Generalized linear models

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS 424/524, SML 302

Lecture 10

Generalized linear models

In previous lectures, we learned about *linear regression* for prediction and *logistic regression* for classification.

- Linear regression and logistic regression are both *linear models*.
- The coefficient β enters the distribution of y_i through a linear combination with x_i.
- The difference is in the conditional distribution of the response.
  - in linear regression the response is real valued and distributed as a Gaussian;
  - in logistic regression the response is binary and distributed as a Bernoulli.

Generalized linear models

Linear and logistic regression are instances of *generalized linear models* (GLMs).

GLMs can handle many types of response:

- real-valued
- binary
- categorical
- positive real valued
- positive integer
- ordinal

How will we generalize the response for linear models?

To do this, our approach will be:

- rewrite conditional distribution, $p(y \mid x, \beta)$, using an appropriate distribution for response variable
- choose an appropriate function to project $\beta^T x \in \Re$ onto the appropriate parameter space

Example: logistic regression

- conditional distribution, $p(y \mid x, \beta)$, is Bernoulli distribution
- the logistic function projects $\beta^T x \in \Re$ onto $(0, 1)$, the Bernoulli bias.

Generalized linear models (GLMs)

- Formalize GLMs by describing a general probability density function for a family of distributions;
- then, define the conditional model $p(y|x, \beta)$ and derive parameter updates with respect to general form of PDF;
- this general PDF suggests a natural function to project our term $\beta^T x$ onto the appropriate space for our chosen distribution.

Do GLMs exist for arbitrary response distributions?

The exponential family

GLMs naturally capture responses from *exponential family* distributions.

The *exponential family* is a family of distributions that includes many useful distributions.

A probability density in the exponential family can be written as

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\},$$

where

- η is the *natural parameter*
- $t(x)$ are *sufficient statistics*
- $h(x)$ is the *underlying measure*, ensures x is in the right space
- $a(\eta)$ is the *log normalizer*

Exponential family: examples

Exponential family members

- Gaussian distribution
- gamma distribution
- Bernoulli distribution
- Poisson distribution
- multinomial distribution
- beta distribution
- Dirichlet distribution

Distributions not in the exponential family

- student-t distribution
- mixtures of distributions

Log normalizer $a(\eta)$

Exponential family form

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\},$$

The log normalizer ensures that the density integrates to 1,

$$a(\eta) = \log \int h(x) \exp\{\eta^T t(x)\} dx$$

This is the negative logarithm of the normalizing constant.

Note that $a(\eta)$ is not a function of x; x is marginalized away

Sufficient statistic $t(x)$

Exponential family form

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\},$$

The sufficient statistic $t(x)$ completely summarizes the data (in the sense of no information loss) for *the chosen distribution*.

When data $\mathcal{D} = \{x_1, \ldots, x_n\}$ are independent, identically distributed, then:

$$t(x) = \sum_{i=1}^{n} t(x_i)$$

Parameter estimates (e.g., MLE, MAP) require $t(x)$; x may be ignored.

Example: Bernoulli

Exponential family form

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\},$$

Bernoulli distribution in exponential family form

Standard Bernoulli distribution:

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x} \quad x \in \{0, 1\}$$

In exponential family form:

$$\begin{aligned} p(x \mid \pi) &= \exp\{\log \pi^x (1-\pi)^{1-x}\} \\ &= \exp\{x \log \pi + (1-x) \log(1-\pi)\} \\ &= \exp\{x \log \pi - x \log(1-\pi) + \log(1-\pi)\} \\ &= \exp\{x \log(\pi/(1-\pi)) + \log(1-\pi)\} \end{aligned}$$

Example: Bernoulli distribution

Exponential family form

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\},$$

Bernoulli distribution in exponential family form

$$p(x \mid \pi) = \exp\{x \log(\pi/(1-\pi)) + \log(1-\pi)\}$$

This reveals the exponential family where:

$$\begin{aligned} \eta &= \log(\pi/(1-\pi)) \\ t(x) &= x \\ a(\eta) &= -\log(1-\pi) = \log(1+e^{\eta}) \\ h(x) &= 1 \end{aligned}$$

Example: Bernoulli distribution

Bernoulli distribution in exponential family form

The relationship between the mean parameter π and the natural parameter η is invertible:

$$\begin{aligned} \eta &= \log\left(\frac{\pi}{1-\pi}\right) \\ \pi &= \frac{1}{1+e^{-\eta}} \end{aligned}$$

This is the log odds ratio and the *logistic function*.

