OptiMod

SCIENTIFIC REPORT

What is associated with daily bike demand?

Real observations. Explicit model choices.
Reproducible statistical evidence.

UCI BIKE SHARING

Capital Bikeshare, Washington, DC
Daily recorded rentals, 1 January 2011 to 31 December 2012

731	547 / 184	95%
observed days	training / holdout	HAC confidence intervals

Question. Which weather and calendar factors are associated with daily recorded rentals after adjustment for time and seasonality?

Key finding. A negative-binomial model handles extra count variation, but Poisson achieves lower held-out point errors in this split. Temperature, humidity and weather categories show conditional associations; serial dependence remains.

Illustrative product report · independently computed reference analysis

Prepared 2 November 2026. Public data and independent statsmodels calculations.

01 / DATA AND INTEGRITY

Start with the observations

The original daily CSV is preserved. Each row represents one calendar day, not a rider or a station.

Scope	Observed data
Source and license	UCI Bike Sharing, day.csv. Hadi Fanaee-T (2013). CC BY 4.0.
Target	cnt: total recorded rental trips per day; casual + registered.
Coverage	731 consecutive dates; 16 original columns; no missing calendar day.
Demand distribution	Mean 4,504.3; median 4,548; range 22 to 8,714 rentals/day.
Total recorded rentals	3,292,679

Checks against the actual file

Check	Finding	Treatment
Blank cells / duplicate rows / duplicate dates	0 / 0 / 0	No deletion or imputation
Target components	casual + registered = cnt on 731/731 rows	Exclude both predictors
Temperature overlap	r(temp, atemp) = 0.9917	Retain temp only
Recorded humidity = 0	1 day: 10 March 2011	Keep, flag, sensitivity refit
Weather support	Category 3: 21 days; category 4: 0	Show uncertainty; no category-4 effect

Meaning of a clean file. Zero blank cells does not establish measurement quality. The zero-humidity value is suspicious but not proven missing. Recorded rentals may also be constrained by available bikes or service interruptions.

Source: original UCI day.csv, downloaded 2 November 2026. Calendar, weekday and count-component consistency were checked directly.

02 / PREPARATION AND DESIGN

Make every transformation explicit

Primary analysis: all 731 rows retained. No imputation, winsorization or outlier removal.

Variable group	Decision and reason
casual, registered	Exclude: exact components of cnt would reveal the target.
instant, atemp	Exclude the record ID and near-redundant feeling-temperature proxy.
dteday	Parse YYYY-MM-DD; sort; derive month and days since 2011-01-01 / 365.25.
yr, mnth, season, workingday	Replace duplicate calendar fields with date trend/month and weekday + holiday.
temp	Retain normalized scale. Set temp10 = (temp - 0.50) / 0.10 and add temp10 squared.
hum, windspeed	Scale each by 1 / 0.10. A humidity unit means +10 percentage points.
weekday, weathersit	Categorical factors. Reference levels: Sunday and clear/few clouds.

Metadata conflict. The archive README and current UCI page describe temperature normalization and season names differently. This report does not infer Celsius values. Month comes from the parsed date; wind also stays in normalized units.

Chronological evaluation, fixed before fitting



Training: 1 January 2011 - 30 June 2012.
Holdout: 1 July - 31 December 2012. No random split and no test-set tuning.

The conditional model uses seven direct source fields, expanded to 26 mean-model parameters. Same-day weather is observed, not known in advance.

