DEPARTMENT OF RESEARCH
GENERAL EDUCATION PROGRAM
MATHEMATICS AREA
College Algebra & Applied Calculus

Polynomial Model Selection with Small Samples: *Fitting, Generalization, and Cross-Validation in the Analysis of International Tourism in Nicaragua (2020–2024)*

Research Project Submitted for
College Algebra (Spring 2026)

Sara Samira Silwany Garcia[1]
Fernando José Hernández Gómez, Ph.D.[1,2,*]

[1] Keiser University Latin American Campus, San Marcos, Nicaragua
[2] Department of Research, Mathematics Area

*Corresponding Author
ORCID: 0000-0002-5132-2769
Email: fernando.hernandez@keiseruniversity.edu

Academic Supervisor

Fernando José Hernández Gómez, Ph.D.
Professor of Mathematics

San Marcos, Nicaragua
July 27, 2026

Abstract

This study addresses model selection in data-scarce regimes using five annual observations of international tourism arrivals in Nicaragua (2020–2024). Four candidate specifications (linear, quadratic, cubic, and exponential) are fitted and evaluated via leave-one-out cross-validation (LOOCV), the PRESS statistic, and the corrected AIC_c. The results demonstrate that R^2 is a severely biased proxy for predictive performance when $p/n \geq 0.4$. The cubic model ($R^2 = 0.971$) exhibits an out-of-sample prediction error six times larger than that of the linear model (PRESS $= 2.61 \times 10^6$ vs. 4.31×10^5) and yields physically impossible forecasts. All criteria converge on the linear model as the optimal specification. Projections for 2025–2030 with 95% parametric bootstrap intervals are provided, together with a fully reproducible Python pipeline.

Keywords: model selection; leave-one-out cross-validation; PRESS; corrected AIC; overfitting; tourism in Nicaragua; polynomial regression; statistical parsimony.

Contents

1 Introduction

1.1 The problem

Curve fitting in introductory courses often stops at the coefficient R^2, presenting it as the definitive quality metric. This practice ignores the fundamental distinction between *in-sample fit* and *out-of-sample predictive ability*. In small-sample settings ($n \lesssim 20$), R^2 systematically selects overfitted models that interpolate the data yet lack predictive validity.

With $n = 5$ observations and models with up to $p = 4$ parameters, this article demonstrates how an $R^2 = 0.971$ for the cubic model conceals predictive instability that only cross-validation uncovers. The principles are transferable to any domain where data are scarce.

1.2 Methodological gap

The gap is twofold: (1) *statistical*—standard expositions do not discuss the effect of p/n on the bias of R^2; and (2) *applied*—tourism time series in developing economies frequently exhibit such small-sample conditions, especially after shocks like the COVID-19 pandemic.

1.3 Research question and objectives

Can a higher-degree polynomial model, despite exhibiting a better R^2, be preferable to a linear model when the selection criterion incorporates out-of-sample predictive ability (LOOCV, AIC_c, parsimony)?

Objectives:

1. Formalize the bias of R^2 and the corrected R^2_{adj}.
2. Compare AIC, AIC_c, BIC, and PRESS-LOOCV for four candidate models.
3. Show why the cubic model produces absurd predictions via the hat matrix.
4. Derive 2025–2030 projections with bootstrap confidence intervals.
5. Link to Tikhonov regularization and the principle of parsimony.
6. Provide a reproducible Python pipeline.

1.4 Contributions

Table 1: Contributions and curricular context

Contribution	Description	Context
1. R^2 bias	Derivation of bias for $p/n \geq 0.4$.	Estimation theory
2. LOOCV / hat matrix	PRESS formula via rank-1 update.	Cross-validation
3. Application	LOOCV and AIC_c applied to Nicaraguan tourism.	Real data
4. Projections	Bootstrap intervals 2025–2030.	Predictive uncertainty
5. Regularization	Ridge with optimal λ by LOOCV.	Inverse problems
6. Reproducibility	Modular Python code.	Scientific standard
7. Interpretability	Linear model from explainable analytics.	Interpretable ML

2 Theoretical Framework

2.1 The R^2 estimator and its bias

Definición 1 (Coefficient of determination)**.**

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Proposición 1 (Bias of R^2)**.** *Under normality and correct specification, $E[R^2] > \rho^2$ for $p > 1$:*

$$\mathrm{Bias}(R^2) \approx \frac{p\,(1-\rho^2)}{n-p-1},$$

which diverges when $n-p-1 \to 0$. The adjusted version $R^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-p-1}(1-R^2)$ reduces the bias but does not estimate out-of-sample predictive ability.

