Exponential Families and Generalized Linear Models

Mark M. Fredrickson (mfredric@umich.edu)

Computational Methods in Statistics and Data Science (Stats 406)

Generalized Linear Models

OLS: A quick review

We've mostly motivated **ordinary least squares** (OLS) as an **optimization problem**.

- We fixed the mean function as being **linear in a single argument**: $\mu(a) = a$
- We combined the p predictors and parameters using a **linear combination**: $\eta(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}^T \boldsymbol{\beta}$
- Then we asked, if I were to use **squared error loss**, what is my **optimal** $\hat{\beta}$?

We found (by taking derivatives) that

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} (\mathbf{y} - \mathbf{X}^T \beta)^T (\mathbf{y} - \mathbf{X}^T \beta) \Rightarrow \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

Before getting any more complicated, let's stop to ask, **why would we want squared error loss?**

Alternative Motivation: Maximum Likelihood

A more typical (at least for a stats class) motivation for OLS usually starts with:

$$y_i \sim N(\mathbf{x_i}^T \boldsymbol{\beta}, \sigma)$$

(all independent with the same variance)

To perform, **maximum likelihood estimation**, we find likelihood for $\boldsymbol{\beta}$ is:

$$L(\beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T \boldsymbol{\beta})^2\right\}$$

Taking the log and discarding terms without β, yields:

$$l^*(\beta) = -\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T \boldsymbol{\beta})^2 = -R(\beta)$$

Maximizing $-R(\beta)$ is the same as **minimizing** $R(\beta)$.

Other issues with OLS

Previously, we asked the question: what if I think $\mathrm{E}(Y \mid \mathbf{x})$ is **not linear in $\mathbf{x}$**?

We found we could get lots of flexibility by replacing $\mathbf{x}$ with $f(\mathbf{x})$ in $f(\mathbf{x})^T\boldsymbol{\beta}$.

Notice that this formulation may not be linear in $\mathbf{x}$, but it **remains linear in $\boldsymbol{\beta}$**.

Implications:

- Cannot model **non-linear parameters**
- Since $\mathbf{x}^T\boldsymbol{\beta} \in (-\infty, \infty)$ we cannot limit $\mathrm{E}(y \mid \mathbf{x})$ to a particular range (e.g. $[0, 1]$).
- OLS $\hat{\boldsymbol{\beta}}$ is not the maximum likelihood estimate for other distributions of $Y \mid \mathbf{x}$.

Generalized Linear Models: Basics

Suppose we are able to write:

$$\mathrm{E}\,(Y \mid \mathbf{x}) = G^{-1}(\mathbf{x}^T\boldsymbol{\beta})$$

- G is the **link function** that relates the conditional mean to **linear predictors** $\mathbf{x}^T\boldsymbol{\beta}$:

$$G(\mathrm{E}\,(y \mid \mathbf{x})) = \mathbf{x}^T\boldsymbol{\beta} \iff G^{-1}(\mathbf{x}^T\boldsymbol{\beta}) = \mathrm{E}\,(Y \mid \mathbf{x})$$

- This allows $\mu(a)$ to be **non-linear** (provided it is invertable)
- If we have a distribution for $Y \mid \mathbf{x}$, **maximum likelihood estimates** can be produced similar to OLS.
- Particular distributions suggest **loss functions** even when $\mu(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$

Exponential Family Distributions

If we want to express the **conditional mean** as a **increasing function**, what distributions have that quality?

The **exponential (dispersion) family** is defined as having densities (or PMFs) of the form:

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

where a, b, c are **functions**.

We say θ is the **parameter of interest** and ψ is a **nuisance parameter**. Both can be vectors.

While we will see many distributions are EDF, **not all distributions can be factored as above**. The Laplace distribution is an example.

