Mid-term Review

Today's Agenda

- Logistic Regression
- Mid-term review

Today's class mandatory steps

- Canvas → Modules → Week6
- Download "**logistics_regression _code_complete.R**"
- Place the file in

  "**oba_455_555_ddpm_r/rproject/ k. logistics_regression**"
- Open RStudio project
- Open "**logistics_regression _code_complete.R**" file within RStudio

Results

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.84157	0.54068	-3.406	0.000659	***
Day_WeekTue	-0.67940	0.25773	-2.636	0.008386	**
Day_WeekWed	-0.47836	0.25075	-1.908	0.056429	.
Day_WeekThu	-0.73454	0.24043	-3.055	0.002250	**
Day_WeekFri	-0.21699	0.22799	-0.952	0.341217	
Day_WeekSat	-1.49640	0.34040	-4.396	1.10e-05	***
Day_WeekSun	-0.20009	0.25419	-0.787	0.431180	
Dep_Hour7	0.04760	0.42763	0.111	0.911363	
Dep_Hour8	0.28277	0.40780	0.693	0.488044	
Dep_Hour9	-0.51082	0.53187	-0.960	0.336842	
Dep_Hour10	-0.61237	0.52950	-1.156	0.247482	
Dep_Hour11	-0.20855	0.57692	-0.361	0.717728	
Dep_Hour12	0.19174	0.41037	0.467	0.640333	
Dep_Hour13	-0.45058	0.44891	-1.004	0.315508	
Dep_Hour14	0.61125	0.36355	1.681	0.092695	.
Dep_Hour15	0.70128	0.38754	1.810	0.070360	.
Dep_Hour16	-0.04023	0.39993	-0.101	0.919865	
Dep_Hour17	0.36409	0.35760	1.018	0.308607	
Dep_Hour18	0.10559	0.53913	0.196	0.844719	
Dep_Hour19	0.80912	0.40411	2.002	0.045260	*
Dep_Hour20	0.84016	0.51545	1.630	0.103110	
Dep_Hour21	0.76004	0.37590	2.022	0.043181	*
OriginBWI	0.58962	0.39020	1.511	0.130772	
OriginDCA	-0.23702	0.35701	-0.664	0.506743	
DestinationEWR	-0.23076	0.30188	-0.764	0.444635	
DestinationJFK	-0.51075	0.24129	-2.117	0.034279	*
CarrierCO	1.45615	0.49514	2.941	0.003273	**
CarrierDH	1.07403	0.47128	2.279	0.022668	*
CarrierDL	0.29343	0.28149	1.042	0.297213	
CarrierMQ	1.34045	0.28232	4.748	2.06e-06	***
CarrierOH	0.16358	0.76850	0.213	0.831439	
CarrierRU	0.98956	0.45567	2.172	0.029881	*
CarrierUA	0.20541	0.80356	0.256	0.798236	
Weather	17.86962	465.82175	0.038	0.969399	

High level Insights & Grouping

- Excessive variables
- Most of the variables are insignificant
- What can be done to improve the model exposition?
- Group into broader categories
    - Day_Week to weekend or weekday
    - Hours to morning (6-12pm), afternoon (12pm – 5pm) and evening (5pm-10pm)
    - Insignificant carriers into one group

Results

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5146	0.2778	-9.051	< 2e-16	***
Day_Typeweekday	0.3494	0.1716	2.036	0.04175	*
Time_Dayafternoon	0.3399	0.1747	1.946	0.05163	.
Time_Dayevening	0.6363	0.1783	3.568	0.00036	***
OriginBWI	0.4554	0.2754	1.653	0.09830	.
OriginDCA	-0.1679	0.1672	-1.004	0.31542	
DestinationEWR	-0.3151	0.1950	-1.616	0.10605	
DestinationJFK	-0.4566	0.2185	-2.089	0.03670	*
Carrier_NewCO_DH_MQ_RU	0.9750	0.2034	4.794	1.63e-06	***
Weather	18.0735	466.1000	0.039	0.96907	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- Flights that operate on **weekdays** have delays with an odds of **1.4182**($2.718^{0.3494}$) relative to Flights that operate on a **weekend**
- Flights that leave during the **evening** have delays with an odds of **1.8894**($2.718^{0.6363}$) relative to Flights that leave during the **morning**
- Flights that arrive at **JFK** have delays with an odds of **0.6334**($2.718^{-0.4566}$) relative to Flights that arrive to **LGA**