Thus, we can map every natural parameter onto the mean parameter space, and vice versa.

Moments of an exponential family

Exponential family form

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\}$$

Hidden in this equation is a way to find the mean and variance of this family of distributions:

$$\begin{aligned}
\nabla_\eta a(\eta) &= \nabla_\eta \{\log \int \exp\{\eta^T t(x)\} h(x) dx\} \\
&= \frac{\nabla_\eta \int \exp\{\eta^T t(x)\} h(x) dx}{\int \exp\{\eta^T t(x)\} h(x) dx} \\
&= \int t(x) \frac{\exp\{\eta^T t(x)\} h(x)}{\int \exp\{\eta^T t(x)\} h(x) dx} dx \\
&= \int t(x) p(x|\eta) dx \\
&= \mathrm{E}_\eta[t(X)]
\end{aligned}$$

Exercise: Higher order derivatives give higher order moments.

Mean parameter

Expectation implies that the *mean parameter* $\mu = \mathrm{E}[t(X)]$ and natural parameter η have a 1-1 relationship.

Mean parameter and natural parameter: Logistic regression

Mean parameter $\pi = \mathrm{E}[X]$ because X is Bernoulli, and $t(x) = x$ in a Bernoulli.

The logistic function and log odds ratio map between the natural and mean parameter.

Mean parameter

To study the $1-1$ relationship between $\mu = \mathrm{E}[t(X)]$ and η, note that

- $\mathrm{Var}(t(X)) = \nabla_\eta^2 a(\eta)$ is positive
- this means that $a(\eta)$ is convex
- For convex functions, there is a 1-1 relationship between argument η and first derivative $\mu = \nabla_\eta a(\eta)$

Example: Poisson distribution

Exponential family form

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\}$$

Poisson distribution in exponential family form (mean parameter is λ)

$$\begin{aligned} p(x \mid \lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \quad x \in 0 \cup \mathcal{Z}^+ \\ &= \exp\left\{\log \frac{\lambda^x e^{-\lambda}}{x!}\right\} \\ &= \exp\{x \log \lambda - \lambda - \log x!\} \\ &= \frac{1}{x!} \exp\{x \log \lambda - \lambda\} \end{aligned}$$

Example: Poisson distribution

Exponential family form

$$p(x \mid \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\},$$

Poisson distribution in exponential family form

$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}$$

This reveals the exponential family form, where:

$$\begin{aligned} \eta &= \log(\lambda) \\ t(x) &= x \\ a(\eta) &= \lambda = e^{\eta} \\ h(x) &= \frac{1}{x!} \end{aligned}$$

Example: Poisson distribution

Poisson distribution in exponential family form

Relationship between mean parameter μ (here, λ) and natural parameter η is invertible:

$$\begin{aligned} \eta &= \log(\lambda) \\ \lambda &= \exp\{\eta\}. \end{aligned}$$

Example: Poisson distribution

Poisson distribution in exponential family form

We can find the moments of the Poisson distribution by taking the derivative of the log normalizer:

$$\begin{aligned} \nabla_\eta a(\eta) &= \nabla_\eta \exp\{\eta\} = \exp\{\eta\} = \lambda = \mathrm{E}[t(X)] \\ \nabla^2_\eta a(\eta) &= \nabla^2_\eta \exp\{\eta\} = \exp\{\eta\} = \lambda = Var[t(X)]. \end{aligned}$$

Example: Poisson distribution

Poisson distribution in exponential family form

Estimates of mean parameter λ fully contained in $t(x) = x$:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} t(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Summary of exponential family

The exponential family allows us to:

- work with a general form of a large family of distributions;
- come up with solutions that hold across these distributions;
- once we have our distribution of interest in the exponential family form, we can just "turn the mathematical crank" to get a solution.

Generalized linear models for prediction

Let's use the exponential family form to build the conditional distribution of a linear model.

