Module 8: Generalized linear regression

Yuan Tian

07/23/2023

Outline

In this module, we will review generalized linear regression.

Exponential family

The Gaussian, Binomial and Poisson distributions are special cases of `exponential family` which assumes the following density function

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

- θ: canonical parameter
- ϕ: dispersion parameter

Gaussian as a special case of exponential family

Assume that $y \sim \mathbf{N}\left(\mu, \sigma^2\right)$. Then

$$
\begin{aligned}
f(y;\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln\left(2\pi\sigma^2\right)\right)\right\} \\
&= \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}
\end{aligned}
$$

- $\theta = \mu, \phi = \sigma^2$
- $a(\phi) = \phi; (\theta) = \theta^2/2; c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \ln(2\pi\phi)\right)$

Binomial as a special case of exponential family

Assume that $z \sim B(m, \pi)$. Define the rate $y = z/m$. Then

$$\begin{aligned} f(y;\theta,\phi) &= \exp\left\{ z \ln \frac{\pi}{1-\pi} + m \ln(1-\pi) + \ln \begin{pmatrix} m \\ z \end{pmatrix} \right\} \\ &= \exp\left\{ m \left(\frac{z}{m} \operatorname{logit}(\pi) + \ln(1-\pi) \right) + \ln \begin{pmatrix} m \\ mz/m \end{pmatrix} \right\} \\ &= \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right\} \end{aligned}$$

- $\operatorname{logit}(\pi) = \ln(\pi/(1-\pi))$
- $\theta = \operatorname{logit}(\pi) \to \pi = e^\theta / \left(1 + e^\theta\right), \phi = 1$
- $a(\phi) = 1/m, b(\theta) = -\ln(1-\pi) = \ln\left(1 + e^\theta\right)$

$$c(y,\phi) = \ln \begin{pmatrix} m \\ my \end{pmatrix}$$

Poisson as a special case of exponential family

Assume that $y \sim P(\mu)$. Then

$$\begin{aligned} f(y;\theta,\phi) &= \mu^y \exp(-\mu)/y! \\ &= \exp\{y \ln \mu - \mu - \ln y!\} \\ &= \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right\} \end{aligned}$$

- $\theta = \ln \mu$
  $a(\phi) = 1$
- $b(\theta) = e^\theta$
  $c(y,\phi) = -\ln y!$

Moment generating function

Assume that

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

The moment generating function of y is

$$\begin{aligned}
M(t) \triangleq \exp(ty) &= \int \exp\left\{ty + \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} dy \\
&= \int \exp\left\{\frac{y[\theta + ta(\phi)] - b(\theta)}{a(\phi)} + c(y,\phi)\right\} dy \\
&= \int \exp\left\{\frac{y\theta' - b\left(\theta' - ta(\phi)\right)}{a(\phi)} + c(y,\phi)\right\} dy \\
&= \exp\left\{\frac{b\left(\theta'\right) - b\left(\theta' - ta(\phi)\right)}{a(\phi)}\right\} \int \exp\left\{\frac{y\theta' - b\left(\theta'\right)}{a(\phi)} + c(y,\phi)\right\} dy \\
&= \exp\left\{\frac{b(\theta + \mathrm{ta}(\phi)) - b(\theta)}{a(\phi)}\right\}
\end{aligned}$$

Mean and variance

Since

$$\ln M(t) = \frac{b(\theta + \mathrm{ta}(\phi)) - b(\theta)}{a(\phi)}$$

We have

$$\begin{aligned} \mathrm{E}(y) &= (\ln M(t))'\big|_{t=0} = b'(\theta) \\ \mathrm{Var}(y) &= (\ln M(t))''\big|_{t=0} = a(\phi)b''(\theta) \end{aligned}$$

Examples of means and variances

Binomial: $a(\phi) = 1/m, b(\theta) = \ln\left(1 + e^{\theta}\right)$

$$\mu \triangleq \mathrm{E}(y) = b'(\theta) = \frac{e^{\theta}}{1 + e^{\theta}} = \pi$$
$$\mathrm{Var}(y) = a(\phi)b''(\theta) = \pi(1-\pi)/m = \mu(1-\mu)/m$$

Variance depends on the mean! It reaches maximum at $\mu = 1/2$.

Poisson: $a(\phi) = 1, b(\theta) = e^{\theta}$

$$\mathrm{E}(y) = b'(\theta) = e^{\theta} = \mu$$
$$\mathrm{Var}(y) = a(\phi)b''(\theta) = \mu$$

Variance = mean!

