Chapter 4

Generalized Linear Models

EXAM PA LEARNING OBJECTIVES

6. Topic: Generalized Linear Models

Learning Objectives

The Candidate will be able to describe and select a Generalized Linear Model (GLM) for a given data set and regression or classification problem.

b) Understand the specifications of the GLM and the model assumptions.

c) Create new features appropriate for GLMs.

d) Interpret model coefficients, interaction terms, offsets, and weights.

e) Select and validate a GLM appropriately.

Chapter overview: Despite their simplicity and the central role they play in predictive analytics, linear models as introduced in Chapter 3 rely heavily on the target variable being normally distributed (as well as a number of restrictive assumptions) and often bear limited applicability to insurance business. Here are some examples of target variables that are commonly encountered in insurance, but clearly defy the normal distributional assumption:

- A binary variable that indicates whether an event of interest takes place over a reference period or whether a subject has a particular characteristic of interest (e.g., yes/no, pass/fail, claim/no claim, alive/dead, solvent/insolvent). Such a variable is often coded as taking on values of either zero, if the event does not occur, or one, if the event does occur, in contrast to a normal random variable, which can assume any real values.

- A count variable that "counts" the number of times a certain event of interest (e.g., number of claims, accidents, deaths) takes place over a reference period. Such a variable can take the value of any non-negative integers.

- A continuous variable which is restricted to take only positive values and may be highly skewed, such as income and insurance claim severity.

While technically a linear model can still be fitted to these highly non-normal variables, the resulting model may make little practical sense and the predictions may hardly be reliable.

Building on the groundwork in Chapter 3, this chapter considerably expands the range of modeling approaches to understanding data by introducing *generalized linear models* (GLMs), which

provide a unifying framework for dealing with a rich class of non-normally distributed target variables and allow for more general and complex relationships between target variables and predictors. All of the released past and sample PA projects test GLMs, speaking to the prominent role that GLMs play in this exam. We begin in Section 4.1, which develops the theoretical underpinnings of GLMs conducive to a conceptual understanding of what GLMs do and practical implementations of GLMs. Sections 4.2 and 4.3 present two GLM-based case studies with an actuarial focus, both of which not only serve to demonstrate how GLMs for claim severity and claim occurrence can be constructed, evaluated, and applied in R to make predictions, but also to help prepare for your future role as an actuary.

4.1 Conceptual Foundations of GLMs

This section lays the conceptual groundwork of GLMs and paves the way for the two case studies in Sections 4.2 and 4.3. In Exam PA, there are often tasks (e.g., executive summary, which is often the most weighty item on the exam) that require you to describe, in high-level terms, what a GLM is and the pros and cons of a GLM relative to other predictive models, so the conceptual aspects of GLMs will be useful not only for understanding the practical implementations of GLMs in the next two sections, but also for tackling exam items. Because all of the feature generation (e.g., binarization of categorical predictors, introduction of polynomial and interaction terms) and feature selection techniques (e.g., stepwise selection algorithms) for linear models presented in Subsections 3.2.3 and 3.2.4 generalize to GLMs in essentially the same way and everything we learned about the bias-variance trade-off for linear models also applies here, our focus in this section is on issues that are specific to GLMs but absent for linear models. These issues include:

- Selection of target distributions and link functions
- Use of offsets and weights
- Interpretation of the results of a GLM
- Evaluation of GLMs for binary target variables

GLMs in a nutshell. Dating back to the 1970s, GLMs are an important predictive analytic tool whose importance is continually rising, especially in insurance settings, where most variables are non-normal in nature, but are amenable to generalized linear modeling. If one word is to summarize the virtues of a GLM relative to a linear model, that word is probably "flexible." Compared to linear models, GLMs provide considerable flexibility and substantially widen the scope of applications in two respects:

1. *Distribution of the target variable:* The target variable in a GLM is no longer confined to the class of normal random variables; it need only be a member of the so-called *exponential family of distributions.* The mathematical details of the exponential family are not important for Exam PA; it is good enough to know that this is a rich class of distributions that include a number of discrete and continuous distributions commonly encountered in practice, such as the normal, Poisson, binomial, and gamma distributions. GLMs therefore provide a unifying approach to modeling binary, discrete and continuous target variables and doing both regression and classification problems, all within the same statistical framework.

2. *Relationship between the target mean and linear predictors:* Instead of equating the mean of the target variable to the linear combination of predictors, a GLM sets a *function* of the target mean to be linearly related to the predictors. This allows us to analyze situations in which the effects of the predictors on the target mean are more complex than merely additive in nature without having to transform the target observations.

   Mathematically, the equation of a GLM is of the form[i]

$$\boxed{g(\mu) = \eta := \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,}$$

   where $g(\cdot)$ is the *link function* "linking" the mean of the target variable μ to the linear combination of the predictors, sometimes referred to as the *linear predictor* η. The link function can be any monotonic function (e.g., identity, log, inverse); its monotonicity of g allows us to invert the link function and make the target mean μ the subject:

$$\mu = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p).$$

Together, the distribution of the target variable and the link function g fully define a GLM:

$$\text{GLM:} \qquad \text{exponential family } \ni \text{ target variable} \quad + \quad \text{Link function.}$$

In the special case when the target variable is normally distributed and the link function is the identity function $g(\mu) = \mu$, we are back to the linear model setting in Chapter 3.

GLMs vs. linear models on transformed data. Before the advent of GLMs, actuaries would attempt to transform the non-normal data so that they can be reasonably described by a normal linear model. If, for example, the target variable is heteroscedastic with its variability increasing with the observed values, then a logarithmic transformation may help restore homoscedasticity and a linear model can be fitted to the log of the target variable (assuming that it is positive).

