ECON 2823 | Spring 2026 | Checkpoint 3

Due: Friday, April 10

Work in groups. Use notebooks to compute answers. Learn something. Ask clarifying questions. Submit your answers to Gradescope. Have fun.

Q1. Hand-Coded MLE vs GLM

In Checkpoint 2, you estimated Liverpool's home scoring rate by writing a Poisson log-likelihood function and maximizing it with `scipy.optimize`. Here we show that `smf.glm()` does this automatically and can extend it to include covariates.

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
import statsmodels.api as sm
from scipy import stats

file_path = 'https://tayweid.github.io/econ-2823/parts/data/'
soccer = pd.read_csv(file_path + 'soccer/soccerData.csv')
```

a) Fit a Poisson GLM with only an intercept to Liverpool's home goals:

```
liverpool_home = soccer[soccer['HomeTeam'] == 'Liverpool']
intercept_model = smf.glm('FTHG ~ 1', data=liverpool_home,
family=sm.families.Poisson()).fit()
```

Compute $\exp(\hat{\beta}_0)$. Verify that this equals the MLE $\hat{\lambda}_H$ you found in Checkpoint 2 (the sample mean of Liverpool's home goals).

b) Now use the full dataset (all teams, all matches) and fit a Poisson GLM for home goals with `HomeTeam` as the predictor:

```
home_model = smf.glm('FTHG ~ C(HomeTeam)', data=soccer,
family=sm.families.Poisson()).fit()
```

What does the intercept represent? What does the coefficient on `C(HomeTeam)[T.Liverpool]` represent? Compute the expected home goals for Liverpool from these coefficients and verify it matches (a).

c) Fit a Gaussian GLM (identity link) with the same specification: `FTHG ~ C(HomeTeam)`. Compare the Liverpool coefficient to the Poisson model. Are the coefficients the same? Are the fitted values (predicted means) the same? Explain the difference in 1-2 sentences.

d) The Poisson model assumes $\mathrm{Var}(Y) = \mu$. For each team, compute the sample mean and sample variance of home goals. Plot variance against mean with a 45-degree line. Does the Poisson variance assumption look reasonable?

Q2. Model Selection and Marginal Effects in a Poisson GLM

Using the full soccer dataset, we compare competing models for home goals and compute marginal effects.

a) Fit three nested Poisson GLMs: 1) intercept only, 2) home team fixed effects, 3) home team FE + away team FE. Report the log-likelihood and number of parameters for each.

b) Conduct a likelihood ratio test of Model A vs Model B. How many degrees of freedom? What is the p-value? Can we reject the null that all home teams score at the same rate?

c) Conduct a likelihood ratio test of Model B vs Model C. What is the null hypothesis in words? What do you conclude?

d) Compare all three models using AIC. Do they agree with each other and with the LRT results?

Q3. Binary Outcome Models: Term Life Insurance

The file `life_insurance.csv` contains survey data on 2,000 households. The binary outcome `Owns.Term.Life` indicates whether the household owns a term life insurance policy.

```
insurance = pd.read_csv(file_path + 'life_insurance.csv')
insurance.head()
```

a) Fit a logistic regression predicting `Owns.Term.Life` using all available predictors. What is the estimated multiplicative effect on the odds ratio from having a high-school education as opposed to a college education? *(Hint: take the exponential of the difference in estimated parameter values.)*

b) Using your estimated model, plot the predicted probability of owning term life as a function of age for two household profiles making $60,000 and $110,000.

c) Fit a probit model with the same specification as in (a). Use AIC to determine which model has a better fit.

d) Compute the AME of `hhsize` from the logit model.

e) Fit an LPM with the same specification. Compare its `hhsize` coefficient to the logit AME. Then check the LPM for any out-of-bounds predictions.

f) If you were trying to target households without term life who you thought would want it in the near future, what characteristics would your analysis suggest you should be looking for?