First, define the conditional mean of y_i with respect to x_i, β:

$$\mathrm{E}[y_i \mid x_i, \beta] = f(\beta^T x_i) = \mu_i.$$

The generic GLM has the conditional probability density function:

$$\begin{aligned} p(y_i \mid x_i) &= h(y_i) \exp\{\eta_i^T t(y_i) - a(\eta_i)\} \\ \eta_i &= \beta^T x_i = \psi(\mu_i) \\ \mu_i &= f(\beta^T x_i) \end{aligned}$$

The input x_i enters the model through $\beta^T x_i$.

Generalized linear models for prediction

The generic GLM has the conditional probability density function:

$$\begin{aligned} p(y_i \mid x_i) &= h(y_i) \exp\{\eta_i^T y_i - a(\eta_i)\} \\ \eta_i &= \psi(\mu_i) \\ \mu_i &= f(\beta^T x_i) \end{aligned}$$

- Response y_i has conditional mean μ_i (prediction).
- Function $\mu_i = f(\beta^T x_i)$ is called the *response function*.
- The natural parameter is $\eta_i = \psi(\mu_i)$
- Function $\eta_i = \psi(\mu_i)$ is the *link function*

Generalized linear model: choices

GLMs let us build probabilistic predictors of many kinds of responses.

There are two choices to make in a GLM:

1. the distribution of the response y_i
2. the response function that gives us its mean $\mu_i = f(\beta^T x_i)$.

How can we make good choices for our data?

Generalized linear model: distribution choice

The distribution is usually determined by the form of y

Conditional distribution

- $y_i \in \Re$, then $y_i \mid x_i$ may be Gaussian
- $y_i \in \{0, 1\}$, then $y_i \mid x_i$ may be Bernoulli
- $y_i \in 0 \cup \mathcal{Z}^+$, then $y_i \mid x_i$ may be Poisson

Generalized linear model: response function choice

The response function is only constrained to give a value in the appropriate mean space for the distribution of y

response function: Bernoulli

For a Bernoulli, $\mu \in (0, 1)$; response function may take many forms

- logistic function
- probit function

How can we make a good choice of response function?

Canonical response function

There is a special response function called the *canonical response function*.

The canonical response function $f = \psi^{-1}$ is the inverse of the link function

With the canonical response function, natural parameter $\eta_i = \beta^T x_i$,

$$p(y_i \mid x_i) = h(y_i) \exp\{(\beta^T x_i) t(y_i) - a(\eta_i)\}.$$

Key point: link function is fixed by distribution; response function is not.

Example: Bernoulli linear model

Logistic regression

Recall the Bernoulli distribution in exponential family form:

$$p(x \mid \lambda) = \exp\{x \log(\pi/(1-\pi)) + \log(1-\pi)\}$$

If we choose the canonical response function, $\mu_i = \frac{1}{1+\exp^{-\eta_i}} = \frac{1}{1+\exp^{-\beta^T x_i}}$:

$$p(y_i \mid x_i) = \exp\left\{(\beta^T x_i) y_i - \log(1 + \exp(\beta^T x_i))\right\}$$

Then, once we fit parameter $\hat{\beta}$ to data, we can predict $y^* \mid x^*$:

$$\hat{y}^* = \mu^* = f(\eta^*) = \frac{1}{1 + \exp^{-\hat{\beta}^T x^*}}$$

What is the range of the predictions? Binary?

Example: Poisson linear model

Poisson linear model

Recall the Poisson distribution in exponential family form:

$$p(x \mid \pi) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}$$

If we choose the canonical response function, $\mu_i = \exp(\eta_i) = \exp(\beta^T x_i)$:

$$p(y_i \mid x_i) = \frac{1}{y_i!} \exp\{(\beta^T x_i) y_i - \exp(\beta^T x_i)\}$$

