Introduction to Generalized Linear Models

1. During the Severe Acute Respiratory Syndrome (SARS) outbreak of 2003, some researchers believed that the treatment Ribavirin may be helpful in preventing death due to SARS. Consider a statistical model with the dosage level of Ribavirin as a continuous predictor and fatality ("Death" or "No Death") as a response. This model violates the standard linear regression assumptions.

Answer:

True

---

2. Generalized linear models (GLMs) extend the linear regression framework to allow for non-normal responses, such as counts.

Answer:

True

---

3. Standard linear regression is a type of generalized linear model (GLM).

Answer:

True

---

4. The response of a generalized linear model (GLM) is the random component.

Answer:

True

---

5. The link function in a generalized linear model (GLM) connects the random response to a linear combination of predictor variables.

Answer:

True

---

6. A generalized linear model (GLM) includes:

Answer:

- A random response component.
- A link function.
- A systematic component consisting of a linear combination of discrete or continuous predictors, and fixed parameters.

---

7. Generalized linear models require that the response comes from the exponential distribution.

Answer:

False

Binomial Regression

1. The range (output) of the binomial regression link function $g(p) = \log\left(\frac{p}{1-p}\right)$ is the interval $[0, 1]$.

Answer:

False

---

2. The "probit" link function is $\eta = \Phi^{-1}(p)$, where Φ^{-1} is the inverse of the standard normal cdf, p is the probability of success from the binomial response, and η is the linear predictor.

Answer:

True

---

3. The likelihood function for binomial regression is the joint pmf of the response, but interpreted as a function of the parameters of the model (with the response data fixed).

Answer:

True

---

4. The likelihood function and the log-likelihood function:

Answer:

Are both maximized at the same input/parameter value.

---

5. Let event E have probability p of occurrence. Then the odds in favor of E is defined as: $\frac{p}{1+p}$.

Answer:

False

---

6. Suppose that the probability of contracting a virus v is $p = 0.1$. What are the odds of contracting v?

Answer:

$\frac{1}{9}$

---

7. Consider data on the survival of patients who had undergone surgery for breast cancer. The data consists of a response (survival status after five years) and two predictors (the age of the patient at the time of the operation, and the number of cancerous auxiliary nodes detected):

- x_1: Age of patient in years at time of operation (predictor)
- x_2: Number of cancerous axillary nodes detected (predictor)
- Y_i: Survival status (response): 0 = the patient survived 5 years or longer; 1 = the patient died within 5 years

Suppose that a logistic regression model, with standardized predictors, correctly fits the data:

$$\eta = \beta_0 + \beta_1 z_1 + \beta_2 z_2 = \log\left(\frac{p}{1-p}\right),$$

where p is the probability of a patient surviving 5 years or longer, and

$$z_j = \frac{x_j - \text{mean}(x_j)}{\text{sd}(x_j)} \quad \text{for } j = 1, 2.$$

Which of the following are correct?

Answer:

- β_0 represents the mean log odds of surviving 5 years or longer for a person of (sample) mean age, and with the (sample) mean number of cancerous axillary nodes detected.
- For a fixed number of cancerous axillary nodes detected, a one standard deviation increase in age increases the odds of survival beyond 5 years by a multiplicative factor of e^{β_1}, on average.
- For a fixed number of cancerous axillary nodes detected, a one standard deviation increase in age increases the log-odds of survival beyond 5 years by β_1, on average.

---

8. Consider a logistic regression model that uses data to estimate the probability that a client will default on a monthly credit card payment (defaulting on a payment means that the client fails to pay their bill by the deadline for the month in question.)

- x_1: credit limit in dollars (predictor)
- x_2: dollar amount of the bill statement one month prior (predictor)
- x_3: dollar amount of the bill statement for two months prior (predictor)
- x_4: dollar amount of the payment one month prior (predictor)
- x_5: dollar amount of the payment two months prior (predictor)
- Y_i: default status (response): 0 = the client did not default on the payment for the month in question; 1 = the client did default on the payment for the month in question.

Suppose that a logistic regression model correctly fits the data:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 = \log\left(\frac{p}{1-p}\right),$$

where p is the probability of default.

Answer:

- β_0 represents the mean log-odds of defaulting on a payment for a person with a 0 credit limit, a 0 bill statement for the last two months, and 0 in payments for the last two months.
- e^{β_0} represents the mean odds of defaulting on a payment for a person with a 0 credit limit, a 0 bill statement for the last two months, and 0 in payments for the last two months.

Binomial Regression Inference

1. The maximum likelihood estimator is unbiased.

Answer:

False

---

2. As the sample size n tends to infinity, the distribution of the maximum likelihood estimator becomes $\hat{\theta}_{ML} \sim N(\theta, I^{-1}(\theta))$.