Sketch. Under H_0, $R^2 \sim \mathrm{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right)$. The bias for $\rho^2 > 0$ is obtained from a second-order expansion of $E[\mathrm{RSS}/\mathrm{TSS}]$ (see [11], Chap. 10). □

Observación 1. For the cubic model ($n = 5$, $p = 4$): $n - p - 1 = 0$. R^2_{adj} is undefined. This invalidates inference based on the F distribution.

Table 2: Expected bias of R^2 ($n = 5$, bound $\rho^2 = 0$)

Model	p	$n-p-1$	Bias	Diagnosis
Linear	2	2	0.667	Moderate
Quadratic	3	1	2.000	Severe
Cubic	4	0	∞	**Collapse**
Exponential	2	2	0.667	Moderate

2.2 LOOCV and the PRESS statistic

Definición 2 (PRESS)**.**

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2,$$

where $\hat{y}_{(i)}$ is the prediction obtained when leaving observation i out.

Teorema 2 (PRESS formula via the hat matrix)**.** *Let* $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$*. Then:*

$$y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}}, \qquad \text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2.$$

Sketch. Application of the Sherman–Morrison formula to $(\mathbf{X}^\top\mathbf{X})^{-1}$ when removing row i (see [7], Sec. 7.10). □

Observación 2. When $h_{ii} \to 1$ (extreme points, $p \approx n$), the LOOCV error explodes. For the cubic model at $x = 0$: $h_{00} \approx 0.952$, amplification $\approx 21\times$, LOOCV error of $+1\,101$ (thousands of tourists).

2.3 Information criteria

$$\text{AIC} = n \ln(\text{RSS}/n) + 2p, \tag{1}$$

$$\text{BIC} = n \ln(\text{RSS}/n) + p \ln n, \tag{2}$$

$$\text{AIC}_c = \text{AIC} + \frac{2p(p+1)}{n-p-1}. \tag{3}$$

The correction AIC_c [9] is mandatory when $n - p - 1 \leq 1$. For the cubic model it diverges to $+\infty$.

2.4 Tikhonov regularization

Ridge regression: $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$, $\lambda > 0$. Guarantees invertibility even when $\mathbf{X}^\top\mathbf{X}$ is singular [7, 16].

2.5 Interpretability and parsimony

The linear model is intrinsically interpretable [12]: (β_0, β_1) carry direct economic meaning without post-hoc approximations [6, 10].

3 Data, Variables, and Context

3.1 Source

Annual international tourist arrivals to Nicaragua (2020–2024), in thousands, from official records [14]. The variable x denotes years since 2020.

Table 3: International tourist arrivals to Nicaragua (thousands), 2020–2024

Year	x	$T(x)$	ΔT	Context
2020	0	474.421	—	COVID-19
2021	1	312.425	−162.0	Pandemic low (−34%)
2022	2	932.700	+620.3	Reopening (+199%)
2023	3	1 202.3	+269.6	Consolidation
2024	4	1 090.0	−112.3	Adjustment

3.2 Finite differences

First differences: −162, +620, +270, −112 (thousands). Second differences: +782, −351, −382. The variability indicates that no fixed polynomial faithfully represents the dynamics (COVID shock + non-linear recovery).

3.3 Dispersion

Mean $\bar{T} = 802$ thousand, $S = 339$ thousand, $CV = 0.42$. High dispersion supports the use of robust metrics over R^2 alone.

4 Candidate Models and Estimation

$$\text{Linear:} \quad T(x) = \beta_0 + \beta_1 x \tag{4}$$

$$\text{Quadratic:} \quad T(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{5}$$

$$\text{Cubic:} \quad T(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \tag{6}$$

$$\text{Exponential:} \quad T(x) = A\, e^{kx} \tag{7}$$

4.1 Estimated coefficients

$$\hat{T}_{\text{lin}}(x) = 378.163 + 212.103\,x$$
$$\hat{T}_{\text{quad}}(x) = 342.265 + 283.899\,x - 17.949\,x^2$$
$$\hat{T}_{\text{cub}}(x) = 458.632 - 550.066\,x + 563.887\,x^2 - 96.973\,x^3$$
$$\hat{T}_{\text{exp}}(x) = 389.2\,e^{0.3011\,x}$$

Linear model: $\hat{\beta}_0 = 378$ (baseline level), $\hat{\beta}_1 = 212$ (thousands/year). The cubic coefficients lack economic interpretation and are highly unstable.

