Food Inspections: Modeling Risk with Ordinal Logistic Regression

MSCA 31010 Linear & Nonlinear Models

Group 2: Christopher Hein, Yuming Liao, Andrew McCurdy, Nina Randorf, Melanie Tran

March 2021

Disclaimer

- This project uses data pulled from the Chicago Data Portal. The City of Chicago voluntarily provides data to its Chicago Data Portal and requires users of the data to include the following disclaimer:

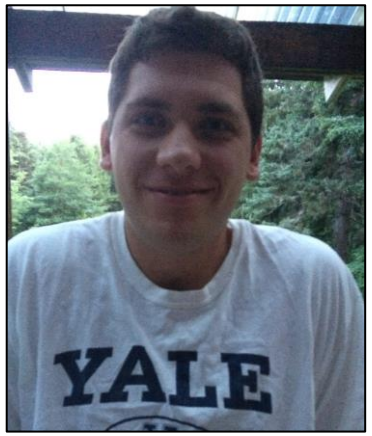
  - "This site provides applications using data that has been modified for use from its original source, www.cityofchicago.org, the official website of the City of Chicago. The City of Chicago makes no claims as to the content, accuracy, timeliness, or completeness of any of the data provided at this site. The data provided at this site is subject to change at any time. It is understood that the data provided at this site is being used at one's own risk."

- Links
  - https://www.chicago.gov/city/en/narr/foia/CityData.html
  - https://www.chicago.gov/city/en/narr/foia/data_disclaimer.html

Our Team

Yuming Liao

- FT Student, 2nd Quarter
- B.S. in Physics from University of California, Irvine



Andrew McCurdy

- FT Student, 2nd Quarter
- B.A. in Psychology from Columbia University in the City of New York

Nina Randorf

- FT Student, 2nd Quarter
- B.S. in Civil Engineering from the University of Michigan (Go Blue!)

Melanie Tran

- PT Student, 2nd Quarter
- B.S. in Supply Chain Management from University of Maryland, College Park

Christopher Hein

- FT Student, 2nd Quarter
- B.S. in Information Systems
- Minor in Computer Science
- DePaul University

Agenda

1. Executive Summary
2. Initial Exploratory Analysis
3. Ordinal Logistic Model
4. Our Model
5. Findings & Insights
6. Lessons Learned
7. Questions

Executive Summary

The Chicago Department of Public Health regularly inspects businesses serving food to ensure restaurants and other food retail outlets are following safe food handling procedures. Inspections determine that safeguards are in place to protect food from contamination by food handlers, cross-contamination, and contamination from other sources in the restaurant.

We created an **Ordinal Logistic Regression Model** to explore our business case research question.

Our Business Case	Application	Data	Our Approach
What factors influence risk assignment for restaurants which pass their inspections?	Stakeholders: Restaurant owners and managers Use Case: Determine which violations to prioritize to minimize the severity of their risk assignment during inspections	• 43,573 Rows, 17 Attributes • July 1, 2018 to February 19, 2021 • Data Types: Categorical and Ordinal	• Initial Exploratory Analysis • Design & Fit the Model • Extract Insights • Document Lessons Learned

I. Initial Exploratory Analysis

Initial exploratory analysis revealed an establishment could pass an inspection but still be categorized as high risk. The dataset recorded Risk as an ordered factor which could be affected by multiple categorical variables.

Categorical Data Types

Ordinal (High > Med > Low)

spection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations
2184275	MCCB	MCCB	2600357.0	Restaurant	Risk 1 (High)	2138 S ARCHER AVE	CHICAGO	IL	60616.0	7/3/2018	License	Pass w/ Conditions	NaN
2184211	JAMES KITCHEN + BAR	JAMES HOTEL CHICAGO	1884681.0	Restaurant	Risk 1 (High)	616 N RUSH ST	CHICAGO	IL	60611.0	7/3/2018	Canvass	Out of Business	NaN
2182184	CAFE EL TAPATIO	CAFE EL TAPATIO	2432567.0	Restaurant	Risk 1 (High)	3400-3402 N ASHLAND AVE	CHICAGO	IL	60657.0	7/3/2018	Canvass Re-Inspection	Pass	NaN
2170218	ROYS LUNCH BAG	ROYS LUNCH BAG	46653.0	Restaurant	Risk 1 (High)	403 E 71ST ST	CHICAGO	IL	60619.0	7/3/2018	Canvass Re-Inspection	Pass	NaN
2182196	MCDONALD'S	MCDONALD'S	1840645.0	Restaurant	Risk 2 (Medium)	6740 N CLARK ST	CHICAGO	IL	60626.0	7/3/2018	Canvass	Pass w/ Conditions	5. PROCEDURES FOR RESPONDING TO VOMITING AND D...

