STATS 212: Generalized Linear Models
Lecture 2: Beyond Ordinary Linear Regression Models

Babak Shahbaba

UCI, Spring 2010

What could go wrong with linear regression models

- In practice, one or more assumptions of linear regression models might be violated.
- This could result in wrong inference.
- Here, we discuss these assumptions can be violated and mention some possible fixes.

Linearity

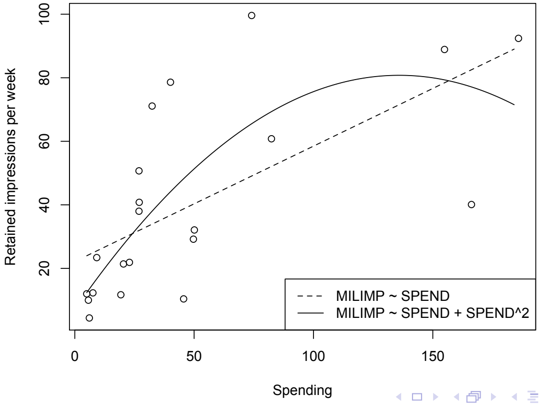
- Using linear models, we implicitly assume that the relationship between x and y is linear (note that this in general is different from the linearity assumption of the function in terms of parameters; i.e., $g(x) = x\beta$).
- If the assumption of linear relationship (with respect to variables) does not hold, we might still be able to use linear regression models after some transformation of original variables.
- Typical transformations are (we could use a combination of these with the original variables)
  - $\log(x)$: For variables with positive values and heavily right-skewed distribution.
  - $\sqrt{x}$:This transformation has milder effect compared to log transformation, and it is usually recommended for counts.
  - $x^2, x^3, \ldots$: To create nonlinear relationships in the form of polynomial function.

TV Ad Yields

- This data appeared in the Wall Street Journal and can be obtained from the “data and story” website.
- The advertisement were selected by an annual survey conducted by Video Board Tests, Inc., a New York ad-testing company.
- The data are based on interviews with 20,000 adults who were asked to name the most outstanding TV commercial they had seen, noticed, and liked.
- The retained impressions were based on a survey of 4,000 adults, in which regular product users were asked to cite a commercial they had seen for that product category in the past week.
- The objective was to investigate the relationship between TV advertising budget (SPEND) and millions retained impressions per week (MILIMP)

TV Ad Yields

- The following graphs shows the results of two regression models:
  1. $E(MILIMP) = \beta_0 + \beta_1 SPEND$
  2. $E(MILIMP) = \beta_0 + \beta_1 SPEND + \beta_2 SPEND^2$



Additivity

- In linear regression models, the effects of explanatory variables on the response variable are assumed to be additive. This means, the expected value of the response variable changes by a fixed amount when one of the explanatory variables is varied, regardless of the values of the other variables

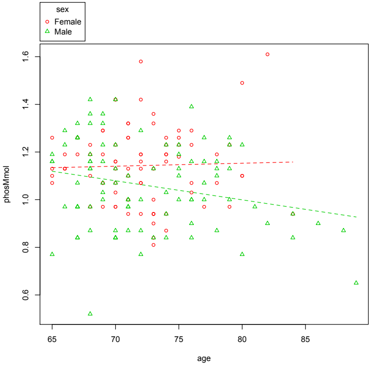
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- If the theory suggests that the effects are not additive, i.e., the effect of one variable on response depends on the value of the other variable, we can still use linear regression models with some minor adjustment.
- For example, we could add interactions terms into our model.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Interaction

- The following graph shows the scatter plot of inorganic phosphorus levels vs. age for sample elderly patients.
- As we see, the relationship seems to be different between men and women.



Additivity

- In some situations, we can use appropriate transformations to create additivity. For example, consider the following model with multiplicative effects:

$$\hat{y} = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$$

- We can use the log-transformation to create an additive model

$$\log(\hat{y}) = \log(\beta_0) + \beta_1 \log(x_1) + \beta_2 \log(x_2)$$

  where the effects (on log scale) are additive.

- When we can use simple transformations to make non-linear/non-additive models linear/additive, we say the models are *intrinsically linear/additive*.

