HKU BUSINESS SCHOOL
港大經管學院

Business Statistics

Topic Review of MSBA7002

Weichen Wang

Assistant Professor

Innovation and Information Management

Lec 1: Linear Regression

Summary

- Coefficients $b_1 = r \times \frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$.
- Predicted value $\hat{y}_i = b_0 + b_1 x_i$.
- Actual value y_i.
- Residual $e_i = y_i - \hat{y}_i$.
- We choose the line to make SSE or RSS as small as possible.
- Both for linear relationship between two variables.
  - Same sign between b_1 and r.
- r does not depend on which is x and which is y.

RMSE

```
> summary(lm(price~weight))

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

RMSE

R^2

```
> summary(lm(price~weight))

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783     Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

R Square

Model Diagnostics

- Make sure there are no gross violations of the model
  - Is the relationship between x and y *linear*?
  - Do the residuals show iid normal behavior (i.e., *independent, equal variance, normality*)?
  - Are there *outliers* that may distort the model fit?
- Crucial steps in checking a model
  - A *y vs. x scatterplot* should reveal a linear pattern, linear dependence.
  - A *Residual vs. x scatterplot* should reveal no meaningful pattern.
  - A *Residual vs. Predicted scatterplot* should reveal no meaningful pattern.
  - A *histogram and normal quantile plot* of the residuals should be consistent with the assumption of normality of the errors.

Inference

1. Is $\beta_1 = 0$? $\Longrightarrow$ t-statistics
2. Are all $\beta_j = 0$? $\Longrightarrow$ F-statistics
3. 95% Confidence interval of β_j? $\hat{\beta}_j \pm 2 * SE(\hat{\beta}_j)$
4. 95% Confidence interval of $\hat{y}_i$? $\hat{y}_i \pm 2 * RMSE$

Colinearity and Interaction

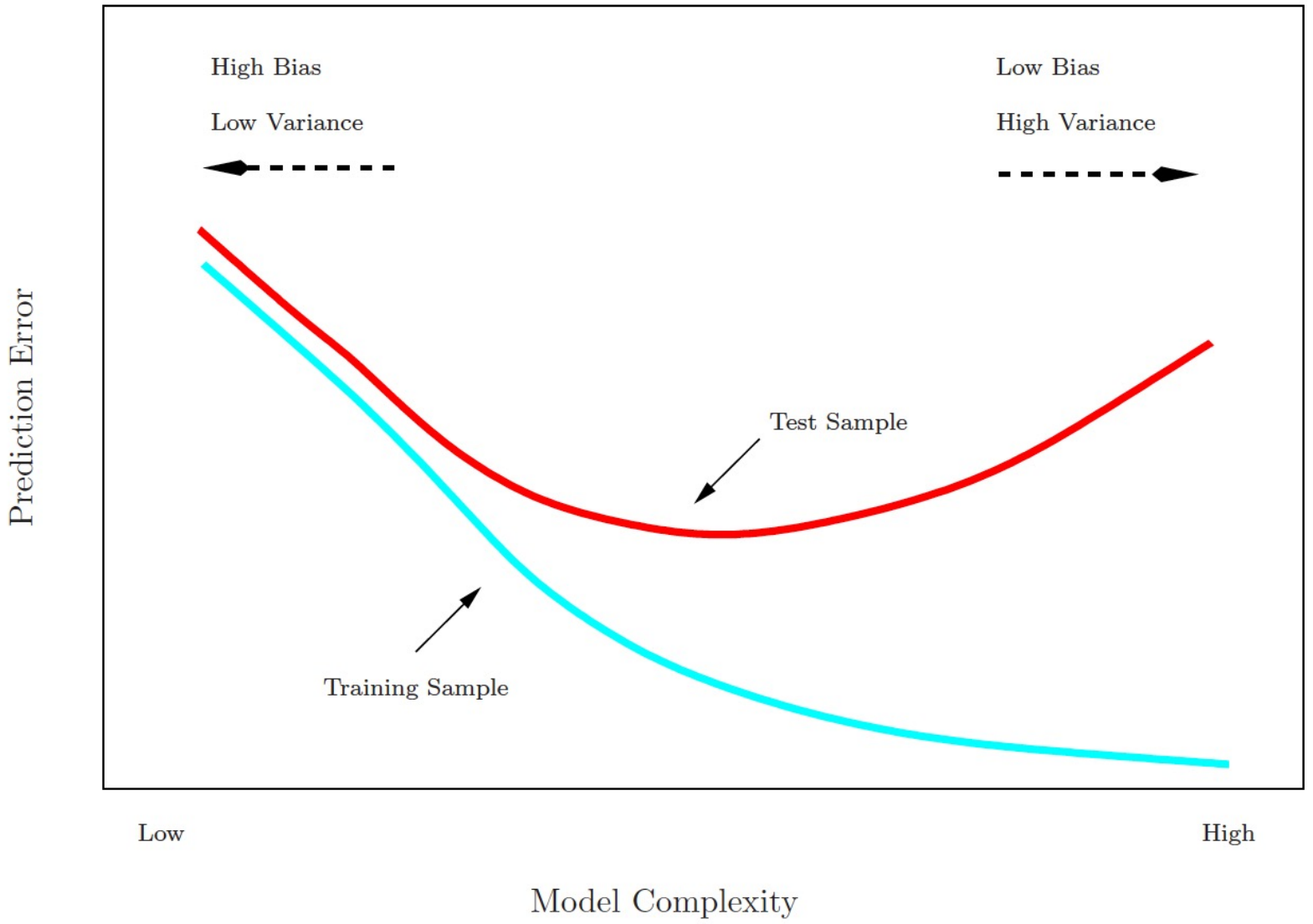
1. What is collinearity? Substantial correlation among predictors.
2. Colinearity violates the linear regression assumption? False.
3. Interaction between a dummy variable A and a continuous variable B affects the slope of B when the dummy is 1 and 0? True.
4. When interaction of A*B is significant, but A is not significant, we should remove A from the model and only keep A*B? False.

