CS 660: Mathematical Foundations of Analytics

Dr. Francis Parisi

Pace University

Spring 2018

Linear Models

- Linear Regression
- Logistic Regression

Linear Models

Linear Regression

- Linear regression is useful for predicting a quantitative response
- Although it has been around for a very long time, it is still one of the most widely used statistical learning methods
- The linear model either assumes that the regression function $\mathbb{E}(Y|X)$ is linear, or that the linear model is a reasonable approximation
- In the linear model, we have an input vector $X' = (X_1, X_2, \ldots, X_p)$, and want to predict a real-valued output Y The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Linear Models

The X_j may be

- quantitative inputs
- transformations of quantitative inputs, such as log, square-root or square
- basis expansions, such as $X_2 = X_1^2, X_3 = X_1^3$, leading to a polynomial representation
- numeric or “dummy” coding of the levels of qualitative inputs
  For example, if G is a three-level factor input, we might create $X_j,\ j = 1, \dots, 3$, such that $X_j = I(G = j)$ where one of the X_j's is one, and the others are zero
- interactions between variables, for example, $X_3 = X_1 \times X_2$

Linear Models

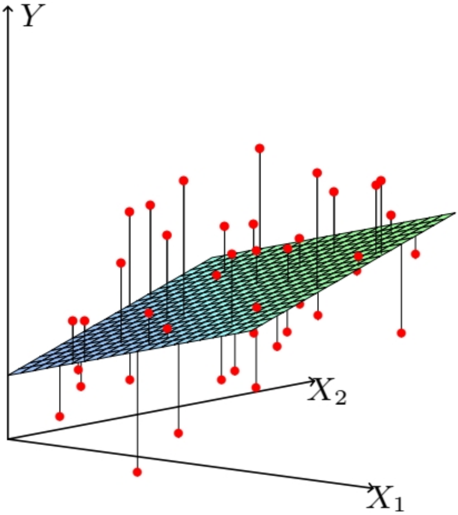
Finding β's

- The training data make up n ordered pairs $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ where the $\mathbf{x}_i$ are vectors of p explanatory variables
- The most popular estimation method is least squares, in which we pick the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ to minimize the residual sum of squares
- The residual sum of squares, $RSS(\beta)$, equals

$$\begin{aligned} RSS(\beta) &= \sum_{i=i}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=i}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} X_j \beta_j)^2 \end{aligned} \tag{6.1}$$

Linear Models

Figure 3.1 from Hastie et al. (2009)



Linear Models

- So how do we minimize (6.1)?
- Let $\mathbf{X}$ be the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position), and let $\mathbf{y}$ be the N-vector of outputs in the training set
- Then the residual sum-of-squares can be written as

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Linear Models

Differentiating with respect to β we get

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)$$

We set the derivative to zero

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = 0$$

to obtain the unique solution

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

Linear Models

The fitted values from the training data are found by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{6.2}$$

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in (6.2) is often called the **hat matrix** because it puts the hat on $\mathbf{y}$

Linear Models

If we know the distributional properties of $\hat{\beta}$ we can test for the significance and construct a confidence interval

Recall the regression model is

$$\begin{aligned} Y &= \mathbb{E}[Y|X_1, \ldots, X_p] + \epsilon \\ &= \beta_0 + \sum_{j=1}^{N} X_j \beta_j + \epsilon \end{aligned}$$

We have the variance-covariance matrix for $\hat{\beta}$ as

$$\mathrm{Var}\left(\hat{\beta}\right) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

Linear Models

Finally we have

$$\epsilon \sim N(0, \sigma^2) \text{ and } \hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$$

We can test the hypothesis that $\beta_j = 0$

First we find

$$Z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

where v_j is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$

Recall from hypothesis testing that $Z_j \sim t_{N-p-1}$

Linear Models

We can test for the significance of a group of coefficients simultaneously

Suppose we have a model ith p_1 parameters and we want to test if some of them, say $p_1 - p_0$, are zero

We can do this using a *partial* F test

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}$$

where RSS_1 is the residual sum of squares from the *full model* and RSS_0 is the residual sum of squares from the *reduced model*

Linear Models

Example (3.2.1 from Hastie et al. (2009))

This example examines the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (`lcavol`), log prostate weight (`lweight`), `age`, log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`), and percent of Gleason scores 4 or 5 (`pgg45`).