4.2 Graphical comparison

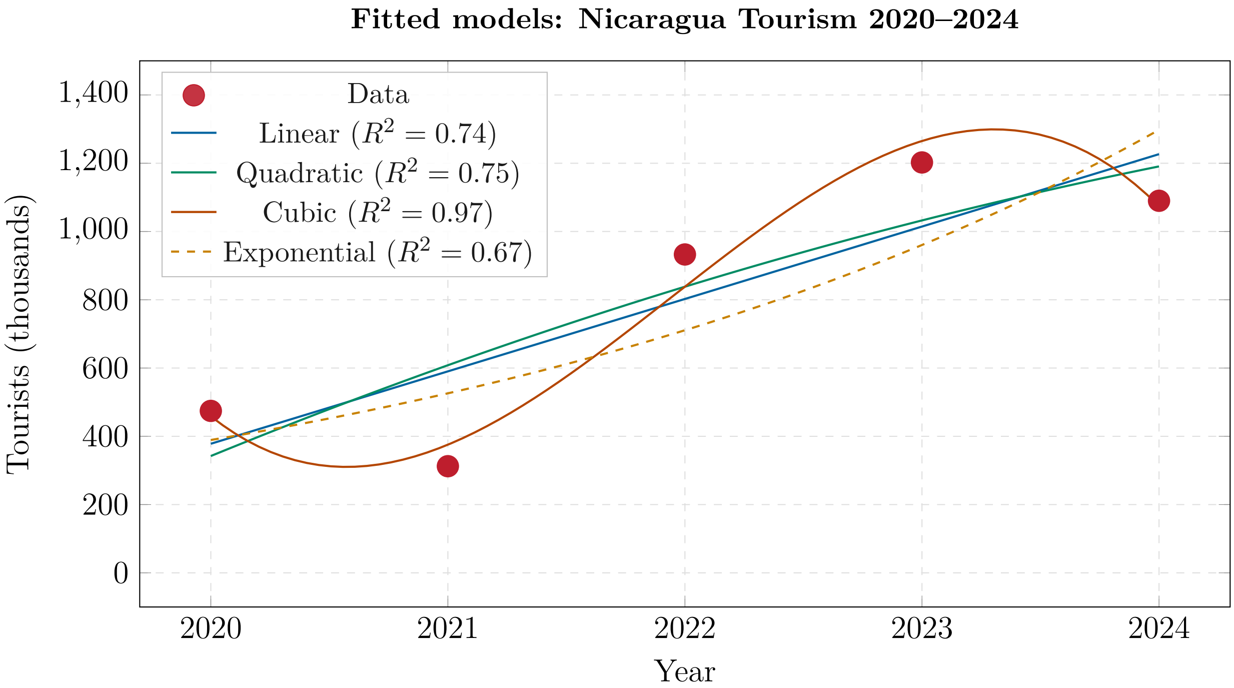

Figure 1: Candidate models. The cubic interpolates almost perfectly, but at a severe predictive cost (section 5).

5 Evaluation Metrics

Table 4: Multi-criteria system

Model	R^2	R^2_{adj}	AIC	AIC$_c$	BIC	PRESS
Linear	0.741	0.654	55.8	**60.5**	55.0	$\mathbf{4.31\times10^5}$
Quadratic	0.748	0.497	57.6	81.6	56.5	2.44×10^6
Cubic	**0.971**	0.886	**48.8**	$+\infty$	**47.2**	2.61×10^6
Exponential	0.665	0.553	57.1	61.8	56.3	—

AIC$_c$ for cubic: $+\infty$ ($n-p-1=0$). PRESS for exponential not comparable (log-linear fit). **Bold** = best per column.

Uncorrected AIC/BIC favor the cubic model. AIC$_c$ reverses the ranking: the cubic diverges, the linear model wins. PRESS confirms this with a factor of 6 difference.