Lec 2: Model Selection

Model Selection

- *Subset Selection*. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
- *Shrinkage*. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as *regularization*) has the effect of reducing variance and can also perform variable selection.
- *Dimension Reduction*. We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different *linear combinations*, or *projections*, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

Model Complexity & Prediction Error



Estimating Test Error: Two Approaches

1. Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.

2. Directly estimate the test error, using either a validation set approach or a cross-validation approach.

Model Comparison Criteria

Adjustment to the Training Error

- Adjusted R^2 $$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$
- Mallow's C_p $$C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2)$$
- Bayesian Information Criterion (BIC) $$BIC = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2)$$
- Akaike Information Criterion (AIC) $$AIC = -2\log L + 2 \cdot p$$

Subset Selection vs Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p? True.
- Best subset selection does not suffer from overfitting when p is large? False.
- Both forward and backward selection approach search through only $1 + p(p + 1)/2$ models? True.
- Forward and backward stepwise selection are not guaranteed to yield the best model containing a subset of the p predictors? True.

K-fold CV Illustration

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train

The Computing Details

- Let the K parts be $C_1, C_2, \ldots C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/K$.
- Compute

$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k$$

  where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.
- Setting $K = n$ yields n-fold or *leave-one out cross-validation* (LOOCV).

Lec 3: Shrinkage Methods

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

The Bias-Variance Decomposition

- Assume that

$$Y = f(X) + \varepsilon$$

where E(ε)=0 and Var(ε)=σ_ε^2.

- At an input point $X = x_0$, the expected squared prediction error is

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}.
\end{aligned}$$

- More complex model leads to higher bias but lower variance? False.

The Lasso

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model
- The *Lasso* is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

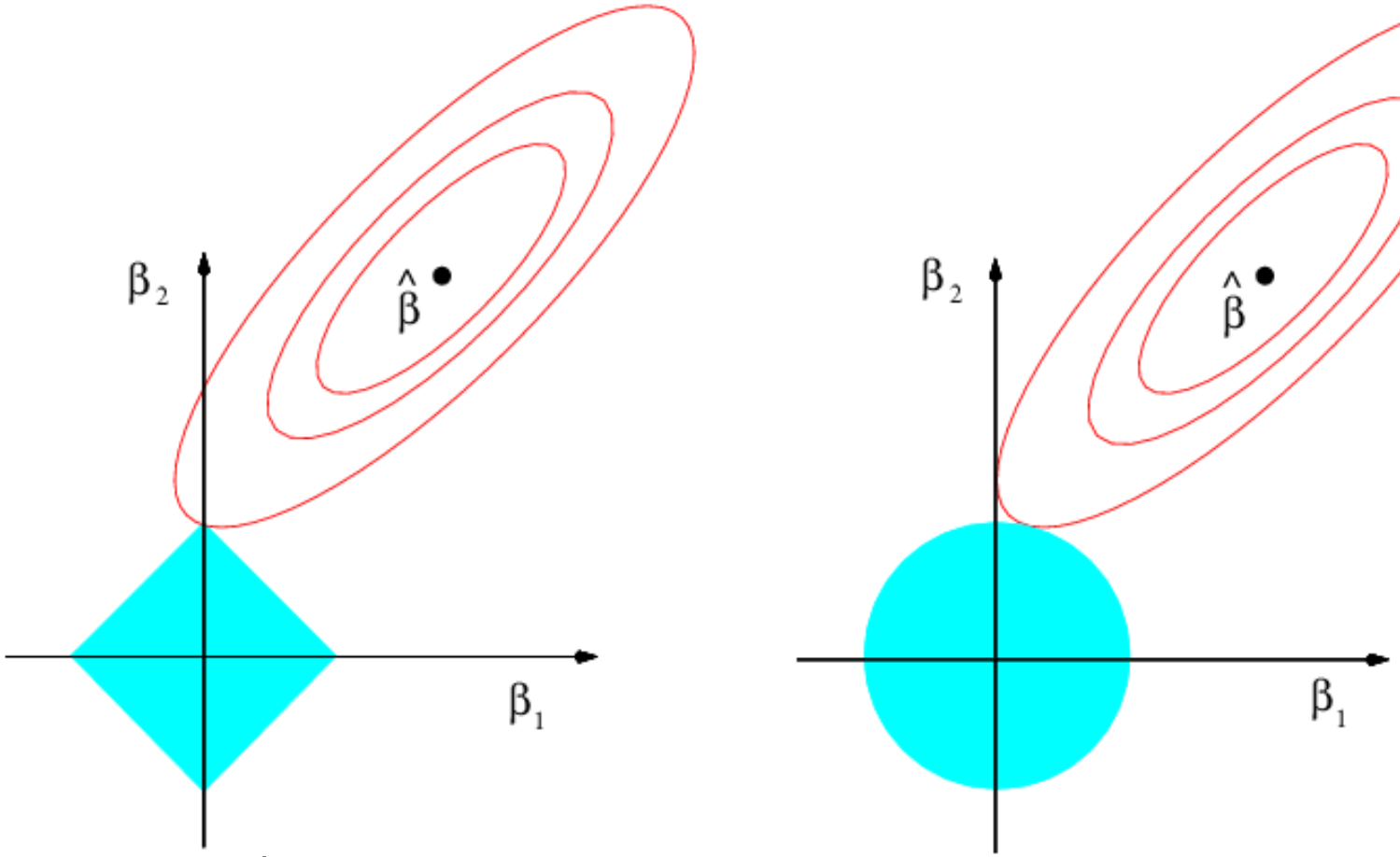
How do we select best λ?