Example: Normal distribution

$$
\begin{aligned}
f(y, \mu, \sigma^2) &= \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{1}{2}\log(2\pi\sigma^2)\right) \exp\left(-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)
\end{aligned}
$$

Then we can translate to the **canonical notation** using:

- $\theta = \mu$, $b(\theta) = \theta^2/2$
- $\psi = \sigma^2$, $a(\psi) = \psi$
- $c(y, \psi) = -(y^2 + \psi \log(2\pi\psi))/(2\psi)$

Example: Bernoulli distribution

$$
\begin{aligned}
P(Y=y) &= \mu^y(1-\mu)^{1-y} \\
&= \exp\{y\log(\mu) + (1-y)\log(1-\mu)\} \\
&= \exp\{y(\log(\mu) - \log(1-\mu)) + \log(1-\mu)\} \\
&= \exp\{y(\log(\mu/(1-\mu)) + \log(1-\mu)\}
\end{aligned}
$$

This suggests $\theta = \log(\mu/(1-\mu))$. Solving for μ, yields $\mu = e^\theta/(1+e^\theta)$:

$$
P(Y=y) = \exp\left\{y\theta + \log\left(1 - \frac{e^\theta}{1+e^\theta}\right)\right\}
$$

So we have

$$
b(\theta) = -\log\left(1 - \frac{e^\theta}{1+e^\theta}\right) = -\log\left(\frac{1+e^\theta-e^\theta}{1+e^\theta}\right) = \log(1+e^\theta)
$$

($a(\psi) = 1$ and $c(y,\psi) = 0$)

Deriving some useful facts about EFDs

Under some mild conditions on the functions a, b, and c, we will state without proof that:

$$E\left[\frac{\partial}{\partial\theta}\log f(Y,\theta,\psi)\right] = 0$$

$$E\left[\frac{\partial^2}{\partial^2\theta}\log f(Y,\theta,\psi)\right] = -E\left[\left(\frac{\partial}{\partial\theta}\log f(Y,\theta,\psi)\right)^2\right]$$

and use these to derive the **mean** and **variance** of EFDs.

EFD means

We previously stated the fact:

$$E\left(\frac{\partial}{\partial\theta}\log f(Y,\theta,\psi)\right) = 0$$

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log f(Y,\theta,\psi) &= \frac{\partial}{\partial\theta}\log\left[\exp\left(\frac{Y\theta - b(\theta)}{a(\psi)} + c(Y,\psi)\right)\right] \\
&= \frac{\partial}{\partial\theta}\frac{Y\theta - b(\theta)}{a(\psi)} + c(Y,\psi) \\
&= \frac{Y - b'(\theta)}{a(\psi)}
\end{aligned}$$

$$E\left(\frac{Y - b'(\theta)}{a(\psi)}\right) = 0 \Rightarrow E(Y) = b'(\theta)$$

EFDs variance

We have the second fact,

$$E\left[\frac{\partial^2}{\partial^2\theta}\log f(Y,\theta,\psi)\right] = -E\left[\left(\frac{\partial}{\partial\theta}\log f(Y,\theta,\psi)\right)^2\right]$$

$$\frac{\partial^2}{\partial^2\theta}\log f(Y,\theta,\psi) = \frac{\partial}{\partial\theta}\frac{Y-b'(\theta)}{a(\psi)} = \frac{-b''(\theta)}{a(\psi)} \Rightarrow E\left[\left(\frac{\partial}{\partial\theta}\log f(Y,\theta,\psi)\right)^2\right] = \frac{b''(\theta)}{a(\psi)}$$

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}\left([Y-\mathrm{E}(Y)]^2\right) = \mathrm{E}\left(\left[Y-b'(\theta)\right]^2\right)\\ &= \mathrm{E}\left(a(\psi)^2\left[\frac{Y-b'(\theta)}{a(\psi)}\right]^2\right)\\ &= a(\psi)^2\frac{b''(\theta)}{a(\psi)} = b''(\theta)a(\psi)\end{aligned}$$

Checking results

Normal:

- $b(\theta) = \theta^2/2$, so $b'(\theta) = \theta$. We defined $\theta = \mu$, so $E(Y) = \mu$.
- $a(\psi) = \psi$, so $b''(\theta)a(\psi) = \psi$. We defined $\psi = \sigma^2$.

