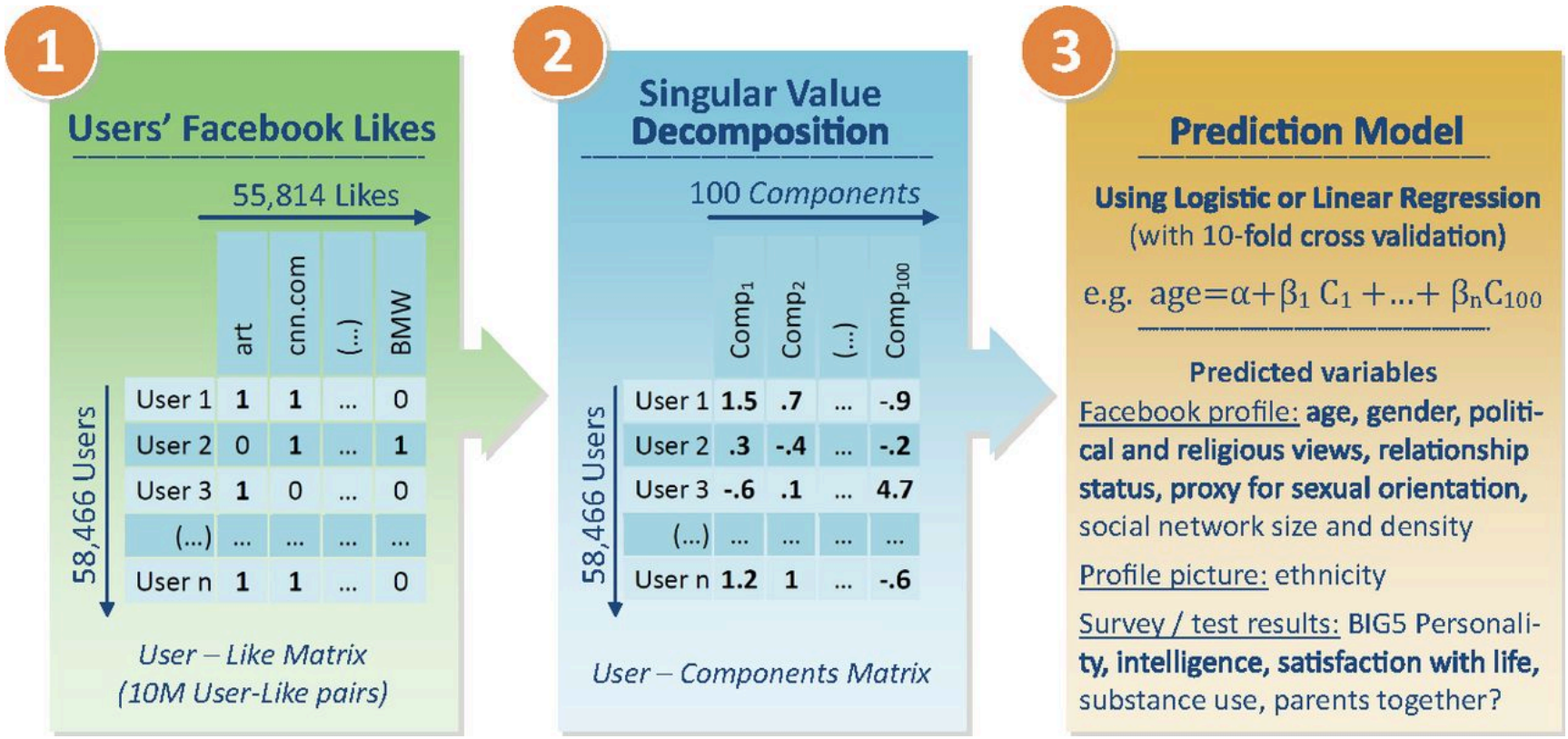
Stat 306:
Finding Relationships in Data.
Lecture 22
6.2 Principal Component Analysis + Count Regression

Kosinski et al. (2013):



