Quantitative Research Methods

GLMs and Splines

Advanced module A4 — optional self-study

Prof. Dr. Christoph Weisser

HSBI

Summer Semester 2026

Where this module fits

This is an **optional advanced module**: nothing in it is required for the exam, and everything in it deepens material you have already met.

	Course chapter	What this module adds
	Ch. 3 — Linear regression	. . . becomes the *Gaussian corner* of one big family
	Ch. 4 — Classification	. . . logistic and Poisson regression re-derived as GLMs
	Ch. 5 — Resampling	. . . CV and GCV choose the spline penalty λ
	Ch. 6 — Model selection	. . . deviance, likelihood-ratio tests and AIC for GLMs
▶	**Ch. 7 — Splines and GAMs**	**. . . penalized splines, and GAMs for count data**

One sentence

Chapter 4's likelihood view and Chapter 7's smooth curves are two halves of the same machine — this module assembles it: **exponential family** → **GLM** → **penalized spline** → **GAM for counts**, all on one real dataset (`Bikeshare`).

Contents

Optional and advanced material is collected in the **appendix** at the end of the deck.

Notation in this module

Symbol	Meaning
Y_i, y_i	response (here: hourly count of bikers); its observed value
$\mu_i = E[Y_i]$	mean of the response for observation i
θ, ϕ	canonical (natural) parameter; dispersion parameter
$b(\theta)$, $a(\phi)$, $c(y, \phi)$	the three building blocks of an exponential-family density
$V(\mu)$	variance function: $\mathrm{Var}(Y) = a(\phi)\, V(\mu)$
$\eta_i = x_i^\top \beta$	linear predictor
$g(\cdot)$	link function, $g(\mu_i) = \eta_i$; canonical link $g = (b')^{-1}$
D, D_0	residual deviance; null deviance (intercept-only model)
$\hat{\varphi}$	Pearson estimate of the dispersion (quasi-Poisson)
α	negative-binomial overdispersion: $\mathrm{Var} = \mu + \alpha\mu^2$
$B_j(x)$, β_j	j-th B-spline basis function; its coefficient
λ, S	smoothing penalty weight; penalty matrix, $\int f''(t)^2 dt = \beta^\top S \beta$
edf	effective degrees of freedom, $\mathrm{tr}\{H(\lambda)\}$
$f_j(\cdot)$	smooth partial effect of predictor j in a GAM

Sources and references

Primary textbook

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning, with Applications in Python.* Springer. — Sections 4.6 (Poisson regression) and Chapter 7 (splines, GAMs).

Going deeper (optional)

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall.
Wood, S. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed., CRC Press.
Eilers, P. & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2).

Learning objectives

By the end of this module you will be able to:

- **Explain** the exponential family and derive a member's mean and variance from $b(\theta)$.
- **Specify** a GLM — random component, linear predictor, link — for a given data type.
- **Interpret** Poisson coefficients as rate ratios, and **use** offsets for exposure.
- **Test** nested GLMs with deviance and likelihood-ratio tests, and **compare** models with AIC.
- **Diagnose** overdispersion and **choose** between quasi-Poisson and negative binomial.
- **Tune** a penalized spline via GCV or CV and **read** its effective degrees of freedom.
- **Fit** and **interpret** a penalized-spline GAM for counts with `statsmodels`.

Roadmap of this module

From a broken linear model to a GAM for counts

1. Why the linear model fails on `Bikeshare` counts.
2. The exponential family: one density form behind Chs. 3 and 4.
3. GLM anatomy: random component + linear predictor + link; Poisson regression, offsets.
4. Inference: deviance, likelihood-ratio tests, AIC.
5. Overdispersion: quasi-Poisson and the negative binomial.
6. Splines deepened: B-spline bases, roughness penalties, effective df, GCV.
7. The meeting point: penalized-spline GAMs for counts.

Where this module is used in industry

Sector	Concrete application	Which decision it drives
Insurance	Claim *frequency* as a Poisson/NB GLM with exposure offsets; age curves as GAM smooths	The premium each tariff cell pays
Retail, e-commerce	Daily demand counts per SKU and store, weather and calendar smooths	Ordering, staffing and markdowns
Energy	Load forecasting with GAMs: smooth effects of temperature and time of day	How much generation to schedule
Public health	Case and admission counts vs exposure population	Staffing, stockpiles, early warnings
Banking	Default (yes/no) as a logistic GLM — the same machinery	Approve or decline; capital held
Mobility platforms	Hourly ride counts vs hour, weather, events (this module's dataset)	Fleet rebalancing and surge staffing

In industry — The common thread

Whenever the response is a **count** or a **rate**, industry reaches for exactly this stack: GLM for interpretability and valid uncertainty, smooth terms for the shapes no one wants to guess, offsets for unequal exposure.

Industry case in depth: pricing motor insurance

In industry — Claim frequency at a motor insurer

A pricing team models the **number of claims** per policy-year for millions of policies. Policies are observed for unequal durations — three months here, a full year there — so the model is a Poisson GLM for the claim *rate* with log(exposure) as an **offset**. Tariff factors (region, vehicle class, bonus–malus) enter as dummies; **driver age** enters as a penalized-spline smooth, because its effect is famously U-shaped and no polynomial degree is defensible to a regulator.

Why the family matters commercially

The premium is roughly frequency $\times$ severity. If the frequency model ignores **overdispersion**, its standard errors are far too small: tariff cells look "significantly" different when they are not, and the tariff fragments into noise. Actuarial practice therefore fits quasi-Poisson or negative binomial by default.

Takeaway

Every ingredient of this module — log link, rate ratios, offsets, overdispersion, spline smooths, edf — appears in one production pricing model.

Outline

The `Bikeshare` data: 8 645 hours of counts

Hourly bike rentals in Washington DC (ISLP): response `bikers`, the number of riders in that hour.

Variable	Type	Range / levels	Role
`bikers`	count	1 to 651	response
`hr`	hour of day	0–23	23 dummies (or a smooth
`temp`	normalised temperature	0.02–1 ($1 \approx 41^\circ$C)	numeric
`workingday`	binary	0/1	dummy
`weathersit`	categorical	clear / cloudy-misty / light rain / heavy rain	3 dummies

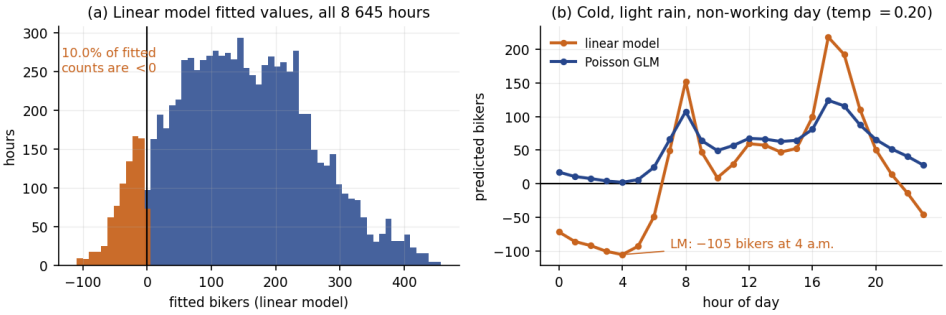
The response at a glance

$n = 8\,645$ hours; mean $\bar{y} = 143.8$ bikers, variance $s^2 = 17\,902$ — the variance is about $124\times$ the mean. Weather levels are very unbalanced: clear 5 645, cloudy/misty 2 218, light rain/snow 781, heavy rain/snow **a single hour**.

Takeaway

A count response lives on $\{0, 1, 2, \dots\}$ and its spread grows with its level. Keep both facts in mind — the linear model respects neither.

Fit the Chapter-3 linear model . . . and watch it fail



Takeaway

OLS on `bikers` ∼ `hr + temp + workingday + weathersit`: **10.0%** of fitted values (861 of 8 645 hours) are *negative*; on a cold rainy night it predicts −105 riders at 4 a.m. — the Poisson GLM (right) predicts 2.2: small, positive, sensible.

Two failures, one diagnosis

What exactly is wrong with the Gaussian linear model here?

Model under indictment: $Y = x^\top \beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

1. **Wrong support.** A Gaussian mean can be negative; a count cannot. The 10.0% negative predictions are the model doing what it was told.
2. **Wrong variance.** Constant σ^2 — but quiet 4 a.m. hours vary by a few riders, busy 5 p.m. hours by hundreds: variance *grows with the mean*.

What we need instead

A model whose **mean respects the data type**, whose **variance is tied to the mean**, and which keeps a **linear predictor** $x^\top \beta$ we know how to estimate and interpret. That is precisely a *generalized linear model*.

