Logistic Regression

Previous class

- Resampling process
  - LOOCV – Leave One Out Cross-Validation
  - K-Fold Cross-Validation
- Logistic Regression

Logistic Regression

- Non-Linear model
- Like Linear Regression, the method fits a relationship between a categorical variable Y and set of "q" predictors $X_1, X_2, X_3, \cdots\cdots X_q$
- The outcome variable Y is categorical
- Predictors $X_1, X_2, X_3, \cdots\cdots X_q$ can be categorical or numerical
- Prediction is a probability that the new record belongs to a category
- What is the difference compared with *k*-NN applied as Classification?
    - *k*-NN prediction is 100% belonging to a class
    - Logistic Regression prediction is probability belonging to a class

Transformation

- Logistic response function

Value of **e** is 2.718

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

- For any values of $X_1, X_2, X_3, \cdots \cdots X_q$, the right-hand side is always between 0 and 1

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

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

- Odds word is much popular in horse races, sports, gambling…

- Instead of using probability of winning, people quote odds of winning

- If $p = 0.5$, then $\text{Odds} = 1$

Estimation

- Log Odds

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

- Information on both $X's$ & Y is available
- $\beta_0, \beta_1, \beta_2 \cdots\cdots \beta_q$ are coefficients
- Required to estimate the coefficients
- Underlying estimation process: **Maximum Likelihood Estimation (MLE)**
  - Find estimates that maximize the chance of obtaining the data we have

Today's class mandatory steps

- Create a folder name "**k. logistics_regression**" within the folder "**oba_455_555_ddpm_r/rproject**"
- Download "**logistics_regression _code.R**", and all **csv** files from canvas
- Place all downloaded files in "**oba_455_555_ddpm_r/rproject/ k. logistics_regression**"
- Open RStudio project
- Open "**logistics_regression_code.R**" file within RStudio

Example : Predicting Flight Delays

- Response: Predicting whether a flight is delayed or not
- Predictors (X)
  - Day of the week
  - Departure time
  - Origin
  - Destination
  - Carrier
  - Weather
- 2,201 flights
- Flights from Washington DC area into New York City area during Jan 2004

Data Description

- Day_Week: coded as 1 = Mon, 2 = Tue, 3 = Wed……
- Departure time: Hour and Minutes
- Origin : BWI (Baltimore-Washington Intl), DCA (Reagan National), IAD (Dulles)
- Destination: JFK, LGA, EWR
- Carrier: CO (Continental), DH (Atlantic Coast), DL (Delta), MQ (American Eagle), OH (Comair), RU (Continental Express), UA (United), US (US Airways)
- Weather: Coded as 1 if there is a weather-related delay

Delays data partition

- Let's us consider 70-30 partition
- **Train**: Randomly filter 70% of the entire data
- **Validation**: Extract the remaining 30% of the entire data

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	

Results : Day_Week

	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	

- Flights that operate on **Thu** have delays with an odds of **0.4979** ($= 2.718^{-0.7345}$) relative to Flights that operate on **Mon**
- Flights that operate on **Sat** have delays with an odds of **0.2239**($= 2.718^{-1.4964}$) relative to Flights that operate on **Mon**

Results : Dep_Hour

	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	

- Flights that depart during **7-8pm** have delays with an odds of **2.245**($= 2.718^{0.8091}$) relative to Flights that depart during **6-7am**
- Flights that depart during **9-10pm** have odds of **2.138**($= 2.718^{0.7600}$) of delays relative to Flights that depart during **6-7am**

Results : Destination

	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	

- Flights that arrive to **JFK** have delays with an odds of **0.6**($= 2.718^{-0.5107}$) relative to Flights that arrive to **LGA**

Confusion Matrix and Accuracy

```
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
```

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

Next Class

- Mid-term Review

Thank You