Extensions of Linear Regression

26th February 2023

Multivariate Linear Regression **

- In last slides, we considered the multiple linear regression, where the predictor is a p-dimensional vector and the response is a univariate random variable.
- Now we consider a slightly complex case where the response Y is a q-dimensional vector.
- The Multivariate (multiple) linear regression assumes that

$$Y = B^T X + E,$$

where $B \in \mathbb{R}^{p\times q}$ is the regression coefficient, $E \in \mathbb{R}^q$ is the error term, and E is uncorrelated (independent) with X.

  - In the model, we omit the intercept since we may let the first elements of X be 1.
- B captures the linear relationship between Y and X.

Is the covariance of E useful?

- Consider n independent samples $\{(Y_i, X_i)\}_{i=1}^n$.
- We further assume that $E \sim N(0, \Sigma)$.
- Question: Will the MLEs of B be different for different Σ? Namely, can the covariance information help to improve the estimation?
- Unfortunately, the MLEs are the same for all Σ. Specifically, $\hat{B}^{MLE} = (\sum_{i=1}^n X_i X_i^T)^{-1}(\sum_{i=1}^n X_i Y_i^T)$.
- Let $\mathbb{X} = \{X_1, \cdots, X_n\}^T \in \mathbb{R}^{n \times p}$ and $\mathbb{Y} = \{Y_1, \cdots, Y_n\}^T \in \mathbb{R}^{n \times q}$ be the stacked sample matrices.
- $\hat{B}_{MLE} = \hat{B}_{OLS} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$.

Is the covariance of E useful?

- It is disappointing that the covariance does not improve the MLE.
- Or, we may say that considering all the responses together is equivalent to considering them separately. (In terms of MLE).
- Methods for considering the responses together:
  - Reduced rank regression
  - Sparse methods
  - Envelope method
  - ...

Reduced Rank Regression

Reduced Rank Regression

- Reduced-rank regression (RRR) is a variant of multiple multivariate regression with an added constraint.
- RRR enforces that $\mathrm{rank}(B) = r$, where $r < \min(p, q)$.
- Intuitively, this constraint enforces the assumption that X and Y are related through a small number of latent factors.
- Free parameters: $pq \rightarrow (p + q - r)r$.

Estimation of RRR

- RRR attempts to solve the following optimization problem:

$$\text{argmin}_B \|\mathbb{Y} - \mathbb{X}B\|_F^2,$$

  where $\|\dot{\|}_F$ is the Frobenius norm.
- Since the rank of B is r, we have

$$B = AC^T$$

  where $A \in \mathbb{R}^{p \times r}$ and $C \in \mathbb{R}^{q \times r}$.
- Notice that this problem is not identifiable. If we consider any nonsingular matrix $M \in \mathbb{R}^{r \times r}$, and set $A' = AM^{-1}$ and $C' = CM^T$, then

$$B = A'C'^T = AM^{-1}(CM^T)^T = AM^{-1}MC^T = AC^T.$$

Estimation of RRR

- The objective function of RRR can be equivalently written as (why?)

$$\operatorname{argmin}_B \left\| \mathbb{Y} - \mathbb{X}\hat{B}_{\mathrm{OLS}} \right\|_F^2 + \left\| \mathbb{X}\hat{B}_{\mathrm{OLS}} - \mathbb{X}B \right\|_F^2$$

.

- Hence,

$$\hat{B}_{\mathrm{RRR}} = \operatorname{argmin}_B \left\| \mathbb{X}\hat{B}_{\mathrm{OLS}} - \mathbb{X}B \right\|_F^2 .$$

- Notice that this is minimized by performing an SVD on $\mathbb{Y}_{\mathrm{OLS}} = \mathbb{X}\hat{B}_{\mathrm{OLS}}$ (why?)
- Specifically $\mathbb{X}\hat{B}_{\mathrm{OLS}} = UDV^T$ and let V_r be matrix stacked by the first r columns of V. Then

$$\hat{B}_{\mathrm{RRR}} = \hat{B}_{\mathrm{OLS}} V_r V_r^T \quad \text{(why?)}$$

.