03 / MODEL COMPARISON

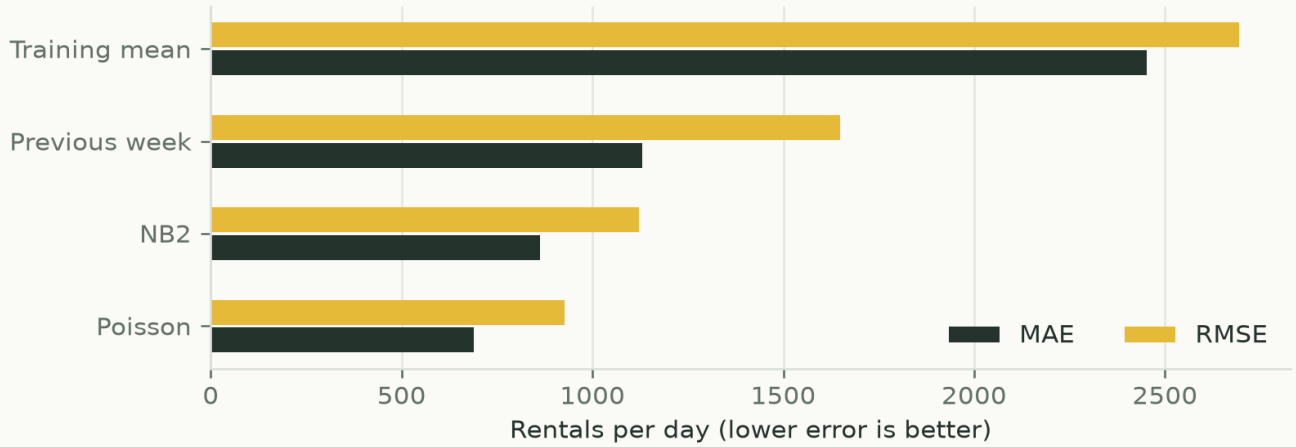
Variance fit and point error differ

Both candidate models use the same mean equation and the same 547 training days.

Training statistic	Poisson	Negative binomial (NB2)
Working-likelihood AIC	66,197.63	8,815.65
Pearson dispersion	109.200	0.894
Mean parameters / total parameters	26 / 26	26 / 27
Extra-variance parameter alpha	0 (fixed)	0.04119

Association-model choice. Training residual variation greatly exceeds the Poisson variance assumption. NB2 estimates Var(Y | X) = mu + alpha × mu² and has lower training AIC. Serial dependence limits a literal independent-observation interpretation of that likelihood.

Conditional holdout error: 184 later days



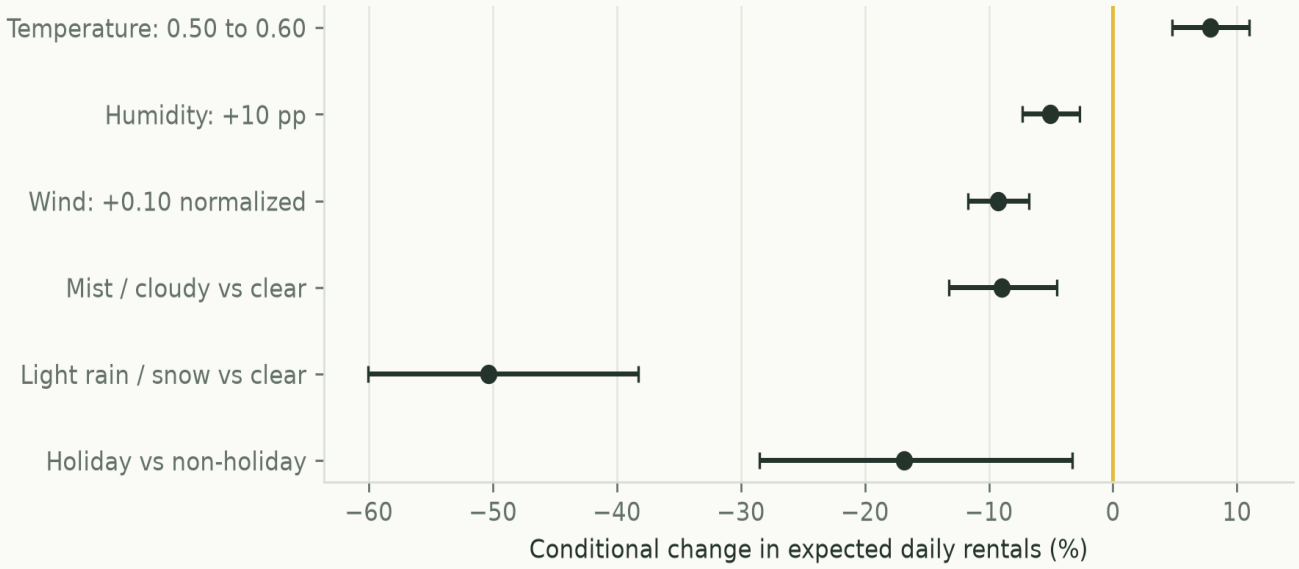
Model	MAE	RMSE	Mean predicted - actual
Poisson	688.6	927.9	+248.0
NB2	863.4	1,122.8	+648.1

