# Week 3 Notes: GLM Theory and Analysis of Deviance

This week, we will learn about some of the basic theory underlying GLMs, as well as how these models tend to be fit in practice. While somewhat abstract, it's worth noting that the topics presented here have highly pragmatic use cases, from constructing hypothesis tests to designing inference algorithms, and I think it's important that one see these things at least once in their life.

After this week, we will look at some applications of GLMs in some simple datasets.

## 1 The Likelihood of a GLM

GLMs are fit in `R` using *likelihood based inference.* The likelihood function for a GLM, given data $\mathcal{D} = \{(Y_i, X_i) : i = 1, \ldots, N\}$ is given by

$$L(\beta, \phi) = \prod_{i=1}^{N} \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(\mathrm{Y}_i; \phi/\omega_i)\right\},$$

where we define $\theta_i \equiv (b')^{-1}(\mu_i)$ and $\mu_i \equiv g^{-1}(X_i^\top \beta)$. We can then derive the *score function* $s(\beta, \phi) = \frac{\partial}{\partial\beta} \log L(\beta, \phi)$ as

$$s(\beta, \phi) = \sum_{i=1}^{N} \frac{\partial}{\partial\beta} \frac{Y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(Y_i; \phi/\omega_i).$$

Again, we will write $\frac{\partial}{\partial\beta} F(\beta)$ for the gradient of $F$ and $\frac{\partial^2}{\partial\beta\partial\beta^\top} F(\beta)$ for the Hessian matrix.

### Exercise 1: Deriving the Score

Using the chain rule $\frac{\partial}{\partial\beta} = \frac{\partial}{\partial\theta} \times \frac{\partial\theta}{\partial\mu} \times \frac{\partial\mu}{\partial\beta}$, show that

$$s(\beta, \phi) = \sum_{i=1}^{N} \frac{\omega_i (Y_i - \mu_i) X_i}{\phi V(\mu_i) g'(\mu_i)}.$$

Show also that, for the canonical link, we have $g'(\mu_i) = V(\mu_i)^{-1}$ so that this reduces to

$$s(\beta, \phi) = \sum_{i=1}^{N} \frac{\omega_i (Y_i - \mu_i) X_i}{\phi}.$$

**Hint:** recall that $\frac{d}{dx} g^{-1}(x) = \frac{1}{g'\{g^{-1}(x)\}}$.

Note for posterity that the MLE is justified, in large part, because it is the sample solution $s(\widehat{\beta}, \phi) = 0$ to the population-level estimating equation $\int s(\beta, \phi)\, f_0(\mathbf{y} \mid \mathbf{X}) d\mathbf{y} = 0$ (here, $\mathbf{X}$ denotes the design matrix and $f_0$ denotes the true conditional density of $\mathbf{Y} = (Y_1, \ldots, Y_N)$. Also note that this estimating equation holds *even when the model is misspecified!* We only need the mean structure (not the exponential dispersion family) for the population-level estimating equation to hold, which suggests that the MLE $\widehat{\beta}$ might still be a good estimator even when the model is misspecified. More on this when we study $M$-estimation.

### Exercise 2: Deriving the Fisher Information

We define the *Fisher Information* to be

$$\mathcal{I}(\beta, \phi) = -\mathbb{E}\left\{ \frac{\partial^2}{\partial\beta \partial\beta^\top} \log L(\beta, \phi) \mid \beta, \phi \right\}.$$

The Fisher information plays an important role in inference for GLMs. The “observed” Fisher information is also used,

$$\mathcal{J}(\beta, \phi) = -\frac{\partial^2}{\partial\beta\partial\beta^\top} \log L(\beta, \phi).$$