Independence and constant variance assumptions for errors

- In linear regression models, the error terms are assumed to be independent.
- In this case, the covariance matrix of error terms is not diagonal ($\sigma^2 I$) anymore, we need to use a full covariance matrix (Σ).
- Moreover, they are assumed to have equal variance. When this is not the case, we could use *weighted* least squares, where the weight of each data point is inversely proportional to its variance.
- The assumption of normality for errors is not as important as the above two, but it should still be tested.

Unbounded response variable

- In linear regression analysis, we model the expected value of the response variable as a function of explanatory variables, $E(y|x) = x\beta$.
- The right had side of this equation is unbounded in general. This could cause a problem, if the left hand side, $E(y|x)$, is bounded.
- For example, if the response variable is binary, $y \in \{0, 1\}$, its expectation is between 0 and 1.
- For count variables, the expectation would be a non-negative number.

Generalized linear model

- To deal with some of these issues, we need a more flexible family of models.
- The class of generalized linear models (GLM), that includes linear models as a special case, provides such flexibility while it is still easy to use.
- Generalized linear models have three components:
    - A random component
    - A systematic component
    - A link function

Generalized linear model

- The random component identifies the response variable and its probability distribution.
- In most situations, we assume parametric model $P(y|\theta)$ for the distribution of y from the exponential family.
- Recall that the exponential family includes most of the well-known distributions such as normal, binomial, multinomial and Poisson.
- In general, if the outcome variable is continuous and real-valued, we use the normal distribution.
- If the outcome is binary, we use the binomial distribution. For outcome variables with multiple categories, we use the multinomial instead.
- If the outcome variable represent counts data, we use the Poisson distribution.

Generalized linear model

- The systematic component specifies the set of predictors (i.e., explanatory variables) $x = (x_1, ..., x_p)$ used in a *linear predictor* function.
- As before, we also append a vector of ones at the beginning of x.
- In the matrix form, the linear predictor function $\eta = x\beta$, where $\beta = (\beta_0, \beta_1, ..., \beta_p)$.
- Alternatively,for each observation i, where $i = 1, ..., n$, the linear predictor function is $\eta_i = \beta_0 + \sum_j^p x_{ij}\beta_j$.
- Also, as before, some of predictors could be a transformation (e.g., x^2) of original predictors.

Generalized linear model

- The link function is a monotonic differentiable function that connects the random and systematic components.
- More specifically, if $\mu = E(y|x)$, the link function g connects μ to η such that $g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$ for each observation i.
- For the ordinary linear model we discussed before, the link function is identity: $g(\mu_i) = \mu_i$. That is $\mu_i = \eta_i = x_i\beta$.

Logistic regression model

- As mentioned before, for binary outcome variable, we use the binomial distribution.

$$y_i | n_i, \mu_i \sim \mathrm{binomial}(n_i, \mu_i)$$

  with the Bernoulli distribution as its special case when $n_i = 1$.
- As usual, we define the systematic part of the model $\eta_i = x_i \beta$ (where x_i is a row vector of all observed values for subject i, and β is a column vector of size $p + 1$).
- A common link function for this model is the *logit* function *logit* and defined as

$$g(\mu_i) = \log(\frac{\mu_i}{1 - \mu_i}) = x_i \beta$$

  where μ_i is the probability of success (i.e., $y_i = 1$).
- As the result

$$\mu_i \quad = \quad \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)}$$

Logistic regression model

- The likelihood is therefore defined in terms of β as follows:

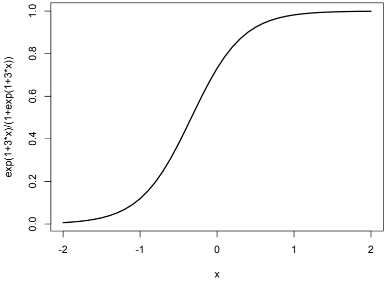
$$\begin{aligned} P(y|\mu) &\propto \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{n_i-y_i} \\ P(y|\beta) &\propto \prod_{i=1}^{n} \left(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)^{y_i}\left(\frac{1}{1+\exp(x_i\beta)}\right)^{n_i-y_i} \end{aligned}$$

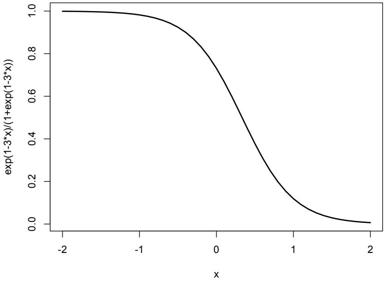
- Note that in this model the variance of $y|x$ depends on the mean and therefore will not be constant

$$var(y_i|x_i) = \mu_i(1-\mu_i)$$

Logistic regression model

- The following two graphs show examples of logistic function with negative and positive β