Bernoulli:

- Recall $b(\theta) = \log(1 + e^\theta)$, so

$$b'(\theta) = \frac{e^\theta}{1 + e^\theta}$$

we defined $\theta = \log(\mu/(1 - \mu))$

$$\frac{\mu/(1-\mu)}{1 + \mu/(1-\mu)} = \frac{\mu/(1-\mu)}{1/(1-\mu)} = \mu$$

Connection to the link function

Recall we want to model:

$$\mathrm{E}\,(y \mid \mathbf{x}) = G^{-1}(\mathbf{x}^T\boldsymbol{\beta})$$

We know that for the exponential family,

$$\mathrm{E}\,(y \mid \theta) = b'(\theta) = G^{-1}(\mathbf{x}^T\boldsymbol{\beta})$$

This is known as the **canonical link function**.

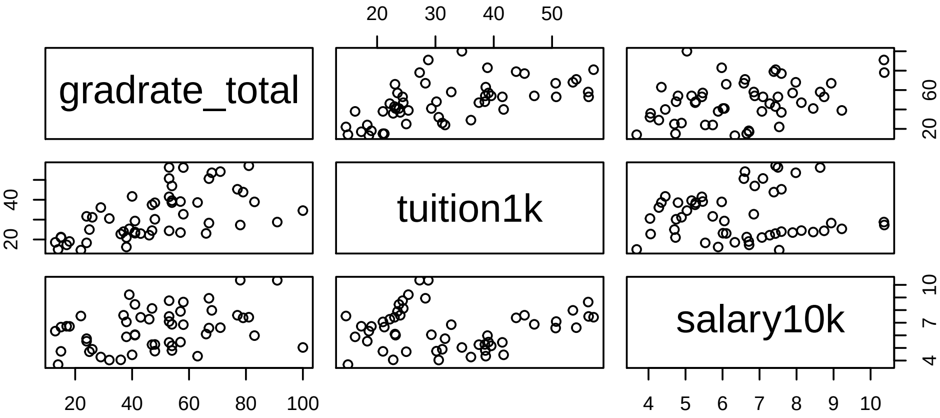
- Normal: $b'(\theta) = \theta$, so $G^{-1}(\mathbf{x}^T\boldsymbol{\beta}) = \mathbf{x}^T\boldsymbol{\beta}$ (we saw this with OLS!).
- Bernoulli: $b'(\theta) = e^{\theta}/(1 + e^{\theta})$, so we have

$$G^{-1}(\mathbf{x}^T\boldsymbol{\beta}) = e^{\mathbf{x}^T\boldsymbol{\beta}}/(1 + e^{\mathbf{x}^T\boldsymbol{\beta}})$$

(this is also called the "logistic link function")

Graduating at Least 50%

Recall our educational data:



Model

Let's focus on graduating more than 50% of students (G):

```
> edu_analysis$grad50 <- edu_analysis$gradrate_total >= 50
```

We'll model the mean of G given **salary** and **tuition**:

$$\mathrm{E}\,(G \mid t, s) = P(G = 1 \mid t, s) = \frac{\exp(\beta_0 + \beta_1 t + \beta_2 s)}{1 + \exp(\beta_0 + \beta_1 t + \beta_2 s)}$$

This is the **canonical link** function for the **binomial distribution**.

R

We can use the glm function with the a binomial family:

```
> mod50 <- glm(grad50 ~ tuition1k + salary10k,
+              data = edu_analysis,
+              family = binomial())
> coef(mod50)
```

```
(Intercept)   tuition1k   salary10k
   -15.4287      0.2866      0.9705
```

Interpreting Parameters

Recall when we discussed OLS, we used the idea of **taking partial derivatives** to understand how the mean function changed with the predictors.

The same idea applies to GLMs, but is often more complicated:

$$\frac{\partial}{\partial x_j} G^{-1}(\mathbf{x}^T \boldsymbol{\beta}) = \beta_j \, \frac{d}{du} G^{-1}(u) \Big|_{u=\mathbf{x}^T \boldsymbol{\beta}}$$

Interpretation: β_j tells us **direction** of mean increase (some care needed depending on G^{-1} monotonically increasing or decreasing), but the **amount of increase** also depends on $\mathbf{x}^T \boldsymbol{\beta}$ and the derivative of $G^{-1}(u)$.