Linear Models

Example (3.2.1 Continued)

```
> prostate <- read.csv(file =
"https://web.stanford.edu/˜hastie/ElemStatLearn/datasets/prostate.data",sep="\t")
> prostate <- prostate[,-1]
> fit1 <-lm(lpsa˜lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=prostate,
```

Linear Models

Example (3.2.1 Continued)

```
> summary(fit1)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.42917    1.55359    0.28   0.7833
lcavol       0.57654    0.10744    5.37  1.5e-06 ***
lweight      0.61402    0.22322    2.75   0.0079 **
age         -0.01900    0.01361   -1.40   0.1681
lbph         0.14485    0.07046    2.06   0.0443 *
svi          0.73721    0.29856    2.47   0.0165 *
lcp         -0.20632    0.11052   -1.87   0.0670 .
gleason     -0.02950    0.20114   -0.15   0.8839
pgg45        0.00947    0.00545    1.74   0.0875 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.712 on 58 degrees of freedom
Multiple R-squared:  0.694,        Adjusted R-squared:  0.652
F-statistic: 16.5 on 8 and 58 DF,  p-value: 2.04e-12
```

Linear Models

```
> fit2<-lm(lpsa~lcavol+lweight+lbph+svi,data=prostate,subset=train)

> anova(fit1, fit2, test="F")
Analysis of Variance Table

Model 1: lpsa ~ lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45
Model 2: lpsa ~ lcavol + lweight + lbph + svi
Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     58 29.43
2     62 32.81 -4    -3.389 1.67  0.169
```

Linear Models

The summary from our full model shows that **`age, lcp, gleason, pp45`** are not significant at $\alpha = 0.05$

We fit another model and leave these variables out and call it `fit2`

We calculate the partial F statistic and find that the difference between the full model and the reduced model is not significant, so the reduced model is appropriate

We don't need to do this by hand, instead we use the `anova()` function in R, pass it the two models, and designate the test as F

Linear Models

Subset Selection

Two key reasons for improving the linear regression model estimates:

1. Prediction accuracy – the least squares estimates often have low bias but large variance. Prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero. We sacrifice a little bit of bias to reduce the variance of the predicted values, which may improve the overall prediction accuracy
2. Interpretation – With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects. In order to get the “big picture,” we’re willing to sacrifice some of the small details.

Linear Models

Best Subset For each $k \in \{0, 1, \ldots, p\}$ find the subset size k that gives the smallest RSS

Forward-Stepwise Start with the intercept then add significant variables one at a time

Backward-Stepwise Start with the full model then remove insignificant variables one at a time

Other Methods

Ridge Regression – shrinks the regression coefficients by imposing a penalty on their size. Shrinks β's closer to zero and to each other proportionately.

The Lasso – is a shrinkage method like ridge, with subtle but important differences. Lasso translates each coefficient by a constant factor λ, truncating at zero.

Linear Models

So far we have been talking about **multiple linear regression**
When $p = 1$ we have the **simple linear regression** model

$$Y = X\beta + \epsilon$$

The estimates and residuals are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

and

$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1$$

and

$$r_i = y_i - x_i\hat{\beta}$$

Linear Models

Model Selection

Let's look a little deeper into model selection measures and methods

- Mallow's C_p

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

where d is the number of predictors, $\hat{\sigma}^2$ is an estimate of $\mathrm{Var}\,(\epsilon)$

1. If $\hat{\sigma}^2$ is an unbiased estimate of σ^2 then C_p is an unbiased estimate of the test MSE
2. Select the model with the lowest C_p

Linear Models

- Akaike Information Criterion (AIC)

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

1. AIC and C_p are proportional to each other
2. Select the model with the lowest AIC

- Bayesian Information Criterion (BIC)

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

1. BIC usually selects a smaller model than C_p
2. Select the model with the lowest BIC

Linear Models

- Adjusted R^2

$$Adj.\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

1. Select the model with the largest R^2

All of these measures penalize for adding “noise” terms to the model – that is, terms that don’t reduce the RSS

Linear Models

Best Subset Algorithm

1. Let $\mathcal{M}_0$ denote the *null* model, with no predictors, which just predicts the mean for each observation.
2. For $k = 1, 2, 3, \ldots, p$:
   - (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   - (b) Pick the best among the $\binom{p}{k}$models and call it $\mathcal{M}_k$. The best in this case has the smallest RSS or largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p,\ AIC,\ BIC,$ or, R^2.

Linear Models

Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   - (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   - (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p,\ AIC,\ BIC,$ or, R^2.

Linear Models

Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$ for a total of $k-1$ predictors.
   (b) Choose the best among these k models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p,\ AIC,\ BIC,$ or, R^2.