One of the most common misconceptions about GLMs is to confuse them with linear models fitted to transformed data; they are fundamentally different modeling approaches. To see this, consider the following two models:

- *Model 1 (Linear model fitted to* $\ln Y$*):* Suppose that we have fitted a simple linear regression model with equation

$$\ln Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim \text{N}(0, \sigma^2),$$

  or

$$Y = \exp(\beta_0 + \beta_1 X + \varepsilon).$$

  As $\exp(\varepsilon)$ is *log*normally distributed, so is Y. In particular, the target variable must take non-negative values only.

- *Model 2 (GLM with normal target and log link):* Now suppose that have fitted a GLM with a normal target variable, a single predictor, and the log link. The model equation is

$$\ln \mu = \beta_0 + \beta_1 X \qquad \text{or} \qquad \mu = \exp(\beta_0 + \beta_1 X).$$

  Thus the target variable Y follows a $\text{N}(\exp(\beta_0 + \beta_1 X), \sigma^2)$ distribution. Although its mean is constrained to be non-negative, the target variable itself can take positive or negative values.

As you can see, the two modeling methods imply entirely different behavior of the target variable. Which model should we recommend? That will depend on which one is easier to interpret or has better predictive performance. None of the two models will yield universally superior answers.

[i]Note that there is no error term in the equation of a GLM because it is an equation for the mean of the target variable.

4.1.1 Selection of Target Distributions and Link Functions

Component 1: Target distributions. A common item in Exam PA is the selection and justification of an appropriate target distribution (i.e., distribution of the target variable) and a link function for a given situation. The nature of the target variable often guides the choice of the target distribution as well as the link function. For most types of target variables, usually one or two target distributions are most appropriate.

- *Continuous, positive data:* In insurance applications, variables that are continuous and strictly positive, such as claim amounts, income, and amount of insurance coverage, play a prominent role. Although the linear model with a log transformation of the variable may work fine in some cases, GLMs supply a variety of target distributions, like *gamma* and *inverse Gaussian*, that more effectively capture the skewness of the target variable. A case study of GLMs that deals with a continuous, positive, and right-skewed target variable will be presented in Section 4.2.

- *Binary data:* When the target variable is binary, which happens in a classification problem with the target variable indicating the occurrence or non-occurrence of an event of interest, the *binomial* (Bernoulli, to be precise) distribution is probably the only reasonable choice. Examples of binary target variables include whether or not a policyholder lapses or submits a claim, whether or not a submitted claim is fraudulent, and whether or not an actuarial student passes Exam PA. In Section 4.3, we will look at a full-length case study involving a binary target variable.

- *Count data:* For target variables that assume only non-negative integers, as is the case for count data, the *Poisson* distribution is one of the most natural candidate distributions. One of its drawbacks is that it requires that its mean and variance be equal. When the variance of the target variable exceeds its mean, a situation known as *overdispersion*, the Poisson distribution can be tweaked to account for this. Alternatively, other count distributions such as negative binomial (not discussed in the PA modules) can be explored.

Example 4.1.1. (CAS Exam 8 Fall 2014 Question 3 (b) (Reworded): Selection of response distributions) For each of the target variables below, identify the response distribution that should be used to model the data. Briefly explain why that response distribution is appropriate.

(a) Severity

(b) Policy renewal retention

Solution. (a) The gamma or inverse Gaussian distributions both work well as they produce only positive outcomes and are skewed in nature, consistent with the character of severity.

(b) The binomial distribution is arguably the only appropriate distribution for policy renewal retention, which is either a yes/no outcome.

□

Component 2: Link functions. The choice of the link function is more controversial than the choice of the target distribution, and often the nature of the target variable as well as the interpretability of the link function are both taken into account. For example:

- If the target mean is a priori known to be positive, then a link function such as the *log link* $g(\mu) = \ln \mu$ is a good candidate link function as it ensures that the predictions of the GLM given by $\hat{\mu} = e^{\hat{\eta}}$ are always positive. The log link also has the advantage of ease of interpretation, as we will see in the "Interpretation of regression coefficients" paragraph below.

- If the target mean is between 0 and 1, as is the case of binary target variables coded as 0 or 1, then the link function should ensure that the predictions of the GLM are unit-valued. A common choice is the *logit* link given by

$$g(\pi) = \ln \frac{\pi}{1-\pi} = \ln(\text{odds}),$$

where π is the mean of the binary target variable and $\pi/(1-\pi)$, known as the *odds* of the event of interest, is the ratio of the probability of occurrence to the probability of non-occurrence and provides a measure of chance valued between 0 and ∞ (not between 0 and 1). In terms of the linear predictor, the target mean is

$$\pi = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}},$$

which is always between 0 and 1. As we will see below, the logit link is also easy to interpret due to its connections with the log link. Terminology-wise, a GLM for a binary target variable with the logit link is called a *logistic regression model*, which Section 4.3 will center on.

One way to help with the specification of the link function is to look at the *canonical link function* that is associated with each target distribution. The canonical link functions for a number of common target distributions are tabulated below:

Target Distribution	Canonical Link Function
Normal	Identity, i.e., $g(\mu) = \mu$
Binomial	Logit, i.e., $g(\pi) = \ln[\pi/(1-\pi)]$
Poisson	Natural log, i.e., $g(\mu) = \ln \mu$
Gamma	Inverse, i.e., $g(\mu) = 1/\mu$
Inverse Gaussian	Squared inverse, i.e., $g(\mu) = 1/\mu^2$

Canonical links have the advantage of simplifying the estimation procedure, but this alone does not mean canonical links should always be used. More important factors to consider are whether the predictions provided by the link align with the characteristics of the target variable and whether the resulting GLM is easy to interpret. In the case of the gamma GLM, for example, the canonical link, which is the inverse link, does not guarantee positive predictions, neither is it easy to interpret. As a result, the log link is much more commonly used.