Confusion Matrix and Accuracy

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 533 120
         1   0   7

               Accuracy : 0.8182
                 95% CI : (0.7866, 0.8469)
    No Information Rate : 0.8076
    P-Value [Acc > NIR] : 0.2624

                  Kappa : 0.0861

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.00000
            Specificity : 0.05512
         Pos Pred Value : 0.81623
         Neg Pred Value : 1.00000
             Prevalence : 0.80758
         Detection Rate : 0.80758
   Detection Prevalence : 0.98939
      Balanced Accuracy : 0.52756

       'Positive' Class : 0
```

Comparison before and after grouping

Before Grouping

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 532 118
         1   1   9

               Accuracy : 0.8197
                 95% CI : (0.7882, 0.8483)
    No Information Rate : 0.8076
    P-Value [Acc > NIR] : 0.2309

                  Kappa : 0.1063

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99812
            Specificity : 0.07087
         Pos Pred Value : 0.81846
         Neg Pred Value : 0.90000
             Prevalence : 0.80758
         Detection Rate : 0.80606
   Detection Prevalence : 0.98485
      Balanced Accuracy : 0.53449

       'Positive' Class : 0
```

After Grouping

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 533 120
         1   0   7

               Accuracy : 0.8182
                 95% CI : (0.7866, 0.8469)
    No Information Rate : 0.8076
    P-Value [Acc > NIR] : 0.2624

                  Kappa : 0.0861

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.00000
            Specificity : 0.05512
         Pos Pred Value : 0.81623
         Neg Pred Value : 1.00000
             Prevalence : 0.80758
         Detection Rate : 0.80758
   Detection Prevalence : 0.98939
      Balanced Accuracy : 0.52756

       'Positive' Class : 0
```

Can we apply Linear Regression to Classification?

- Technically YES
- Treating Y (which is 0 or 1) as continuous
- Often referred to as "Linear Probability Model."
- What is the problem with this model?
- The predictions can be beyond the range of 0 to 1
- What does it mean to have probability beyond the range of 0 to 1?

Midterm2 (20%)

- Canvas quiz
  - **Thursday 12th May 2022, 8 am - 9:45 am (105 minutes)**
  - 49 questions, 60 points
  - Path: Canvas → Assignments → Midterm2
- Content
  - Linear regression, Logistics regression
  - Model evaluation (classification & regression) and Cross-validation
- Open book
- Exam in class

Linear Regression

- Rudimentary model in Supervised Learning
- Predicting a numeric response
- Goal : Fit a relationship between
    - numeric output variable Y & set of "p" input variables $X_1, X_2, X_3, \cdots\cdots X_p$
- Output variable Y is also referred as
    - Response / Target / Outcome variable
- Input variables $X_1, X_2, X_3, \cdots\cdots X_p$ are also referred as
    - Predictors / Independent variables / Regressors / Covariates

Linear Regression

- Predict "Y" using a linear combination of predictors $X_1, X_2, X_3, \cdots \cdots X_p$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Noise or Unexplained part

- Information available on both $X's$ & Y
- $\beta_0, \beta_1, \beta_2 \cdots \cdots \beta_p$ are coefficients
- Required to estimate the coefficients
- Underlying estimation process : **Ordinary Least Squares (OLS)**

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

- Estimated values are generally represented by hat $\widehat{\ }$

Types

- Simple Linear Regression (p = 1)

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- Multiple Linear Regression (p > 1)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Regression modeling includes **estimating coefficients**, and **choosing which predictors (X's) to include and in what form**
- E.g., A transformed numerical predictor can be included (E.g., $\log X_1$) in the regression

Multiple Linear Regression model

$$\text{price} = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ km} + \beta_3 \text{ fuel\_type} + \beta_4 \text{ hp} + \beta_5 \text{ metcolor} + \beta_6 \text{ automatic} + \beta_7 \text{ cc} + \beta_8 \text{ doors} + \beta_9 \text{ quarterly tax} + \beta_{10} \text{ weight} + \epsilon$$

Data Partition : Training & Validation

Entire Data

Train
- Model building

- Use **train** data for model building
- Employ the model to predict the **validation** data output
- Generate error/accuracy measures from **validation** data

Validation
- Assessment

- Quantify & assess the model performance
- Run the model on the **entire** data
- Employ the model to predict for new data

New data
- Predict output

Why this process?

Model built directly using the **entire** data does not demonstrate/quantify the predictive ability of new data

Is the Regression overall significant?