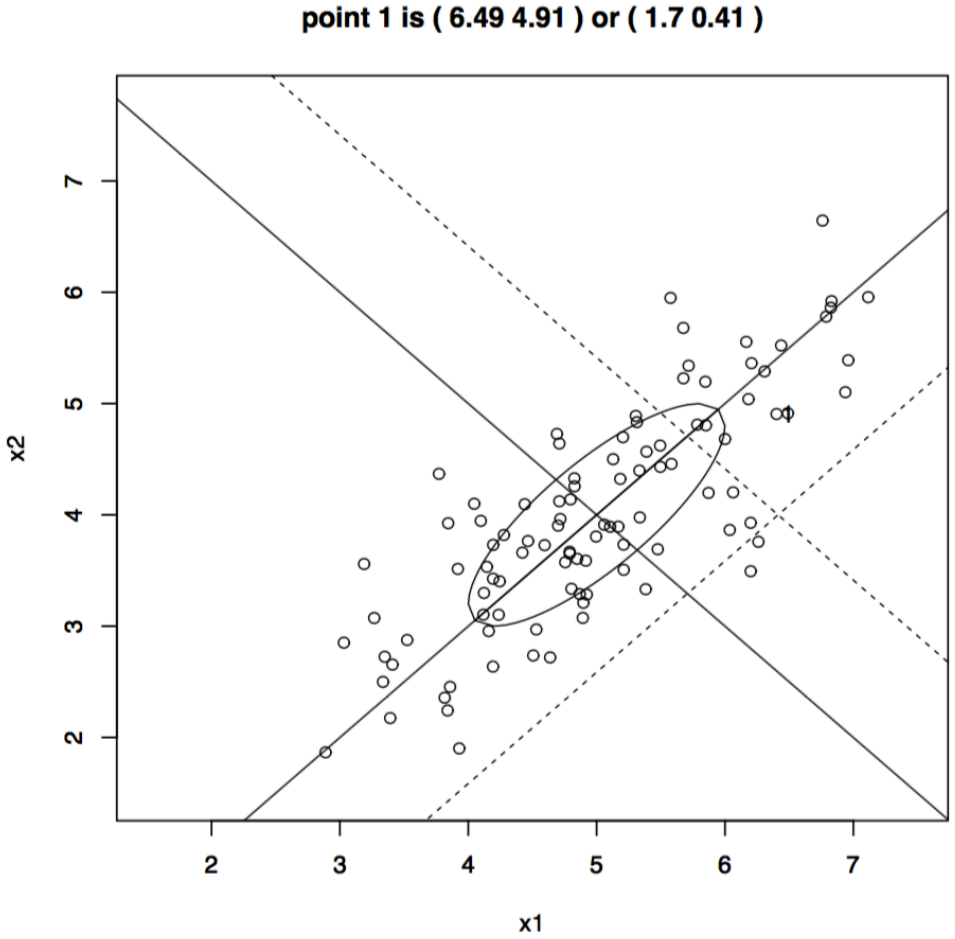
Principal component analysis.

- Principal component analysis (PCA) is a dimensionality reduction technique
  - many statistical models suffer from high correlation between covariates.
  - PCA can be used to produce linear combinations of the covariates that are uncorrelated between each other.
  - Often there are too many potential variables.
  - Sometimes there are more variables than observations (p>n)!
  - PCA can be used to reduce the number of variables in the model while maintaining as much information as possible.

Excellent explanation:

http://www.milanor.net/blog/performing-principal-components-regression-pcr-in-r/

Figure 7.1: Two-dimensional data in original coordinates and rotated coordinates. The solid lines show the axes of the rotated coordinates centred at the vector of sample means. The dashed lines show one positive unit in the new coordinates. The point labelled with "1" is $(x_1 = 6.49, x_2 = 4.91)$ and it becomes $(1.70, 0.41)$ in the new coordinates. The first axis (principal component) of the rotated coordinates is the major axis of the "ellipsoid" of points. See Section 7.1.1 for further explanations of the ellipse/ellipsoid.



