CLASSIFICATION

Part III

Outline

- A comparison of classifiers
- Generalized linear models

A comparison of classifiers

An analytical comparison

$$\max_{k} \log\left\{\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right\}$$

- Logistic regression

$$\log\left\{\frac{p_k(x)}{p_K(x)}\right\} = \beta_{k0} + \sum_j \beta_{kj} x_j$$

- LDA

$$\log\left\{\frac{p_k(x)}{p_K(x)}\right\} = b_{k0} + \sum_j b_{kj} x_j$$

- QDA

$$\log\left\{\frac{p_k(x)}{p_K(x)}\right\} = c_{k0} + \sum_j c_{kj}x_j + \sum_{j,l} c_{kjl}x_j x_l$$

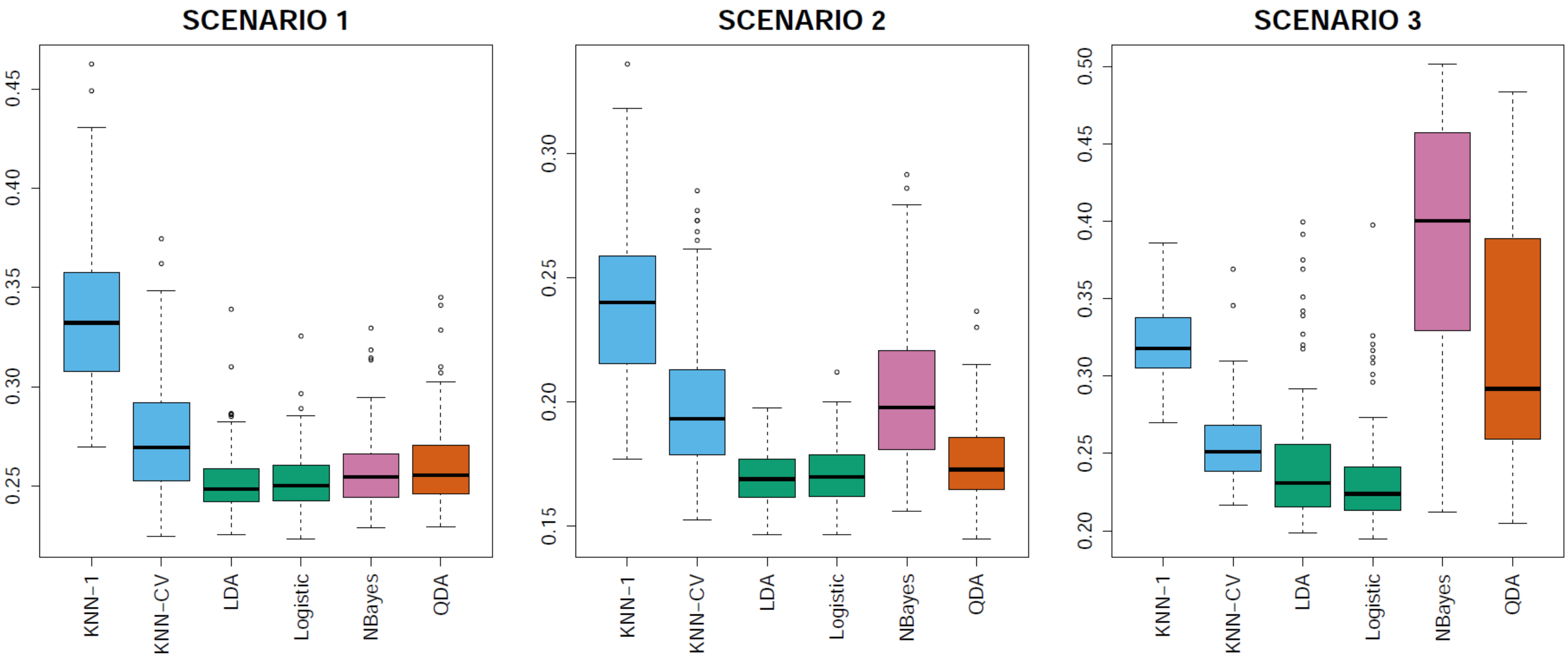
- Naive Bayes

$$\log\left\{\frac{p_k(x)}{p_K(x)}\right\} = a_k + \sum_j g_{kj}(x_j)$$

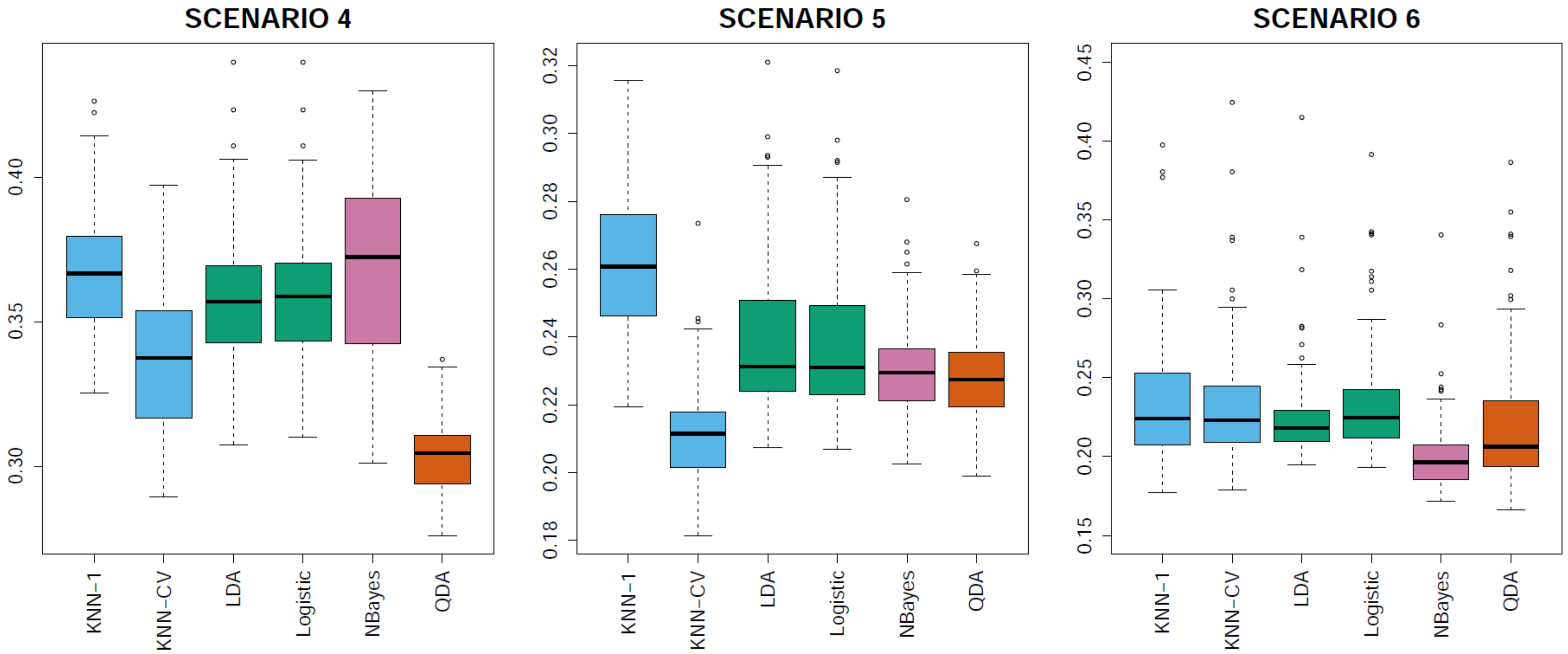
  - An additive model

- LDA is a special case of QDA
- LDA is a special case of naive Bayes
- LDA versus logistic regression
- Neither QDA nor naive Bayes is a special case of the other
- Empirically none of these methods uniformly dominates the others

