Logistic Regression

DSBA/MBAD 6211

Logistic Regression Overview

- Logistic regression is a predictive technique used when the dependent variable has binary outcomes (e.g., outage or no outage; 1 or 0)
- The logistic regression estimate calculates the probability of the event occurring based on past data
- Logistic regression is considered “supervised learning technique” since the data includes actual outcome values from past observations
- Examples of logistic regression applications
  - Predicting the probability of a system or part failing
  - Predicting the probability of a subscriber cancelling a service (churn)
  - Predicting the probability of a patient readmission
  - Predicting the probability of an outage

Multiple Linear Regression Prediction

prediction estimate

$$\xrightarrow{} \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

intercept estimate

parameter estimate

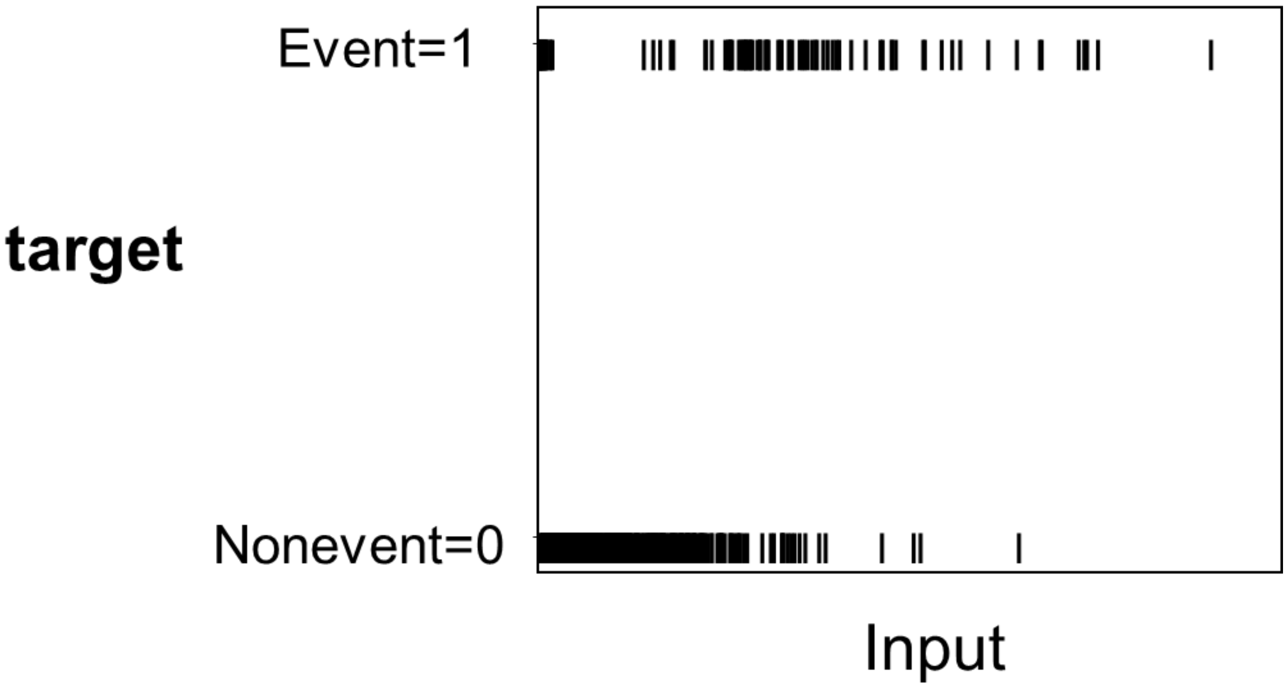
Choose intercept and parameter estimates to *minimize*:

squared error function

$$\sum_{\text{training data}} (y_i - \hat{y}_i)^2$$

Binary Dependent Variable

- When the dependent variable is binary (e.g., values 1 or 0), linear regression does not work directly, since the results are generally unbounded.
- Instead, we use the probability *p* that the event will occur rather than directly using 1 or 0 .



Why Linear Regression Will Not Work?

- If linear regression is used, the predicted values will become greater than one and less than zero if you move far enough on the X-axis. Such values are theoretically inadmissible.
- An assumption of regression is that the variance of Y is constant across values of X (homoscedasticity). This cannot be the case with a binary variable, because the variance is p*(1-p), where p is the probability of the event. When 50 percent of the observations are 1s, then the variance is .25, its maximum value. As we move to more extreme values, the variance decreases. When p=.10, the variance is .1*.9 = .09, so as p approaches 1 or zero, the variance approaches zero.
- Another assumption is that errors of prediction (Y-Y') are normally distributed. Because Y only takes the values 0 and 1, this assumption is pretty hard to justify, even approximately.

