COS513: FOUNDATIONS OF PROBABILISTIC MODELS
LECTURE 9: LINEAR REGRESSION

SEAN GERRISH AND CHONG WANG

1. WAYS OF ORGANIZING MODELS

In probabilistic modeling, there are several ways of organizing models:

(1) Bayesian vs. Frequentist.
(2) Discriminative vs. Generative.
   (a) Discriminative: conditioned on some information, e.g.regression models, classification models.
   (b) Generative: we fit a probability distribution to every part of the data, e.g. clustering, naive Bayesian classification.

   Is discriminative better or generative better? A common myth is that one of these is always the appropriate solution.
(3) Per-data point prediction vs. Data set density estimation.
(4) Supervised vs. Unsupervised models.
   (a) Supervised: given $\{(x_i, y_i)\}_{i=1}^N$ in training, predict y given x in testing (e.g. classification).
   (b) Unsupervised: given data, we seek to find structure in it. Clustering is an example.

All of these models involve

(1) treating observations as random variables in a probability distribution; and
(2) computing something about the distribution.

2. LINEAR REGRESSION

In this section, we will talk about the basic idea of linear regression and then study how to fit a linear regression.

2.1. **Overview of linear regression.** Linear regression is a method to predict a real valued response y from covariates x using linear models. See Figure 1 shows an example. Usually, we have multiple *covariates* $x =< x_1, x_2, \ldots, x_p >$, where p is the number of covariates.

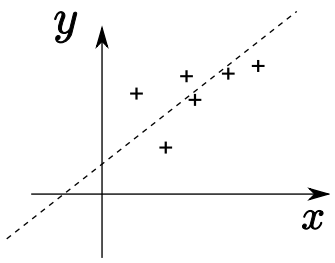


FIGURE 1. Linear regression. '+'s are data points and the dashed line is the output of fitting the linear regression.

In linear regression, we fit a linear function of covariates

$$f(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \beta_0 + \beta^T x. \tag{1}$$

Note that $\beta^T x = 0$ is a hyperplane.

Many candidate features can be used as the input x:

(1) any raw numeric data;
(2) any transformation, e.g. $x_2 = \log x_1$ and $x_3 = \sqrt{x_1}$;
(3) basis expansions, e.g. $x_2 = x_1^2$ and $x_3 = x_1^3$;
(4) indicator functions of qualitative inputs, e.g. 1[the subject has brown hair]; and
(5) interactions between other covariates, e.g. $x_3 = x_1 x_2$.

2.2. **Fitting a linear regression.** Suppose we have a dataset $D = \{(x_n, y_n)\}_{n=1}^N$. In the simplest form of a linear regression, we assume $\beta_0 = 0$ and $p = 1$. So the function to be fitted is just

$$f(x) = \beta x. \tag{2}$$

To fit a linear regression in this simplified setting, we minimize the sum of the distances between fitted values and the truth. Thus, the objective function is (if we use Euclidean distance)

$$\text{RSS}(\beta) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \beta x_n)^2, \tag{3}$$

where RSS stands for *Residual Sum of Squares*. Figure 2 illustrates this. To minimize $\text{RSS}(\beta)$, we take its derivative,

$$\frac{d\,\text{RSS}(\beta)}{d\,\beta} = -\sum_{n=1}^{N} (y_n - \beta x_n) x_n. \tag{4}$$

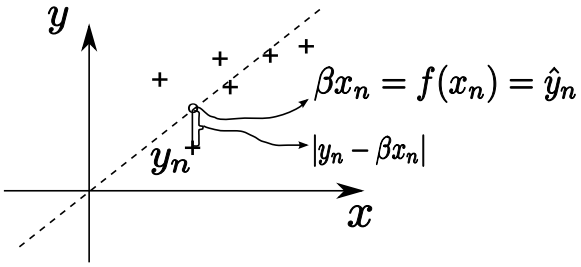


FIGURE 2. Linear regression. '+'s are data points and the dashed line is the output of fitting the linear regression.

