# Regression

*Nishanth Nair*

*01 Nov,2019*

## Overview

### Types of Learning

- Unsupervised learning models a set of inputs. like clustering.
  - Here we learn P(x), the probability distribution of x
- Supervised learning generates a function that maps inputs to desired outputs.
  - For example, in a classification problem, the learner approximates a function mapping a vector into classes by looking at input-output examples of the function.
  - Here we learn P(Y|x)
- Semisupervised learning combines both labeled and unlabeled examples to generate an appropriate function or classifier.
- Reinforcement learning:
  - One-shot decision making versus lifetime-value modeling.
  - Learns how to act given an observation of the world.
  - Every action has some impact in the environment, and the environment provides feedback in the form of rewards that guide the learning algorithm.
  - Transduction tries to predict new outputs based on training inputs, training outputs, and test inputs.

### Supervised Learning family

- Generative Classifier (Bottom-up learning)
  - Build model of each class.
  - Assume the underlying form of the classes, and estimate their parameters (e.g., a Gaussian).
  - Solver more complex problems
  - More flexible predictions
  - e.g., Bayesian networks, HMM, naive Bayes
- Discriminative Classifier (Top down)
  - Build model of boundary between classes.
  - Assume the underlying form of the discriminant, and estimate its parameters (e.g., a hyperplane)
  - Better with small datasets
  - Faster to train

    - e.g, linear regression, ANN, SVM

## Parametric vs Non Parametric

- Parametric ML algorithms (e.g., OLS, decision trees; SVMs, NNs)
    - Model-based methods, such as neural networks and the mixture of Gaussians, use the data to build a parameterized model.
    - After training, the model is used for predictions, and the data are generally discarded.
- Nonparametric (lowess(); knn; some flavors SVMs)
    - "memory-based" methods are nonparametric approaches that explicitly retain the training data and use it each time a prediction needs to be made.
    - The term "nonparametric" (roughly) refers to the fact that the amount of stuff we need to keep in order to represent the hypothesis/model grows linearly with the size of the training set.
    - Very little learning; hyperparameter tuning.
    - Hyperparameter tuning can be readily done in MapReduce.
    - Prediction is challenging.

## Regression

Regression is a method for modeling the relationship between one or more (independent/input) variables and one (dependent/output) variable.

- **The Statistician's Paradigm**:
    - Used by economists, social scientists, sociologists, and other quantitative modelers
    - Goal: understand underlying causal relationships
        - What is the impact of an additional year of schooling on later wages?
        - What is the effect of a $X discount on customers' propensity to adopt a product?
    - General question: What is the causal effect of changes in an independent variable on changes in a dependent variable?
    - Estimated coefficients typically linked to meaning in the real world
    - Consistency and bias are essential
- **The Predictive Paradigm**:
    - More common in computer science and engineering
    - Goal: prediction and pattern recognition
        - How accurately can we predict a person's wages given his years of schooling?
        - How accurately can we predict whether a customer is likely to adopt a product?
    - General question: How accurately can we predict a dependent variable based on independent variables?
    - Goodness-of-fit takes precedence over bias and efficiency.

- Types of regression:
  - Linear regression: assumes a linear relation between independent and dependent variables
  - Bivariate: exactly one independent and dependent variable
  - Multiple: linear regression with multiple independent variables
  - Logistic regression: binary dependent variable
  - Other types:
    - Polynomial regression
      - Defines a linear relationship between y and polynomials of x
    - Kernel regression
      - Kernel regression is nonparametric.
      - y is conditional not only on x but on values of y near x.
      - Essentially an average of outputs for inputs near x.
    - Generalized least squares (GLS)
    - Difference/double difference regression
      - A way of transforming data
      - Best for time series data
      - Finds relationship between changes in independent variables and changes in dependent variable
    - Fixed effects/normalization
      - Similar to difference regressions
      - Implements individual-specific dummy variables on the right hand side of the regression
      - This effectively subtracts the mean of the outcome variable from each observation.