Odds and Probability

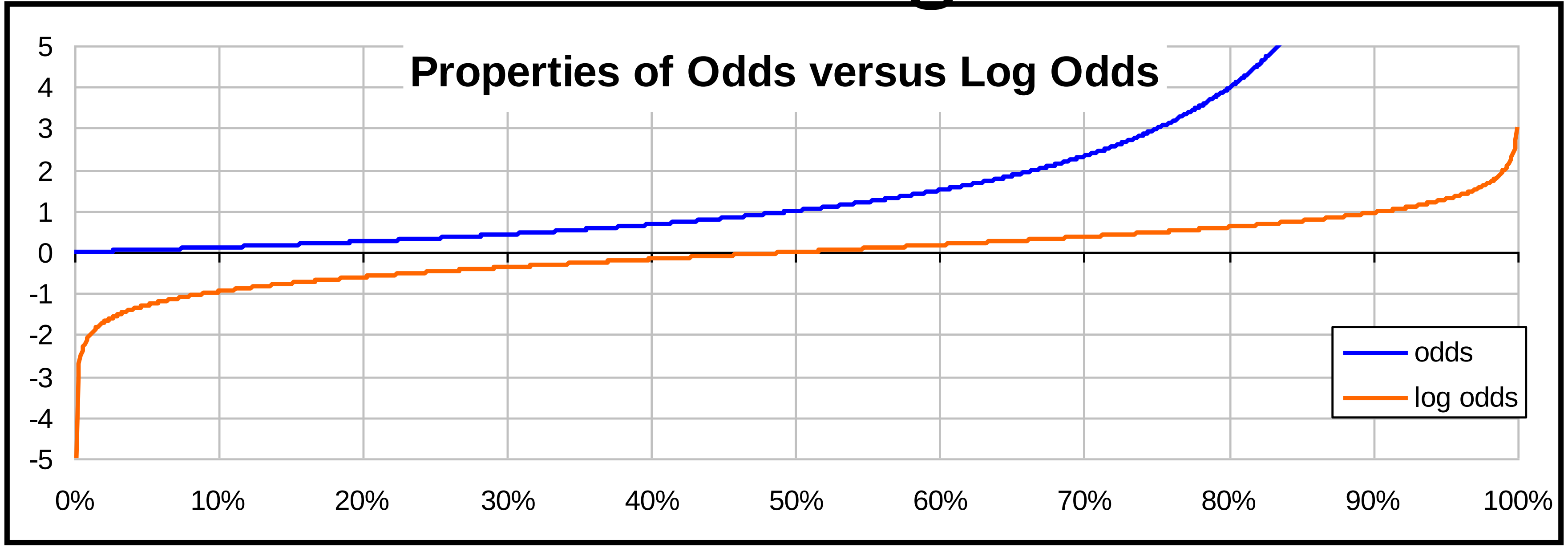
- Consider the probability *p* of an event (such as an outage) occurring.
- The probability of the event not occurring is *1-p*.
- The odds of the event happening are *p*:(1-*p*)

$$odds = \frac{p_{win}}{p_{loss}} = \frac{p}{1-p}$$

- For example, if the probability of outage p=0.2, the odds are $\frac{0.2}{0.8} = 0.25$
- Odds may also be expressed as integers such as 1:4, which means that in every 5 events, 1 event is an outage and 4 events are non-outages
- From the above odds, the probability of an outage is computed as