In industry — Mobility platforms: why negative forecasts hurt

A fleet-rebalancing optimiser fed "-105 rides expected" will happily truck bikes *away* from a station that will see riders at dawn.

Exercise A4.1 — Choosing the random component [Concept]

Exercise

1. Name the **two** distinct failures of the Gaussian linear model on hourly bike counts, and say which *observable symptom* each produces.
2. For each response below, propose a distribution and a link that respect the data type: (a) number of insurance claims per policy-year; (b) whether a loan defaults; (c) a person's log-wage.
3. A colleague proposes fixing the negative predictions by modelling $\log(y + 1)$ with OLS. Give one reason this is *not* equivalent to a Poisson GLM.

Solution — Exercise A4.1

Solution

Method. Match the *support* and the *mean–variance relation* of the response to a distribution; the link keeps the mean inside the support.

Part 1. (i) *Wrong support*: fitted means can be negative — symptom: 10.0% negative predictions on `Bikeshare`. (ii) *Wrong variance*: constant σ^2 — symptom: residual spread visibly grows with the fitted level (a funnel in the residual plot).

Part 2. (a) counts $\Rightarrow$ Poisson with log link (offset for exposure); (b) binary $\Rightarrow$ Bernoulli with logit link — exactly Chapter 4's logistic regression; (c) continuous, roughly symmetric $\Rightarrow$ Gaussian with identity link — Chapter 3.

Part 3. OLS on $\log(y+1)$ models $E[\log(Y+1)]$, not $\log E[Y]$: by Jensen's inequality these differ, so back-transformed predictions are biased for the mean count; the "$+1$" is arbitrary; and the implied variance is again constant on the log scale rather than Poisson-like. The GLM models $\log E[Y]$ directly.

Common mistake

"Transform y, then use OLS" changes *what is being estimated*. A GLM transforms the **mean**, not the **data**.

Outline

One density to rule them all

Definition (exponential dispersion family)

A response Y belongs to the exponential family if its density (or pmf) can be written

$$f(y;\theta,\phi) \;=\; \exp\!\left\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}.$$

How to read this

- θ — the **canonical parameter**: the natural scale on which y enters the exponent linearly.
- $b(\theta)$ — the **cumulant function**: its derivatives hand us mean and variance for free.
- $a(\phi)$ — the **dispersion**: a scale factor ($\phi = 1$ for Poisson and Bernoulli).
- $c(y,\phi)$ — normalisation without θ: never touches the estimation of β.

Takeaway

Gaussian (Ch. 3), Bernoulli (Ch. 4) and Poisson (Ch. 4.6) are all of this form — one estimation theory covers all three.

The mean and the variance come from $b(\theta)$

Rule (mean–variance identities)

$$E[Y] = b'(\theta), \qquad \mathrm{Var}(Y) = b''(\theta)\, a(\phi) = a(\phi)\, V(\mu).$$

How to read this

- Differentiate the log-likelihood in θ, take expectations, and both identities fall out of the score equations (full derivation in the **appendix**).
- $V(\mu) = b''(\theta)$ rewritten as a function of μ is the **variance function**: it states *how the spread must move with the mean* — the very thing OLS got wrong on counts.
- $a(\phi)$ scales the whole variance but leaves its *shape* $V(\mu)$ alone.

Takeaway

Choosing a family = choosing a mean–variance relation. That is the modelling decision; everything else is shared machinery.

Worked derivation: the Poisson is a member

Rewrite the Poisson pmf into family form

$$P(Y=y)=\frac{e^{-\mu}\mu^{y}}{y!}=\exp\{\,y\log\mu\;-\;\mu\;-\;\log y!\,\},\qquad y=0,1,2,\ldots$$

Match against $\exp\{(y\theta-b(\theta))/a(\phi)+c(y,\phi)\}$: $\theta=\log\mu$, $b(\theta)=e^{\theta}$, $a(\phi)=1$, $c(y,\phi)=-\log y!$. Cash in the identities: $E[Y]=b'(\theta)=e^{\theta}=\mu$ ✓ and $\mathrm{Var}(Y)=b''(\theta)\cdot 1=\mu$ — the famous **mean = variance** property.

Sanity check with numbers

With $\mu=2$: $P(Y=3)=e^{-2}\,2^{3}/3!=0.1804$; canonical parameter $\theta=\log 2=0.693$ — the *log-mean* scale on which Poisson regression will be linear.

Takeaway

The canonical parameter of the Poisson is $\log\mu$ — the log link is not a convention, it is **built into the distribution**.

The three members you already know

Family	$\theta(\mu)$	$b(\theta)$	$a(\phi)$	$E[Y] = b'(\theta)$	$\mathrm{Var}(Y) = b''(\theta)a(\phi)$
Gaussian	μ	$\theta^2/2$	σ^2	$\theta = \mu$	σ^2
Bernoulli	$\log \frac{\mu}{1-\mu}$	$\log(1 + e^\theta)$	1	$\frac{e^\theta}{1+e^\theta} = \mu$	$\mu(1-\mu)$
Poisson	$\log \mu$	e^θ	1	$e^\theta = \mu$	μ

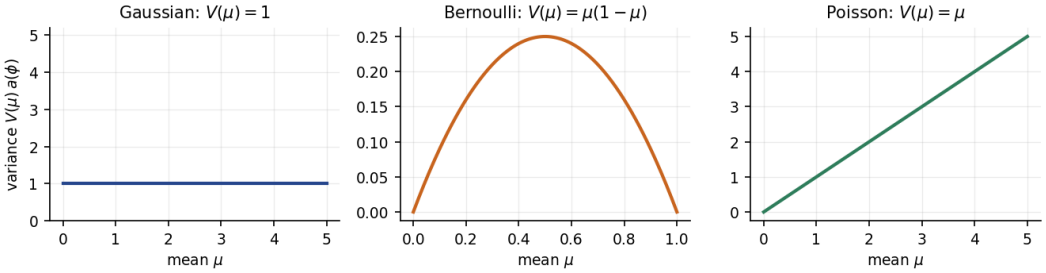
Read each row the same way

Column 2 names the *natural scale* of the mean (identity, log-odds, log); the last column is the **variance function** — constant, quadratic in μ with a maximum at $\mu = \frac{1}{2}$, and proportional to μ, respectively.

Takeaway

The Gaussian derivation is in the appendix; the Bernoulli one is **Exercise A4.2** — next slide.

Each family fixes how spread follows the mean



Takeaway

Gaussian: flat variance, whatever the mean. Bernoulli: variance peaks at 0.25 when $\mu = 0.5$ and vanishes at the ends. Poisson: variance climbs one-for-one with the mean — at $\mu = 5$ the variance *must* be 5. Data that violate the chosen line (as `Bikeshare` will, spectacularly) violate the family.

Exercise A4.2 — The Bernoulli as a family member [Math]

Exercise

Let $Y \in \{0, 1\}$ with $P(Y = 1) = \mu$, so $f(y; \mu) = \mu^y (1 - \mu)^{1-y}$.

1. Rewrite f in exponential-family form and identify $\theta(\mu)$, $b(\theta)$ and $a(\phi)$.
2. Verify $b'(\theta) = \mu$.
3. Verify $b''(\theta) = \mu(1 - \mu)$, and state in one sentence what this variance function implies for observations with μ near 0.5 versus near 0.99.

Solution — Exercise A4.2

Solution

Method. Take logs, isolate the coefficient of y, and match term by term.
Part 1. $\log f = y \log \mu + (1-y)\log(1-\mu) = y \log \frac{\mu}{1-\mu} + \log(1-\mu)$, hence

$$\theta = \log \frac{\mu}{1-\mu} \ \text{(the \textbf{logit})}, \qquad a(\phi) = 1, \qquad c(y, \phi) = 0.$$

Inverting the logit gives $\mu = e^\theta/(1+e^\theta)$, so $\log(1-\mu) = -\log(1+e^\theta)$ and therefore $b(\theta) = \log(1+e^\theta)$.
Part 2. $b'(\theta) = \dfrac{e^\theta}{1+e^\theta} = \mu$ ✓— the sigmoid of Chapter 4.
Part 3. $b''(\theta) = \dfrac{e^\theta}{(1+e^\theta)^2} = \mu(1-\mu)$ ✓. Near $\mu = 0.5$ the variance is maximal (0.25): most informative about β; near $\mu = 0.99$ it is 0.0099 — nearly deterministic, little information.

Common mistake

$b(\theta)$ must be written **as a function of** θ, not of μ — differentiating $b = -\log(1-\mu)$ with respect to θ as if μ were constant gives nonsense.