I. Initial Exploratory Analysis

We transformed the dataset so each violation would be listed with an inspection was recorded as its own categorical variable ("0 " indicated a pass and "1" indicating a hit or violation in that inspection area). This new format would then be compatible with building an ordinal logistic regression model.

	Risk	Facility Type	Inspection Date	Inspection Type	Violations
0	Risk 2 (Medium)	Restaurant	7/3/2018	Canvass	58. ALLERGEN TRAINING AS REQUIRED - Comments: ...
1	Risk 2 (Medium)	Restaurant	7/6/2018	License	57. ALL FOOD EMPLOYEES HAVE FOOD HANDLER TRAIN...
2	Risk 1 (High)	Restaurant	7/9/2018	Complaint	58. ALLERGEN TRAINING AS REQUIRED - Comments: ...
3	Risk 1 (High)	Restaurant	7/9/2018	Canvass	49. NON-FOOD/FOOD CONTACT SURFACES CLEAN - Com...
4	Risk 1 (High)	Restaurant	7/10/2018	Canvass	5. PROCEDURES FOR RESPONDING TO VOMITING AND D...

	Risk	Facility Type	Inspection Date	Inspection Type	Supervision	EmployeeHealth	HygenicPractices	HandContamination	ApprovedSource	Contamination
0	Med	Restaurant	Summer	Canvass	0.0	0.0	0.0	0.0	0.0	0.0
1	High	Restaurant	Summer	Canvass	0.0	0.0	0.0	0.0	0.0	0.0
2	High	Restaurant	Summer	Canvass	0.0	1.0	0.0	0.0	0.0	0.0
3	High	Restaurant	Summer	Canvass	0.0	0.0	0.0	0.0	0.0	0.0
4	High	Restaurant	Summer	Canvass	0.0	0.0	0.0	1.0	0.0	0.0

Ordinal Logistic Regression (OLR)

Ordinal Logistic Regression is used to predict an ordinal dependent variable given one or more independent variables.

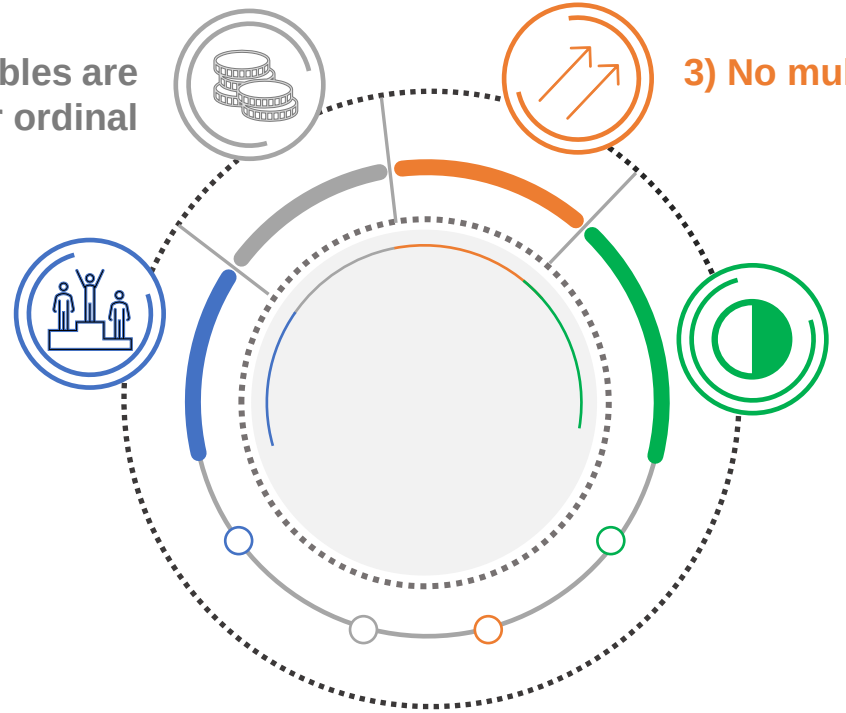
The Ordinal Logistic Model must meet four assumptions:

2) One or more predictor variables are continuous, categorical, or ordinal

3) No multicollinearity

1) Dependent variable are ordinal

4) Proportional odds



Ordinal Logistic Regression (OLR)

- **Multinomial** (MLR) **vs. Ordinal Logistic Regression Models**
  - Key difference: the response variable is nominal vs ordinal
  - mlogit(y ~., data=learn)
  - In a MLR model the coefficients represent intercepts for one response variable against each other.
  - In an OLR model the coefficients represent "cut points", which represent where one variable switches to the next in a hierarchical order.

Multinomial

$$P(Y \leq j|x_i) = \sum_{k \leq j} P(Y = j|x_i)$$

$$\log\left(\frac{P(Y_i \leq j|\mathbf{X}_i)}{1 - P(Y_i \leq j|\mathbf{X}_i)}\right) = \theta_j - \mathbf{X}_i\beta$$

Ordinal

$P(Y \leq j)$ is the cumulative probability of Y less than or equal to a specific category *j*.

$$\log\frac{P(Y \leq j)}{P(Y > j)} = logit(P(Y \leq j)$$

$$logit(P(Y \leq j) = \beta_{0j} + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

```
> summary(m1)
Call: nnet::multinom(formula = V1 ~ GarbageInfo)

Coefficients:
       (Intercept) GarbageInfo
red     -0.6011338  -0.1331142
yellow  -0.3221203  -0.5995860

Std. Errors:
       (Intercept) GarbageInfo
red      0.5164521   0.8448432
yellow   0.4932972   0.8380932
...


> summary(m2)
...
Call: MASS::polr(formula = V2 ~ GarbageInfo)

Coefficients:
             Value Std. Error t value
GarbageInfo -0.2181     0.6431 -0.3391

Intercepts:
                Value   Std. Error t value
green|yellow -0.1541  0.3910    -0.3941
yellow|red    0.9856  0.4045     2.4364
...
```

Image from StackExchange

P = probability of a response
j = response level
β = coefficient of explanatory variable
x = explanatory (independent) variable
p = number of predictor

Our Model

Our ordered response variable can be modeled using the cumulative logistic model:

$$logit(P(Y \leq j) = \beta_{0j} + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Where the response Y (Risk) has j-ordered levels (low, med, high) described by a combination of explanatory variables (inspection season and types of health inspection focus areas).

The model will be described by two equations:

$$\log \frac{P(Y \leq L)}{P(Y > M)} = logit(P(Y \leq L) = \beta_{0L|M} + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$$\log \frac{P(Y \leq M)}{P(Y > H)} = logit(P(Y \leq M) = \beta_{0M|H} + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

And the coefficients of the model can be interpreted as an odds ratio holding all other variables constant, $e^{x_P \beta_j}$.

```
Call:
polr(formula = Risk ~ ., data = dta_mini, Hess = T)

Coefficients:
                          Value Std. Error t value
Inspection.DateWinter   0.38477    0.15640  2.4602
Inspection.DateSpring  -0.40910    0.15307 -2.6727
Inspection.DateSummer  -0.61075    0.15436 -3.9567
HandContamination.L     0.27371    0.12982  2.1084
Contamination.L         0.20936    0.17876  1.1712
FoodTemperature.L      -0.42486    0.14030 -3.0283
FoodIdentification.L    0.28734    0.13189  2.1787
FoodContamination.L     0.10263    0.09566  1.0728
ProperUtensils.L        0.52630    0.16659  3.1593
Equipment.L             0.10667    0.07631  1.3978
PhysicalFacilities.L    0.06233    0.09528  0.6542
EmployeeTraining.L      0.74643    0.14920  5.0030

Intercepts:
         Value    Std. Error t value
Low|Med   -7.3898   0.5110   -14.4608
Med|High  -2.6670   0.2511   -10.6206

Residual Deviance: 2312.469
AIC: 2340.469
```