Poisson has the lower holdout MAE and RMSE. NB2 is not presented as the universal predictive winner. The previous-week benchmark is rolling: it uses the actual count seven days earlier, including earlier holdout days.

04 / CONDITIONAL ASSOCIATIONS

Effects with stated comparisons

Separate association fit on all 731 days. Approximate 95% pointwise HAC intervals, seven daily lags.



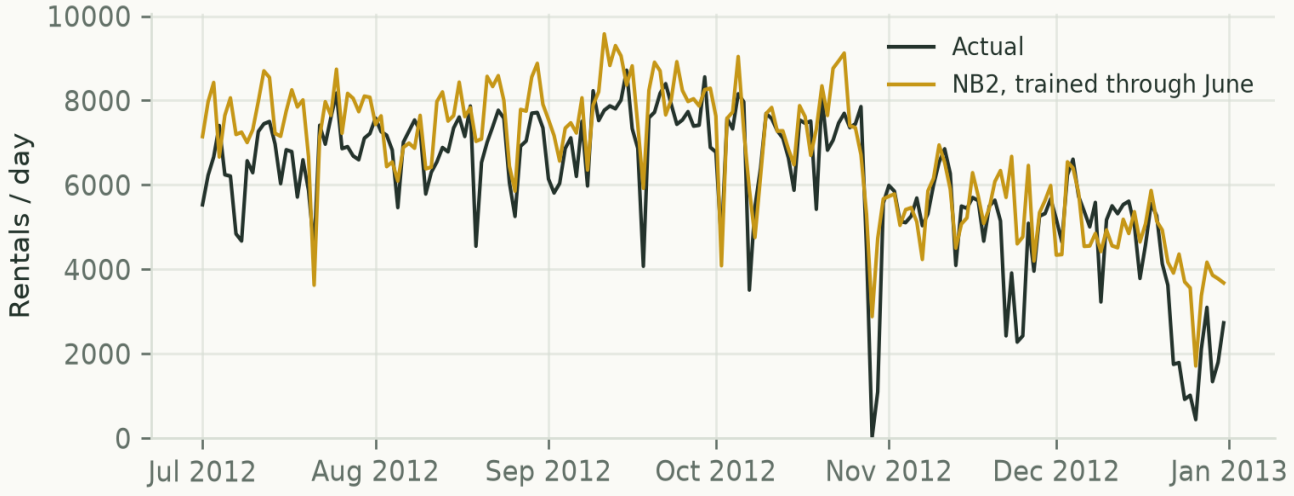
Comparison	Change	95% confidence interval
Temperature 0.50 to 0.60 normalized	+7.8%	+4.8% to +11.0%
Humidity +10 percentage points	-5.1%	-7.3% to -2.7%
Wind speed +0.10 normalized	-9.3%	-11.7% to -6.8%
Mist / cloudy weather vs clear / few clouds	-9.0%	-13.2% to -4.5%
Light rain / snow vs clear / few clouds	-50.4%	-60.1% to -38.3%
Holiday Holiday vs non-holiday	-16.8%	-28.5% to -3.3%

Interpretation. Each comparison holds the remaining covariates fixed. Temperature is nonlinear: the displayed 0.50-to-0.60 contrast uses both temperature terms and their covariance. These are exploratory associations, with no causal or multiplicity-adjusted interpretation.

