Ollscoil na Gaillimhe
University of Galway

MS5105 – Statistical Techniques for Business Analytics

Week 7 – **Logistic Regression**

Dr Mona Isazad Mashinchi– Academic Year 2024/25
Mona.IsazadMashinchi@universityofgalway.ie

University *of* Galway.ie

J. Bruce Ismay

Former managing director of the White Star Line

Born
12 Dec 1862
Crosby, United Kingdom

Died
17 Oct 1937
Mayfair, London, Unite...

Jonathan Hyde

Australian actor

Reddit

What are your thoughts on Bruce ismay? : r/titanic - Reddit

I don't think Ismay was a coward. No one on the ship actually WANTED to go into t…

21 Jan 2024

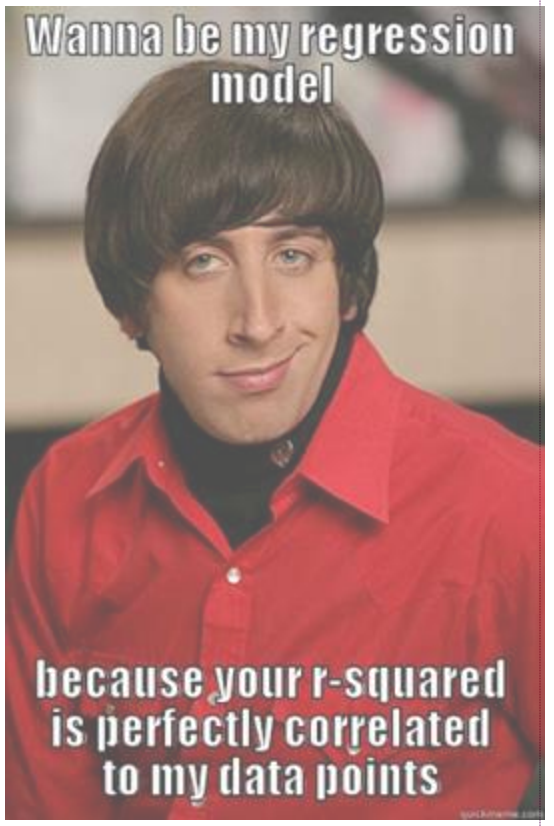
Student survey ☺

Your voice matters!

Recap of week 4- Regression

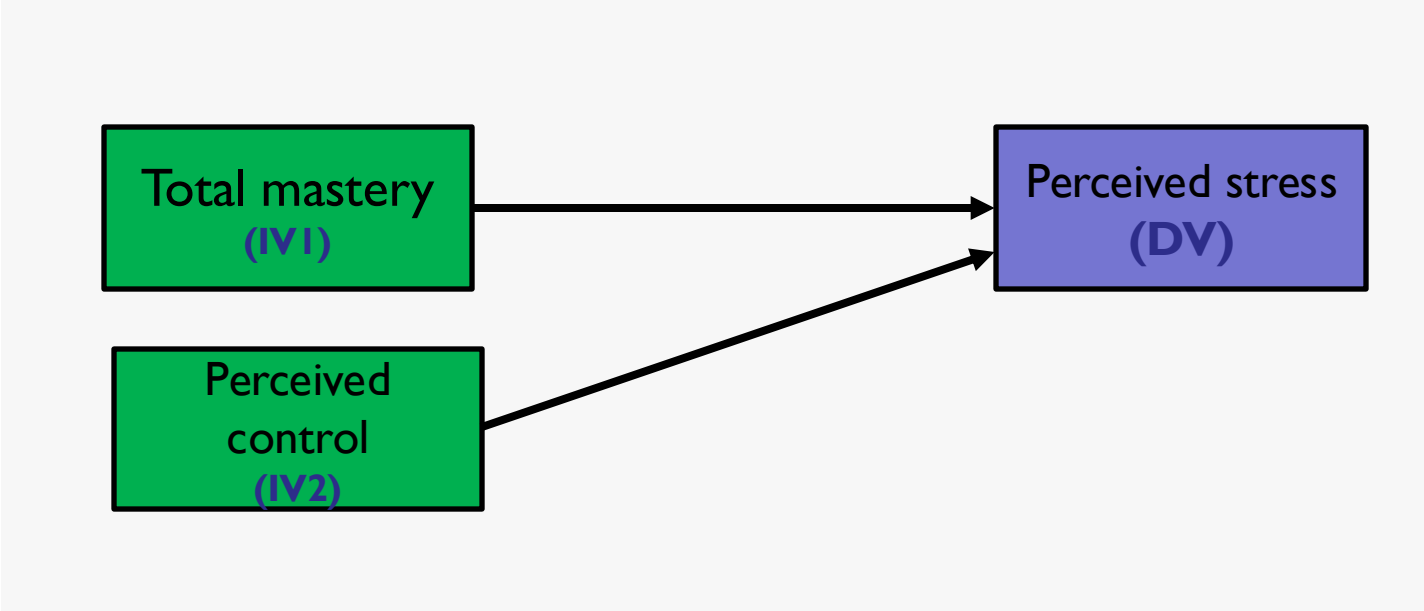
Key characteristics

- Predicts a continuous outcome: The model is designed to **predict a continuous numerical value** as the **dependent variable**.
- Assumes a linear relationship: It assumes that the **relationship** between the dependent and independent variables is **linear**.
- Independent variables can be continuous or categorical: Independent variables can be numerical or, **if categorical, must be converted to dummy variables**.
- Sensitive to outliers: Linear regression can be **sensitive to outliers**, as extreme values can disproportionately affect the model's predictions.
- Additive effects: Each predictor **has an independent and linear contribution** to the outcome.



Recap of week 4- Regression

- How well do the 2 measures of control (total mastery, pcoiss) predict perceived stress?



Let's start with a few questions☺

- What is logistic regression?
- When it can be used?
- How should I interpret the results?
- How should I write the results?

Regression

Regression is a statistical method used to understand the **relationship between variables** (i.e., **modelling relationships between variables**). In simple terms, it helps us predict the value of one variable (i.e., dependent variable) based on the value of another variable (i.e., independent variable).