II. Model Building – Assumptions

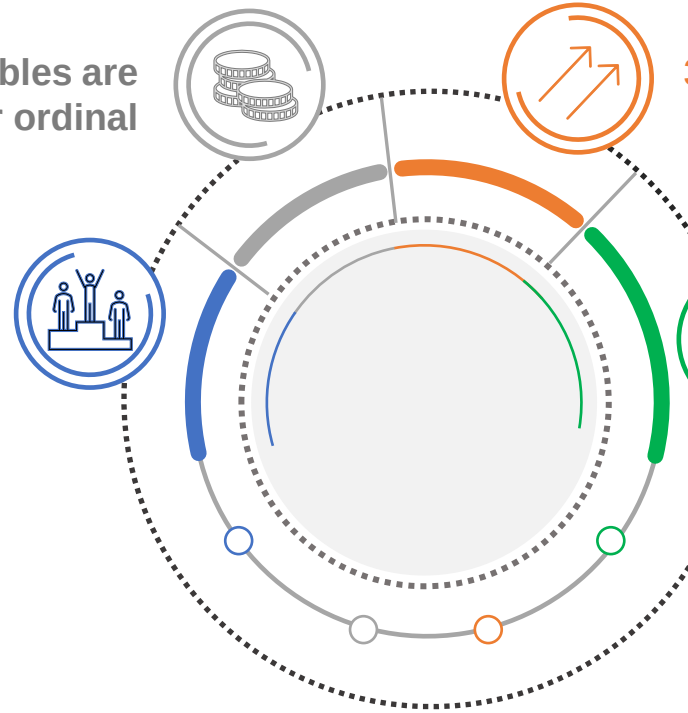
The Ordinal Logistic Model must meet four assumptions:

2) One or more predictor variables are continuous, categorical, or ordinal

3) No multicollinearity

1) Dependent variable are ordinal

4) Proportional odds



II. Model Building – Assumptions 1 & 2

Assumption	Description	Pass
1) The dependent variable are ordered	Risk is ordered Low < Med < High	✓
2) One or more predictor variables are continuous, categorical, or ordinal	All predictor variables are categorical (Date: Seasons, Violations: Pass and Hit)	✓

	Risk	Inspection.Date	HandContamination	Contamination	FoodTemperature	FoodIdentification	FoodContamination
1	Med	Summer	Pass	Pass	Pass	Pass	Pass
2	High	Summer	Pass	Pass	Pass	Pass	Pass
3	High	Summer	Pass	Pass	Pass	Pass	Pass
4	High	Summer	Pass	Pass	Pass	Pass	Pass
5	High	Summer	Hit	Pass	Pass	Pass	Pass
6	High	Summer	Pass	Pass	Pass	Pass	Pass
7	Med	Summer	Pass	Pass	Pass	Pass	Pass
8	High	Summer	Pass	Pass	Pass	Pass	Pass

II. Model Building – Assumption 3

Assumption	Description	Pass
3) No multicollinearity	Using the VIFfit() test, we can conclude that no variables have a VIF score > 5, therefore there is no multicollinearity.	✓

Variance Inflation Factor

The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

- 1 = Not correlated
- 1 - 5 = Moderately correlated
- > 5 = Highly correlated

```
numRisk <- as.numeric(dta$Risk)
VIFfit <- lm(scale(numRisk) ~ ., data = dta_mini)
VIF(VIFfit)
```

	GVIF	Df	GVIF^(1/(2*Df))
Risk	1.048278	2	1.011857
Inspection.Date	1.050227	3	1.008201
HandContamination	1.012984	1	1.006471
Contamination	1.005303	1	1.002648
FoodTemperature	1.013250	1	1.006603
FoodIdentification	1.023206	1	1.011536
FoodContamination	1.019121	1	1.009515
ProperUtensils	1.029838	1	1.014809
Equipment	1.008589	1	1.004285
PhysicalFacilities	1.020507	1	1.010202
EmployeeTraining	1.024369	1	1.012111

II. Model Building – Assumption 4

Assumption	Description	Pass
4) Proportional odds	Brant Test & PoTest indicates that the probabilities are > than our alpha, 0.05	✓