Since RSS(β) is convex, setting Equation 4 to zero and solving for $\hat{\beta}$ leads to an algebraic version of the optimal solution,

$$\hat{\beta} = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} x_n^2}. \tag{5}$$

When we have a new input x_{new}, then, the prediction is simply $\hat{y}_{new} = \hat{\beta} x_{new}$. We can generalize Equation 2 by allowing for a constant offset in the predictions:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i.$$

Note that solving for β using the setup above does not determine the fixed offset β_0. We can get around this by setting $\beta_{p+1} = \beta_0$ and $x_{p+1} = 1$. Then we have,

$$y = \beta^T x.$$

As noted above, the RSS gives a sense of how accurate our estimate is. In many situations (such as our current one), we are interested in minimizing the RSS. One approach to find the minimum is to perform gradient ascent. Equation 4 gives the gradient of interest,

$$\nabla_\beta \, \mathrm{RSS}(\beta) = -\sum_{n=1}^{N} (y_n - \beta^T x_n) x_n, \tag{6}$$

of our objective function. With a convex objective function such as this RSS, the gradient is generally sufficient; we could stick this into a black box gradient descent algorithm and have a solution relatively efficiently. In this case, as noted in Equation 5, there fortunately also exists an explicit solution.

2.3. **Concise form of the exact solution.** We can more concisely describe the exact solution with a bit of linear algebra. The *design matrix* X contains a set of n observations in p dimensions. Using the constant-offset trick above, we also append 1 to each row of X:

$$X = \begin{bmatrix} x_{1,1} & \dots & x_{1,p} & 1 \\ x_{2,1} & \dots & x_{2,p} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,p} & 1 \end{bmatrix} \tag{7}$$

The *response vector* y describes the corresponding set of labels for these observations:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{8}$$

Combining Equations (6), (7), and (8) above, we can write the gradient of β concisely as

$$\nabla_\beta \operatorname{RSS}(\beta) = -X^T(Y - X\beta).$$

Setting this to 0 and solving for β, we have

$$-X^T(Y - X\hat{\beta}) = 0 \tag{9}$$

$$\implies \quad X^T X\hat{\beta} = X^T Y \tag{10}$$

$$\implies \quad \hat{\beta} = (X^T X)^{-1} Y. \tag{11}$$

We observe a couple of things about the equations above. First, Equations 9-11 are sometimes referred to as the *normal equations*. In addition, note that the matrix $X^T X$ is invertible as long as X has rank $p + 1$, which requires that our covariates not be linearly dependent.

3. Probabilistic Interpretation

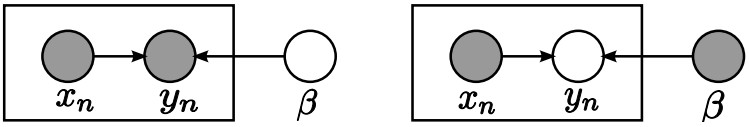
We can also frame linear regression using the probabilistic tools we have developed so far.

When fitting the model, we have access to the observations (x_n, y_n), and we seek to determine β. When we are predicting, our goal is to determine y_n, given x_n and β:

$$y_n \sim \mathcal{N}(\beta^T x_n, \sigma^2),$$

or, equivalently,

1. Draw $\epsilon \sim \mathcal{N}(0, \sigma^2)$
2. Set $y_n = \beta^T x_n + \epsilon$.



(a) Fitting (b) Predicting

FIGURE 3. In modeling linear regression with graphical models, (a). we first fit the parameter β, then (b). predict values y_n given a set of covariates x_n.

Note that we can interpret this as a discriminative model, because we always condition on x_n. Because of this, we don't need to specify the distribution of x_n in the model. Figure 3 illustrates this process.

Given this model, then, our goal is to find the conditional MLE of $\hat{\beta}$. The likelihood of β given our training data x and y is

$$l(\beta|x_{1:N}, y_{1:N}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_n - \beta^T x_n)^2}{2\sigma^2}\right).$$

Taking the logarithm and maximizing the log likelihood, we get

$$\beta_{\text{MLE}} := \underset{\beta}{\text{argmax}} \left(-\frac{1}{2}\sum_{n=1}^{N} \frac{(y_n - \beta^T x_n)^2}{\sigma^2}\right) = \underset{\beta}{\text{argmin}} \left(\frac{1}{2}\sum_{n=1}^{N} (y_n - \beta^T x_n)^2\right),$$

which is exactly the objective function we found our earlier treatment of the problem.

