

STATS 211: Statistical Methods II

Lecture 1: A brief review of linear regression models

Babak Shahbaba

UCI, Winter 2012

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 think about the change in the conditional distribution of y given x , i.e., changes in $P(y|x)$ 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 .
- 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 values of the p explanatory variables)

- β is a vector of $r = p + 1$ 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*.
- 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 one 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

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

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

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

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

Least squares method

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

$$\begin{aligned}\frac{\partial RSS}{\partial \beta} &= -2x^T(y - x\beta) \\ \frac{\partial^2 RSS}{\partial \beta \partial \beta'} &= 2x^T x\end{aligned}$$

- To have a unique solution for β , $x'x$ needs to be positive definite; x has to be full column rank: $r = p + 1$.
- If this holds, the unique solution is obtained by setting the first derivative to zero

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

Geometrical view of least squares

- The least squares estimate for the response variable is

$$\hat{y} = xb = x(x^T x)^{-1} x^T y$$

- $H = x(x^T x)^{-1} x^T$ is a projection matrix applied to y . But projection of y onto which subspace?
- Consider the n observed data points as vectors in \mathcal{R}^n
- The column vectors of x span a subspace of \mathcal{R}^n – the column space of x denoted as $\mathcal{C}(x)$
- Each vector in this subspace can be presented as a linear combination of column vectors x_0, x_1, \dots, x_p

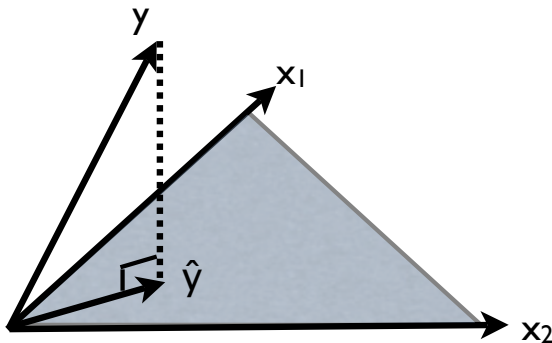
$$\mathcal{C}(x) = \{b_0 x_0 + b_1 x_1 + \dots + b_p x_p \mid b_0, b_1, \dots, b_p\}$$

- Or in matrix form

$$\mathcal{C}(x) = \{xb \mid b \in \mathcal{R}^r\}$$

Geometrical view of least squares

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



Geometrical view of least squares

- The projection matrix, H , is also called the *hat matrix* since it puts a hat on y .
- H is symmetric ($H^T = H$) and idempotent ($H^2 = H$).
- Now we can find the residual vector, e , as follows:

$$e = (I - H)y$$

- $I - H$ is also symmetric and idempotent. This is the projection matrix onto the null space of x^T (a.k.a., left null space), which is orthogonal to $\mathcal{C}(x)$, and is denoted as $\mathcal{C}^\perp(x)$.

Geometrical view of least squares

- In other words, $x^T e = 0$; that is, the residual vector is independent of x .
- As a result, $x_0^T e = 0$, where x_0 is the first column of x , and all of its elements are 1. This means $\sum_i e = 0$.
- Note that we are in fact decomposing $y \in \mathcal{R}^n$ onto two orthogonal spaces

$$\begin{array}{rcl} y = & xb & + (y - xb) \\ \text{space} & \mathcal{C}(x) & \mathcal{C}^\perp(x) \\ \text{dimension} & r & n - r \end{array}$$

Prediction

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

$$\tilde{y} = \tilde{x}b = \tilde{x}(x^T x)^{-1}x^T y$$

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 $r > n$ (more covariates than observations).
- In the first case, we can of course remove the redundant covariate. In the second scenario, we can use regularization, which we will discuss later.

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 do not need to make such assumptions if all we want are point estimates of regression parameters.
- Usually, we want more than point estimates; we, for example, want to know about the variability (e.g., standard error) of the estimates.

Sampling distribution of parameters

- We now 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,

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

- 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

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

We use $n - r$ (where $r = p + 1$) instead of n to make the estimate unbiased.

- The fit of the model can be measured based on $\hat{\sigma}^2$.
- For this, we can 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

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

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} diagonal element of the covariance matrix $(x^T x)^{-1} \sigma^2$.
- 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., $H_0 : \beta_1 = \beta_2 = \dots, \beta_s = 0$, 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.

Likelihood function

- Alternatively, we can use the likelihood function to estimate model parameters.
- 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 (e.g., β in regression models).
- 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., $f(\theta)$.

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 given 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 that $P(y|x)$ 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

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

- 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).

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

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., $L(\beta) = \log(f(\beta))$.
- In general, this is a convex optimization problem.
- To maximize the likelihood function, we obviously need to focus on the part of the function that is related to the parameter.
- 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

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

- 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 $n \rightarrow \infty$: 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, $\text{Cov}(\hat{\beta})$, is the inverse of *Fisher information matrix*, $i(\beta)$, where the (j, k) element of $i(\beta)$ is

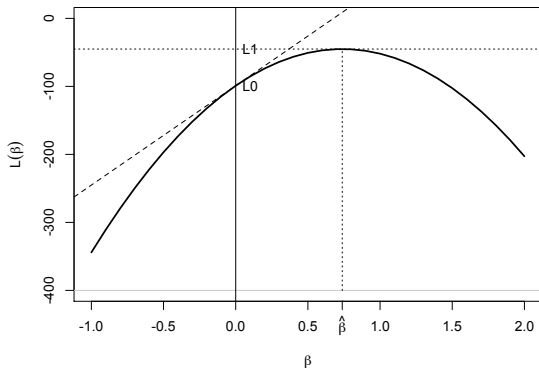
$$\text{Cov}\left[\frac{\partial L(\beta)}{\partial \beta_j}, \frac{\partial L(\beta)}{\partial \beta_k}\right]$$

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 graph 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 for performing 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 approximate $N(0, 1)$ 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

$$w^2 = (\hat{\beta} - \beta_0)^T [\text{Cov}(\hat{\beta})]^{-1} (\hat{\beta} - \beta_0)$$

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

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 $Cov(u) = i(\beta)$)

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

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

- The advantage of score test is that we do not need to estimate the maximum likelihood estimate.

Wald, score, and likelihood ratio tests

- 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, $-2 \log(LR) = -2(L_0 - L_1)$ 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.

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 briefly discuss how some of these assumptions can be violated.
- Addressing these issues is the focus of this course.

Nonlinearity

- Using linear models, we implicitly assume that the relationship between x and y is linear (note that this 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 does not hold, we might still be able to use linear regression models after some transformation (log, square root, polynomial) of original variables.
- In this course, we will discuss more complex strategies for nonlinear modeling using basis expansion and generalized additive models.

Dependent (clustered) observations

- In linear regression models, the observations are assumed to be independent.
- In this course, we discuss modeling of clustered data, where the independence assumption does not hold.
- We are especially interested in situations where there are multiple observations for the same subject over time (longitudinal data).
- We also discuss strategies to deal with the non-constant variance problem.

Bounded 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 hand 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.
- To deal with these issues, we need a more flexible family of models.
- The class of generalized linear models (discussed in Stats 212), that includes linear models as a special case, provides such flexibility while it is still easy to use.
- In this course, we discuss some alternative strategies specific to modeling categorical response variables.