    - Quantile regression: linear regression estimating averages other than the mean
    - Locally weighted regression: kernel regression taking mean of y near x

- **Linear Regression in ML**: For input $x \in R^n$, predict a scalar $y \in R$ as output.So, we need to find parameters $\beta$ that satisfy:

  $$\hat{y} = \beta^T.x$$

  Different ways to approach the problem:

  - OLS: Error Minimization (frequentist approach)
    - Use mean squared error(MSE) as performance measure or objective function. Then the goal is to minimize MSE.
    - This gives us: $\beta = (X^T.X)^-1.(X^T.Y)$
    - This can be used to determine the OLS parameters for the regression.
  - OLS: Maximum Likelihood (Bayesian approach)
    - Try all different values of parameters, and for each value, figure out what the actual error is.
    - Search the space of parameters until you figure out which pairing minimizes the actual error.

- **Frequentist statistics**: Estimate a single value for a parameter and make all predictions based on that estimate.
- **Bayesian statistics**: Consider all possible values for the parameter when making a prediction.

- **Logistic Regression**
  - A way to model the linear relationship between one or more arbitrary independent variables and binary dependent variables
  - can be used for inference and prediction
  - Ordinary least squares regression: continuous outcome variable versus Logistic regression: binary outcome variable
  - **Multinomial logistic regression**: used when the outcome variable can take one of a set of categorical values whose size is greater than two.
  - **Ordered logistic regression**: used when the outcome variable is categorical with rank order (e.g., socioeconomic status)

  Ways to approach the problem:

  - Logit function transforms a continuous infinite scale into a scale between 0 and 1.
  - No easy way to solve for parameters.
  - Use Maximum Likelihood estimation:
    - Picks initial parameter values.
    - Determine likelihood of data, given chosen parameters.
    - Improve parameter estimates incrementally (e.g., Newton's method or gradient descent).
    - Recomputes likelihood of data, given these new parameters.
    - When parameters cease to change significantly, presume we have reached a minimum or maximum.
  - Interpreting parameter $\beta$: For a one unit increase in x, we expect to see a $1 - e^{\beta}$ % change in y.

  What do the predictions represent:

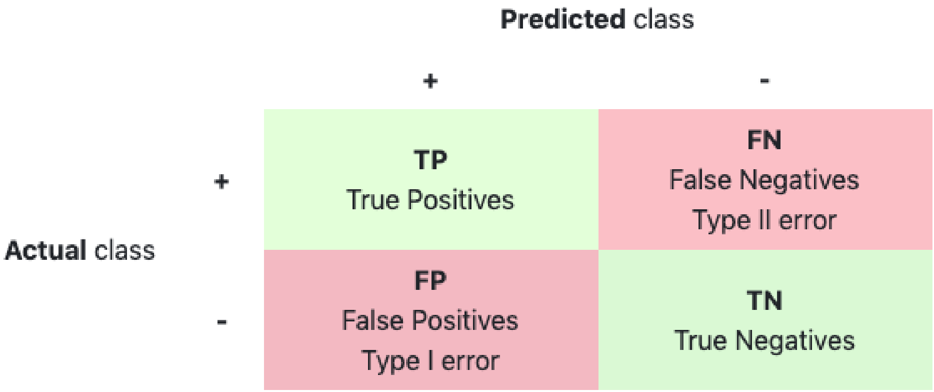
  - They represent the likelihood the actual value is 1.
  - The predicted outcome must be transformed into a binary outcome to be operationalized. This is done by using thresholds.
  - Threshold Operationalization:
    - Increased thresholds are more restrictive.
      - False positives less likely
      - False negatives more likely
    - Decreased thresholds are less restrictive.
      - False positives more likely
      - False negatives less likely
    - Context dictates whether false positive or false negative is preferable. (spam vs terrorist)
    - **Precision** (% positives correctly identifed) : $\frac{TP}{TP+FP}$