Outline

The three components of a GLM

Rule (GLM specification)

1. Random component: Y_i exponential-family with mean μ_i. **2. Systematic component:** linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. **3. Link:** $\boxed{g(\mu_i) = \eta_i = x_i^\top \beta}$

How to read this

- The link g maps the mean's natural range onto the whole real line, where a linear predictor can roam freely: $\log : (0, \infty) \to \mathbb{R}$, $\text{logit} : (0, 1) \to \mathbb{R}$.
- The **canonical link** sets $g = (b')^{-1}$, i.e. $\theta_i = \eta_i$: identity for Gaussian, logit for Bernoulli, log for Poisson — concave log-likelihoods, simple score equations.
- Estimation is by **maximum likelihood**, via iteratively reweighted least squares (IRLS — sketch in the appendix).

Takeaway

Chapter 3 = Gaussian + identity. Chapter 4 = Bernoulli + logit. Today's workhorse = Poisson + log. Same anatomy, three organs swapped.

Logistic regression, re-derived in one line

Chapter 4 falls out as a special case

Take the Bernoulli member ($\theta = \log \frac{\mu}{1-\mu}$) and impose the canonical link $\theta_i = \eta_i$:

$$\log \frac{p(x_i)}{1 - p(x_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad \Longleftrightarrow \quad p(x_i) = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}}.$$

This *is* equation (4.6) of ISLP — and "e^{β_j} multiplies the odds" is the Bernoulli twin of the rate ratios two slides ahead.

Takeaway

Nothing about logistic regression was ad hoc: sigmoid, log-odds reading, ML fitting — all forced by "Bernoulli + canonical link".

In industry — Credit scoring: the regulated GLM

Default models in banking are Bernoulli GLMs: every e^{β_j} has a documented odds-ratio meaning — a key reason GLMs survive in the deep-learning era.

Poisson regression: the model for `Bikeshare`

Rule (Poisson GLM with log link)

$$Y_i \sim \text{Poisson}(\mu_i), \qquad \log \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \iff \mu_i = e^{\beta_0}\, e^{\beta_1 x_{i1}} \cdots e^{\beta_p x_{ip}}.$$

The multiplicative reading

- $\mu_i > 0$ automatically — no negative bikers, ever.
- One unit of x_j, rest fixed: the expected count is *multiplied* by e^{β_j} — the **rate ratio** (> 1 more bikers, < 1 fewer).
- The variance follows along: $\text{Var}(Y_i) = \mu_i$ — busy hours are automatically noisier.

Takeaway

Coefficients act on the **log of the mean**; effects on the count scale are **multiplicative**, never additive.

Fitting it in statsmodels

```
import numpy as np, pandas as pd
import statsmodels.api as sm, statsmodels.formula.api as smf

bike = pd.read_csv('Bikeshare.csv', index_col=0) # 8 645 hourly rows
bike['hr'] = bike['hr'].astype(int) # hour 0..23
form = 'bikers ~C(hr) + temp + workingday + C(weathersit)'
pois = smf.glm(form, data=bike,
               family=sm.families.Poisson()).fit() # log link = canonical
print(pois.deviance, pois.df_resid) # 269 006 on 8 616 df
print(np.exp(pois.params['temp'])) # rate ratio 4.79
```

Run this live in the lab notebook

Notebook `advanced_04_glms_splines_lab.ipynb`, §3 *Poisson regression on Bikeshare*: runs this fit, prints the full summary table and reproduces every rate ratio quoted on the next slide.

Reading the fitted coefficients as rate ratios

Term	$\hat{\beta}$	SE	$e^{\hat{\beta}}$ (rate ratio)
Intercept	3.042	0.009	20.9 bikers (baseline hour)
`temp`	1.567	0.005	4.79 per full unit of temp
`workingday`	0.002	0.002	1.002
`weathersit=cloudy/misty`	−0.051	0.002	0.951
`weathersit=light rain/snow`	−0.505	0.004	0.603
`weathersit=heavy rain/snow`	−1.349	0.167	0.260 (!) — from $n = 1$ hour
`hr=8` / `hr=17` / `hr=4`	1.826 / 1.970 / −2.045		6.21 / 7.17 / 0.129 vs midnight

Three concrete readings

- **Light rain** cuts the expected count to $e^{-0.505} = 0.603$ of the clear-sky rate — **39.7% fewer riders**, all else equal.
- $+0.1$ `temp` ($\approx 4^\circ$C) multiplies the rate by $e^{0.1 \times 1.567} = 1.17$ — **+17% riders**.
- Rush hour vs night: $e^{1.970-(-2.045)} = e^{4.015} = 55.4$ — 17:00 carries 55× the 04:00 rate.

Takeaway

Report $e^{\hat{\beta}}$, not $\hat{\beta}$ — and distrust coefficients built on one observation (heavy rain).

Offsets: modelling rates when exposure varies

Rule (offset)

If Y_i is a count accumulated over exposure t_i (hours open, policy-years, population), model the *rate* μ_i/t_i:

$$\log \frac{\mu_i}{t_i} = x_i^\top \beta \iff \log \mu_i = \underbrace{\log t_i}_{\text{offset: coefficient fixed at 1}} + x_i^\top \beta.$$

Why raw counts mislead

Station A: 30 rentals in 2 hours $\Rightarrow$ 15.0/hour. Station B: 60 rentals in 6 hours $\Rightarrow$ 10.0/hour. B has *twice* A's count but only $\frac{2}{3}$ of its rate; the offset makes the GLM compare 15.0 vs 10.0 — the comparison you actually want.

Takeaway

An offset is a covariate with β **frozen at 1**: `smf.glm(..., offset=np.log(t))`. (`Bikeshare` needs none — every row is exactly one hour.)

In industry — Insurance and epidemiology: the offset is the law of the land

Claim counts per policy-year and disease counts per 100 000 population are *always* fitted with exposure offsets — otherwise big portfolios and big cities dominate for size alone.

Exercise A4.3 — Rate ratios on `Bikeshare` [Python]

Exercise

Using the fitted Poisson GLM (`bikers ~ C(hr) + temp + workingday + C(weathersit)`):

1. Compute the rate ratio for `light rain/snow` versus `clear`, and phrase it as a percentage change in expected riders.
2. The `temp` coefficient is $\hat{\beta} = 1.567$. By what factor does the expected count change when `temp` rises by 0.2 (about 8°C)? Write the one line of Python that computes it from `pois.params`.
3. Why would it be wrong to say "a 0.2 rise in temp *adds* $0.2 \times 1.567 = 0.31$ bikers"?

Run this live in the lab notebook

The worked solution sits in `advanced_04_glms_splines_lab.ipynb` under *Lecture exercises — worked Python solutions*: the cell headed *Exercise A4.3 — Rate ratios on Bikeshare*.

Solution — Exercise A4.3

Solution

Method. All effects live on the log-mean scale; exponentiate to get multiplicative effects on the count scale.
Part 1. $\hat{\beta}_{\text{light rain}} = -0.505$, so the rate ratio is $e^{-0.505} = 0.603$: light rain/snow is associated with $1 - 0.603 = 39.7\%$ *fewer* expected riders than clear weather, holding hour, temperature and day type fixed.
Part 2. Factor $= e^{0.2 \times 1.567} = e^{0.313} = 1.368$ — about $+36.8\%$.

```
float(np.exp(0.2 * pois.params['temp']))  # 1.368
```

Part 3. 0.31 is a change in $\log \mu$, not in μ. The additive reading confuses the two scales: the actual change in expected riders depends on the starting level — $+36.8\%$ of 10 riders is ≈ 3.7 riders, of 400 riders ≈ 147. Multiplicative effects have no single additive equivalent.

Common mistake

Treating GLM coefficients like OLS slopes. On the log link, $\hat{\beta}_j$ is a **log rate ratio**; only $e^{\hat{\beta}_j}$ has a direct count-scale meaning.

Outline

Deviance: the GLM's residual sum of squares

Rule (deviance)

With $\hat{\ell}_{\text{sat}}$ the log-likelihood of the *saturated* model (one parameter per observation, each y_i fitted exactly):

$$D \;=\; 2\left(\hat{\ell}_{\text{sat}} - \hat{\ell}\right) \qquad (\phi = 1;\ \text{Poisson: } D = 2\textstyle\sum_i \left[\, y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right]).$$

How to read this

- Smaller D = closer to a perfect fit; for the Gaussian, D *is* the RSS — deviance generalises Chapter 3.
- The **null deviance** D_0 (intercept-only model) plays the role of the total sum of squares.
- `Bikeshare`: $D_0 = 1\,052\,921$ on 8 644 df $\to$ $D = 269\,006$ on 8 616 df — the 28 predictors remove $\sim$74% of the null deviance.

Takeaway

Deviance is to GLMs what RSS is to OLS: the currency in which fits are compared.