- *Scenario 1:* $n_1 = n_2 = 20$ and $p = 2$. The observations in each class were uncorrelated normal variables, but with different means
- *Scenario 2:* Details are as in Scenario 1, except that in each class, the two predictors had a correlation of -0.5
- *Scenario 3:* $n_1 = n_2 = 50$. The observations in each class were generated from the t-distribution. The responses were sampled from the logistic function


SCENARIO 1
SCENARIO 2
SCENARIO 3
0.25
0.30
0.35
0.40
0.45
0.15
0.20
0.25
0.30
0.20
0.25
0.30
0.35
0.40
0.45
0.50
KNN−1
KNN−CV
LDA
Logistic
NBayes
QDA

- *Scenario 4:* The data were generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and correlation of -0.5 in the second class
- *Scenario 5:* Within each class, the observations were generated from a normal distribution with uncorrelated predictors. The responses were sampled from the logistic function applied to a non-linear function of the predictors
- *Scenario 6:* $n_1 = n_2 = 6$. The data were generated from a normal distribution with a different diagonal covariance matrix in each class


SCENARIO 4
0.30
0.35
0.40
SCENARIO 5
0.18
0.20
0.22
0.24
0.26
0.28
0.30
0.32
SCENARIO 6
0.15
0.20
0.25
0.30
0.35
0.40
0.45
KNN−1
KNN−CV
LDA
Logistic
NBayes
QDA