- **Recall** (% existing positives identified) : $\frac{TP}{TP+FN}$
- **Optimal point** on ROC (precision/recall) curve
- **Accuracy** : $\frac{TP+TN}{TP+TN+FP+FN}$
- **F score** : $\frac{\text{2.Precision.Recall}}{\text{Precision+Recall}}$

## Classification Metrics

### Confusion Matrix

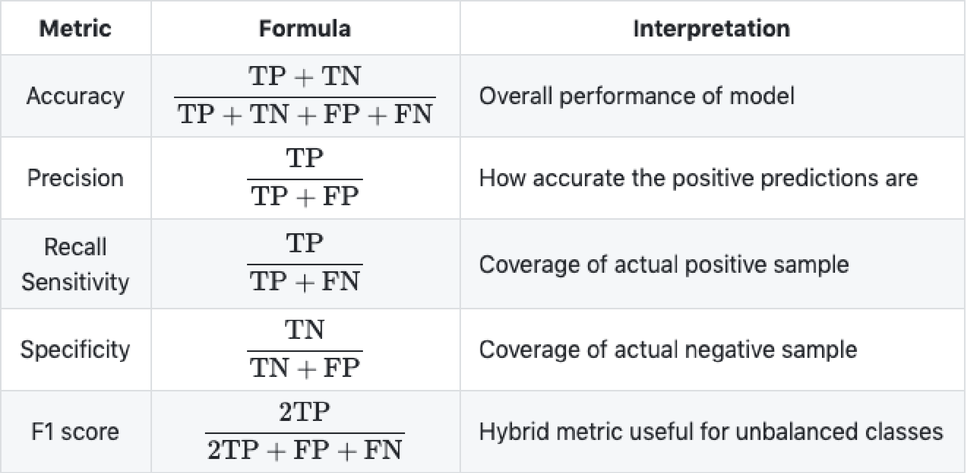
The confusion matrix is used to have a more complete picture when assessing the performance of a model. It is defined as follows:



| | | Predicted class | |
|---|---|---|---|
| | | + | - |
| Actual class | + | **TP**<br>True Positives | **FN**<br>False Negatives<br>Type II error |
| | - | **FP**<br>False Positives<br>Type I error | **TN**<br>True Negatives |

Figure 1: Confusion Matrix

### Performance Metrics

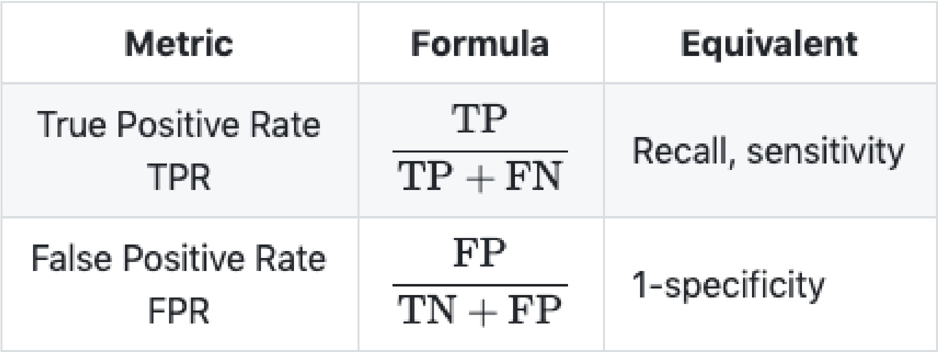


| Metric | Formula | Interpretation |
|---|---|---|
| Accuracy | $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$ | Overall performance of model |
| Precision | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$ | How accurate the positive predictions are |
| Recall<br>Sensitivity | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ | Coverage of actual positive sample |
| Specificity | $\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$ | Coverage of actual negative sample |
| F1 score | $\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$ | Hybrid metric useful for unbalanced classes |

