**Model Selection**

Model Building Concepts

* **Info criteria** 🡪 commonly used to **select vars** for the model
* **Selection algorithm** 🡪 automated technique to **evaluate vars** on some **selection criteria**
  + Step wise selection (forward, backwards, step wise)

Fear of Overfitting

* **Model selection** should **ALWAYS** be done with **training data**
* Will hold out **validation/testing data** to **evaluate** (honest assessment) if there is **overfitting**

i. Information Criteria

AIC and BIC

* Both approximate out-of-sample prediction error by applying penalty for model complexity
  + AIC (Akaike info criterion) 🡪 crude, large-sample approx of leave-one-out cross-validation
  + BIC (Bayesian info criterion) 🡪 favors smaller models/penalizes model complexity more
  + Lower values = BETTER than higher values
  + No amount of lower is “better” enough.
  + May not always agree, but neither is necessarily better

ii. Forward Selection

Process

* Start with a null/empty model (only intercept) and systematically build model one var at a time
  + 2: run all 2 models conditioned that X5 is in the model

A grid of squares with a number in the middle

Description automatically generated

1. Start with EMPTY intercept only model (base model)
2. For each var not in model, create a linear regression model with base model plus this var
3. See which linear regression is best (based on criterion)
4. If better than base model, continue to next step…otherwise STOP (this is your model)
   * Model stops when no other variable added improves the model
5. Update base model to this new model, repeat whole process…

iii. Backward Selection

Process

* Systematically removes variables “not informative” in the model (one variable at a time)
  + Remove one var of the 12, run model with 11 vars and see which has the lowest criterion

1. A blue and white squares

   Description automatically generatedStart with FULL model with all variables (this is the base model)
2. Create models such that each model has exactly one predictor variable removed from it and calculate the criterion for each model
3. See which linear regression is best (based on criterion)
   1. Remove vars with the lowest criterion (listed at the top of the R output)
4. If better than base model, continue to next step… otherwise STOP (this is your model)
5. Update base model to this new model, repeat whole process…
   1. See which model is better as variables are being removed until <none>is listed at the top instead of a var

iv. Step-Wise

Process

* Start with a “null” model (just the intercept) and build the model (one variable at a time), but can also delete variables
  + Combo of forward and backward

A grid of squares with a number in the middle

Description automatically generated

1. Start with EMPTY intercept only model (like in forward selection)
2. For each variable not in model, create a linear regression model with the base model plus this variable
3. For each variable in the model, create models with the base model taking away one variable at a time
   * At each step, you can add or take away a var already there (unlike forward - once a variable is added it cannot be removed)
4. See which linear regression is best (based on criterion)
5. If better than base model, continue to next step… otherwise STOP (this is your model)
6. Update base model to this new model, repeat whole process…
   * After <none>, the next better criterion can be to delete a var already added in the model

Issues with Automatic Search Algorithms

* Automated model selection results in the following:
* Biases in parameter estimates, predictions, and standard errors
* Incorrect calculation of degrees of freedom (p-value method)
* P-values that tend to err on the side of overestimating significance
* (increasing Type I Error probability)
* Can result in locally best model (not global)
  + DO NOT blindly use result from automatic search algorithm as final model!

v. Significance Levels

* A green and white table with numbers

  Description automatically generatedAutomatic stepwise search algorithms can help provide a subset of potential variables
* NO model chosen from a algorithm should be blindly selected as final // always explore other potential models and investigate model assumptions
* If you use p-values for your selection, be sure to adjust your p-values if you have a large sample size

**Diagnostics**

* Visualize data 1st before diagnostics

i. Examining Residuals

Linear Regression Assumptions

* Mean of the Ys is accurately modeled by a linear function of the Xs

Violation of Model Assumptions

* Linear in parameters 🡪 indicates a misspecified model, results are not meaningful
  + If the mean doesn’t follow a linear function of Y, it is a misspecified model
* Constant variance 🡪 does not affect the parameter estimates, but standard errors are compromised
  + Can’t talk about p-values
* Normality 🡪 does not affect the parameter estimates, but affects test results
  + Can still make predictions, but can’t make inferences, use p-value, hypothesis testing
* Independent observations 🡪 does not affect the parameter estimates, but standard errors are compromised
  + Can still make predictions with parameter estimates, but can’t get not hypothesis testing???
* Many assumptions are investigated using residuals
  + Residual = actual Y – Predicted Y
  + How far off (vertical distance) is actual Y from the line