Answer:

True

---

3. Let β_j be the parameter associated with predictor x_j in a binomial regression model. For a reasonably large sample size n, a standard normal "z-test" can be used to test $H_0 : \beta_j = 0 \ \ vs. \ \ H_1 : \beta_j \neq 0$.

Answer:

True

---

4. Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \theta)$, and let $\hat{\theta}$ be the maximum likelihood estimator of θ. Then

$$\left(\hat{\theta}_{ML} - \frac{1.96}{\sqrt{nI(\theta)}}, \hat{\theta}_{ML} + \frac{1.96}{\sqrt{nI(\theta)}}\right)$$

is an approximate 95% confidence interval for θ.

Answer:

True

---

5. Goodness of fit metrics—such as the residual deviance—are only useful for binomial regression with a relatively large number of trials (e.g., $n > 5$).

Answer:

True

---

6. Consider a logistic regression fit to an independent response $Y_i \sim Binomial(1, p)$ and a single predictor variable x. The linear predictor is:

$$\eta_i = \beta_0 + \beta_1 x_i.$$

Test $H_0 : \beta_1 = 0 \ vs \ H_1 : \beta_1 \neq 0$ by computing the appropriate p-value, rounded to the hundredths place.

Coefficients	Estimate	Std. Error
(Intercept)	-1.2467	0.6347
x	1.4224	1.1541

Answer:

0.22

You can estimate the t-test (t-stat) as:

$$t_{\text{stat}} = \frac{\text{Estimate}}{\text{Std. Error}}$$

Computing p-value when t-stat or z-score is given:

$$p_{\text{value}} = 2 \times (1 - \text{pnorm}(|t_{\text{stat}}|))$$

---

7. Consider a logistic regression fit to an independent response $Y_i \sim Binomial(1, p)$ and a single predictor variable x. The linear predictor is:

$$\eta_i = \beta_0 + \beta_1 x_i.$$

Use maximum likelihood theory to construct an approximate 95% confidence interval for β_0. Round all values to the hundredths place.

Coefficients	Estimate	Std. Error
(Intercept)	-1.2467	0.6347

Coefficients	Estimate	Std. Error
x	1.4224	1.1541

Answer:

(-2.48, -0.02)

Poisson Regression Basics

1. Consider the following modeling scenario: For a single year, researchers measure the number of motor vehicle accidents that result in death in each of the 50 states in the United States. They also record each state's speed limit laws over the same time period, and each state's population. They are interested in the following research question: Are the number of motor vehicle deaths in a given state related to a state's speed limit laws?

Based on the information given, a reasonable first attempt at answering this question would include:

Answer:

A Poisson regression model without an offset term.

---

2. Consider the following modeling scenario: researchers would like to construct a model that can predict the number of times an individual would be admitted to a hospital (y_i). The covariate class—the set of predictors—might include age, gender, and other health conditions (e.g., heart conditions, diabetes). Let λ_i be the average number of times individual i was admitted to the hospital. Individuals were observed for different periods of time (e.g., some for one year, others for two years).

The correct link function for this model is $\eta_i = \log(\lambda_i)$, where η_i is the linear predictor.

Answer:

False

---

3. For Poisson regression with $Y_i \sim Poisson(\lambda_i)$,

$$\exp(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}) = \lambda_i.$$

Answer:

True

---

4. e^{β_0} is the mean of the response when each predictor is set to zero.

Answer:

True

---

5. e^{β_j} is the amount added to the mean of the response for a one unit increase in $x_{i,j}$, fixing (or adjusting for) all other predictors.

Answer:

False

---

6. Consider the following modeling scenario: For an entire year, researchers collect data on fraudulent credit card transactions, including whether or not a particular transaction was ruled as fraudulent, the amount of each purchase, the distance from the card holder's zip code, whether the purchase was online or not, and several other variables. The goal is to use this data to construct a model that will help flag future purchases as potentially fraudulent.

Based on the information given, a reasonable first attempt at a model would be:

Answer:

A binomial regression, with the fraudulent/not fraudulent variable as the response and all other variables as predictors.

---

7. Consider a model that attempts to explain the number of awards earned by students at a high school in a year based on their math final exam score and the type of program that they are enrolled in. The categorical predictor variable has three levels indicating the type of program in which the student is enrolled. The categorical predictor levels are "Remedial", "Standard," and "Honors". Here's some output from a Poisson regression.

```
glm(formula = num_awards ~ prog + math, family = "poisson", data = p)
```

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.2471	0.6585	-7.97	1.6e-15 ***

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
progStandard	1.0839	0.3583	3.03	0.0025 **
progHonors	0.3698	0.4411	0.84	0.4018
math	0.0702	0.0106	6.62	3.6e-11 ***

Which of the following statements are correct? (Choose all that apply.)

