Introduction to regression models

May 28, 2023

Overview

1. Conditional modeling
2. Bayesian analysis of classical regression
3. Regression for causal inference: incumbency and voting

Conditional modeling

- Denotes the explanatory variables X as a $n \times k$ matrix, the conditional mode of response variable y is

$$\mathrm{E}(y_i|\beta, X) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

  for $i = 1, \ldots, n$.
- The intercept term is usually treated as the first term x_1 which is a one vector with length n.
- An ordinary linear regression assume the response y_is are i.i.d. condition on x_is, so that $var(y_i|\theta, X) = \sigma^2$ for all i.
- The parameters are $\theta = \{\beta_1, \ldots, \beta_k, \sigma\}$

Formal Bayesian justification of conditional modeling

- A full Bayesian model includes a distribution for X, $p(X|\psi)$, indexed by a parameter vector ψ.
- The joint likelihood, $p(X, y|\psi, \theta)$, has a prior distribution, $p(\psi, \theta)$.
- ASSUMPTION: X is assumed to provide no information about the conditional distribution of y given X; that is, we assume prior independence of the parameters θ determining $p(y|X, \theta)$ and the parameters ψ determining $p(X|\psi)$.
- Also, we suppose ψ and θ are independent in their prior distribution, $p(\psi, \theta) = p(\psi)p(\theta)$

Formal Bayesian justification of conditional modeling

- The posterior distribution factors,

$$p(\psi, \theta|X, y) = p(\psi|X)p(\theta|X, y),$$

- and we can analyze the second factor by itself (that is, as a standard regression model), with no loss of information:

$$p(\theta|X, y) \propto p(\theta)p(y|X, \theta).$$

Bayesian analysis of classical regression

- Under a standard noninformative prior distribution, the Bayesian estimates and standard errors coincide with the classical results.
- However, even in the noninformative case, posterior simulations are useful for predictive inference and model checking.
- The linear regression should be defined as

$$y|\beta, \sigma, X \sim \mathrm{N}(X\beta, \sigma^2 I),$$

where I is the $n \times n$ identity matrix.

Bayesian analysis of classical regression

- The normal regression convenient noninformative prior distribution is uniform on $(\beta, \log \sigma)$

$$p(\beta, \sigma^2|X) \propto \sigma^{-2}.$$

- The joint posterior distribution
  $p(\beta, \sigma^2|X, y) = p(\beta|\sigma^2, X, y)p(\sigma^2|X, y)$
- The conditional posterior distribution of the (vector) parameter β given σ, is the exponential of a quadratic form in β and hence is normal.

$$\beta|\sigma, y \sim \mathrm{N}(\hat{\beta}, V_\beta \sigma^2),$$

where

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y, \\ V_\beta &= (X^T X)^{-1}. \end{aligned}$$

Bayesian analysis of classical regression

- The marginal posterior distribution of σ^2 is a scaled inverse-χ^2

$$p(\sigma^2|X,y) = \frac{p(\beta,\sigma^2|X,y)}{p(\beta|\sigma^2,X,y)} = Inv-\chi^2(n-k,s^2)$$

where

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T(y - X\hat{\beta}).$$

Sampling from the posterior distribution

- Computing $\hat{\beta}$ and V_β
- Computing s^2
- Drawing σ^2 from the scaled inverse-χ^2 distribution
- Drawing β from the multivariate normal distribution.

hint: if $\gamma^2 \sim \chi^2(n-k)$, then $1/\gamma^2 \sim inv\chi^2(n-k)$, then $s^2(n-k)/\gamma^2 \sim$ Scale-$inv\chi^2(n-k, s^2)$

Sampling from the posterior distribution

Computational efficiency is important for large datasets and also with the iterative methods required to estimate several variance parameters simultaneously.
To be computationally efficient, the simulation can be set up as follows,

- Compute the QR factorization, $X = QR$, where Q is an $n \times k$ matrix of orthonormal columns and R is a $k \times k$ upper triangular matrix.
- Compute R^{-1}. R^{-1} is a Cholesky factor of the covariance matrix V_β, since $R^{-1}R^{-T} = (X^T X)^{-1} = V_\beta$.
- Compute $\hat{\beta}$ by solving the linear system $R\hat{\beta} = Q^T y$, using the fact that R is upper triangular.