Linear regression

- It predicts a continuous numerical outcome.
  - Example: predicting sales based on the amount of spend on digital marketing

Logistics regression

- It is a statistical method used to **predict a categorical outcome**, usually **binary** (e.g., yes/no, true/false, 0/1). It estimates the **probability that a certain event will occur** based on one or more independent variables. For example, we can use logistic regression to study the influence of factors such as **income**, **age**, **sex**, **education level,** and **marital status** on whether a **certain condition exists** or not (e.g., whether a person defaults on a loan).

Logistic Regression

Logistics regression

- Example: A bank is trying to predict whether a customer will default (0/1) on a loan using their income, credit score and employment status. In this case, the response would have two possible outcomes:
  - **0 = No Default (pay back)**
  - **1 = Default (fails to pay it back)**
  - ➢ **In this example you might want to estimate the probability for the occurrence of the characteristic 1(1 = Default) .**

The **independent variables** in logistic regression can be;

- **Continuous** (e.g., income, age, total spent).
- **Categorical** (e.g., job role, sex, marital status).
- **Binary** (e.g., smoker/non-smoker, rent a house/owner).

Customer ID	Income	Credit score	Employment status	Default
181987	25000	351	Unemployed	1
231879	89000	556	Full- time	0
67102	47800	876	Part- time	1
98222	68000	600	Unemployed	0
16989	59000	300	Full time	1

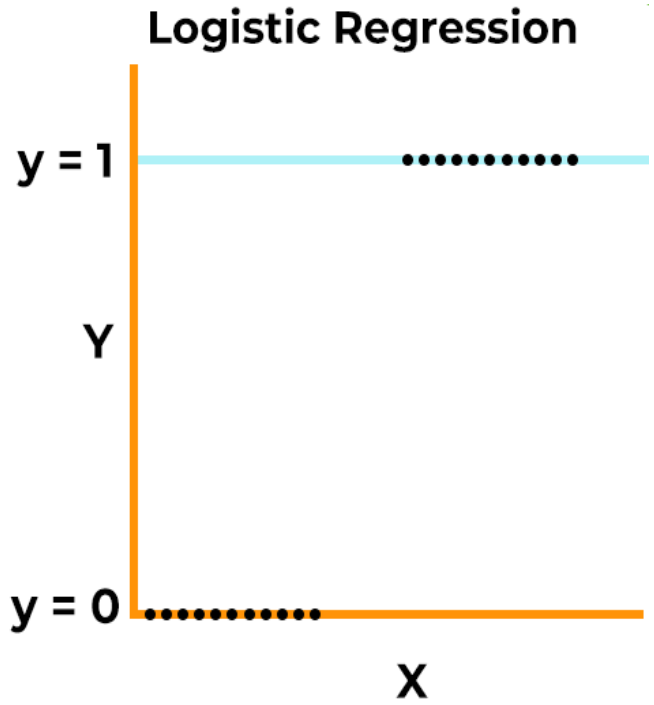
Logistic Regression

- We want to know what influence the independent variables have on the default.
- If there is an influence, then we can predict how likely a customer is to default on a loan.

- **So now why we cannot just use linear regression?**
- In linear regression, we use a regression equation with a **dependent variable**, **independent variables**, **and regression coefficients**.
- However, in some cases, the dependent variable is binary (either 0 or 1).
- In this scenario, linear **regression would simply fit a straight line** through the data points, potentially resulting in values that **extend beyond the range of 0 and 1**. But in **logistic regression**, the **goal is to estimate the probability of an event occurring**, which must be between 0 and 1.
- Therefore, **we need a function that limits the prediction values** to the **range between 0 and 1**, which is achieved through the logistic function.

Logistic Regression

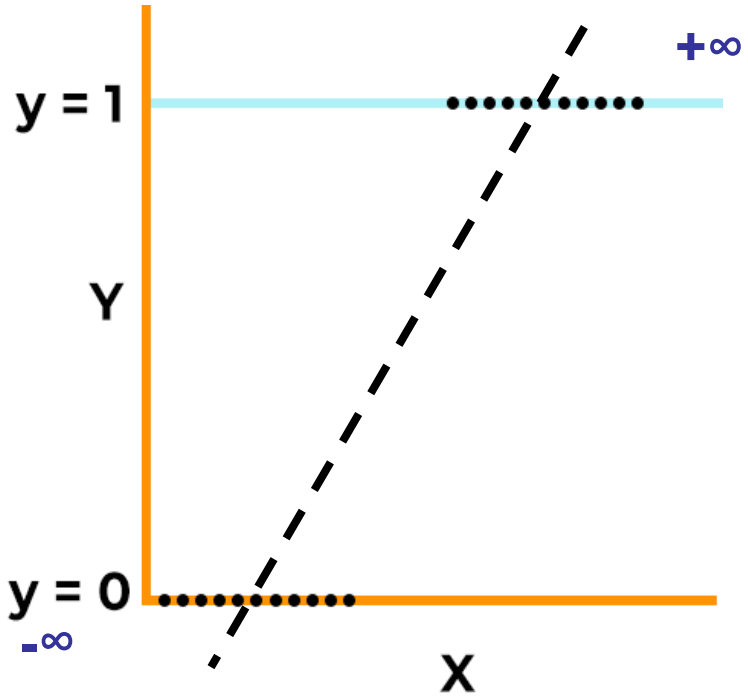
In logistic regression, our dependent variable can only take on the values 0 or 1.