Answer:

- A one-unit increase in a student's math final exam score is associated with a multiplicative change of approximately 1.07 in the number of awards, adjusting for program type.
- The average number of awards for a student in the "Standard" program and with a zero math final exam score is approximately 0.016.

8. Like standard linear regression, we can estimate the Poisson regression model parameters using least squares.

Answer:

False

9. Consider a model that attempts to explain the number of awards earned by students at a high school in a year based on their math final exam score and the type of program that they are enrolled in. The categorical predictor variable has three levels indicating the type of program in which the student is enrolled. The categorical predictor levels are “Remedial”, “Standard,” and “Honors.” Here's some output from a Poisson regression:

```
glm(formula = num_awards ~ prog + math, family = "poisson", data = p)
```

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.2471	0.6585	-7.97	1.6e-15 ***
progStandard	1.0839	0.3583	3.03	0.0025 **
progHonors	0.3698	0.4411	0.84	0.4018
math	0.0702	0.0106	6.62	3.6e-11 ***

What is the expected number of awards for a student who is in the honors program and whose math final exam score is set to the mean of the sample: math = 53? Round to the nearest hundredth place.

Answer:

0.31

Poisson Regression Inference and Goodness of Fit

1. The "deviance" of a Poisson regression model is -2 times the log likelihood of the Poisson regression model evaluated at the maximum likelihood estimates.

Answer:

True

2. The null deviance is the deviance for the model with just an intercept term and a single predictor.

Answer:

False

3. The saturated model is the model that includes all of the predictors in the dataset.

Answer:

False

4. The residual deviance can be used to test the hypotheses:

- H_0: The model with p predictors fits well enough.
- H_1: The model with p predictors does not fit well enough.

Answer:

True

5. A plot of the deviance residuals against the linear predictor (η_i) can provide evidence of a lack of fit of a Poisson regression model.

Answer:

True

6. Consider a model that attempts to explain the number of awards earned by students at a high school in a year based on their math final exam score and the type of program that they are enrolled in. The categorical predictor variable has three levels indicating the type of program in which the student is enrolled. The categorical predictor levels are ”Remedial”, “Standard,” and ”Honors.” Here's some output from a Poisson regression.

Consider fitting two models, one with both predictors, and one with just math final exam score as a predictor.

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
198	204			
196	189	2	14.6	0.00069 ***

Answer:

- The test performed was a χ^2 test.
- The hypotheses under consideration are:
  - H_0: The model with just math score fits well enough.
  - H_1: The model with just math score does not fit well enough.
- The conclusion of this test is that the program variable is statistically significant.

7. Consider a Poisson regression model with the response of the total number of cyclist counts at Manhattan Bridge in a 24-hour period. But suppose that cyclist counts on this bridge are such that, if an individual cycles over the Manhattan Bridge on a particular day, that individual will be more likely to cycle over the Manhattan Bridge the following day. So, an event occurring on one day impacts the probability of the event occurring on the next day. The distribution of the number of cyclists over the Manhattan Bridge will then be overdispersed with respect to the Poisson model.

Answer:

True

8. Which of the following are potential causes of overdispersion?

Answer:

- Outliers.
- A dependent response variable.
- A missing predictor variable.
- Having many zeros recorded for the response.

Nonparametric Regression: Theory

2. Parametric modeling is more efficient when the relationship between variables is unknown.

Answer:

False

3. Binomial regression is a type of nonparametric model.

Answer:

False

4. In the context of kernel estimation, the smaller the bandwidth, the rougher the fit.

Answer:

True

5. Consider the following modeling scenario: Ethanol fuel was burned in a single-cylinder engine. For various settings of the engine compression, the emissions of nitrogen oxides were recorded:

- **NOx**: Concentration of nitrogen oxides (NO and NO_2) in micrograms/J.

- **C**: Compression ratio of the engine.

Researchers would like to understand how NOx is related to C. It is quite plausible that the relationship is nonlinear.

Based on the information given, a reasonable first attempt at answering this question would include:

Answer:

- A loess model.
- A smoothing spline.
- A kernel regression.

---

6. The parametric approach assumes far less about the form of the model and so it is less liable to make major mistakes that result in bias.

Answer:

False

---

7. Nonparametric models often don't have a formulaic way of describing the relationship between the predictors and response.

Answer:

True

---

8. In smoothing spline regression, we estimate our model f by minimizing the mean squared error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2.$$

Answer:

False

---

9. In the context of smoothing spline regression, as $\lambda \to \infty$, the fit converges to $\hat{f}(x_i) = 0$ for all i.

Answer:

True

---

10. Since loess models rely on theory for weighted regression, it is not possible to quantify uncertainties in the model, in much the same way as is done for, e.g., linear regression.

