LECTURE 11: GENERALIZED LINEAR MODELS AND GENERALIZED ADDITIVE MODELS

Text reference: Shalizi Chapter 12

Introduction: Two Familiar Regression Models

So far we have seen two parametric models for regression. Let $X \in \mathbb{R}^p$ be a vector of predictors.

1. In **linear regression**, we observe a real-valued response $Y \in \mathbb{R}$ and assume a linear model:

   ---

   for some coefficients $\beta \in \mathbb{R}^p$. Often we also assume that the data are normal:

   $$Y_i | X_i \sim \text{Normal}(\mu_i, \sigma^2)$$

   .

   ---

2. In **logistic regression**, we observe a binary response $Y \in \{0, 1\}$, and assume a logistic model:

   Because the $Y_i's$ are binary, the data are Bernoulli

   $$Y_i | X_i \sim \text{Bernoulli}(\mu_i).$$

- *What is the similarity between these two models?*

  Note that in the logistic setting, $\mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$. Therefore, in both settings, we are assuming that **a transformation of the conditional expectation** $\mathbb{E}(Y|X)$ is a **linear function of** x, i.e.,

  ---

  for some function g. The function g is called the **link** function. The function $\eta = \beta^T X$ is called the **linear predictor.**

- *What are the link functions in linear regression and logistic regression, respectively?*

Different transformations might be appropriate for different types of data. E.g., the identity transformation $g(u) = u$ is not really appropriate for logistic regression (why? Hint: Demo 10.1), and the logit transformation $g(u) = \log(u/(1-u))$ not appropriate for linear regression (why?), but each is appropriate in their own intended domain.

For a third data type, it is entirely possible that transformation neither is really appropriate. What do we do then? We think of another transformation g that is in fact appropriate, and this is the basic idea behind a generalized linear model.

Generalized Linear Models

Recall: We can write the logistic regression model as

$$\begin{aligned} Y_i|x_i &\sim \text{Bernoulli}(\mu_i) \\ g(\mu_i) &= X_i^T\beta \end{aligned}$$

where $g(z) = \text{logit}(z)$. The function g is an example of a **link** function and the Bernoulli is an example of an **exponential family**, which we explain below. We define a **generalized linear model** as any model in which

Note the GLMs includes both a *random component* that specifies a distribution for the outcome variable (conditional on X), and a *systematic (non-random) component* that relates a parameter to the predictors X through a link function.

A probability function (or probability density function) is said to be in the *exponential family* if there are functions $\eta(\theta)$, $B(\theta)$, $T(y)$ and $h(y)$ such that

$$f(y;\theta) = h(y)e^{\eta(\theta)T(y)-B(\theta)}.$$

Example 11.1. Let $Y \sim \text{Poisson}(\theta)$. Then

$$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!} = \frac{1}{y!} e^{y \log \theta - \theta}$$

and hence, this is an exponential family with $\eta(\theta) = \log \theta$, $B(\theta) = \theta$, $T(y) = y$, $h(y) = 1/y!$.

Example 11.2. Let $Y \sim \text{Binomial}(n, \theta)$. Then

$$f(y;\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} = \binom{n}{y} \exp\left\{ y \log\left(\frac{\theta}{1-\theta}\right) + n \log(1-\theta) \right\}.$$

In this case,

$$\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right), B(\theta) = -n \log(\theta)$$

and

$$T(y) = y, h(y) = \binom{n}{y}.$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector,then we say that $f(y;\theta)$ has exponential family form if

$$f(y;\theta) = h(y) \exp\left\{ \sum_{j=1}^{k} \eta_j(\theta) T_j(y) - B(\theta) \right\}.$$

Example 11.3. Consider the Normal family with $\theta = (\mu, \sigma)$. Now,

$$f(y;\theta) = \exp\left\{ \frac{\mu}{\sigma^2} y - \frac{y^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\}.$$

This is exponential with

$$\eta_1(\theta) = \frac{\mu}{\sigma^2},\ T_1(y) = y$$

$$\eta_2(\theta) = -\frac{1}{2\sigma^2},\ T_2(y) = y^2$$

$$B(\theta) = \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right),\ h(y) = 1.$$

Example 11.4. [Normal Regression] Here, $Y_i|X_i \sim N(\mu_i, \sigma^2)$ and the link $g(\mu_i) = \mu_i$ is the indentity function.

Example 11.5. [Logistic Regression] Here, $Y_i|X_i \sim \text{Bernoulli}(\mu_i)$ and $g(\mu_i) = \text{logit}(\mu_i)$.

Example 11.6. [Poisson Regression] This is often used when the outcomes are counts. Here, $Y_i|X_i \sim \text{Poisson}(\mu_i)$ and the usual link function is $g(\mu_i) = \log(\mu_i)$.