Click here to explore

Logistic Regression

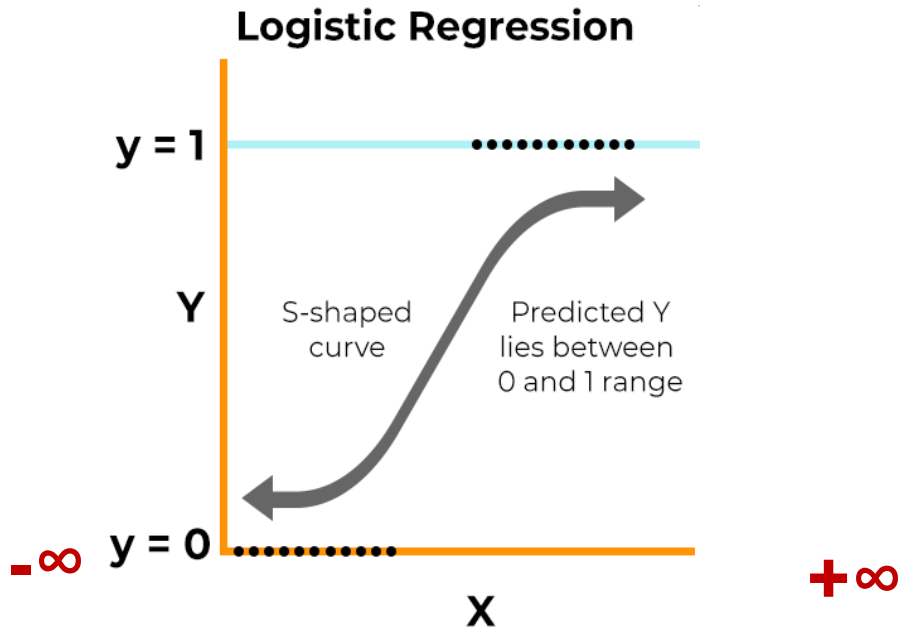
If we use linear regression, it would simply place a straight line through the points, allowing values between negative and positive infinity. However, the goal of logistic regression is to estimate the probability of occurrence, so the predicted values should fall between 0 and 1.



Click here to explore

Logistic Regression

- In logistic regression we want to **estimate the probability of occurrence.**
- The value range for **prediction should be between 0 and 1.**



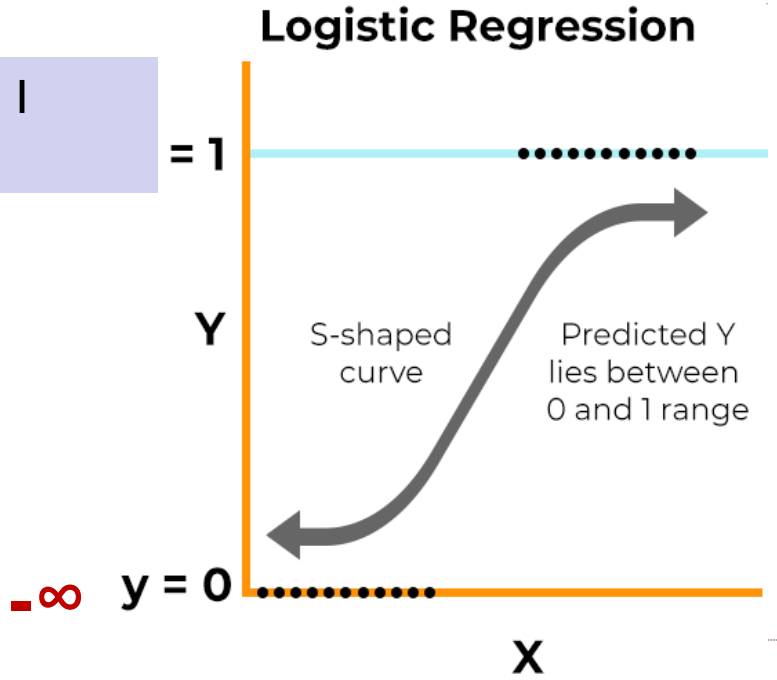
Click here to explore

Logistic Regression

So, we need a function that only takes value between 0 and 1 and that is exactly what logistic function does:

$$P(y=1 \mid x_1+x_2 \ldots x_n) = \frac{1}{1+e^{-(b_0+b_1 \cdot x_1+b_2 \cdot x_2+\cdots+bk \cdot xk)}}$$

Which calculate the probability that our dependent variable is 1 given values of the independent variables:



Logistic Regression

- Example: A bank is trying to predict whether a customer will default (0/1) on a loan using their income, credit score and employment status. In this case, the response would have two possible outcomes:
- Based on our example this is the equation. And we need to identify coefficients.

$$P(y=1 \mid x_1+x_2 \ldots x_n) = \frac{1}{1+e^{-(b_0+b_1 \cdot \text{income}+b_2 \cdot \text{credit score}+b_3 \cdot \text{eployment status})}}$$

SPSS

- Analyse → Regression → Binary Logistic → select dependent variable in dependent box and independent variables in block 1 → Select categorical → Bring the categorical variable in covariate and choose reference category as last then click continue → make sure the method is Enter and then select ok

Output

- **Observed**: These are the actual observed outcomes. It shows how many individuals in the dataset actually defaulted (Yes) or did not default (No).
- **Predicted**: These are the predictions made by the logistic regression model. The model predicts whether an individual will default or not, based on their characteristics (income, credit score, employment status, etc.).

Classification Table[a]

Observed			Predicted		
			Default		Percentage Correct
			No	Yes	
Step 1	Default	No	13	23	36.1
		Yes	5	159	97.0
	Overall Percentage				86.0

a. The cut value is .500

Output

B values represent the change in log-odds for a unit increase in the predictor variable. A positive B value means that as the predictor increases, the odds of the event happening go up. A negative B value means that as the predictor increases, the odds of the event happening go down.

- **Income**: For every increase of one unit in income, which in this case is 1,000 euros, the log-odds of default decrease by 0.084. So, this indicates that higher income reduces the likelihood of default.
- **Credit Score**: For every one-point increase in credit score, the log-odds of default increase by 0.010. This coefficient is positive, indicating that as the credit score increases, there is a slight increase in the likelihood of default, although it is a small effect.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Income	-.084	.018	22.981	1	<.001	.919
	Credit_Score	.010	.003	14.491	1	<.001	1.010
	Employment_Status			13.441	2	.001	
	Employment_Status(1)	2.146	.641	11.199	1	<.001	8.549
	Employment_Status(2)	1.381	.633	4.761	1	.029	3.980
	Constant	-.736	1.505	.239	1	.625	.479

a. Variable(s) entered on step 1: Income, Credit_Score, Employment_Status.