$$p(outage) = \frac{1}{(1+4)} = 0.2$$

Properties of Odds and Log Odds

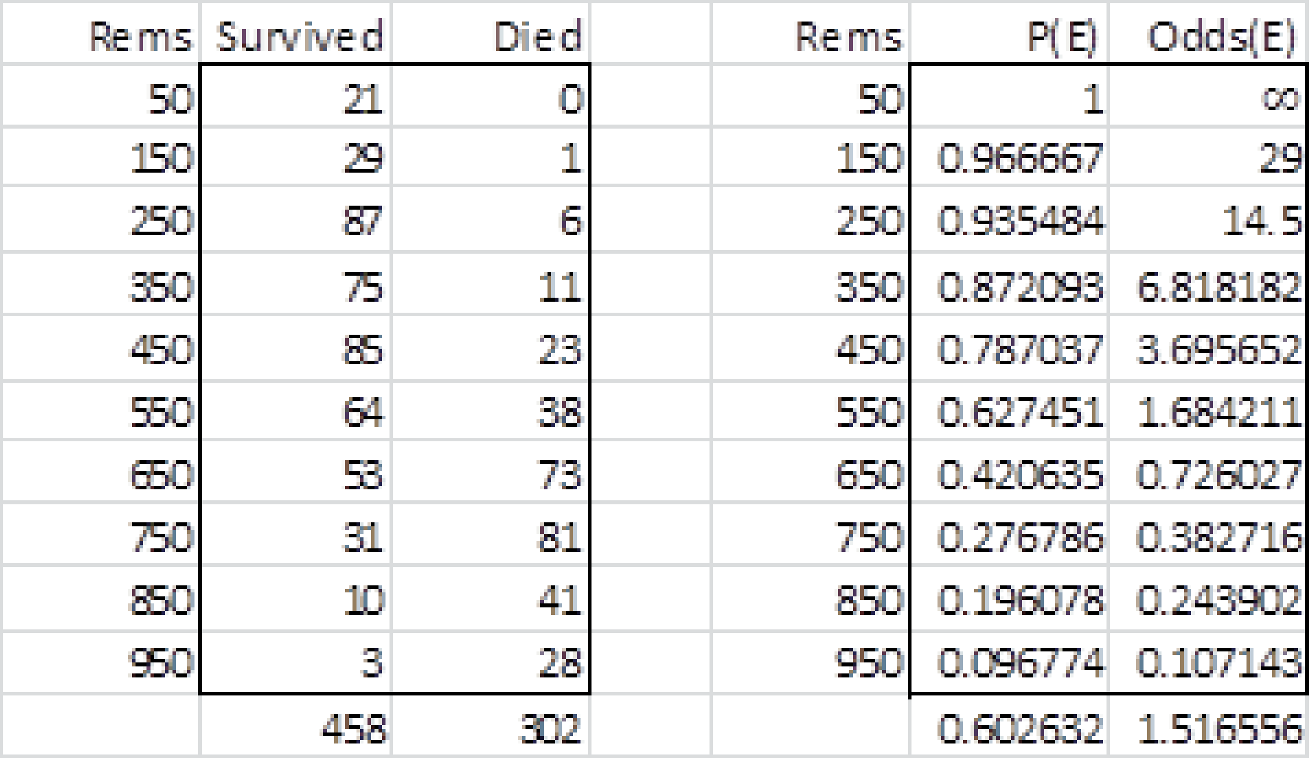


- Odds is not symmetric, varying from 0 to infinity.
- Odds is 1 when the probability is 50%.

- Log Odds is symmetric, going from minus infinity to positive infinity, like a line.
- Log Odds is 0 when the probability is 50%.
- It is highly negative for low probabilities and highly positive for high probabilities.

A Sample Dataset

- Following a nuclear accident, individual’s radiation exposure and survival were captured in a dataset. The objective is to estimate the relationship between the radiation exposure (in Rems) and survival.