```
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624
automatic        1.320e+02  1.880e+02   0.702   0.4827
cc              -3.994e-02  9.185e-02  -0.435   0.6638
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

Regression on training data

If p-value < 0.05, then at minimum one of the predictors impacts price

Significance of individual predictors

```
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624
automatic        1.320e+02  1.880e+02   0.702   0.4827
cc              -3.994e-02  9.185e-02  -0.435   0.6638
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

Effect of predictors are **insignificant** if you see "." or no stars

Impact of individual predictors

```
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624
automatic        1.320e+02  1.880e+02   0.702   0.4827
cc              -3.994e-02  9.185e-02  -0.435   0.6638
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

Coefficients (All β^{s})

Interpreting character predictor

```
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624
automatic        1.320e+02  1.880e+02   0.702   0.4827
cc              -3.994e-02  9.185e-02  -0.435   0.6638
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

What is the base category in the **fuel_type** predictor?

Model fit

```
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age            -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km             -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp              2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color      -4.311e+00  9.143e+01  -0.047   0.9624
automatic       1.320e+02  1.880e+02   0.702   0.4827
cc             -3.994e-02  9.185e-02  -0.435   0.6638
doors          -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax   8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight          2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

Multiple R-Square (R^2) : Proportion of variation in price explained by predictors

Predictor selection in Linear Regression

- Kitchen-Sink approach
    - Use all the variables
- Problems
    - Expensive and Time consuming
    - Unstable (Multi-collinearity, large standard errors……)
    - Including uncorrelated predictors (insignificant) can increase the variance of predictions
    - Dropping correlated predictors (significant) can increase the average bias of predictions

How to reduce number of predictors ?

- Domain knowledge
  - Experienced individuals in the industry sometimes can provide a more valuable information
- Computational power
  - Exhaustive search
  - Subset selection algorithms

Exhaustive Search

- Evaluate all combinations of predictors
- For "n" predictors, how many models can you run with different combinations of X's
  - 2^n-1
- Three predictors $\mathrm{X}_1, \mathrm{X}_2$, X_3
  - 7 models
  - $\mathrm{Y}\sim\mathrm{X}_1$, $\mathrm{Y}\sim\mathrm{X}_2$, $\mathrm{Y}\sim\mathrm{X}_3$, $\mathrm{Y}\sim\mathrm{X}_1 + X_2$,$\mathrm{Y}\sim\mathrm{X}_1 + X_3$,$\mathrm{Y}\sim\mathrm{X}_2 + X_3$, $\mathrm{Y}\sim\mathrm{X}_1 + X_2 + X_3$
- Choose the model based on one of the performance measures
  - High Adjusted R-Square (R^2)
  - Akaike Information Criterion (AIC) , Bayesian Information Criterion (BIC)
  - Mallow's $\mathrm{C_p}$

Algorithms

- Backward Elimination
  - Step 1 : Run a regression with all the predictor variables
  - Step 2 : Drop the insignificant predictor with the highest p-value
  - Step 3 : Run a regression model with the remaining predictors
  - Step 4 : Repeat steps 2 & 3 until all the predictors are significant
- Forward Selection
  - Step 1 : Run list of regression models with each individual predictor separately
  - Step 2 : Choose the model among the list with highest R^2
  - Step 3 : Run list of regression models by incrementally advancing Step 2 model by adding remaining predictors individually
  - Step 4 : Repeat steps 2 & 3 until all predictors are significant in the model and all exhaustive combinations are executed

