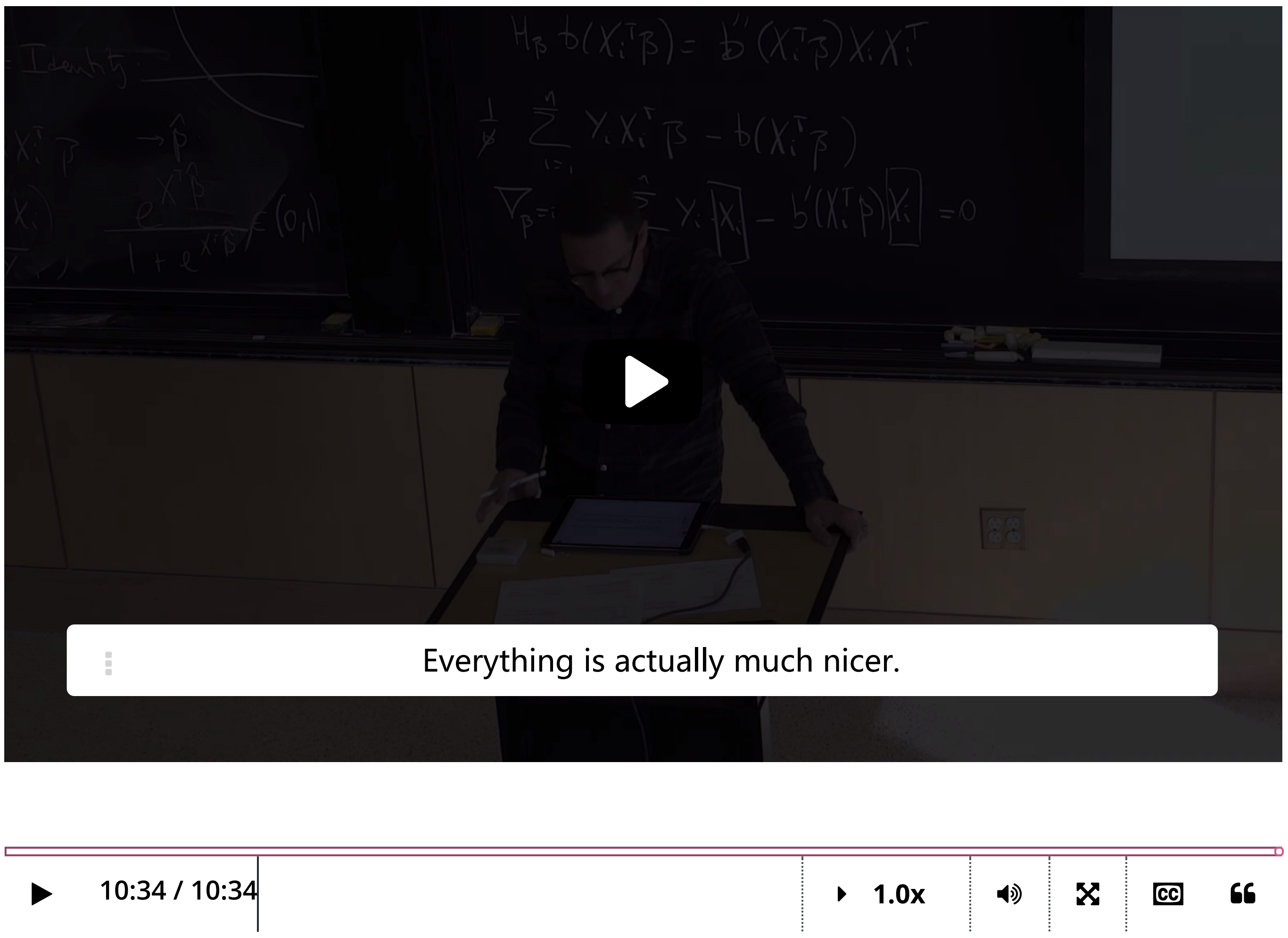
7. Concluding Remarks on MLE for GLMs

Optimization for General Link Functions and Prediction using Estimated Beta



it's much easier if you actually compute your Bayesian

estimator.

And it looks like it's actually xn bar plus something that

goes to zero, then, yes, you actually know a bunch of things

about this estimator.

If you know it's xn bar plus 1 over n,

then you know it's asymptotically normal,

and you can apply the central limit theorem.