Variance function

In general,

$$\mathrm{Var}(y) = a(\phi)b''(\theta)$$

- $a(\phi)$ does not depend on μ
- $b''(\theta)$ depends on μ only

We define

$$V(\mu) \triangleq b''(\theta)$$

as the `variance function`.

- Gaussian: $V(\mu) = 1$
- Binomial: $V(\mu) = \mu(1-\mu)$
- Poisson: $V(\mu) = \mu$

Systematic component

Same as the LM, for covariates $x_1, \cdots, x_p$, a `linear predictor` is

$$\eta \triangleq \sum_{j=1}^{p} \beta_j x_j$$

Link function

A link function g describes how the mean μ depends on the `linear predictor` $\eta : \eta = g(\mu)$. We assume that g is `monotone` (therefore it is one-to-one) and differentiable.

A canonical link is the link function such that $\eta = \theta$.

Link function

Why do we need more complicated link functions than the simple identity link function? In other words, why can't we model the mean directly as a function of covariates using an additive linear function?

- the linear predictor $\eta = \sum_{j=1}^{p} \beta_j x_j$ can take any value in $(-\infty, \infty)$.
- the link function is to define the scale over which the systematic component is additive.

Common link functions for Binomial data

Binomial: assume that $0 < \mu < 1$

1. logit: $g(z) = \ln \frac{z}{1-z}$. It is the canonical link since $\eta = \ln \frac{\mu}{1-\mu} = \theta$
2. probit: $g(z) = \Phi^{-1}(z)$, where Φ is the standard Gaussian CDF. Then $\eta = \Phi^{-1}(\mu)$
3. complementary log-log: $g(z) = \ln\{-\ln(1-z)\}$, i.e. $\eta = \ln\{-\ln(1-\mu)\}$

Comparison of links for Binomial data

- The logit and the probit links are almost linearly related i the middle (specifically when $.1 \leq \mu \leq .9$). For this reason it is usually difficult to discriminate between these two links on the grounds of goodness-of-fit.
- The complementary log-log link approaches to infinity slower and approaches to minus infinity faster than the logit and probit links.
- Different link functions can be motivated from latent variable models. Logit, probit and complementary log-los links correspond to logistic, Gaussian and extreme value CDFs for the latent variable.
- The choice of link is usually made based on assumptions derived from physical knowledge or simple convenience.
- The logit link is the most popular choice because it is the canonical link and parameters have nice interpretations based on odds ratio. No simple interpretations for other links.

Interpretations of parameters in a LM - review

Consider the following simple LM

$$y = \beta_0 + \beta_1 x + \epsilon.$$

The parameter β_1 has the simple interpretation that the effect of a unit change in x is to increase the expected response by β_1.

Interpretations of parameters - Binomial with logit link

Now for Binomial data with logit link, consider a similar model

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$$

or equivalently,

$$\text{odds}(x) \triangleq \frac{\pi}{1-\pi} = \exp\left(\beta_0 + \beta_1 x\right)$$

Then

$$\text{odds}(x+1) = \text{odds}(x) \times \exp\left(\beta_1\right).$$

Interpretation of β_1 : the effect of a unit change in x is to increase the odds by a factor $\exp\left(\beta_1\right) \cdot \exp\left(\beta_1\right)$ is often called odds ratio. When there are multiple independent variables, the interpretation remains the same with other independent variables being fixed.

Link function for Poisson data

Poisson: assume that $\mu > 0$. The canonical link is

$$g(z) = \ln z$$

That is

$$\eta = \ln \mu$$

Summary

A GLM has three components: - Random component: exponential family

$$y \sim \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

- Systematic component: linear predictor $\eta = \sum_{j=1}^{p} \beta_j x_j$
- Link g : depends on the type of data, e.g. logit link for Binomial data and log link for Poisson data

Two key characteristics of a LM are

- linear dependence on unknown parameters
- additive random error

Neither one is true for the GLM (except for the Gaussian case). Therefore, we don't write the model in the form of `observation $=$ linear predictor + random error`

Some common GLMs

	Gaussian	Binomial	Poisson
Notations	$\mathrm{N}\left(\mu, \sigma^2\right)$	$B(m, \pi)/m$	$P(\mu)$
Range of y	$(-\infty, \infty)$	$\{0, 1\}$	$\{0, 1, 2, \cdots\}$
Dispersion parameter	σ^2	1	1
Canonical parameter θ	μ	logit (π)	ln μ
Canonical link	identity	logit	ln
Mean $\mu(\theta)$	μ	$e^\theta / \left(1 + e^\theta\right)$	e^θ
Variance function $V(\mu)$	1	$\mu(1-\mu)$	μ

Maximum likelihood estimation of parameters

For a GLM, the log-likelihood of a single observation is (subscript i omitted for now)

$$l = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

Using the chain rule,

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta} \frac{d\theta}{d\mu} \frac{d\mu}{d\eta} \frac{\partial \eta}{\partial \beta_j}$$

We need to compute each component.

