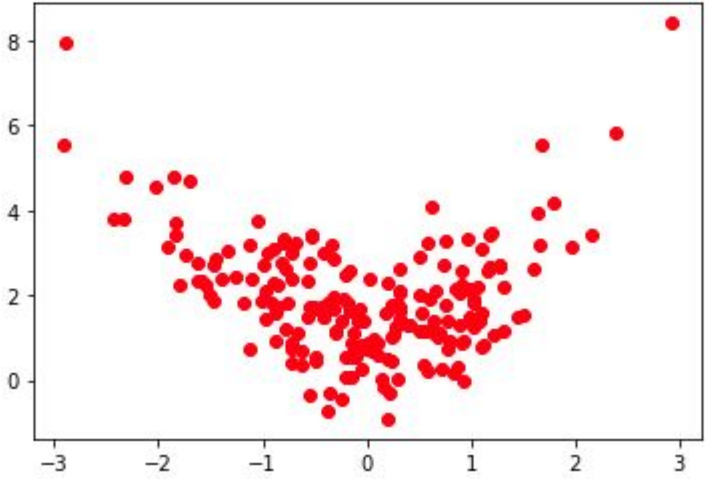
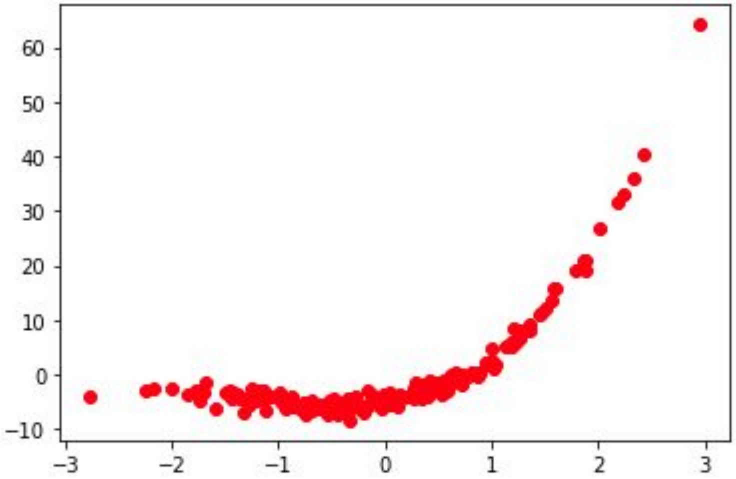
Generalized Linear Regression Models

GR 5205 / GU 4205
Section 3

Columbia University
Xiaofei Shi

Recalling adding curvature





Change the design matrix to incorporate the nonlinear function of predictors

Recalling adding categorical predictors

```
levels(mobility$State)

##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID"
## [15] "IL" "IN" "KS" "KY" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC"
## [29] "ND" "NE" "NH" "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC" "SD"
## [43] "TN" "TX" "UT" "VA" "VT" "WA" "WI" "WV" "WY"
```

Change the design matrix to incorporate the categorical predictors

Model after transformations

Relationship: $Y = x\beta + \epsilon;$

Assumptions: $\mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \sigma^2 I_n \, .$

- Only changes the design matrix.
- Parameter estimations and inferences remain the “same” as in multivariate linear regression models.
- But sometimes transformations on the predictors are not enough…

Transforming the response

- Another way to accommodate nonlinearity: transform the response variables
- We assume the model as:

$$g(Y) = x\beta + \epsilon \quad \Leftrightarrow \quad Y = g^{-1}\left(x\beta + \epsilon\right)$$

- Even we assume Gaussian noise, the distribution of the response variable is non-Gaussian.

The noise around the mean of response variable is not additive

Choice of transformation

- Log, polynomials, sine and cosine, exponential, etc
- Always choose the model first on their physical meaning
- Any other way?

Choice of transformation

$$h(Y) = g(x)\beta + \epsilon$$

There are too many possibilities for $g(\cdot)$ and $h(\cdot)$, so let's consider just a few special cases.

The *power transformation family*, defined for strictly the positive variable U, is

$$\psi(U,\lambda) = \begin{cases} U^{\lambda} & \lambda \neq 0 \\ \log U & \lambda = 0 \end{cases}$$

With the 0th power understood to represent a logarithm, we try to find λ_1 and λ_2 so that

$$\mathrm{E}(Y^{\lambda_2}|X = x) \approx \beta_0 + \beta_1 x^{\lambda_1}$$

This is a more manageable problem.