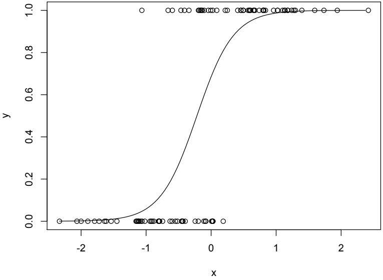




Logistic regression model

- Fitting logistic model to simulated data from

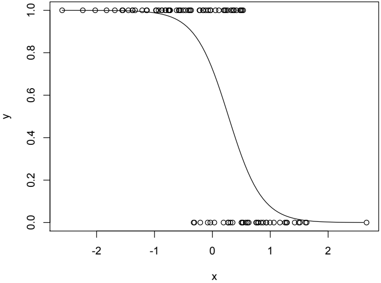
$$
\begin{aligned}
x_i &\sim N(0,1) \\
y_i &\sim \mathrm{Bernoulli}\left(\frac{\exp(1+3x)}{1+\exp(1+3x)}\right)
\end{aligned}
$$



Logistic regression model

- Fitting logistic model to simulated data from

$$
\begin{aligned}
x_i &\sim N(0,1) \\
y_i &\sim \mathrm{Bernoulli}\Big(\frac{\exp(1-3x)}{1+\exp(1-3x)}\Big)
\end{aligned}
$$



Multinomial logistic model

- This is a generalization of logistic regression when the outcome could have multiple values (i.e., could belong to one of K classes).

$$y_i|n_i, \mu_{i1}, ..., \mu_{iK} \sim \mathrm{multinomial}(n_i, \mu_{i1}, ..., \mu_{iK})$$

  where μ_{ik} is the probability of class k for observation i such that $\sum_{k=1}^{K} \mu_{ik} = 1$.

- y_i is also a vector of K elements with $\sum_{k=1}^{K} y_{ik} = n_i$.
- The systematic part is now a vector $\eta_{ik} = x_i\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a matrix of size $(p+1) \times K$.

Multinomial logistic model

- Each column k (where $k = 1, ..., K$) corresponds to a set of $p + 1$ parameters associated with class k.
- This representation is redundant and results in nonidentifiability, since one of the β_k's (where $k = 1, ..., J$) can be set to zero without changing the set of relationships expressible with the model.
- Usually, either the parameters for $k = 1$ (the first column) or for $k = K$ (the last column) would be set to zero.
- In Bayesian models, removing this redundancy would make it difficult to specify a prior that treats all classes symmetrically. Therefore, we do not remove redundancy (in general, nonidentifiability does not create problem for Bayesian models). In this case, what matters is the difference between the parameters of different classes.

Multinomial logistic model

- For the multinomial logistic model, we use a generalization of the link function we used for the binary logistic regression

$$\mu_{ik} = \frac{\exp(x_i \boldsymbol{\beta}_k)}{\sum_{k'=1}^{K} \exp(x_i \boldsymbol{\beta}_{k'})}$$

- The likelihood in terms of β is as follows:

$$\begin{aligned} p(y|\mu) &\propto \prod_{i=1}^{n} \prod_{k=2}^{K} \mu_{ik}^{y_{ik}} \\ P(y|x, \boldsymbol{\beta}) &\propto \prod_{i=1}^{n} \prod_{k=1}^{K} \Big(\frac{\exp(x \boldsymbol{\beta}_k)}{\sum_{k'=1}^{K} \exp(x \boldsymbol{\beta}_{k'})} \Big)^{y_{ik}} \end{aligned}$$

- Here $\boldsymbol{\beta}_k$ is a column vector of $p+1$ parameters corresponding to class k.

Multinomial logistic model

- $\boldsymbol{\beta}$ in general is a $(p+1) \times K$ matrix. The first row, $(\beta_{01}, ..., \beta_{0K})$ are intercepts, and $(\beta_{j1}, ..., \beta_{jK})$ in row $j+1$ are regression parameters associated with the j^{th} predictor.
- x_i is the row vector of predictors value for observation i (including the constant 1 at the beginning).
- y_{ik} is the number of cases in observation i that are in class k.