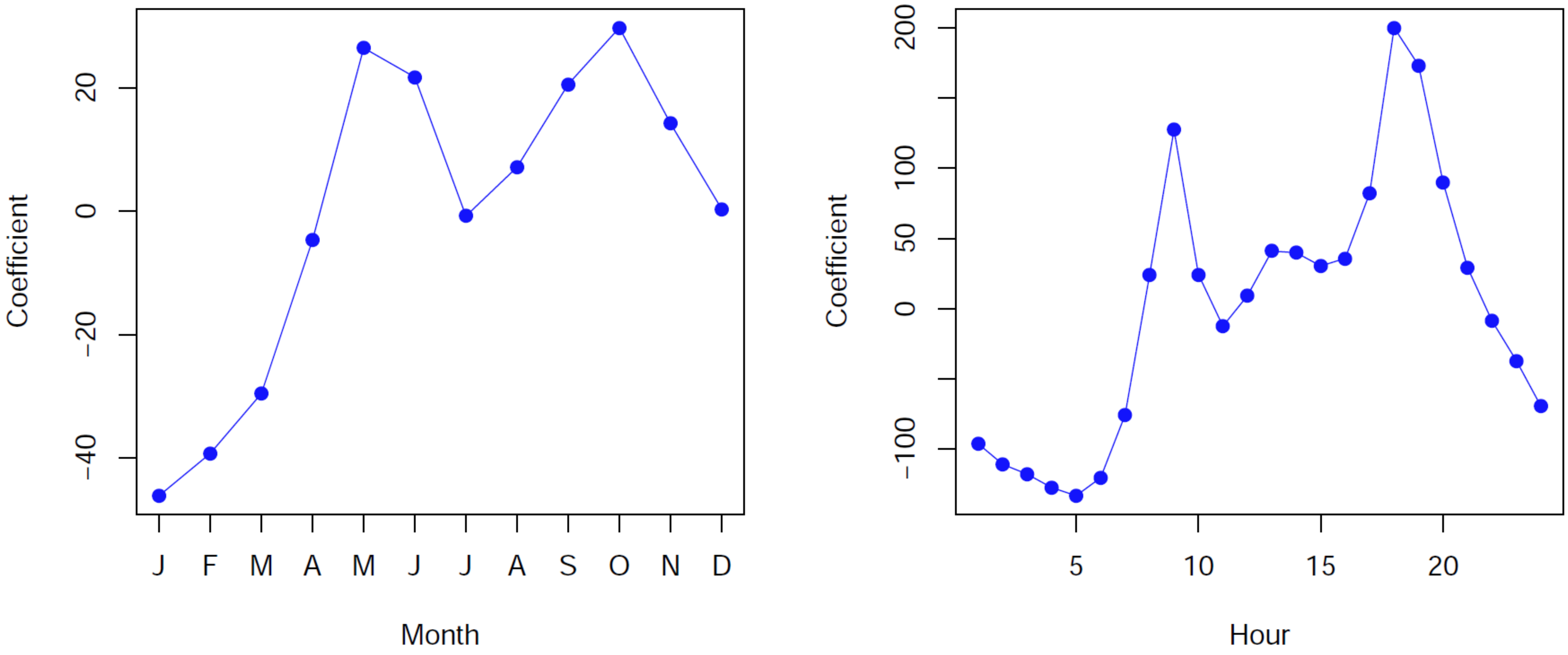
Generalized linear models

Bikeshare data

- Hourly usage of a bike sharing program in Washington, DC (Bikeshare)
- Predict bikers (the number of hourly users) using mnth (month of the year), hr (hour of the day), workingday (an indicator variable), temp (the normalized temperature), and weathersit (clear; misty/cloudy; light rain/light snow; or heavy rain/heavy snow)

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	73.60	5.13	14.34	0.00
`workingday`	1.27	1.78	0.71	0.48
`temp`	157.21	10.26	15.32	0.00
