STATS 212: Generalized Linear Models
Lecture 1: A Review of Linear Regression Models

Babak Shahbaba

UCI, Spring 2010

Linear models

- Linear models have been extensively used in practice.
- They include a large class of models such as ANOVA and linear regression.
- They owe their popularity mostly to the fact that they are easy to interpret. (The computational aspect was also used to be a factor in the past, but it is less crucial these days.)
- We use these models to capture the relationship between the response variable, y, and a set of explanatory variables (predictors, covariates, ...), x.
- What does it mean for two random variables to be related?
- When we talk about relationship between y and x, we usually think about the change in the conditional distribution of y given x, i.e., $P(y|x)$, as x changes.

Relationship

- Regression models are based on the assumption that the only change in the conditional distribution we are interested in is the change in the expectation of the distribution, $E(y|x)$ (note that this by itself imposes limitations on the type of relationships we can detect).
- In general, this means $E(y|x) = g(x)$, and the relationship between x and y exists if $g(x)$ is not a constant function.
- In this setting, $g(x)$ also defines the type of relationship between x and y.

Linear regression models

- For linear regression models, $g(x)$ is a linear function in terms of model parameters, β.
- Recall that a function $f : \mathcal{R}^n \to \mathcal{R}^m$ is linear if
  - $f(z+t) = f(z) + f(t), \quad \forall z, t \in \mathcal{R}^n$
  - $f(az) = af(z), \quad \forall z \in \mathcal{R}^n, \forall a \in \mathcal{R}$
- The function $g(x)$ has the following general form:

$$g(x) = x\beta$$

  where x is a $n \times (p+1)$ matrix (the first column is the constant 1, and the remaining p columns are the observed value of p explanatory variables)
- β is a $(p+1)$-vector of parameters. The first element of this vector is the intercept, and the remaining parameters are called regression coefficients.

Linear regression models

- In regression terminology, $\epsilon = y - g(x)$ is called the *error*, which is a random variable assumed to be independent of x.
- We can therefore write the relationship between the response variable y and the explanatory variables x as follows:

$$y = g(x) + \epsilon$$

- For the observed data, we usually refer to the corresponding values of ϵ as *residuals*.

Least squares method

- There are many ways to estimate β, one of the most popular approach is the method of *least squares*, which is in general an optimization problem with no constraints

$$\text{minimize}||y - x\beta||_2^2$$

- Recall that ℓ_2-norm (Euclidean norm) is defined as

$$||z||_2 = (|z_1|^2 + |z_2|^2 + ... + |z_n|^2)^{1/2}$$

  In general, the ℓ_p norm ($p \geq 1$) is as follows:

$$||z||_p = (|z_1|^p + |z_2|^p + ... + |z_n|^p)^{1/p}$$

- $||y - x\beta||_2^2 = \sum_{i=1}^{n}(y_i - x_i\beta)^2$ is called residual sum of squares, *RSS*, which is a quadratic function of regression parameters, $RRS(\beta)$.

Least squares method

- To find the value of β that minimizes $RSS(\beta)$, we set the first derivative to zero,

$$\begin{aligned} \frac{\partial RRS}{\partial \beta} &= -2x'(y - x\beta) \\ \frac{\partial^2 RSS}{\partial \beta \partial \beta'} &= 2x'x \end{aligned}$$

- To have a unique solution for β, $x'x$ needs to be positive definite (x has to be full column rank).
- If this holds, the unique solution is obtained by setting the first derivative to zero

$$\begin{aligned} -2x'(y - x\beta) &= 0 \\ \hat{\beta} &= (x'x)^{-1}x'y \end{aligned}$$

Geometrical view of least squares

- The least squares estimate for the response variable is

$$y = x\hat{\beta} = x(x'x)^{-1}x'y$$

- Consider the n observed data points as vectors in $\mathcal{R}^n$
- The column vectors of x span a subspace of $\mathcal{R}^n$
- Denote the linear subspace of $\mathcal{R}^n$ as $\mathcal{L}(x)$
- Each point in this linear subspace can be presented as a linear function of column vectors $x_0, x_1, ..., x_p$