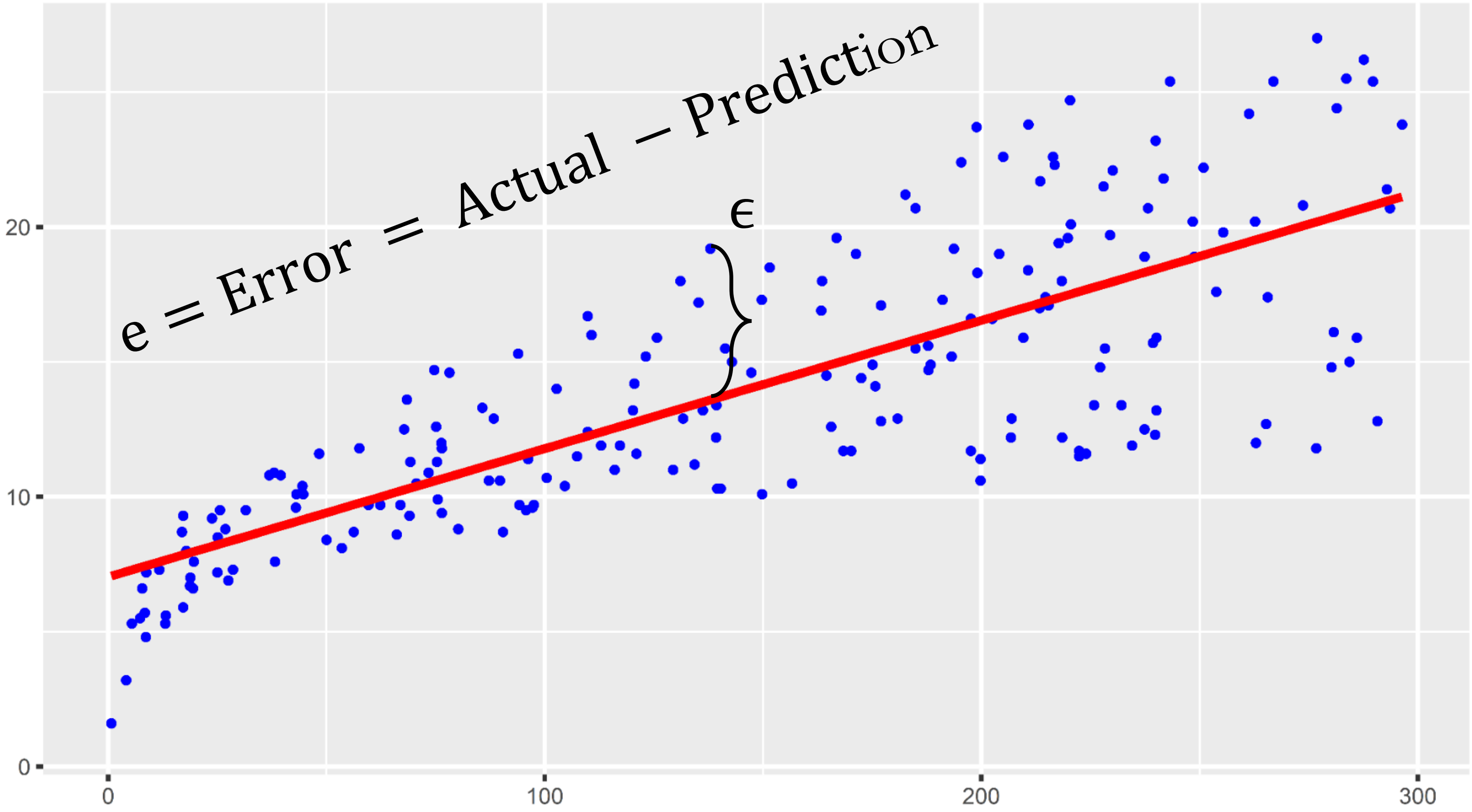
Steps for building Regression model

- Step 1 : Partition the data into training and validation
- Step 2 : Build the Regression model on the training data
- Step 3 : Use the model from Step 2 to predict the output in validation data
- Step 4 : Compute error as difference between actual output and predicted output in the validation data
- Step 5 : Develop accuracy measures using errors