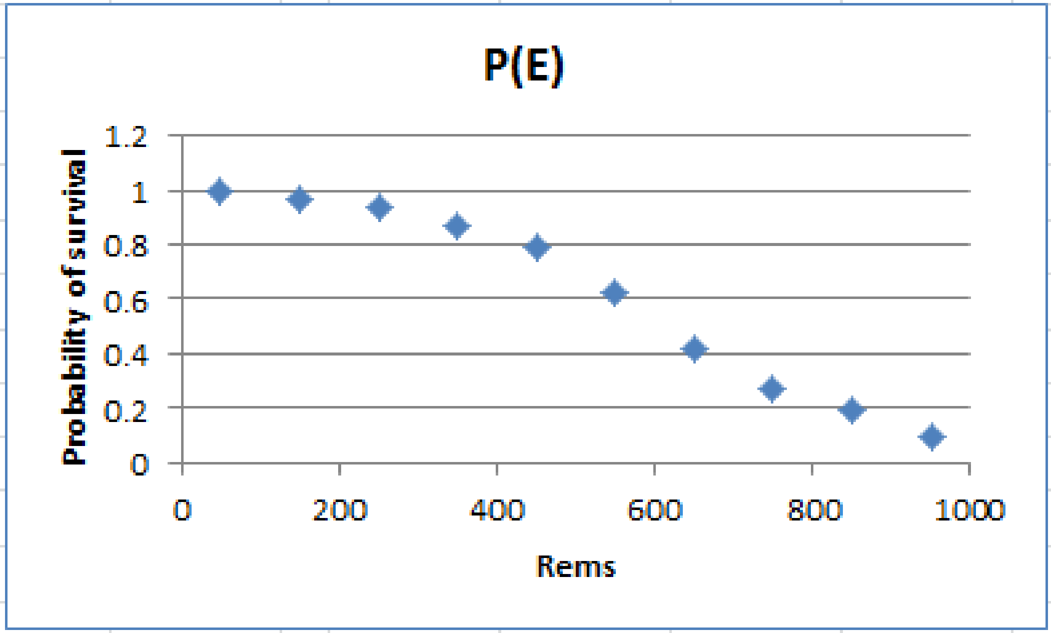


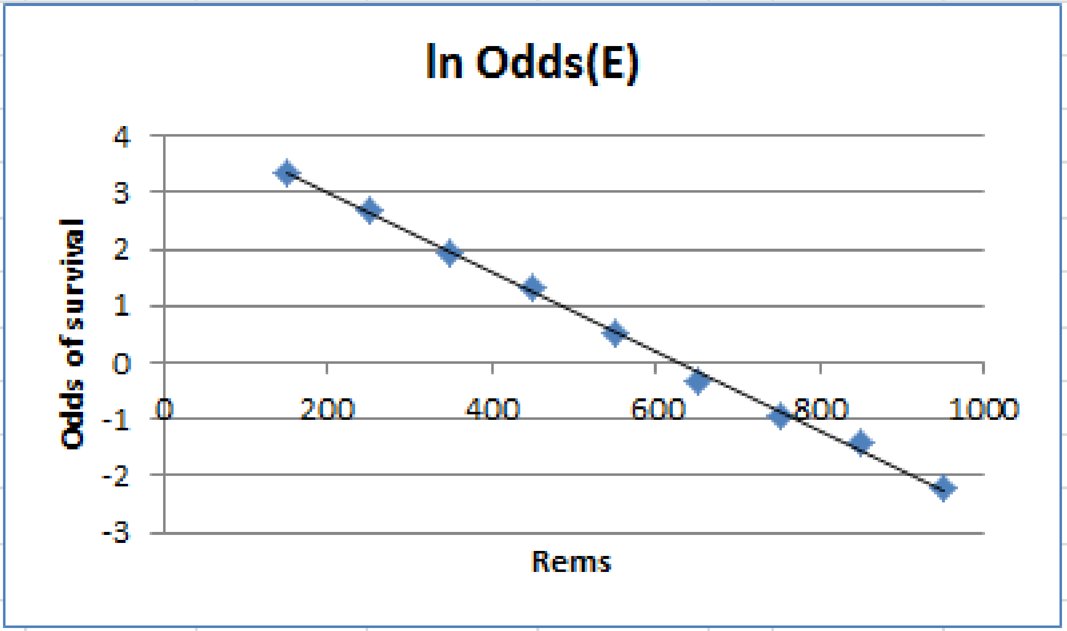
Rems	Survived	Died		Rems	P(E)	Odds(E)
50	21	0		50	1	∞
150	29	1		150	0.966667	29
250	87	6		250	0.935484	14.5
350	75	11		350	0.872093	6.818182
450	85	23		450	0.787037	3.695652
550	64	38		550	0.627451	1.684211
650	53	73		650	0.420635	0.726027
750	31	81		750	0.276786	0.382716
850	10	41		850	0.196078	0.243902
950	3	28		950	0.096774	0.107143
	458	302			0.602632	1.516556

Source: http://www.real-statistics.com/logistic-regression/basic-concepts-logistic-regression/

Comparing p(E) vs ln(odds(E))

E = event, which in this case is the individual survived





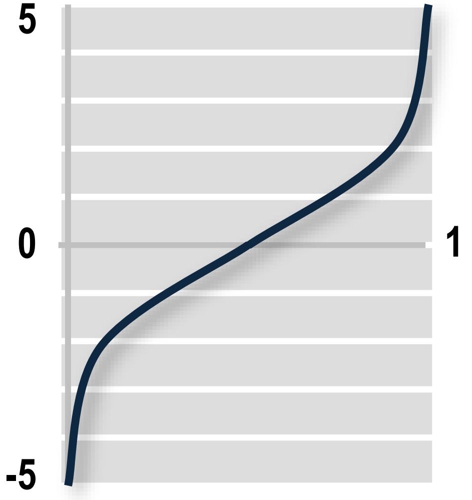
Source: http://www.real-statistics.com/logistic-regression/basic-concepts-logistic-regression/

Logistic Regression Prediction Formula

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

The left-hand side of the above expression is called the logit link function. The logit link function transforms probabilities (between 0 and 1) to logit scores (between −∞ and +∞).

logit
link function



Logistic Regression Prediction Formula

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = \text{logit}(\hat{p})$$

To obtain prediction estimates, the logit equation is solved for $\hat{p}$.

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

Estimation

- Maximum likelihood estimation
  - An iterative procedure that successively works to get closer and closer to the correct fit.
  - A model that fits the data well will have a small absolute value of log-likelihood*. A perfect model would have a log-likelihood value of zero.
    - In the statsmodels Python library, Log-Likelihood (negative) value will be reported in the output. The larger this value(i.e., smaller absolute value), the better is the model. LL-Null is the value when no predictors are used.
- Larger sample size requirements than linear regression
  - Linear: 10 * number of IV
  - Logistic: 20 * number of IV

Estimation

$$P(Y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}}$$

Likelihood function, where Y_i = observed i^{th} outcome; p_i = predicted probability for Y_i=1

$$L(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n}[p_i^{Y_i} \times (1 - p_i)^{1-Y_i}]$$