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

CV.

- In statistical parlance, the lasso uses an ℓ_1 (pronounced "ell 1") penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$.

The Geometry of Ridge and Lasso



$\hat{\beta}$: the least square estimate
red ellipse: region of constant RSS (increasing)
blue region: lasso/ridge constraint (driven by s)

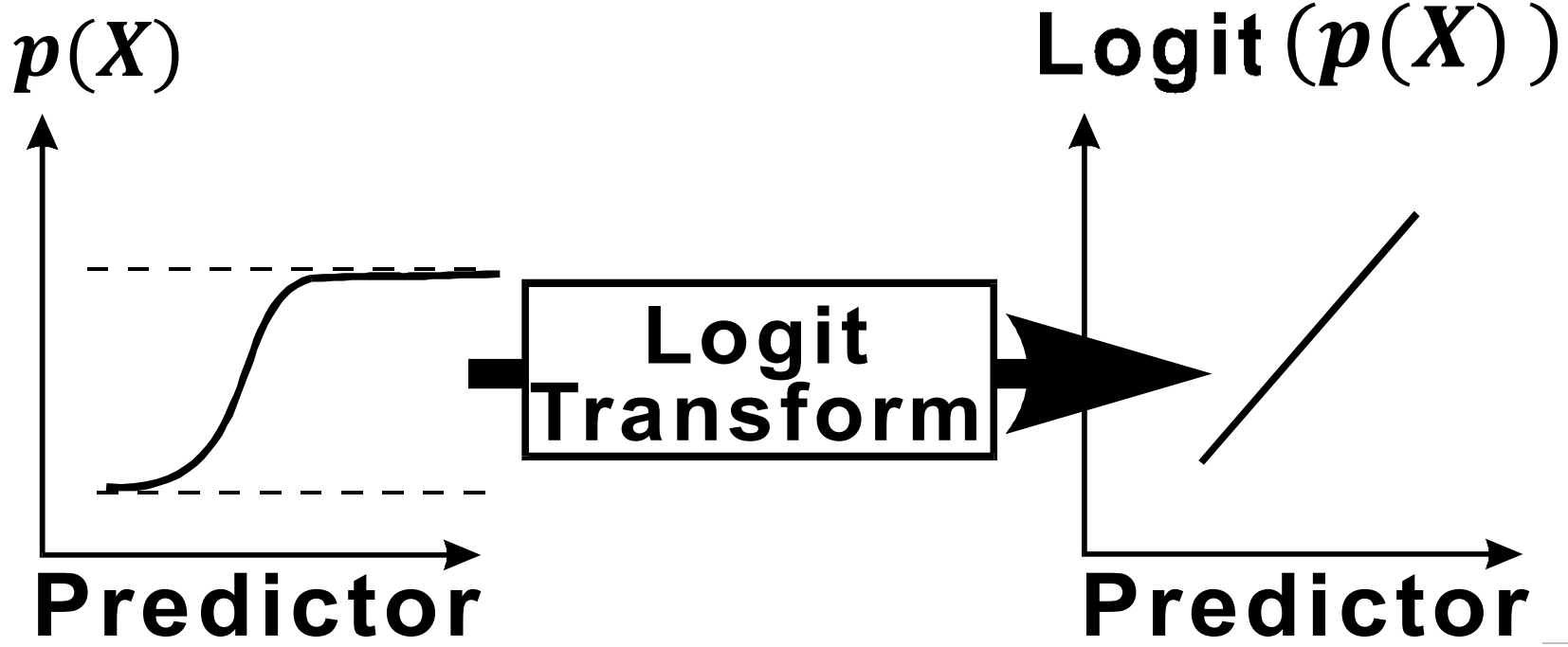
Lec 4: Logistic Regression

Logistic Regression

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 $p(X)$.



Odds Ratio

- From $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, we have $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x}$

- The odds ratio increases **multiplicatively** by e^{β_1} for every 1-unit increase in x
  - The odds at $X = x + 1$ are e^{β_1} times the odds at $X = x$
  - $\frac{odds(x+1)}{odds(x)} = e^{\beta_1}$
- Therefore, e^{β_1} **is an odds ratio!**

- e^{β_1} represents the change in the odds of the outcome (multiplicatively) by increasing x by 1 unit
  - If $\beta_1 > 0$, the odds and probability increase as x increases ($e^{\beta_1} > 1$)
  - If $\beta_1 < 0$, the odds and probability decrease as x increases ($e^{\beta_1} < 1$)
  - If $\beta_1 = 0$, the odds and probability are the same at all x levels ($e^{\beta_1}=1$)

Logistic Regression

1. How do we find β? Maximum Likelihood Estimation

2. Which of the following are correct for interpreting $\beta_j > 0$? a), b), c)
   a) A one unit increase in X_j is associated with an increase of β_j in the log odds (logit) of $P(Y = 1)$.
   b) A one unit increase in X_j is associated with an increase of a multiplicative factor of e^{β_j} in the odds of $P(Y = 1)$.
   c) The odds ratio of X_j is $e^{\beta_j} > 1$.
   d) A one unit increase in X_j is associated with an increase of $e^{\beta_j}/(1 + e^{\beta_j})$ in $P(Y = 1)$.