Although many link functions could be used, there are default link functions that are standard for each family. Here they are:

Distribution	Link	Inverse Link (Regression Function)
Normal	Identity $g(\mu) = \mu$	$\mu = x^T\beta$
Bernoulli	Logit $g(\mu) = \text{logit}(\mu)$	$\mu = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$
Poisson	Log $g(\mu) = \log(\mu)$	$\mu = e^{x^T\beta}$
Gamma	Inverse $g(\mu) = 1/\mu$	$\mu = \frac{1}{x^T\beta}$

Fitting Generalized Linear Models

We usually fit GLMs by maximum likelihood. In general, there is no closed form expression for the solution to the maximization problem (as there is for linear regression under the assumption of normality). Fortunately, we can as for logistic regression use iteratively reweighted least squares (IRLS).

In R you type:

```
glm(y ~ x, family= xxxx)
```

where xxxx is Normal, binomial, poisson etc. R will assume the default link. Inference is carried out by using (large-sample) normal theory for MLEs.

R Demo 11.1: Does smoking increase the risk of coronary heart disease?

This is a famous data set collected by Sir Richard Doll in the 1950's. I am following example 9.2.1 in Dobson. The data are on smoking and number of deaths due to coronary heart disease. Here are the data:

Age	Smokers		Non-smokers	
	Deaths	Person-years	Deaths	Person-years
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

Questions of interest are:

1. Is the death rate higher for smokers than non-smokers?
2. If so, by how much?
3. Is the differential effect related to age?

(a) Visualize the data and find a good model for the relationship between the rate of deaths due to coronary heart disease and the other covariates (age and smoking status).

(b) Is the death rate higher for smokers than non-smokers, and if so by how much? Include measures of uncertainty. Discuss your results.

(c) Assess the model's goodness of fit. Discuss your results. □

Poisson regression is also appropriate for rate data, where the rate is a count of events occurring for a particular unit of observation, divided by some measure of that unit's exposure. For example, biologists might count the number of tree species in a forest, and the rate would be the number of species per square kilometre. Demographers may model death rates in geographic areas as the count of deaths divided by "person-years". This is defined as the number of persons at risk over the number of years. It is important to note that event rates can be calculated as events per units of varying size. For instance, in these data, person-years vary. As the people get older there are fewer at risk. At the time the data were collected most people smoked, so there were fewer person-years for non-smokers. In general, exposure is with respective to unit (be it area, person-years or time). In Poisson regression this is handled using an **offset**, where the exposure variable enters on the right-hand side of the equation, but with the coefficient (for log(exposure)) constrained to 1.

This makes sense because our model is that the expected rate of deaths equals $\exp(x^t\beta)$, i.e.,

---

Thus we pull $\log(\text{exposure})$ to the right-hand-side of the equation and fit a

Poisson regression model. The offset option allows us to include $\log(\text{exposure})$ in the model without estimating a β coefficient for exposure.

We use the midpoint of each age group as the age of everyone in the group. A plot of deaths by age, exhibits an obvious increasing relationship with age which shows some hint of nonlinearity. The increase may differ between smokers and non-smokers so we will include an interaction term.

```
> ### page 155 dobson
> deaths = c(32,104,206,186,102,2,12,28,28,31)
> age    = c(40,50,60,70,80,40,50,60,70,80)
> py     = c(52407,43248,28612,12663,5317,
+             18790,10673,5710,2585,1462)
> smoke  = c(1,1,1,1,1,0,0,0,0,0)
> agesq  = age*age
> sm.age = smoke*age
#
#Notice the use of the offset for person-years below
#
> out = glm(deaths~smoke+age+agesq+sm.age,offset=log(py),family=poisson)
> summary(out)

Deviance Residuals:
       1         2         3         4         5         6         7         8         9
 0.43820  -0.27329  -0.15265   0.23393  -0.05700  -0.83049   0.13404   0.64107  -0.41058
      10
-0.01275

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.970e+01  1.253e+00 -15.717  < 2e-16 ***
smoke        2.364e+00  6.562e-01   3.602 0.000316 ***
age          3.563e-01  3.632e-02   9.810  < 2e-16 ***
agesq       -1.977e-03  2.737e-04  -7.223 5.08e-13 ***
sm.age      -3.075e-02  9.704e-03  -3.169 0.001528 **
```

```
---

    Null deviance: 935.0673  on 9  degrees of freedo
Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 66.703
```

Based on the p-value from the Wald tests above, smoking appears to be quite important (but keep the usual causal caveats in mind).