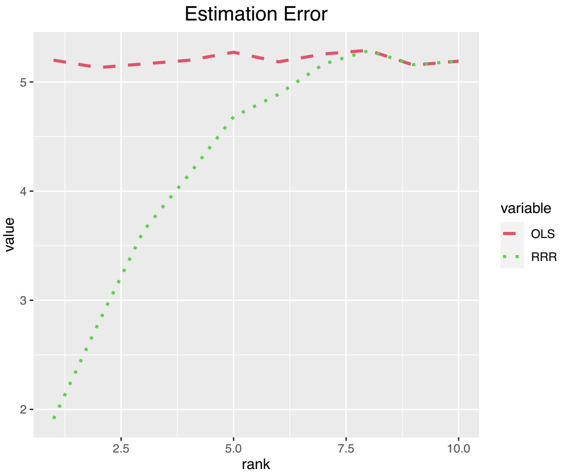
Estimation of RRR

- We first state the following conclusion (Matrix approximation lemma): Suppose that $A = UDV^T$, where $D = \text{diag}(d_1, \cdots, d_s, 0, \cdots, 0)$, and $r \leq s$. Then the solution of
$$\text{argmin}_{\text{rank}(X) \leq r} \|A - X\|_F$$
is UD_rV^T, where $D = \text{diag}(d_1, \cdots, d_r, 0, \cdots, 0)$.
- Let $\mathbb{Y}_{\text{OLS}} = \sum_{i=1}^s d_i u_i v_i^T$. The best rank-$r$ approximation of $\mathbb{Y}_{\text{OLS}}$ is $\sum_{i=1}^r d_i u_i v_i^T$. Define $P_r = \sum_{i=1}^r v_i v_i^T$ and $\hat{B}_{\text{RRR}} = \hat{B}_{\text{OLS}} P_r$. Then $\mathbb{X}\hat{B}_{\text{RRR}} = \mathbb{X}\hat{B}_{\text{OLS}} P_r = \left(\sum_{i=1}^s d_i u_i v_i^T\right) \sum_{i=1}^r v_i v_i^T = \sum_{i=1}^r d_i u_i v_i^T$. Hence, $\hat{B}_{\text{RRR}}$ is the minimizer of $\left\|\mathbb{X}\hat{B}_{\text{OLS}} - \mathbb{X}B\right\|_F^2$.

A Simulation Example

- We generate a data set from the multivariate linear regression model.
- $p = q = 10$, r takes value in $\{1, 2, \cdots, 10\}$.
- Each element of X_i is generated from $U(0, 1)$ and E_i is generated from standard normal distribution independently for $i = 1, \cdots, n$.
- For each rank, we generate 100 replicates.
- We report the estimation error $\|\hat{B} - B\|_F$.

A Simulation Example



Application in Chemometrics Example

- There are $n = 56$ observations with $p = 22$ and $q = 6$. The data is generated from a simulation of a low density tubular polyethylene reactor.
- The predictor variables consists of 20 temperature measurements at equal distance along the reactor along with the wall temperature and the feed rate.
- The responses are output characteristics of the polymers produced, namely, Number avg. molecular weight. (Y_1), Weight avg. molecular weight (Y_2), Long chain branching (Y_3), Short chain branching (Y_4), content of vinyl group (Y_5) and content of vinyledene $\mathrm{group}\,(Y_6)$.
- As the responses were all right skewed we applied log transformation, and finally standardized them.

Application in Chemometrics Example

- Consider the leave-one-out prediction error.

	OLS	RRR	RRR+Ridge
Y_1	0.49	0.44	0.15
Y_2	1.12	0.46	0.22
Y_3	0.53	0.65	0.39
Y_4	0.24	0.14	0.24
Y_5	0.30	0.18	0.27
Y_6	0.28	0.16	0.27
Avg	0.50	0.34	0.26

- Performance comparison for the chemometrics data

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) $\cancel{\times}\cancel{\times}$

Motivation:

- Recall that the goal of the multivariate linear regression is capturing the linear relationship between $\mathbf{x}$ and $\mathbf{y}$.
- Is there other ways to maximize the "linear relationship" between $\mathbf{x}$ and $\mathbf{y}$.
- We may consider the correlation between them.
- Find two directions $\mathbf{a}$ and $\mathbf{b}$ such that $\mathrm{Cor}(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{y})$ attains its maximum.