Accuracy Measures
Regression

Error

- Error (e_i) for each observation i
- Error (e_i) : Difference between actual (Y_i) and predicted outcome ($\hat{Y}_i$)



Error measures for Regression

- Mean Error (ME) : $\frac{1}{n}\sum_{i=1}^{n} e_i$
  - ➢ Indicates on-average predictions are over or under the outcome
- Mean Absolute Error (MAE) : $\frac{1}{n}\sum_{i=1}^{n}|e_i|$
  - ➢ Magnitude of average absolute error
- Mean Percentage Error (MPE) : $\left(\frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{Y_i}\right) * 100$
  - ➢ Measure relative to the size of outcome Y_i
- MAPE (Mean Absolute Percentage Error) : $\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{Y_i}\right|\right) * 100$
- Root Mean Square Prediction Error (RMSE) : $\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$
  - ➢ Similar to standard error and has same units as outcome Y_i

Error/Accuracy Measures

- We computed error measures for **validation** data
- Can they be computed for **training** data?
- What do the measures infer for each data?

Training

- Goodness-of-fit
- Additional measures - R^2, standard error
- Does not indicate predictive abilities

Validation

- Indicates predictive abilities
- Used to <u>**compare across models**</u> to assess their degree of prediction accuracy

- **Overfitting** can be detected by comparing the error measures between **training** and **validation** data
- Greater the difference in train & validation data error measures, greater the overfitting

Logistic Regression Model

- Predict a categorical outcome

- Logistic response function

Value of **e** is 2.718

$$p = \Pr(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q)}}$$

- Odds : Ratio of probability of belonging to class 1 to probability of belonging to class 0

$$\text{Odds}(Y = 1) = \frac{p}{1 - p} \qquad \text{Odds}(Y = 0) = \frac{1 - p}{p}$$

$$\log(\text{Odds}(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q$$

- Estimation methodology : **Maximum Likelihood Estimation**

Steps for building Logistics Regression Model

- Step 1 : Partition the data into training and validation
- Step 2 : Build the Logistics Regression model on the training data
- Step 3 : Use the model to predict the probability that each observation in validation data belongs to a Class1 (assume the data has two classes)
- Step 4 : Set the cutoff value (0.5) and classify the record into a class
  - If $p \geq 0.5$, observation is classified to category "Class1"
  - If $p < 0.5$, observation is classified to category "Class2"
- Step 5 : Develop accuracy measures based on actual output class and predicted output class

Example : Acceptance of Personal Loan

- Response: Bank customer accepting a loan (1) or not (0)

- Predictors (X)
  - Age (years), Experience (years), Income($000s)
  - Family Size
  - Education (undergrad, graduate, advanced)
  - Ccavg (Spending on Credit cards)
  - Mortgage (value of house mortgage in $000s)
  - Securities account (1 if the customer has securities account with the bank)
  - CD account ((1 if the customer has a certificate of deposit account with the bank)
  - Online banking (1 if the customer uses Internet banking facilities)
  - Credit card (1 if the customer uses credit card issued by the bank)

- 5000 customers, 480 accepted (9.8%)

Logistic Regression on training data

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

logistic regression is run on training data

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Higher income, family
- Higher ccavg
- Graduate
- Advanced degree
- Holding a cd account

Associated with a higher probability of accepting a loan offer

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Holding securities account
- Holding a credit card

Associated with a lower probability of accepting a loan offer

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- A $1000 increase in income, holding others constant increases the odds that the customer accepts the loan offer by a factor of 1.061($2.718^{0.05939}$)

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Customer who has cd account will accept the offer with an odds of 46.2 ($2.718^{3.833}$) relative to a customer who does not have a cd account holding all other variables

Accuracy Measures Classification

Confusion/Classification Matrix

		Actual/Reference	
		C_1	C_2
Prediction	C_1	Correct Classification (n_{11})	Incorrect Classification (n_{12})
	C_2	Incorrect Classification (n_{21})	Correct Classification (n_{22})

- Total observations in **validation** data, $n = n_{11} + n_{12} + n_{21} + n_{22}$
- Estimated misclassification rate, $err = \frac{n_{12}+n_{21}}{n}$
- Accuracy $= 1 - err = \frac{n_{11}+n_{22}}{n}$