5.1 PRESS decomposition: cubic model

Table 5: PRESS-LOOCV by observation (cubic)

x	Year	y_i	$\hat{y}_{(i)}$	Error	%
0	2020	474	−627	+1 101	**62%**
1	2021	312	571	−258	17%
2	2022	933	960	−28	0%
3	2023	1 202	1 354	−151	6%
4	2024	1 090	1 360	−270	15%
Total PRESS					2.61×10^6

Point $x=0$ accounts for 62% of the error. LOOCV prediction: −627 thousand (negative tourism), a physical absurdity invisible to R^2.

6 Residual Analysis

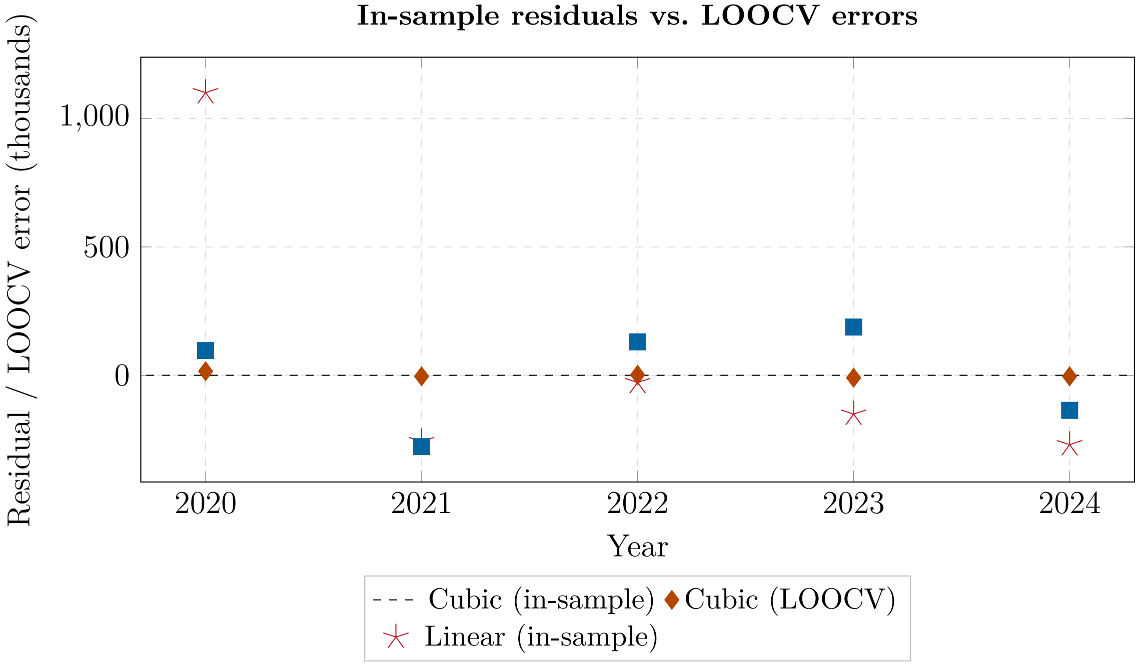

Figure 2: Cubic residuals are nearly zero in-sample, yet the LOOCV error at $x = 0$ reaches +1 101. The linear model is stable.

6.1 Hat matrix

Table 6: Leverage of the cubic model

i	x_i	h_{ii}	$(1-h_{ii})^{-1}$
1	0	0.952	21.3
2	1	0.619	2.6
3	2	0.486	1.9
4	3	0.619	2.6
5	4	0.952	21.3

Cook's distance at $x = 0$: $D_0 \approx 2.4$. Excluding this observation completely alters the predictions.

7 Projections 2025–2030

Selected linear model: $\hat{T}(x) = 378.163 + 212.103\,x$. 95% intervals via parametric bootstrap (10 000 replicates).

Table 7: Projections with 95% bootstrap CI

Year	x	$\hat{T}$	Lower	Upper	Width
2025	5	1 439	689	2 214	106%
2026	6	1 651	728	2 615	114%
2027	7	1 863	748	3 028	122%
2028	8	2 075	753	3 456	130%
2029	9	2 287	747	3 895	138%
2030	10	2 499	730	4 342	145%

Width = (upper−lower)/$\hat{T}$. The cubic model predicts −4 600 thousand for 2030.