Canonical Correlation Analysis (CCA)

- Canonical correlation analysis (CCA) is a classical method to analyze the relationship between two multivariate measurements.
- Consider random vectors $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$.
- Define $\Sigma_{yx} = \mathrm{cov}(y, x), \Sigma_{xx} = \mathrm{cov}(x)$ and $\Sigma_{yy} = \mathrm{cov}(y)$.
- For a positive integer $k < \min\{p, q\}$, CCA finds canonical directions $\{\mathbf{a}_i, \mathbf{b}_i\}_{i=1}^k$ that sequentially maximize the correlation between $\mathbf{a}_i^T\mathbf{x}$ and $\mathbf{b}_i^T\mathbf{y}$.
- Let $S_{\mathbf{yx}}, S_{\mathbf{xx}}$ and $S_{\mathbf{yy}}$ be the sample estimates of Σ_{yx}, Σ_{xx} and Σ_{yy}.

The first pair of canonical variables

- We want to find the linear combination of the X-variables and the linear combination of the Y-variables which is most highly correlated.
- Find a and b which maximize

$$\mathrm{Cor}\left(\mathbf{a}^\top \mathbf{x}, \mathbf{b}^\top \mathbf{y}\right) = \frac{\mathbf{a}^\top \mathbf{S_{xy}} \mathbf{b}}{\left(\mathbf{a}^\top \mathbf{S_{xx}} \mathbf{a}\right)^{1/2} \left(\mathbf{b}^\top \mathbf{S_{yy}} \mathbf{b}\right)^{1/2}}$$

- In other words: Maximise $\mathrm{Cor}\left(\mathbf{a}^\top \mathbf{x}, \mathbf{b}^\top \mathbf{y}\right)$ for non-zero vectors $\mathbf{a}(p \times 1)$ and $\mathbf{b}(q \times 1)$.
- Intuitively, this objective makes sense, because we want to find the linear combination of the $\mathbf{x}$-variables and the linear combination of the $\mathbf{y}$-variables which are most highly correlated.

The first pair of canonical variables

- However, note that for any $\gamma > 0$ and $\delta > 0$,

$$\begin{aligned}\mathrm{Cor}\left(\gamma \mathbf{a}^\top \mathbf{x}, \delta \mathbf{b}^\top \mathbf{y}\right) &= \frac{\gamma\delta}{\sqrt{\gamma^2\delta^2}} \mathrm{Cor}\left(\mathbf{a}^\top \mathbf{x}, \mathbf{b}^\top \mathbf{y}\right) \\ &= \mathrm{Cor}\left(\mathbf{a}^\top \mathbf{x}, \mathbf{b}^\top \mathbf{y}\right)\end{aligned}$$

- There will be an infinite number of solutions to this optimization problem, because if $\mathbf{a}$ and $\mathbf{b}$ are solutions, then so are $\gamma\mathbf{a}$ and $\delta\mathbf{b}$, for any $\gamma > 0$ and $\delta > 0$.
- A more useful way to formulate this optimisation problem is

$$\begin{array}{ll}\text{Maximize} & \mathbf{a}^\top \mathbf{S}_{\mathbf{xy}} \mathbf{b} \\ \text{subject to} & \mathbf{a}^\top \mathbf{S}_{\mathbf{xx}} \mathbf{a} = \mathbf{b}^\top \mathbf{S}_{\mathbf{yy}} \mathbf{b} = 1.\end{array}$$

The first pair of canonical variables

- Assume that $\mathbf{S_{xx}}$ and $\mathbf{S_{yy}}$ are both non-singular, and consider the singular value decomposition of the matrix $\mathbf{Q} := \mathbf{S_{xx}}^{-1/2}\mathbf{S_{xy}}\mathbf{S_{yy}}^{-1/2}$

$$\mathbf{Q} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top = \sum_{j=1}^{t} \sigma_j \mathbf{u}_j \mathbf{v}_j^\top$$

where $t = \operatorname{rank}(\mathbf{Q})$ and $\sigma_1 \geq \cdots \geq \sigma_t > 0$. Then the solution to the constrained optimization problem is

$$\mathbf{a} = \mathbf{S_{xx}}^{-1/2}\mathbf{u}_1 \quad \text{and} \quad \mathbf{b} = \mathbf{S_{yy}}^{-1/2}\mathbf{v}_1.$$

The maximum value of the correlation coefficient is given by the largest singular value σ_1 :

$$\max_{\mathbf{a},\mathbf{b}} \mathbb{C}\mathrm{or}\left(\mathbf{a}^\top\mathbf{x}, \mathbf{b}^\top\mathbf{b}\right) = \sigma_1$$