Lec 5: Multinomial Regression

Nominal and Ordinal Responses

- Nominal response
  - Red, green, blue
  - Yes, no
  - Sick, healthy
- Ordinal response
  - Young, middle aged, old
  - Dislike very much, dislike, no opinion, like, like very much

Nominal Logistic Regression

- Choose one category as the reference category, say the 1st category
- Define the logits for the other categories as

$$\text{logit}(\pi_j) \equiv \log(\frac{\pi_j}{\pi_1}) = x^T\beta_j, \quad \text{for } j = 2, \ldots, J.$$

- Since the probabilities add up to 1, we have

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(x^T\hat{\beta}_j)}$$

$$\hat{\pi}_j = \frac{\exp(x_j^T\hat{\beta}_j)}{1 + \sum_{j=2}^{J} \exp(x^T\hat{\beta}_j)}, \quad \text{for } j = 2, \ldots, J.$$

- Changing the reference category will change the above probabilities? False.

Ordinal Logistic Regression

- Cumulative logit model

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = x^T \beta_j.$$

- Special case: Proportional odds model

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.$$

Lec 6: Poisson Regression

Poisson Regression

- Three ingredients:
  - Likelihood of response = Poisson density

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!},$$

  - Find expectation of y condition on x

$$\mathrm{E}(Y|X_1, \ldots, X_p) = \lambda(X_1, \ldots, X_p)$$

  - Link expectation with the linear function of predictors

$$\log(\lambda(X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Estimate β by maximizing the likelihood.

Poisson vs Linear Regression

- An increase in X_j by one unit is associated with a change in $E(Y) = \lambda = e^{X\beta}$ by a factor of $\exp(\beta_j)$.
  - By contrast, in linear regression, an increase in X_j by one unit is associated with an increase of β_j in $E(Y) = \mu = X\beta$.
- Poisson regression implicitly assumes that mean equals variance.
  - By contrast, linear regression assumes the variance takes on a constant value.
- Poisson regression gives non-negative predictions.
  - By contrast, linear regression predictions can be negative.

Lec 7: Discriminant Analysis

Bayes Theorem for Classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probablistic modeling. Here we focus on a simple result, known as Bayes theorem:

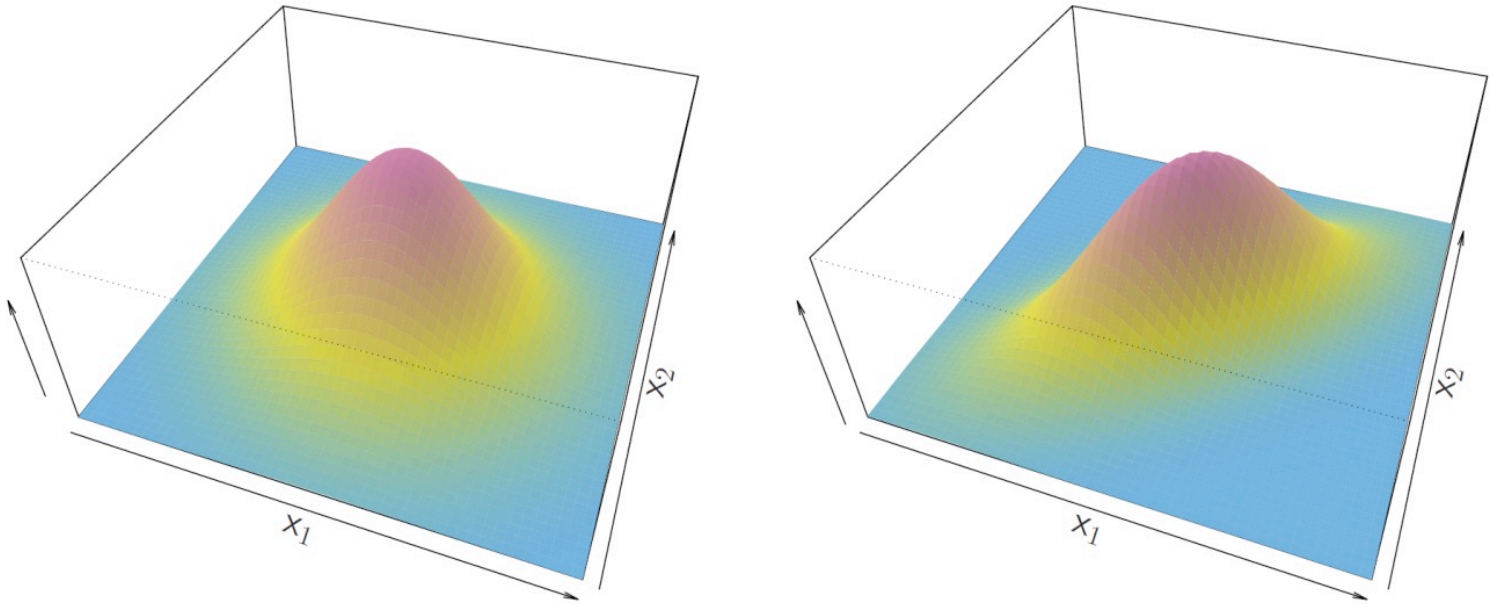
$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

One writes this slightly differently for discriminant analysis:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}, \quad \text{where}$$

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

Linear Discriminant Analysis when $p > 1$



$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$$

$$\text{Discriminant function: } \delta_k(x) = x^T \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \log \pi_k$$

Despite its complex form,

$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \ldots + c_{kp}x_p$ — a linear function.

From $\delta_k(x)$ to Probabilities

Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}.$$

So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{\Pr}(Y = k|X = x)$ is largest.

When $K = 2$, we classify to class 2 if $\widehat{\Pr}(Y = 2|X = x) \geq 0.5$, else to class 1.

We can compare the probability to a different threshold to balance Type I and Type II error?

True.