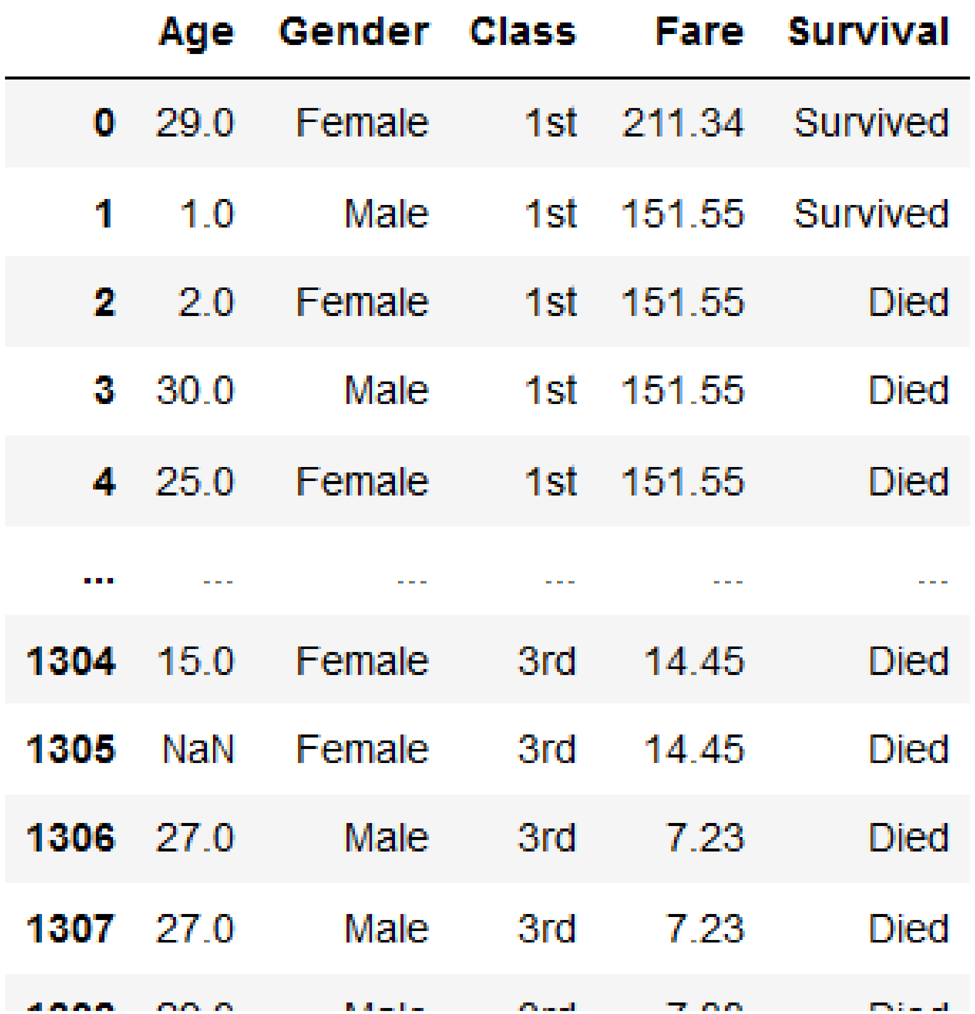
Log of the Likelihood function

$$\log(L) = \sum_{i=1}^{n}[Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)]$$

The values of the betas that maximizes the log(likelihood) function are the best estimates for the model

