ECON 140: Econometrics

Homework: Predicting Survival on the *Titanic*

Demian Pouzo — UC Berkeley

Background

In this assignment, you will work with data from passengers of the *RMS Titanic*. Your goal is to use regression analysis to predict the likelihood that a passenger survived the disaster, based on their characteristics.

The dataset provided is `titanic.csv`.

Answer every question in a report-style format and attach the respective code (R, Python, etc.) at the end. Submit a single, unified PDF or document via Gradescope.

1. Choosing the Model

The outcome variable is

$$Y_i = \begin{cases} 1 & \text{if passenger } i \text{ survived,} \\ 0 & \text{otherwise.} \end{cases}$$

Choose **four attributes** $X_{1i}, X_{2i}, X_{3i}, X_{4i}$ from the dataset that you believe help predict survival. Possible candidates include:

- `sex` (male/female)
- `age` (in years)
- `pclass` (1st, 2nd, 3rd class)
- `fare` (ticket price)
- `sibsp` (number of siblings/spouses aboard)
- `parch` (number of parents/children aboard)
- `embarked` (port of embarkation: C, Q, S)

2. Creating Dummy Variables

Some regressors are categorical (for example, `sex` or `embarked`), so you must create **dummy variables**. For instance, to create a dummy for being female:

$$\text{female}_i = \begin{cases} 1 & \text{if } \texttt{sex} = \text{female}, \\ 0 & \text{if } \texttt{sex} = \text{male}. \end{cases}$$

If a variable has K categories, create $K-1$ dummy variables to avoid perfect multicollinearity. For example, if $\texttt{embarked} \in \{C, Q, S\}$, you could define

$$\text{embarked\_C}_i = 1(\text{embarked} = C), \qquad \text{embarked\_Q}_i = 1(\text{embarked} = Q),$$

and use "S" as the omitted (baseline) category.

3. Estimating the Model

Estimate the model:

$$E[Y_i \mid X_{1i}, X_{2i}, X_{3i}, X_{4i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}.$$

Questions:

1. What is $E[Y_i \mid X_{1i}, \ldots, X_{4i}]$ in this context? What is it conceptually measuring? (Hint: think about the probability that a passenger with those attributes survives.)
2. What are you estimating when you run this regression?

4. Interpretation

After estimating the model:

1. Interpret each coefficient $\hat{\beta}_j$ (including the intercept). What is the expected change in the probability of survival when X_j increases by one unit, holding other attributes constant?
2. Test for statistical significance:
   - Individually: test $H_0 : \beta_j = 0$ for each $j = 1, \ldots, 4$.
   - Jointly: test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

5. Prediction

Generate five hypothetical passengers with different combinations of your chosen attributes. For example:

Passenger	sex	pclass	age	fare	sibsp	parch
1	female	1	25	100	0	0
2	male	3	30	10	1	0
3	female	2	40	50	0	2
4	male	1	60	80	0	0
5	female	3	22	15	1	1

Use your estimated model to compute the predicted probability of survival:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i}.$$

Then classify each passenger as:

$$\text{Predicted survive} = \begin{cases} 1 & \text{if } \hat{Y}_i > 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

6. Discussion

- Which attributes were the most important predictors of survival?
- Did your model correctly classify your five new passengers?
- How would you improve the model? (e.g., adding interaction terms or nonlinearities)