Example 4.1.2. (SOA Exam SRM Sample Question 7 (Reworded): Selection of target distribution and link function) Determine which of the following pairs of distribution and link function is the most appropriate to model if a person is hospitalized or not.

(A) Normal distribution, identity link function

(B) Normal distribution, logit link function

(C) Binomial distribution, identity link function

(D) Binomial distribution, logit link function

(E) It cannot be determined from the information given.

Solution. Whether a person is hospitalized or not is a binary variable, which is best modeled by a binomial (more precisely, Bernoulli) distribution, leaving only Answers (C) and (D). The link function should be one that restricts the Bernoulli target mean to the range zero to one. Among the identity and logit links, only the logit link has this property. **(Answer: (D))** □

Example 4.1.3. (CAS Exam MAS-I Fall 2019 Question 25: Logistic regression) A bank uses a logistic model to estimate the probability of clients defaulting on a loan, and it comes up with the following parameter estimates:

j	Variable	$\hat{\beta}_j$
0	Intercept	−1.6790
1	Income (in 000's)	−0.0294
2	Student [Yes]	−0.3870
3	Number of credit cards	0.7710

The following four clients applied for loans from the bank:

Client	Income	Student	# of credit cards
1	25,000	Y	1
2	10,000	Y	3
3	20,000	N	0
4	75,000	N	3

The bank will reject any loan if the probability of default is greater than 10%. Calculate the number of clients whose loan requests are rejected.

(A) 0

(B) 1

(C) 2

(D) 3

(E) 4

Solution. Recall that in terms of the linear predictor, the mean of a Bernoulli random variable with a logit link is

$$\pi = \frac{1}{1+e^{-\eta}}.$$

Applying this formula four times, we can calculate the fitted linear predictor and the predicted probability of default for the four clients, and decide whether to reject their loan requests:

Client	$\hat{\eta} = \mathbf{x}^\top \hat{\boldsymbol{\beta}}$	$\hat{\pi} = \frac{1}{1+e^{-\hat{\eta}}}$	Prediction
1	$-1.6790 - 0.0294(25) - 0.3870 + 0.7710(1) = -2.03$	0.1161	Reject
2	$-1.6790 - 0.0294(10) - 0.3870 + 0.7710(3) = -0.047$	0.4483	Reject
3	$-1.6790 - 0.0294(20) + 0.7710(0) = -2.267$	0.0939	Not reject
4	$-1.6790 - 0.0294(75) + 0.7710(3) = -1.571$	0.1721	Reject

In conclusion, three clients have their loan requests rejected. **(Answer: (D))** □

Interpretation of regression coefficients. How do we interpret the coefficients of a GLM? That depends crucially on the choice of the link function. For some link functions, the coefficients of a GLM may be difficult to explain and interpret. Two commonly used link functions that are particularly appealing in terms of interpretability are:

- *Log link:* When the log link is used, the target mean and linear predictor are related via

$$\mu = e^{\eta} = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}.$$

  In this case, a unit change in a continuous predictor with coefficient β_j is associated with a *multiplicative* increase in the target mean by a factor of e^{β_j}, holding all other variables fixed. If β_j is positive (resp. negative), then the target mean gets amplified (resp. shrunken). Equivalently, the *algebraic* change in the target mean is $(e^{\beta_j} - 1)\mu$, where μ is the target mean before the change, and the *percentage* change in the target mean is $e^{\beta_j} - 1$.

  The coefficients of a categorical predictor admits a similar interpretation: The target mean when the categorical predictor lies in a certain level with coefficient β_j is e^{β_j} times of that when the categorical predictor is in the baseline level, holding all other variables fixed.

- *Logit link:* When the logit link is used, usually for binary data, the model equation is

$$\ln(\text{odds}) = \eta \quad \Leftrightarrow \quad \text{odds} = e^{\eta},$$

  which is just another form of the log link, so the interpretations above based on algebraic or percentage changes apply.

Sidebar: Weights and offsets. Many feature generation and selection techniques that apply to linear models carry over easily to GLMs to set up the right-hand side of the model equation $g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. Weights and offsets, however, are modeling tools that are much more commonly used with GLMs (although technically they can also be used with linear models) and thus are discussed here.

- *Weights:* So far we have assumed that every observation of the target variable exhibits the same amount of variability.[ii] As you can expect, this constant variability assumption among the target variable observations may not be always realistic. In many situations, the datasets provided by an insurance company store the information not of individual policyholders, but of a *group* of policyholders of similar or the same characteristics. In the latter case, each observed value of the target variable may be an *average* of the target variable across policyholders in the same group and thus is exposed to a smaller amount of variation due to the averaging. A GLM can take advantage of the fact that different observations have different degrees of precision to improve the fitting procedure.