A graph with orange dots

Description automatically generatedExamining Residual Plots

* Y-axis 🡪 plot residuals // X axis 🡪 plot each X or the predicted values
* Residuals should be randomly scattered about 0 reference line
  + NO PATTERN
  + If there is a pattern = misspecified model

ii. Misspecified Model

Examining Residuals Plots for Misspecified Model

* Pattern is detected in residuals // model form is incorrect
* Possible remedies depending on pattern: polynomial terms, interactions, splines …

Polynomial Regression models

* Quadratic polynomial model:
* Cubic polynomial model:
* Polynomial model with cross-product term:

Model Hierarchy

* When adding higher order terms (powers/interactions), should have ALL lower terms
  + If not, all signals from the lower order is put into the higher order
  + E.g. forces everything (squared) into the cubic term
  + If is in the model, should also include and X
* For interaction between and (), then and should also be included
* Polynomial regression example:
  + A researcher is interested in studying the effect of a chemical additive on paper strength. The response var is the amount of force required to break the paper (strength) and the explanatory var is the amount of chemical additive (amount)

When a straight line is inappropriate

* Fit a polynomial/more complex regression model
* Transform the dependent and/or independent var to obtain linearity
* Fit a nonlinear regression model, if appropriate
* Fit a nonparametric regression model

iii. Lack of Constant Variance

* Constant variance assumption is violated
* Possible remedy is transforming vars to stabilize the variance
* Procedures that model the nonconstant variance can be used

A graph with orange dots

Description automatically generated

Homoscedasticity

* Random error term =, , is assumed to have a constant variance,

Heteroscedasticity

* Hypothesis tests and CI based on the t, F, distributions will not be valid

A graph with blue dots

Description automatically generated

* + Does not affect calculation of parameter estimates
  + Does affect standard errors of parameter estimates

Detecting Heteroscedasticity

1. Plotting residuals and looking for patterns
2. Spearman Rank Correlation: uses ranks of data (still between -1 and 1)

Spearman Rank Correlation

* If the Spearman rank correlation coefficient between the **absolute value of the residuals** and the **predicted values** is
  + Close to **zero**, then the variance is potentially **homoscedastic**
  + **Positive**, as the **mean increases** the **variance increases**
  + **Negative**, as the **mean increases** the **variance decreases**
  + Can perform a test:
    - **H0** : variance is **homoscedastic**
    - **HA** : variance is **heteroscedastic**
  + Spearman will NOT discover if there **is a relationship** between the absolute value of residuals and predicted value but **it is not linear**

Accounting for Heteroscedasticity

1. Weighted Least Squared (WLS) or reweighted least squares (IRLS)
2. Transform data
3. Use different distribution (e.g. poisson distribution)

iv. Lack of Normality

Detecting Lack of Normality

* Check that the error terms are Normally distributed by examining:
  + Histogram of the residuals
  + QQ plot (normal probability plot) of the residuals
  + Formal tests for normality
    - H0 : normality
    - HA : not normality

Tests for Normality

1. Anderson-Darling is based on the empirical cumulative distribution function of the data and gives more weight to the tails
2. Shapiro-Wilk test uses the idea of correlation between the sample data and normal scores
   * Better for smaller data sets

Accounting for Lack of Normality

* Depends on why the lack of Normality occurred:
  + Outliers → Robust Regression
  + Nonnormal→ Transformation Needed
  + Can try Box-Cox transformation
    - Transformation on the Y, Box-Cox test helps determine what transformation to use to get normality