Notice: $\beta_j = 0$ if and only if the conditional mean does not depend on x_j.

Canonical Binomial Link

For the binomial distribution, we have the inverse link function:

$$G^{-1}(u) = \frac{e^u}{1+e^u} \Rightarrow \frac{d}{du}G^{-1}(u) = \frac{e^u}{(1+e^u)^2}$$

So a one unit change in x_j leads to

$$\beta_j \frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{(1+e^{\mathbf{x}^T\boldsymbol{\beta}})^2}$$

Not the most easily interpreted quantity.

Working on the link scale

We've been mostly using **the inverse link function**, but we could also consider the **link function**:

$$g(\mathrm{E}(Y \mid \mathbf{x})) = \mathbf{x}^T \boldsymbol{\beta}$$

In this case, a one unit change in $\mathbf{x}$ leads to β_j **change in** $g(\mu(\mathbf{x}))$.

E.g., binomial case the link function works on the **log-odds of** $P(Y = 1)$:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

Interpreting Graduation Model

```
> coef(mod50)
(Intercept)   tuition1k   salary10k
   -15.4287      0.2866      0.9705
```

Comparing two possible x

For OLS, the mean function was linear. E.g., $E(Y \mid x) = \beta_0 + \beta_1 x$. So two points that differed by $\delta = x_2 - x_1$ would have an **expected difference**

$$\mathrm{E}\left(Y_2 \mid x_2\right) - \mathrm{E}\left(Y_1 \mid x_1\right) = (\beta_0 + \beta_1 x_2) - (\beta_0 + \beta_1 x_1) = \beta(x_2 - x_1) = \delta\beta_1$$

(likewise for other β_j for $p > 1$) Idea: Express **change in conditional mean** at different vectors of predictors.

Usual technique:

- For all predictors, compute the sample means $\bar{\mathbf{x}}$
- Let $\mathbf{s_j}$ be 0 expect for the jth entry, which is the s. d. of x_j
- Compute: $\hat{\mu}(\bar{\mathbf{x}} + \mathbf{s_j}) - \hat{\mu}(\bar{\mathbf{x}})$

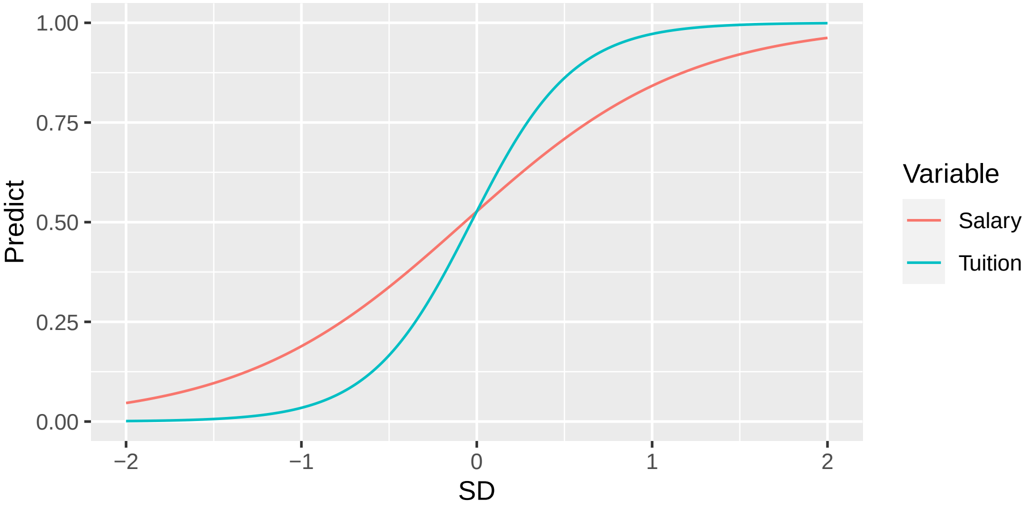
Other options include comparing specific quantiles or one unit changes in particular predictors.