Brant Test

Test for	X2	df	probability
Omnibus	1.49	12	1
Inspection.DateWinter	0	1	1
Inspection.DateSpring	0	1	1
Inspection.DateSummer	0	1	0.97
HandContamination.L	0	1	1
Contamination.L	0	1	1
FoodTemperature.L	1.04	1	0.31
FoodIdentification.L	0	1	1
FoodContamination.L	0.11	1	0.74
ProperUtensils.L	0	1	1
Equipment.L	0.64	1	0.42
PhysicalFacilities.L	0	1	0.96
EmployeeTraining.L	0	1	1

Proportional Odds Test

Tests for Proportional Odds
polr(formula = Risk ~ ., data = dta_mini, Hess = T)

	b[polr]	b[>Low]	b[>Med]	Chisquare	df	Pr(>Chisq)
Overall				1.49	12	1.00
Inspection.DateWinter	3.85e-01	1.86e+01	3.81e-01	0.00	1	1.00
Inspection.DateSpring	-4.09e-01	1.87e+01	-4.15e-01	0.00	1	1.00
Inspection.DateSummer	-6.11e-01	-6.38e-01	-6.07e-01	0.00	1	0.97
HandContamination.L	2.74e-01	1.27e+01	2.71e-01	0.00	1	1.00
Contamination.L	2.09e-01	1.30e+01	2.06e-01	0.00	1	1.00
FoodTemperature.L	-4.25e-01	-1.26e+00	-4.17e-01	1.04	1	0.31
FoodIdentification.L	2.87e-01	1.26e+01	2.84e-01	0.00	1	1.00
FoodContamination.L	1.03e-01	-1.67e-01	1.02e-01	0.11	1	0.74
ProperUtensils.L	5.26e-01	1.27e+01	5.24e-01	0.00	1	1.00
Equipment.L	1.07e-01	-4.26e-01	1.07e-01	0.64	1	0.42
PhysicalFacilities.L	6.23e-02	2.02e-02	6.02e-02	0.00	1	0.96
EmployeeTraining.L	7.46e-01	1.27e+01	7.45e-01	0.00	1	1.00

III. Model Testing (Train-Test Split)

Now that we know our model passes the Ordinal Logistic Regression model assumptions, we tested our model.

- Training and testing **60:40**

- While an **82% accuracy** is relatively high, we can see the model is only predicting "High" because of the underlying imbalance in the dataset.

Training set: **83.4% accuracy**

```
#view frequency table
table(datatrain$Risk, pred1)
```

```
##       pred1
##         Low  Med High
##   Low     0    0    3
##   Med     0    0  257
##   High    0    0 1308
```

```
#model test accuracy
sum(diag(table(datatrain$Risk, pred1)  ))/sum(table(datatrain$Risk, pred1))
```

```
## [1] 0.8341837
```

Testing set: **82% accuracy**

```
#view frequency table
table(datatest$Risk, pred_test)
```

```
##       pred_test
##         Low Med High
##   Low     0   0    2
##   Med     0   0  181
##   High    0   0  863
```

```
#model test accuracy
sum(diag(table(datatest$Risk, pred_test)))/sum(table(datatest$Risk, pred_test))
```

```
## [1] 0.8250478
```

Full dta_mini dataset: **83% accuracy**

```
#view frequency table
table(dta_mini$Risk, pred_dta_mini)
```

```
##       pred_dta_mini
##         Low  Med High
##   Low     0    0    5
##   Med     0    0  438
##   High    0    0 2171
```

```
#model test accuracy
sum(diag(table(dta_mini$Risk, pred_dta_mini)))/sum(table(dta_mini$Risk, pred_dta_mini))
```

```
## [1] 0.8305279
```

III. Model Testing (Oversampling)

We can examine **random oversampling** where duplicates are randomly selected from the minority class (any Risk factor other than "High").

The overall model accuracy **decreases to 49%** so our OLR model works best by just predicting "High" Risk factors for all inspections because so "Med" and "Low" assignments are less frequent.

Given the drop in accuracy with resampling, we continued using our "m_mini" model in upcoming analysis. The "m_mini" OLR model had an **82%** accuracy.

```
#creating predictions based on test set
random.oversample.pred <- predict(m_resample, datatest, type = "class")

#recall the original confusion matrix using the dta_mini_train model and test set
table(datatest$Risk, pred_test)
```

```
##       pred_test
##        Low Med High
##   Low    0   0    2
##   Med    0   0  181
##   High   0   0  863
```

```
#confusion matrix after random oversampling
oversample.table = table(datatest$Risk, random.oversample.pred)
oversample.table
```

```
##       random.oversample.pred
##        Low Med High
##   Low    1   1    0
##   Med   58  56   67
##   High 181 228  454
```

```
#viewing accuracy and confusion matrix summary after random oversampling
sum(diag(oversample.table))/sum(oversample.table)
```

```
## [1] 0.4885277
```

IV. Model Interpretation & Results

The log odd coefficients are not easily interpretable from the model summary. Converting the coefficients into an odds ratio lets us describe the affect of a variable unit change on the response.