05 / RESIDUAL DIAGNOSTICS

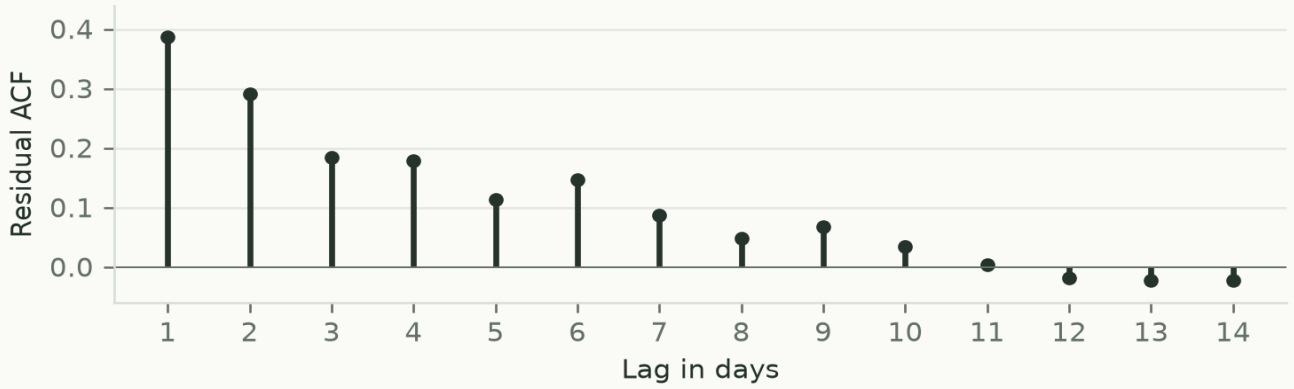
Dependence remains visible

The holdout series and full-model residual checks reveal limitations that fit statistics alone can hide.



Actual daily rentals and NB2 conditional predictions for July-December 2012. Predictions use the model trained through June and the recorded weather on each day.

Autocorrelation of full-model Pearson residuals



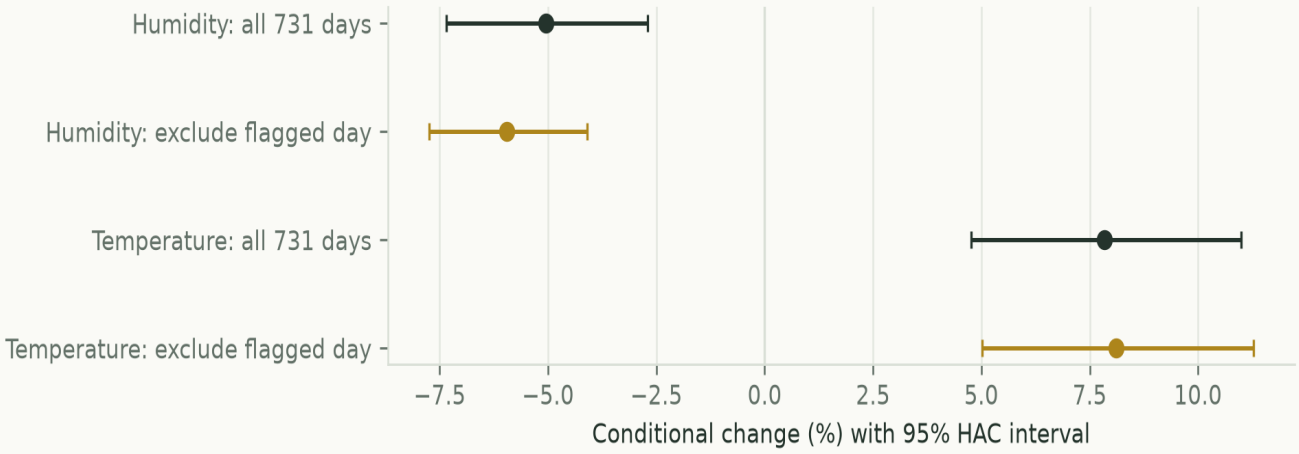
Lag-1 residual correlation: 0.387. Full-model Pearson dispersion is 0.754. Positive short-lag correlations remain. The Ljung-Box lag-7 diagnostic gives $p < 0.001$; this is a descriptive dependence check, not a definitive count-model specification test.