1. Extrapolation 6 years beyond the range; width reaches 145%.
2. Assumes constant growth ($\approx$ 212 thousand/year) without macroeconomic covariates.
3. The cubic model predicts negative tourism for 2030: total uselessness.

8 Discussion

8.1 The illusion of a high R^2

An $R^2 = 0.971$ with $n-p-1 = 0$ does not indicate predictive quality: a polynomial of degree $n-1$ exactly interpolates n points (Lagrange) [2]. The cubic model is one degree away from that perfect interpolation.

8.2 Parsimony and interpretability

AIC$_c$ and PRESS converge on the linear model. Its parameters have economic meaning; the cubic yields absurd predictions: a case of the overfit-model trap [12].

8.3 KULAC ecosystem

- **Borge, Salgado & Zelaya [3]:** pharmacokinetic modeling with finite differences.
- **Cruz Treminio & Hernández [5]:** quadratic cost modeling ($R^2 = 0.913$, $n = 7$).
- **Santo [13]:** `algebra.py` for pharmacokinetics, extended here.

9 Limitations and Future Work

1. $n = 5$: results are illustrative; prospective validation with 2025 data is necessary.
2. Exponential model excluded from LOOCV (log-linear fit); resolve via Levenberg–Marquardt.
3. Possible autocorrelation cannot be estimated with 5 data points.
4. Omitted variables (GDP, air connectivity).
5. Extensions: validation with 2025 data, ARIMA when $n \geq 15$, formal Ridge analysis.

10 Pedagogical Implications

L1. R^2 does not measure predictive ability (inadequate when $p/n \geq 0.4$).

L2. LOOCV is the standard for $n < 20$.

L3. Each additional parameter has a cost (factor 6 in this case).

L4. AIC_c is mandatory when $n/p < 40$.

L5. Interpretability is a scientific property, not only a pedagogical one.

11 Conclusions

1. **Theoretical:** R^2 bias of order $O((1-\rho^2)p/(n-p-1))$, divergent.
2. **Computational:** PRESS in $O(n^3)$ without refitting n times.
3. **Empirical:** Cubic $R^2 = 0.97$, PRESS$= 2.6\times10^6$. Linear $R^2 = 0.74$, PRESS$= 4.3\times10^5$ (factor 6).
4. **Diagnostic:** 62% of error at $x = 0$ ($h_{00} \approx 1$, prediction -627 thousand).
5. **Predictive:** Linear $\rightarrow$ 2 500 (2030); cubic $\rightarrow$ $-4\,600$.
6. **Pedagogical:** Analysis with $n < 20$ without generalization metrics is incomplete.

Acknowledgments

The authors sincerely thank Laura Cruz Treminio, José Santo, Kevin Borge, Reymar Salgado, and Junior Zelaya, whose enthusiastic participation in College Algebra (Spring 2026, KULAC) provided the foundational classroom environment for this investigation.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

[2] Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, **27**(1), 17–21. https://doi.org/10.1080/00031305.1973.10478966

[3] Borge, K., Salgado, R., & Zelaya, J. (2026). Therapeutic window adjustment through pharmacokinetic modeling. College Algebra Project, Keiser University Latin American Campus.

[4] Burden, R. L., & Faires, J. D. (2011). *Numerical Analysis* (9th ed.). Brooks/Cole.

[5] Cruz Treminio, L., & Hernández Gómez, F. J. (2026). *Interpretable Quadratic Cost Functions for Data-Driven Manufacturing: A Finite-Difference, Calculus-Based Framework Integrating Economic Theory and Explainable Business Analytics* (Version 4) [Preprint]. Zenodo. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6902899

[6] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://arxiv.org/abs/1702.08608

[7] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.

[8] Hernández Gómez, F. J. (2026). *Study Guide: Pharmacokinetic Modeling* [Instructional material]. Keiser University Latin American Campus.

[9] Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

[10] Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, **16**(3), 31–57. https://doi.org/10.1145/3236386.3241340

[11] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6th ed.). Wiley.