Confusion Matrix for validation data

		Actual/Reference	
		Nonowner	Owner
Prediction	Nonowner	**4 (n_{11})**	**1 (n_{12})**
	Owner	**2 (n_{21})**	**3 (n_{22})**

- Total observation in **validation** data $n = n_{11} + n_{12} + n_{21} + n_{22} = 10$
- Estimated misclassification rate, $err = \frac{n_{12}+n_{21}}{n} = \frac{3}{10} = 30\%$
- Accuracy $= 1 - err = \frac{n_{11}+n_{22}}{n} = \frac{7}{10} = 70\%$

Unequal importance of classes

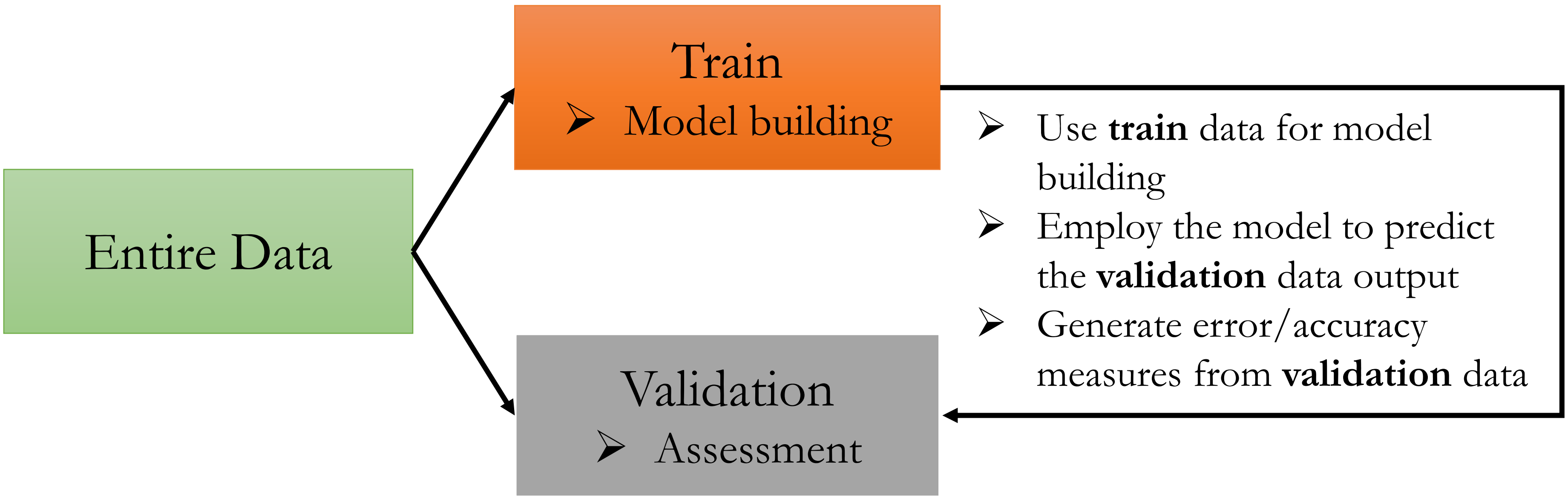
- Sometimes it is **more important** to predict a membership correctly in class C_1 than in class C_2
- Example : Predicting financial status (bankrupt/solvent) of firms
- Predicting **bankrupt** status is more important than **solvent**
- Overall Accuracy is not a good measure under unequal importance of classes
- Measures : **Sensitivity** and **Specificity**

Confusion Matrix

		Actual/Reference	
		C_1	C_2
Prediction	C_1	Correct Classification (n_{11})	Incorrect Classification (n_{12})
	C_2	Incorrect Classification (n_{21})	Correct Classification (n_{22})

- Lets say the important class is C_1
- **Sensitivity** : Ability to **detect** the important class members correctly
  - ➢ $\dfrac{n_{11}}{n_{11} + n_{21}}$
- **Specificity** : Ability to **rule out** non-important class members correctly
  - ➢ $\dfrac{n_{22}}{n_{22} + n_{12}}$

Data Partition : Training & Validation



- Assuming 80-20 partition, how many exhaustive partitions are possible for a dataset with 100 rows?
- $\binom{100}{80} = \frac{100!}{80! * 20!} = 5.36 * 10^{20}$
- We are analyzing only one partition of $5.36 * 10^{20}$
- What about other partitions?

