1 Likelihood function for linear regression

Consider the log likelihood for linear regression:

$$\ell(\mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_n) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right] \quad (1)$$

Define

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (2)$$

$$s^2 = (\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}}) \quad (3)$$

Show that

$$\ell(\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{s^2}{2\sigma^2}\right] \exp\left[-\frac{1}{2\sigma^2}(\mathbf{w}-\hat{\mathbf{w}})^T(\mathbf{X}^T\mathbf{X})(\mathbf{w}-\hat{\mathbf{w}})\right] \quad (4)$$

This expresses the log likelihood in terms how far $\mathbf{w}$ deviates from the OLS estimate $\hat{\mathbf{w}}$ (where distance is measured using a Mahalanobis distance with weight matrix $(\mathbf{X}^T\mathbf{X})^{-1}$). We will use this result when we discuss Bayesian inference for linear regression.

2 Multiple-output linear regression

(Source: Jaakkola)

Multiple output linear regression, which is an example of a **multiple response model**, is just like "regular" linear regression, except the output is a vector. (The term **multivariate linear regression** or **multiple linear regression** refers to the case where the *input* is a vector; since this is nearly always the case in practice, we just call this linear regression.) Hence we replace the weight vector with a weight matrix:

$$\mathbf{y}_i = \tilde{\mathbf{W}}\mathbf{x}_i + \boldsymbol{\epsilon}_i \quad (5)$$

where $\mathbf{x}_i$ is a column vector of p inputs (covariates), $\mathbf{y}_i$ is a column vector of q outputs (responses), and $\tilde{\mathbf{W}}$ is a $q \times p$ matrix. We assume the noise is uncorrelated, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \mathbf{I}_q)$.

Let $\tilde{\mathbf{Y}}$ be a matrix whose *columns* are $\mathbf{y}_1$ to $\mathbf{y}_n$, and $\tilde{\mathbf{X}}$ be a matrix whose *columns* are $\mathbf{x}_1$ to $\mathbf{x}_n$, and let $\mathbf{E}$ be a matrix whose *columns* are $\boldsymbol{\epsilon}_1$ to $\boldsymbol{\epsilon}_n$. Then we can write the above in matrix form as

$$\begin{pmatrix} | & & | \\ \mathbf{y}_1 & \cdots & \mathbf{y}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} - & \mathbf{w}_1 & - \\ & \vdots & \\ - & \mathbf{w}_q & - \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{pmatrix} + \begin{pmatrix} | & & | \\ \boldsymbol{\epsilon}_1 & \cdots & \boldsymbol{\epsilon}_n \\ | & & | \end{pmatrix} \quad (6)$$

or

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{W}}\tilde{\mathbf{X}} + \tilde{\mathbf{E}} \quad (7)$$

Since it is traditional to store the input examples $\mathbf{x}_i$ along the *rows* of the design matrix, we define $\mathbf{X} = \tilde{\mathbf{X}}^T$. Similarly, we define $\mathbf{Y} = \tilde{\mathbf{Y}}^T$, $\mathbf{E} = \tilde{\mathbf{E}}^T$ and $\mathbf{W} = \tilde{\mathbf{W}}^T$. With this, we get

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{E} \tag{8}$$

or

$$\begin{pmatrix} - & \mathbf{y}_1 & - \\ & \cdots & \\ - & \mathbf{y}_n & - \end{pmatrix} = \begin{pmatrix} - & \mathbf{x}_1 & - \\ & \cdots & \\ - & \mathbf{x}_n & - \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{w}_1 & \cdots & \mathbf{w}_q \\ | & & | \end{pmatrix} + \begin{pmatrix} - & \boldsymbol{\epsilon}_1 & - \\ & \cdots & \\ - & \boldsymbol{\epsilon}_n & - \end{pmatrix} \tag{9}$$

1. Consider the objective

$$J(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{y}_i - \tilde{\mathbf{W}}\mathbf{x}_i||^2 \tag{10}$$

   Show that the minimal least squares estimator is given by

$$\hat{\mathbf{W}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{11}$$

   Hint: show that the objective decomposes into q independent single-output least squares problems.