- For restaurants with **Inspection Dates in Winter**, the odds of getting a higher risk inspection rating (ex. Med or High vs Low) **is 1.47 times** that of restaurants who get Inspections in other seasons, holding constant all other variables.
- For restaurants with **Inspection Dates in Spring**, the odds of getting a higher risk inspection rating (ex. Med or High vs Low) **is 0.66 times** that of restaurants who get Inspections in other seasons, holding constant all other variables.

Model Summary

```
## polr(formula = Risk ~ ., data = dta_mini, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Inspection.DateWinter  0.38477    0.15640  2.4602
## Inspection.DateSpring -0.40910    0.15307 -2.6727
## Inspection.DateSummer -0.61075    0.15436 -3.9567
## HandContamination.L    0.27371    0.12982  2.1084
## Contamination.L        0.20936    0.17876  1.1712
## FoodTemperature.L     -0.42486    0.14030 -3.0283
## FoodIdentification.L   0.28734    0.13189  2.1787
## FoodContamination.L    0.10263    0.09566  1.0728
## ProperUtensils.L       0.52630    0.16659  3.1593
## Equipment.L            0.10667    0.07631  1.3978
## PhysicalFacilities.L   0.06233    0.09528  0.6542
## EmployeeTraining.L     0.74643    0.14920  5.0030
##
## Intercepts:
##          Value    Std. Error t value
## Low|Med   -7.3898   0.5110   -14.4608
## Med|High  -2.6670   0.2511   -10.6206
##
## Residual Deviance: 2312.469
## AIC: 2340.469
```

Odds Ratio and Confidence Intervals

```
##                           OR  2.5 % 97.5 %
## Inspection.DateWinter 1.4693 1.0806 1.9967
## Inspection.DateSpring 0.6642 0.4911 0.8954
## Inspection.DateSummer 0.5429 0.4002 0.7334
## HandContamination.L   1.3148 1.0276 1.7116
## Contamination.L       1.2329 0.8821 1.7843
## FoodTemperature.L     0.6538 0.4994 0.8667
## FoodIdentification.L  1.3328 1.0377 1.7425
## FoodContamination.L   1.1080 0.9214 1.3411
## ProperUtensils.L      1.6927 1.2400 2.3898
## Equipment.L           1.1126 0.9583 1.2927
## PhysicalFacilities.L  1.0643 0.8807 1.2801
## EmployeeTraining.L    2.1094 1.5957 2.8705
```

IV. Model Interpretation & Results

Key Takeaways:

- Variables that Increased the Odds of Higher Risk Assignment
  - Inspections in Winter
  - Violations in Hand Contamination, Food Identification, Proper Utensils, and Employee Training
- Inspections in the Spring or Summer reduced the odds of higher risk assignment

Interpretation:

- Inspection date impacts the odds of risk assignment.
  - Odds of risk assignment might increase in Winter because restaurants are more likely to lose power (impacting the facility's ability to maintain health code standards)
  - Restaurants may also be more prepared for inspections in Spring and Summer as those are traditional busy seasons
- The log odds of food temperature violations is < 0, but this might be tied to less frequent power outages outside of winter months (freezers maintain temperatures more reliably)

Odds Ratio and Confidence Intervals

##	OR	2.5 %	97.5 %
## Inspection.DateWinter	1.4693	1.0806	1.9967
## Inspection.DateSpring	0.6642	0.4911	0.8954
## Inspection.DateSummer	0.5429	0.4002	0.7334
## HandContamination.L	1.3148	1.0276	1.7116
## Contamination.L	1.2329	0.8821	1.7843
## FoodTemperature.L	0.6538	0.4994	0.8667
## FoodIdentification.L	1.3328	1.0377	1.7425
## FoodContamination.L	1.1080	0.9214	1.3411
## ProperUtensils.L	1.6927	1.2400	2.3898
## Equipment.L	1.1126	0.9583	1.2927
## PhysicalFacilities.L	1.0643	0.8807	1.2801
## EmployeeTraining.L	2.1094	1.5957	2.8705