Poisson model

- When the outcome variable, y, represents counts, we use the Poisson model.

$$y_i | \mu_i \sim \mathrm{Poisson}(\mu_i)$$

- The systematic components are defined as before: $\eta_i = x_i \beta$.
- The usual link function for this model is the log link:

$$g(\mu_i) = \log(\mu_i) = \eta_i$$

- We therefore have

$$\mu_i = \exp(\eta_i) = \exp(x_i \beta)$$

Poisson model

- The likelihood in terms of β can obtained as follows:

$$
\begin{aligned}
p(y_i|\mu_i) &\propto \prod_i^n \exp(-\mu_i)\mu_i^{y_i} \\
p(y_i|\beta) &\propto \prod_i^n \exp[-\exp(x_i\beta)][\exp(x_i\beta)]^{y_i}
\end{aligned}
$$

- Similar to logistic and multinomial models, the variance of $y|x$ in Poisson model depends on the mean and therefore will not be constant

$$var(y_i|x_i) = \mu_i$$

Fitting GLMs in R

- In R, we use the function `glm` to fit generalized linear models.
- The function has the following format:

```
glm( formula, family = gaussian, data )
```

- `formula`- This specifies the systematic component, for example, we could have

  $y \sim x_1 + x_2$, or $y \sim .$

- `family`- This specifies the stochastic part of the model, i.e., probability of the response variable. The type of link function could be given within our specification of family, for example,

```
family = binomial(link="logit")
```

Fitting GLMs in R

- Some of the default links are

```
   binomial(link = "logit")
 gaussian(link = "identity")
   Gamma(link = "inverse")
    poisson(link = "log")
```

- For multinomial logit model, we can use the function `multinom` in `nnet` package.

Exponential family of distributions

- For the random component of models we discussed so far, we assumed distributions such as normal, binomial, and Poisson.
- These distributional forms are members of the exponential family.
- A single parameter distributional form belongs to the exponential family if the distribution has the following form

$$P(y_i|\theta) = \exp\{g(\theta)T_i(y_i) + c_i(\theta) + h_i(y_i)\}$$

  where P is the density function for continuous random variables and probability mass function for discrete variables.
- For example, for Poisson distributions, we have

$$\begin{aligned} P(y|\theta) &= e^{-\theta}\theta^{y}/y! \\ &= \exp\{\log(\theta)y - \theta - \log(y!)\} \end{aligned}$$

- Here, $g(\theta) = \log(\theta)$, $T(y) = y$, $c(\theta) = -\theta$, and $h(y) = -\log(y!)$

Sufficiency in exponential family

- For a vector of independent observations, $y = (y_1, y_2, ..., y_n)$, we have

$$P(y|\theta) = \exp\{g(\theta) \sum T_i(y_i) + \sum c_i(\theta) + \sum h_i(y_i)\}$$

- If the observations are identically distributed, we can drop the index i, and present the exponential family as

$$P(y|\theta) = \exp\{g(\theta) T(y) + c(\theta) + h(y)\}$$

- Based on the factorization theorem, $T(y)$ is sufficient statistic for θ.
- Recall that sufficiency in classical sense (and assuming that the distribution is parametric) means that $P(y|T, \theta)$ does not depend on θ.
- Sufficiency in Bayesian sense means that for any prior $P(\theta)$, there exist versions of posterior $P(\theta|y)$ and $P(\theta|T)$ such that $P(\theta|y) = P(\theta|T)$.

Multiparameter exponential family

- In general, a distribution in exponential family can have multiple parameters.
- In this case, g, T, and c would be vectors, and instead of $g(\theta)T(y)$, we have their dot product, $g^T(\theta)T(y)$.

$$P(y|\theta) = \exp\{\sum_{k=1}^{K} g_k(\theta)T_k(y) + c(\theta) + h(y)\}$$

  where $T = (T_1, T_2, ..., T_k)$ is sufficient for θ.
- Note that while the dimension of T, which is K, is usually the same as the dimension of θ, this does not have to be the case in general.
- Also note that g and T are not unique. We can for example use their linear transformation such that $g^* = A^T g$ and $T^* = A^{-1}T$.
- The dot product of $g^*(\theta)$ and $T^*(y)$ is the same as $g^T(\theta)T(\theta)$.