Suppose we want to compare smokers to non-smokers for people of age 40 years. The estimated Poisson model is

---

---

and hence to compare rates of smokers to non-smokers with $x = 40$ in our population

$$\begin{aligned}\frac{\mathbb{E}(Y|\text{smoker}, \text{age} = 40)}{\mathbb{E}(Y|\text{non} - \text{smoke}, \text{age} = 40)} &= \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 + 40\widehat{\beta}_2 + 1600\widehat{\beta}_3 + 40\widehat{\beta}_4 + \log(52407)\}}{\exp\{\widehat{\beta}_0 + 40\widehat{\beta}_2 + 1600\widehat{\beta}_3 + \log(18790)\}} \\ &= e^{\widehat{\beta}_1 + 40\widehat{\beta}_4 + \log(52407) - \log(18790)}\end{aligned}$$

This gives us the ratio of the expected number of deaths in a population similar to this study. But we are only interested in the rate parameter relevant to an individual, so we now want to drop the person-years terms. The appropriate ratio of rates is

$$e^{\widehat{\beta}_1 + 40\widehat{\beta}_4} = 3.1.$$

suggesting that smokers in this group have a death rate due to coronary heart disease that is 3.1 times higher than non-smokers.

Let's get a confidence interval for this:

We are interested in $\psi = e^{\gamma}$. The confidence interval is

$$(e^a, e^b).$$

In R:

```
> summ = summary(out)
> v    = summ$dispersion * summ$cov.unscaled
#summ$dispersion is 1 unless we allow "over dispersion"
#relative to the model. This is a topic I skipped over.
#
> print(v)
              (Intercept)         smoke           age         agesq        sm.age
(Intercept)  1.5711934366 -4.351992e-01 -4.392812e-02  2.998070e-04  6.445856e-03
smoke       -0.4351992084  4.306356e-01  7.424636e-03 -1.601373e-05 -6.280941e-03
age         -0.0439281178  7.424636e-03  1.318857e-03 -9.633853e-06 -1.144205e-04
agesq        0.0002998070 -1.601373e-05 -9.633853e-06  7.489759e-08  2.700594e-07
sm.age       0.0064458558 -6.280941e-03 -1.144205e-04  2.700594e-07  9.416983e-05
> ell  = c(0,1,0,0,40)
> gam  = sum(ell*out$coef)
> print(exp(gam))
[1] 3.106274
> se   = sqrt(ell %*% v %*% ell)
> ci   = exp(c(gam- 2*se, gam+ 2*se))
> print(round(ci,2))
[1]  1.77 5.45
```

The result is that the rate is 2 to 5 times higher for smokers than non-smokers at age 40. Since heart disease is much more common than lung cancer, the risk of smoking has a bigger impact on public health for heart disease than smoking. □

There is a formal way to test the model for goodness of fit. As with logistic regression we can compute the deviances. Recall that the log likelihood for a Poisson is of the form

$$\ell(\theta) = y \log(\theta) - \theta.$$

The deviance residuals are defined as

$$\begin{aligned} d_i &= \text{sign}(Y_i - \widehat{Y}_i)\sqrt{2[\ell(Y_i) - \ell(\widehat{Y}_i)]} \\ &= \text{sign}(Y_i - \widehat{Y}_i)\sqrt{2[(Y_i \log(Y_i/\widehat{Y}_i) - (Y_i - \widehat{Y}_i)}. \end{aligned}$$

The **deviance** is defined as

$$D = \sum_i d_i^2.$$

This statistic is approximately distributed χ^2_{n-p-1} where p is the number of covariates. If D is larger than expected (i.e., the p-value is small) this means that the Poisson model with the covariates included is not sufficient to explain the data.

For these data the model appears to fit well.

```
> print(1-pchisq(out$deviance,df=5))
[1] 0.8969393
```

Generalized Additive Models

- Recall that we could augment the standard linear model

$$\mathbb{E}(Y_i|X_i = x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots n,$$

  with the additive model

$$\mathbb{E}(Y_i|X_i = x_i) = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}),$$

  where each r_j is an arbitrary (univariate) regression function, $j = 1, \ldots p$

- The same extension can be applied to the generalized linear model, yielding a **generalized additive model**. The only change is in the function η, which we now define as:

- That is, the link function g now connects the the mean of the exponential family distribution $\mu_i = \mathbb{E}(Y_i|X_i = x_i)$ to the function η_i via:

- In the Gaussian case, the above reduces to the additive model that we have already studied. In the Bernoulli case, with canonical link, this model becomes

$$\log \frac{p_i}{1-p_i} = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}), \quad i = 1, \ldots n,$$

were $p_i = \mathbb{P}(Y_i = 1|X_i = x_i)$, $i = 1, \ldots n$, which gives us **additive logistic regression**. In the Poisson case, with canonical link, this becomes

$$\log \mu_i = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}), \quad i = 1, \ldots n,$$

which is called **additive Poisson regression**

- Generalized additive models are a marriage of two worlds: *generalized linear models* and *additive models*, which together offer a very flexible, powerful platform for modeling. The generalized linear model allows us to account for different types of outcomes Y_i, $i = 1, \ldots n$, and the additive model allows us to consider a transformation of the mean $\mathbb{E}(Y_i|x_i)$ as a nonlinear (but additive) function of $x_i \in \mathbb{R}^p$, $i = 1, \ldots n$