Classification Terminology

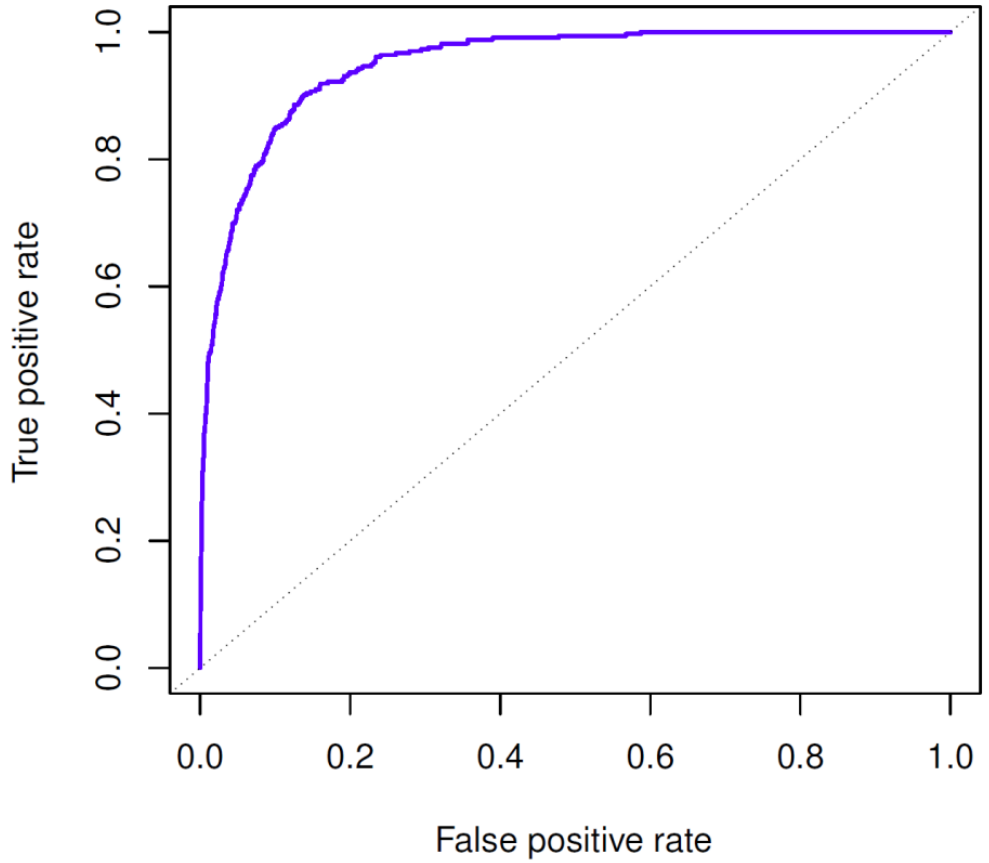
		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

The Receiver Operating Characteristics (ROC) Curve



- The ROC plot
  - False positive rate: 1-Specificity; healthy identified as not
  - True positive rate: Sensitivity; sick identified as so
- Higher area under the curve (AUC) is better

Other Forms of Discriminant Analysis

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different $\mathbf{\Sigma}_k$ in each class, we get *quadratic discriminant analysis*.
- With $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$ (conditional independence model) in each class we get *naive Bayes*. For Gaussian this means the $\mathbf{\Sigma}_k$ are diagonal.
- Many other forms, by proposing specific density models for $f_k(x)$, including nonparametric approaches.

Comparison of LDA, QDA, NB, Logistic

- LDA is a special case of QDA? True.
- Any linear classifier is a special case of NB? True.
- Discriminant Analysis is a type of discriminative learning? False.

	LDA	QDA	NB
Gaussian $f_{kj}(x_j)$?	Yes	Yes	Not necessarily
Diagonal Σ_k?	No	No	Yes
Shared $\Sigma_k = \Sigma$?	Yes	No	No

Lec 8: Support Vector Machine

The Non-separable Case: Soft Margin

- The non-separable case pursues:
  - Greater robustness to individual observations,
  - Better classification for most of the training observations.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

- Slack variables ϵ_i allow observations to be on the wrong side.
- Cost C is a budget for the amount that the margin can be violated.
- "Support Vectors": observations that lie on the margin, or on the wrong side of the margin.

SVM Facts

1. The soft margin SVM optimization for the non-separable case has no solution with $M > 0$? False.

2. For observations that are not support vectors, they must be at least a distance M from the hyperplane? True.

3. If C is integer, it is not possible to have more than C observations on the wrong side of the margin? False.

4. Smaller C leads to wider margin, thus a classifier that is more biased but has lower variance? False.

5. SVM is very robust to outliers that are not support vectors? True.

SVMs with More than Two Classes

Two popular ideas:

- One-Versus-One Approach:
  - Construct $\binom{K}{2}$ SVMs for each pair of classes.
  - Assign a test observation to the class to which it was most frequently assigned in these $\binom{K}{2}$ pairwise classifications.
- One-Versus-All Approach:
  - Construct K SVMs, each time comparing one of the K classes to the remaining $K-1$ classes.
  - Assign a test observation x to the class for which $\beta_{0k} + \beta_{1k}x_1 + \cdots + \beta_{pk}x_p$ is largest.