Box-Cox power transformation

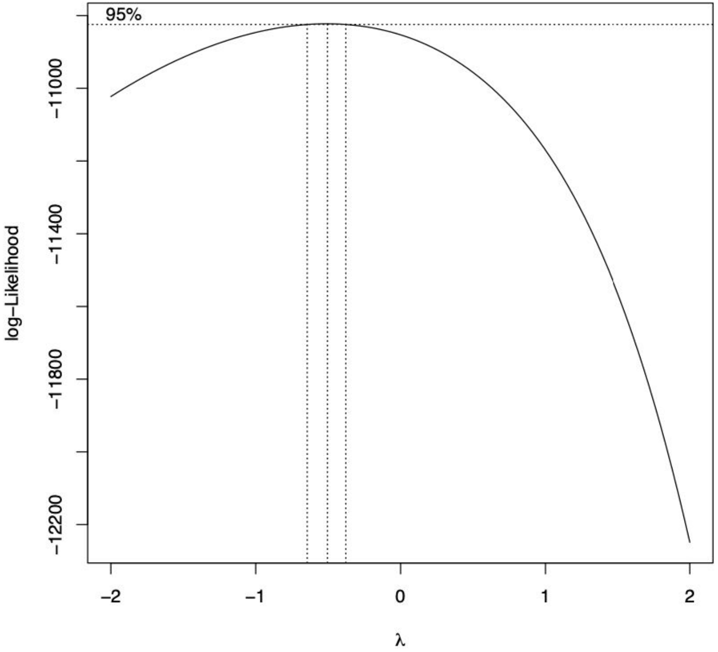
$$b_\lambda(Y) := \frac{Y^\lambda - 1}{\lambda} = x\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

Based on maximum likelihood theory. Finds the power transformation $\psi_M(Y, \lambda)$ that makes the residuals as closely as possible resemble a random sample from a Normal population.

The output is a *profile log-likelihood* like

Box-Cox power transformation

$$b_\lambda(Y) := \frac{Y^\lambda - 1}{\lambda} = x\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$



- Python:

  https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html

  R:

  https://www.rdocumentation.org/packages/EnvStats/versions/2.4.0/topics/boxcox

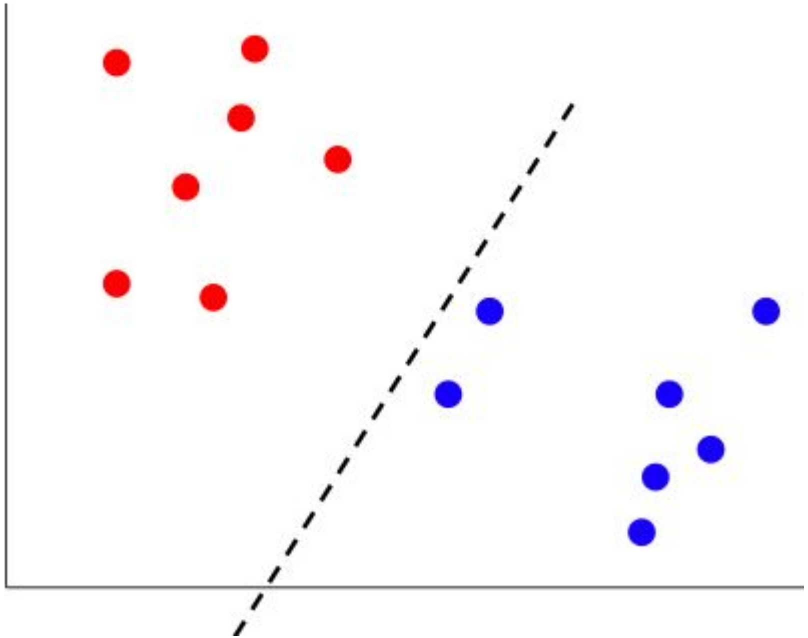
Special case:
What if your response variable is categorical?

Instead of modeling the relationship of Y wrt x,

we model the conditional probability of $Y|X = x$, i.e.