Graduation Model

```
> mean_sd <- summarize(edu_analysis,
+   tuition1k_sd = sd(tuition1k),
+   tuition1k_mean = mean(tuition1k),
+   salary10k_sd = sd(salary10k),
+   salary10k_mean = mean(salary10k),
+ )
```

Notice the use of the **type argument**:

```
> sds <- seq(-2, 2, length.out = 1000)
> predict_tuition1k <- with(mean_sd,
+ predict(mod50, type = "response",
+   newdata = data.frame(tuition1k = tuition1k_mean + sds * tuition1k_sd,
+                        salary10k = salary10k_mean)))
> predict_salary10k <- with(mean_sd,
+ predict(mod50, type = "response",
+   newdata = data.frame(tuition1k = tuition1k_mean,
+                        salary10k = salary10k_mean + sds * salary10k_sd))
```


1.00
0.75
0.50
0.25
0.00
Predict
−2
−1
0
1
2
SD
Variable
Salary
Tuition

Example: Exponential Distribution

Suppose we think that $Y \mid \mathbf{x}$ is **Exponential with mean μ**.

The density function is:

$$\begin{aligned} f(y;\mu) &= \frac{1}{\mu}\exp\left\{-\frac{y}{\mu}\right\} \\ &= \exp\left\{-\frac{y}{\mu} - \log(\mu)\right\} \end{aligned}$$

To write in canonical form, let $\theta = -\frac{1}{\mu}$,

$$f(y;\theta) = \exp\{\theta y - \log(-1/\theta)\} \quad (a(\psi) = 1, c(y,\psi) = 0)$$

Cannonical link for Exponential Distribution

From the previous slide, we have

$$b(\theta) = \log(-1/\theta)$$

Recall that the mean of this distribution will be equal to $b'(\theta)$:

$$\frac{d}{d\theta} b(\theta) = \frac{d}{d\theta} - \log(-\theta) = \frac{1}{\theta}$$

Suppose we have predictors $\mathbf{x}$ and we want to model

$$\mathrm{E}\left(y \mid \mathbf{x}\right) = G^{-1}(\mathbf{x}^T \boldsymbol{\beta})$$

Relating $\theta = \mathbf{x}^T\boldsymbol{\beta}$, the **cannonical inverse link function** is

$$G^{-1}(\mathbf{x}^T \boldsymbol{\beta}) = b'(\mathbf{x}^T \boldsymbol{\beta}) = \left[\mathbf{x}^T \boldsymbol{\beta}\right]^{-1}$$

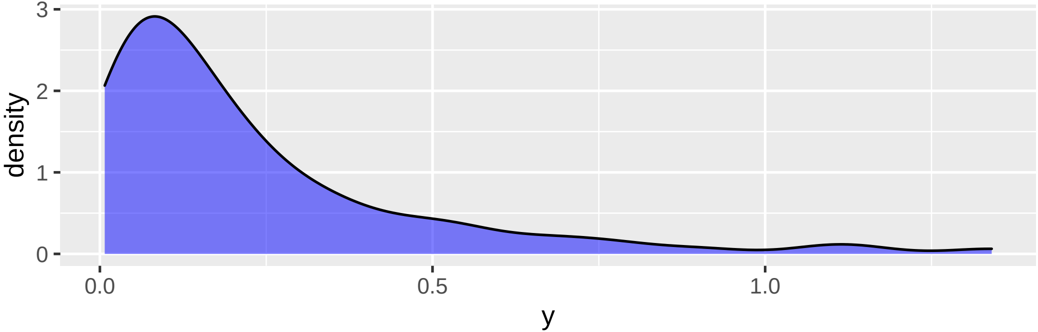
Simulating from the conditional-Exponential