Steps to obtain principal components:

1. Calculate the estimated covariance matrix of X (N observations, K variables), is the matrix Σ with entries:

$$\Sigma_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_{ij} - \bar{X}_j\right) \left(X_{ik} - \bar{X}_k\right)$$ for k in 1,…K, and j in 1,...,J

https://en.wikipedia.org/wiki/Covariance#Calculating_the_sample_covariance

2. Obtain eigenvalues and their corresponding eigenvectors for this covariance matrix.

https://www.youtube.com/watch?v=IdsV0RaC9jM&ab_channel=patrickJMT

3. Reduce X to a lower-dimensional projection: the eigenvectors corresponding to the largest eigenvalues.

4. Center the eigenvectors. This is the "best" projection of X onto the lower-dimension.

5. Determine what percentage of the original variance is maintained in the lower-dimensional projection.

Why does this work?

Suppose $\mathbf{X} = (X_1, \ldots, X_p)^T$ is a random vector with covariance matrix $\mathbf{\Sigma} = (\sigma_{jk})_{1 \le j,k \le p}$. Data are $(x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, considered as n independent realizations of $(X_1, \ldots, X_p)$. The sample covariance matrix is denoted as $\mathbf{S}$, and this is an estimate of $\mathbf{\Sigma}$.

From Appendix A, $\mathbf{w}^T\mathbf{X} = \sum_{j=1}^{p} w_j X_j$ is a linear combination with variance $\mathrm{Var}\,(\mathbf{w}^T\mathbf{X}) = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = \sum_{j=1}^{p}\sum_{k=1}^{p} w_j w_k \sigma_{jk}$.

For the *first principal component*, we want to find a coefficient vector $\mathbf{w}$ with unit length such that

$$\mathrm{Var}\,(\mathbf{w}^T\mathbf{X}) = \mathrm{Var}\,(\mathbf{w}^T(\mathbf{X} - \mathbf{a})) = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \tag{7.1}$$

is maximized. Why is unit length $\mathbf{w}^T\mathbf{w} = 1$ condition imposed?

Why does this work?

To solve this problem using calculus, we introduce the Lagrangian function:

$$H(\mathbf{w}) = \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1), \tag{7.2}$$

$$\frac{\partial H}{\partial \mathbf{w}} = 2\boldsymbol{\Sigma}\mathbf{w} - 2\lambda\mathbf{w}. \tag{7.3}$$

Set to zero and let the solution involve $(\mathbf{w}^*, \lambda^*)$:

$$2\boldsymbol{\Sigma}\mathbf{w}^* - 2\lambda^*\mathbf{w}^* = 0, \text{ or } \boldsymbol{\Sigma}\mathbf{w}^* = \lambda^*\mathbf{w}^*. \tag{7.4}$$

Equation (7.4) is the equation for an eigensystem (review your linear algebra textbook if needed). So $\mathbf{w}^*$ is a unit-length eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue λ^*. Because $\boldsymbol{\Sigma}$ is symmetric, the eigenvalues are real. Because $\boldsymbol{\Sigma}$ is non-negative definite, the eigenvalues are non-negative.

However….

Instead of using the covariance matrix of X to do PCA, we should use the correlation matrix of X. In this way all the variables in the X matrix are on a similar scale.

To go from a Covariance matrix, Σ, to a correlation matrix ρ:

$$D = \sqrt{diag(\Sigma)}$$

$$\rho = D^{-1}\Sigma D^{-1}$$

Then obtain eigenvalues and their corresponding eigenvectors for ρ as before.

PCA Example 1: Wine data.

The variables come from chemical analyses of wines grown in the same region in Italy but derived from three different cultivars or classes, with respectively sample sizes of 59, 71, 48 for a total of $n = 178$.