  Recall from elementary statistics that the variance of the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ coming from an i.i.d. sample $X_1, \ldots, X_n$ is given by $\text{Var}(\bar{X}) = \text{Var}(X_1)/n$, with a division by the sample size n. The higher the value of n, the smaller the variance of $\bar{X}$ and the more reliable $\bar{X}$ becomes as an estimator. The same is true for a GLM, with the number of "policyholders" (the group size) corresponding to each observation playing the same role as n. To refine the model fitting, we can attach a higher *weight* to observations that are averaged across more subjects of similar characteristics and are thus more precise. In R, this can be easily achieved by adding a simple command of the form `weight = <weight_variable>` as an argument of the `lm()` function or `glm()` function (which we will learn in Section 4.2) to specify the weight given to each observation, but it is still good to know the rationale behind the use of weights. Also, in Exam PA the project may not specifically ask you to apply weights when faced with a dataset where the individual outcomes are averaged across a group of independent subjects, but still expect you to realize that the use of weights is desirable.

- *Offsets:* Another common way for the number of subjects underlying each observation to affect the probabilistic behavior of the target variable is when we are dealing with count data. That is, the target variable counts how many times an event of interest (e.g., death, claim) happened over a certain period in a group of subjects (e.g., policyholders) with a known size. Other things equal, we can expect that the more people are in the group, the more events we observe. This motivates the use of the group size as an additional predictor, called the *offset*, to account for the different means of different observations.

  Technically, an offset is a special predictor whose regression coefficient is known a priori to be one; no estimation is needed. When including an offset term in the equation of a GLM, it is important to make it on the same scale as the linear predictor. In the case of the log link, with which the offset is most commonly used, we should use the log of the group size, $\ln n_i$, as the offset term:

$$\ln \mu_i = \boxed{\ln(\text{group size})} + \eta_i = \boxed{\ln n_i} + (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}),$$

  which is equivalent to

$$\ln\left(\frac{\mu_i}{n_i}\right) = \eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}.$$

  In this form, we are modeling the occurrence rate of the event of interest, which is the expected number of events per subject in the group.

[ii] To be precise, we have implicitly assumed that all observations of the target variable of a GLM share the same dispersion parameter, but not necessarily the same variance. If you recall what you learned in Exam SRM, we have $\text{Var}(Y) = \phi v(\mu)$, where ϕ is the dispersion parameter and $v(\mu)$ is the variance function, which varies across different observations.

Note that offsets and weights, from a certain perspective, both involve a kind of average and, at first sight, they seem so similar that they are synonyms of each other. They, in fact, impose vastly different structures on the mean and variance of the target variable. The key differences lie in the way the observations of the target variable are recorded (average vs. aggregate) and whether the group size affects the mean or variance of the target variable. The two modeling approaches are not consistent with each other.

- *Weights:* To use weights properly, the observations of the target variable should be *averaged* across members of the same group. Due to the averaging, the variance of each observation is inversely related to the group size, which serves as the weight for that observation. However, the weights do not affect the mean of the target variable.
- *Offsets:* For offsets, the observations are values *aggregated* over members of the same group. The group size, when serving as an offset, is positively related to the mean of the target variable, but leaves its variance unaffected.

Example 4.1.4. (Specifying an appropriate GLM) You are given an insurance dataset with the following variables:

- `NumClaims`: Number of claims of a policyholder
- `Time`: The length of time (in years) a policyholder is observed
- `Age`: Age of a policyholder
- `Gender`: Gender of a policyholder
- `RiskCat`: Risk category to which a policyholder belongs

Set up a GLM to determine the effects of various factors on `NumClaims`. Specify the distribution of the target variable, the predictors, the link function and the offset (if any).

Solution. This dataset has no "group size" at work, but it is reasonable to expect that `NumClaims` will vary in direct proportion to `Time`, which serves as an offset term to adjust the different exposure to which different policyholders are exposed. The GLM can be specified as follows:

- *Target distribution:* Poisson
- *Predictors:* `Age` (continuous), `Gender` (binary), `RiskCat` (categorical)
- *Link:* Log
- *Offset:* ln(`Time`)

In R, such a Poisson GLM can be fitted with the command

```
GLMPoisson <- glm(NumClaims ~ Age + Gender + RiskCat,
                  family = poisson, offset = log(Time))
```

or

```
GLMPoisson <- glm(NumClaims ~ . - Time,
                  family = poisson, offset = log(Time))
```

Notice that the `offset` argument is set to `log(Time)`, not `Time`. □

4.1.2 Model Statistics

Parameter estimates. As soon as the target distribution and the link function have been specified, a GLM can be constructed and fitted. Instead of the ordinary least squares method that is used in a linear model, the method of *maximum likelihood estimation* (MLE) is commonly employed to estimate the unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$ of a GLM (recall that for linear models, maximum likelihood estimates coincide with the least squares estimates). If we denote the coefficient estimates by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, then the estimated linear predictor and predicted target mean are, respectively,

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip} \quad \text{and} \quad \hat{\mu}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}).$$

The mathematical details of MLE are not the focus of Exam PA (you should have learned MLE in Exam STAM/C!). Loosely speaking, we choose the parameter estimates in such a way to maximize the likelihood of observing the given data. Statistical theory asserts that estimates found this way possess some desirable statistical properties, such as asymptotic unbiasedness, efficiency and normality. A downside to MLE is that occasionally the optimization process involved may be plagued by convergence issues, which may happen when a non-canonical link is used. No estimates may be produced and the GLM cannot be fitted as a result.

Deviance. For GLMs, goodness-of-fit measures which are based in part on the normal distribution such as R^2 do not apply. The liberation from the normal distribution framework calls for more general, likelihood-based goodness-of-fit measures that serve to evaluate the model fit on the training set appropriately.