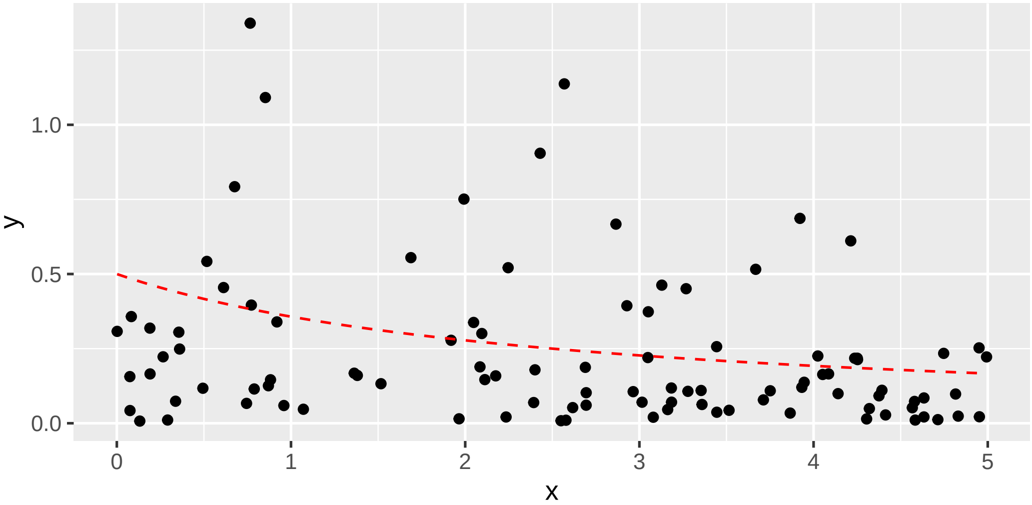
Let's pick some values for β_0 and β_1 to simulate from this distribution.

```
> b0 <- 2
> b1 <- 0.8
> mu <- function(x) {  1 / (b0 + b1 * x) } # inv. link
> x <- runif(100, 0, 5)
> y <- rexp(100, rate =  1 / mu(x)) # rate is 1/mean
```

Marginal distribution

Marginal distribution is not exponential!




1.0
0.5
0.0
y
0
1
2
3
4
5
x

Fitting the model

The exponential distribution is a special case of the **Gamma distribution**, with the shape parameter set to 1.

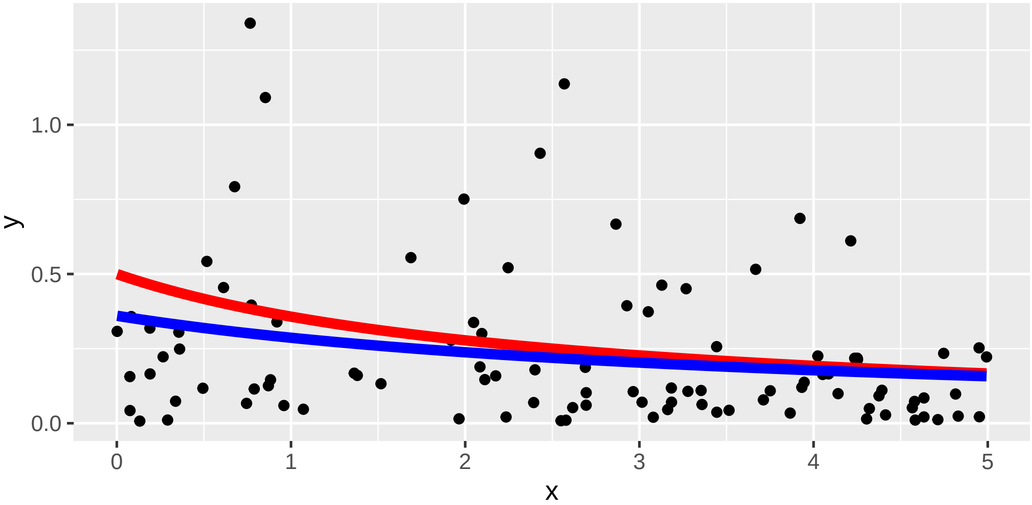
The mean of the Gamma distribution **does not depend on** k, so fitting a Gamma GLM is **equivalent** to fitting an exponential distribution.

```
> exp_mod <- glm(y ~ x, family = Gamma(link = "inverse"))
```

```
Call:  glm(formula = y ~ x, family = Gamma(link = "inverse"))

Coefficients:
(Intercept)            x
      2.777        0.714

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:              120
Residual Deviance: 114        AIC: -94
```