Likelihood-ratio tests: comparing nested GLMs

Rule (LRT)

For nested models (reduced $\subset$ full, differing by q parameters), under H_0 "the extra parameters are zero":

$$\text{LR} \;=\; D_{\text{reduced}} - D_{\text{full}} \;=\; 2\,(\hat{\ell}_{\text{full}} - \hat{\ell}_{\text{reduced}}) \;\overset{H_0}{\approx}\; \chi^2_q \quad (n \to \infty).$$

Does weather matter once hour + temp + day type are in?

Drop `weathersit` ($q = 3$ dummies): $D_{\text{reduced}} = 287\,263$, $D_{\text{full}} = 269\,006$, so LR $= 18\,256.8$ on 3 df against a 5% critical value of $\chi^2_{3,0.95} = 7.81$; p is numerically 0. Weather stays in.

Takeaway

The GLM analogue of Chapter 3's F-test for a group of coefficients — same logic, χ^2 instead of F.

Common mistake

The χ^2 result needs *nested* models fitted to the *same rows* by *maximum likelihood* — it does not apply to quasi-likelihood fits.

AIC: model selection beyond nesting (bridge to Ch. 6)

Rule (Akaike information criterion)

$\text{AIC} = -2\hat{\ell} + 2p$ (smaller is better; p = number of parameters).

On `Bikeshare`

Full Poisson model: $\hat{\ell} = -161\,022$, $p = 29 \Rightarrow \text{AIC} = 322\,102$. Without `weathersit`: AIC $= 340\,352$. The full model wins by 18 251 AIC points — consistent with the LRT.

How to read this

- Chapter 6's AIC for regression was the Gaussian special case of this one.
- AIC compares **non-nested** models too (Poisson-with-splines vs Poisson-with-dummies) — as long as both are true likelihoods *for the same response data*.

Takeaway

Deviance for nested tests, AIC for menus of candidates — both are likelihood arithmetic, both carry over to GAMs.

Exercise A4.4 — Deviance, LRT and AIC by hand [Math]

Exercise

For the `Bikeshare` Poisson GLM you are given: full model $D = 269\,005.9$ on $8\,616$ df, $\hat{\ell} = -161\,021.8$, $p = 29$; reduced model (no `weathersit`) $D = 287\,262.7$, $p = 26$.

1. Compute the likelihood-ratio statistic for `weathersit` and its degrees of freedom.
2. Compare with $\chi^2_{3,0.95} = 7.81$ and state the conclusion.
3. Compute the full model's AIC from $\hat{\ell}$ and p.
4. The reduced model's AIC is $340\,352.4$. Recover its log-likelihood.

Solution — Exercise A4.4

Solution

Method. Deviance differences are likelihood-ratio statistics; AIC is $-2\hat{\ell} + 2p$ — everything is plug-in arithmetic.

Part 1. LR $= D_{\text{red}} - D_{\text{full}} = 287\,262.7 - 269\,005.9 = 18\,256.8$, with $q = 29 - 26 = 3$ df (three weather dummies).

Part 2. $18\,256.8 \gg 7.81$: reject H_0 decisively — weather adds explanatory power far beyond chance, given hour, temperature and day type.

Part 3. AIC $= -2(-161\,021.8) + 2 \times 29 = 322\,043.7 + 58.0 = 322\,101.7$ ✓(matches `statsmodels`).

Part 4. $\hat{\ell}_{\text{red}} = -(\text{AIC} - 2p)/2 = -(340\,352.4 - 52)/2 = -170\,150.2$. Check: $2(\hat{\ell}_{\text{full}} - \hat{\ell}_{\text{red}}) = 2(-161\,021.8 + 170\,150.2) = 18\,256.8$ ✓— the same LR as Part 1.

Common mistake

Comparing AICs of models fitted to **different responses** (e.g. y vs $\log y$) or different subsets of rows. AIC differences are only meaningful on identical data.

Outline

The Poisson straitjacket: variance = mean, no exceptions

The Poisson family has **no free variance parameter**: choosing it *asserts* $\mathrm{Var}(Y_i) = \mu_i$.

`Bikeshare` violates this on every shelf

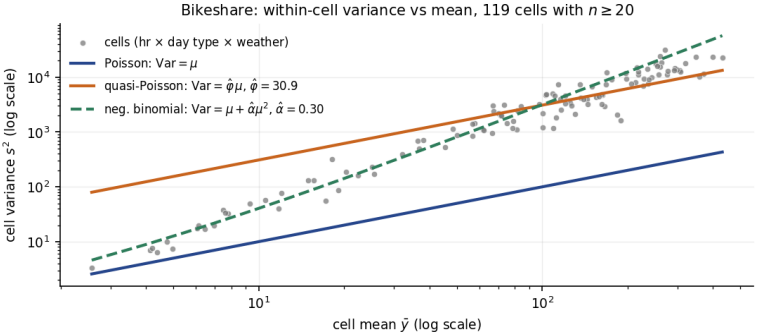
Group the hours into cells of (hour $\times$ day type $\times$ weather), keep the 119 cells with $n \geq 20$:

- 5 p.m., working day, clear: mean 431.5, variance 23 085 — ratio 53.5.
- 4 a.m., working day, clear: mean 4.97, variance 7.4 — ratio 1.5.
- **All 119 cells** have variance above their mean; the median ratio is 27.7.

Takeaway

Counts in the wild are almost always **overdispersed**: unmodelled heterogeneity (events, marketing, school holidays…) inflates the variance beyond Poisson. The mean model can still be fine — the *uncertainty* is not.

Seeing the violation: variance vs mean across cells



Takeaway

The Poisson line Var $= \mu$ runs far below every cell. The quasi-Poisson line $\hat{\varphi}\,\mu$ with $\hat{\varphi} = 30.9$ matches mid-range cells; the negative-binomial curve $\mu + 0.30\,\mu^2$ bends up to catch the busy, high-variance cells. The data choose the curved law.

Fix 1 — quasi-Poisson: rescale the uncertainty

Rule (quasi-likelihood with dispersion)

Keep the mean model and $V(\mu) = \mu$, but allow $\mathrm{Var}(Y_i) = \varphi\,\mu_i$, estimating

$$\hat{\varphi} = \frac{X^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \qquad \widehat{\mathrm{SE}}_{\mathrm{quasi}}(\hat{\beta}_j) = \sqrt{\hat{\varphi}}\,\widehat{\mathrm{SE}}_{\mathrm{Pois}}(\hat{\beta}_j).$$

Bikeshare

$X^2 = 266\,229$ on $n - p = 8\,616$ df $\Rightarrow$ $\hat{\varphi} = 30.90$: every standard error grows by $\sqrt{30.90} = 5.56$. **Point estimates do not move.** Consequences: `temp` z: $330 \to 59.4$ (still overwhelming); `workingday` z: $1.19 \to 0.21$ — its apparent effect evaporates.

Takeaway

Overdispersion rarely changes *what* the effects are — it changes *how sure you are*. Borderline stars are the first casualties.

Fix 2 — negative binomial: model the extra variance

Rule (NB2 negative binomial)

Let $Y_i \mid G_i \sim \text{Poisson}(\mu_i G_i)$ with a gamma-distributed multiplicative effect G_i ($E[G_i] = 1$, $\text{Var}(G_i) = \alpha$). Then $E[Y_i] = \mu_i, \quad \text{Var}(Y_i) = \mu_i + \alpha\,\mu_i^2$.

Bikeshare

ML estimate $\hat{\alpha} = 0.305$. Log-likelihood improves from $-161\,022$ (Poisson) to $-45\,266$ (NB); AIC drops from 322 102 to 90 590 — not a close call. At $\mu = 400$ the implied variance is $400 + 0.305 \times 400^2 \approx 49\,200$, versus Poisson's 400.

Quasi vs NB in one line each

Quasi-Poisson: same estimates, honest SEs, **no likelihood** (no AIC, no LRT). *Negative binomial*: a genuine likelihood — estimates shift (it reweights observations), AIC/LRT available, variance grows *quadratically*.

Takeaway

Only need honest standard errors? Quasi suffices. Need likelihood machinery or predictive distributions? Fit the negative binomial.

Poisson vs quasi-Poisson vs negative binomial, side by side

	Poisson		quasi-Poisson	Negative binomial	
Term	$\hat{\beta}$	SE	SE ($\times 5.56$)	$\hat{\beta}$	SE
`temp`	1.567	0.005	0.026	1.893	0.032
`workingday`	0.002	0.002	0.011	−0.195	0.013
`light rain/snow`	−0.505	0.004	0.022	−0.538	0.022
AIC	322 102		— (no likelihood)	90 590	

Which numbers changed — and why

Quasi keeps every $\hat{\beta}$ and multiplies SEs by $\sqrt{\hat{\varphi}} = 5.56$. NB *refits*: its likelihood downweights huge counts, so estimates move — `workingday` even flips sign, because its true effect is heterogeneous across hours (positive at commuter peaks, negative at night) and the two likelihoods average that heterogeneity with different weights.