The first pair of canonical variables

Proof: If we let

$$\tilde{\mathbf{a}} = \mathbf{S}_{\mathrm{xx}}^{1/2}\mathbf{a} \quad \text{and} \quad \tilde{\mathbf{b}} = \mathbf{S}_{\mathrm{yy}}^{1/2}\mathbf{b}$$

we may write the constraints $\mathbf{a}^\top \mathbf{S}_{\mathbf{xx}}\mathbf{a} = \mathbf{b}^\top \mathbf{S}_{\mathbf{yy}}\mathbf{b} = 1$ as

$$\tilde{\mathbf{a}}^\top \tilde{\mathbf{a}} = 1 \quad \text{and} \quad \tilde{\mathbf{b}}^\top \tilde{\mathbf{b}} = 1.$$

If we write

$$\mathbf{a} = \mathbf{S}_{\mathbf{xx}}^{-1/2}\tilde{\mathbf{a}} \quad \text{and} \quad \mathbf{b} = \mathbf{S}_{\mathbf{yy}}^{-1/2}\tilde{\mathbf{b}}$$

then the optimization becomes

$$\max_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}} \tilde{\mathbf{a}}^\top \mathbf{S}_{\mathbf{xx}}^{-1/2}\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yy}}^{-1/2}\tilde{\mathbf{b}}$$

subject to

$$\|\tilde{\mathbf{a}}\| = 1 \quad \text{and} \quad \|\tilde{\mathbf{b}}\| = 1.$$

The first pair of canonical variables

Then we can see that

$$\tilde{\mathbf{a}} = \mathbf{u}_1 \quad \text{and} \quad \tilde{\mathbf{b}} = \mathbf{v}_1$$

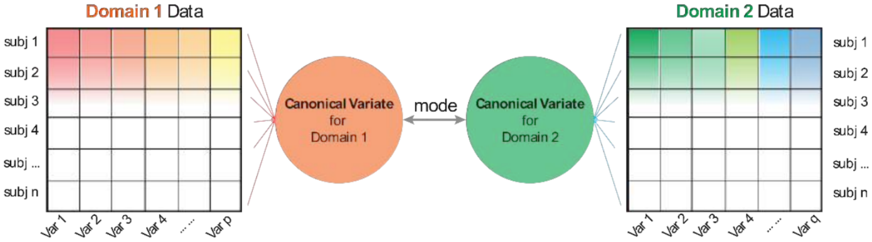
and the result follows.

- We will label the solution found as

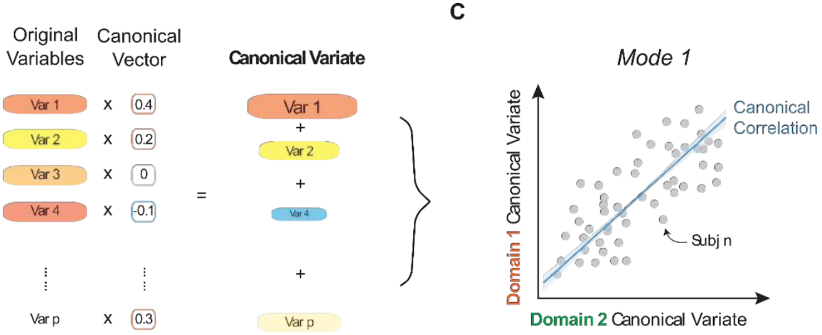
$$\mathbf{a}_1 := \mathbf{S}_{xx}^{-\frac{1}{2}}\mathbf{u}_1 \quad \text{and} \quad \mathbf{b}_1 := \mathbf{S}_{yy}^{-\frac{1}{2}}\mathbf{v}_1$$

to stress that $\mathbf{a}_1$ and $\mathbf{b}_1$ are the first pair of canonical correlation (CC) vectors. The variables $\eta_1 = \mathbf{a}_1^\top(\mathbf{x} - \overline{\mathbf{x}})$ and $\psi_1 = \mathbf{b}_1^\top(\mathbf{y} - \overline{\mathbf{y}})$ are called the first pair of canonical correlation variables, and $\sigma_1 = \mathrm{Cor}\left(\eta_1, \psi_1\right)$ is the first canonical correlation.

