) model.
   a. If you find no significant serial correlation, use the AR(2) model.
   b. If you find significant serial correlation of the residuals, keep increasing the order of the AR model until the residual serial correlation is no longer significant.
7. Your next move is to check for seasonality. You can use one of two approaches:
   a. Graph the data and check for regular seasonal patterns.
   b. Examine the data to see whether the seasonal autocorrelations of the residuals from an AR model are significant (for example, the fourth auto correlation for quarterly data) and whether the autocorrelations before and after the seasonal autocorrelations are significant. To correct for seasonality, add seasonal lags to your AR model. For example, if you are using quarterly data, you might add the fourth lag of a time series as an additional variable in an AR(1) or an AR(2) model.
8. Next, test whether the residuals have autoregressive conditional heteroskedasticity. To test for ARCH(1), for example, do the following:
   a. Regress the squared residual from your time-series model on a lagged value of the squared residual.
   b. Test whether the coefficient on the squared lagged residual differs significantly from 0.
   c. If the coefficient on the squared lagged residual does not differ significantly from 0, the residuals do not display ARCH and you can rely on the standard errors from your time-series estimates.
   d. If the coefficient on the squared lagged residual does differ significantly from 0, use generalized least squares or other methods to correct for ARCH.
9. Finally, you may also want to perform tests of the model's out-of-sample forecasting performance to see how the model's out-of-sample performance compares to its in-sample performance

▼ Learning Module 6: Machine learning

1. 机器学习模型 Evaluation
   a. overfitting & underfitting
   b. in-sample errors & out of sample errors
   c. out of sample errors
      i. bias error：模型在训练样本中的偏差，模型假设过多时可能导致
      ii. variance error：验证样本以及测试样本中表现，过高则过拟合
      iii. base error：数据自身随机性
   d. 一般模型越复杂，bias error越低，variance error越高

Exhibit 4: Learning Curves: Accuracy in Validation and Training Samples

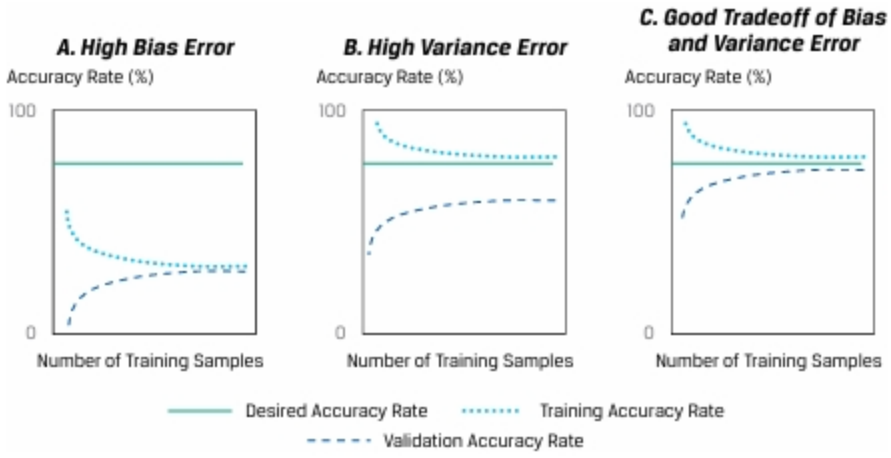
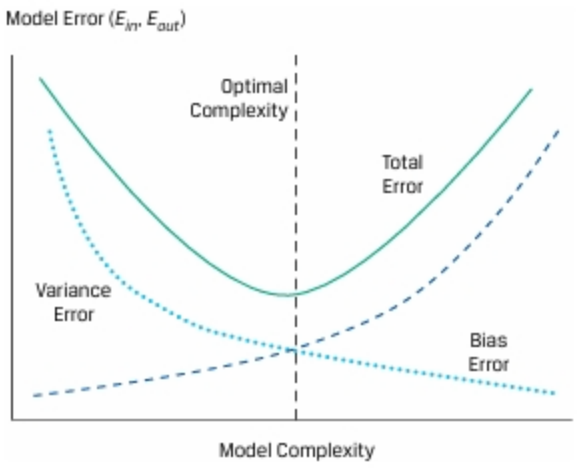