One of the most important goodness-of-fit measures for GLMs is the *deviance*. Unlike the R^2, which measures how a linear model compares with the most primitive linear model (the intercept-only linear model), the deviance of a GLM measures the extent to which the GLM departs from the most elaborate GLM, which is known as the *saturated model* and allows for perfect fit. Mathematically, if we denote the loglikelihood of the saturated model by l_{SAT} and that of a given GLM by l, then the deviance of the GLM is given by[iii]

$$D = 2(l_{\text{SAT}} - l).$$

The lower the deviance, the closer the GLM is to the model with perfect fit, and the better its goodness of fit on the training set.

Although the deviance is part of the output when the summary of a GLM is returned and there will hardly be any instances in Exam PA where you will manipulate the deviance of a GLM, this model statistic provides the foundation of an important model diagnostic tool for GLMs: The deviance residual.

Deviance residuals. In a GLM, the raw residuals $e_i = y_i - \hat{\mu}_i$, where $\hat{\mu}_i$ is the fitted target mean for the ith observation in the training set, are not as useful as they are in a linear model because they are no longer normally distributed (not even approximately, because the target distribution may not be normal), nor do they possess constant variance (because their variance varies with the target mean, which varies across different observations). This makes it difficult to find a universal benchmark against which the raw residuals of all GLMs can be reliably compared. This is where deviance residuals play their role.

[iii] The first term of the deviance formula, $2l_{\text{SAT}}$, does not vary with the fitted values of the GLM in question, so some define the deviance as simply $D' = -2l$. With this definition, the AIC and BIC become AIC $= D' + 2p$ and BIC $= D' + \ln(n_{\text{tr}})p$, respectively.

The *deviance residual* for the ith observation, denoted by d_i, is defined as the *signed*[iv] square root of the contribution of the ith observation to the deviance. Mathematically, we have $\sum_{i=1}^{n} d_i^2 = D$. In contrast to the raw residuals, the deviance residuals are, provided that the assumed GLM is adequate, approximately normally distributed and homoscedastic (i.e., same variance), for most target distributions (which need not be a normal distribution). This allows us to compare the distribution of the deviance residuals with the normal distribution (e.g., by examining Q-Q plots of deviance residuals) and evaluate whether the fitted GLM is an accurate description of the given data.

It is true that the formula and calculations for deviance residuals are rarely tested in Exam PA, but it is important to know what we can learn by looking at deviance residuals and why they are needed in place of their raw counterparts.

Example 4.1.5. (CAS Exam S Spring 2017 Question 35 (Adapted): Calculation of deviance residual) You have fit a Poisson GLM. One observation of the target variable is 59. The corresponding GLM fitted value is 71.

Calculate the deviance residual corresponding to this observation.

(Note: The formula for the deviance of a Poisson GLM based on a training sample of size n is

$$D = 2\sum_{i=1}^{n}\left[y_i \ln\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right].)$$

Solution. Taking the signed square root of the individual term in the deviance formula, we get the following expression for the ith deviance residual:

$$d_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2\left[y_i \ln\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right]},$$

where $\text{sign}(y_i - \hat{\mu}_i) = 1$ if $y_i > \hat{\mu}_i$ and $\text{sign}(y_i - \hat{\mu}_i) = -1$ if $y_i < \hat{\mu}_i$. Because $59 = y_i < \hat{\mu}_i = 71$, the deviance residual should be negative, so

$$d_i = -\sqrt{2\left[59\ln\left(\frac{59}{71}\right) - (59 - 71)\right]} = \boxed{-1.4674}.$$

□

Example 4.1.6. (A protest!?!?) An actuarial student, Samson, was asked to examine the deviance residuals of a gamma GLM via a Q-Q plot and look for any anomalies. Defying his supervisor's order, Samson protested and said:

> It makes completely no sense to look at the Q-Q plots for GLMs! The residuals for a GLM, except for a normal linear model, are not normally distributed, so diagnostic tools based on the normal distribution are misleading, if not flawed. I need a more competent supervisor!

[iv] The ith deviance residual is taken to be positive if $y_i > \hat{\mu}_i$ and negative if $y_i < \hat{\mu}_i$, and has the same sign as the corresponding raw residual.

Would you recommend that Samson leave his company in search of a more competent supervisor?

Solution. To begin with, let's make one thing absolutely clear:

> The residuals in the diagnostic plots for a GLM are not the *raw* residuals, but the (standardized) deviance residuals.

While the raw residuals of a GLM are non-normally distributed (except for a normal GLM), the deviance residuals of a GLM are, in most cases (binary targets being an exception), approximately normally distributed, provided that the model is correctly specified. This fact, which is pointed out earlier, is exactly the reason why deviance residuals are used in place of raw residuals and justifies looking at the Q-Q plot for a GLM, which displays the standardized deviance (not raw) residuals plotted against the standard normal quantiles. If the observed points in the Q-Q plot are close to the reference straight line, then the fitted GLM is perhaps adequate; if the points are far away from the reference straight line, then the target distribution used and/or the form of the model may not be appropriate. Note that it will be rare for the points to fall almost exactly on the reference line (which is merely a benchmark) because results for GLMs are not exact, but asymptotic, requiring a large sample to hold true.

Back to Samson's protest, a good response is:

> Please learn predictive analytics well before you make unfounded, hyperbolic statements (the ACTEX PA manual seems a good resource for this purpose ☺).

□

EXAM NOTE

The June 2019 PA exam model solution says that when asked to interpret diagnostic plots for a GLM,

> "[s]ome candidates did not understand that the Q-Q plot checks the normality of the standardized *deviance* residuals. If an appropriate model is being used, this should be the case *regardless of the distribution used* in the model."

While you need not know the technical details behind the output you see in R, you have to understand, at a conceptual level, what the output is and what it is for. This explains why this introductory section on GLMs is useful!