In addition to being easier to evaluate, using $\mathcal{J}$ has been argued to be the right-thing-to-do™. In any case, show that

$$\langle \mathcal{J}(\beta, \phi) \rangle_{jk} = \frac{1}{\phi} \sum_{i=1}^{N} X_{ij} X_{ik} \left\{ \frac{\omega_i}{V(\mu_i) g'(\mu_i)^2} - \frac{\omega_i (Y_i - \mu_i)}{g'(\mu_i)} \left( \frac{\partial}{\partial\mu_i} \frac{1}{V(\mu_i) g'(\mu_i)} \right) \right\}$$

and

$$\langle \mathcal{I}(\beta, \phi) \rangle_{jk} = \frac{1}{\phi} \sum_{i=1}^{N} X_{ij} X_{ik} \frac{\omega_i}{V(\mu_i) g'(\mu_i)^2}$$

where $\langle A \rangle_{ij}$ denotes the $(i, j)^{\text{th}}$ element of the matrix $A$. Show also that $\mathcal{I}(\beta, \phi) = \mathcal{J}(\beta, \phi)$ when the canonical link is used.

From the above exercise, notice that the Fisher information has the familiar form

$$\mathcal{I}^{-1} = \phi(\mathbf{X}^\top D \mathbf{X})^{-1}$$

where $D$ is a diagonal matrix with entries $\omega_i / \{V(\mu_i) g(\mu_i)^2\}$. Similarly, we can write $\mathcal{J}^{-1} = \phi(\mathbf{X}^\top \widetilde{D} \mathbf{X})^{-1}$ for some diagonal matrix $\widetilde{D}$. Compare this with the linear model, which has inverse Fisher information $\mathcal{I}^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

## 2 Aside: Likelihood-Based Inference

In this section, we will briefly refresh our memories on the theory underlying likelihood methods. For simplicity, consider data $\mathcal{D} = \{Z_i : i = 1, \ldots, N\}$ with $Z_i$'s iid according to some density $f(z \mid \theta_0)$ where $\{f(\cdot \mid \theta : \theta \in \Theta)\}$ is a family of distributions satisfying some (unstated) regularity conditions. We define the log-likelihood, score, and Fisher information with the equations

$$\ell(\theta) = \sum_{i=1}^N \log f(Z_i \mid \theta), \quad s(\theta) = \frac{\partial}{\partial \theta} \ell(\theta), \quad \text{and} \quad \mathcal{I}(\theta) = -\mathbb{E}\left\{ \frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta) \mid \theta \right\}.$$

The following identities are fundamental to likelihood inference

$$\mathbb{E}\{s(\theta) \mid \theta\} = \mathbf{0}, \qquad \text{and} \qquad \operatorname{Var}\{s(\theta) \mid \theta\} = \mathcal{I}(\theta).$$

We will study three types of methods for constructing inference procedures from the likelihood: Score methods, Wald methods, and likelihood ratio (LR) methods.

**Exercise 3: Score Methods**

Using the multivariate central limit theorem, show that

$$s(\theta_0) \stackrel{\bullet}{\sim} \text{Normal}\{0, \mathcal{I}(\theta_0)\},$$

but only if we plug in the true value $\theta_0$ *Note:* this asymptotic notation means that $X \stackrel{\bullet}{\sim} \text{Normal}(\mu, \Sigma)$ if-and-only-if $\Sigma^{-1/2}(X - \mu) \to \text{Normal}(0, \mathrm{I})$ in distribution.

**Exercise 4: Wald Methods**

Using Taylor's theorem, we have

$$s(\theta_0) = s(\widehat{\theta}) - \mathcal{J}(\theta^\star)(\theta_0 - \widehat{\theta}) = -\mathcal{J}(\theta^\star)(\theta_0 - \widehat{\theta}).$$

where $\theta^\star$ lies on the line segment connecting $\theta_0$ and $\widehat{\theta}$. Now, assume that we know somehow that $\widehat{\theta}$ is a *consistent* estimator of $\theta_0$. Show that

$$\widehat{\theta} \stackrel{\bullet}{\sim} \text{Normal}(\theta_0, \mathcal{I}(\theta_0)^{-1}).$$