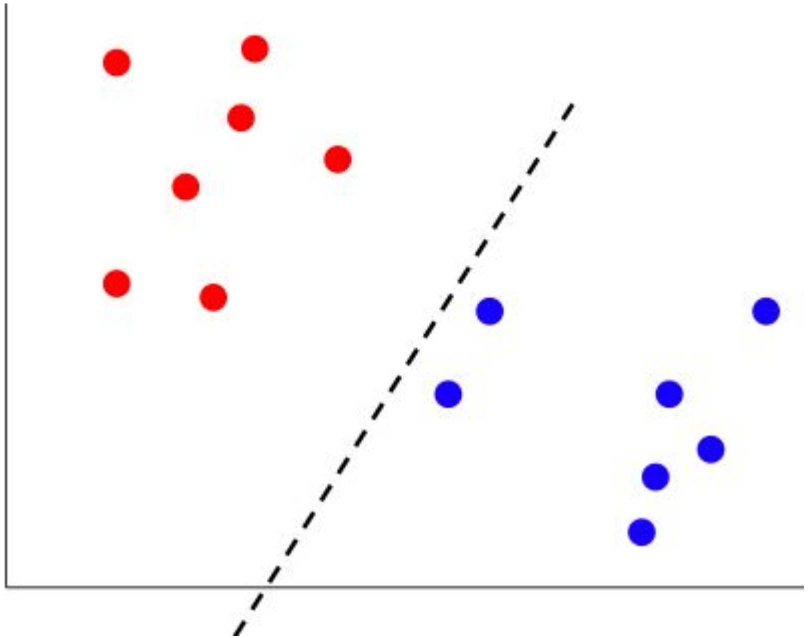
$$\mathbb{P}[Y|X = x] = x\beta$$

Binary response



- Support vector machine (SVM)
- Logistic regression

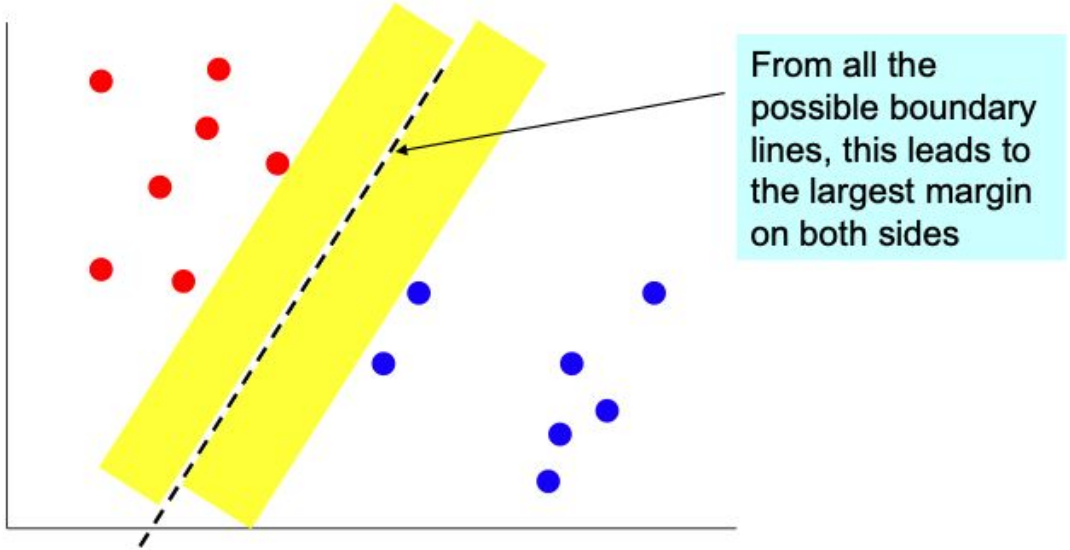
SVM



SVM

- Instead of fitting all points, focus on boundary points

•Learn a boundary that leads to the largest **margin** from both sets of points (that is, largest distance to the closest point on either side)

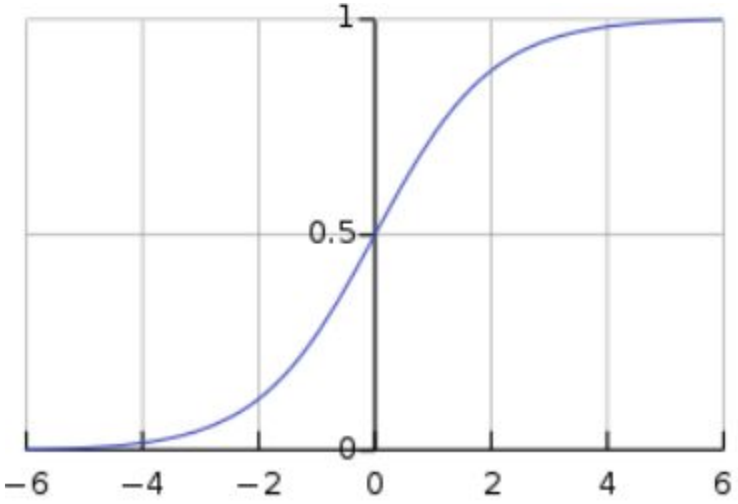


Logistic regression

- Instead of modeling Y =0 or 1 directly, we modify the probability of P(Y=0|x) as

$$P[Y=0;x] = \frac{1}{1+\exp(x\beta)}$$

$$P[Y=1;x] = 1 - \frac{1}{1+\exp(x\beta)} = \frac{\exp(x\beta)}{1+\exp(x\beta)}$$



Expressing conditional likelihood

More than 2 categories

- Logistic regression in more general case, where $Y \in \{y_1,\ldots,y_K\}$

for *k<K*

$$P(Y = y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki}X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

for *k=K* (normalization, so no weights for this class)

$$P(Y = y_K|X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

Predict $f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			