IV. Model Interpretation & Results

Business and stakeholders can't control the date of a random inspection, but could reduce their odds of higher risk assignment by passing all checks in hand contamination, food identification, proper utensils, and employee training:

Preventing Contamination by Hands (HandContamination)

- Hands clean & properly washed
- No bare hand contact with RTE food or pre-approved alternative procedure properly allowed
- Adequate handwashing sinks properly supplied and accessible

Food Identification

- Food properly labeled; original container

Proper Use of Utensils (ProperUtensils)

- In-use utensils: properly stored
- Utensils, equipment & linens: properly stored, dried, & handled
- Single-use/single-service articles: properly stored & used
- Gloves used properly

Employee Training

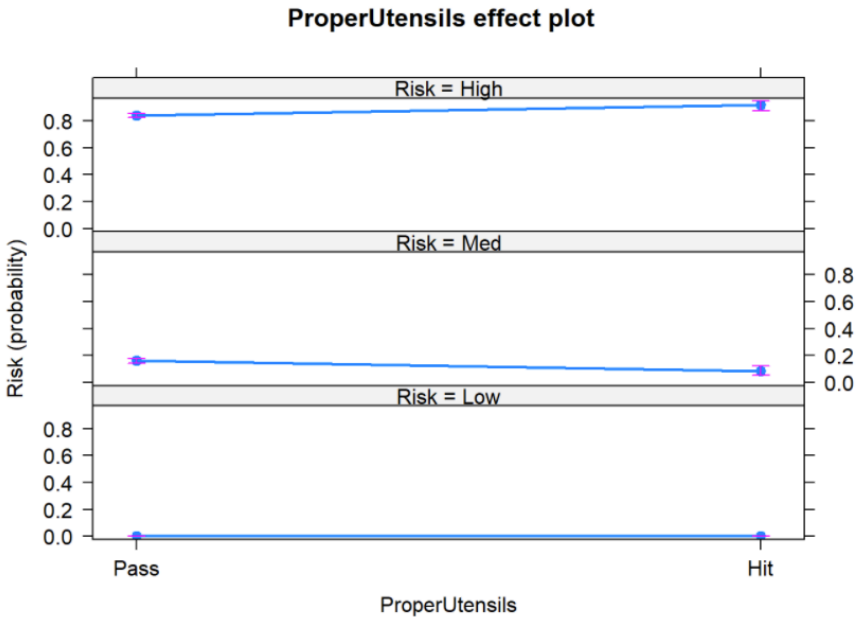
- All food employees have food handler training
- Allergen training as required

Odds Ratio and Confidence Intervals

	OR	2.5 %	97.5 %
Inspection.DateWinter	1.4693	1.0806	1.9967
Inspection.DateSpring	0.6642	0.4911	0.8954
Inspection.DateSummer	0.5429	0.4002	0.7334
HandContamination.L	1.3148	1.0276	1.7116
Contamination.L	1.2329	0.8821	1.7843
FoodTemperature.L	0.6538	0.4994	0.8667
FoodIdentification.L	1.3328	1.0377	1.7425
FoodContamination.L	1.1080	0.9214	1.3411
ProperUtensils.L	1.6927	1.2400	2.3898
Equipment.L	1.1126	0.9583	1.2927
PhysicalFacilities.L	1.0643	0.8807	1.2801
EmployeeTraining.L	2.1094	1.5957	2.8705

IV. Model Interpretation & Results

Another way to visualize the change in risk assignment is to inspect a plot of the probability changes across each explanatory variable. The plots below illustrate how the probability of "high" risk assignment increases (holding all other variable constant), if a restaurant receives a hit on an inspection area.