[12] Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1**(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[13] Hernández Gómez, F. J. (2026). *Comparative Rational Pharmacokinetic Modeling for Therapeutic Window Adjustment: A Computational-Analytical Framework Integrating College Algebra, Applied Calculus, and Model Selection Theory* (Version 4) [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.20632076

[14] Silwany García, S. S. (2026). *International Tourist Arrivals in Nicaragua (2020–2024)* [Data set]. Keiser University Latin American Campus. https://doi.org/10.5281/zenodo.21383993

[15] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, **36**(2), 111–147.

[16] Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. *Soviet Mathematics Doklady*, **4**, 1035–1038.

[17] Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2

A Python Pipeline

Listing 1: Pipeline: fit, LOOCV, AIC_c, bootstrap

```python
import numpy as np
from sklearn.model_selection import LeaveOneOut

# Data
x = np.array([0, 1, 2, 3, 4], dtype=float)
y = np.array([474.421, 312.425, 932.700, 1202.300, 1090.000])
n = len(y); ybar = np.mean(y); TSS = np.sum((y - ybar)**2)

def analyze_poly(degree, label):
    p = degree + 1
    coef = np.polyfit(x, y, degree)
    y_hat = np.polyval(coef, x)
    RSS = np.sum((y - y_hat)**2)
    R2 = 1 - RSS/TSS
    dof = n - p - 1
    R2a = 1 - (n-1)/dof*(1-R2) if dof > 0 else float('nan')
    AIC = n*np.log(RSS/n) + 2*p
    BIC = n*np.log(RSS/n) + p*np.log(n)
    AICc = AIC + 2*p*(p+1)/dof if dof > 0 else float('inf')

    loo = LeaveOneOut()
    press = 0.0; preds = []
    for tr, te in loo.split(x):
        c = np.polyfit(x[tr], y[tr], degree)
        yp = np.polyval(c, x[te])[0]
        press += (y[te[0]] - yp)**2
        preds.append(yp)

    print(f"\n{'='*50}")
    print(f"  {label} (degree {degree}, p={p})")
    print(f"  R2={R2:.4f}  R2adj={R2a:.4f}")
    print(f"  AIC={AIC:.2f}  AICc={AICc:.2f}  BIC={BIC:.2f}")
    print(f"  PRESS={press:.2e}")
    print(f"  LOOCV predictions: {[round(v,1) for v in preds]}")
    return coef, press

for deg, lbl in [(1,"Linear"),(2,"Quadratic"),(3,"Cubic")]:
    analyze_poly(deg, lbl)

# Exponential
ce = np.polyfit(x, np.log(y), 1)
A, k = np.exp(ce[1]), ce[0]
```

```
ye = A*np.exp(k*x)
print(f"\nExponential: A={A:.1f}, k={k:.4f}, R2={1-np.sum((y-ye)**2)/TSS:.4f}")

# Bootstrap (linear model)
np.random.seed(42)
cl = np.polyfit(x, y, 1); b1, b0 = cl
s2 = np.sum((y - np.polyval(cl, x))**2)/(n-2)
xf = np.arange(5, 11, dtype=float)
B = 10000; bp = np.zeros((len(xf), B))
for b in range(B):
    yb = np.random.normal(b0 + b1*x, np.sqrt(s2))
    cb = np.polyfit(x, yb, 1)
    bp[:, b] = np.polyval(cb, xf)

print("\nProjections 2025-2030 (95% CI):")
for i, (xi, yr) in enumerate(zip(xf, range(2025, 2031))):
    yp = b0 + b1*xi
    lo, hi = np.percentile(bp[i], [2.5, 97.5])
    print(f"  {yr}: {yp:.0f} [{lo:.0f}, {hi:.0f}]")
```

B Ridge Regression

Listing 2: Ridge with λ selected by LOOCV

```
from sklearn.linear_model import RidgeCV
from sklearn.preprocessing import PolynomialFeatures

poly = PolynomialFeatures(degree=3, include_bias=True)
X3 = poly.fit_transform(x.reshape(-1,1))
lambdas = np.logspace(-2, 4, 100)
ridge = RidgeCV(alphas=lambdas, cv=LeaveOneOut(),
                scoring='neg_mean_squared_error')
ridge.fit(X3, y)
yr = ridge.predict(X3)
print(f"Optimal lambda = {ridge.alpha_:.2f}, R2 = {1-np.sum((y-yr)**2)/TSS:.4f}")
```