**Note:** you do not need to give a completely rigorous proof. In particular, you can assume that, if $\theta_N \to \theta_0$, then $\frac{1}{N}\mathcal{J}(\theta_N) \to i(\theta_0)$ where

$$i(\theta_0) = -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta\partial\theta^\top}\log f(Z_i \mid \theta)\right\} = \mathcal{I}(\theta_0)/N.$$

### Exercise 5: Likelihood Ratio Methods

Consider the Taylor expansion

$$\ell(\theta_0) = \ell(\widehat{\theta}) + s(\widehat{\theta})^\top(\theta_0 - \widehat{\theta}) - \frac{1}{2}(\theta_0 - \widehat{\theta})^\top \mathcal{J}(\theta^\star)(\theta_0 - \widehat{\theta})$$

where $\theta^\star$ lies on the line segment connecting $\widehat{\theta}$ and $\theta_0$. Using Exercise 4, show that

$$-2\{\ell(\theta_0) - \ell(\widehat{\theta})\} \to \chi^2_P.$$

in distribution, where $P = \dim(\theta)$. Recall here that the $\chi^2_P$ distribution is the distribution of $\sum_{i=1}^P U_i^2$ where $U_1, \ldots, U_P \overset{\text{iid}}{\sim} \text{Normal}(0,1)$.

Exercise 4 provides the basis for "Wald" methods, while Exercise 3 provides the basis for "score" methods, and Exercise 5 provides the basis for "likelihood ratio" (LR) methods.

More generally one can show, using the same sorts of Taylor expansions used above, the following result.

### Theorem 1: Wilk's Theorem

Suppose that $\{f_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is a parametric family satisfying certain regularity conditions. Consider the null hypothesis $H_0 : \eta = \eta_0$, let $\widehat{\theta}_0$ denote the MLE obtained under the null model, and let $(\widehat{\theta}, \widehat{\eta})$ denote the MLE under the unrestricted model. Then, if $(\theta_0, \eta_0)$ denote the values of the parameters that generated the data (so that $H_0$ is true) then

$$-2\{\ell(\widehat{\theta}_0, \eta_0) - \ell(\widehat{\theta}, \widehat{\eta})\} \overset{\bullet}{\sim} \chi^2_D$$

where $D = \dim(\eta)$, as the amount of data tends to $\infty$.

The statement of the result above is deliberately vague; the regularity conditions are unstated, and what it means for the "amount of data" to tend to $\infty$ is not made precise. If our data is of the form $\mathcal{D} = \{(X_i, Y_i) : i = 1, \ldots, N\}$ then taking $N \to \infty$ will suffice, but there are other situations where the result will hold even if $N$ is fixed (with the data "tending to infinity" in other ways). An example in which this is the case is analysis of contingency tables, where $N$ denotes the number of cells. Another important requirement to get a result like this is that we should have the dimension of the parameters fixed, which is sometimes not the case even for old-school GLMs.

The above result will let us do things like perform hypothesis tests and construct confidence intervals.

**Note:** It is also possible to extend the score method to the case with parameter constraints, but we don't have time to do so; in this case, we expect a result of the form. My experience is that, for the most part, LR methods tend to be better than score methods (both are better than Wald methods), although this isn't universally true. Notice that LR methods require fitting both a constrained and unconstrained model. It turns out that score methods only require fitting the constrained model, which can occasionally be useful if the unconstrained model is difficult to fit.

## 3 Likelihood-Based Inference for GLMs

Bayesian inference for GLMs is quite straightforward — just fit the model with `stan_glm`, get the posterior samples, and interpret the results as you normally would.

For Frequentist inference, we can use likelihood-based methods to do things like conduct hypothesis tests. A convenient quantity to use is the *deviance* of a GLM.