CCA Illustration



CCA Illustration



The full sets of canonical variables

- We now repeat this process to find the next most important linear combination, subject to being uncorrelated with the first linear combination.
- For $\mathbf{a}^\top \mathbf{x}$ to be uncorrelated with $\eta_1 = \mathbf{a}_1^\top \mathbf{x}$ we require

$$0 = \mathrm{Cov}\left(\mathbf{a}_1^\top \mathbf{x}, \mathbf{a}^\top \mathbf{x}\right) = \mathbf{a}_1^\top \mathbf{S}_{xx} \mathbf{a},$$

  and similarly we require the condition $\mathbf{b}_1^\top \mathbf{S}_{yy} \mathbf{b} = 0$ for $\mathbf{b}$.
- Thus, we need to solve the following optimization problem:

$$\max_{\mathbf{a},\mathbf{b}} \mathbf{a}^\top \mathbf{S}_{\mathbf{xy}} \mathbf{b}$$

  subject to the constraints

$$\mathbf{a}^\top \mathbf{S}_{\mathbf{xx}} \mathbf{a} = \mathbf{b}^\top \mathbf{S}_{\mathbf{yy}} \mathbf{b} = 1,$$
$$\mathbf{a}_1^\top \mathbf{S}_{\mathbf{xx}} \mathbf{a} = \mathbf{b}_1^\top \mathbf{S}_{\mathbf{yy}} \mathbf{b} = 0.$$

The full sets of canonical variables

Proposition:

- For $k = 1, \ldots, r = \operatorname{rank}(\mathbf{S}_{xy})$, the solution to sequence of optimization problems

$$
\begin{aligned}
&\text{Maximize } \mathbf{a}^\top \mathbf{S}_{xy} \mathbf{b} \\
&\text{subject to } \mathbf{a}^\top \mathbf{S}_{xx} \mathbf{a} = \mathbf{b}^\top \mathbf{S}_{yy} \mathbf{b} = 1 \\
&\qquad\qquad \text{and } \mathbf{a}_i^\top \mathbf{S}_{xx} \mathbf{a} = \mathbf{b}_i^\top \mathbf{S}_{yy} \mathbf{b} = 0 \text{ for } i = 1, \ldots, k-1
\end{aligned}
$$

is achieved at $\mathbf{a}_k = \mathbf{S}_{xx}^{-1/2} \mathbf{u}_k$ and $\mathbf{b}_k = \mathbf{S}_{yy}^{-1/2} \mathbf{v}_k$ with $\mathbf{a}_k \mathbf{S}_{xy} \mathbf{b}_k = \sigma_k$.

CCA Example

Team	W	D	L	G	GA	GD
Liverpool	32	3	3	85	33	52
Manchester City	26	3	9	102	35	67
Manchester United	18	12	8	66	36	30
Chelsea	20	6	12	69	54	15
Leicester City	18	8	12	67	41	26

- We shall treat W and D, the number of wins and draws, as the $\mathbf{x}$-variables. The number of goals for and against, G and GA, will be treated as the $\mathbf{y}$-variables.
- We shall consider the questions:
  - how strongly associated are the match outcome variables, W and D, with the goals for and against variables, G and GA ?
  - what linear combination of W and D, and of G and GA are most strongly correlated?

CCA Example

$$\mathbf{S}_{xx} = \begin{pmatrix} 40.4 & -9.66 \\ -9.66 & 10.7 \end{pmatrix}, \quad \mathbf{S}_{yy} = \begin{pmatrix} 354 & -155 \\ -155 & 141 \end{pmatrix},$$

$$\mathbf{S}_{xy} = \mathbf{S}_{yx}^{\top} = \begin{pmatrix} 108 & -60 \\ -28.9 & -2.36 \end{pmatrix}.$$

$$\mathbf{S}_{xx} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\top} = \begin{pmatrix} -0.959 & -0.285 \\ 0.285 & -0.959 \end{pmatrix} \begin{pmatrix} 43.2 & 0 \\ 0 & 7.82 \end{pmatrix} \begin{pmatrix} -0.959 & 0.285 \\ -0.285 & -0.959 \end{pmatrix}$$

$$\begin{aligned} \mathbf{S}_{\mathbf{xx}}^{-1/2} &= \mathbf{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^{\top} \\ &= \begin{pmatrix} -0.959 & -0.285 \\ 0.285 & -0.959 \end{pmatrix} \begin{pmatrix} 0.152 & 0 \\ 0 & 0.357 \end{pmatrix} \begin{pmatrix} -0.959 & 0.285 \\ -0.285 & -0.959 \end{pmatrix} \\ &= \begin{pmatrix} 0.169 & 0.0561 \\ 0.0561 & 0.341 \end{pmatrix}. \end{aligned}$$