Output

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
Employment_Status	Unemployed	58	1.000	.000
	Part_time	47	.000	1.000
	Full_time	95	.000	.000

Dummy variables?

When conducting regression analysis with a dataset that contains non-numerical (categorical) variables, it is necessary **to create dummy variables for these non-numerical factors**.

- This process involves the creation of **n-1 columns**, where **'n' represents the number of categories (levels) within the variable**.
- For example, if a variable has three levels, you would need to create two (3-1) dummy variables.

Each new column corresponds to one of the levels, except one (as one level is represented by all zeroes in these dummy variables). **In each of these columns, a value of 1 is assigned to represent the presence of a particular level, and a 0 is assigned otherwise.** This conversion enables the inclusion of categorical data in regression analysis by converting it into a numerical format. After this step, the dataset is prepared for analysis.

Dummy variables?

- Let's take the variable **'Education Level'** with five levels: **'Primary', 'Secondary', 'High School', 'Undergraduate', and 'Postgraduate'**. Since this variable has five levels, we need to create n-1 = 5-1 = 4 dummy variables to include it in a regression analysis. Let's name these dummy variables as follows:

- Primary_Education (This represents whether the individual has primary education. It takes a value of 1 if this is the case, and 0 otherwise.)
- Secondary_Education (This represents whether the individual has secondary education. It takes a value of 1 if this is the case, and 0 otherwise.)
- High_School_Education (This represents whether the individual has high school education. It takes a value of 1 if this is the case, and 0 otherwise.)
- Undergraduate_Education (This represents whether the individual has undergraduate education. It takes a value of 1 if this is the case, and 0 otherwise.)

Sample	Education Level
1	Primary
2	Secondary
3	High School
4	Undergraduate
5	Postgraduate
6	Primary
7	High School
8	Secondary
9	Undergraduate
10	Postgraduate

Sample	Education Level	Secondary	High School	Undergraduate	Postgraduate
1	Primary	0	0	0	0
2	Secondary	1	0	0	0
3	High School	0	1	0	0
4	Undergraduate	0	0	1	0
5	Postgraduate	0	0	0	1
6	Primary	0	0	0	0
7	High School	0	1	0	0
8	Secondary	1	0	0	0
9	Undergraduate	0	0	1	0
10	Postgraduate	0	0	0	1

Output

- **Employment Status**: The reference category is **full time** (which does not appear in the table), and **two dummy variables are created** for **unemployed** and **part time.**
- Being unemployed **increases** the **log-odds of default by 2.146** compared to being full time employed.
- Being part time employed **increases** the **log-odds of default by 1.381** compared to being full time employed.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Income	-.084	.018	22.981	1	<.001	.919
	Credit_Score	.010	.003	14.491	1	<.001	1.010
	Employment_Status			13.441	2	.001	
	Employment_Status(1)	2.146	.641	11.199	1	<.001	8.549
	Employment_Status(2)	1.381	.633	4.761	1	.029	3.980
	Constant	-.736	1.505	.239	1	.625	.479

a. Variable(s) entered on step 1: Income, Credit_Score, Employment_Status.

Output

Exp(B)=Odds Ratio= The odds ratio provides an easier interpretation of the coefficients by converting the log-odds into an odds ratio. The Exp(B) shows the exponentiated value of the B coefficients. This gives the odds ratio, which tells you how the odds change for each unit increase in the predictor.
Exp(B) > 1: Increases the odds.
Exp(B) < 1: Decreases the odds.

Income: For every increase of 1,000 euros in income, the odds of defaulting decrease by a factor of 0.919. This means higher income reduces the likelihood of default.

Credit Score: For each one-point increase in credit score, the odds of defaulting increase slightly by a factor of 1.010.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Income	-.084	.018	22.981	1	<.001	.919
	Credit_Score	.010	.003	14.491	1	<.001	1.010
	Employment_Status			13.441	2	.001	
	Employment_Status(1)	2.146	.641	11.199	1	<.001	8.549
	Employment_Status(2)	1.381	.633	4.761	1	.029	3.980
	Constant	-.736	1.505	.239	1	.625	.479

a. Variable(s) entered on step 1: Income, Credit_Score, Employment_Status.

Output

Employment Status(1) (unemployed): Being unemployed increases the odds of defaulting by a factor of 8.549 compared to being full-time employed. This indicates that unemployment is a very strong predictor of default.
Employment Status(2) (part time): Being part-time employed increases the odds of defaulting by a factor of 3.980 compared to being full-time employed.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Income	-.084	.018	22.981	1	<.001	.919
	Credit_Score	.010	.003	14.491	1	<.001	1.010
	Employment_Status			13.441	2	.001	
	Employment_Status(1)	2.146	.641	11.199	1	<.001	8.549
	Employment_Status(2)	1.381	.633	4.761	1	.029	3.980
	Constant	-.736	1.505	.239	1	.625	.479

a. Variable(s) entered on step 1: Income, Credit_Score, Employment_Status.

Is this familiar?

You guess now? ☺

A	B	C	D	E	F	G	H	I	J	K	L	M	N
pclass	**survived**	**name**	**sex**	**age**	**sibsp**	**parch**	**ticket**	**fare**	**cabin**	**embarked**	**boat**	**body**	**home.dest**
1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
1	1	Allison, Master. Hudson Trevor	male	0.917	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville, ON
1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville, ON
1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville, ON
1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville, ON
1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
1	1	Andrews, Miss. Kornelia Theodosia	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
1	0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
1	0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
1	1	Aubart, Mme. Leontine Pauline	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
1	1	Barber, Miss. Ellen "Nellie"	female	26	0	0	19877	78.8500		S	6		
1	1	Barkworth, Mr. Algernon Henry Wilson	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
1	0	Baumann, Mr. John D	male		0	0	PC 17318	25.9250		S			New York, NY
1	0	Baxter, Mr. Quigg Edmond	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
1	1	Bazzani, Miss. Albina	female	32	0	0	11813	76.2917	D15	C	8		
1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
1	1	Beckwith, Mr. Richard Leonard	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
1	1	Bidois, Miss. Rosalie	female	42	0	0	PC 17757	227.5250		C	4		
1	1	Bird, Miss. Ellen	female	29	0	0	PC 17483	221.7792	C97	S	8		

Let's be a detective ☺

Let's be a detective and see what we can learn.