`weathersit[cloudy/misty]`	-12.89	1.96	-6.56	0.00
`weathersit[light rain/snow]`	-66.49	2.97	-22.43	0.00
`weathersit[heavy rain/snow]`	-109.75	76.67	-1.43	0.15

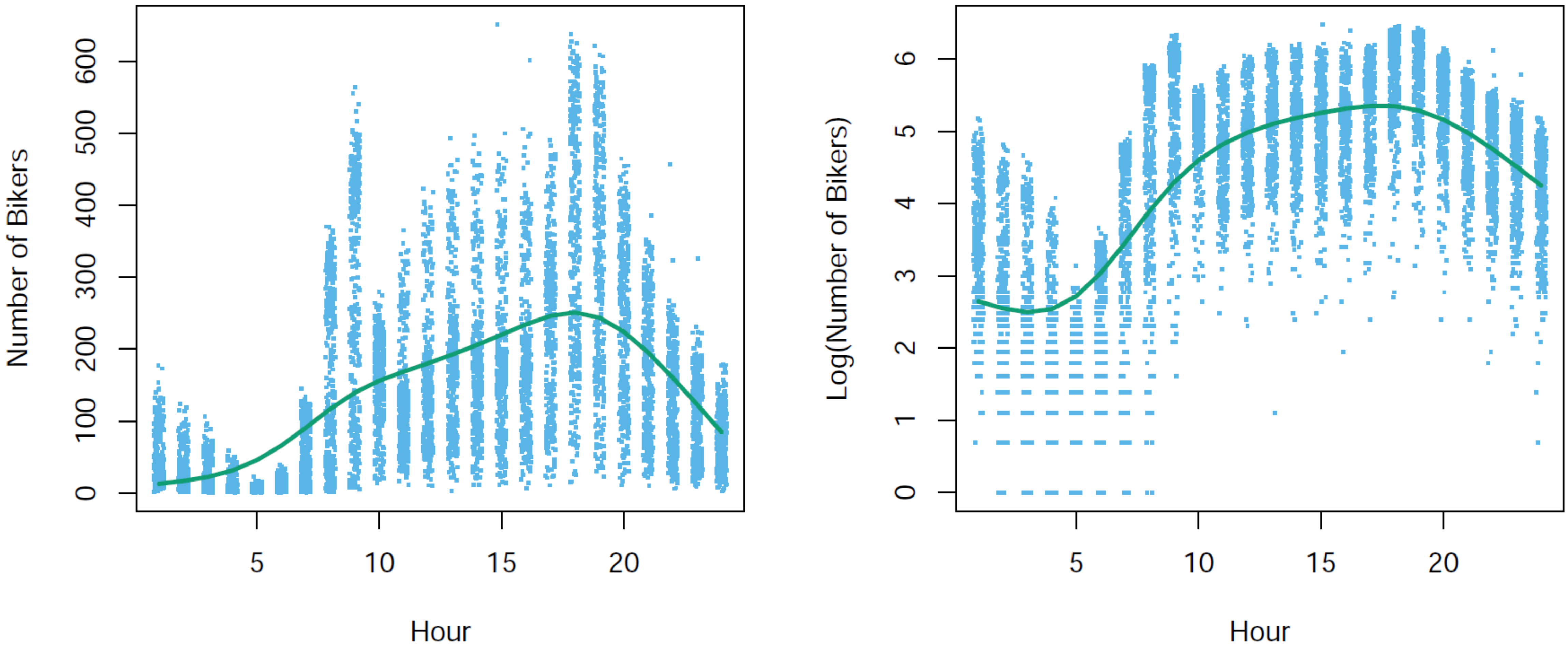
TABLE 4.10. *Results for a least squares linear model fit to predict* `bikers` *in the* `Bikeshare` *data. The predictors* `mnth` *and* `hr` *are omitted from this table due to space constraints, and can be seen in Figure 4.13. For the qualitative variable* `weathersit`*, the baseline level corresponds to clear skies.*