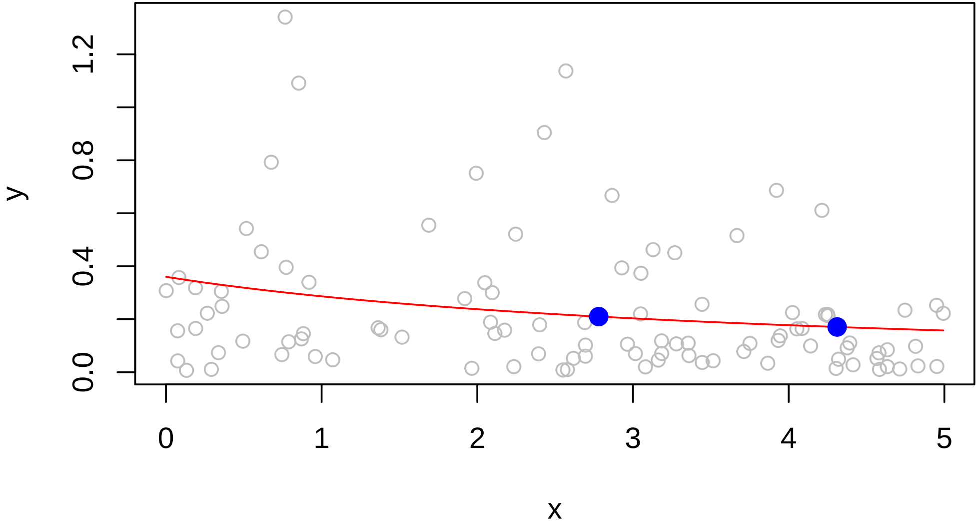
1.0
0.5
0.0
y
0
1
2
3
4
5
x

Example: Exponential Regression

```
> medx <- median(x)
> sdx <- sd(x)
> predict(exp_mod, newdata = data.frame(x = medx + sdx), type = "response")
+     predict(exp_mod, newdata = data.frame(x = medx), type = "response")

       1
-0.03921
```

NB: use the "response" type, otherwise you get predicted $\mathbf{x}^T\hat{\beta}$ ("linear predictors")


y
0.0
0.4
0.8
1.2
0
1
2
3
4
5
x

Large Sample Inference for β

One nice feature of **maximum likelihood estimators** is that they are **asymptotically Normal** (i.e., in large samples β is approximately multivariate Normal).

```
> summary(exp_mod)$coefficients

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   2.7771     0.7158   3.879 0.0001898
x             0.7141     0.3021   2.364 0.0200656

> confint(exp_mod)

             2.5 % 97.5 %
(Intercept) 1.5394  4.344
x           0.1216  1.312
```

Bootstrap Inference for GLMs

Of course, we will **emphasize computational approaches**. As usual, these fall into:

- Non-parametric bootstrap: Sample (without replacement) $(y, \mathbf{x})$ and refit model.
- Parametric bootstrap: Fit once, then sample from the model using $\hat{\beta}$.

The first can also be viewed from our **loss functions** perspective as estimating the parameter we would find by applying a loss function to a **population** (relaxes EDF assumption).

The second has the advantage that if the model is true, we can get **finite sample** distributions instead of the asymptotic MLE distributions.

NP Bootstrap

Use `index` argument to pick which coefficient you want:

```
> library(boot)
> boot_glm_np <- function(y, idx, x) {
+     ystar <- y[idx]
+     xstar <- x[idx]
+     coef(glm(ystar ~ xstar, family = Gamma(link = "inverse")))
+ }
> boot_exp_mod <- boot(y, boot_glm_np, R = 1000, x = x)
```

NP CIs

```
> boot.ci(boot_exp_mod, index = 1, type = "basic")$basic[, 4:5]
1.336 3.819
> boot.ci(boot_exp_mod, index = 2, type = "basic")$basic[, 4:5]
0.1611 1.2072
```

Parametric method

The `simulate` function will generate new samples:

```
> newy <- simulate(exp_mod, 1000)
> bscoefs <- apply(newy, 2, function(newy) {
+   coef(glm(newy ~ x, family = Gamma())) })
> quantile(bscoefs[1, ], c(0.025, 0.975)) ## percentile interval
```

```
 2.5% 97.5%
1.741 4.464
```

```
> quantile(bscoefs[2, ], c(0.025, 0.975)) ## percentile interval
```

```
  2.5%  97.5%
0.1441 1.2840
```

Other Link Functions

So far, we've been focusing on the **cannonical (inverse) link function**:

$$G^{-1}(\mathbf{x}^T\boldsymbol{\beta}) = b'(\mathbf{x}^T\boldsymbol{\beta})$$

This had the advantage of being **natural from the model assumptions** and **motivated choice of loss functions**.