In industry — Retail demand: the dispersion audit

Demand-forecasting teams sanity-check $\hat{\varphi}$ per store cluster. A jump of $\hat{\varphi}$ from 3 to 30 after a promotion week is a red flag that the mean model is missing a driver — not a licence for tighter intervals.

Exercise A4.5 — Dispersion arithmetic [Math]

Exercise

The `Bikeshare` Poisson fit has Pearson statistic $X^2 = 266\,229$ with $n = 8\,645$ observations and $p = 29$ parameters. The `temp` coefficient is $\hat{\beta} = 1.567$ with Poisson SE 0.0047.

1. Estimate the dispersion $\hat{\varphi}$ and the factor by which all standard errors must grow.
2. Compute the 95% confidence interval for $\beta_{\texttt{temp}}$ under the Poisson fit and under quasi-Poisson (use ± 1.96 SE).
3. Convert both intervals to rate-ratio scale.
4. A colleague argues: "the quasi interval is wider, so the Poisson one is more precise — keep it." Rebut in one sentence.

Solution — Exercise A4.5

Solution

Method. $\hat{\varphi} = X^2/(n-p)$; quasi SEs scale by $\sqrt{\hat{\varphi}}$; exponentiate CI endpoints for rate ratios.
Part 1. $\hat{\varphi} = 266\,229/(8\,645 - 29) = 266\,229/8\,616 = 30.90$; SEs grow by $\sqrt{30.90} = 5.56$.
Part 2. Poisson: $1.567 \pm 1.96 \times 0.0047 = [1.558,\ 1.576]$. Quasi: SE $= 5.56 \times 0.0047 = 0.0264$, so $1.567 \pm 1.96 \times 0.0264 = [1.515,\ 1.619]$.
Part 3. Rate-ratio scale: Poisson $[e^{1.558}, e^{1.576}] = [4.75,\ 4.83]$; quasi $[e^{1.515}, e^{1.619}] = [4.55,\ 5.05]$.
Part 4. The Poisson interval is not *more precise*, it is *wrong*: its claimed coverage assumes Var $= \mu$, which the data contradict by a factor of $\sim$31, so the narrow interval covers the truth far less often than 95%.

Common mistake

Reading interval width as a quality metric. Width is only meaningful *after* the variance assumption survives a check like $\hat{\varphi} \approx 1$.

Extended Exercise A4.1 — The overdispersion pipeline [Integrative]

Extended exercise (15 min) — from suspicion to model choice

Work through the full overdispersion protocol on `Bikeshare` (`bikers ~ C(hr) + temp + workingday + C(weathersit)`):

1. **Diagnose.** Compute $\hat{\varphi}$ from the Poisson fit's Pearson statistic. What values would have let you keep plain Poisson?
2. **Repair the SEs.** Which of the following change under quasi-Poisson, and how: point estimates, standard errors, fitted means, AIC?
3. **Re-model.** The negative binomial fit gives $\hat{\alpha} = 0.305$ and AIC 90 590 vs Poisson's 322 102. Why is comparing these two AICs legitimate, while "quasi-Poisson AIC" does not exist?
4. **Decide.** `workingday` has $\hat{\beta} = 0.002$ (Poisson), $z_{\text{quasi}} = 0.21$, but $\hat{\beta} = -0.195$ (NB, $z \approx -14.7$). What should an analyst conclude and report about the working-day effect?

Run this live in the lab notebook

Notebook §5 *Overdispersion: quasi-Poisson and negative binomial* fits all three models and prints the comparison table from the previous slide.

Solution — Extended Exercise A4.1 (1/2)

Solution — diagnosis and repair

Method. One dispersion statistic drives the whole protocol; then separate what quasi-likelihood rescales from what a new likelihood refits.

Part 1. $\hat{\varphi} = 266\,229/8\,616 = 30.90$. Values near 1 (rule of thumb $\lesssim 1.5$, given the huge n here even tighter) would justify plain Poisson; 30.9 is a categorical rejection — the variance is $\sim 31\times$ the Poisson claim on average.

Part 2. Under quasi-Poisson: point estimates **unchanged** (same estimating equations); fitted means **unchanged**; standard errors **multiplied by** $\sqrt{30.90} = 5.56$; AIC **undefined** — quasi-likelihood is not a likelihood, so there is no $\hat{\ell}$ to plug into $-2\hat{\ell} + 2p$.

Solution — Extended Exercise A4.1 (2/2)

Solution — re-model and decide

Part 3. Poisson and negative binomial are both *genuine likelihoods for the same response vector*, so their AICs are comparable: the NB's $90\,590 \ll 322\,102$ says the quadratic variance law fits these counts far better. Quasi-Poisson specifies only mean and variance, never a density, hence no likelihood and no AIC.

Part 4. The honest conclusion: *averaged over all hours*, working days do not add a stable multiplicative effect — quasi says the average effect is indistinguishable from zero, and the NB's -0.195 reflects its different weighting of hours, not a newly discovered negative effect. The effect is **heterogeneous** (strongly positive at 8 a.m. and 5–6 p.m., negative at night and midday); report it via an *interaction* (workingday $\times$ hour) or a GAM with separate hour curves, not a single number.

Common mistake

Reading the sign flip as "the NB found the true effect". When two reasonable likelihoods disagree this much, the model is mis-specified — the fix is a richer mean model, not a coin flip between packages.

Outline

Chapter 7 recap: regression splines and the knot dilemma

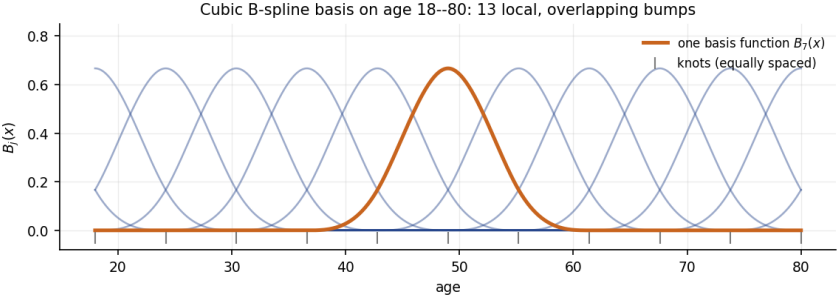
What you already know

- A **regression spline** of degree 3 is a piecewise cubic, forced to be continuous with continuous first and second derivatives at the **knots**.
- Fit by OLS on a spline **basis**: flexibility is controlled by the *number and placement of knots* (equivalently, the degrees of freedom).
- Chapter 7's recipe: place knots at quantiles, pick df by cross-validation.

The dilemma this section resolves

Knot choice is **discrete and awkward**: 5 vs 6 knots is a lumpy, all-or-nothing decision, and knots in the wrong place cannot be undone by fitting. The penalized view replaces "how many knots?" by a single **continuous** dial λ: use *plenty* of knots, then penalise wiggliness.

The B-spline basis: local, overlapping bumps



Takeaway

Each cubic B-spline $B_j(x)$ is positive on only 4 adjacent knot intervals and zero elsewhere; at any x they sum to 1. The fit $f(x) = \sum_j \beta_j B_j(x)$ is a **local** blend — moving one β_j reshapes the curve only near its bump, which is why the basis is numerically stable where raw polynomials (x^7, x^8, ...) explode.

Penalized splines: many knots, one dial

Rule (penalized fitting)

Use a rich B-spline basis (20–40 knots) and minimise

$$\underbrace{\text{Deviance}(\beta)}_{\text{fit the data}} + \lambda \underbrace{\int f''(t)^2\, dt}_{\text{punish wiggliness}} \qquad \text{with } \int f''(t)^2 dt = \beta^\top S \beta.$$

Gaussian case: $\hat{\beta}_\lambda = (B^\top B + \lambda S)^{-1} B^\top y$ — ridge regression (Ch. 6) with a structured penalty.

How to read this

$\lambda = 0$: unpenalized — as wiggly as the basis allows. $\lambda \to \infty$: $f'' \equiv 0$ forced — a **straight line** (zero curvature is invisible to the penalty). In between: the bias–variance dial of Chapter 2, turned *continuously*.

Takeaway

Knot placement stops mattering once knots are plentiful; **all** structural choice is concentrated in the single number λ.

Effective degrees of freedom: how much did you really fit?