4. PREDICTION FROM EXPECTATION

4.1. **The Bias-Variance Tradeoff.** In many statistical applications, it is useful to understand how close our *empirical* distribution's mean $\hat{\beta}$ is to the true mean β of some underlying distribution. Such questions are central to countless applications of statistics, motivating metrics such as *standard error* used frequently in nearly every branch of science.

To formalize this a bit more, consider a dataset with N i.i.d. observations $\{(x_i, y_i)\}_{i=1}^N$ drawn randomly from some true distribution $\mathcal{D}$. The empirical MLE estimate $\hat{\beta}_{\text{MLE}}$, which we can compute from these observations, is then a random variable, with its randomness arising from the fact that the observations are themselves drawn from $\mathcal{D}$. Given an unseen datum $(x_{new}, y_{new}) \sim \mathcal{D}$, we seek to find how close $\mathbb{E}[y_{new}|x_{new}, \hat{\beta}_{\text{MLE}}]$ is to $\mathbb{E}[y_{new}|x_{new}, \beta] = \beta x$.

The *mean squared error* (MSE) is one measure of how close our estimator $\hat{\beta}_{\text{MLE}}$ is to the truth. We can decompose the MSE into both a variance

term and a bias term:

$$\begin{aligned}
\text{MSE} &= \mathbb{E}_{\mathcal{D}}[(\hat{\beta}X - \beta X)^2] \\
&= \mathbb{E}_{\mathcal{D}}[(\hat{\beta}X)^2] - 2\mathbb{E}_{\mathcal{D}}[\hat{\beta}X]\beta X + (\beta X)^2 \\
&= \mathbb{E}_{\mathcal{D}}[(\hat{\beta}X)^2] - 2\mathbb{E}_{\mathcal{D}}[\hat{\beta}X]\beta X + (\beta X)^2 + \left(\mathbb{E}_{\mathcal{D}}[\hat{\beta}X]^2 - \mathbb{E}_{\mathcal{D}}[\hat{\beta}X]^2\right) \\
(12) &= \mathbb{E}_{\mathcal{D}}[(\hat{\beta}X)^2] - \mathbb{E}_{\mathcal{D}}[\hat{\beta}X]^2 \\
(13) &\quad + (\mathbb{E}_{\mathcal{D}}[\hat{\beta}X] - \beta X)^2,
\end{aligned}$$

where Equation 12 is the variance of our estimator $\hat{\beta}$ and Equation 13 is its squared bias. When this bias is zero, the estimator is called an unbiased estimator; least squares is an example of such an estimator.

For many years, statisticians cared only about unbiased estimators. Recently, however, biased estimators have become more popular because it is sometimes possible to significantly decrease variance at the expense of a little bit of bias. This will be the topic of our next section. Figure 4 illustrates this.

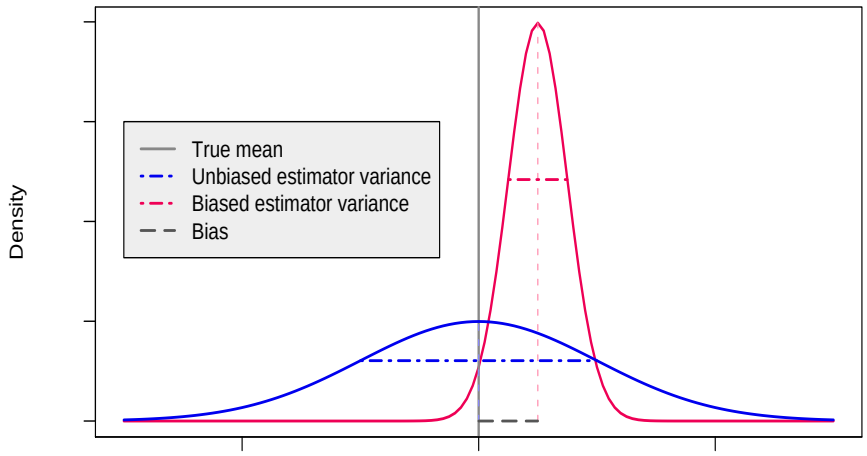


FIGURE 4. When attempting to determine the true parameter β of a distribution, we can use biased and unbiased estimators. Many estimators, such as standard least-squares, lead to unbiased estimates of the response means (blue). At other times, we may wish to use biased estimators, which may decrease variance of the estimate of the response mean at the expense of bias (red). (Here we have abused terminology and blurred the distinction between bias and squared bias.)