Computation of components

From the fact that $b'(\theta) = \mu$, we have

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} = \frac{y - \mu}{a(\phi)}$$

Again, using the fact that $b'(\theta) = \mu$,

$$\frac{d\mu}{d\theta} = b''(\theta) = V$$

where V is the variance function. Therefore,

$$\frac{d\theta}{d\mu} = \frac{1}{V}$$

Computation of components

Since $\eta = g(\mu)$, we have

$$\frac{d\mu}{d\eta} = \frac{1}{g'(\mu)}$$

Since $\eta = \sum_{j=1}^{p} \beta_j x_j$, we have

$$\frac{\partial \eta}{\partial \beta_j} = x_j$$

First derivative of the log-likelihood

Putting the pieces together, we have

$$
\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \frac{y-\mu}{a(\phi)}\frac{1}{V}\frac{1}{g'(\mu)}x_j \\
&= \frac{\varpi}{a(\phi)}(y-\mu)g'(\mu)x_j
\end{aligned}
$$

where

$$\varpi = \frac{1}{V\left(g'(\mu)\right)^2}$$

Full likelihood

The log-likelihood of all n observations is

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \left\{ \frac{y_i\theta - b(\theta)}{a_i(\phi)} + c\left(y_i, \phi\right) \right\}$$

Note: we allow a different function $a_i(\phi)$ for each observation.

The score statistic

$$\begin{aligned} u_j &\triangleq \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\varpi_i}{a_i(\phi)} \left(y_i - \mu_i\right) g'\left(\mu_i\right) x_{ij} \end{aligned}$$

Matrix form

$$
\begin{aligned}
\boldsymbol{u} &\triangleq \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix} = \frac{\partial l}{\partial \beta} \\
&= \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \vdots & \vdots \\ x_{1p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \frac{\varpi_1}{a_1(\phi)} & & \\ & \ddots & \\ & & \frac{\varpi_n}{a_n(\phi)} \end{pmatrix} \begin{pmatrix} (y_1 - \mu_1)\, g'(\mu_1) \\ \vdots \\ (y_n - \mu_n)\, g'(\mu_n) \end{pmatrix} \\
&= X^T W \begin{pmatrix} (y_1 - \mu_1)\, g'(\mu_1) \\ \vdots \\ (y_n - \mu_n)\, g'(\mu_n) \end{pmatrix}
\end{aligned}
$$

where

$$
W \triangleq \begin{pmatrix} \frac{\varpi_1}{a_1(\phi)} & & \\ & \ddots & \\ & & \frac{\varpi_n}{a_n(\phi)} \end{pmatrix}
$$

Score equation

We want to estimate β by solving the score equation

$$\boldsymbol{u} = X^T W \begin{pmatrix} (y_1 - \mu_1)\, g'(\mu_1) \\ \vdots \\ (y_n - \mu_n)\, g'(\mu_n) \end{pmatrix} = \mathbf{0}$$

Numerical solution

In the score equation, the weight matrix W^{-1} is unknown and may depend on β. Therefore, the score equation is a non-linear system of equations and can't be solved analytically. We need to compute them numerically using an iterative scheme. A common approach is the Newton-Raphson procedure.

Newton-Raphson procedure

Suppose that we want to find the maximizer of a function $f(z)$. Using Taylor expansion, we approximate f near z_0 by

$$f(\boldsymbol{z}) \approx f(\mathbf{z}_0) + \boldsymbol{u}^T(\boldsymbol{z} - \mathbf{z}_0) + \frac{1}{2}(\boldsymbol{z} - \mathbf{z}_0)^T H (\boldsymbol{z} - \mathbf{z}_0) \triangleq h(\boldsymbol{z})$$

- $\boldsymbol{u} = (\partial f / \partial \mathbf{z})|_{\boldsymbol{z}=\mathbf{z}_0}$: gradient
- $H = \left(\partial^2 f / \partial \mathbf{z} \partial \mathbf{z}^T\right)|_{\mathbf{z}=\mathbf{z}_0}$: Hessian

Note that $h(z)$ is a quadratic function since $\boldsymbol{u}$ and H are fixed. We can maximize $h(z)$ by solving

$$\frac{\partial h(\boldsymbol{z})}{\partial \boldsymbol{z}} = \boldsymbol{u} + H(\boldsymbol{z} - \mathbf{z}_0) = \mathbf{0}.$$