What HAC changes. A sandwich covariance estimate with seven lags adjusts coefficient uncertainty for short-run dependence and heteroskedasticity. It does not remove autocorrelation, correct the conditional mean or create operational forecast validity.

06 / SENSITIVITY AND INTERPRETATION

Keep the anomaly, test its influence

The primary result retains the zero-humidity record on 10 March 2011. A separate refit excludes only that day.



Conditional comparison	All 731 days	Without flagged day (730)
Humidity: +10 percentage points	-5.05%	-5.94%
Temperature: 0.50 to 0.60	+7.84%	+8.10%

The humidity estimate changes, while its direction remains negative in this specification. The sensitivity fit is not substituted for the primary analysis. With one date removed, its HAC lags refer to ordered observations; the primary series has a complete daily grid.

Boundaries for a mobility decision

Recorded trips are not fleet capacity. No station inventory, stock-outs, prices, closures or operating capacity is modelled.

Observed weather is not a forecast input. These results cannot establish advance demand forecasting performance.

Transfer is untested. One system, two historical years; no claim about another city or current operations.

Rare weather is weakly supported. Only 21 category-3 days and no category-4 observations; avoid severe-weather extrapolation.

Association is not intervention. Time trend and calendar controls do not eliminate confounding or guarantee causality.

07 / STATISTICAL SPECIFICATION

The fitted model, in full

NB2 log link; maximum likelihood with estimated alpha. Full-data association model: 731 observations.

log(mu) = intercept + trend + month + weekday + holiday

+ temp10 + temp10² + hum10 + wind10 + weather

Var(Y | X) = mu + alpha × mu²

Term	Estimate	HAC SE	95% lower	95% upper
Intercept	8.4773	0.1467	8.1898	8.7648
Month 2 vs January	-0.0152	0.0640	-0.1407	0.1103
Month 3 vs January	0.0448	0.0718	-0.0960	0.1856
Month 4 vs January	0.0907	0.0786	-0.0634	0.2447
Month 5 vs January	0.1513	0.0861	-0.0174	0.3200
Month 6 vs January	0.1138	0.0876	-0.0579	0.2855
Month 7 vs January	0.0932	0.0900	-0.0832	0.2695
Month 8 vs January	0.0317	0.0868	-0.1385	0.2018
Month 9 vs January	0.0649	0.0785	-0.0889	0.2188
Month 10 vs January	0.0285	0.0761	-0.1206	0.1777
Month 11 vs January	-0.0201	0.0696	-0.1564	0.1162
Month 12 vs January	-0.1651	0.1131	-0.3869	0.0566
Monday vs Sunday	0.0675	0.0281	0.0123	0.1226
Tuesday vs Sunday	0.0839	0.0254	0.0341	0.1336
Wednesday vs Sunday	0.0750	0.0266	0.0229	0.1272
Thursday vs Sunday	0.0951	0.0256	0.0450	0.1453
Friday vs Sunday	0.1032	0.0250	0.0543	0.1522
Saturday vs Sunday	0.0842	0.0233	0.0385	0.1299
Weather 2 vs category 1	-0.0940	0.0244	-0.1419	-0.0461
Weather 3 vs category 1	-0.7005	0.1112	-0.9185	-0.4826
trend_years	0.4514	0.0298	0.3931	0.5098
holiday	-0.1845	0.0770	-0.3353	-0.0336
temp10	0.1206	0.0152	0.0908	0.1504
temp10_sq	-0.0451	0.0054	-0.0557	-0.0345
hum10	-0.0518	0.0124	-0.0762	-0.0274
wind10	-0.0973	0.0139	-0.1246	-0.0700
alpha	0.0535	0.0121	0.0297	0.0773

Coefficients are on the log-mean scale. Alpha is the NB2 dispersion parameter, not a rate ratio. Reference levels: January, Sunday and weather category 1. temp10 = (temp - 0.50) / 0.10; hum10 = hum / 0.10; wind10 = windspeed / 0.10.