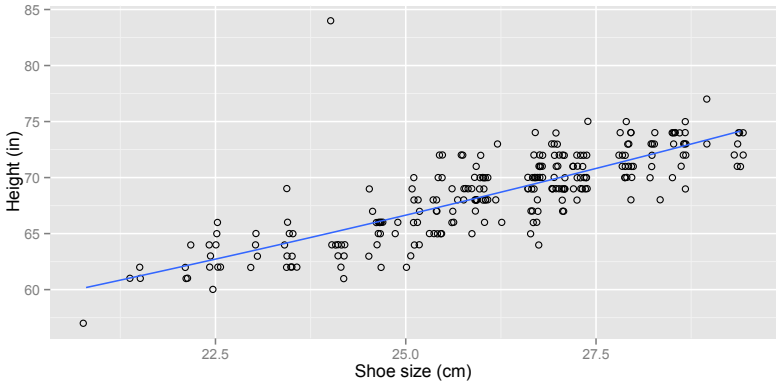
Then, once we fit parameter $\hat{\beta}$ to data, we can predict $y^* \mid x^*$:

$$\hat{y}^* = \mu^* = f(\eta^*) = \exp(\hat{\beta}^T x^*)$$

What is the range of the predictions? Non-negative integers?

Example: Poisson linear regression

Fit Poisson regression to data, where $y =$ height and $x =$ shoe size.

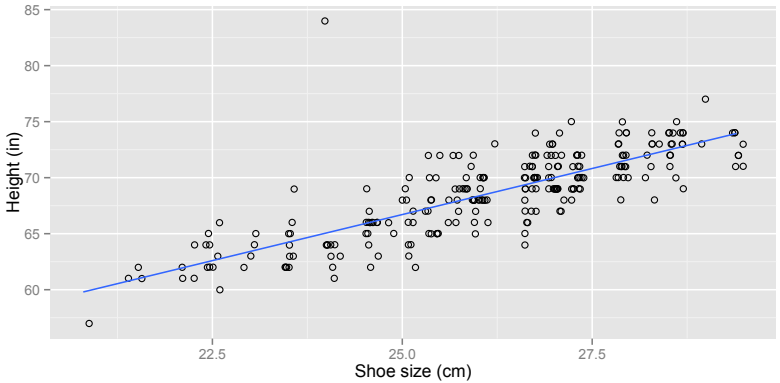


Here, $\beta_0 = 3.59$, $\exp\{\beta_0\} = 36.2$, and $\beta = 0.024$ (non-zero)

What is the function in blue here?

Example: linear regression

Fit linear regression to data, where $y =$ height and $x =$ shoe size.



Here, $\beta_0 = 25.6$ and $\beta = 1.65$ (non-zero)

Summary of generalized linear models

Generalized linear models allow us to create a linear model for a response with any exponential family distribution

As with linear and logistic regression, our GLM models can include:

- an intercept term: $\beta_0 + \beta x$;
- p predictors $\mathbf{x}^\mathbf{T}\beta$: *multivariate GLM*;
- sparse priors on the predictors;
- kernelized predictors $\phi(x_i)$.

Practically, if you have a response with a specific exponential family distribution, you can find a model that will allow you to fit these data.

Fitting the parameters of a GLM

We can fit GLMs with gradient descent methods.

As with linear and logistic regression, we will examine the conditional likelihood and its derivative.

The data are predictor and response pairs $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

The conditional log likelihood is

$$\mathcal{L}(\beta; \mathcal{D}) = \sum_{i=1}^{n} \log h(y_i) + \eta_i^T t(y_i) - a(\eta_i),$$

and recall that η_i is a function of β and x_i (via f and ψ)

Fitting a GLM: Gradient descent

The conditional log likelihood

$$\mathcal{L}(\beta; \mathcal{D}) = \sum_{i=1}^{n} h(y_i) + \eta_i^T t(y_i) - a(\eta_i),$$

Define log likelihood for each sample i as $\mathcal{L}_i$. The gradient with respect to β is

$$\begin{aligned}
\nabla_\beta \mathcal{L} &= \sum_{i=1}^{n} \nabla_{\eta_i} \mathcal{L}_i \nabla_\beta \eta_i \\
&= \sum_{i=1}^{n} (t(y_i) - \nabla_{\eta_i} a(\eta_i)) \nabla_\beta \eta_i \\
&= \sum_{i=1}^{n} (t(y_i) - \mathrm{E}[y_i \mid x_i, \beta])(\nabla_{\mu_i} \eta_i)(\nabla_{\beta^T x_i} \mu_i) x_i
\end{aligned}$$

