Regression Functional Form

PSCI 1801 · Class Handout

Dr. Marc Trussler

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Categorical Variables in Regression

Any numeric variable can go into a regression. If a categorical variable has a natural ordering and roughly equal gaps between levels, you can just treat it as continuous. But when the gaps aren't equal — or the variable has no ordering at all — you need dummy variables.

Dummy variables

Create a 0/1 indicator for each category. Then omit one — the reference category *— and include the rest.*

Example: predicting income by region (Northeast, Midwest, South, West). Include three dummies; leave one out.

R does this automatically if the variable is a `factor`*:*

Interpretation

- The intercept is the predicted y for the reference (omitted) category, when other predictors are 0.
- Each dummy coefficient is the difference in y between that category and the reference, holding other predictors constant.

You cannot include all *four dummies at once — that would create perfect collinearity (they sum to 1). R would refuse.*

Regression Assumptions (Gauss-Markov)

OLS gives unbiased and efficient estimates under five assumptions. Violations vary in severity.

GM 1: Full rank (no perfect collinearity)

Predictors must not be linear combinations of each other. R will refuse to estimate a model that violates this — nothing to do.

GM 2: Correct functional form

The relationship between x and y must match the model you specified. Two big violations: nonlinear relationships and unmodeled interactions. Covered below.

GM 3: No heteroskedasticity

Errors should have constant variance across levels of x. When variance grows with x (or shrinks), the coefficient is still unbiased but the standard errors are wrong — *often too small.*

Diagnosis: plot residuals vs. fitted values, or use `lmtest::bptest()`. *Fixes:*

1. Add explanatory variables that reduce unmodeled variance (best)
2. Use robust (Huber-White) standard errors: `sandwich::vcovHC(model, type = "HC1")` + `lmtest::coeftest()`

GM 4: No autocorrelation

Observations should be independent. Violated by:

- **Time series**: adjacent time points share unmodeled shocks
- **Clustered data**: observations within groups (students in schools, counties in states) share unmodeled traits

Fixes: cluster-robust SEs, mixed models, or explicit time-series methods. Beyond this course.

GM 5: Causality (zero conditional mean errors)

$$E[\epsilon_i \mid X] = 0$$

Formally: errors uncorrelated with predictors. In plain English: no confounders. This is what "correlation causation" is really about.

Handling Nonlinear Relationships

Some relationships aren't lines. Example: household income vs. test scores — big returns at the low end, flatter at the high end.

Polynomial terms

Add x^2 (and possibly higher powers) as a predictor:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

R syntax:

Use `raw = TRUE` with `poly()` — omitting it produces orthogonal polynomials with weird coefficients.

Interpreting a polynomial

The effect of x on y now depends on the current value of x:

$$\frac{\partial y}{\partial x} = \hat{\beta}_1 + 2\hat{\beta}_2 x$$

So a single coefficient no longer tells the whole story. Report the effect at meaningful values of x (e.g., at the 25th, 50th, 75th percentiles).

Categorical Variables Reframed

A categorical variable with k levels acts like $k-1$ separate slope tests. Region dummies, for instance, are asking "how much higher/lower is the outcome in each region compared to the reference," net of other predictors.

The choice of reference category doesn't change the model's fit or predictions — only how the coefficients are labeled. Pick a reference that makes for clean interpretation.