Multiparameter exponential family

- Let's consider a normal distribution with unknown mean and unknown variance.

$$
\begin{aligned}
P(y|\mu,\sigma^2) &= \exp\{\frac{-(y-\mu)^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\} \\
&= \exp\{-\frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\}
\end{aligned}
$$

- Here,

$$
\begin{aligned}
g(\mu,\sigma^2) &= (\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}) \\
T(y) &= (-\frac{y^2}{2}, y) \\
c(\mu,\sigma^2) &= -\frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2} \\
h(y) &= 0
\end{aligned}
$$

Degenerate forms of exponential family

- Sometimes, the dimensionality of g and T can be reduced.
- For example, consider the Bernoulli distribution

$$\begin{aligned} P(y|\theta) &= \theta^y(1-\theta)^{1-y} \\ &= \exp\{\log(\theta)y + \log(1-\theta)(1-y)\} \end{aligned}$$

- It seems that $g(\theta)$ and $T(y)$ are vectors

$$\begin{aligned} g(\theta) &= (\log(\theta), \log(1-\theta)) \\ T(y) &= (y, 1-y) \\ c(\theta) &= 0 \\ h(y) &= 0 \end{aligned}$$

- However, there is only one parameter in this model.

Degenerate forms of exponential family

- We can modify the form of the distribution as follows:

$$P(y|\theta) \;=\; \exp\{\log(\frac{\theta}{1-\theta})y + \log(1-\theta)\}$$

- This way, $g(\theta) = \log(\frac{\theta}{1-\theta}),\ T(y) = y$, and $c(\theta) = \log(1-\theta)$.
- A member of exponential family is called degenerate if
  a. Some linear combination of $T_k(y)$ is constant. For example, in the case of Bernoulli distribution, $(y) + (1-y) = 1$.
  b. Some linear combination of $g_k(\theta)$ is constant.
- As we saw in the case of Bernoulli distribution, we can modify the form of the distribution so it is not degenerate any more.
- In this case, $\phi = g(\theta) = \log(\frac{\theta}{1-\theta})$ is called *natural parameter*.

Natural parameter

- In general, the form of exponential family does not change if we perform one-to-one transformations of variable or parameter.
- More specifically, notice that

$$\int_{\chi} \exp[h(y)] \exp\{\sum_{k=1}^{K} g_k(\theta) T_k(y)\} = 1/\exp[c(\theta)]$$

- Therefore, to define the distribution, all we need to specify is $g(\theta)$ such that $1/\exp[c(\theta)]$ is finite.
- $g(\theta)$, which in general is a vector of size K, is called the natural parameter.
- We can change the parameter using the transformation $\phi_k = g_k(\theta)$

$$P(y|\theta) = \exp\{\sum_{k=1}^{K} \phi_k T_k(y) + c^*(\phi) + h(y)\}$$

Distribution of natural sufficient statistic

- As we can see, we can change the random variable from y to $t = T(y)$.
- The sufficient statistic T also has an exponential family distribution, and the natural parameter of its distribution is the same as that y.
- That is,

$$\begin{aligned} P(y|\theta) &= \exp\{\sum_{k=1}^{K} g_k(\theta) T_k(y) + c(\theta) + h(y)\} \\ P(t|\theta) &= \exp\{\sum_{k=1}^{K} g_k(\theta) t_k + c(\theta) + h^*(t)\} \end{aligned}$$

Score function and information

- Recall that the first derivative of log-likelihood function, $L(\theta)$, is called the *score function*

$$u(\theta) \quad = \quad \frac{\partial L(\theta)}{\partial \theta}$$

- For single parameter exponential family, the score function is

$$u(\theta) \quad = \quad T(y)\frac{\partial g(\theta)}{\partial \theta} + \frac{\partial c(\theta)}{\partial \theta}$$

- The value of score function for a given θ_0 (e.g., $H_0 : \theta = \theta_0$) is called the *efficient score*, $u(\theta_0)$.
- In terms of natural parameter ϕ,

$$u(\phi) \quad = \quad T(y) + \frac{\partial c^*(\phi)}{\partial \phi}$$

Score function and information

- Under some regularity conditions (mainly to make it possible to interchange integration and differentiation), for a given value of θ we have

$$E_\theta[u(\theta)] \quad = \quad 0$$

- As the result

$$var_\theta[u(\theta)] \quad = \quad E[u^2(\theta)] = i(\theta)$$

- $i(\theta)$ is called *Fisher information* about θ given y.
- Under the regularity conditions assumed above,

$$i(\theta) \quad = \quad E[u^2(\theta)] = E[-\frac{\partial^2 L(\theta)}{\partial\theta^2}]$$