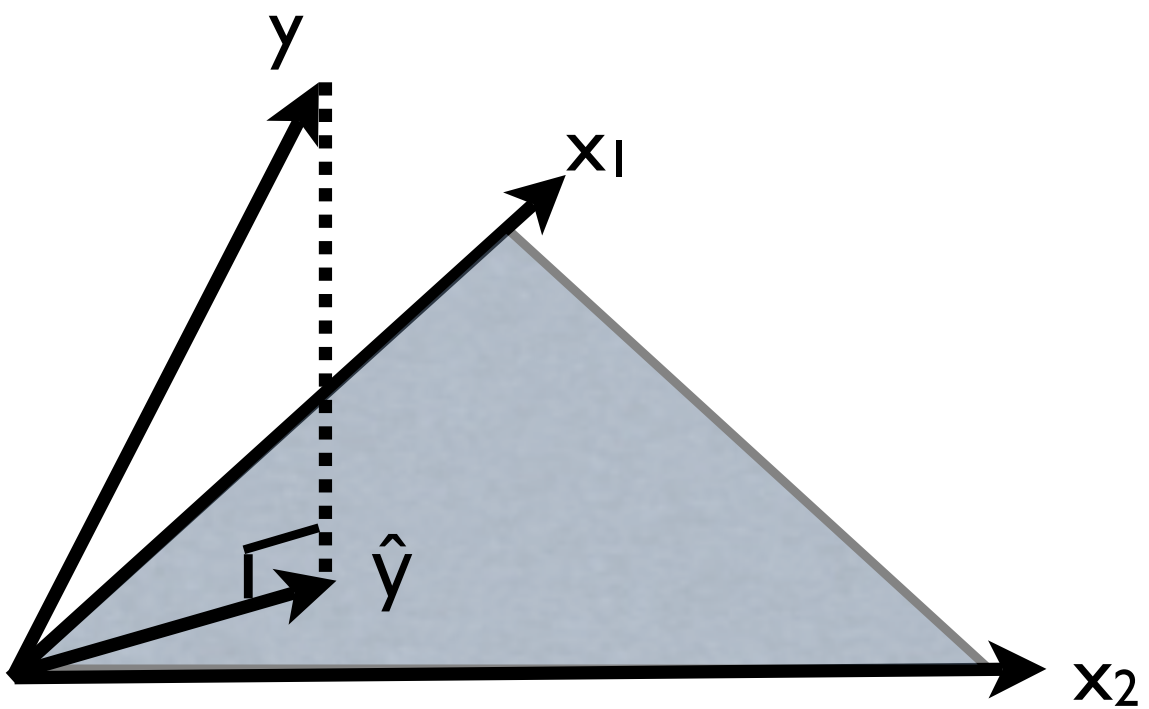
$$\mathcal{L}(x) = \{b_0x_0 + b_1x_1 + ... + b_px_p | b_0, b_1, ..., b_p \in \mathcal{R}\}$$

- Or in matrix form

$$\mathcal{L}(x) = \{xb | b \in \mathcal{R}^{p+1}\}$$

Geometrical view of least squares

- The least squares method provides the point in $\mathcal{L}(x)$, denoted as $\hat{y} = x\hat{\beta}$, that has the closesst Euclidean distance to $y \in \mathcal{R}^n$.
- This point is obtained by the orthogonal projection of y onto $\mathcal{L}(x)$ using the projection matrix $H = x(x'x)^{-1}x'$.



Geometrical view of least squares

- The projection matrix, H, is also called the *hat matrix* since puts a hat on y.
- H is symmetric ($H' = H$) and idempotent ($H^2 = H$).
- $I - H$ is also symmetric and idempotent. This is the projection matrix onto $\mathcal{L}^{\perp}(x)$, where

$$\epsilon = (I - H)y$$

Geometrical view of least squares

- In other words, $x'\epsilon = 0$; that is, the residual vector is independent of x, and it is orthogonal to $\mathcal{L}(x)$.
- Note that we are in fact decomposing $y \in \mathcal{R}^n$ onto two orthogonal spaces

$$
\begin{array}{ccccc}
y = & x\beta & + & (y - x\beta) \\
\text{space} & \mathcal{L}(x) & & \mathcal{L}^{\perp}(x) \\
\text{dimension} & \text{p} + 1 & & \text{n - p - 1}
\end{array}
$$

Prediction

- For a future observation whose values of explanatory variables are $\tilde{x}$, the *predicted* value of the response variable is

$$\tilde{y} = \tilde{x}\hat{\beta} = \tilde{x}(x'x)^{-1}x'y$$

- What is the 95% confidence interval for $\tilde{y}$?

Limitations of least squares

- In general, the least squares method would not work if the column vectors of x are not linearly independent (i.e., there are redundancy), or $p > n$ (more covariates than observations).
- In the first case, we can of course remove the redundant covariate. In the second scenario, we can use regularization.