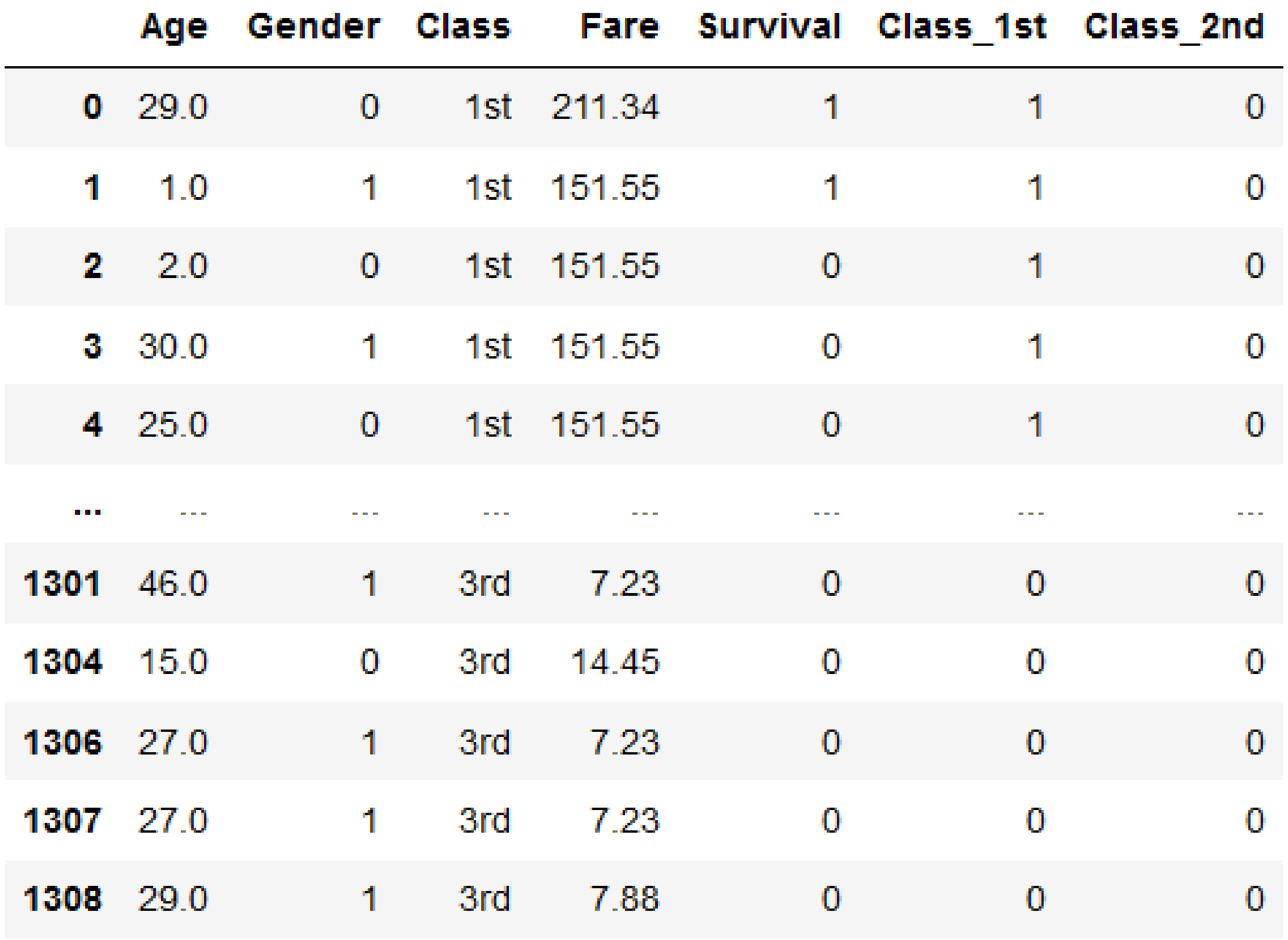
A logistic regression example

- The titanic dataset



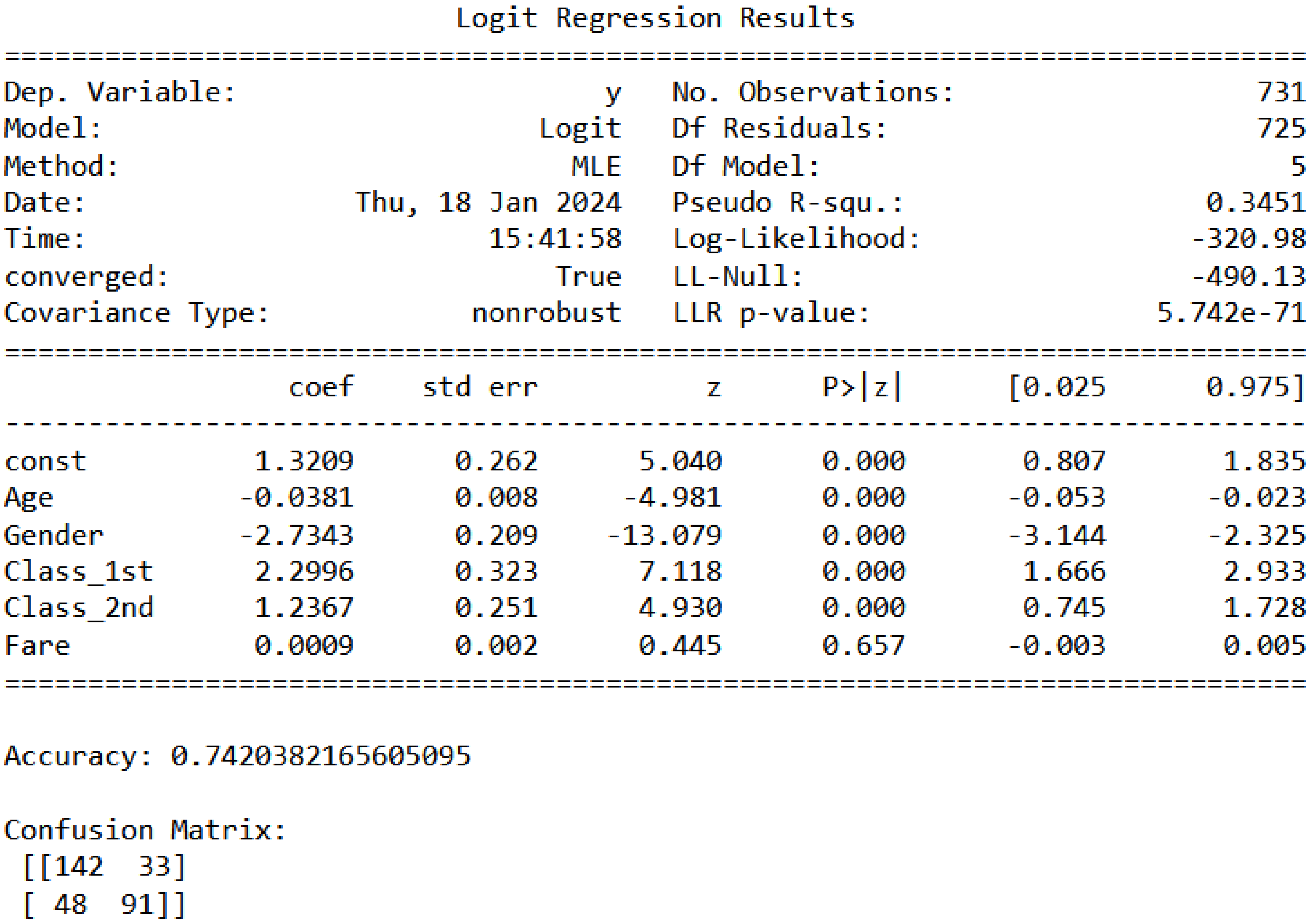
	Age	Gender	Class	Fare	Survival
0	29.0	Female	1st	211.34	Survived
1	1.0	Male	1st	151.55	Survived
2	2.0	Female	1st	151.55	Died
3	30.0	Male	1st	151.55	Died
4	25.0	Female	1st	151.55	Died
...	...	...	...	...	...
1304	15.0	Female	3rd	14.45	Died
1305	NaN	Female	3rd	14.45	Died
1306	27.0	Male	3rd	7.23	Died
1307	27.0	Male	3rd	7.23	Died
1308	29.0	Male	3rd	7.88	Died