Fitting a GLM: Gradient descent

The conditional log likelihood

$$\mathcal{L}(\beta; \mathcal{D}) = \sum_{i=1}^{n} \log h(y_i) + \eta_i^T t(y_i) - a(\eta_i),$$

The gradient is

$$\begin{aligned} \nabla_\beta \mathcal{L} &= \sum_{i=1}^{n} (t(y_i) - \mathrm{E}[y_i \mid x_i, \beta])(\nabla_{\mu_i} \eta_i)(\nabla_{\beta^T x_i} \mu_i) x_i \\ &= \sum_{i=1}^{n} (t(y_i) - \mathrm{E}[y_i \mid x_i, \beta]) \left(\frac{\partial \eta_i}{\partial \mu_i}\right) \left(\frac{\partial \mu_i}{\partial \beta^T x_i}\right) x_i \end{aligned}$$

In a canonical GLM, $\eta_i = \beta^T x_i$ and

$$\nabla_\beta \mathcal{L} = \sum_{i=1}^{n} (t(y_i) - \mathrm{E}[y_i \mid x_i, \beta]) x_i$$

MLE of β parameter

We can use this in a general approach to gradient descent for GLMs.

Let's choose a canonical response, and then $\eta_i = \beta^T x_i$

We have, for step size parameter ρ:

$$\beta^{(t+1)} = \beta^{(t)} + \rho \left(\sum_{i=1}^{n} (t(y_i) - \mathrm{E}[y_i \mid x_i, \beta^{(t)}]) x_i \right)$$

Note that the term $(t(y_i) - \mathrm{E}[y_i \mid x_i, \beta^{(t)}])$ is the residual of the GLM.

This is a *batch update*: we touch every sample at each iteration.

Stochastic gradient descent for GLMs

We can extend this to an online method using stochastic gradient descent.

Stochastic gradient descent with GLMs

Until convergence: at each iteration, pick a single sample i:

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + \rho(t(y_i) - \mathrm{E}[y_i \mid x_i, \beta^{(t)}])x_i.$$

Step size ρ should be chosen carefully and convergence monitored closely.

Example: SGD and logistic regression

Bernoulli in exponential family form with canonical response

$$\begin{aligned} t(y_i) &= y_i \\ \mathrm{E}[y_i \mid x_i, \beta] &= \frac{1}{1+e^{-\eta_i}} \\ &= \frac{1}{1+e^{-x_i^T\beta}}. \end{aligned}$$

We can write the stochastic gradient descent update

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + \rho\left(y_i - \frac{1}{1+e^{-x_i^T\beta^{(t)}}}\right)x_i.$$

This is exactly what we derived in the last lecture.

Example: SGD and Poisson regression

Poisson in exponential family form with canonical response

$$\begin{aligned} t(y_i) &= y_i \\ \mathrm{E}[y_i \mid x_i, \beta] &= \exp(\eta_i) \\ &= \exp(x_i^T \beta). \end{aligned}$$

We can write the stochastic gradient descent update

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + \rho \left(y_i - \exp(x_i^T \beta^{(t)}) \right) x_i.$$

We will estimate parameters this way because, as in logistic regression, the closed form solution is not available.

Generalized linear models: interpretations and assumptions

Interpretations:

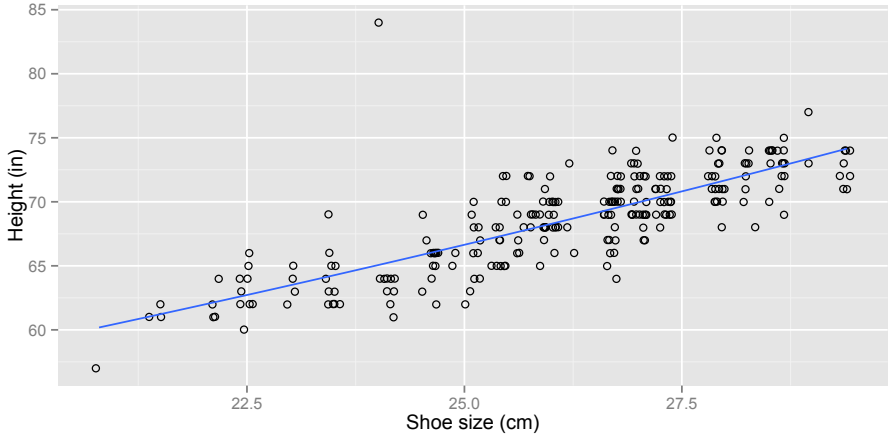
- coefficients that are zero remove corresponding predictor from model;
- larger (relative) magnitude coefficients have greater influence on response
- As with all models, GLMs may overfit data, making interpretations useless.