Figure 2: Performance Metrics

### AUC - ROC Curve

ROC (Receiver Operating Curve) is the plot of TPR (True positive rate) versus FPR(False Positive Rate) by varying the threshold where TPR and FPR are given by:



| Metric | Formula | Equivalent |
|---|---|---|
| True Positive Rate<br>TPR | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ | Recall, sensitivity |
| False Positive Rate<br>FPR | $\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}$ | 1-specificity |

Figure 3: Definition

AUC (Area Under the Curve): The area under the receiving operating curve, also noted AUC or AUROC, is the area below the ROC

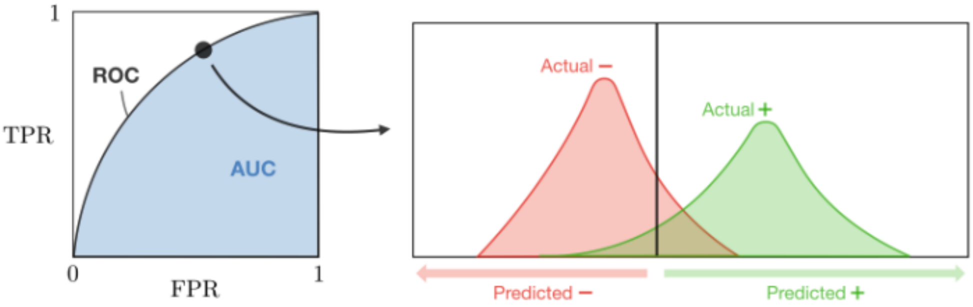
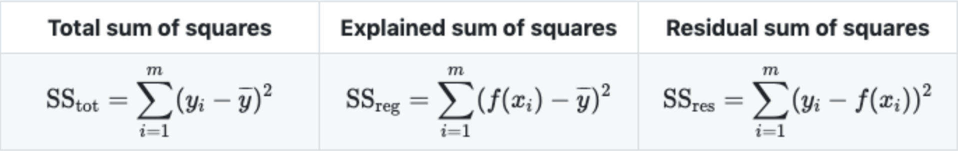


Figure 4: AUC or AUROC curve

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

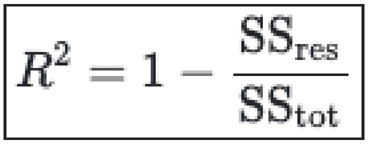
## Regression Metrics



| Total sum of squares | Explained sum of squares | Residual sum of squares |
|---|---|---|
| $\mathrm{SS}_{\text{tot}} = \sum_{i=1}^{m}(y_i - \overline{y})^2$ | $\mathrm{SS}_{\text{reg}} = \sum_{i=1}^{m}(f(x_i) - \overline{y})^2$ | $\mathrm{SS}_{\text{res}} = \sum_{i=1}^{m}(y_i - f(x_i))^2$ |

Figure 5: Performance Metrics

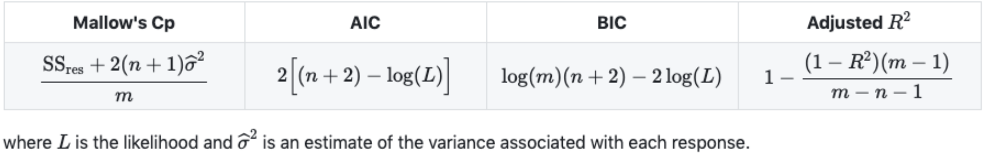
**Coefficient of determination**: The coefficient of determination, often noted $R^2$,sprovides a measure of how well the observed outcomes are replicated by the model.



$$\boxed{R^2 = 1 - \frac{\mathrm{SS}_{\text{res}}}{\mathrm{SS}_{\text{tot}}}}$$

Figure 6: Definition

The following metrics are commonly used to assess the performance of regression models, by taking into account the number of variables n that they take into consideration:



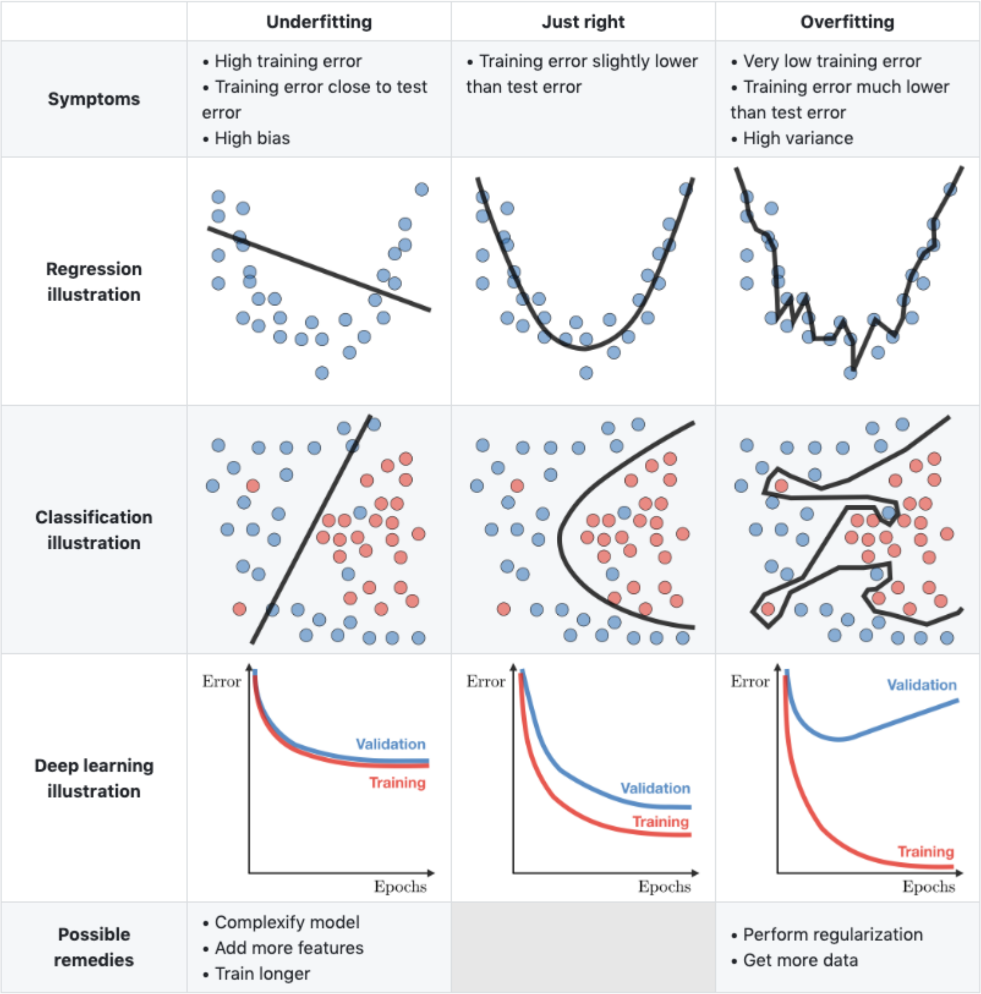
| Mallow's Cp | AIC | BIC | Adjusted $R^2$ |
|---|---|---|---|
| $\frac{\mathrm{SS}_{\text{res}} + 2(n+1)\widehat{\sigma}^2}{m}$ | $2\Big[(n+2) - \log(L)\Big]$ | $\log(m)(n+2) - 2\log(L)$ | $1 - \frac{(1-R^2)(m-1)}{m-n-1}$ |

where $L$ is the likelihood and $\widehat{\sigma}^2$ is an estimate of the variance associated with each response.