Box-Cox Transformation

* Method to determine the best (power) transformation to induce normality
* The Box-Cox transformation has the following form:
  + - (𝑦 𝜆 − 1) / 𝜆 ≠ 0
    - log(𝑦) 𝜆 = 0
* 𝜆 is the power in which the response variable y is raised to
  + - If 𝜆 = 2, then we would square y
    - Exception: 𝜆 = 0 log transformation

v. Correlated Error Terms

Independence

* Know the source of your data:
  + - Clustered/Grouped data
    - Observations connected in some way
    - Complex survey designs
    - Repeated measures
    - Data gathered over time
* Data dependent on time

A graph with blue dots and white text

Description automatically generated

* Obs are not independent
* Residuals follow cyclic pattern
  + - Errors in prediction
* Errors are impacting the next error
  + - Obs are dependent o each other
* Data is collected over time

Detect Non-Independence

* Plots of residuals versus time or other ordering component
* Durbin-Watson can catch collinearity (2/+ vars are closely related)
  + Linear regression will not be appropriate then
* Durbin-Watson statistic or the first-order autocorrelation statistic for time-series data
  + H0: no residual correlation
  + HA: residual correlation
* Durbin-Watson statistic:
  + Bounded between 0 and 4
  + d=2 🡪 FTR H0 , assume there is not enough evidence supporting autocorrelation
  + d < 2 🡪 possible positive autocorrelation (this is the one usually used)
  + d > 2 🡪 possible negative autocorrelation

vi. Influential Point and Outliers

Anomalous Observations

* 2 types of anomalous observations that will be discussed

1. Outliers 🡪 point with large standardized residual (lie far away from the fitted line in the Y-direction)
2. Leverage Points 🡪 point that falls outside the normal range (far from the mean) in the X-space (possible values of the predictors) and have a large “influence” on the regression line

* Observations could be one or both

Residual Analysis

* Don’t only focus efforts on residuals of data.
* Residual analysis only tends to discover outliers instead of leverage points

Diagnostic Statistic

* Standardized residuals (good for detecting outliers)
* Studentized residuals (good for detecting outliers)
  + Studentized residuals are obtained by dividing the residuals by their standard errors (standard errors are found by deleting that one observation).
    - |SR| > 2 for data sets with a relatively small number of observations
    - |SR| > 3 for data sets with a relatively large number of observations
* Cook’s Distance (good for detecting influential observations)
  + Measures the difference in regression estimates when the ith observation is left out
    - Suggested cutoff:
* DFFITS (good for detecting influential observations)
  + DFFITS 🡪 measures the impact that the ith observation has on the predicted value
* DFBETAS (good for detecting influential observations)
  + Measure of change in the jth parameter estimate with deletion of the ith observation
  + One DFBETA per parameter per observation
  + Helpful in explaining on which parameter coefficient the influence most lies
* Hat values (good for detecting influential observations)
  + Using matrix notation, the estimate of the parameters is:
  + Which means the estimated line is:
  + And the hat values:
  + Suggested cutoff is:

vii. Collinearity

Illustration of Collinearity

A diagram of a diamond with lines and points

Description automatically generated A diagram of a diagram

Description automatically generated

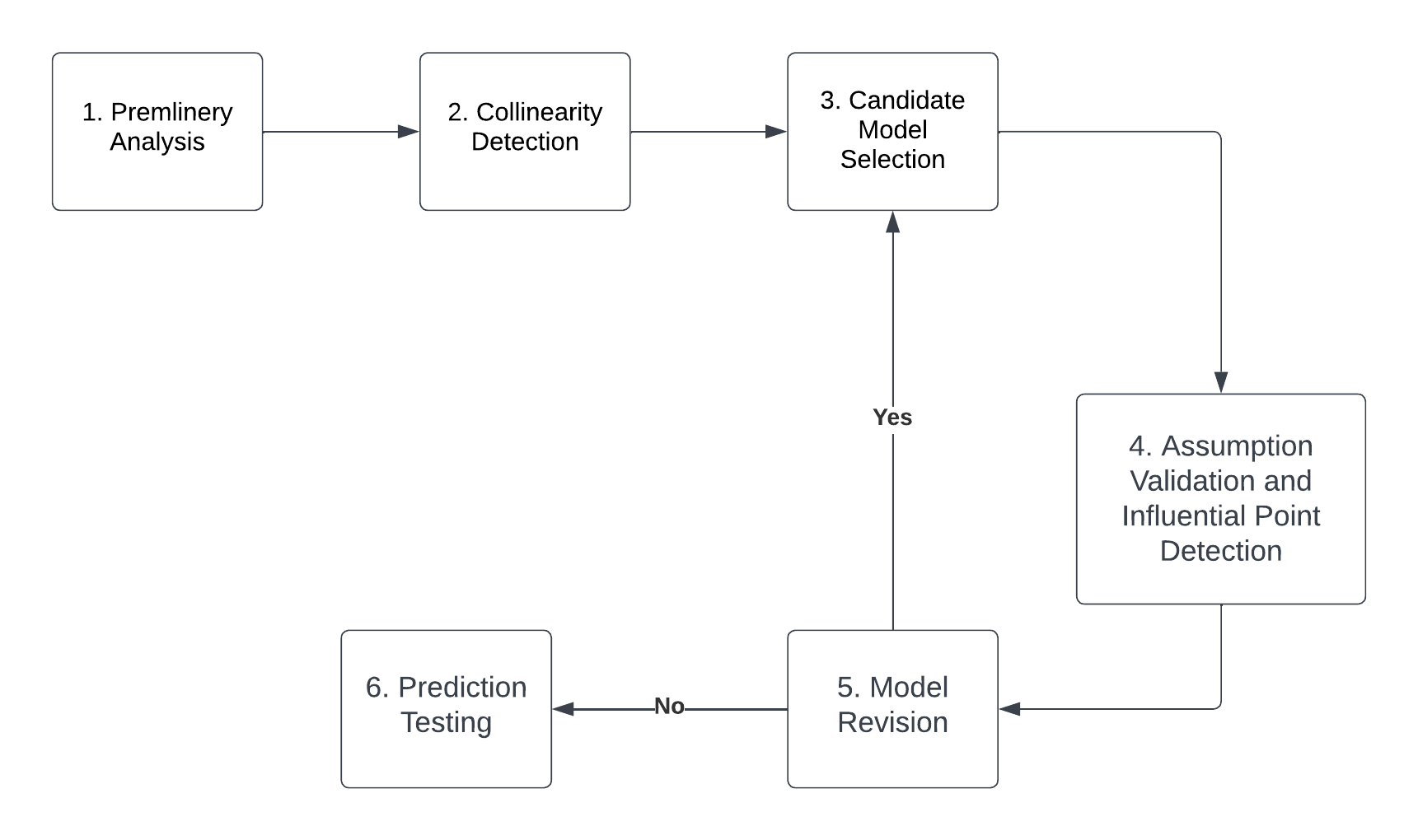
Collinearity Diagnostics

* Looking at correlation matrix of predictors
* One of the most commonly used measures is thenvariance inflation factor (VIF).
* VIF is calculated by:
  + Values of VIF greater than 10 indicate potential collinearity (if using GVIF, then use a cutoff of 5)

Dealing with Multicollinearity

* Exclude redundant independent variables
* Redefine variables
* Use biased regression techniques (e.g. LASSO)
* Center the independent variables in polynomial regression models or models with interaction terms (in R, you can easily do this with the scale(data, scale=F)

Effective Modeling Cyle

[](https://lucid.app/lucidchart/4c8289d0-daae-4f68-a73c-87e8ffd517ea/edit?crop=content&page=0&signature=a80cb2cc497282fc42229dcf35fb8b5cfbaf5067cc186e9579bee4ed739d1899)

**Model Building and Scoring for Prediction**

i. Model Building

Linear Regression Assumptions

* Linear regression is the **best linear unbiased estimator** (BLUE)
  + Estimator 🡪 slope you use to help estimate the predictor