Support Vector Classification vs Regression

- Classification:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

- Regression:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max[0, |y_i - f(x_i)| - \epsilon] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Lec 9: Principal Component Analysis

The First Principal Component

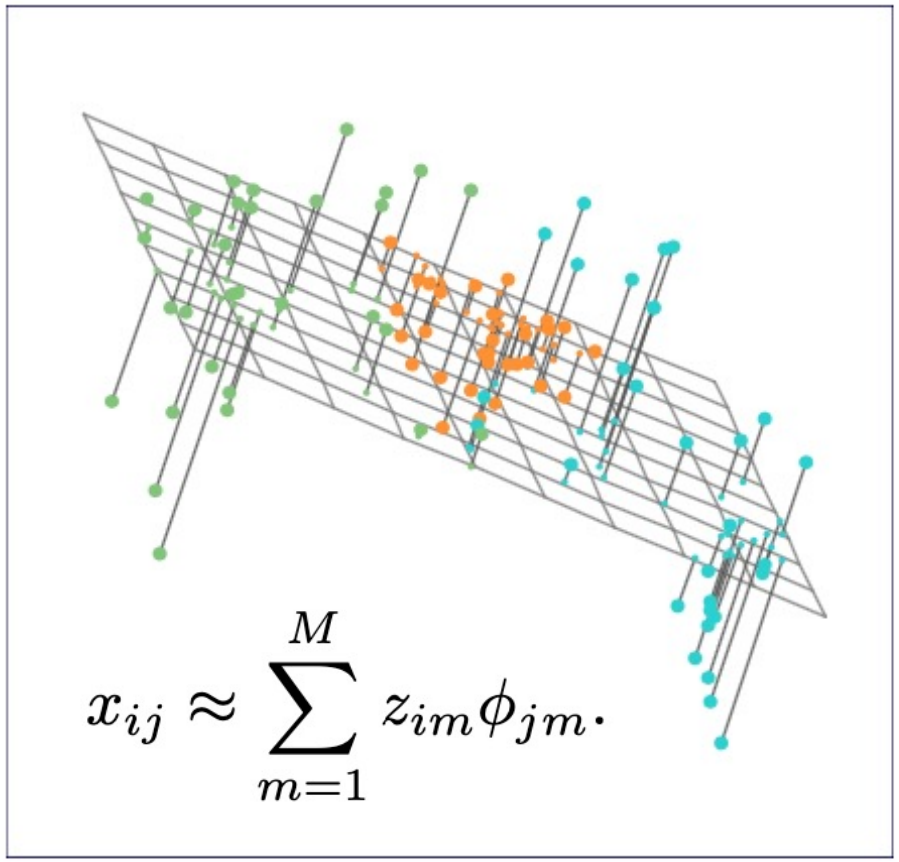
- Consider a set of features: $X_1, X_2, \ldots, X_p$ on n individuals. The first principal component (PC) is the linear combination $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$ that has the largest variance

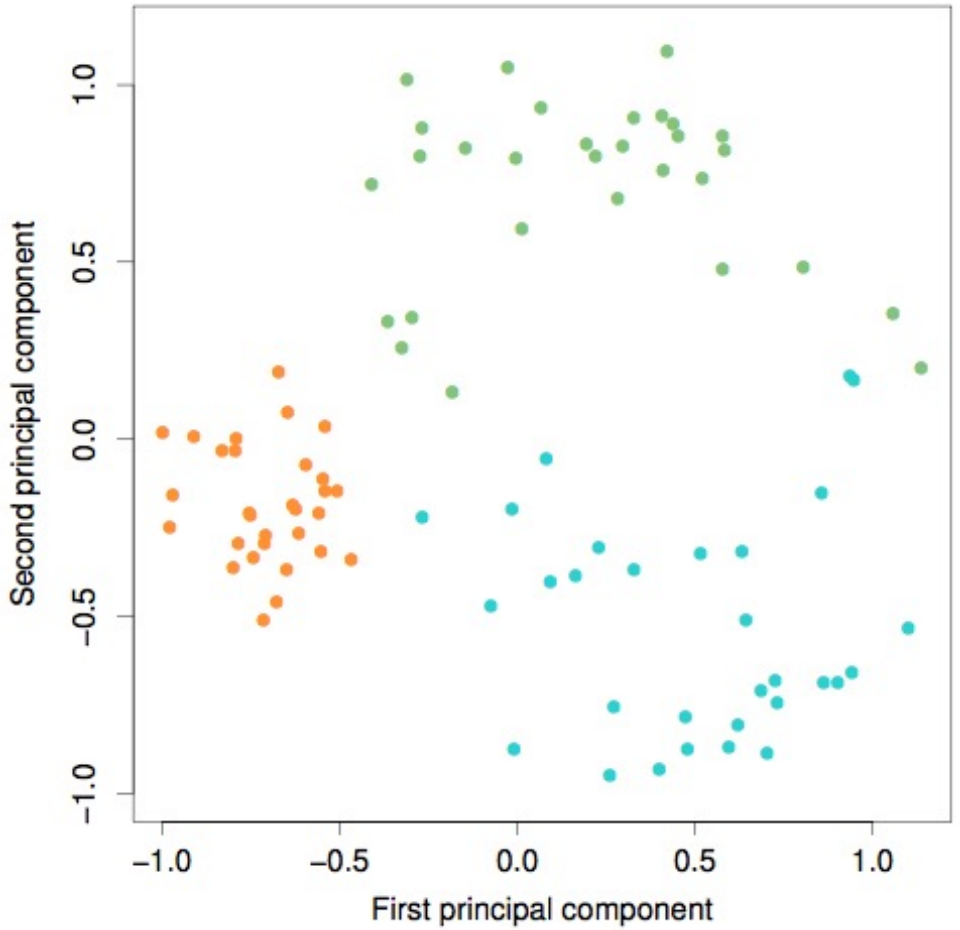
$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \right\} \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1.$$

- The entries $z_{11}, z_{21}, \ldots, z_{n1}$ are the PC scores.
- The PC loading vector: $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$

- The second principal component Z_2 is the linear combination that has the maximal variance such that ϕ_1 and ϕ_2 are orthogonality.