Figure 7: Definition

## Diagnostics

- **Bias**: The bias of a model is the difference between the expected prediction and the correct model that we try to predict for given data points.
- **Variance**: The variance of a model is the variability of the model prediction for given data points.
- **Bias-variance tradeoff**:
  - The simpler the model, the higher the bias,
  - The more complex the model, the higher the variance.



| | Underfitting | Just right | Overfitting |
|---|---|---|---|
| **Symptoms** | • High training error<br>• Training error close to test error<br>• High bias | • Training error slightly lower than test error | • Very low training error<br>• Training error much lower than test error<br>• High variance |
| **Regression illustration** | | | |
| **Classification illustration** | | | |
| **Deep learning illustration** | | | |
| **Possible remedies** | • Complexify model<br>• Add more features<br>• Train longer | | • Perform regularization<br>• Get more data |

Figure 8: Bias-Variance Tradeoff

# Statistical interpretations

## Simple Regression Model

- Simple Linear Regression model or two-variable Linear regression Model takes the form

$$y = \beta_0 + \beta_1 x + u$$

where, y is called Dependent\Explained\Response\Predicted variable or Regressand and x is called Independent\Explanatory\Control\Predictor variable or regressor. u is the error or disturbance. $\beta_0$ is the intercept parameter or constant term. $\beta_1$ is the slope parameter

- Assumption to derive a ceteris paribus (all other conditions remaining identical) relation between x and y
  - E(u) = 0
  - E(u|x) = E(u) ( $\implies$ u is mean independent of x)
  - from above 2, E(u|x) = 0 (zero conditional mean)
- Population Regression Function(PRF)

$$E(y|x) = \beta_0 + \beta_1 x$$

- OLS Estimates (Ordinary Least Squares)

$$\hat{\beta_1} = \frac{cov(x_i, y_i)}{var(x_i)} = \hat{\rho_{xy}}.\frac{\hat{\sigma_x}}{\hat{\sigma_y}}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}.\bar{x}$$

- Fitted value of y

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$$

- Residual

$$\hat{u_i} = y_i - \hat{y_i} = y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)$$

- OLS regression Line or Sample Regression Function(SRF).This is the estimated version of PRF

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$$

  From this, $\hat{\beta_1} = \Delta\hat{y}/\Delta x$

- Properties of OLS statistics
  - First Order condition(Method of moments estimation)
    - $\sum_{i=1}^{n} \hat{u_i} = 0$
    - $\sum_{i=1}^{n} x_i.\hat{u_i} = 0$ OR $cov(x_i, \hat{u_i}) = 0$

- $(\bar{x}, \bar{y})$ is always on the OLS regression line.So if we plug in $\bar{x}$, the estimated y value will be $\bar{y}$.
- Total sum of squares, SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Explained sum of squares, SSE = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Residual sum of squares, SSR = $\sum_{i=1}^{n} \hat{u_i}^2$
- SST = SSE + SSR
- Goodness of fit, **R-Squared** or coefficient of determination: Measures the fraction of total variation explained by the regression

  $$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST}$$

  - Always between 0 and 1
  - 100.$R^2$ is the percentage of the sample variation in y explained by x
  - High R-squared only tells us that a lot of the variation in our Y variable is explained by our model
  - R-squared can be considered a measure of predictive accuracy.
  - If looking for inference, understanding an effect, or testing a theory, R-Squared should not be used and will be misleading if used.
  - R-Square is equal to the square of the sample correlation coefficient between $y_i$ and $\hat{y}$.

- Changing units of measurement:
  - If y is multiplied by ‘c’, then OLS intercept and slope are also multiplied by ‘c’
  - If x is multiplied by ‘c’, then OLS slope is divided by ‘c’ and vice versa.OLS intercept is not affected.
  - R-Squared is not affected by change in units.
  - If y is in log form, say log(y), if y is multiplied by ‘c’, then OLS slope remains unchanged but the OLS intecept is now increased by log(c) , i.e it becomes log(c)+ $\beta_0$
  - If x is in log form, say log(x), if x is multiplied by ‘c’, then OLS slope remains unchanged but the OLS intecept is now decreased by log(c) , i.e it becomes $\beta_0$-log(c)
- Specifications: The process of deciding what variables should be in a regression model and what form they should take is called specification.This changes the interpretation of the beta’s.