Remark. This example is motivated by this popular post on Reddit about two so-called "mistakes" in the June 2019 PA exam model solution: (last accessed on February 6, 2020).

> `https://www.reddit.com/r/actuary/comments/dvtb85/mistake_in_the_june_exam_pa/`

I will not call the first "mistake" (uploading the wrong solution file) a mistake on the part of the PA exam committee. More importantly, the second "mistake" is not a mistake at all and reflects a serious misconception of the original poster. Unaware of the fact that the Q-Q plot for

a GLM involves deviance, not raw residuals, he irresponsibly brands the SOA's model solution as flawed (when it is not) in an attempt to promote his own product and says "Let's pray that the December exam was written by more competent individuals." *(wow!)* This makes a mockery of someone who claims to be doing predictive modeling for a living and helping PA candidates to pass the exam.

I rarely comment on other exam preparation resources, but this post is to the detriment of many PA candidates as well as the SOA and has gone way too far. While the original poster eventually made a brief, less than apologetic clarification, the number of upvotes received by that post suggests that many readers may have been misled and I find it imperative to clarify unequivocally, hence this newly created example.

Without feeling guilty, the original poster asked me to refrain from referring to this post because my clarification will damage his "reputation." I instead suggested that he reflect on the unwarranted damage and injustice his post has done to the SOA and issue a new public post apologizing to the SOA and readers on Reddit and clarifying that the (second) "mistake" is a mistake on his part. Regrettably, the original poster does not have the courage to do so.

Penalized likelihood measures: AIC and BIC. The deviance of a GLM parallels the residual sum of squares or R^2 for a linear model in the sense that it is merely a goodness-of-fit measure on the training set and always decreases when new predictors are added, even if they are not predictive of the target variable. More meaningful measures of model quality balance the goodness of fit with model complexity. Two common examples are the AIC and BIC, which are defined in the same way as for a linear model:

$$\text{AIC} = -2l + 2p \qquad \text{and} \qquad \text{BIC} = -2l + \ln(n_{\text{tr}})p,$$

where l is the loglikelihood of the GLM on the training set, p is the number of parameters of the GLM, and n_{tr} is the size of the training set. Using these information criteria, we can perform stepwise feature selection in the same way as a linear model (refer to Subsection 3.2.4) and drop non-predictive features to prevent overfitting.

Example 4.1.7. (CAS Exam MAS-I Fall 2019 Question 32: Comparing models with respect to AIC and BIC) You have three competing GLMs that each predict the number of claims under an insurance policy, and are evaluating the models using AIC and BIC. All models are trained on the same dataset of 300 observations. These models are summarized below:

Model	Likelihood	Number of Parameters
1	0.0456	4
2	0.0567	5
3	0.0575	6

The following are three statements about the fit of these models:

I. Model #1 is best based on BIC

II. Model #2 is best based on AIC

III. Model #3 is best based on BIC

Determine which of the above statements is/are true.

(A) I only

(B) II only

(C) III only

(D) I, II, and III

(E) The answer is not given by (A), (B), (C), or (D)

Solution. Let's calculate the AIC and BIC for each of the three GLMs:

Model	L	p	AIC $= -2\ln L + 2p$	BIC $= -2\ln L + p\ln 300$
1	0.0456	4	14.1757	28.9908
2	0.0567	5	15.7400	34.2589
3	0.0575	6	17.7119	39.9346

Since Model 1 has the lowest AIC as well as the lowest BIC, it is the best model among the three models, so only I is true. **(Answer: (A))** □

4.1.3 Performance Metrics for Classifiers

Recall from Subsection 3.1.2 that RMSE is the most commonly used performance metric for numeric target variables. Since GLMs accommodate both numeric and categorical target variables, we need performance metrics for categorical variables as well. We briefly mentioned in Subsection 3.1.2 the classification error rate as a measure of the performance of a classifier. This is just one of a few commonly used measures. To perform a more comprehensive analysis of a classifier, we can employ a confusion matrix, which bears a direct relation to how a classifier is used in practice.

Confusion matrices. A *confusion matrix* is a tabular display of how the predictions of a binary classifier line up with the observed classes. From a confusion matrix, a lot of useful performance metrics can be derived.

Recall that a binary classifier merely produces a prediction of the probability that the event of interest occurs, given a set of feature values. It does not directly say whether the event is predicted to happen or not. To translate the predicted probabilities into classifications, we need a pre-specified *cutoff* (or threshold) to decide on the class assignment. If the predicted probability $\hat{\pi}_i$ for an observation is higher (resp. lower) than the cutoff, then the event is predicted to occur (resp. not to occur). This simple classification rule results in four possibilities exhibited in a confusion matrix: (much like the Type I and Type II errors in hypothesis testing)

	Reference (= Actual)	
Prediction	No	Yes
No	TN	FN
Yes	FP	TP

where:

- *TP = true positive:* The classifier predicts that the event occurs and indeed it does.
- *TN = true negative:* The classifier predicts that the event does not occur and indeed it does not.
- *FP = false positive:* The classifier predicts that the event occurs, but it does not.
- *FN = false negative:* The classifier predicts that the event does not occur, but it does.

Given a confusion table corresponding to a certain cutoff, a lot of useful performance metrics of a classifier can be computed. The classification error rate mentioned earlier is the proportion of misclassifications, which are on the off-diagonal of the confusion matrix:

$$\text{classification error rate} = \frac{\text{FN} + \text{FP}}{n},$$

where n is the total number of observations (on the training or test set where the confusion matrix is constructed). The following two measures are also used frequently to summarize the predictive performance of a classifier:

- *Sensitivity (also called recall):* This is the relative frequency of correctly predicting an event of interest when the event does take place, or equivalently, the ratio of TP to the total positive events. In symbols,

  $$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

  It is a measure of how "sensitive" a classifier is to identify positive cases.
- *Specificity:* This is the relative frequency of correctly predicting a non-event when there is indeed no event, or the ratio of TN to the total negative events:

  $$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

The higher the sensitivity and specificity, the more attractive a classifier.