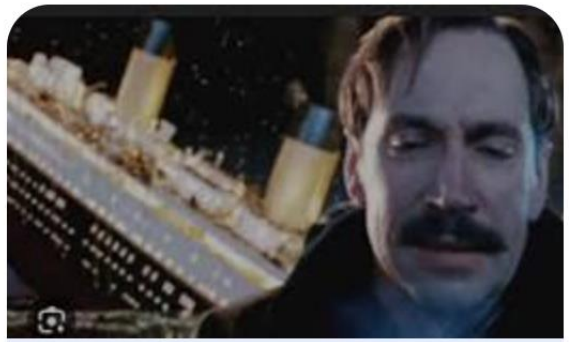
You do not need them, you have the data, you now know logistic regression, so you are a detective now, you can solve it … ☺

Let's focus on only one person…

- Let's be a detective and see what we can learn.

J. Bruce Ismay

Former managing director of the White Star Line



Reddit

What are your thoughts on Bruce ismay? : r/titanic - Reddit

I don't think Ismay was a coward. No one on the ship actually WANTED to go into t…

21 Jan 2024

Born

12 Dec 1862

Crosby, United Kingdom

Died

17 Oct 1937

Mayfair, London, Unite...

After the Titanic sank, the ship's owner hid away in Co Galway

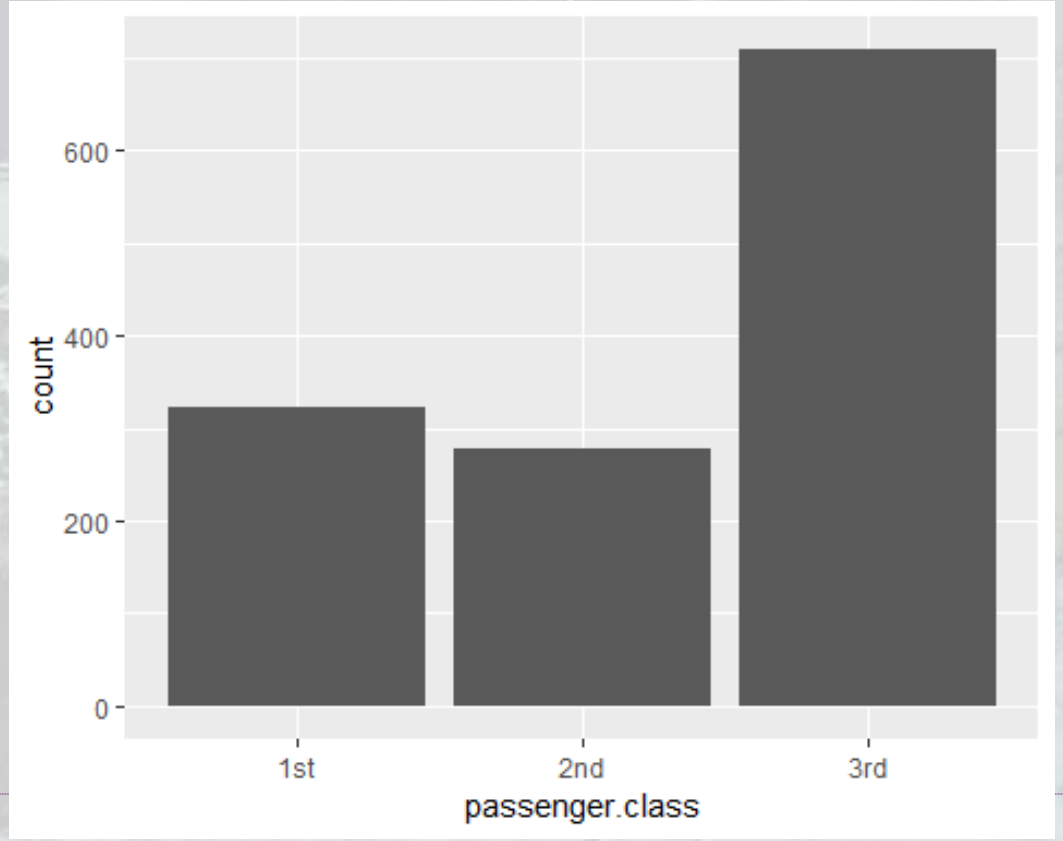
J. Bruce Ismay, the chairman of White Star Line who made the fatal decision to reduce the number of lifeboats on Titanic, settled in Co Galway after the 1912 disaster.

Our person of interest!