Rule (edf)

The penalized fit is linear in y: $\hat{y} = H(\lambda)\, y$ with smoother matrix $H(\lambda) = B\,(B^\top B + \lambda S)^{-1} B^\top$. Define

$$\mathrm{edf}(\lambda) \;=\; \mathrm{tr}\{H(\lambda)\} \;=\; \sum_k \frac{1}{1+\lambda d_k}, \qquad d_k \geq 0 \text{ the penalty's eigenvalues in the basis.}$$

Toy computation

Four basis directions with $d = (0,\, 0,\, 1,\, 4)$ and $\lambda = 1$: $\mathrm{edf} = 1 + 1 + \frac{1}{2} + \frac{1}{5} = 2.7$. The two $d = 0$ directions (constant, linear — zero curvature) are never shrunk; the wiggly ones are partially switched off.

Takeaway

edf interpolates smoothly between 2 (a line) and K (the full basis) — "this smooth is worth 6.2 parameters" is a meaningful, comparable statement.

Choosing λ: cross-validation or GCV (bridge to Ch. 5)

Rule (generalized cross-validation)

$$\mathrm{GCV}(\lambda) \;=\; \frac{n\,\mathrm{RSS}(\lambda)}{\big(n - \mathrm{edf}(\lambda)\big)^2} \qquad \text{— minimise over } \lambda.$$

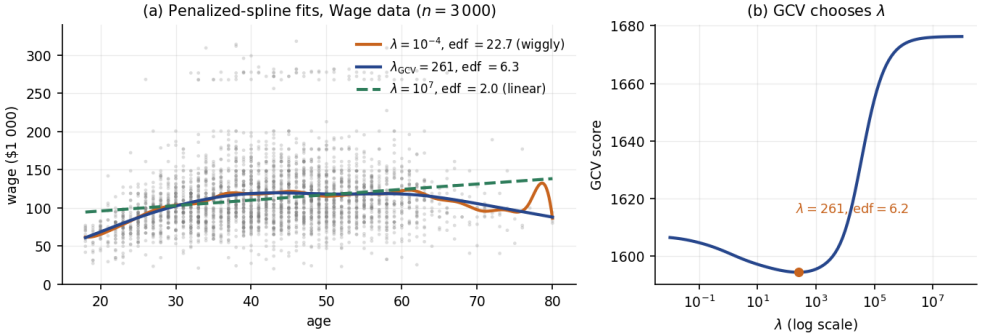
How to read this

- GCV is a shortcut to leave-one-out CV: no n refits, just the single fit's edf — the same idea as Chapter 5's LOOCV formula for linear smoothers.
- k-fold CV on a λ-grid works too — the honest default when the model is not a linear smoother (e.g. GLM deviance losses).
- Watch **edf**, not λ: λ's scale depends on the basis, edf does not.

Takeaway

λ is a tuning parameter like any in Chapters 5–6: chosen by (generalized) cross-validation, never by eye.

One picture: Wage across three orders of magnitude of λ



Takeaway

23 cubic B-splines on age ($n = 3\,000$). At $\lambda = 10^{-4}$ the fit burns 22.7 edf and chases noise around age 78; at $\lambda = 10^{7}$ it collapses to a line (2.0 edf) and misses the plateau. GCV picks $\lambda = 261$, edf $= 6.2$ — close to the df Chapter 7 found by CV for this very curve.

Exercise A4.6 — Reading the λ dial [Math]

Exercise

A penalized cubic spline uses $K = 23$ B-spline basis functions with a second-derivative penalty S whose null space contains exactly the constant and linear functions.

1. State edf in the limits $\lambda \to 0$ and $\lambda \to \infty$, with one sentence of justification each.
2. In a small example the penalty eigenvalues are $d = (0, 0, 1, 4)$. Compute edf for $\lambda = 1$ and for $\lambda = 4$.
3. Your colleague reports "$\lambda = 261$" as the headline of their smooth. Why is "edf $= 6.2$" the better headline?

Solution — Exercise A4.6

Solution

Method. Use $\text{edf}(\lambda) = \sum_k 1/(1+\lambda d_k)$ and think about which directions the penalty can and cannot touch.
Part 1. $\lambda \to 0$: no shrinkage, every basis direction survives, edf $\to K = 23$ (an unpenalized regression spline). $\lambda \to \infty$: every direction with $d_k > 0$ is annihilated; only the null space of S — constant + linear — survives, so edf $\to 2$ (a straight line).
Part 2. $\lambda = 1$: $1 + 1 + \frac{1}{1+1} + \frac{1}{1+4} = 2 + 0.5 + 0.2 = 2.70$.
$\lambda = 4$: $1 + 1 + \frac{1}{5} + \frac{1}{17} = 2 + 0.200 + 0.059 = 2.26$ — quadrupling λ removed less than half an edf: the dial is logarithmic.
Part 3. λ's numerical value depends on the basis size, the knot spacing, even the units of x — 261 is meaningless across studies. edf is invariant to all of that and reads directly as model complexity: "a curve worth 6.2 parameters".

Common mistake

Comparing λ values across different bases or datasets. Only **edf** travels.

Outline

The meeting point: GAM = GLM + smooth terms

Rule (penalized-spline GAM for counts)

$$Y_i \sim \text{Poisson}(\mu_i), \qquad \log \mu_i \;=\; \beta_0 + f_1(\texttt{hr}_i) + f_2(\texttt{temp}_i) + \beta_w\, \texttt{workingday}_i + \sum_k \gamma_k\, \texttt{weather}_{ik}$$

each f_j a penalized B-spline with its own λ_j; fit by **penalized maximum likelihood** (penalized IRLS).

How to read this

- Everything from this module composes: exponential family (random part), log link, deviance-based fitting, roughness penalties per smooth.
- Chapter 7 introduced GAMs with fixed-df smoothers; the upgrade here is that each curve's complexity is *estimated* via λ_j, not preset.
- Additivity is retained: each f_j is a one-dimensional curve you can *plot and defend* — the interpretability contract of Ch. 7 survives.

Fitting the GAM with statsmodels

```
from statsmodels.gam.api import GLMGam, BSplines

xs = bike[['hr', 'temp']].astype(float)
bs = BSplines(xs, df=[12, 6], degree=[3, 3]) # rich bases for hr, temp
Xpar = ... # intercept, workingday, 3 weather dummies (parametric part)

gam = GLMGam(bike.bikers, exog=Xpar, smoother=bs,
             family=sm.families.Poisson(), alpha=[0.319, 0.00423])
res = gam.fit()
# the two penalty weights above were chosen by the built-in criterion:
# gam.select_penweight()[0] -> [0.319, 0.00423]
print(res.deviance) # 259 650
```

Run this live in the lab notebook

Notebook §7 *A penalized-spline GAM for Bikeshare* builds `Xpar`, runs `select_penweight`, and draws the partial-effect plots on the next slide with confidence bands.

What the machine chose

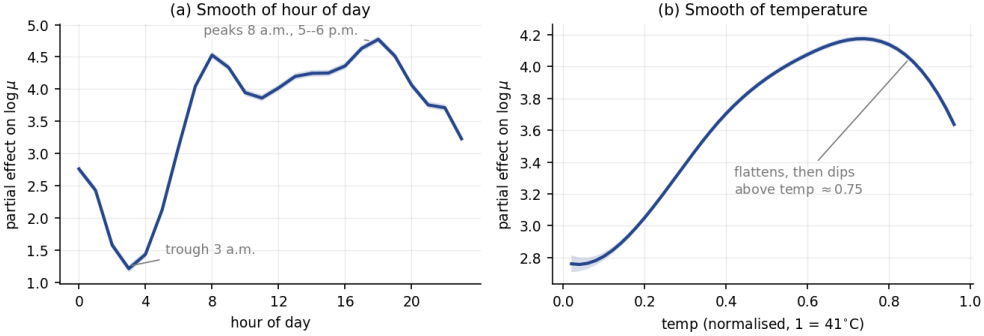
Selected penalties and complexity on `Bikeshare`

- `select_penweight` (GCV-type criterion) returns $\hat{\alpha} = (0.319,\ 0.00423)$ for (`hr`, `temp`).
- Resulting complexity: hour smooth ≈ 11 edf, temperature smooth ≈ 5 edf — the light penalties keep almost all of the 12 and 6 basis df: with 8 645 observations the data can afford it.
- Deviance 259 650 vs 269 006 for the dummy-`hr` GLM with linear `temp`; AIC 312 730 vs 322 102 — the smooth-temp model wins by $\sim$9 400 AIC points with *fewer* effective parameters.

Takeaway

The gain comes almost entirely from letting `temp` **bend**: the linear-temp assumption of the earlier GLM was the binding constraint, not the hour dummies.

The fitted smooths: hour of day and temperature



Takeaway

Hour: trough at 3 a.m., twin commuter peaks at 8 a.m. and 5–6 p.m.; peak-to-trough ratio $e^{4.77-1.21} = 35\times$. **Temp**: riders triple ($\times 3.1$) from temp 0.2 to 0.7, then *fall* beyond ≈ 0.75 ($\sim 31^\circ$C) — a hot-weather dip a linear term is structurally unable to show.