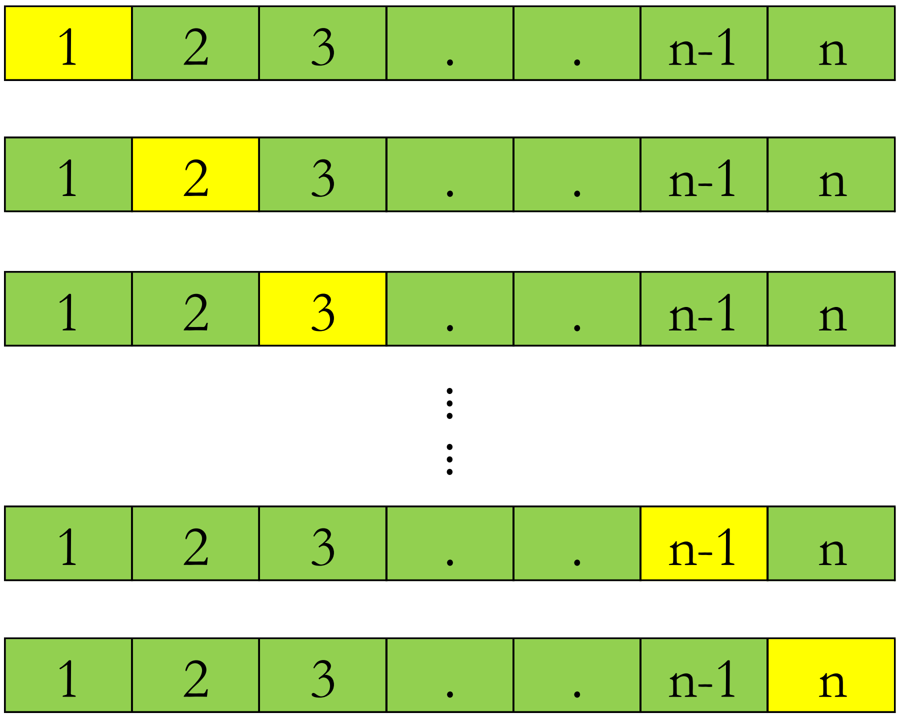
Drawbacks

- What are the drawbacks of analyzing one randomly partition ?
  - Model fit is analyzed on **one** training data partition
  - Error/Accuracy measures are evaluated on **one** validation data partition
  - Likelihood of an excellent model fit and performance on this **one** partition is possible
- Analyzing on a different partition can lead to an unfavorable end result
- How to overcome this drawback?

Resampling

- Indisspensable tool in Statistics/Machine Learning
- Idea
  - Repeatedly draw sample from the data
  - Fit model of interest on each sample
- Example
  - Fit Linear Regression on each repeated sample
  - Examine the extent to which results/accuracy measures differ across multiple validation datasets
- Computationally expensive
- Methods : **Cross-Validation** and **Bootstrap**

Leave-One-Out Cross-Validation (LOOCV)



Accuracy measure for observation 1

Accuracy measure for observation 2

Accuracy measure for observation 3

⋮

⋮

Accuracy measure for observation n-1

Accuracy measure for observation n

Report the Mean/Standard deviation of the accuracy measures

K-fold Cross-Validation

Fold 1	Fold 2	Fold 3	.	.	Fold K-1	Fold K
Validation	Training	Training	.	.	Training	Training
Training	Validation	Training	.	.	Training	Training
Training	Training	Validation	.	.	Training	Training
			⋮			
Training	Training	Training	.	.	Validation	Training
Training	Training	Training	.	.	Training	Validation