Sampling distribution of parameters

- So far, we have not made any assumption regarding the distributional form of the random variables (more specifically for the response variable since x is assumed to be fixed).
- We did not need to make such assumptions if all we wanted was point estimates of regression parameters.
- Usually, we want more than point estimates; we, for example, want to know about variability (e.g., standard error) of the estimates.

Sampling distribution of parameters

- For this, we assume that x are fixed at the observed value and y's are uncorrelated with a constant variance; i.e., $cov(y|x) = \sigma^2 I$ (note that we have not fully specified the distribution yet).
- As the result,

$$\begin{aligned} cov(\hat{\beta}) &= (x'x)^{-1}x'[(x'x)^{-1}x']'\sigma^2 \\ &= (x'x)^{-1}x'x(x'x)^{-1}\sigma^2 \\ &= (x'x)^{-1}\sigma^2 \end{aligned}$$

- We also have

$$\begin{aligned} E(\epsilon) &= E(y) - E(E(y|x)) = E(y) - E(y) = 0 \\ var(\epsilon) &= \sigma^2 \end{aligned}$$

Estimating σ

- σ itself is almost always unknown and needs to be estimated based on the data.
- To estimate σ, we usually use the following unbiased estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y - x_i\hat{\beta})^2}{n - p - 1}$$

  We use $n - p - 1$ instead of n to make the estimate unbiased.
- The fit of the model can be measured based on $\hat{\sigma}^2$.
- For this, we use $R^2 = 1 - \frac{\hat{\sigma}^2}{S_y^2}$, which is the fraction of variance of response variable explained by the model. Here, S_y^2 is the observed variance of y.

Inference

- Note that while we could provide a measure of variability for the estimator of regression parameters, to perform statistical inference about these parameters, we need to make more assumptions about the distribution of y.
- We assume that

$$y|x, \beta, \sigma \sim N(x\beta, \sigma^2 I)$$

- Therefore,

$$\epsilon|\sigma \sim N(0, \sigma^2 I)$$

- As the result, we have

$$\begin{aligned} \hat{\beta}|\sigma &\sim N(\beta, (x'x)^{-1}\sigma^2) \\ \frac{n\hat{\sigma}^2}{\sigma^2} &\sim \chi^2(n-p-1) \end{aligned}$$

- Moreover, we can show that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Inference

- Using the sampling distribution of β, we can obtain the confidence interval for a given confidence level c.
- For each individual β_j (corresponding to x_j), the standard error is the square-root of the i^{th} element of the covariance matrix $(x'x)^{-1}\sigma^2$.
- The c level confidence interval for β_j can be obtained as

$$\hat{\beta}_j \pm t_c^* se(\hat{\beta}_j)$$

where t_c^* is the corresponding t-critical value based on $t(n-p-1)$ distribution.

Inference

- To test the null hypothesis $H_0 : \beta_j = 0$, we can use the following T-statistics:

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- Under H_0, T has a $t(n - p - 1)$ distribution.
- If we want to test the null hypothesis with respect to a group of coefficients, i.e., $H_0 : \beta_1 = \beta_2 = ..., \beta_s = 0$, we use the F statistic

$$F = \frac{(RSS_r - RSS)/s}{RSS/(n - p - 1)} \sim \mathcal{F}(s, n - p - 1)$$

where RSS_r is the residual sum of squares for the reduced model.

Gauss-Markov Theorem

- Suppose $var(y) = \sigma^2 I$.
- Let $\tilde{\beta} = cy$ be an unbiased estimator of β.
- Then, the variance of linear functions of $\tilde{\beta}$ is at least as great as the variance of linear functions of $\hat{\beta}$
- That is, the ordinary least squares estimate is the best linear unbiased estimator (BLUE) of β.

Likelihood function

- An alternative approach for estimating the parameters of linear regression model (an in general, all statistical models) is based on likelihood function.
- To find the likelihood function, we first need to assume a probability distribution for the data, i.e., $P(y|\theta)$, where θ are unknown parameters.
- This distribution is based on our opinion regarding the mechanism that generates the data.
- The likelihood function is defined by plugging-in the observed data in the probability distribution and expressing it as a function of model parameters, i.e., $f(\theta, y)$.

Likelihood function

- For linear regression models, the data include the response variables y and the explanatory variables x. Therefore, in general we need to specify $P(x, y)$.
- However, since x are assumed to be fixed at their observed value, $P(x) = 1$, the joint distribution reduces to the conditional distribution of $y|x$.

$$P(x, y) = P(x)P(y|x) = P(y|x)$$

- Therefore, we only need to specify the conditional distribution of y given x.