Assumptions:

- coefficients always enter model in a linear way
- predictors are additive in terms of their relationship on the response
- samples are IID
- response has a specific conditional distribution

Example revisited: Poisson linear regression

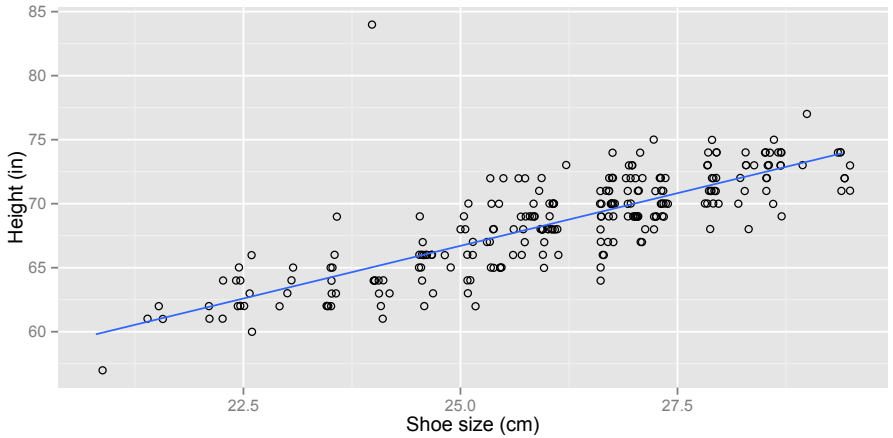
Fit Poisson regression to data, where $y =$ height and $x =$ shoe size.



Here, $\beta_0 = 3.59$, $\exp\{\beta_0\} = 36.2$, $\beta = 0.024$ (non-zero), and RSS: 1300.305

Example revisited: linear regression

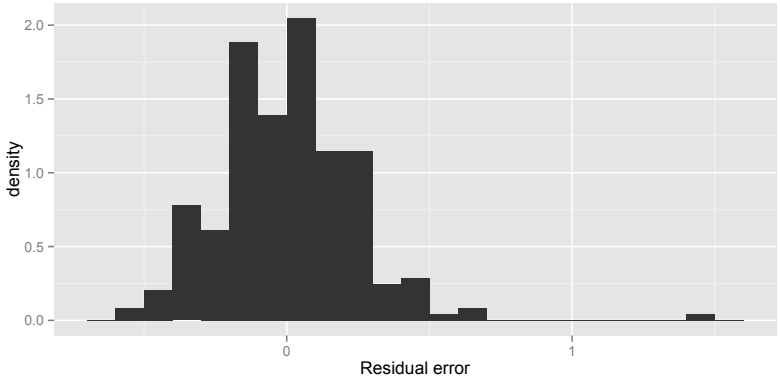
Fit linear regression to data, where $y =$ height and $x =$ shoe size.



Here, $\beta_0 = 25.6$, $\beta = 1.65$ (non-zero), and RSS: 1294.341

Example: Poisson regression

Using *glm* in R for Poisson GLM, we fit a model with y = height and x = all other covariates.



What distribution is this residual error?