Interpreting the hour curve — and one honest caveat

Reading a partial effect

Each curve shows the contribution to $\log \mu$ *holding the other terms fixed*. Differences along one curve exponentiate to rate ratios: $f(18) - f(3) = 4.77 - 1.21 = 3.56 \Rightarrow e^{3.56} \approx 35\times$ more riders at 6 p.m. than at 3 a.m., same weather, same temperature, same day type.

Smooths smooth — also where the truth is spiky

The dummy model rates 17:00 at 55.4× the 04:00 level; the smooth version says 24.6×. The penalty shrinks the razor-sharp night dip and commuter spikes toward their neighbours. For genuinely *discontinuous* patterns (timetables, opening hours), factors or more basis df beat a heavily penalized smooth — check edf and residuals by hour before trusting the curve.

In industry — Energy: GAMs run the grid

European system operators forecast national electricity load with exactly this structure — smooth(hour), smooth(temperature), calendar dummies — because engineers must *see and sign off* each curve. The hot-weather bend (air conditioning) is the industry's textbook example of why temp must be a smooth.

Pitfalls when GLMs meet splines

Don't

1. **Extrapolate a spline.** Outside the knot range the fit continues as a low-order polynomial with *no data behind it*; on the log scale a mild linear overshoot exponentiates into absurd counts.
2. **Read smooths causally.** $f(\texttt{temp})$ is an association given the other terms; hot hours also differ in season, daylight and tourism. Nothing here is a do-operator.
3. **Ignore overdispersion because the mean is now flexible.** The GAM's deviance is still Poisson: with $\hat{\varphi} \approx 31$ on this data, its naive confidence bands are $\sim 5.6\times$ too narrow too. Rescale or go NB.

Takeaway

Flexibility fixes the *mean*; it does not buy causality and it does not fix the *variance*. Keep the three audits separate.

Extended Exercise A4.2 — Build and defend a count GAM [Python]

Extended exercise (15 min) — the full modelling arc

Using `GLMGam` with `BSplines(df=[12, 6], degree=[3, 3])` on `hr` and `temp`, plus parametric `workingday` and weather dummies:

1. Write down the model equation for $\log \mu_i$ and say which parts are penalized.
2. The selected penalties give edf ≈ 11 (hr) and ≈ 5 (temp). What would edf ≈ 2 for `temp` have told you?
3. From the fitted hour smooth, $f(18) = 4.77$ and $f(3) = 1.21$. Compute the 6 p.m. vs 3 a.m. rate ratio and interpret it in one sentence.
4. The GAM's weather coefficient is -0.569 (light rain) vs the GLM's -0.505. Give one reason the estimates differ.
5. Name the single most important robustness check before this model ships to the fleet-rebalancing team.

Run this live in the lab notebook

Notebook §7 closes with this exercise: the cell headed *Extended Exercise A4.2* reproduces every number.

Solution — Extended Exercise A4.2 (1/2)

Solution — model / edf / rate ratio

Method. Assemble the GAM equation, then read complexity from edf and effects from differences on the log scale.

Part 1. $\log \mu_i = \beta_0 + f_1(\texttt{hr}_i) + f_2(\texttt{temp}_i) + \beta_w \texttt{workingday}_i + \gamma_1 \texttt{cloudy}_i + \gamma_2 \texttt{light}_i + \gamma_3 \texttt{heavy}_i$. Only f_1 and f_2 are penalized (each with its own α); intercept, dummies and β_w are ordinary ML parameters.

Part 2. edf ≈ 2 for `temp` would mean the penalty had shrunk the smooth to (near-)linearity — i.e. the data see no curvature and the Ch. 4 linear-temp GLM would have been adequate. The actual edf ≈ 5 says the bend (including the hot-weather dip) is real structure worth $\sim$5 parameters.

Part 3. $e^{4.77-1.21} = e^{3.56} \approx 35$: holding temperature, weather and day type fixed, the expected count at 6 p.m. is about 35 times the 3 a.m. count.

Solution — Extended Exercise A4.2 (2/2)

Solution — coefficient shift and the shipping check

Part 4. The two models adjust for temperature differently: the GLM forces a linear `temp` effect, the GAM lets it bend. Rainy hours are also cool hours, so when the temperature adjustment changes shape, the weather dummies absorb a different share of the rain–cold overlap (-0.569 vs -0.505). Coefficients of correlated regressors always move when a co-regressor's functional form changes.

Part 5. An **overdispersion audit** of the GAM itself: compute the Pearson dispersion of the GAM fit (it remains $\gg 1$ on this data) and widen the uncertainty accordingly (quasi rescaling or an NB-GAM). Shipping the $35\times$ headline with Poisson bands would overstate certainty roughly 5–6-fold — the fleet team would over-trust hour-level differences that are partly noise.

Common mistake

Declaring victory after the mean model improves. The variance audit (§Overdispersion) applies to GAMs verbatim — flexible means do not repair broken variances.

Outline

Module A4 in one slide

What we accomplished

- Diagnosed why OLS fails on counts: negative means, constant variance — 10% negative predictions on `Bikeshare`.
- Unified Chs. 3–4 in the **exponential family**: mean $b'(\theta)$, variance $b''(\theta)a(\phi)$.
- Assembled the **GLM**: family + linear predictor + link; Poisson coefficients as **rate ratios**; offsets for exposure.
- Ran GLM **inference**: deviance, LRT (18 257 on 3 df for weather), AIC.
- Confronted **overdispersion** ($\hat{\varphi} = 30.9$): quasi-Poisson rescales SEs $\times 5.56$; negative binomial ($\hat{\alpha} = 0.305$) refits.
- Rebuilt splines as **penalized** fits: one dial λ, complexity read as edf, tuned by GCV/CV.
- Met in the middle: a **penalized-spline Poisson GAM** with interpretable hour and temperature curves.

Key formulas at a glance

Exponential family $f(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\}$

Mean and variance $E[Y] = b'(\theta), \quad \mathrm{Var}(Y) = b''(\theta)\,a(\phi) = a(\phi)V(\mu)$

GLM link $g(\mu_i) = \eta_i = x_i^\top \beta$; canonical link $g = (b')^{-1}$

Poisson rate ratio one unit of x_j multiplies μ by e^{β_j}; offset: $\log \mu_i = \log t_i + x_i^\top \beta$

Deviance / LRT $D = 2(\hat{\ell}_{\text{sat}} - \hat{\ell}); \quad D_{\text{red}} - D_{\text{full}} \sim \chi^2_q$ under H_0

AIC $-2\hat{\ell} + 2p$

Dispersion $\hat{\varphi} = X^2/(n-p)$; quasi SE $= \sqrt{\hat{\varphi}} \times$ Poisson SE

Negative binomial $\mathrm{Var}(Y) = \mu + \alpha\mu^2$

Penalized spline minimise Deviance $+ \lambda \int f''(t)^2 dt = \mathrm{Dev} + \lambda\, \beta^\top S \beta$

Effective df $\mathrm{edf} = \mathrm{tr}\{B(B^\top B + \lambda S)^{-1} B^\top\} = \sum_k (1 + \lambda d_k)^{-1}$

GCV $n\,\mathrm{RSS}/(n - \mathrm{edf})^2$

Count GAM $\log \mu_i = \beta_0 + \sum_j f_j(x_{ij}) +$ (parametric terms)

Vocabulary checklist

Term	Meaning
Canonical parameter θ	natural scale on which y enters the density linearly
Cumulant function $b(\theta)$	generates mean (b') and variance (b'')
Variance function $V(\mu)$	mandated shape of variance as a function of the mean
Link function g	maps the mean's range to $\mathbb{R}$; canonical if $g = (b')^{-1}$
Linear predictor η	$x^\top \beta$, the part shared with Chs. 3–4
Rate ratio	e^{β_j}: multiplicative effect on the expected count
Offset	covariate with coefficient fixed at 1, encodes exposure
Saturated model / deviance	perfect-fit benchmark; $2\times$ log-likelihood gap to it
Overdispersion	Var $>$ family's claim; here $\hat{\varphi} = 30.9$
Quasi-likelihood	mean–variance assumptions only; no density, no AIC
Negative binomial	gamma-mixed Poisson; Var $= \mu + \alpha\mu^2$
B-spline basis	local overlapping polynomial bumps; numerically stable
Roughness penalty	$\lambda \int f''^2$: price tag on curvature
Effective df (edf)	tr $H(\lambda)$: parameters actually spent
GCV	leave-one-out shortcut for linear smoothers
GAM / partial effect	additive smooth terms inside a GLM; one curve per predictor

Decision rules of thumb

- Response is a **count** $\Rightarrow$ start with Poisson + log link; **binary** $\Rightarrow$ Bernoulli + logit; **continuous symmetric** $\Rightarrow$ Gaussian + identity.
- Counts accumulated over unequal windows $\Rightarrow$ **offset** log(exposure), always.
- After any Poisson fit, compute $\hat{\varphi} = X^2/(n-p)$. If $\hat{\varphi} \gtrsim 1.5$: quasi-Poisson for quick honest SEs; negative binomial when you need AIC, LRTs or predictive distributions.
- Effect shape unknown $\Rightarrow$ penalized spline with generous basis; let GCV/CV set λ; **report edf**, not λ.
- Smooth's edf $\approx 2 \Rightarrow$ simplify to a linear term; edf near the basis size $\Rightarrow$ consider a larger basis or a factor.
- Report $e^{\hat{\beta}}$ (rate/odds ratios) with intervals from the *dispersion-corrected* fit.