Likelihood function

- We assume this $P(y|x)$ is a normal distribution.
- As we mentioned, we model the expectation of this distribution as a linear function of x, i.e., $E(y|x) = x\beta$, and we assume the variance of this distribution is σ^2 (which is independent of x and β).
- Therefore, assuming that the observations are independent, we have

$$y|x, \beta \sim (2\pi\sigma^2)^{-n/2} \exp(-\frac{\sum_{i=1}^{n}(y_i - x_i\beta)^2}{2\sigma^2})$$

- The likelihood function is specified by plugging-in the observed values of x and y in the probability distribution and expressing the result as a function of β (for now, we assume σ is fixed).

$$f(\beta) = (2\pi\sigma^2)^{-n/2} \exp(-\frac{\sum_{i=1}^{n}(y_i - x_i\beta)^2}{2\sigma^2})$$

Maximum likelihood estimation

- To estimate model parameters, we can find their values such that the probability of the observed data is maximum.
- For this, we maximize the likelihood function with respect to model parameters. Of course, it is easier to maximize the log of likelihood function, i.e., $L(\beta) = log(f(\beta))$.
- This in general is convex optimization assuming that the function is log-concave in β for a fixed x and y.
- To maximize the likelihood function, we obviously need to focus on the part of the function that is related to the parameter (this part of the likelihood function is called *kernel*).
- For linear regression models,

$$L(\beta) = -\sum_{i=1}^{n}(y_i - x_i\beta)^2 - log(2\sigma^2)$$

Maximum likelihood estimation

- For simplicity, we can also remove all the constant (not related to the parameters) parts;

$$L(\beta) = -\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

- Now we can simply set the first derivative to zero (likelihood equation) to obtain the maximum likelihood estimate

$$\begin{aligned} \frac{\partial L(\beta)}{\partial \beta} &= 2\sum_{i=1}^{n} x_i(y_i - x_i\beta) \\ x'(y - x\beta) &= 0 \\ \hat{\beta} &= (x'x)^{-1}x'y \end{aligned}$$

- In this case, MLE is the same as the least squares estimate.

Maximum likelihood estimation

- Under weak regularity conditions, the MLE demonstrates attractive properties as $n \to \infty$: the asymptotic distribution of MLE is normal, MLE is asymptotically consistent and efficient.
- Under some regularity conditions (Rao, 1973), the asymptotic covariance matrix for MLE, $cov(\hat{\beta})$ is the inverse of *Fisher information matrix*, $i(\beta)$, where the (j, k) element of $i(\beta)$ is

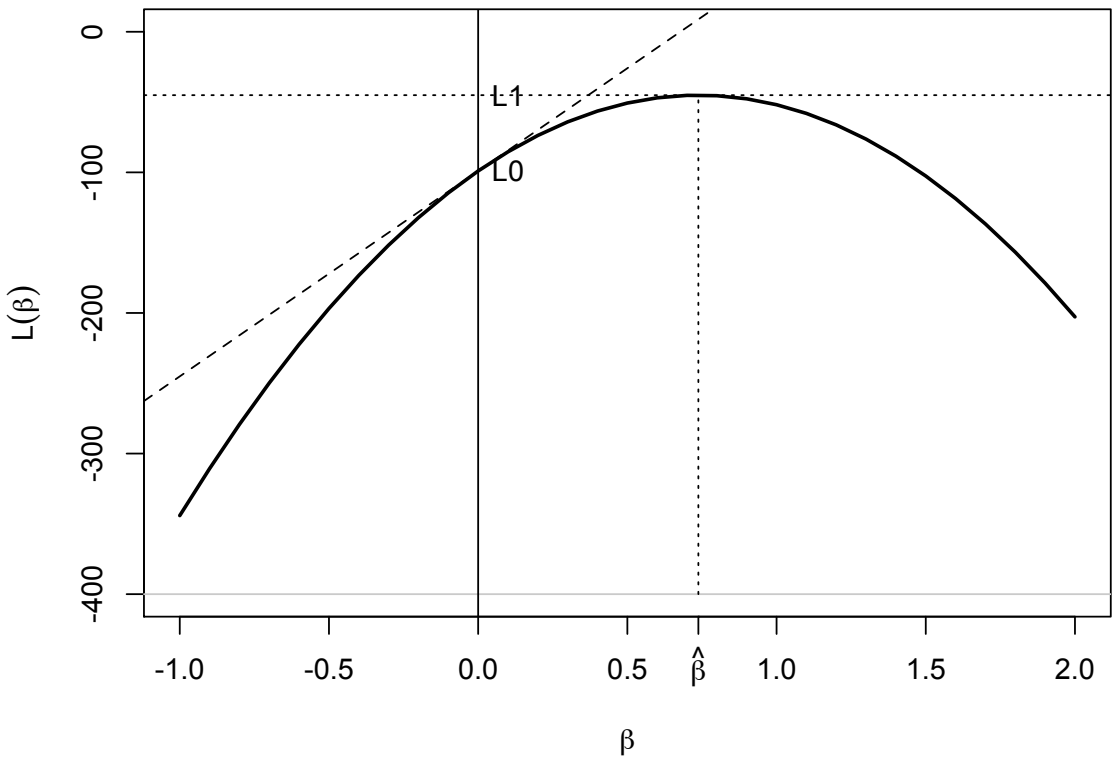
$$cov[\frac{\partial L(\beta)}{\partial \beta_j}, \frac{\partial L(\beta)}{\partial \beta_k}]$$