**Definition 1: Deviance of a GLM**

Let $\mathcal{D} = \{(Y_i, X_i) : i = 1, \ldots, N\}$ be modeled with a GLM in the exponential dispersion family with canonical parameters $\theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(g^{-1}(X_i^\top \beta))$. Let $\theta = (\theta_1, \ldots, \theta_N)$ and let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_N)$ be the MLE of the $\theta$'s under our model. We define the *saturated model* as the model which has a separate parameter for all unique values of $x$ in $\mathcal{D}$:

$$f(y \mid x, \phi/\omega) = \exp\left\{\frac{y\theta_x - b(\theta_x)}{\phi/\omega} + c(y; \phi/\omega).\right\}.$$

The *residual deviance* of a model is defined by

$$D = -2\phi\left\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_x)\right\}$$

where $\ell(\theta) = \sum_{i=1}^N \dfrac{\omega_i(Y_i\theta_i - b(\theta_i))}{\phi}$ is the log-likelihood of $\theta$ and $\widehat{\theta}_{xi} = (b')^{-1}(Y_i)$. The *scaled deviance* is $D^\star = D/\phi$. **Note:** if $\phi$ is unknown, we generally replace $\phi$ with an estimate $\widehat{\phi}$ given by

$$\widehat{\phi} = \frac{1}{N-P}\sum_{i=1}^N \frac{\omega_i(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

### Exercise 6: Estimating the Dispersion

Show that the quantity

$$\widetilde{\phi} = \frac{1}{N} \sum_i \frac{\omega_i (Y_i - \mu_i)^2}{V(\mu_i)}$$

is unbiased for $\phi$. We don't use $\widetilde{\phi}$ because we don't know the $\mu_i$'s, so the modified denominator in $\widehat{\phi}$ compensates for the "degrees of freedom" used to estimate $\beta$.

**In words:** the scaled deviance is the LRT statistic for comparing the model with the saturated model which has the maximal number of model parameters in the GLM.

**What is the deviance good for?** The deviance is commonly used for two purposes.

1. It can be used as a goodness-of-fit statistic, testing to see whether the model under consideration would be rejected in favor of the saturated model. *From a modern perspective, this might be construed as a test of the parametric model against a nonparametric alternative, without making any smoothness assumptions on $\theta_x$.* This is a little bit tricky, however, since the saturated model has $P = N$ parameters, so the usual asymptotics (where $N$ is large relative to $P$) do not apply. In certain situations, however, one can show that $D^\star \overset{\bullet}{\sim} \chi^2_{N-P}$, as would be suggested by a naive application of Theorem 1.

2. If model $\mathcal{M}_0$ is a submodel of $\mathcal{M}_1$ then the LRT statistic for comparing these models is $D_0^\star - D_1^\star$. Under very weak conditions, we have $D_0^\star - D_1^\star \overset{\bullet}{\sim} \chi^2_K$ where $K$ is the difference in the number of parameters between the two models.

### Example 1: More Ships

The LRT in Theorem 1 can be performed using the `anova` function. We illustrate on the `ships` dataset from the previous week.

```r
## Load
ships <- MASS::ships

## Fit GLM (see previous notes)
ships_glm <- glm(
  incidents ~ type + factor(period) + factor(year),
  family = poisson,
  offset = log(service),
  data = dplyr::filter(ships, service > 0)
)

anova(ships_glm, test = "LRT")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: incidents

Terms added sequentially (first to last)

               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              33    146.328
type            4   55.439        29     90.889 2.629e-11 ***
factor(period)  1   20.786        28     70.103 5.135e-06 ***
factor(year)    3   31.408        25     38.695 6.975e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This table gives an *analysis of deviance*, which should feel quite similar to the analysis of variance that you are already familiar with. Let's first discuss the columns of this table:

- The first (unlabeled) column gives different terms of the model that describe the tests we will be doing. The `NULL` model refers to a model with only an intercept.
- The second column gives the degrees of freedom associated to each term. For example, there are five types, so there are four degrees of freedom for type.
- The third column, which is labeled `Deviance` is not the deviance as we have discussed it, but is the difference in the deviance of the current row and the previous row. Under the null hypothesis, `Deviance` will have a $\chi^2$ distribution with `Df` degrees of freedom.
- The third column, labeled `Resid. Df` gives the degrees of freedom for the comparison of the current model to the saturated model.
- The fourth column `Resid. Dev` gives the residual deviance, as we have defined it.
- The fifth column is the $P$-value of the test that takes the model from the previous row as the null and the current row as the alternative. It is computed using the first two columns.

Now, let's discuss the different models being compared.

(a) The `NULL` row corresponds to an intercept-only model.

(b) The `type` row corresponds to a model including only `type` and an intercept.

(c) The `period` row corresponds to a model including only `period`, `type`, and an intercept.

(d) The `year` row corresponds to a model with all terms included in the model.

The results of the analysis of deviance point clearly to the relevance of the individual terms. Unlike the output of `summary`, this provides the likelihood ratio test rather than Wald tests for the parameters, and the different variables have been helpfully grouped together (i.e., we have a single term of `type` with four degrees of freedom, rather than four separate coefficients).

The last residual deviance for the full model is 38.695, with a (naive) null distribution of $\chi^2_{25}$. This corresponds to a $P$-value of