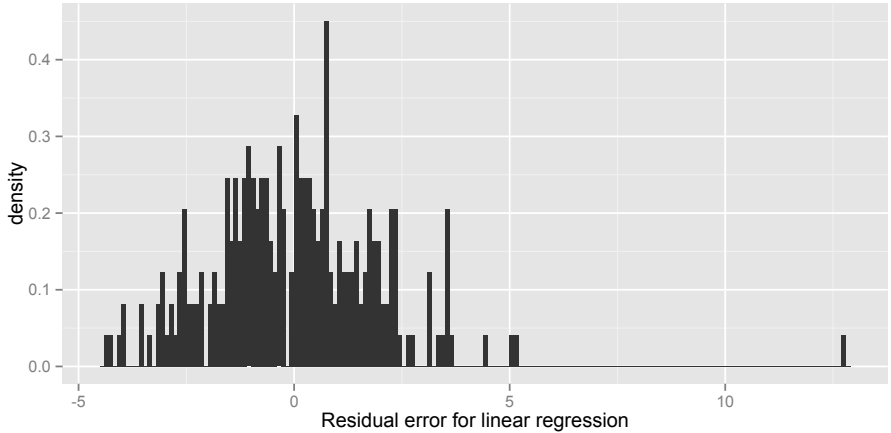
Example: Poisson regression to predict height

We can examine the coefficients of the unregularized Poisson regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8026	0.2301	16.53	0.0000
gender	-0.0318	0.0281	-1.13	0.2579
shoe_size	0.0177	0.0070	2.54	0.0112
month	-0.0010	0.0027	-0.37	0.7101
digit	0.0009	0.0035	0.26	0.7911
sleep	-0.0008	0.0072	-0.12	0.9080
siblings	0.0003	0.0076	0.04	0.9671
handed	-0.0042	0.0224	-0.19	0.8495
thumb	-0.0020	0.0172	-0.12	0.9059
registered	0.0005	0.0245	0.02	0.9831
The_Imitation_Game	0.0025	0.0162	0.15	0.8768
Pulp_Fiction	0.0048	0.0136	0.36	0.7206
Gone_Girl	-0.0098	0.0147	-0.67	0.5048
Avatar	-0.0005	0.0114	-0.05	0.9625
Matrix	-0.0006	0.0113	-0.05	0.9567
Frozen	-0.0005	0.0113	-0.05	0.9636
Silver_Linings_Playbook	0.0005	0.0129	0.04	0.9702
The_Hunger_Games	0.0016	0.0118	0.14	0.8922
Slumdog_Millionaire	-0.0008	0.0137	-0.06	0.9509
The_Princess_Bride	-0.0002	0.0113	-0.02	0.9849
Monty_Python_and_the_Holy_Grail	-0.0067	0.0126	-0.53	0.5952
Ferris_Buellers_Day_Off	0.0015	0.0144	0.11	0.9163
Love_Actually	0.0073	0.0120	0.61	0.5417
Fight_Club	-0.0054	0.0134	-0.40	0.6879
Shawshank_Redemption	-0.0018	0.0173	-0.10	0.9167
The_Social_Network	0.0003	0.0118	0.02	0.9802
The_Dark_Knight	0.0046	0.0113	0.41	0.6805

Example: linear regression

Using *lm* in R for linear model, we can fit a model with $y =$ height and $x =$ all of the other covariates.



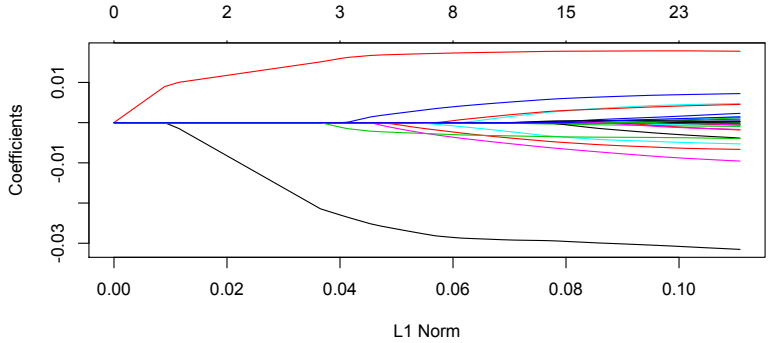
Example: linear regression to predict height