CCA Example

$$\mathbf{a}_1 = \mathbf{S}_{\mathbf{xx}}^{-1/2}\mathbf{u}_1 = \begin{pmatrix} 0.169 & 0.0561 \\ 0.0561 & 0.341 \end{pmatrix} \begin{pmatrix} -0.99 \\ -0.143 \end{pmatrix} = \begin{pmatrix} -0.175 \\ -0.104 \end{pmatrix}$$

$$\mathbf{b}_1 = \mathbf{S}_{\mathbf{yy}}^{-1/2}\mathbf{v}_1 = \begin{pmatrix} -0.0234 \\ 0.0541 \end{pmatrix}$$

This leads to the first pair of CC variables, obtained using these CC vectors/weights:

$$\eta_1 = -0.175(W - \bar{W}) + -0.104(D - \bar{D})$$

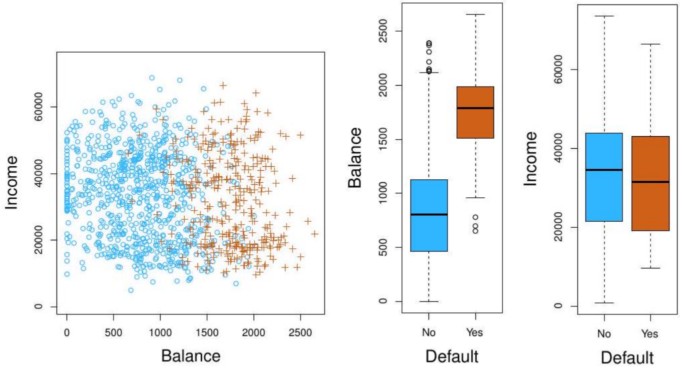
$$\psi_1 = -0.0234(G - \bar{G}) + 0.0541(GA - \overline{GA}).$$

We can see that ψ_1 is measuring something similar to goal difference $G - GA$, as usually defined, but it gives higher weight to goals conceded than goals scored (0.0541 versus 0.0234).

Logistic Regression

Qualitative variables

- Recall that in linear regression model, our response Y is usually a continuous random variable.
- What if Y is categorical, namely it takes values in a finite set $\mathcal{C}$.
- Linear regression is not appropriate for this scenario.
- This scenario is usually described as classification task: build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the probabilities that X belongs to each category in $\mathcal{C}$.
  - For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.



- The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.

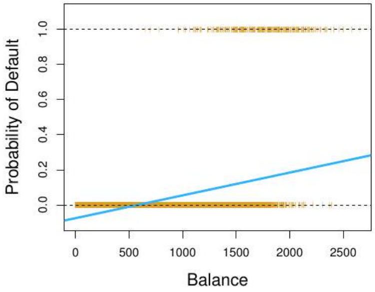
Can we use Linear Regression?

- Suppose for the Default classification task that we code

$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes.} \end{cases}$$

- Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?
- linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

Linear versus Logistic Regression



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1 \mid X)$ well. Logistic regression seems well suited to the task.

Linear versus Logistic Regression

- Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

- This coding suggests an ordering, and in fact implies that the difference between stroke and drug overdose is the same as between drug overdose and epileptic seizure.
- Linear regression is not appropriate here.
- Multiclass Logistic Regression is more appropriate.

Logistic Regression

Let's write $p(X) = \Pr(Y = 1 \mid X)$ for short and consider using balance to predict default. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

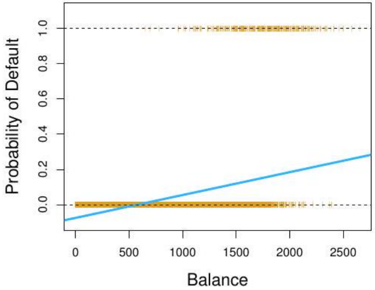
$e \approx 2.71828$ is a mathematical constant [Euler's number.]) It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1 .

A bit of rearrangement gives

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

This monotone transformation is called the log odds or logit transformation of $p(X)$. (by log we mean natural log: ln.)

Linear versus Logistic Regression



Logistic regression ensures that our estimate for $p(X)$ lies between 0 and 1.