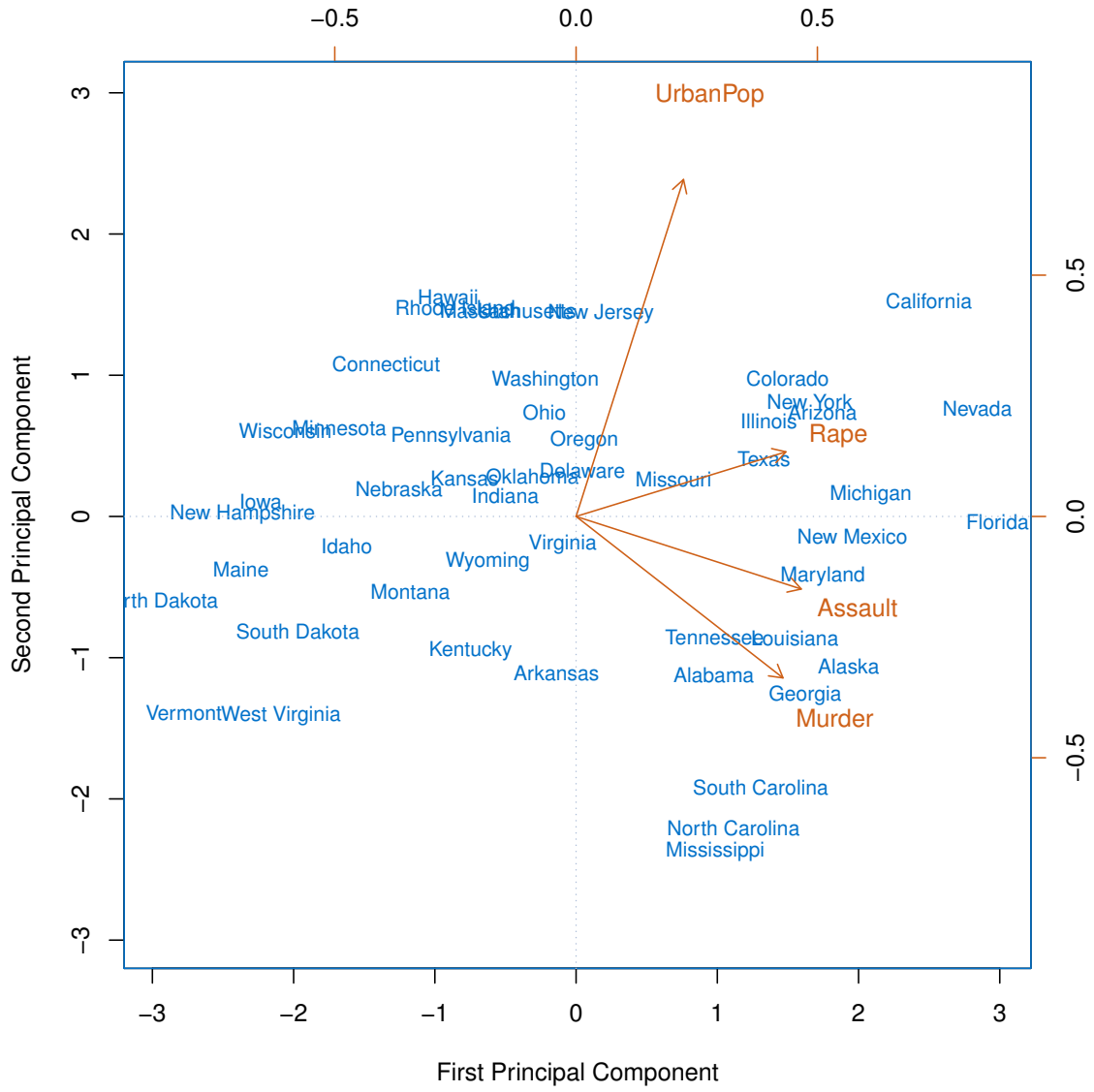
Another Interpretation of Principal Components





- PCA finds hyperplanes closest to the observations? True.