BLUE

* Statistical/estimator biased // does not mean there is sampling bias
* Estimate the true s (can't see) - if on avg is aiming towards the true , it is unbiased
  + Take many samples and find avg
  + That avg would be the truth of what we are estimating
* The statistics have been proven as best linear estimator
* True s2 is a good estimator of σ2 🡪 shooting at the target in the center, not un-centered
* s are centered at the truth // aiming at the true slope that exists
* **What does it mean to be unbiased?**
  + On avg, coefficients from all samples are centered around the true coefficients
* **What does it mean to be best?**
  + IF assumptions hold, is the minimum variance of all unbiased estimators
  + **If assumptions hold**, the spread of s (every sample gives a different sampling distribution spread) won’t be overly wide
  + Best - spread of guesses are as narrow as it get
  + Best & unbiased - how good is your guess
    - If it is aiming at the middle of the right target, the spread isn't wide
  + Might make prediction better at the cost of being unbiased
  + Not aiming at the true slope (getting biased): at the cost of interpretation
  + Can I find a way to make a different relationship between x and y more predictive at the cost of the true relationship --> regularized regression
* What if assumptions don’t hold?
  + What if biased estimators had smaller variance? (you lose interpretability for better predictions)

ii. Regularized Regression

Potential Problems

* As the number of variables increases, more problems tend to arise
  + Assumptions start to fail, multicollinearity concerns
  + Coefficients are changing and not trustworthy // multicollinearity can lead to overfitting
    - Choose prediction or interpretation
  + Regularized regression - can have multicollinearity
    - **Assumption**: at least **10 obs** per var
* Multicollinearity problems 🡪 coefficients vary widely
  + Variations lead to overfitting (only predicting the training data well, but not generalizing to the test dataset)
  + Higher variance than desired
* More vars than observations (genetic modeling)

Regularized Regression

* Regularized regression 🡪 (penalized / shrinkage regression) puts constraints on the estimated coefficients in our model and shrink these estimates to 0
  + Regularized 🡪 penalizing the model, changing the coefficients by shrinking them (making them closer to 0)
  + Moving s bc think might get a better prediction
  + Giving up interpretability of for the hope of being able to predict better
* Coefficients become biased, but potentially improve variance of the model
  + Changing coefficient and makes it biased (no longer aiming for true relationship)
  + Willing to give up interpretability to relax assumptions to make prediction better
  + A diagram of a diagram of a certain type of matter

    Description automatically generated
  + Aim for the center of the board: true β
  + Each sample = single dart thrown at board
  + Aiming for center, but probably won't fit the center
  + Unbiased estimator 🡪 shoot towards the middle, but there is a spread going on
    - Good: least amount of spread you can have when aiming for the center
  + Regularized regression: aiming slightly off target, but hoping of spread of possible errors is smaller // predictions would potentially be better
  + Guess is closer to the truth in an biased way = better prediction
    - If we give up some interpretability, we might be able to predict better

Penalties in Models

* Ordinary least squares (OLS) minimizes the sum of squared errors (SSE)
  + Take all vertical distances, square the, = best line // find minimum of the errors
* Regularized regression introduces a penalty term to the minimization:
  + Dart board: off center = biased // less spread = trade off
  + Still minimizing MSE, and penalize the model
* Penalty 🡪 slightly moving off the center of the target, but still get a better prediction overall // moves too much - aim is farther away from center
  + Penalty is moving the close knit darts farther from the middle
  + Want to make it so there is less spread, but that it is still close to the middle
  + Box-cox - OLS is the best Y to find X and Y relationship
    - Sticking to OLS as best shot, blend and change model to fix it
  + Regularized: don't changed the Y