– Semi-log form: $log(y) = \beta_0 + \beta_1 x + u$. Here,

$$\beta_1 = \frac{\partial log(y)}{\partial x} = \frac{1}{y}.\frac{\partial y}{\partial x} = \frac{\frac{\partial y}{y}}{\partial x}$$

This represents a percentage change in y in response to a unit change in x.So essentially, the percent change is fixed, but the actual value is increasing.To make it clear,the equation can be written as:

$$y = e^{\beta_0 + \beta_1 x + u}$$

– log-log form: In this form, the coefficient would represent the percent change in y given a percent change in x.

– Summary of functional forms:

| Model | Dependent Variable | Independent | Interpretation of $\beta_1$ |
|---|---|---|---|
| Level-Level | y | x | $\Delta y = \beta_1.\Delta x$ |
| Level-Log | y | log(x) | $\Delta y = (\beta_1/100).\%\Delta x$ |
| log-level | log(y) | x | $\%\Delta y = (100\beta_1)\Delta x$ |
| log-log (Constant Elasticity model) | log(y) | log(x) | $\%\Delta y = \beta_1.\%\Delta x$ |

## Multiple Regression Model

- Given by:

  $$y = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + .... + \beta_k.x_k + u$$

  where,

  $$\beta_j = \frac{\partial y}{\partial x_j}$$

- **Regression Anatomy Formula**

Given

$$y = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + .... + \beta_k.x_k + u$$

we can write $x_1$ as

$$x_1 = \delta_0 + \delta_2.x_2 + ... + \delta_k.x_k + r_1$$

, where $r_1$ is the unique variation in x, other variables being "partialled out".

This gives us:

$$y = \gamma_0 + \gamma_1.r_1 + v$$

From the above, we get the regression anatomy formula:

$$\beta_1 = \frac{cov(r_1, y)}{var(r_1)}$$

- **Measure of Fit**($R^2$)
  - Simple $R^2$ gives us the fit for each independent variable.
  - Multiple $R^2$ is the the square of the correlation between the observed Y values and the Y values predicted by the multiple regression model.
  - Interpretation of multiple $R^2$: amount of variation in the outcome variable that is accounted for by the model.
  - Problem with this: As number of variables increase, the value of $R^2$ increases.
  - **Akaike Information Criterion (AIC) or parsimony-adjusted measure of fit**
    - AIC penalizes the model as the number of variables increases.
    - Larger/higher AIC values = worse fit
    - AIC: a way to look at several models (with same data and same dependent variable) to find the most parsimonious model that has a good fit
- Assumptions for OLS regression to work:
  - **Gauss-Markov theorem**: Under certain assumptions, OLS is **BLUE**.
  - BLUE: Best Linear Unbiased Estimator