Note that a confusion matrix can be computed on both the training and test sets, but it is the confusion matrix on the test set that we care more about. (If desirable, we can compare the training confusion matrix and the test confusion matrix to detect overfitting. A large discrepancy in the classification performance on the two sets is indicative of overfitting.)

Effect of changing the cutoff. Ideally, the classifier and the cutoff should be such that both sensitivity and specificity are close to 1. Except for artificial situations, having such a perfect classifier is almost always impossible. In general, the selection of the cutoff involves a trade-off between having high sensitivity and high specificity.

- If the cutoff is set to 1, meaning that everyone is predicted to be "No," then the sensitivity and specificity of the classifier are 0 and 1, respectively.
- As we drop the cutoff, we are having more true positives and fewer false negatives, at the cost of having more false positives and fewer true negatives, so the sensitivity increases but the specificity decreases.

- If the cutoff is set to 0, meaning that everyone is predicted to be "Yes," then the sensitivity and specificity are 1 and 0, respectively.

The determination of the precise cutoff requires weighing the benefits of correctly catching positive cases and costs of missing positive cases. This cost-benefit analysis is often a business decision that requires subject matter expertise and practical considerations. If, for instance, you are using a classifier to identify fraudulent claims, which will impose a huge unwarranted cost on an insurance company, then we may choose a relatively low cutoff to mitigate the cost of undetected fraud (though some genuine claims will be misclassified as fraud!). Another example is Task 9 of the Hospital Readmissions sample project, which will be discussed in Component 3 of this study manual.

Example 4.1.8. (CAS Exam 8 Fall 2019 Question 5 (a), (b), & (d): Given a confusion matrix) The following confusion matrix shows the result from a claim fraud model with a discrimination threshold of 25%:

		Actual
Predicted	No	Yes
No	1203	162
Yes	63	72

(a) Identify a link function that can be used for a generalized linear model that has a binary target variable and briefly explain why this link function is appropriate.

(b) Calculate the sensitivity and specificity from the above data.

(c) Briefly describe how the severity of claims will impact the selection of the model threshold.

Solution. (a) The logit link $g(\pi) = \ln \frac{\pi}{1-\pi}$ is the most appropriate for linking the mean of a binary target variable to the linear predictor as it ensures that the predictions are always valued between 0 and 1. It also makes the model easily interpretable.

(b) The sensitivity is the proportion of fraudulent claims correctly predicted to be fraudulent:

$$\text{sensitivity} = \frac{72}{72+162} = 0.3077.$$

The specificity is the proportion of non-fraudulent claims correctly predicted to be non-fraudulent:

$$\text{specificity} = \frac{1203}{1203+63} = 0.9502.$$

(c) With high severity claims, we should lower the threshold so that more claims are predicted to be fraudulent and fewer fraudulent claims go undetected. When the claims are more severe, the benefit of identifying fraudulent claims outweighs the cost of having false negatives.

□

ROC curves. Instead of manually choosing a particular cutoff, the *Receiver Operating Characteristic* (ROC) *curve* is a graphical tool plotting the sensitivity against the specificity of a given

classifier for each cutoff in the range $[0, 1]$. Conventionally, one minus the specificity is plotted on the horizontal axis and sensitivity on the vertical axis. Each point on the ROC curve corresponds to a certain cutoff.

All ROC curves begin at $(0, 0)$ and end at $(1, 1)$, which, from the discussions above, correspond to a cutoff of 1 and 0, respectively. For a classifier with perfect predictive performance, its sensitivity and specificity should be both equal to 1, which gives rise to the point $(0, 1)$ in the top-left corner. The ROC curve of a good classifier should rise quickly to 1 as soon as it leaves the origin $(0, 0)$. The closer the curve is to the top-left corner, the better the predictive ability. With this line of reasoning, the predictive performance of a classifier can be summarized by computing the *area under the curve* (AUC). The highest possible value of the AUC is 1 while that for a classifier whose predictive ability is no better than chance (represented by the 45° diagonal line) is 0.5. These two extreme values provide a benchmark for judging the absolute and relative performance of different classifiers.

Example 4.1.9. (A toy classification example) A bank adopts the following default classification rule with cutoff l for any firm with predicted probability of default $\hat{\pi}$:

- If $\hat{\pi} > l$, then the firm is classified as a bad firm likely to default.
- If $\hat{\pi} \leq l$, then the firm is classified as a good firm unlikely to default.

A test set has six firms under investigation. Their predicted probabilities of default and real default statuses are summarized in the following table:

Firm	Predicted Probability	Status (1 = default, 0 = not default)
1	0.25	0
2	0.40	0
3	0.50	1
4	0.60	0
5	0.75	1
6	0.80	1

(a) Construct the confusion matrix of the classifier for $l = 0.7$.

(b) Plot the ROC curve for the above classifier.