A linear regression model was fit to predict bikers in the Bikeshare data set. Shown are the coefficients associated with mnth and the coefficients associated with hr

Drawbacks of linear regression

- Linear regression predicts a negative number of users during 9.6% of the hours
  - The response takes on non-negative integer values, or *counts*
- The error of a linear model is constant and not a function of the predictors
  - Data heteroscedasticity, or the mean-variance relationship, violates this assumption



Left: bikers is displayed on the y-axis, and hr is displayed on the x-axis. As the mean of bikers increases, so does its variance. Right: The log of bikers is displayed on the y-axis

Poisson regression

- Suppose that the response Y takes on nonnegative integer values
- If Y follows the Poisson distribution, then

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \dots$$

  - $\lambda = E(Y) > 0$ and $Var(Y) = \lambda$

- Poisson regression assumes a Poisson distribution for Y, with the mean response being a non-linear function of the predictors

$$\mathrm{E}(Y|X) = \lambda(X) = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$$

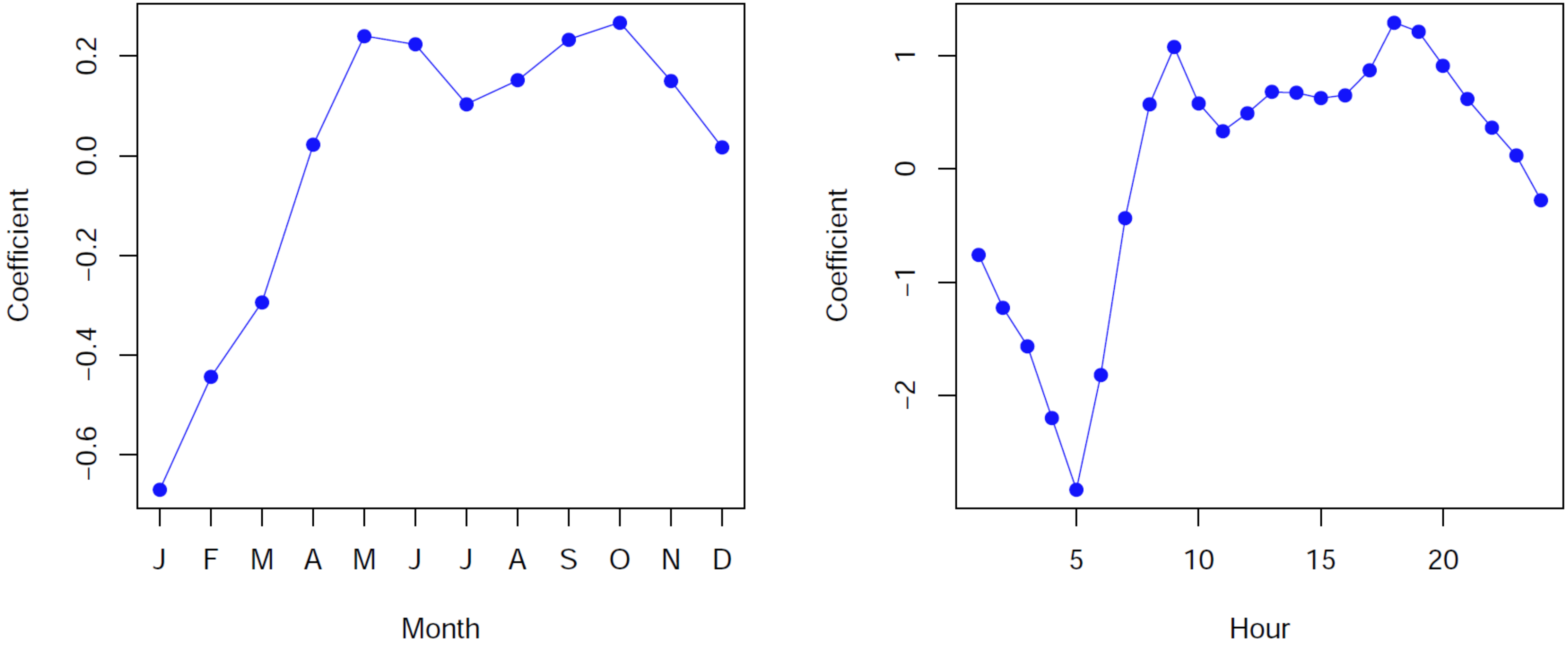
  - A one-unit increase in X_j has a multiplicative impact of e^{β_j} on $\lambda(X)$

- Again, we use the maximum likelihood approach to estimate the parameters
- The likelihood function

$$l(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!}$$

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	4.12	0.01	683.96	0.00
`workingday`	0.01	0.00	7.5	0.00
`temp`	0.79	0.01	68.43	0.00
`weathersit[cloudy/misty]`	-0.08	0.00	-34.53	0.00
`weathersit[light rain/snow]`	-0.58	0.00	-141.91	0.00
`weathersit[heavy rain/snow]`	-0.93	0.17	-5.55	0.00

TABLE 4.11. *Results for a Poisson regression model fit to predict* `bikers` *in the* `Bikeshare` *data. The predictors* `mnth` *and* `hr` *are omitted from this table due to space constraints, and can be seen in Figure 4.15. For the qualitative variable* `weathersit`*, the baseline corresponds to clear skies.*



A Poisson regression model was fit to predict bikers in the Bikeshare data set. Shown are the coefficients associated with mnth and the coefficients associated with hr

Generalized linear models

- Linear, logistic, and Poisson models share some common characteristics
  - Each uses predictors X to predict a response Y
  - Conditioning on X, Y belongs to a certain family of distributions
  - Each models the mean response as a function of the predictors

$$E(Y|X) = g\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}$$

- The Gaussian, Bernoulli and Poisson distributions are all members of the exponential family
  - Other members are the exponential distribution, the Gamma distribution, and the negative binomial distribution

- We can perform a regression by
  - modeling the response as coming from a member of this family
  - transforming the mean of the response so that the transformed mean is a linear function of the predictors
- Any approach that follows this recipe is known as a generalized linear model (GLM)

- All GLMs have three components
  - The random component identifies the response and assumes a probability distribution for it
  - The systematic component specifies the predictors for the model
  - The link function specifies a function of the mean of the response, which the GLM relates to the predictors

$$\eta\{E(Y|X)\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- A GLM generalizes ordinary linear models in two ways
  - It allows the response to have a distribution other than the normal
  - It allows modeling some function of the mean
- The choice of link function is separate from the choice of random component