Report the Mean/Median of the accuracy measures obtained for **K** iterations

Generally K is chosen 5 or 10

Comparison

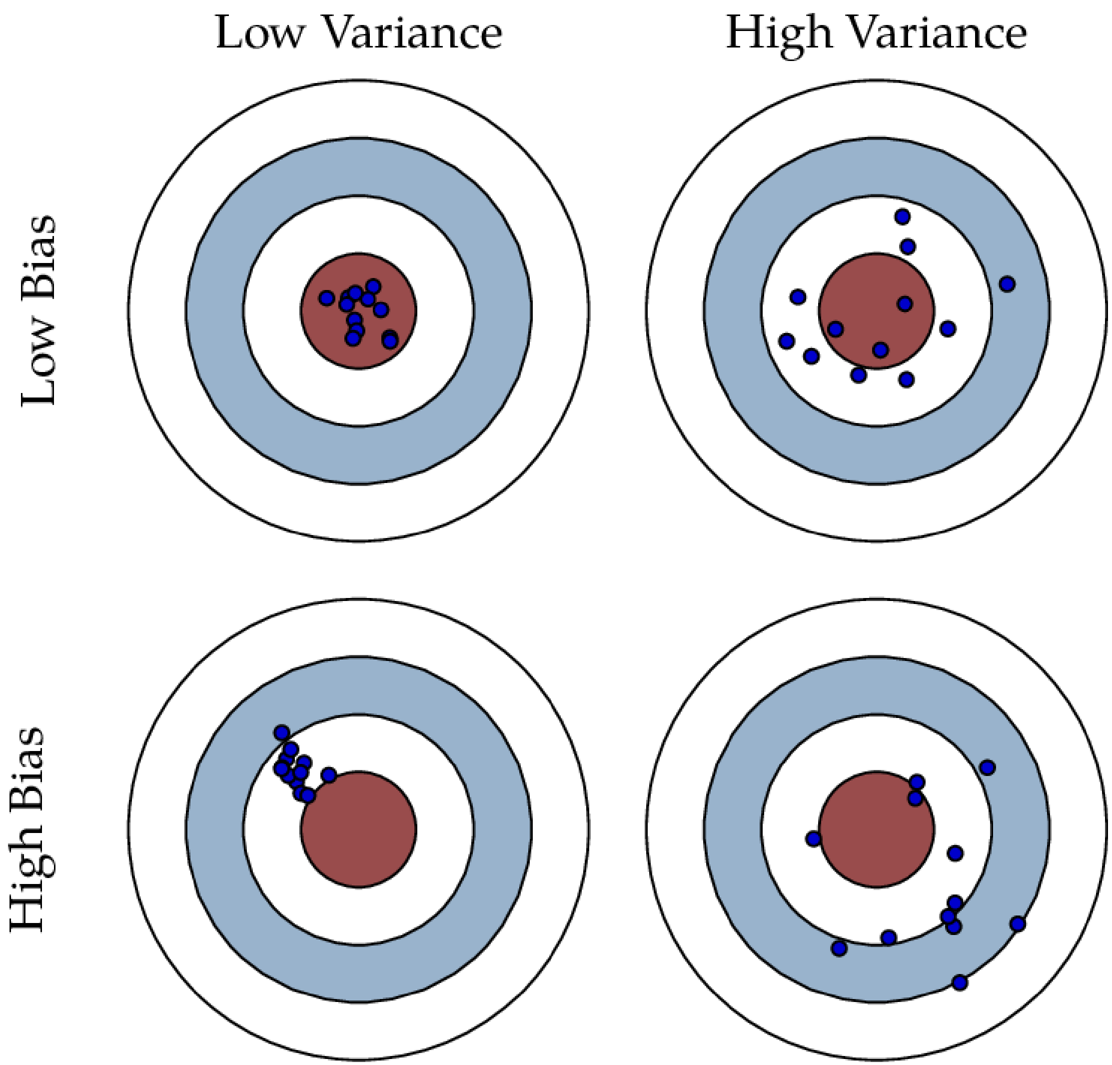
LOOCV

- No randomness in the process
- Time-consuming when "n" is large
- Special case of K-fold CV when K = n
- Less bias compared to **true** validation error measures
- Higher variance

K-fold CV

- Incorporates randomness
- Less time consuming as the process requires to run only K times
- More bias compared to **true** validation error measures
- Less variance

Bias-Variance Trade-off



LOOCV

K-fold CV

Thank You