Solution. (a) When the cutoff is $l = 0.7$, only Firms 5 and 6 are predicted to default. Here is the confusion matrix: (We will discuss the code in Section 4.3)

```
library(caret)
obs <- c(0, 0, 1, 0, 1, 1)
pred <- c(0.25, 0.40, 0.50, 0.60, 0.75, 0.80)
confusionMatrix(factor(1*(pred > .7)), factor(obs), positive = "1")

## Confusion Matrix and Statistics
##
##           Reference
## Prediction 0 1
##          0 3 1
##          1 0 2
##
##                Accuracy : 0.8333
##                  95% CI : (0.3588, 0.9958)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.1094
##
##                   Kappa : 0.6667
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.6667
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.7500
##              Prevalence : 0.5000
##          Detection Rate : 0.3333
##    Detection Prevalence : 0.3333
##       Balanced Accuracy : 0.8333
##
##        'Positive' Class : 1
##
```

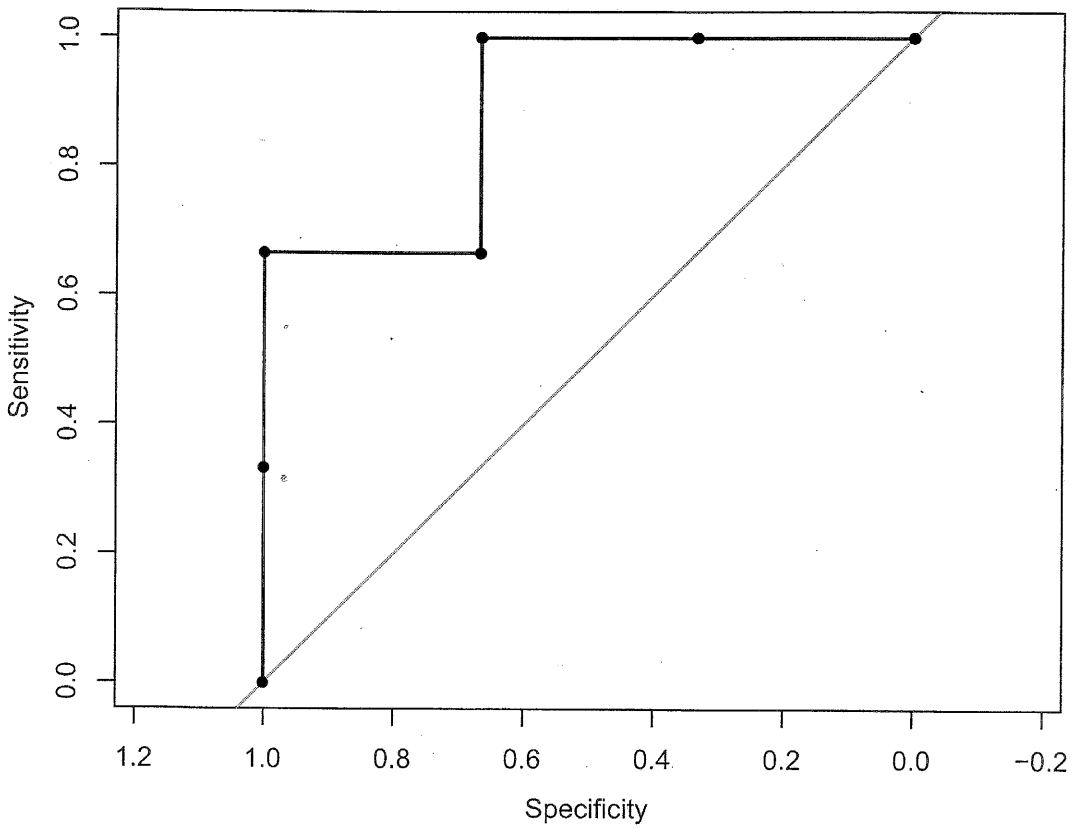
In particular, the sensitivity and specificity of the classifier are 0.6667 and 1, respectively.

(b) We can repeat the process in part (a) for each possible cutoff in the unit interval. The details are shown in Table 4.1. The ROC curve of the classifier is plotted in Figure 4.1.1 (with the seven points connected by straight lines) and the AUC is calculated to be 0.8889. □

Case	Firms Predicted to Default	Confusion Matrix			Point on ROC curve: (1 − Specificity, Sensitivity)
$0.80 < l \leq 1$	None		Reference		$(0, 0)$
		Prediction	0	1	
		0	3	3	
		1	0	0	
$0.75 \leq l < 0.80$	6		Reference		$(0, 1/3)$
		Prediction	0	1	
		0	3	2	
		1	0	1	
$0.60 \leq l < 0.75$	5, 6		Reference		$(0, 2/3)$
		Prediction	0	1	
		0	3	1	
		1	0	2	
$0.50 \leq l < 0.60$	4, 5, 6		Reference		$(1/3, 2/3)$
		Prediction	0	1	
		0	2	1	
		1	1	2	
$0.40 \leq l < 0.50$	3, 4, 5, 6		Reference		$(1/3, 1)$
		Prediction	0	1	
		0	2	0	
		1	1	3	
$0.25 \leq l < 0.40$	2, 3, 4, 5, 6		Reference		$(2/3, 1)$
		Prediction	0	1	
		0	1	0	
		1	2	3	
$0 \leq l < 0.25$	All		Reference		$(1, 1)$
		Prediction	0	1	
		0	0	0	
		1	3	3	

Table 4.1: A series of confusion matrices corresponding to different cutoffs in Example 4.1.9.

```
library(pROC)
roc.model <- roc(obs, pred)
plot(roc.model)
points(c(1, 1, 1, 2/3, 2/3, 1/3, 0), c(0, 1/3, 2/3, 2/3, 1, 1, 1),
       pch = 16)
```




```
auc(roc.model)

## Area under the curve: 0.8889
```

Figure 4.1.1: The ROC curve in Example 4.1.9.