We can examine the coefficients of the unregularized linear regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.8758	4.0212	9.92	0.0000
gender	-2.1165	0.4892	-4.33	0.0000
shoe_size	1.2028	0.1219	9.87	0.0000
month	-0.0681	0.0477	-1.43	0.1547
digit	0.0614	0.0612	1.00	0.3162
sleep	-0.0632	0.1255	-0.50	0.6149
siblings	0.0272	0.1327	0.20	0.8378
handed	-0.2915	0.3911	-0.75	0.4569
thumb	-0.1489	0.3015	-0.49	0.6219
registered	0.0254	0.4259	0.06	0.9525
The_Imitation_Game	0.1581	0.2834	0.56	0.5777
Pulp_Fiction	0.3400	0.2368	1.44	0.1525
Gone_Girl	-0.6567	0.2556	-2.57	0.0109
Avatar	-0.0411	0.1984	-0.21	0.8359
Matrix	-0.0462	0.1976	-0.23	0.8152
Frozen	-0.0317	0.1978	-0.16	0.8728
Silver_Linings_Playbook	0.0409	0.2234	0.18	0.8548
The_Hunger_Games	0.1123	0.2049	0.55	0.5841
Slumdog_Millionaire	-0.0612	0.2408	-0.25	0.7997
The_Princess_Bride	-0.0168	0.1969	-0.09	0.9321
Monty_Python_and_the_Holy_Grail	-0.4557	0.2212	-2.06	0.0406
Ferris_Buellers_Day_Off	0.1078	0.2529	0.43	0.6704
Love_Actually	0.4970	0.2076	2.39	0.0175
Fight_Club	-0.3708	0.2356	-1.57	0.1170
Shawshank_Redemption	-0.1371	0.3041	-0.45	0.6526
The_Social_Network	0.0230	0.2062	0.11	0.9112
The_Dark_Knight	0.3263	0.1963	1.66	0.0979

Example: sparse Poisson regression

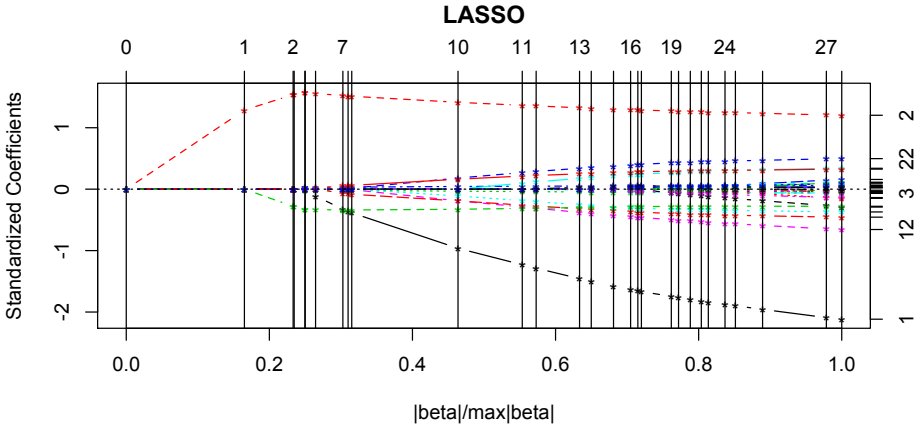
Using *glmnet* package in R, ℓ_1 regularized Poisson GLM, we can fit a model with $y =$ height and $x =$ all of the other covariates.



The trace shows that *shoe size*, *gender* enter regression first, but *gender* (-) and *Gone Girl* (-) ratings have large magnitude coefficients at the end.

Example: sparse linear regression

Using *lars* package in R, ℓ_1 regularized linear regression, we can fit a model with $y =$ height and $x =$ all of the other covariates.



Summary

- The exponential family of distributions contains a number of useful distributions and has important mathematical properties.
- Generalized linear models (GLMs) take advantage of these properties to allow linear regression for exponential family-distributed responses.
- While interpretation of parameters/residuals is different across GLMs, basic machinery is identical.

Additional Resources

- McCullagh & Nelder (1989) *Generalized Linear Models*
- MLAPA: Chapter 9
- *Elements of Statistical Learning*: Chapter 4
- *Pattern Recognition and Machine Learning*: Chapter 4
- *Stochastic Gradient Descent* [Robbins & Monro, 1951]

- (video) Alex Smola: *Exponential Families, Part I*
- Metacademy: *exponential families*
- Metacademy: *generalized linear models*