Practicals and Assignments: Regression in Agricultural Data

Prepared by

Aditya Korekallu Srinivasa

Scientist (Senior Scale)

ICAR-Indian Agricultural Research Institute

New Delhi

April 24, 2025

Contents

1 Simulated Agriculture Datasets

1.1 Dataset 1: Crop Yield and Farm Inputs (Cross-sectional)

Description: 200 farms, variables:

- `Yield`: Wheat yield (quintals/ha)
- `Fertilizer`: Nitrogen fertilizer used (kg/ha)
- `Irrigation`: Irrigation applied (mm)
- `SeedRate`: Seed rate (kg/ha)
- `SoilQuality`: Soil quality index (1=poor, 2=average, 3=good)
- `FarmerEdu`: Farmer's education (years)
- `Region`: Region (factor: North, South, East, West)
- `AdoptTech`: Adoption of new technology (0=No, 1=Yes)

1.2 Dataset 2: Panel Data on Milk Yield (Farms over 5 Years)

Description: 50 dairy farms, 5 years (250 obs), variables:

- `FarmID`: Farm identifier
- `Year`: Year (2018–2022)
- `MilkYield`: Milk yield per cow (liters/day)
- `Feed`: Feed supplied (kg/cow/day)
- `VetVisits`: Veterinary visits per year
- `Breed`: Breed (0=local, 1=improved)
- `FarmerAge`: Farmer's age (years)
- `AdoptAI`: Adoption of artificial insemination (0=No, 1=Yes)
- `Rainfall`: Annual rainfall (mm)

See R code for simulation details.

2 Unit-wise Practicals and Assignments

Unit 1: Classical Linear Regression

Exercises

1. Using Dataset 1, regress `Yield` on `Fertilizer`, `Irrigation`, `SeedRate`, and `SoilQuality`. Interpret the coefficients.
2. Check the assumptions of linear regression using diagnostic plots.
3. Using Dataset 2, regress `MilkYield` on `Feed`, `VetVisits`, and `Breed`. Interpret results.

Assignment

1. For Dataset 1, create a new variable `Experience` (simulate as 5–35 years), add it to the regression, and discuss changes in results.
2. For Dataset 2, fit a fixed effects model (farm-level) for `MilkYield` and interpret the farm effects.

Unit 2: Multicollinearity

Exercises

1. For Dataset 1, calculate the correlation matrix among `Fertilizer`, `Irrigation`, and `SeedRate`.
2. Compute VIF for all regressors in the yield regression.
3. Simulate a new variable `FertIrrig` = `Fertilizer` + 0.8*`Irrigation`, add to the regression, and observe VIF and coefficient changes.

Assignment

1. For Dataset 2, add a new variable `FeedPlus` = `Feed` + 0.9*`VetVisits` and examine multicollinearity.
2. Discuss remedies for multicollinearity in both datasets.

Unit 3: Heteroscedasticity

Exercises

1. For Dataset 1, plot residuals from the yield regression against fitted values and `Fertilizer`.
2. Perform Breusch-Pagan and White tests for heteroscedasticity.
3. Transform `Yield` using log and re-run the regression. Compare residual plots.

Assignment

1. For Dataset 2, test for heteroscedasticity in the `MilkYield` regression and apply robust standard errors.
2. Discuss the impact of heteroscedasticity on inference.

Unit 4: Autocorrelation

Exercises

1. For Dataset 2, regress `MilkYield` on `Feed`, `VetVisits`, and `Year`. Plot residuals by year for a selected farm.
2. Compute the Durbin-Watson statistic for the regression.
3. Simulate AR(1) errors in `MilkYield` for one farm and estimate the autocorrelation coefficient.

Assignment

1. Fit a model correcting for autocorrelation (e.g., Cochrane-Orcutt or using nlme::gls).
2. Discuss how autocorrelation affects standard errors and inference.

Unit 5: Model Misspecification

Exercises

1. For Dataset 1, regress `Yield` on `Fertilizer` only. Compare with the full model using RESET test.
2. For Dataset 2, fit a model for `MilkYield` omitting `Feed`. Compare results and discuss omitted variable bias.
3. Test for functional form (add quadratic term for `Fertilizer`) and compare models.

Assignment

1. For Dataset 1, simulate measurement error in `Fertilizer` and examine its impact on OLS estimates.
2. For Dataset 2, compare AIC/BIC for different model specifications.

Unit 6: Other Issues in Regression

Exercises

1. For Dataset 1, identify outliers and high-leverage points using Cook's distance and leverage plots.
2. Simulate missing values in `Yield` and discuss handling approaches.
3. For Dataset 2, check for non-normality of residuals and apply a transformation if needed.

Assignment

1. For Dataset 1, simulate endogeneity by making `Fertilizer` correlated with the error term. Estimate using IV (2SLS).
2. For Dataset 2, create a hierarchical structure (e.g., farms within districts) and discuss how to model using mixed effects.

Unit 7: Qualitative Variables in Regression

Exercises

1. For Dataset 1, use `Region` and `SoilQuality` as dummy variables in the yield regression. Interpret coefficients.
2. For Dataset 2, use `Breed` and `AdoptAI` as dummies in the milk yield regression.
3. Run a logit model for `AdoptTech` (Dataset 1) as a function of `Education`, `Income`, and `Region`.

Assignment

1. For Dataset 1, estimate and interpret a probit and LPM for `AdoptTech`.
2. For Dataset 2, estimate a logit model for `AdoptAI` and interpret odds ratios and marginal effects.

Unit 8: Simultaneous Equation Models

Exercises

1. For Dataset 1, suppose `Yield` and `AdoptTech` are jointly determined. Specify a simultaneous system (e.g., yield depends on adoption, adoption depends on yield and other factors).
2. Simulate data for such a system and estimate using 2SLS.
3. Test identification using order and rank conditions.

Assignment

1. For Dataset 2, suppose `MilkYield` and `AdoptAI` are jointly determined. Specify and estimate a simultaneous system.
2. Discuss the economic interpretation and identification.