Everything is actually much nicer.

End of transcript. Skip to the start.

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Gradient Descent

Let $\ell_n(\boldsymbol{\beta})$ be the likelihood function as a function of $\boldsymbol{\beta}$ for a given $\mathbb{X}, \mathbf{Y}$. Recall from Lecture 9 the gradient of a real-valued function $f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$.

We can use **gradient descent** to find a local minimum of the negative of the log-likelihood function. The gradient descent optimization algorithm, in general, is used to find the local minimum of a given function $f(\mathbf{x})$ around a starting initial point $\mathbf{x}_0$.

Let $\ell_{n,1}(\boldsymbol{\beta}) = -\ell_n(\boldsymbol{\beta})$.

Given a starting point $\boldsymbol{\beta}$, **repeat**

1. $\Delta\boldsymbol{\beta} = -\nabla\ell_{n,1}(\boldsymbol{\beta})$.

2. *Choose* step size t.

3. *Update* $\boldsymbol{\beta} = \boldsymbol{\beta} + t\Delta\boldsymbol{\beta}$.

until a stopping criterion is satisfied.

The **stopping criterion** for gradient descent is usually of the form $\|\nabla\ell_n(\boldsymbol{\beta})\| \leq \epsilon$ for some very small ϵ.

The analysis of gradient descent and the choice of step size t in every iteration is beyond the scope of this class, but the implementation of this algorithm requires one to compute gradients of the function $\ell_n(\boldsymbol{\beta})$ at various points as given in Step 1 of the algorithm. Hence, the computational complexity of gradient descent boils down to the complexity of evaluating the gradient of the function $\ell_n(\boldsymbol{\beta})$.

Note: The above algorithm is a **descent** algorithm to **minimize** and find a local minimum of a given function. This is the reason why we used the conversion $\ell_{n,1}(\boldsymbol{\beta}) = -\ell_n(\boldsymbol{\beta})$. If one were to re-write the algorithm without this conversion, we would have maximized $\ell_n(\boldsymbol{\beta})$ and Step 1 of the algorithm would be $\Delta\boldsymbol{\beta} = \nabla\ell_n(\boldsymbol{\beta})$. Such an algorithm is called a **gradient ascent** algorithm. It is more common in literature in optimization to use the descent version rather than the ascent version.

One Step of Gradient Ascent for the Poisson GLM

2/2 points (graded)

Let $\boldsymbol{\beta} \in \mathbb{R}^1$ and let $\ell_2(\boldsymbol{\beta}) = \sum_{i=1}^{2} \frac{Y_i X_i^T \boldsymbol{\beta} - e^{X_i^T \boldsymbol{\beta}}}{\phi} + c$, for some constant c. For the Poisson GLM, recall that $\phi = 1$ and we have $b(\theta) = e^{\theta}$.

What is $\nabla\ell_2(\beta)$ for any β?

Use **X_i** for X_i and **Y_i** for Y_i.

$\sum_{i=1}^{2}$ X_i*Y_i - X_i*exp(X_i*bet ✓ **Answer:** Y_i*X_i-X_i*exp(X_i*beta)

$X_i \cdot Y_i - X_i \cdot \exp(X_i \cdot \beta)$

STANDARD NOTATION

Let

1. $X_1^T = X_1 = 0.1$, $Y_1 = 1$,
2. $X_2^T = X_2 = 0.2$, $Y_2 = 2$,
3. $\boldsymbol{\beta} = 0$,
4. Step size $t = 0.01$.

What is the new $\beta = \beta + t \cdot \nabla\ell_2(\beta)$?

0.002 ✓ **Answer:** 0.002

Solution:

The gradient $\nabla\ell_2(\beta)$ is given as

$$\sum_{i=1}^{2} Y_i X_i - X_i e^{X_i \beta}.$$

The value of β after a step is

$$\beta = 0 + 0.01 \times (0.1 - 0.1 + 0.4 - 0.2) = 0.002.$$

Submit You have used 1 of 3 attempts

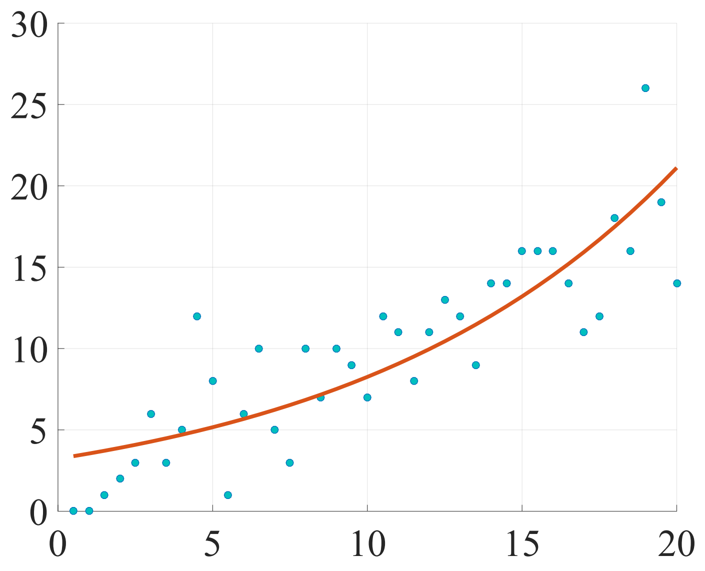
Answers are displayed within the problem

Prediction

1/1 point (graded)

The following figure shows

- scatter plot of (x_i, y_i),
- a generalized linear model assuming Y is Poisson and using the canonical link function for the Poisson exponential family.



The estimated vector $\widehat{\boldsymbol{\beta}} = [1.1723\ \ 0.0939]$, where we assume $g\left(\mu\left(x\right)\right) = \beta_1 + x\beta_2$.

What is the predicted value $\hat{\mu}\left(x\right)$ for a new $x = 22$? Provide an answer with at least 3 decimals.

25.4852537 ✓ **Answer:** 25.4853

Solution:

The canonical link function for the Poisson exponential family is the log link $g\left(\mu\left(x\right)\right) = \ln\left(\mu\left(x\right)\right)$. Given that $\widehat{\boldsymbol{\beta}} = [1.1723\ \ 0.0939]$,

$$\hat{\mu}\left(22\right) = e^{(1.1723+22\times 0.0939)} = 25.4853.$$

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Why Choose the Canonical Link?

1/1 point (graded)
Each choice is an optimistic statement about the canonical link function in the context of GLMs and parameter estimation. Which of the following statements are **correct** ? Choose all that apply.

☑ Given an exponential family, there is always a parametrization that gives a canonical link that is increasing and invertible. ✓

☑ The log-likelihood function is concave if $\phi > 0$. ✓

☑ If the family is Gaussian, then the MLE for $\boldsymbol{\beta}$ is the LSE for linear regression, $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbf{Y}$. ✓

□ There is always a nice formula for the maximum likelihood estimator of $\boldsymbol{\beta}$.

✓

Solution:

The only false statement here is the last one: '**There is always a nice formula for the maximum likelihood estimator of $\boldsymbol{\beta}$.** Of the examples we have seen so far, the Gaussian is the only nice one (as demonstrated by the third option). In general, one needs to use optimization algorithms on the computer to numerically compute the MLE.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Asymptotic Normality

Let $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$ be iid with a distribution from an exponential family such that

$$g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta},$$

where $\mu_i = \mathbb{E}[Y_i | \mathbf{X}_i]$, $\mu_i = b'(\theta_i)$, and $g(\cdot)$ is a link function. θ_i is the canonical parameter of the exponential family for each i. With the transformation that $\theta_i = h(\mathbf{X}_i^T \boldsymbol{\beta}^*)$, let $\boldsymbol{\beta}^*$ denote the true underlying parameter of the observed iid samples $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$.

Let $\widehat{\boldsymbol{\beta}}_{n,\mathrm{MLE}}$ denote the maximum likelihood estimator of $\boldsymbol{\beta}^*$.

Then, $\widehat{\boldsymbol{\beta}}_{n,\mathrm{MLE}}$ is asymptotically normal if the statistical model and the parameter space of $\boldsymbol{\beta}$ satsify conditions required for asymptotic normality of the ML estimator as stated previously in the lecture on maximum likelihood estimation.

Discussion

Show Discussion

Topic: Unit 7 Generalized Linear Models:Lecture 22: GLM: Link Functions and the Canonical Link Function / 7. Concluding Remarks on MLE for GLMs