pclass	survived	name	sex	age	ticket	fare	cabin	boat	home.dest
1st	1	Allen, Miss. Elisabeth Walton	female	29	24160	211.337	B5	2	St Louis, MO
1st	1	Allison, Master. Hudson Trevor	male	0.9167	113781	151.55	C22 C26	11	Montreal, PQ / Chesterville, ON
1st	0	Allison, Miss. Helen Loraine	female	2	113781	151.55	C22 C26		Montreal, PQ / Chesterville, ON
1st	0	Allison, Mr. Hudson Joshua Crei	male	30	113781	151.55	C22 C26		Montreal, PQ / Chesterville, ON
1st	0	Allison, Mrs. Hudson J C (Bessi	female	25	113781	151.55	C22 C26		Montreal, PQ / Chesterville, ON
1st	1	Anderson, Mr. Harry	male	48	19952	26.55	E12	3	New York, NY
1st	1	Andrews, Miss. Kornelia Theodos	female	63	13502	77.9583	D7	10	Hudson, NY
1st	0	Andrews, Mr. Thomas Jr	male	39	112050	0	A36		Belfast, NI
1st	1	Appleton, Mrs. Edward Dale (Cha	female	53	11769	51.4792	C101	D	Bayside, Queens, NY
1st	0	Artagaveytia, Mr. Ramon	male	71	PC 17609	49.5042			Montevideo, Uruguay
1st	**1**	**Ismay, Mr. Joseph Bruce**	**male**	**49**	**112058**	**0**	**B52 B54 B**	**C**	**Liverpool**
1st	0	Astor, Col. John Jacob	male	47	PC 17757	227.525	C62 C64		New York, NY
1st	1	Astor, Mrs. John Jacob (Madelei	female	18	PC 17757	227.525	C62 C64	4	New York, NY
1st	1	Aubart, Mme. Leontine Pauline	female	24	PC 17477	69.3	B35	9	Paris, France
1st	1	Barber, Miss. Ellen \"Nellie\"	female	26	19877	78.85		6	