```
pchisq(38.695, df = 25, lower.tail = FALSE)
```

```
[1] 0.0395148
```

This gives some evidence that the model lacks fit, and the model could be disproved in favor of the saturated model. As mentioned above, however, the deviance may not be close to a $\chi^2_{N-P}$ distribution in this case. For Poisson data, the key is that the individual counts for each observation should be largeish. But

```
print(ships$incidents)
```

```
 [1]  0  0  3  4  6 18  0 11 39 29 58 53 12 44  0 18  1  1  0  1  6  2  0  1  0
[26]  0  0  0  2 11  0  4  0  0  7  7  5 12  0  1
```

Hence, use of the deviance in this situation as a goodness of fit test seems questionable.

We can also construct likelihood-based confidence intervals for the parameters. If I want a confidence interval for (say) $\beta_1$, I can get one by *inverting* the test $H_0 : \beta_1 = \beta_{01}$ to get a confidence set

$$\{\beta_{01} : \text{The LRT fails to reject } H_0 : \beta_0 = \beta_{01}\}.$$

If the LRT has Type I error rate $\alpha$ for all $\beta_{01}$ then the above set is guaranteed to be a $100(1 - \alpha)\%$ confidence set. This is implemented in `R` with the function `confint`:

```
confint(ships_glm)
```

```
Waiting for profiling to be done...

                       2.5 %      97.5 %
(Intercept)      -6.84305161 -5.98968373
```

```
typeB            -0.88135891 -0.18353080
typeC            -1.37649167 -0.07452031
typeD            -0.67151807  0.47524605
typeE            -0.14346972  0.78520455
factor(period)75  0.15339419  0.61740478
factor(year)65    0.40752296  0.99512708
factor(year)70    0.48728088  1.15369754
factor(year)75   -0.01234169  0.90386446
```

**In general, you should use `anova` and `confint` rather than the output of `summary`.**

Lastly, rather than doing sequential tests like `anova` does, you can do a “leave-one-out” test using `drop1` as follows.

```
drop1(ships_glm, test = "LRT")
```

```
Single term deletions

Model:
incidents ~ type + factor(period) + factor(year)
               Df Deviance    AIC    LRT  Pr(>Chi)    
<none>              38.695 154.56                     
type            4   62.365 170.23 23.670 9.300e-05 ***
factor(period)  1   49.355 163.22 10.660  0.001095 ** 
factor(year)    3   70.103 179.97 31.408 6.975e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This tests each model term under the assumption that the other terms are in the model. For example, the row corresponding to `type` performs an LRT that tests a model with all the terms against an alternative that uses only `period` and `year`.

### Exercise 7: Challenger Revisit

Try these ideas out on the Challenger dataset. How do your conclusions differ (or not differ) from the Bayesian analysis?