Maximum Likelihood

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This likelihood gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.
Most statistical packages can fit linear logistic regression models by maximum likelihood. In R we use the glm function.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Making Predictions

What is our estimated probability of default for someone with a balance of $\$1000$?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

With a balance of $\$2000$?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Lets do it again, using student as the predictor.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student [Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\text{ default } = \text{ Yes } \mid \text{ student } = \text{ Yes }) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.043$$

$$\widehat{\Pr}(\text{ default } = \text{ Yes } \mid \text{ student } = \mathrm{No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Logistic regression with several variables

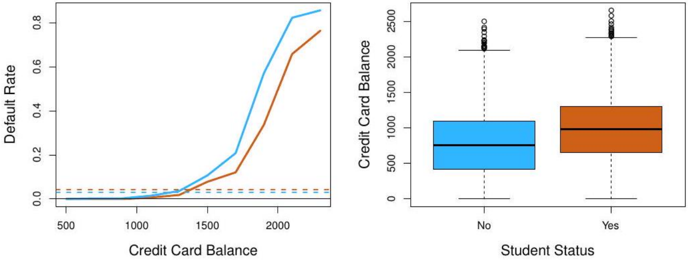
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Why is coefficient for student negative, while it was positive before?

Confounding



- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

Logistic regression with more than two classes

- So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package glmnet) has the symmetric form

$$\Pr(Y = k \mid X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

- Here there is a linear function for each class. (The mathier students will recognize that some cancellation is possible, and only $K-1$ linear functions are needed as in 2-class logistic regression.)
- Multiclass logistic regression is also referred to as multinomial regression.

Fitting Logistic Regression Models

- Recall that logistic regression models can be fitted by maximum likelihood, using the conditional likelihood of Y given X.
- The log-likelihood for N observations is

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{y_i}(x_i; \theta),$$

where $p_k(x_i; \theta) = \Pr(Y = k \mid X = x_i; \theta)$.
- We discuss in detail the two-class case.
- Note that $p_1(x_i; \theta) = 1 - p_2(x_i; \theta)$.
- We denote $p_1(x_i; \theta) = p(x_i; \theta)$ for short. By definition $p(x_i; \theta) = \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)}$. We set the first element of x_i to be 1, which makes the intercept term disappear.

Fitting Logistic Regression Models

- The log-likelihood can be written as

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{N} \{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\} \\ &= \sum_{i=1}^{N} \left\{ y_i \beta^T x_i - \log\left(1+e^{\beta^T x_i}\right)\right\}\end{aligned}$$

- We consider the Newton-Raphson algorithm to solve the MLE. We have

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i;\beta)) = 0,$$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N} x_i {x_i}^T p(x_i;\beta)(1-p(x_i;\beta)).$$

Fitting Logistic Regression Models

$\left\{\begin{array}{l}\text{Newton} \\ \text{Stocastic Gradient.} \\ \text{Gradient descent.}\end{array}\right.$

- Starting with β^{old}, a single Newton update is

$$\beta^{\text{new}} = \beta^{\text{old}} - \left(\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial \ell(\beta)}{\partial\beta},$$

where the derivatives are evaluated at β^{old}.

- Let $\mathbf{y}$ denote the vector of y_i values, $\mathbf{X}$ the $N \times (p+1)$ matrix of x_i values, $\mathbf{p}$ the vector of fitted probabilities with i th element $p\left(x_i; \beta^{\text{old}}\right)$ and $\mathbf{W}$ a $N \times N$ diagonal matrix of weights with i th diagonal element $p\left(x_i; \beta^{\text{old}}\right)\left(1 - p\left(x_i; \beta^{\text{old}}\right)\right)$. Then we have

$$\frac{\partial \ell(\beta)}{\partial\beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

$$\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

The Newton step is thus

$$\begin{aligned}
\beta^{\text{new}} &= \beta^{\text{old}} + \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\
&= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})\right) \\
&= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}.
\end{aligned}$$

In the second and third line we have re-expressed the Newton step as a weighted least squares step, with the response

$$\mathbf{z} = \mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p}),$$

sometimes known as the adjusted response. These equations get solved repeatedly, since at each iteration $\mathbf{p}$ changes, and hence so does $\mathbf{W}$ and $\mathbf{z}$. This algorithm is referred to as iteratively reweighted least squares or IRLS, since each iteration solves the weighted least squares problem:

$$\beta^{\text{new}} \leftarrow \arg\min_{\beta}(\mathbf{z}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{z}-\mathbf{X}\beta).$$