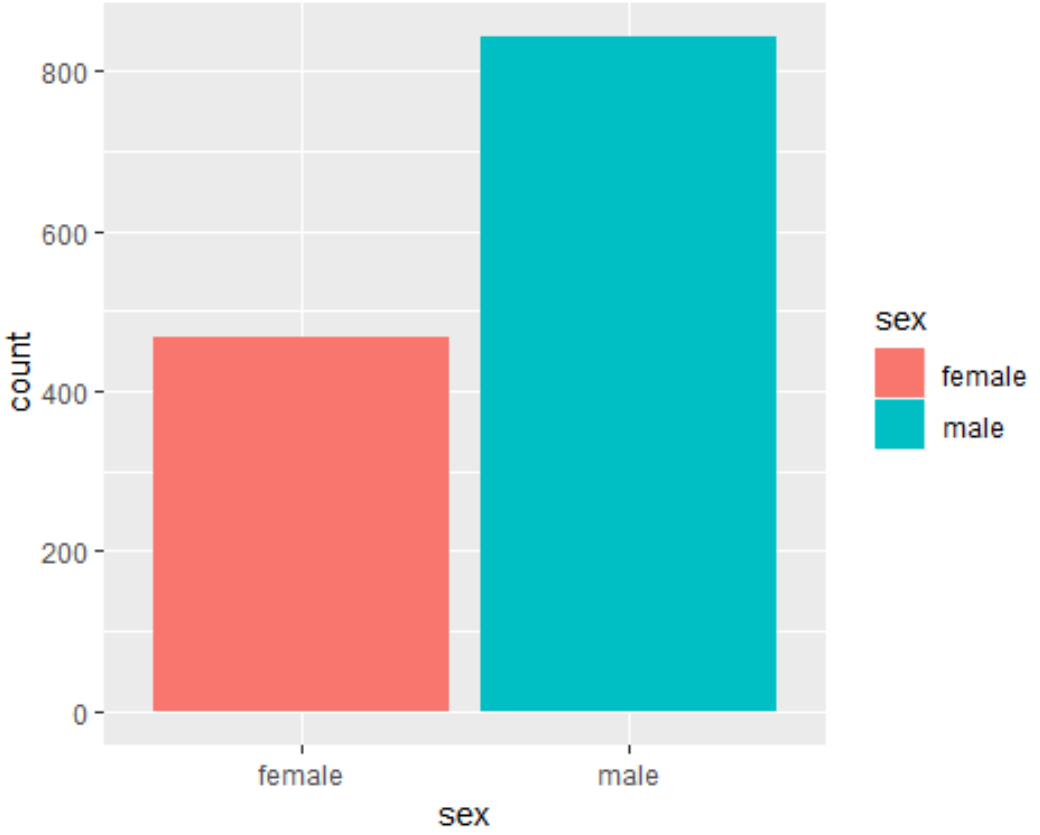
Titanic dataset

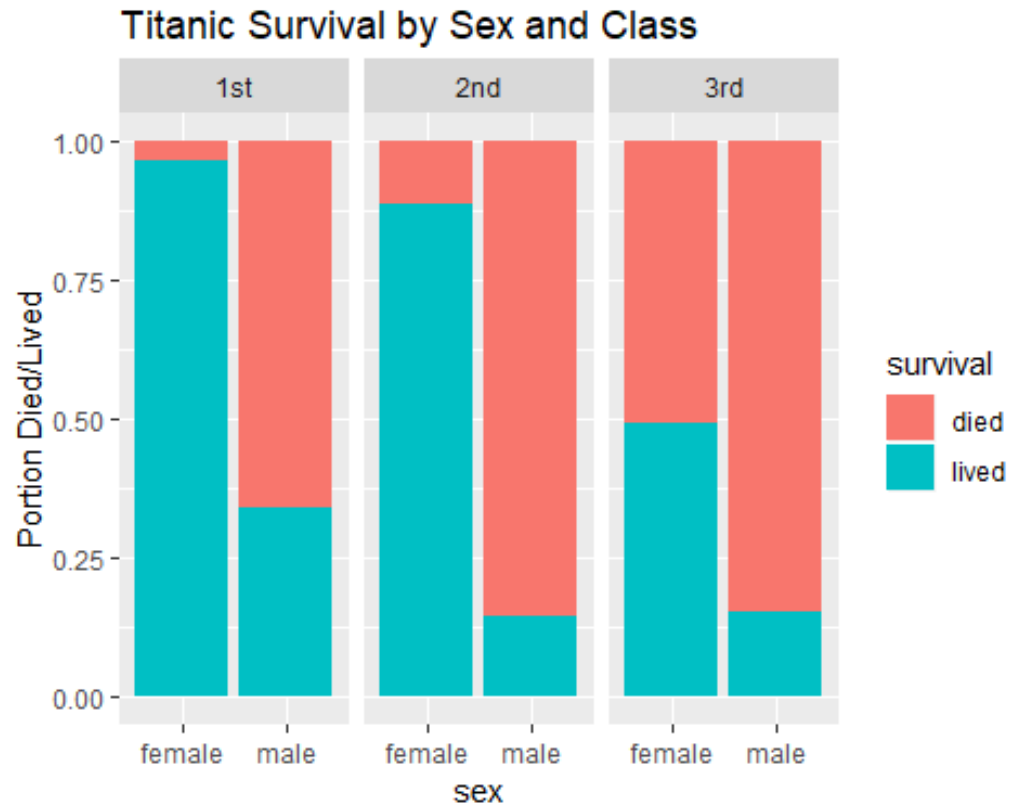
died	809
lived	500



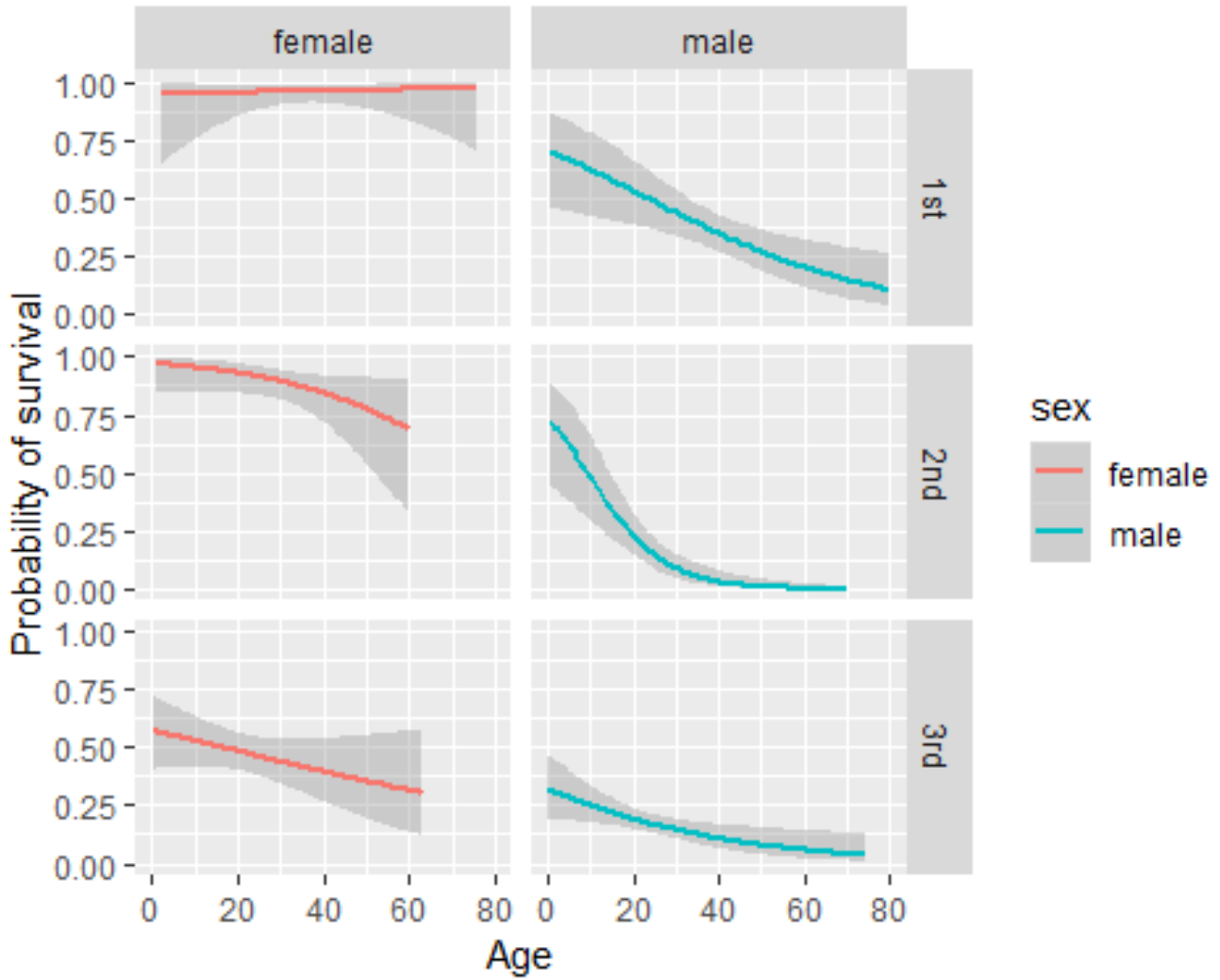
Titanic dataset

died	809
lived	500

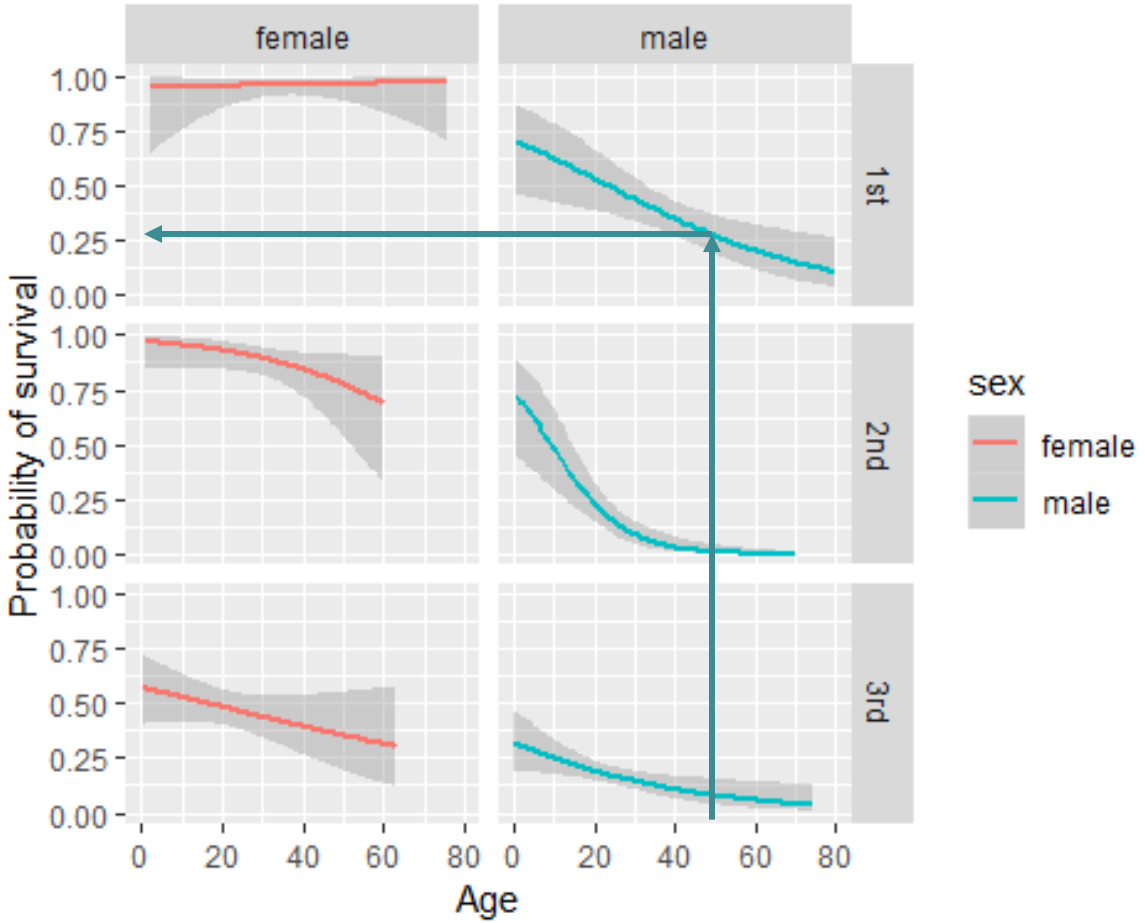




Survival by ticket class, sex and age!