Titanic dataset – Dummy coded



	Age	Gender	Class	Fare	Survival	Class_1st	Class_2nd
0	29.0	0	1st	211.34	1	1	0
1	1.0	1	1st	151.55	1	1	0
2	2.0	0	1st	151.55	0	1	0
3	30.0	1	1st	151.55	0	1	0
4	25.0	0	1st	151.55	0	1	0
...	...	...	...	...	...	...	...
1301	46.0	1	3rd	7.23	0	0	0
1304	15.0	0	3rd	14.45	0	0	0
1306	27.0	1	3rd	7.23	0	0	0
1307	27.0	1	3rd	7.23	0	0	0
1308	29.0	1	3rd	7.88	0	0	0

Logistic Regression Estimates & Metrics



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  731
Model:                          Logit   Df Residuals:                      725
Method:                           MLE   Df Model:                            5
Date:                Thu, 18 Jan 2024   Pseudo R-squ.:                  0.3451
Time:                        15:41:58   Log-Likelihood:                -320.98
converged:                       True   LL-Null:                       -490.13
Covariance Type:            nonrobust   LLR p-value:                 5.742e-71
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3209      0.262      5.040      0.000       0.807       1.835
Age           -0.0381      0.008     -4.981      0.000      -0.053      -0.023
Gender        -2.7343      0.209    -13.079      0.000      -3.144      -2.325
Class_1st      2.2996      0.323      7.118      0.000       1.666       2.933
Class_2nd      1.2367      0.251      4.930      0.000       0.745       1.728
Fare           0.0009      0.002      0.445      0.657      -0.003       0.005
==============================================================================