Takeaway

Family first, link second, dispersion audit third, smoothness by CV fourth — in that order, every time.

Common pitfalls

Don't

1. **Fit OLS to raw counts** (or to $\log(y+1)$ and call it equivalent) — wrong support, wrong variance, biased back-transformed means.
2. **Interpret GLM coefficients additively** — on a log link only $e^{\hat{\beta}}$ speaks the count language.
3. **Trust Poisson standard errors without a dispersion check** — at $\hat{\varphi} = 30.9$ they are $5.6\times$ too small; borderline "discoveries" evaporate.
4. **Use AIC/LRT on quasi-likelihood fits** — there is no likelihood there.
5. **Extrapolate splines beyond the knots** or compare raw λ values across bases — use edf and stay in range.
6. **Read GAM smooths as causal effects** — they are adjusted associations, nothing more.

Self-check questions

Can you answer these without the slides?

1. Derive the mean and variance of the Poisson from $b(\theta) = e^{\theta}$.
2. Why is the logit the *canonical* link for the Bernoulli, and what does canonicality buy?
3. Weather's LRT was 18 257 on 3 df. Which two models were compared, and what assumption makes the χ^2 reference valid?
4. Quasi-Poisson and negative binomial both "fix" overdispersion — name two practical situations that force the NB choice.
5. Why does $\lambda \to \infty$ leave a straight line rather than a constant, for a second-derivative penalty?
6. Your GAM's temperature smooth has edf 5.0 and dips after temp 0.75. What would you check before presenting the dip as real?

Takeaway

If question 1 takes more than two lines, revisit the worked Poisson derivation — it is the

Appendix — what is in here, and why

Nothing in the appendix is needed to follow the chapter: each slide is a formal derivation, a longer worked exercise, or a side topic. Read it when you want the full story, or when a later chapter sends you back here.

Contents of the appendix

- **Score identities** — the two-line proof of $E[Y] = b'(\theta)$ and $\mathrm{Var}(Y) = b''(\theta)a(\phi)$; moved out because the main deck only needs the result.
- **The Gaussian as a family member** — completes the table of three; moved out because it mirrors the Poisson derivation done live.
- **IRLS in one slide** — how GLMs are actually fitted; moved out because no lecture decision depends on the algorithm's internals.
- **Deviance formulas per family** — reference table; moved out because it is look-up material, not narrative.
- **Drill exercise on offsets** — extra practice with full solution; moved out to keep the main flow at two extended exercises.

Score identities: where mean and variance come from

For one observation, $\ell(\theta) = \dfrac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$.

Derivation

First identity. Under regularity, $E[\partial \ell / \partial \theta] = 0$:

$$E\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = 0 \implies E[Y] = b'(\theta).$$

Second identity. The Bartlett identity $E\left[\partial^2 \ell / \partial \theta^2\right] + E\left[(\partial \ell / \partial \theta)^2\right] = 0$ gives

$$-\frac{b''(\theta)}{a(\phi)} + \frac{\mathrm{Var}(Y)}{a(\phi)^2} = 0 \implies \mathrm{Var}(Y) = b''(\theta)\, a(\phi). \qquad \blacksquare$$

Takeaway

Both facts used throughout the module are two lines of calculus on the score — no distribution-specific work needed.

The Gaussian as a family member (completing the table)

Match the normal density to the family form

$$f(y;\mu,\sigma^2) = \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \tfrac{1}{2}\log(2\pi\sigma^2)\right\}.$$

Matching: $\theta = \mu$, $b(\theta) = \theta^2/2$, $a(\phi) = \sigma^2$, $c(y,\phi) = -y^2/(2\sigma^2) - \frac{1}{2}\log(2\pi\sigma^2)$.
Identities: $E[Y] = b'(\theta) = \theta = \mu$ ✓; $\mathrm{Var}(Y) = b''(\theta)a(\phi) = 1 \cdot \sigma^2$ ✓.

Takeaway

$b'' \equiv 1$: the Gaussian is the *only* member whose variance ignores the mean — exactly the property that made OLS fail on counts, and exactly why Chapter 3 never needed a variance function.

IRLS: how a GLM is actually fitted

Algorithm (iteratively reweighted least squares)

Repeat until convergence, starting from any sensible $\hat{\mu}^{(0)}$:

1. Working response: $z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\, g'(\hat{\mu}_i)$.
2. Working weights: $w_i = \left[\, g'(\hat{\mu}_i)^2\, V(\hat{\mu}_i) \,\right]^{-1}$.
3. Update β by weighted least squares of z on X with weights w; set $\hat{\eta} = X\hat{\beta}$, $\hat{\mu} = g^{-1}(\hat{\eta})$.

How to read this

- IRLS is Newton–Raphson with the expected information (Fisher scoring), rearranged as repeated WLS — which is why GLM software is built on the Chapter 3 machinery.
- Penalized GAM fitting only changes step 3: WLS becomes *penalized* WLS with λS added — the ridge update from Chapter 6.

Takeaway

A GLM is "OLS in a loop"; a GAM is "ridge in a loop". Nothing beyond this course's toolbox.

Deviance formulas per family (reference)

Family	Unit deviance contribution d_i	Note
Gaussian	$(y_i - \hat{\mu}_i)^2$	D = RSS: Ch. 3 recovered
Bernoulli	$-2\big[\, y_i \log \hat{\mu}_i + (1 - y_i) \log(1 - \hat{\mu}_i) \,\big]$	= the cross-entropy loss
Poisson	$2\big[\, y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i) \,\big]$	terms with $y_i = 0$: $2\hat{\mu}_i$
Neg. binomial	$2\big[y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i + \frac{1}{\alpha}) \log \frac{1+\alpha y_i}{1+\alpha \hat{\mu}_i} \big]$	reduces to Poisson as $\alpha \to 0$

How to read this

$D = \sum_i d_i$ in each case. Residual checks for GLMs use *deviance residuals* $\text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$, which are the closest GLM analogue of the OLS residuals from Chapter 3.

Drill Exercise A4.A — Offsets in anger [Math]

Exercise

A city compares cycling at two counting stations using a Poisson GLM with log link. Station 1 recorded $y_1 = 180$ riders over $t_1 = 12$ hours; Station 2 recorded $y_2 = 96$ riders over $t_2 = 4$ hours.

1. Compute each station's empirical rate per hour.
2. In the model $\log \mu_i = \log t_i + \beta_0 + \beta_1 \mathbb{1}\{i = 2\}$, express e^{β_1} in terms of the two rates and compute its ML estimate.
3. What would the same model *without* the offset (raw counts) have estimated for e^{β_1}, and why is that misleading?

Solution — Drill Exercise A4.A

Solution

Method. With a saturated two-group Poisson model, ML fits each group's rate exactly — so everything reduces to ratios of empirical rates.

Part 1. Station 1: $180/12 = 15.0$ riders/hour. Station 2: $96/4 = 24.0$ riders/hour.

Part 2. With the offset, $\mu_i = t_i\, e^{\beta_0 + \beta_1 \mathbb{1}\{i=2\}}$, so e^{β_0} is Station 1's rate and e^{β_1} the **rate ratio**:

$$e^{\hat{\beta}_1} = \frac{y_2/t_2}{y_1/t_1} = \frac{24.0}{15.0} = 1.60$$

— Station 2 sees 60% *more* riders per hour.

Part 3. Without the offset the model compares raw counts: $e^{\hat{\beta}_1} = 96/180 = 0.533$ — Station 2 would look 47% *quieter*, purely because it was observed for a third of the time. Same data, opposite conclusion; the offset is what makes the comparison fair.

Common mistake

Adding exposure as an ordinary covariate (β estimated) "to be safe". That lets the model bend the exposure effect away from proportionality and destroys the rate interpretation; the offset's whole point is $\beta_{\text{exposure}} \equiv 1$.