Odds Ratio and Confidence Intervals

	OR	2.5 %	97.5 %
Inspection.DateWinter	1.4693	1.0806	1.9967
Inspection.DateSpring	0.6642	0.4911	0.8954
Inspection.DateSummer	0.5429	0.4002	0.7334
HandContamination.L	1.3148	1.0276	1.7116
Contamination.L	1.2329	0.8821	1.7843
FoodTemperature.L	0.6538	0.4994	0.8667
FoodIdentification.L	1.3328	1.0377	1.7425
FoodContamination.L	1.1080	0.9214	1.3411
ProperUtensils.L	1.6927	1.2400	2.3898
Equipment.L	1.1126	0.9583	1.2927
PhysicalFacilities.L	1.0643	0.8807	1.2801
EmployeeTraining.L	2.1094	1.5957	2.8705

IV. Comparing Methods

- Fit the models using probit or complementary log-log (cloglog) methods
- The probit and cloglog model coefficients are not in terms of log odds and therefore, more difficult to interpret relative to the default logistic method that our model uses

Anova of Default and Probit Methods

```
#update and refit our model with the probit method
m_mini_probit <- update(m_mini, method = "probit", Hess = TRUE)
```

```
#compare models
anova(m_mini, m_mini_probit)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: Risk
##
Model
## 1 Inspection.Date + HandContamination + Contamination + Food
Temperature + FoodIdentification + FoodContamination + ProperUt
ensils + Equipment + PhysicalFacilities + EmployeeTraining
## 2 Inspection.Date + HandContamination + Contamination + Food
Temperature + FoodIdentification + FoodContamination + ProperUt
ensils + Equipment + PhysicalFacilities + EmployeeTraining
##   Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1      2600   2312.469
## 2      2600   2309.677 1 vs 2     0  2.79238        0
```

Anova of Default and Cloglog Methods

```
#update and refit model using complementary log log
m_mini_clog <- update(m_mini, method = "cloglog", Hess = TRUE)
```

```
#compare models
anova(m_mini, m_mini_clog)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: Risk
##
Model
## 1 Inspection.Date + HandContamination + Contamination + Food
Temperature + FoodIdentification + FoodContamination + ProperUt
ensils + Equipment + PhysicalFacilities + EmployeeTraining
## 2 Inspection.Date + HandContamination + Contamination + Food
Temperature + FoodIdentification + FoodContamination + ProperUt
ensils + Equipment + PhysicalFacilities + EmployeeTraining
##   Resid. df Resid. Dev   Test    Df  LR stat. Pr(Chi)
## 1      2600   2312.469
## 2      2600   2313.683 1 vs 2     0 -1.213399        1
```

Given the residual deviance and AIC values for the probit, cloglog, and logistic (the default OLR model which we used to create our m_mini model) are similar, we recommend using our original m_mini model for this analysis.

Lessons Learned

- OLS requires a large sample size to avoid pitfalls of imbalanced datasets; this problem is compounded during train-test splits.
- If the data is imbalanced, it is more common for OLS variables to fail the Proportional Odds Assumptions.
- Accuracy is not always be the best performance metric to evaluate a model.
- Dropping certain datapoints due to scoping limitations is acceptable and does not adversely affect the model and its outcome.
- The logit model is easier to interpret due to the odds ratio compared to probit and complimentary log-log models if the performance is similar.

Questions

References

- Chicago Data Portal. *Food Inspections – 7/1/2018 - Present*, 2021, data.cityofchicago.org/Health-Human-Services/Food-Inspections-7-1-2018-Present/qizy-d2wf/data. Accessed 19 Feb 2021.
- City of Chicago. *Understand Health Code Requirements for Food Establishments*, 2021, chicago.gov/city/en/depts/cdph/provdrs/healthy_restaurants/svcs/understand_healthcoderequirementsforfoodestablishments. Accessed 19 Feb 2021.
- Lee, Evangeline. *Ordinal Logistic Regression and its Assumptions – Full Analysis.* Medium, 25 May 2019, medium.com/evangelinelee/ordinal-logistic-regression-on-world-happiness-report-221372709095.
- Perceptive Analytics. *How to Perform Ordinal Logistic Regression in R.* R-bloggers, 18 June 2019, r-bloggers.com/2019/06/how-to-perform-ordinal-logistic-regression-in-r/.
- UCLA: Statistical Consulting Group. *Ordinal Logistic Regression | R Data Analysis Examples,* stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/
- StackExchange. *What is the difference between multinomial and ordinal logistic regression?*, stats.stackexchange.com/questions/155737/what-is-the-difference-between-multinomial-and-ordinal-logistic-regression. Accessed 14 Mar 2021.

END