        - Best: Relative Efficiency
            - OLS coefficients are random variables, and we want them to be as precise as possible.
            - OLS coefficients have smallest variance of all linear unbiased estimators.
        - Linear: OLS estimates are a linear function of the y's.
        - Unbiased: Each $\hat{\beta_j}$ is an unbiased estimator of population parameter $\beta_j$. i.e $E(\hat{\beta_j}) = \beta_j$
    - Assumptions from Gauss-Markov Theorem to show the estimators are unbiased ( i.e the U in BLUE)
        - **MLR.1**: Linear in parameters
            - y is a linear function of x's
            - i.e Any population distribution could be represented as a linear model plus some error (error might be poorly behaved).
        - **MLR.2**: Random Sampling
            - The data is a random sample drawn from the population.
            - All data points follow the population distribution
            - Data points must be iid-independently and identically distributed.
        - **MLR.3**: No perfect collinearity
            - In the sample (and population), none of the independent variables are constant and there are no exact relationships among the independent variables.
            - Rules out only perfect collinearity/correlation between explanatory variables-imperfect correlation is allowed.
            - If an explanatory variable is a perfect linear combination of other explanatory variables, it is superfluous and may be eliminated.
            - Constant variables are also ruled out (collinear with the intercept term).
        - **MLR.4**: Zero-Conditional Mean
            - The value of the explanatory variables must contain no information about the mean of the unobserved factors.
            - $E(u|x_{i1}, x_{i2}...x_{ik}) = 0$
            - This assumption enforces linearity.
            - MLR.1 establishes a linear population model, but MLR.4 ensures that the population actually follows that linear model.
        - **MLR.4'**: Exogeneity (If MLR.4 fails)
            - $Cov(x_j, u) = 0$ for all j
    - **Theorem 3.1**: Unbiasedness of OLS
        - Under MLR.1-4, OLS estimates are unbiased.i.e $E(\hat{\beta_j}) = \beta_j$
            - Note: unbiasedness is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values.

- Troubleshooting theorem 3.1:
  - Random Sampling: Two ways assumption of random sampling can fail
    - **Clustering**: when individuals are collected into groups, and researchers can only access a limited number of these groups, known as clusters. Even with clustering, OLS coefficients are unbiased.Estimates are much less precise under clustering.
    - **Autocorrelation or serial correlation**:This occurs when the error for one data point is correlated with the error for the next data point.The **Durbin-Watson** statistic compares the differences between successive data points to the magnitude of the data points.
  - Multi-Collinearity: This assumption only rules out perfect multicollinearity.
    - The response is simple: drop redundant variables.
    - When variables are highly correlated but not perfectly collinear, OLS will still work but estimates will be much less precise.
  - Zero-conditional Mean:
    - Plot residuals vs predictor.This should ideally be flat.
    - Doing this however will end up creating many plots.A better way would be to plot residuals versus the fitted value.y-axis would have residuals and x-axis the fitted y values.
    - If the conditional mean of the error is not constant, we may be able to change functional form.Adding new variables may fix the zero-conditional mean assumption.
    - If these options fail, we may not be able to meet zero-conditional mean.
  - Exogeneity:
    - If the assumption of zero-conditional mean fails, we may be able to meet a weaker assumption called exogeneity.
    - Explanatory variables that are correlated with the error term are called **Endogenous** (i.e "originates within the system.").
    - Endogeneity is a violation of zero-conditional mean, and its presence implies that OLS coefficients are biased and inconsistent.
    - Explanatory variables that are uncorrelated with the error term are called **Exogenous**.
    - If $x_j$ is exogenous, $Cov(x_j, u) = 0$.
- **Theorem 3.1'**: Consistency of OLS

- - Under MLR.1-3 and MLR.4', the OLS estimators are consistent. i.e $\lim_{n\to\infty}(\hat{\beta_j}) = \beta_j$
  - Our estimators are no longer unbiased, but consistency means the bias goes to zero for large sample size.
  - As long as we have a dataset of a few hundred or thousand observations, researchers generally focus on achieving consistency.
- Causality
  - The population model, $y = \beta_0 + \beta_1.x + u$, is causal if manipulations to x do not affect u. i.e $\beta_1 = \frac{\partial y}{\partial_x}$ as long as $\frac{\partial u}{\partial_x} = 0$
  - Causality vs Exogeneity: They are different
- **MLR.5** (Homoskedasticity)
  - aka Homogenaity of Variance
  - Variance of error term is constant
  - $var(u_i|x_1, x_2...x_k) = \sigma^2$
- **Theorem 3.5(Gauss Markov Theorem)**
  - Under assumptions MLR.1-MLR.5 the OLS estimators are the Best Linear Unbiased Estimators(BLUE) of the regression co-efficients.
- OLS sampling Distributions
  - Normality Assumption
  - Large Sample sizes (OLS Asymptotics)