Moments of natural sufficient statistics

- For exponential family,

$$E_\theta[T(y)] \quad = \quad -\frac{\partial c(\theta)}{\partial \theta} \Big/ \frac{\partial g(\theta)}{\partial \theta}$$

- In terms of the natural parameter $\phi = g(\theta)$

$$E_\phi[T(y)] \quad = \quad -\frac{\partial c^*(\phi)}{\partial \phi}$$

- For the second moment, note that

$$var_\phi[T(y)] \quad = \quad var[u(\phi)] = i(\phi)$$

- Therefore, we need to find the Fisher information about the natural parameter $\phi = g(\theta)$ given y.

Moments of natural sufficient statistics

- To find the Fisher information we have

$$-\frac{\partial^2 L(\theta)}{\partial\theta^2} \quad = \quad -T(y)\frac{\partial^2 g(\theta)}{\partial\theta^2} - \frac{\partial^2 c(\theta)}{\partial\theta^2}$$

- Therefore, Fisher information about θ is

$$\begin{aligned} i(\theta) \quad &= \quad -E[T(y)]\frac{\partial^2 g(\theta)}{\partial\theta^2} - \frac{\partial^2 c(\theta)}{\partial\theta^2} \\ &= \quad \frac{\partial^2 g(\theta)}{\partial\theta^2}\frac{\partial c(\theta)}{\partial\theta}\Big/\frac{\partial g(\theta)}{\partial\theta} - \frac{\partial^2 c(\theta)}{\partial\theta^2} \end{aligned}$$

- With respect to the natural parameter $\phi = g(\theta)$

$$i(\phi) \quad = \quad -\frac{\partial^2 c^*(\phi)}{\partial\phi^2}$$

- Therefore,

$$var_\phi[T(y)] \quad = \quad -\frac{\partial^2 c^*(\phi)}{\partial\phi^2}$$

Results for multiparameter exponential family

- The above results easily generalize to K-dimensional multiparameter exponential family distributions with parameters $\theta = (\theta_1, ..., \theta_K)$
- The score function in this case is a vector of size K

$$u_k(\theta) = \frac{\partial L(\theta)}{\partial \theta_k}, \qquad k = 1, ..., K$$

- As before,

$$E_\theta[u(\theta)] = \mathbf{0}$$

- Fisher information is $K \times K$ matrix whose $(j,k)^{th}$ element is

$$i_{jk}(\theta) = E[u_j u_k] = E[-\frac{\partial^2 L(\theta)}{\partial \theta_j \partial \theta_k}]$$

Results for multiparameter exponential family

- For the multiparameter case, we have

$$
\begin{aligned}
E_\phi[T_k(y)] &= -\frac{\partial c^*(\phi)}{\partial \phi_k} \\
cov_\phi[T_j T_k] &= -\frac{\partial^2 c^*(\phi)}{\partial \phi_j \phi_k}
\end{aligned}
$$

- Note that ϕ and T are vectors.

Maximum likelihood estimate

- **Assignment**: For single parameter exponential family, find the maximum likelihood estimate (MLE) and show that MLE is a method of moments estimator. (Similar results hold for multi-parameter case, but you do not need to show it here.) Show this is in fact true for Poisson models.

Hypothesis testing

- For a single parameter exponential family, the likelihood ratio test for $H_0 : \theta = \theta_0$ vs. $H_A : \theta = \theta_1$ is as follows

$$LR = \exp\{[g(\theta_1) - g(\theta_0)]T(y) + c(\theta_1) - c(\theta_0)\}$$

- Note that the test involve the data only through the sufficient statistic $T(y)$.
- If we assume (without loss of generality) that $\theta_0 < \theta_1$ and g is strictly increasing in θ, the likelihood ration test is an increasing function of $T(y)$, and we can reject the null hypothesis for large values of $t = T(Y)$.
- More specifically, we reject the null if $T(y) \geq k$, where $P[T(y)|\theta = \theta_0] = \alpha$.