Sampling from the posterior distribution

- For some large problems involving thousands of data points and hundreds of explanatory variables, even the QR decomposition can require substantial computer storage space and time, and methods such as conjugate gradient, stepwise ascent, and iterative simulation can be more effective.

Posterior predictive simulation

The posterior predictive distribution of unobserved data, $p(\tilde{y}|y)$, has two components of uncertainty:

- the fundamental variability of the model, represented by the variance σ^2 in y not accounted for by $X\beta$
- the posterior uncertainty in β and σ due to the finite sample size of y.
- as $n \to \infty$, the variance due to posterior uncertainty in β, σ^2 decreases to zero, but the predictive uncertainty remains.
- To draw a random sample $\tilde{y}$ from its posterior predictive distribution, we first draw (β, σ) from their joint posterior distribution, then draw $\tilde{y} \sim N(\tilde{X}\beta, \sigma^2 I)$

Analytic form of the posterior predictive distribution

We can gain useful insight by studying the predictive uncertainty analytically

- The expectation of $\tilde{y}$ given σ is

$$\begin{aligned} \mathrm{E}(\tilde{y}|\sigma, y) &= \mathrm{E}(\mathrm{E}(\tilde{y}|\beta, \sigma, y)|\sigma, y) \\ &= \mathrm{E}(\tilde{X}\beta|\sigma, y) \\ &= \tilde{X}\hat{\beta}, \end{aligned}$$

- The variance of $\tilde{y}$ given σ is

$$\begin{aligned} \mathrm{var}(\tilde{y}|\sigma, y) &= \mathrm{E}[\mathrm{var}(\tilde{y}|\beta, \sigma, y)|\sigma, y] + \mathrm{var}[\mathrm{E}(\tilde{y}|\beta, \sigma, y)|\sigma, y] \\ &= \mathrm{E}[\sigma^2 I|\sigma, y] + \mathrm{var}[\tilde{X}\beta|\sigma, y] \\ &= (I + \tilde{X}V_\beta\tilde{X}^T)\sigma^2. \end{aligned}$$

conditional on σ, the posterior predictive variance has two terms: $\sigma^2 I$ sampling variation, and $\tilde{X}V_\beta\tilde{X}^T\sigma^2$ uncertainty about β

Analytic form of the posterior predictive distribution

- Given σ, the future observations have mean does not dependent on σ, and variance is proportional to σ^2
- When average over the marginal posterior distribution of σ^2, the resulting posterior predictive distribution, $p(\tilde{y}|y)$, is multivariate t with center $\tilde{X}\hat{\beta}$, squared scale matrix $s^2(I + \tilde{X}V_\beta\tilde{X}^T)$, and $n - k$ degrees of freedom.
- An advantage of the Bayesian approach is that we can compute, using simulation, the posterior predictive distribution for any data summary, so we do not need to put a lot of effort into estimating the sampling distributions of test statistics.

Regression for causal inference: incumbency and voting

- Observers of legislative elections in the United States have often noted that incumbency— that is, being the current representative in a district—is an advantage for candidates
- Use linear regression to study the advantage of incumbency in elections for the U.S. House of Representatives in the past century.
- Every two years, the members of the U.S. House of Representatives are elected by plurality vote in 435 single-member districts (about 100 to 150 of the district elections are uncontested).

Units of analysis, outcome, and treatment variables

- y_i: the proportion of the vote received by the incumbent party in district i
- R_i: the decision of the incumbent to run for reelection

$$R_i = \begin{cases} 1 & \text{if the incumbent officeholder runs for reelection} \\ 0 & \text{otherwise.} \end{cases}$$

Units of analysis, outcome, and treatment variables

- We define the theoretical incumbency advantage for an election in a single district i as

$$\text{incumbency advantage}_i = y^I_{\text{complete}\,i} - y^O_{\text{complete}\,i},$$

where

$y^I_{\text{complete}\,i}$ = proportion of the vote in district i received by the incumbent *legislator*, if he or she *runs for reelection* against major-party opposition in district i (thus, $y^I_{\text{complete}\,i}$ is unobserved in an open-seat election),
$y^O_{\text{complete}\,i}$ = proportion of the vote in district i received by the incumbent *party*, if the incumbent legislator *does not run* and the two major parties compete for the open seat (thus, $y^O_{\text{complete}\,i}$ is unobserved if the incumbent runs for reelection).

a real election in a single district will reveal only one of these.