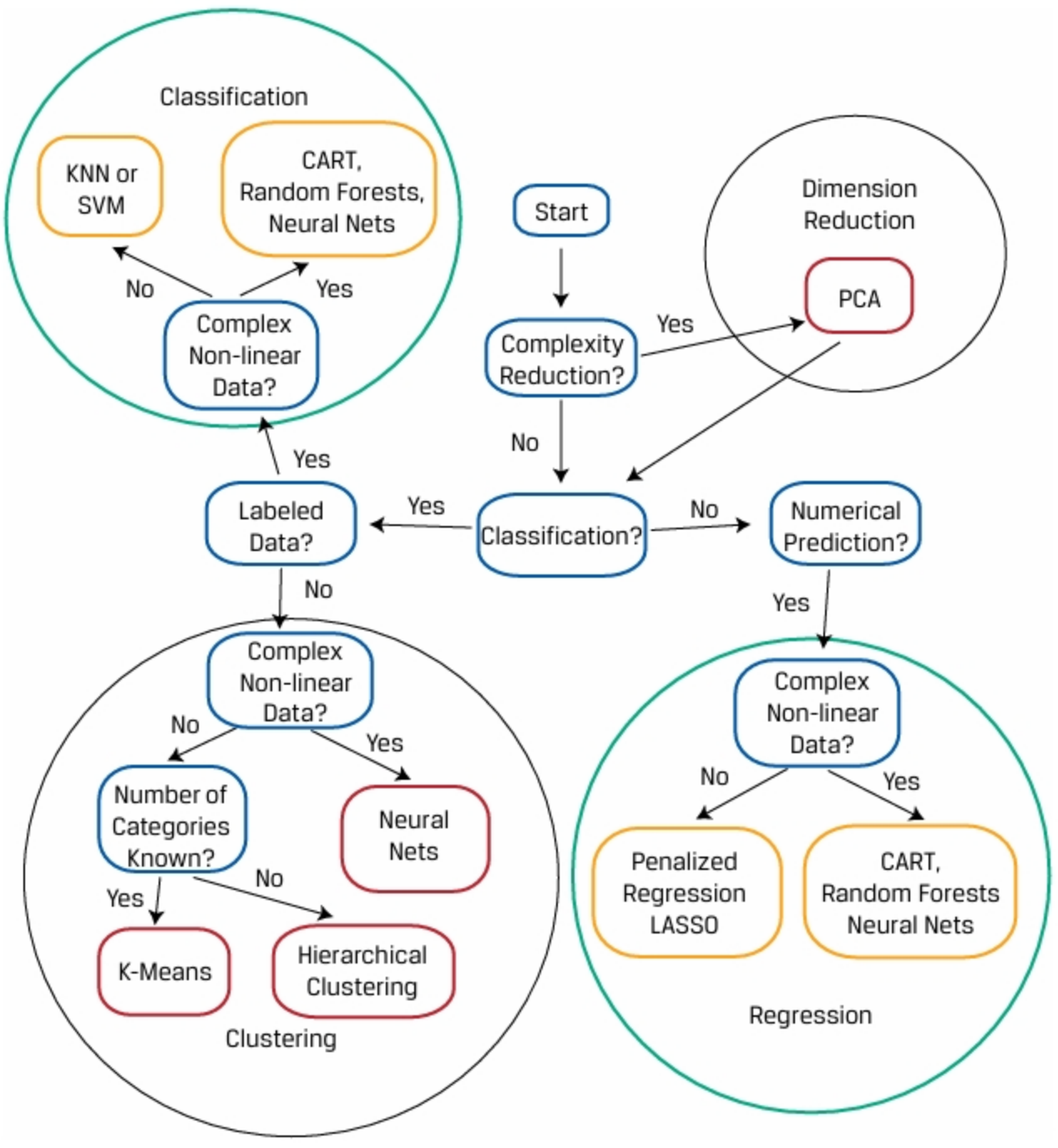
Exhibit 5: Fitting Curve Shows Trade-Off between Bias and Variance Errors and Model Complexity



2. Supervised learning model
   - Penalized regression
     - LASSO：目标函数 $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda \sum_{i=1}^{n} |\hat{b}_i|$
   - SVM
     - maximum margin
     - 分类 or 回归
   - KNN
     - 考察距离新样本点最近的K个样本点，并将新样本点归类为K个样本点中出现次数最多的类别
     - K 的取值不易太低也不宜过高。更适合较少的解释变量。
     - 分类
   - CART 分类回归树
     - root node → decision node → terminal node
     - 构建CART的关键步骤是 bifurcate 分支，将一个节点拆分为两个子节点。当子节点的误差与父节点的误差小于预先设定的阈值时，不再进行分支。
     - 分类 or 回归
   - Ensemble learning and Random forest
     - Voting classifiers：由不同算法的模型组成，通过投票来进行决策
     - Bootstrap aggregating，Bagging：有相同的学习模型组成，通过 bootstrap 得到组合模型
       - bootstrap 有助于防止过度拟合
     - Random forest：对随机森林中的分类树进行分支时，遍历所有尚未进入分支的解释变量的一个随机子集
3. Unsupervised learning model
   - 降维：不影响数据解释能力的情况下降低数据维度
     - PCA：把所有解释变量综合在一起进行正交分解，按照解释力度从高到低逐一分解
   - 聚类
     - Hierarchical Clustering 分层聚类