Accuracy: 0.7420382165605095

Confusion Matrix:
 [[142  33]
 [ 48  91]]
```

Logistic Regression Estimates

- The above results give the estimated model as:

$$\text{logit}(p) = 1.3209 - 0.0381*(\textbf{Age}) - 2.7343*(\textbf{Gender}) + 2.2996*(\textbf{Class\_1st}) + 1.2367*(\textbf{Class\_2nd}) + 0.0009*\textbf{(Fare)}$$

- Find the probability of a severe outage given
    - Age = 25; Gender = Female; Class = 2nd; Fare = 35
- Substitute logit(p) value with the numbers above in the following formula to get probability of survival for the above passenger

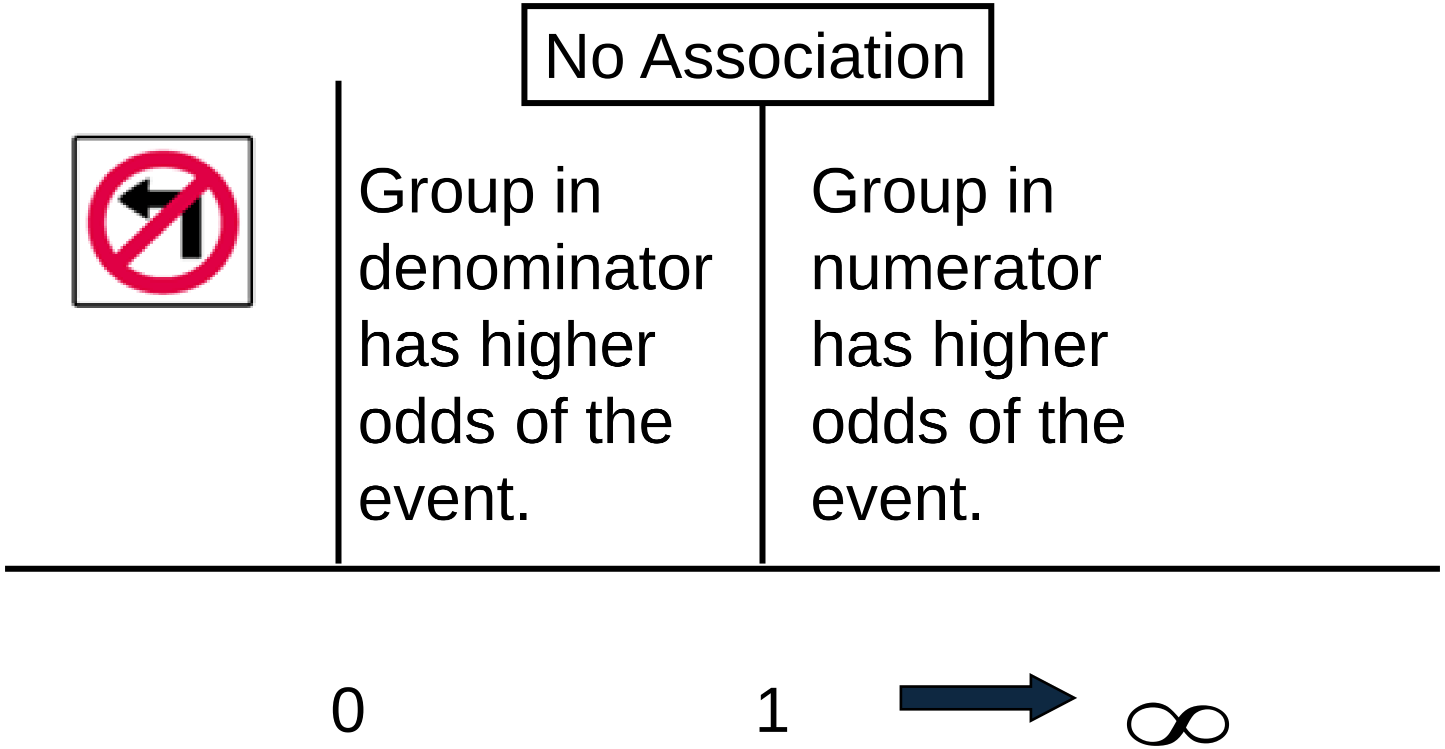
$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

Interpreting Logistic Regression Estimates

- Since we converted the odds to *log(odds)* to estimate the coefficients, we need to convert the coefficient of each independent variable by raising it to the power of **e** to get **odds ratio**
- In the above example, estimated odds ratio for variable **Age** =

  $e^{-0.0381}$= 0.96
- The odds ratio of 0.96 means that for **every year increase in passenger's age, the odds of survival decreases by 0.96 times** (or decreases by 4%).
- The estimated odds ratio for **Gender** = $e^{-2.7343}$ = 0.065
- The odds ratio of 0.065 means that the odds of survival for a **Male passenger is 0.065 times lower compared to a Female passenger** (or is 93.5% lower compared to a Female passenger)

Properties of the Odds Ratio



No Association

Group in denominator has higher odds of the event.

Group in numerator has higher odds of the event.

0

1

∞

Logistic Regression Fit Measures

- The logistic regression coefficients are estimated using maximum likelihood estimation (MLE).
- Unlike the Least Squares Estimation used for linear regression models, the MLE begins with a tentative solution, revised it slightly to see if it can be improved, and repeats until the results have converged.
- Logistic regression's goodness of fit is measured as (-2*(log likelihood of the fitted model)).
- Other measures are likelihood ratio, Cox and Snell R^2 and Nagelkerke R^2.
- Wald Statistic, which is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient, is used to assess the significance of each coefficient.

Logistic Regression – Problem

Assume that a dependent variable purchase is represented as **Z** with *1=purchased* and *0=not purchased*; $\mathbf{X_1}$ = Last purchase ($); $\mathbf{X_2}$ = customer (*1=business and 0=residential*)

The logit model is $\log\left(\frac{p_z}{1-pz}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Assume that with our data, the estimated results are as follows

$$\log\left(\frac{p_z}{1-pz}\right) = 3.5 + 2.1 * X_1 – 0.7 * X_2$$

From the above result, interpret the impacts of independent variable **Last purchase** and **customer** on the **odds of purchase**

Logistic Regression Estimation Tools

- There are different software options to estimate logistic regression
- The following are examples of add-ins available to solve logistic regression in Excel
  - **Real Statistics Resource Pack**: An Excel add-in available for free download from http://www.real-statistics.com
  - **XLSTAT-Base**: An Excel add-in available from https://www.xlstat.com/en/solutions/base with paid license
- **SAS**, **Stata**, and **SPSS Modeler** are examples of non-Excel software that has logistic regression modeling capabilities
- **R** is an open-source and popular statistical computing software that has logistic regression modeling capabilities

Sources

- “Business Intelligence and Analytics” by Sharda et al., Pearson Education, 2015
- “Advanced Business Analytics” Educator Training by SAS Inc.
- http://www.real-statistics.com
- http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/
- Miller, “Modeling Techniques in Predictive Analytics with R” FT Press