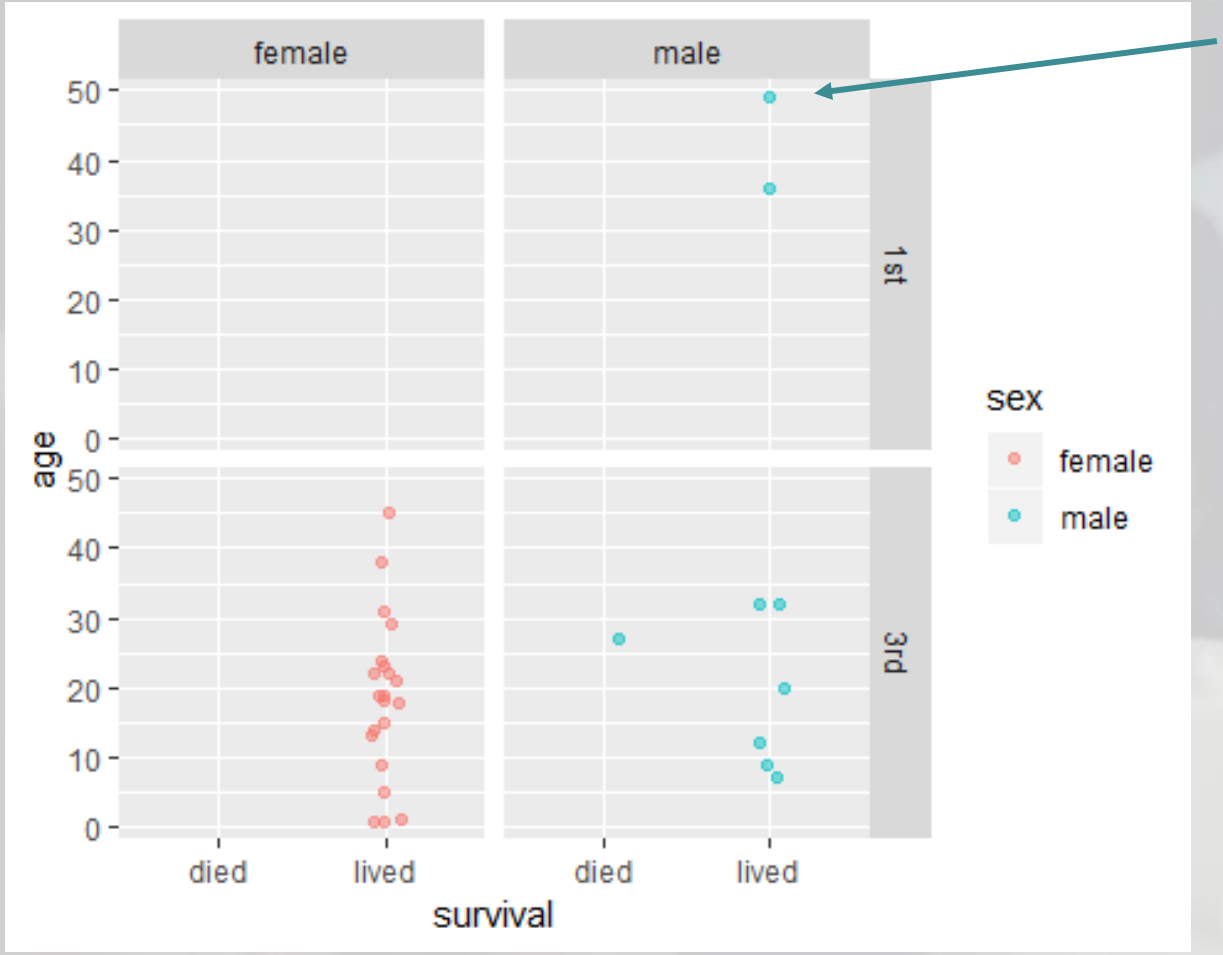


Estimate the probability of a 49 years old male from first class surviving !!! .



Estimated probability of surviving for J. Bruce Ismay ~ 0.25

Closer look at his lifeboat- lifeboat C



40 on board, while the capacity was 60?