Linear Models

Two other model assessment measures are the **Validation Set Approach** and the **Leave-One-Out Cross Validation** (LOOCV)

Validation Set Approach Split the data equally into a training set and a validation set. Fit a model on the training set and predict the outcomes on the validation set, assessing the MSE, as an estimate of the test error rate

Linear Models

LOOCV Take one observation out (x_1, y_1) for the validation set and fit the model on the remaining $n-1$ observations, and make a prediction on the one validation observation, and find $MSE_1 = (y_1 - \hat{y}_1)^2$. Repeat holding out (x_2, y_2), then (x_3, y_3) and so on. The LOOCV estimate for the test MSE is the average of the n test error estimates

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Linear Models

The LOOCV approach has two advantages over the validation set approach

- The LOOCV approach uses almost the entire data set ($n - 1$ observations) at each iteration so there's less bias
- The validation set approach has greater randomness in the training/validation split and will yield different results when applied repeatedly

Linear Models

Logistic Regression

- In our discussion of linear regression models, the response variable is some real valued variable
- Some problems are such that the response is not a real valued outcome but instead a binary result
- For example:
  1. On/Off
  2. Yes/No
  3. Default/No Default
  4. Disease/No Disease, and so ...
- In cases such as these we often make use of **logistic regression**

Linear Models

Logistic Regression models the probability of an observation belonging to one of k classes as a linear function of the X

The probabilities for the k classes sum to one, and are in $[0, 1]$

The model has the general form

$$\log \frac{P\left[Y = 1|X = x\right]}{P\left[Y = K|X = x\right]} = \hat{\beta}_{10} + \hat{\beta}_1' x \tag{6.3}$$

$$\log \frac{P\left[Y = 2|X = x\right]}{P\left[Y = K|X = x\right]} = \hat{\beta}_{20} + \hat{\beta}_2' x \tag{6.4}$$

$$\vdots \qquad \vdots \qquad \vdots \tag{6.5}$$

$$\log \frac{P\left[Y = K-1|X = x\right]}{P\left[Y = K|X = x\right]} = \hat{\beta}_{(K-1)0} + \hat{\beta}_{K-1}' x \tag{6.6}$$

$$\tag{6.7}$$

Linear Models

The model in (6.7) can be re-written as

$$P\left[Y = k|X = x\right] = \frac{\exp[\hat{\beta}_{k0} + \hat{\beta}_k' x]}{1 + \sum_{j=1}^{K-1} \exp[\hat{\beta}_{j0} + \hat{\beta}_j' x]}$$

for $k = 1, 2, \ldots, K - 1$ and

$$P\left[Y = K|X = x\right] = \frac{1}{1 + \sum_{j=1}^{K-1} \exp[\hat{\beta}_{j0} + \hat{\beta}_j' x]}$$

A very common case is when $k = 2$ and we have a binary outcome, (6.7) becomes

$$P\left[Y = k|X = x\right] = \frac{\exp[\hat{\beta}_0 + \hat{\beta}_1' x]}{1 + \exp[\hat{\beta}_0 + \hat{\beta}_1' x]}$$

Linear Models

The estimates for the coefficients are found via maximum likelihood estimation

Linear Models

The ratio of the probabilities in (6.7) (without the $\log$) are called the *odds* – that is, the probability of being in one class vs. another for example

Taking the [natural] $\log$ gives us the *log-odds* aka the *logit*

Finally, the ratio of the conditional probability of being in category 1 to category 0 is the *relative risk*

Linear Models

For example in the binary case...

$$\text{ODDS} = \frac{P\left[Y = 1|X = x\right]}{1 - P\left[Y = 1|X = x\right]}$$

$$\text{RR} = \frac{P\left[Y = 1|X = 1\right]}{P\left[Y = 1|X = 0\right]}$$

$$\text{OR} = \frac{\text{ODDS P}\left[\text{Y} = 1|\text{X} = \text{x}\right]}{\text{ODDS P}\left[\text{Y} = 0|\text{X} = \text{x}\right]}$$

References

Hastie, T., Tibshirani, R. and Friedman, J. (2009).
The Elements of Statistical Learning.
Springer, New York, second edition.
12th printing 2017.

Hogg, R. V., McKean, J. W. and Craig, A. T. (2018).
Introduction to Mathematical Statistics.
Pearson, Boston, eighth edition.

Ross, S. (2010).
A First Course in Probability.
Prentice Hall, Upper Saddle River, eigth edition.

Wu, J. and Coggeshall, S. (2012).
Foundations of Predictive Analytics.
CRC Press, Boca Raton, FL.