        - Divisive Clustering 分裂聚类：自上而下
        - Agglomerative Clustering 合并聚类：自下而上
            - dendrograms 图
    - K-means ：自下而上
        1. 随机产生 K 个 centroids 质心
        2. 计算每个数据点到 K 个质心的距离，将每个数据点归类为距离最近的质心所对应的类
        3. 将新的质心定义为归类到原质心的数据点的中点
        4. 如果新质心相比原质心变化很小，则停止算法；否则回到第二步
4. Deep learning model
    - Neural networks 神经网络
        - 如果 hidden layer 很多（如超过20层），那么也叫做 deep learning nets，DLNs
    - Deep learning and reinforcement learning
        - reinforcement learning 是指人工智能（agent）根据周围的环境采取行动，模型根据行动的结果给予人工智能奖励（reward）或惩罚，从而学习行动方式。e.g. AlphaGo

Exhibit 37: Stylized Decision Flowchart for Choosing ML Algorithms



模型类别	分类	连续	分类或连续
Supervised	Logistic SVM KNN CART	Linear regression Panelized regression Logistic regression CART Random forest	Neural network Deep learning Reinforcement learning
Unsupervised	PCA 分层聚类	PCA 分层聚类	Neural network Deep learning

	K-means	K-means	Reinforcement learning

▼ Learning Module 7: Big data projects

1. 大数据特征：volume, variety, velocity, veracity（可靠性）
2. 结构化数据与非结构化数据

	结构化数据	非结构化数据
明确建模的目标	确认模型的输入和输出	文本分析(text problem formulation)，确认模型的输入和输出
数据收集		数据护理(data curation)
数据的准备与整理 1. data preparing/cleaning	- incompleteness error 数据不完整 - invalidity error 无效错误值 - inaccuracy error 数据不准确 - inconsistency error 数据不一致 - non-uniformity error 非标准错误 - duplication error 重复错误	- 删除 html 的标识符 - 删除断点符号 punctuations - 删除数字 - 删除空白
2. data wrangling/data preprocessing	- extraction：从已有特征中构造新的变量 - aggregation：将两个或更多变量加总后得到类似的变量 - filtration：去掉不需要的行 - selection：去掉不需要的列 - conversion：将数据转换为合适的类型	- 将所有文本转化为小写 - 删除停止词 stop words，例如 the, is, a - 词干提取 stemming - 词形还原 lemmatization → bag-of-words
	异常值 trimming or wisorization normalization：$\frac{X_i - X_{min}}{X_{max} - X_{min}}$ standardization：$\frac{X_i - \mu}{\sigma}$	
数据探索 1. 探索性数据分析 Exploratory data analysis, EDA	通过可视化图表发现数据关联	文本探索(text exploration) - 统计单文本词频
2. 特征选择 Feature selection	- 反复迭代的过程 - 在提高模型解释力度和加快算法运行速度上进行抉择	精简文本标记符，噪声通常是出现频率最高或最低的词 - 利用频率删除噪声特征 - 卡方检验筛选特征 - 利用 mutual information 筛选特征
3. 特征工程 Feature Engineer	- 通过已有特征来构建新的特征	- 标记数字 - N-gram：词组 - 命名实体技术 name entity recognition, NER：识别专有名词 - 词性 parts of speech, POS
训练模型 1. 方法选择 method selection	- 监督模型与非监督模型的选择 - 数据的类型 - 数据的大小：观测值数据量较大时用神经网络模型，特征值较多时用支持向量机	
2. 模型表现评估 performance evaluation	- 错误分析 Error analysis（下面表格） - ROC， Receiver operating characteristic - RMSE = $\sqrt{\sum_{i=1}^{n} \frac{Predicted_i - Actual_i}{n}}$	同左
3. 模型调试 tuning	bias: 模型过于简单，欠拟合 variance：模型过于复杂，过度拟合	

- Error analysis
  - confusion matrix

预测\真实	1	0
1	TP	FP (Type I error)

0	FN (Type II error)	TN

- Precision (P) = TP / (TP + FP)
  - 模型预测为1的样本中，有多少真的是1
- Recall (R) = TP / (TP + FN)
  - 真的是1的样本中，被预测为1的有多少
- Accuracy = (TP + TN) / (TP + FP + TN + FN)
- F1 score = 2*P*R/(P+R)
  - P 和 R 的调和平均值
  - 当数据分类分布不均匀时，F1比accuracy更适用，分数越高模型表现越好

- ROC
  - False positive rate FPR = FP / (TN + FP)
  - True positive rate TPR = TP / (TP + FN) = Recall
  - ROC 越往左上凸越好

Exhibit 25: ROC Curves and AUCs