Covariance: HAC sandwich with Bartlett weights, maximum lag 7 and finite-sample correction. Intervals use estimate ± 1.96 × standard error. The temperature contrast uses joint covariance. NB2 optimizer: BFGS. Complete 27 × 27 matrix: coefficient-covariance.csv.

08 / REPRODUCIBILITY AND PROVENANCE

From original data to every result

The accompanying archive contains the original daily observations and the exact analysis specification.

Reproduce the analysis. Extract the archive into a fresh Python 3.12 environment. Run the commands below from the extracted folder. The script checks the input SHA-256, fits the models and writes analysis.json and report.txt.

```
python -m pip install -r requirements.txt
python analyse.py
```

Input fingerprint

SHA-256 / day.csv

`a6bcf826782d3c0fbfdcbeead17cd0884185a0dafe8ff10cd48a874ee7ba18be`

Recorded software versions

Component	Version
python	3.12.14
numpy	2.5.4
pandas	3.1.0
scipy	1.18.1
statsmodels	0.14.6

Package contents. day.csv; original source-readme.txt; source.txt; README.txt; analyse.py (also supplied as analysis.py); requirements.txt; analysis.json; coefficient-covariance.csv; report.txt; this PDF. No vendor libraries are bundled. Floating-point differences may occur across platforms.

Sources and license

Dataset. Fanaee-T, H. (2013). Bike Sharing. UCI Machine Learning Repository. doi.org/10.24432/C5W894. Licensed CC BY 4.0. Original archive downloaded 2 November 2026.

Associated paper. Fanaee-T, H., and Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. doi.org/10.1007/s13748-013-0040-3.

Methods. statsmodels NegativeBinomial and HAC covariance documentation. Full URLs are provided in source.txt.

Data credit: Hadi Fanaee-T / UCI. Analysis and presentation added for this reference workflow. Calculations use statsmodels, not an OptiMod production engine. No endorsement by UCI or Capital Bikeshare is implied.

09 / STATISTICAL TEST RECORD

Statistical test record

Exact detail from the existing calculation. No refitting or additional tests were performed.

Residual dependence: Ljung-Box reference tests

Null hypothesis: Pearson residual autocorrelations through lag h are jointly zero. Full-data NB2 association fit, n = 731.

Lag h	Q statistic	df	Reference p-value	Finding
7	251.91044086566615	7	1.0862846732672167e-50	Dependence remains
14	258.99118696116716	14	3.9574122927910813e-47	Dependence remains

The calculation used **model_df = 0**, so reference df equals h (7 or 14). These very small p-values indicate remaining dependence. For fitted count-model residuals, they are descriptive chi-square reference checks, not definitive specification tests. No p-value pass threshold was encoded.

Parameter inference: two HAC Wald-test examples

Null hypothesis: the stated log-mean contrast equals zero (count ratio = 1). Two-sided, asymptotic standard-normal reference; df not applicable. These test parameters, not model assumptions.

Contrast, full fit n = 731	Wald z	p-value
Temperature: normalized 0.50 to 0.60	5.121842683926757	3.025642713833995e-07
Humidity: +10 percentage points	-4.162821533364263	3.1433904958171295e-05

Covariance: HAC with seven lags and finite-sample correction. Temperature uses joint covariance of both temperature terms. Intervals and p-values are exploratory and pointwise, without multiple-testing adjustment.

Diagnostics without a computed p-value

Diagnostic	Value	Definition / interpretation
Design-matrix rank	26 of 26 mean columns	Full rank; not a low-collinearity guarantee
Poisson dispersion, training	109.200167973	Pearson sum of squares / 521
NB2 dispersion, training	0.893879934	Pearson sum of squares / 521
NB2 dispersion, full fit	0.754477367	Pearson sum of squares / 705

No global acceptance guarantee. Dispersion has no formal test p-value or calibrated pass band here. HAC changes estimated uncertainty; it does not fix the mean model or remove dependence. Exact source checks, criteria and all six selected Wald tests are recorded in **test-details.json**.