The maximizer is

$$\boldsymbol{z} = \boldsymbol{z}_0 - H^{-1}\boldsymbol{u}$$

Newton-Raphson procedure

Newton-Raphson Algorithm

(1) Select a starting point $\boldsymbol{z}^{(0)}$

(2) At iteration $l+1$,

$$\boldsymbol{z}^{(l+1)} = \boldsymbol{z}^{(l)} - \left(H^{(l)}\right)^{-1} \boldsymbol{u}^{(l)}$$

or equivalently, solve the equation

$$H^{(l)} z^{(l+1)} = H^{(l)} z^{(l)} - u^{(l)}$$

(3) Iterate the second step until convergence

Binomial cases

We have independent observations

$$z_i \sim B(m_i, \pi_i), \quad y_i = z_i/m_i, \quad i = 1, \cdots, n$$

with density function

$$f(y_i; \theta, \phi) = \exp\left\{ m_i \left[y_i \operatorname{logit}(\pi_i) + \ln(1 - \pi_i) \right] + \ln \begin{pmatrix} m_i \\ m_i y_i \end{pmatrix} \right\}$$

Since $a_i(\phi) = 1/m_i$, we have

$$w_i = m_i$$

Since $V_i = \pi_i (1 - \pi_i)$, we have

$$\varpi_i = 1/\left(\pi_i (1 - \pi_i) \left[g'(\mu_i) \right]^2 \right).$$

Binomial cases

Furthermore, for the logit link,

$$g(\mu) = \ln \frac{\mu}{1-\mu}.$$

Then

$$g'(\mu) = \frac{1}{\mu(1-\mu)}$$

Thus,

$$\varpi_i = \pi_i \left(1 - \pi_i\right)$$

and

$$W^{(l)} = \begin{pmatrix} m_1 \pi_1^{(l)} \left(1 - \pi_1^{(l)}\right) & & \\ & \ddots & \\ & & m_n \pi_n^{(l)} \left(1 - \pi_n^{(l)}\right) \end{pmatrix}.$$

Goodness-of-fit

We now introduce the concept of deviance: a measure of goodness-of-fit. Let us consider two extreme models:

- Null model: $\mu =$ constant (equivalently, $\eta =$ constant, i.e. intercept only). All variations in observations are due to random component. This model is usually too simple.
- Saturated model: n parameters which leads to interpolation $\hat{\mu}_i = y_i$. All variations in observations are due to systematic component. Simply repeating the data, this model is uninformative.

Scaled deviance

For a model M with p parameters, we define the scaled deviance as

$$D_M^* \triangleq -2 \ln \frac{\text{maximum likelihood under model } \mathrm{M}}{\text{maximum likelihood under the saturated model}}$$

Assume that $a_i(\phi)$ has the special form $a_i(\phi) = \phi / w_i$. Denote $\hat{\theta}_i$ and $\tilde{\theta}_i$ as estimates under model M and the saturated model. Then

$$l_M = \sum_{i=1}^{n} \left\{ w_i \left[y_i \hat{\theta}_i - b\left(\hat{\theta}_i\right) \right] / \phi + c\left(y_i, \phi\right) \right\}$$

$$l_S = \sum_{i=1}^{n} \left\{ w_i \left[y_i \tilde{\theta}_i - b\left(\tilde{\theta}_i\right) \right] / \phi + c\left(y_i, \phi\right) \right\}$$

Deviance

Therefore,

$$\begin{aligned} D_M^* &= 2\left(l_S - l_M\right) \\ &= 2\sum_{i=1}^{n} w_i \left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right]/\phi \\ &\triangleq D_M/\phi \end{aligned}$$

where

$$D_M \triangleq 2\sum_{i=1}^{n} w_i \left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right]$$

is defined as the deviance.

Examples of deviances

	Deviance
Gaussian	$\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$ (RSS)
Binomial	$2\sum_{i=1}^{n} \{y_i \ln (y_i/\hat{\mu}_i) + (m_i - y_i) \ln ((m_i - y_i) / (m_i - \hat{\mu}_i))\}$
Poisson	$2\sum_{i=1}^{n} \{y_i \ln (y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$

Distribution of deviance

For Gaussian data,

$$D_M^* = D_M/\phi \sim \chi^2_{n-p}.$$

For non-Gaussian data, the χ^2 distribution holds approximately under certain conditions:

- Binomial: $m_i \pi_i (1 - \pi_i) \to \infty$
- Poisson: $\mu_i \to \infty$

Standard asymptotic argument with $n \to \infty$ requires the number of parameters being fixed. It does not apply to deviance since the number of parameters in the saturated model is n. Therefore, the χ^2 approximation do not hold as $n \to \infty$.