which is equal to the following (assuming that we can take differentiate twice inside integral)

$$-E\left(\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k}\right)$$

Maximum likelihood estimation

- This graphs shows the log-likelihood function and the location of MLE for randomly simulated data.



Wald, score, and likelihood ratio tests

- Wald, score, and likelihood ratio are three standard tests based likelihood function to perform statistical inference.
- Consider the null hypothesis $H_0 : \beta = \beta_0$, where β_0 is the value of β under the null.
- Due to large-sample normality of MLE, we have

$$w = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

  where w has an approximately $N(0, 1)$ distribution.
- This type of statistics where we use the standard error of the estimator (as opposed to standard deviation of the null distribution) is referred to as *Wald statistic*.

Wald, score, and likelihood ratio tests

- The multivariate version of this statistic is

$$w^2 = (\hat{\beta} - \beta_0)'[cov(\hat{\beta})]^{-1}(\hat{\beta} - \beta_0)$$

- Asymptotically, w^2 has χ^2 distribution with df equal to the rank of $cov(\hat{\beta})$.

Wald, score, and likelihood ratio tests

- Score test on the other hand is based on the slope at β_0.
- This is in fact the value of *score* function, $u(\beta) = \partial L(\beta)/\partial\beta$, evaluated at β_0.
- The dashed line in the above graph shows the slope at $\beta_0 = 0$.
- As we expect, the further β_0 is away from the MLE, the larger this slope becomes in absolute value (i.e., we can reject the null hypothesis more confidently).

Wald, score, and likelihood ratio tests

- The score statistic is obtained by dividing the $u(\beta_0)$ by its corresponding standard error, $\sqrt{i(\beta_0)}$
- Therefore,

$$s = \frac{u(\beta_0)}{\sqrt{i(\beta_0)}} \sim N(0,1)$$

- Alternatively,

$$s^2 = \frac{[u(\beta_0)]^2}{i(\beta_0)} \sim \chi^2(1)$$

Wald, score, and likelihood ratio tests

- For the multi parameter case, the score test has the following form (note that in general, $E(u) = 0$ and $cov(u) = i(\beta)$)

$$u'(\beta_0) i^{-1}(\beta_0) u(\beta_0)$$

This has an asymptotic χ^2 distribution with the the df equal to the number of constraints.

Wald, score, and likelihood ratio tests

- The advantage of score test is that we do not need to estimate the maximum likelihood estimate.
- The third test statistic is the likelihood ratio test.
- Here, we maximize the likelihood function under H_0 and under $H_0 \cup H_a$ (where H_a is the alternative hypothesis).
- The ratio of these two maximums is called the likelihood ratio test. In general,

$$LR = \frac{\sup_{\theta \in \Omega_0} f(\theta)}{\sup_{\theta \in \Omega} f(\theta)}$$

where Ω_0 is the parameter space under to H_0.

Wald, score, and likelihood ratio tests

- In general, the likelihood ratio cannot exceed 1, since the maximized value under H_0 would be less than or equal to the maximum value under $H_0 \cup H_a$.
- For hypothesis testing, we have $-2\log(LR) = -2(L_0 - L_1)$ has asymptotic χ^2 distribution with the degrees of freedom equal to the difference between the dimension of parameter space under $H_0 \cup H_a$ and under H_0.
- Here L_1 is the maximum value of log-likelihood under $H_0 \cup H_a$, and L_0 is the maximum value of log-likelihood under H_0.
- For the simple linear regression, when testing the null hypothesis, $H_0 : \beta = \beta_0$, $L_1 = L(\hat{\beta})$ and $L_0 = L(\beta_0)$.
- L_1 and L_0 (assuming $H_0 : \beta = 0$) are shown in the above figure.