Setting up control variables so that data collection is approximately ignorable

- Since incumbency is not a randomly assigned experimental treatment, incumbents and nonincumbents no doubt differ in important ways other than incumbency.
- For example, suppose that incumbents tend to run for reelection in 'safe seats' that favor their party, but typically decline to run for reelection in 'marginal seats' that they have less chance of winning.
- A partial solution is to include the vote for the incumbent party in the previous election as a control variable.

Setting up control variables so that data collection is approximately ignorable

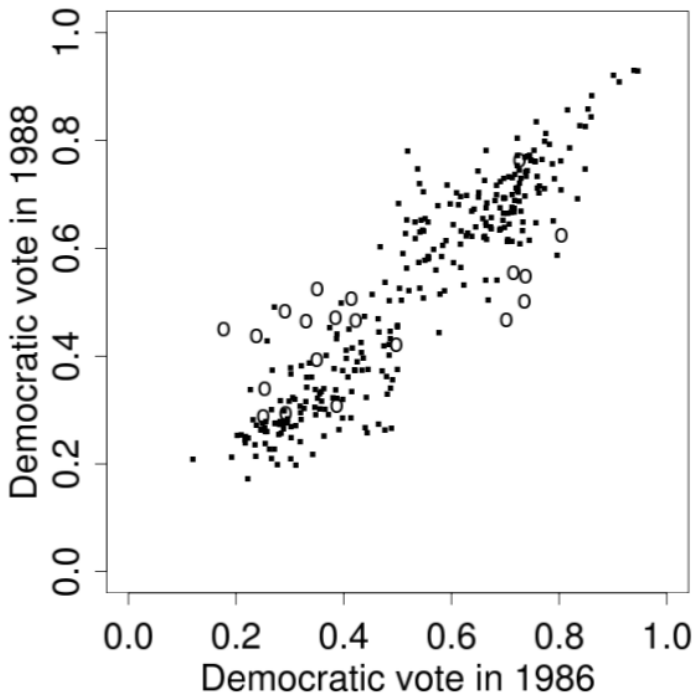

Figure 14.1 *U.S. congressional elections: Democratic proportion of the vote in contested districts in 1986 and 1988. Dots and circles indicate districts that in 1988 had incumbents running and open seats, respectively. Points on the left and right halves of the graph correspond to the incumbent* party *being Republican or Democratic.*

Setting up control variables so that data collection is approximately ignorable

- The strong correlation confirms both the importance of using the previous election outcome as a control variable and the rough linear relation between the explanatory and outcome variables.
- Add variables: Vote proportion in 1986, Incumbent party.

Posterior inference

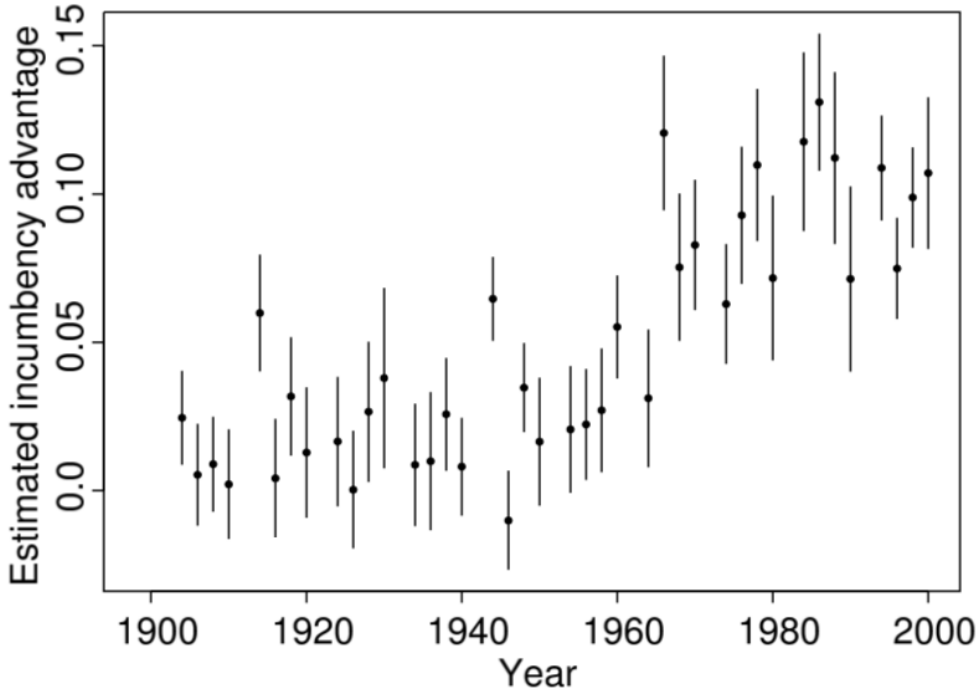

Figure 14.2 *Incumbency advantage over time: posterior median and 95% interval for each election year. The inference for each year is based on a separate regression. As an example, the results from the regression for 1988, based on the data in Figure 14.1, are displayed in Table 14.1.*