But we are not limited to $b'(\theta)$. Another link function that we could use is the "log link" such that:

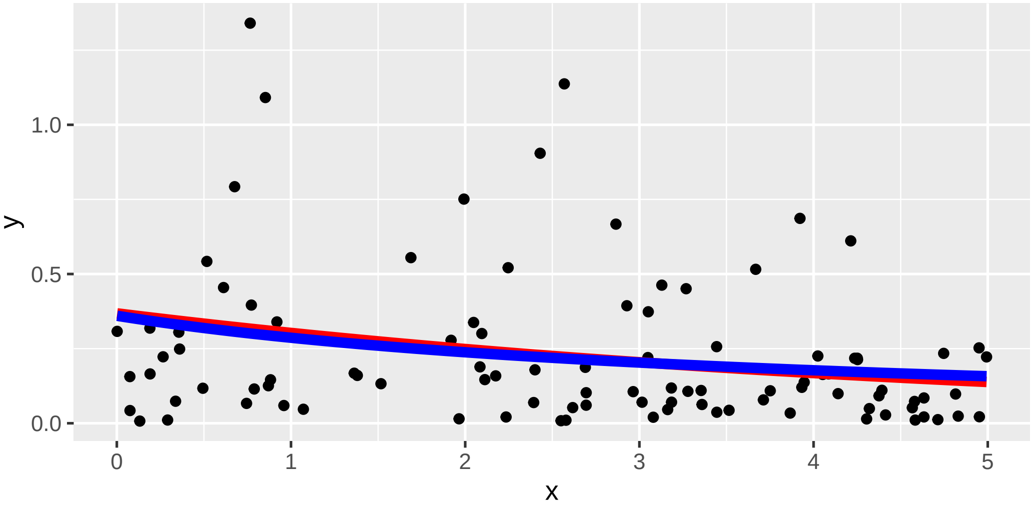
$$\log(\mu(\mathbf{x})) = \mathbf{x}^T\boldsymbol{\beta} \iff \mathrm{E}\left(y \mid \mathbf{x}\right) = \exp\left\{\mathbf{x}^T\boldsymbol{\beta}\right\} = G^{-1}(\mathbf{x}^T\boldsymbol{\beta})$$

```
> ## xy is the exponential data
> (modlog <- glm(y ~ x, data = xy, family = Gamma(link = "log")))

Call:  glm(formula = y ~ x, family = Gamma(link = "log"), data = xy)

Coefficients:
(Intercept)            x
     -0.999       -0.196

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:           120
Residual Deviance: 113        AIC: -95.3
```



y
1.0
0.5
0.0
0
1
2
3
4
5
x

Summary: GLMs

- Expand modeling of $\mathrm{E}\left(y \mid \mathbf{x}\right) = \mu(\eta(\mathbf{x}; \boldsymbol{\beta}))$
- Keep $\eta(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}^T\boldsymbol{\beta}$ (**linear predictors**)
- Let μ be **non-linear**.
- We connect the conditional mean with the linear predictors using a **(inverse) link function)**.

Exponential Family Distributions have nice connections between the functional form of the distribution and the link functions:

$$f(y, \theta, \psi) = \exp\left(\frac{y\theta - b(\theta)}{a(\psi)} + c(y, \psi)\right)$$
$$\mathrm{E}(Y) = b'(\theta)$$
$$\mathrm{Var}(Y) = b''(\theta)a(\psi) = V(\theta)$$

Convenient connection to GLMs.

R implementations:

`binomial`, `gaussian`, `Gamma`, `inverse.gaussian`, `poisson`,
`quasi`, `quasibinomial`, `quasipoisson`

`quasi` allow modifying variance for those distributions