iii. Ridge Regression

Penalties in Models

* Ridge regression introduces an “L2” penalty term to the minimization:
  + Taking all , sum the effects of those coefficients // make the all positive by squaring the betas (L2 penalty)
  + Maybe make SSE smaller if playing around with betas
* Penalty is controlled by tuning parameter, 𝜆
  + if , the its OLS (there is no penalty)
  + 𝜆 🡪 ∞, coefficients shrink to 0
    - 𝜆 = bias, the bigger the number the more biased the regression will be
    - If , there is no 2 half of the equation, there is just OLS
    - As 𝜆 gets bigger, no longer just OLS
    - Only thing that can change is the slope (X and Y are fixed)
    - Make smaller as 𝜆 gets bigger to counteract

iv. LASSO Regression

Penalties in Models

* Least absolute shrinkage and selection operator (LASSO) regression introduces an “L1” penalty term to the minimization:
* Penalty is controlled by tuning parameter, 𝜆
  + if , the its OLS (there is no penalty)
  + 𝜆 🡪 ∞, coefficients shrink to 0
    - Ridge: approaches 0 asymptotically
* Take all vertical distances, square the, = best line // find minimum of the errors

A graph of a number and a line

Description automatically generated

* Estimates of MSE
* Red dot = avg error across folds
* Interval = variability of the error
* Minimum MSE, find beta that minimizes the error // mathematically best uses all vars
* One standard error above the minimum 🡺 if willing to account for variability pf the dots, you can get something close to the minimum error
  + Can get rid of vars to get close to minimum 0, one SE above to get the close error so model can have less var (trade off for min error)

v. Elastic Net Regression

Penalties in Models

* Both ridge and LASSO have advantages and disadvantages
  + LASSO does variable selection
  + Ridge keeps all variables (LASSO drops arbitrarily)
* Elastic net regression combines both penalty terms in the minimization:
* The glmnet function in R takes slightly different approach:
  + Any value of α between 0 and 1 gives a combination of both penalties (elastic net)

|  |  |
| --- | --- |
| A graph of different colored lines  Description automatically generated | * What happens to betas/coefficient when 𝜆 changes   + As 𝜆 increase, decreases   + All lines (all 36 vars/slopes) get closer to 0, can't interpret s from regularized regression   + Rigged regression helps with multicollinearity, forces it to look more reasonable before it reaches 0 |

* Ridge, shrink vars // Lasso can remove var // Elastic net - best between

vi. Optimizing Penalties

Fear of Overfitting

* Need to select 𝜆 for any of the regularized regression approaches
* Don’t want to minimize variance to the point of overfitting our model to the training data

Cross Validation

* Cross-validation (CV) 🡪 approach to prevent overfitting when tuning a parameter
  + Tune models - change some component of model and see how well it fits after
  + Almost like step-wise, ask enough question, figure out the answer on that dataset, not generalizable
  + Split data into many piece, protecting self from over-fitting on one slice of data
* Concept:

1. Split training data into multiple pieces
2. Build model on majority of pieces
3. Evaluate on remaining piece
4. Repeat process with switching out pieces for building and evaluation

k-fold Cross-Validation

A screenshot of a graph

Description automatically generated

* Hold out 1st 10% of data and build model on the 90%
* Put 10% back in and take out another 10% put and build it on the 90% until completed for all slices

vii. Model Comparison

Comparing Models

* The model results in a formula or rules.
* The data require modifications:
  + Derived inputs
  + Transformations
  + Missing value imputation
* To score/compare, you do not rerun the algorithm!
* Apply score code (equations) obtained from final model to the test data for comparing
  + Get to test data set, score the equation // plug in Xs to predict Y to get predictions
  + Plug in new x values for the equation you get to score // don’t rerun, can overfit
* Test dataset is for comparing final models and reporting final metrics
* DO NOT GO BACK AFTER TO REBUILD MODEL!
* DO NOT JUST BUILD 1000’s OF MODELS TO COMPARE IN THE TEST SET!
* We do not want to fit to the test dataset as it is our honest assessment of how good our models can do

Model Metrics

* Root MSE:
  + Note easily interpretable
* Mean absolute error (MAE):
  + Not scale invariant
* Mean absolute percentage error (MAPE):
  + Not symmetric