Posterior inference

Variable	Posterior quantiles				
	2.5%	25%	median	75%	97.5%
Incumbency	0.084	0.103	0.114	0.124	0.144
Vote proportion in 1986	0.576	0.627	0.654	0.680	0.731
Incumbent party	−0.014	−0.009	−0.007	−0.004	0.001
Constant term	0.066	0.106	0.127	0.148	0.188
σ (residual sd)	0.061	0.064	0.066	0.068	0.071

Table 14.1 *Inferences for parameters in the regression estimating the incumbency advantage in 1988. The outcome variable is the incumbent party's share of the two-party vote in 1988, and only districts that were contested by both parties in both 1986 and 1988 were included. The parameter of interest is the coefficient of incumbency. Data are displayed in Figure 14.1. The posterior median and 95% interval for the coefficient of incumbency correspond to the bar for 1988 in Figure 14.2.*

Posterior inference

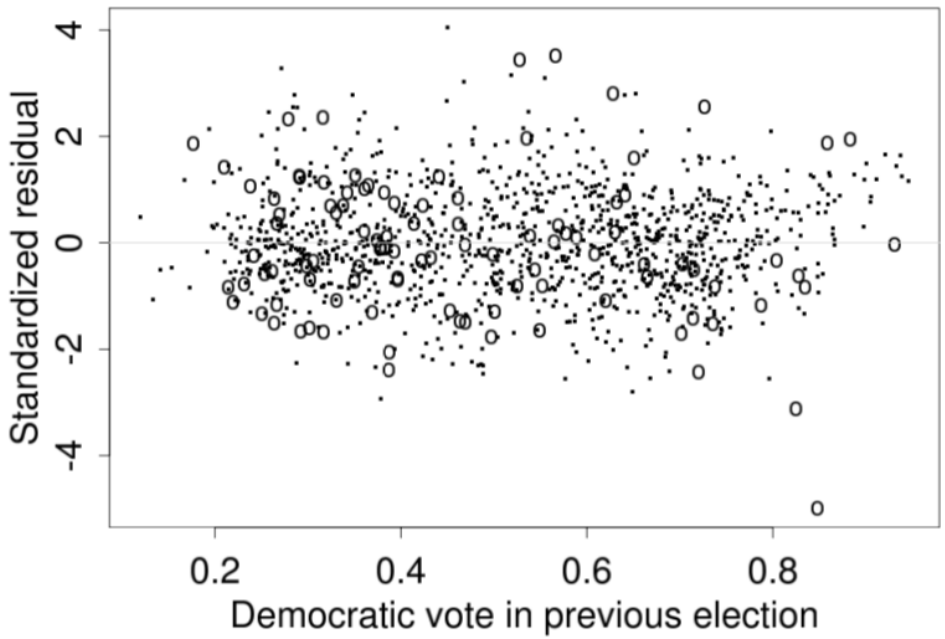

Figure 14.3 *Standardized residuals,* $(y_{it} - X_{it}\hat{\beta})/s_t$*, from the incumbency advantage regressions for the 1980s, vs. Democratic vote in the previous election. (The subscript* t *indexes the election years.) Dots and circles indicate district elections with incumbents running and open seats, respectively.*

Posterior inference

	Observed proportion of outliers	Posterior predictive dist. of proportion of outliers		
		2.5%	median	97.5%
Open seats	$41/1596 = 0.0257$	0.0013	0.0038	0.0069
Incumbent running	$84/10303 = 0.0082$	0.0028	0.0041	0.0054

Table 14.2 *Summary of district elections that are 'outliers' (defined as having absolute (unstandardized) residuals from the regression model of more than 0.2) for the incumbency advantage example. Elections are classified as open seats or incumbent running; for each category, the observed proportion of outliers is compared to the posterior predictive distribution. Both observed proportions are far higher than expected under the model.*