Answer:

False

---

11. Disadvantages of the loess fit include:

Answer:

- Loess can be computationally expensive.
- Loess is not as easy to interpret as standard linear regression.
- Loess requires fairly large, densely sampled data in order to produce good models.

Generalized Additive Models: Basics

1. When compared to some common machine learning techniques, such as random forests, generalized additive models have the advantage of clearly showing the contribution of each predictor to the response.

Answer:

True

---

2. Generalized additive models can be thought of as a way to estimate nonlinear relationships between a response and several predictors simultaneously.

Answer:

True

---

3. Generalized additive models strike a nice balance between the interpretable, yet biased, linear model, and the extremely flexible, "black

box" machine learning algorithms.

Answer:

True

4. Which of the following are additive models?

Answer:

- $f(x_1, x_2) = \pi + \beta_1 e^{x_1} - 5x_2$
- $f(x_1, x_2, x_3) = \beta_0 + \beta_1 \log(x_1 x_3^{\beta_2}) - \beta_2 5x_3$ (Log laws!)
- $f(x_1, x_2, x_3) = \beta_0 x_1 + \beta_1 x_1 + \cos(\pi x_2) + \beta_3 x_2^2 + \sin^2(x_3)$

5. Additive models will work well when strong interactions between predictors exist.

Answer:

False

6. Generalized additive models have trouble incorporating non-normal (e.g., binomial) responses.

Answer:

False

7. Generalized additive models are typically more biased than standard linear regression models.

Answer:

False

8. that a response y is related nonlinearly to a (continuous) predictor x_1, linearly to a (continuous) predictor x_2, and linearly to a three-level factor x_3. Then:

Family: gaussian
Link function: identity

Formula:
y ~ s(x1) + x2 + x3

Parametric Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.49974	0.36072	4.158	9.25e-05 ***
x2	2.81405	0.12020	23.412	< 2e-16 ***
x3B	1.02847	0.02579	39.875	< 2e-16 ***
x3C	1.94796	0.02498	77.973	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

Term	edf	Ref.df	F	p-value
s(x1)	8.283	8.853	600.9	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- **R-sq.(adj)** = 0.994
- **Deviance explained** = 99.5%
- **GCV** = 0.0091693
- **Scale est.** = 0.0077614
- **n** = 80

Answer:

- For a one-unit increase in x_2, the mean change in y is approximately 2.81, adjusting for other predictors.
- The mean change in y for one-unit increase in x_1, adjusting for other predictors, depends on the value of x_1.

Generalized Additive Models: Inference and Data Analysis

1. In a generalized additive model, if a smooth term has an effective degrees of freedom close to 1, then that term should enter linearly into the model.

Answer:

True

---

2. The tests in the summary of the `gam()` function (in the `mgcv` package) associated with the smooth terms are (approximately) F-tests that test the hypothesis that the given smooth term is zero vs nonzero.

Answer:

True

---

3. The adjusted R^2 is reported as a percentage, and is equal to $R^2 = 1 - \frac{RD}{ND}$, where RD is the residual deviance and ND is the null deviance.

Answer:

False

---

4. n the context of generalized additive models, the adjusted R^2 is reasonable to use for model comparisons.

Answer:

True

---

5. The `trees` data frame has 31 observations on 3 variables:

6. **Girth**: Tree diameter in inches

7. **Height**: Height in ft

8. **Volume**: Volume of timber in cubic ft

Consider a GAM fit to the data, where Volume is the response and Girth and Height are predictors.

Family: Gamma
Link function: log

Formula:
Volume ~ s(Height) + s(Girth)

Parametric coefficients:

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.27570	0.01492	219.6	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

Term	edf	Ref.df	F	p-value
s(Height)	1.000	1.000	31.32	3.92e-06 ***
s(Girth)	2.422	3.044	219.28	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- **R-sq.(adj)** = 0.973
- **Deviance explained** = 97.8%
- **GCV** = 0.0080824
- **Scale est.** = 0.006899
- **n** = 31

Height should enter the model parametrically.

Answer:

True

6. The trees data frame has 31 observations on 3 variables.

7. **Girth**: Tree diameter in inches

8. **Height**: Height in feet

9. **Volume**: Volume of timber in cubic feet

Consider a GAM fit to the data, where Volume is the response and Girth and Height are predictors.

Family: Gamma
Link function: log

Formula:
Volume ~ s(Height) + s(Girth)

Parametric coefficients:

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.27570	0.01492	219.6	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

Term	edf	Ref.df	F	p-value
s(Height)	1.000	1.000	31.32	3.92e-06 ***
s(Girth)	2.422	3.044	219.28	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- **R-sq.(adj)** = 0.973
- **Deviance explained** = 97.8%
- **GCV** = 0.0080824
- **Scale est.** = 0.006899
- **n** = 31

The small p-value associated with Girth suggests that it should enter the model nonparametrically.

Answer:

False