2. Consider the following dataset where $p = 1$, $q = 2$ and $n = 6$:

x	y
0	$(-1,-1)^T$
0	$(-1,-2)^T$
0	$(-2,-1)^T$
1	$(1,1)^T$
1	$(1,2)^T$
1	$(2,1)^T$

   If the input $\mathbf{x}_i$ is transformed through a set of basis functions, $\boldsymbol{\phi}(\mathbf{x}_i)$, we can write

$$\hat{\mathbf{W}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{Y} \tag{12}$$

   where $\boldsymbol{\Phi}$ is the modified design matrix. For example, for $x_i \in \{0, 1\}$ we can use $\boldsymbol{\phi}(0) = (1, 0)^T$ and $\boldsymbol{\phi}(1) = (0, 1)^T$. Thus we encode the binary input as a 2-dimensional vector, so $p = q = 2$. Compute $\hat{\mathbf{W}}$ from the above data using this set of basis functions.

3 Deriving the offset term

Let

$$J(\mathbf{w}, w_0) \quad = \quad (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n)^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n) \tag{13}$$

By solving $\frac{\partial}{\partial w_0}J(\mathbf{w}, w_0) = 0$, show that

$$\hat{w}_0 \quad = \quad \overline{y} - \overline{\mathbf{x}}^T\mathbf{w} \tag{14}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$.

4 Sufficient statistics for linear regression

(Source: Jaakkola) Consider fitting the model $\hat{y} = w_0 + w_1 x$ using least squares. Unfortunately we did not keep the original data, x_i, y_i, but we do have the following functions (statistics) of the data:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{15}$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{16}$$

$$C_{xx} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{17}$$

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{18}$$

$$C_{yy} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{19}$$

1. What are the minimal set of statistics that we need to estimate w_1?
2. What are the minimal set of statistics that we need to estimate w_0?
3. Suppose a new data point, x_{n+1}, y_{n+1} arrives, and we want to update our sufficient statistics without looking at the old data, which we have not stored. (This is useful for **online learning**.) We can do this for $\overline{x}$ as follows.

   $$\overline{x}^{(n+1)} \overset{\text{def}}{=} \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}\left(n\overline{x}^{(n)} + x_{n+1}\right) \tag{20}$$

   Show that this can be rewritten in a slightly "prettier" form as follows

   $$\overline{x}^{(n+1)} = \overline{x}^{(n)} + \frac{1}{n+1}(x_{n+1} - \overline{x}^{(n)}) \tag{21}$$

   This has the form: new estimate is old estimate plus correction. We see that the size of the correction diminishes over time (i.e., as we get more samples).
4. Derive a similar recursive update equation for $C_{xy}^{(n+1)}$ in terms of the new data, x_{n+1}, y_{n+1}, the old sufficient statistics $n, \overline{x}^{(n)}, \overline{y}^{(n)}, C_{xy}^{(n)}$, and the newly computed means, $\overline{x}^{(n+1)}$, $\overline{y}^{(n+1)}$. Simplify as much as possible.

5 Linear regression on prostate cancer data (Matlab)

Consider the prostate cancer dataset discussed in [HTF01]. There are 8 continuous inputs and 1 continuous response, namely lpsa, which stands for log of prostate-specific antigen. The (standardized) data is in the file `prostate.mat` which contains the following variables (amongst others)

Listing 1: :

```
  Name          Size            Bytes  Class      Attributes

  Xtest        30x8              1920  double
  Xtrain       67x8              4288  double
  names         1x9               624  cell
  ytest        30x1               240  double
  ytrain       67x1               536  double
```

Fit a simple linear model $\hat{y}(x) = w_0 + w_1 x_1 + \ldots + w_8 x_8$ by maximum likelihood on the training set. What coefficients $\mathbf{w}$ do you get? What is the mean squared error and its standard error on the test set? Turn in your numbers and code. (You should get the same results as Table 1, left column.

Term	LS	ridge
intercept	2.480	2.472
lcavol	0.676	0.366
lweight	0.303	0.228
age	-0.141	-0.021
lbph	0.209	0.151
svi	0.304	0.207
lcp	-0.287	0.039
gleason	-0.021	0.044
pgg45	0.266	0.117
Test MSE	0.586	0.541
SE	0.184	0.170

Table 1: Coefficients and accuracy of least squares and ridge regression on the prostate cancer data. Based on Table 3.3 of [HTF01]. Produced by Exercise 5.

References

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.