Generalized Pearson X^2 statistic

Another measure of the goodness-of-fit is the generalized Pearson X^2 statistic

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

For Gaussian data, it is again the RSS and

$$X^2/\phi \sim \chi^2_{n-p}$$

For non-Gaussian data, the χ^2 distribution hold asymptotically.

Estimates of the dispersion parameter

When ϕ is unknown, based on above χ^2 approximations, two approximately unbiased estimate of the dispersion parameter ϕ are

$$\hat{\phi} = \frac{D_M}{n - p}$$
$$\tilde{\phi} = \frac{X^2}{n - p}$$

Note: for binary data, $\tilde{\phi}$ is consistent while $\hat{\phi}$ is not. $\tilde{\phi}$ usually has smaller bias than $\hat{\phi}$.

Over- and Under-dispersion

Over-dispersion occurs when variance of the response variable exceeds the nominal value. That is,

$$\mathrm{Var}(y) > a(\phi)b''(\theta)$$

Similarly, under-dispersion occurs when variance of the response variable falls short of the nominal value. That is,

$$\mathrm{Var}(y) < a(\phi)b''(\theta)$$

Binomial: over-dispersion means that

$$\mathrm{Var}(y) > m\pi(1-\pi)$$

Poisson: over-dispersion means that

$$\mathrm{Var}(y) > \mu$$

Over- and Under-dispersion

- Over-dispersion is quite common in practice. It is wise to be cautious and assume that over-dispersion is present unless it is shown to be absent.
- Over-dispersion can arise in a number of ways. One common situation will be given as an illustration.
- Under-dispersion is less common.
- Note that over-dispersion and under-dispersion are defined in terms of parameters. How do we check them from data?
- One simple (naive) approach is compare $\hat{\phi}$ and/or $\tilde{\phi}$ with the nominal values ($\phi = 1$ for both Binomial and Poisson data) to find signs of over-dispersion.

How to deal with over-dispersion?

There are two general approaches: seak and model the extra variation

- Binomial: Bete-Binomial model
- Poisson: negative-Binomial model
- In general, generalized linear mixed effects models ignore the underlying mechanism and find a way to account for its effect. For example, quasi-likelihood. The second approach is preferable unless either the mechanism that produces over-dispersion is of interest, or there are strong reasons to assume a particular form of random effects. We will discuss the second approach (briefly) in this class.

Quasi-likelihood

We do not have the likelihood since the distribution is unknown. Quasi-likelihood is a technique which allow us to draw inference based on the first two moments only.

The `quasi-likelihood` for observation y is defined as

$$Q(\mu; y) \triangleq \int_y^\mu \frac{w(y-t)}{\phi V(t)} dt$$

The quasi-likelihood score function is

$$q = \frac{\partial Q}{\partial \mu} = \frac{w(y-\mu)}{\phi V(\mu)}$$

For a GLM with $a(\phi) = \phi/w$, q is the same as the score function $\left(\frac{\partial l}{\partial \mu}\right)$.

Properties of quasi-likelihood

Under above assumptions about the first two moments, q satisfies the following properties

$$\mathrm{E}(q) = 0$$
$$\mathrm{Var}(q) = \frac{W}{\phi V(\mu)}$$
$$-\mathrm{E}\left(\frac{\partial q}{\partial \mu}\right) = \frac{w}{\phi V(\mu)}$$

Most first-order asymptotic theory connected with likelihood is based on these three properties. Therefore, same asymptotic theory applies to the quasi-likelihood.

Dealing with over-dispersion

Quasi-likelihood is a general tool with many applications. Here we apply it to deal with the problem of over-dispersion. For Binomial and Poisson data, the dispersion parameter $\phi = 1$. When there are signs of over-dispersion, we may use the corresponding quasi-likelihood models where the dispersion parameter ϕ is estimated.

GLMs in R

"glm" has several options for family:

binomial (link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi (link = "identity", variance = "constant")
quasibinomial (link = "logit")
quasipoisson(link = "log")

Utility functions for glm

- Summary statement: `summary`
- Fits: `coefficients`, `fitted.values`
- Model building: `step`, `add1`, `drop1` and `stepAIC` in library `MASS`
- Diagnostics: `residuals`, `influence.measures`. The `glm.diag.plots` function in the boot library is very useful for constructing diagnostic plots
- Inference: `anova`
- Prediction: `predict` Type `help (function.name)` in R to find out more information about these functions.

Exercise

More math derivation exercises of inference of GLMs are in this week's exercises.