STATS 205 Notes

Ryan Anderson

2024-04-01

Lecture 4/1: Simple Linear Models

Let $p = 1$. Consider the case of the *simple linear model* on pairs of observations (X_i, Y_i) which gives

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The $X_i\beta_1 + \beta_0$ component can be alternately thought of as the "mean" or "signal" and the ϵ_i component can be thought of as the "noise" or "error".

One critical thing in linear modeling is to be able to identify where the randomness comes from. In particular, here the randomness comes from the error term $\epsilon_i \sim^{iid} N(0, \sigma^2)$ - this is a *fixed design* model. By contrast, if we allowed the X_i to drive the randomness we would have a *random design* model.

One justification for the choice of the fixed design model is the ancillarity principle. Consider a parametrization of the above model $\theta = (\beta_0, \beta_1, \sigma^2)$. The joint distribution of (X_i, Y_i) is parametrized by both θ and ψ, which governs the marginal distribution of X_i:

$$P_{(\theta,\psi)}(X_i, Y_i) = P_\psi(X_i)P_\theta(Y_i|X_i)$$

But we're only interested in estimating $P_\theta(Y_i|X_i)$. The marginal distribution. of X_i has nothing to do with θ, so we are justified in ignoring it and treating the data as fixed.

Estimating β_0 and β_1

Given the goal of such linear models is to maximize the interpretability of the relationship between X and Y, we want to infer the values of β_0 and β_1 that best explain the data. We obtain point estimates of the pair $(\hat{\beta}_0, \hat{\beta}_1)$ via least squares.

We also want to estimate the variance of the point estimates, $var(\hat{\beta}_0)$ and $var(\hat{\beta}_1)$. The variance here will be driven again by the randomness due to the error term, so each of the variances will be a function of σ^2. However, we don't even know σ^2! So we estimate the variance via $v\hat{a}r(\hat{\beta}_1)$ and $v\hat{a}r(\hat{\beta}_0)$ - doing so in this fashion is a *plug-in estimate.*

Multiple Linear Regression

Now let $p > 1$ and consider tuples of observations $(Y_i, X_{i1}, \ldots, X_{ip})$. The multiple linear regression model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i$$

In matrix form we have $Y = X\beta + \epsilon$ where

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times(p+1)}$, $\beta \in \mathbb{R}^{p+1}$, and $\epsilon \in \mathbb{R}^n$.

Our least squares estimate of β is given by

$$\begin{aligned}\sum(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2 &= \sum(Y_i - X_i^T\beta)^2 \\ &= (Y - X\beta)^T(Y - X\beta)\end{aligned}$$

Differentiating with respect to β and setting equal to zero gives the normal equations

$$\begin{aligned}\frac{d}{d\beta}(Y - X\beta)^T(Y - X\beta) &= 0 \\ \frac{d}{d\beta}Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta &= 0 \\ -2X^TY + 2X^TX\beta &= 0 \\ \Rightarrow \hat{\beta}_{LS} &= (X^TX)^{-1}X^TY\end{aligned}$$

This famous least squares estimate for β can be viewed as a projection from the space of Y onto the space spanned by the columns of X.

We also obviously need X^TX to be invertible, which is equivalent to the columns of X being linearly independent.

The error in the multiple linear regression framework still comes from the error vector ϵ, so we have $E[\hat{\beta}_{LS}] = (X^TX)^{-1}X^TE[Y] = (X^TX)^{-1}X^T(X\beta)$, as ϵ has 0 mean. This gives us $E[\hat{\beta}_{LS}] = \beta$.

Similarly, we can estimate the variance of $\hat{\beta}_{LS}$ via $var(\hat{\beta}_{LS}) = ((X^TX)^{-1}X^T)var(Y)(X(X^TX)^{-1})$. If $var(\epsilon)$ is isotropic, i.e., $var(\epsilon) = \sigma^2 I$, then we have $v\hat{a}r(\hat{\beta}_{LS}) = \sigma^2(X^TX)^{-1}$.

Estimating Parameters via MLE

Above we did not need to specify parametric models for the error term. If we instead do so, we can obtain maximum likelihood estimates of the parameters. They will turn out to coincide with the least squares estimates in the case of a Gaussian error term.

Let $Y_i|X_i \sim^{iid} N(X_i^T\beta, \sigma^2)$ and let the likelihood function be $L(\beta) = \prod_{i=1}^n f(Y_i|X_i;\beta)$ where f is the density of the error term. We have

$$\begin{aligned}L(\beta) &= \prod_{i=1}^n f(Y_i|X_i;\beta) \\ L(\beta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - X_i^T\beta)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - X_i^T\beta)^2\right)\end{aligned}$$

Maximizing $L(\beta)$ then will be done by minimizing the term in the exponential, which is the same as minimizing the least squares criterion.

$$\arg\max_{\beta} L(\beta) = \arg\min_{\beta} \sum_{i=1}^n (Y_i - X_i^T\beta)^2$$

Evaluating Fitted Values

Given a model $Y = X\beta + \epsilon$, we can evaluate the fitted values $\hat{Y} = X\hat{\beta}$. The residuals are then given by $e = Y - \hat{Y} = Y - X\hat{\beta}$.

Moreover, we have

$$\begin{aligned} e &= (Y - X\hat{\beta}) = Y - X(X^TX)^{-1}X^TY \\ e &= (I_n - X(X^TX)^{-1}X^T)Y \end{aligned}$$

The matrix $H = X(X^TX)^{-1}X^T$ is called the *hat matrix* and it is the projection matrix that projects Y onto the space spanned by the columns of X. H is symmetrical and idempotent and moreover, $tr(H) = tr(X^TX(X^TX)^{-1}) = p + 1$. We also have $tr(I_n - H) = n - (p + 1)$.

Note this gives a way to estimate another parameter. $\hat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}e^Te$.

What distribution does the MLE of σ^2 follow? We know that if we were instead estimating an object of the form $\sum \frac{e_i^2}{\sigma^2} \sim \chi_n^2$, as we have a sum of squared standard normals.

Instead we have $\sum \frac{e_i^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$, as we have $n - (p + 1)$ degrees of freedom. This is because the residuals are not independent, but they are orthogonal to the fitted values. $E[\sum \frac{e_i^2}{\sigma^2}] = n - (p + 1)$, which allows us to say that $E[\hat{\sigma^2}_{MLE}] = \frac{n-(p+1)}{n}\sigma^2$. Obviously this is not unbiased!