Biplot

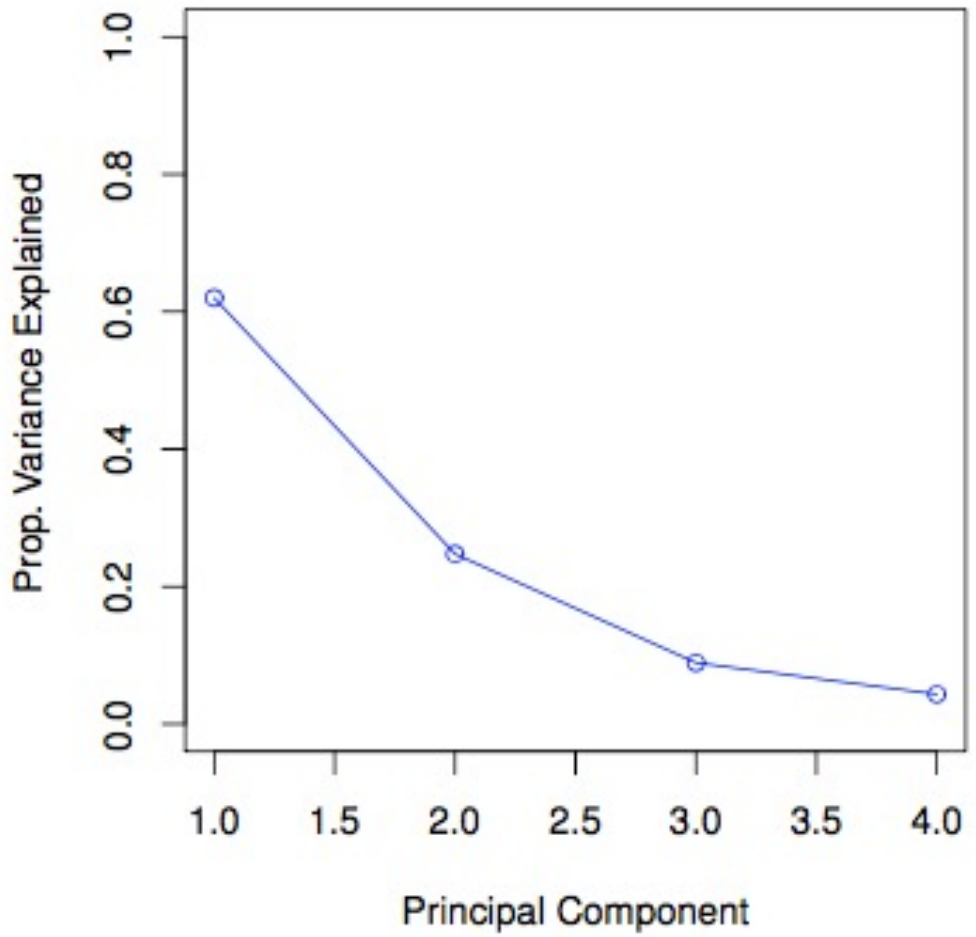


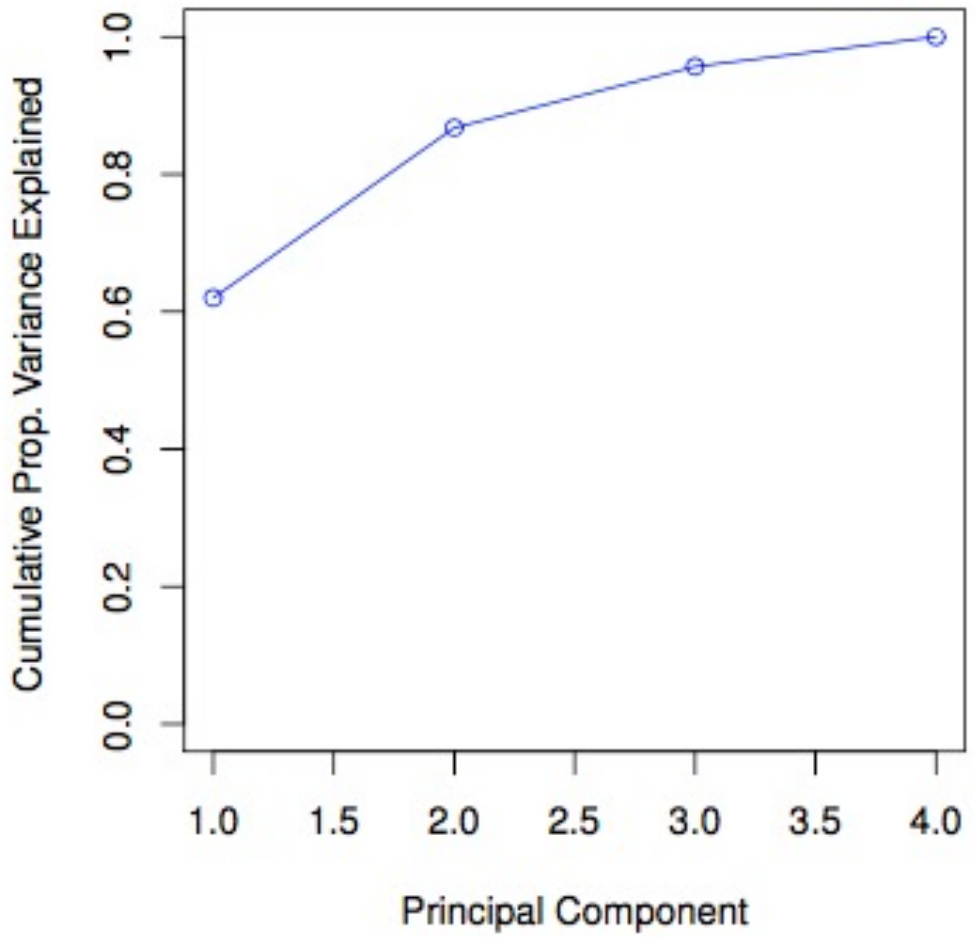
	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Principal components are only unique up to sign? True.

We can read PC scores from left and bottom axis and PC loadings from right and top axis? True.

Scree Plot





Scree plot can be used to determine the necessary number of PCs? True.

Usage of PCA

- Data visualization
- EDA
- Clustering
- Dimension reduction for PCR
- Missing data imputation

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i,$$

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{(i,j)\in\mathcal{O}} \left(x_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2 \right\},$$

$$\hat{x}_{ij} = \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm}$$

Roadmap and Big Picture

What Have We Learnt?

- Supervised learning:
  - Regression
    - Linear / Transformation
    - Forward / Backward / Subset selection
    - Ridge / Lasso / CV
    - Poisson / GLM
    - PCR / PLS
  - Classification
    - Logistic / Multinomial
    - LDA / QDA / Naive Bayes
    - SVM
- Unsupervised learning:
  - PCA
    - Data visualization
    - Dimension reduction
    - Matrix completion
  - Clustering
    - K-means
    - Hierarchical

More Advanced Future Topics

- Machine Learning (Nonlinearity)
  - Splines, Local regression
  - Decision tree, Random forest, Boosting
  - Kernel method, SVM with kernels
  - Deep learning, CNN, RNN
- Econometrics
  - Time series models
  - Instrumental variables for endogeneity
  - Casual inference
- Big Data Analysis
  - High dimensional inference
  - Multiple testing
  - Optimization methods
- Biostatistics / Bioinformatics
  - Survival analysis
  - Experimental design
  - Epidemiology, SEIR model