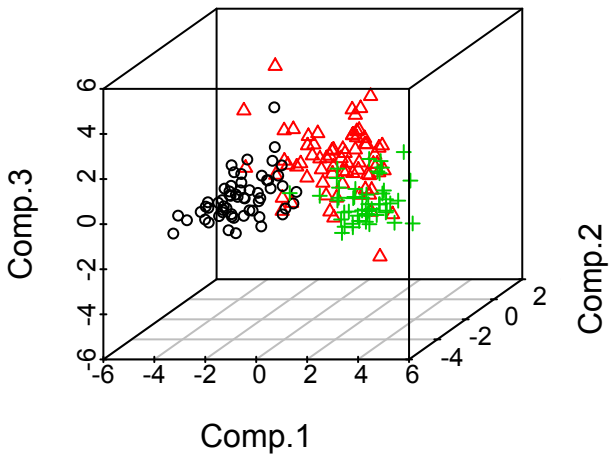
Table 7.1: Variables in wine classification data set

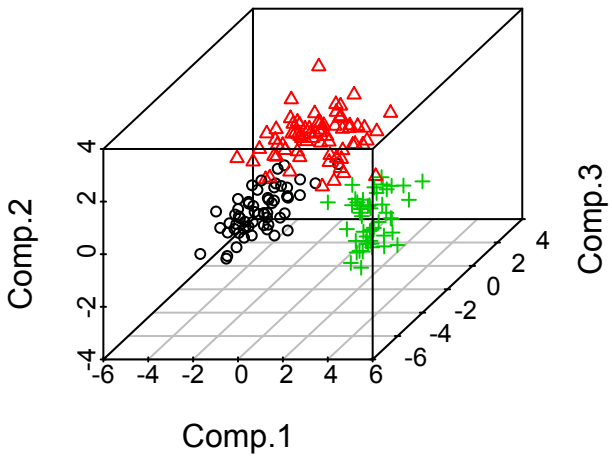
index	variable	index	variable
1	Alcohol	8	Nonflavanoid phenols
2	Malic acid	9	Proanthocyanins
3	Ash	10	Color intensity
4	Alcalinity of ash	11	Hue
5	Magnesium	12	OD280/OD315 of diluted wines
6	Total phenols	13	Proline
7	Flavanoids		

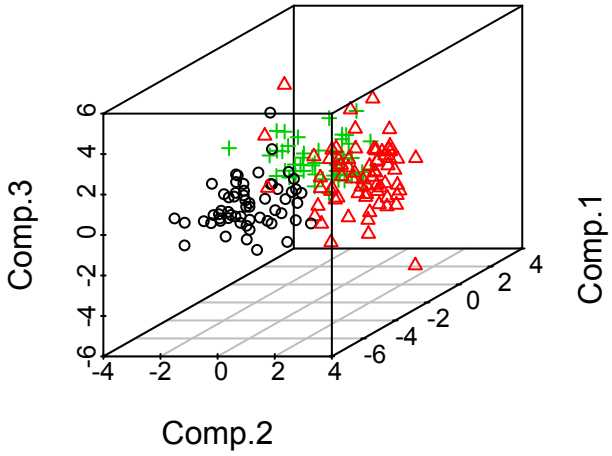
PCA Example 1: Wine data.

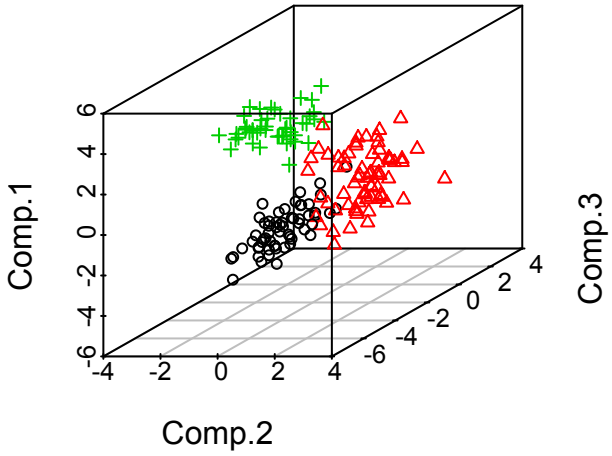
Table 7.3: Loading vectors for first 5 components of the wine data set.

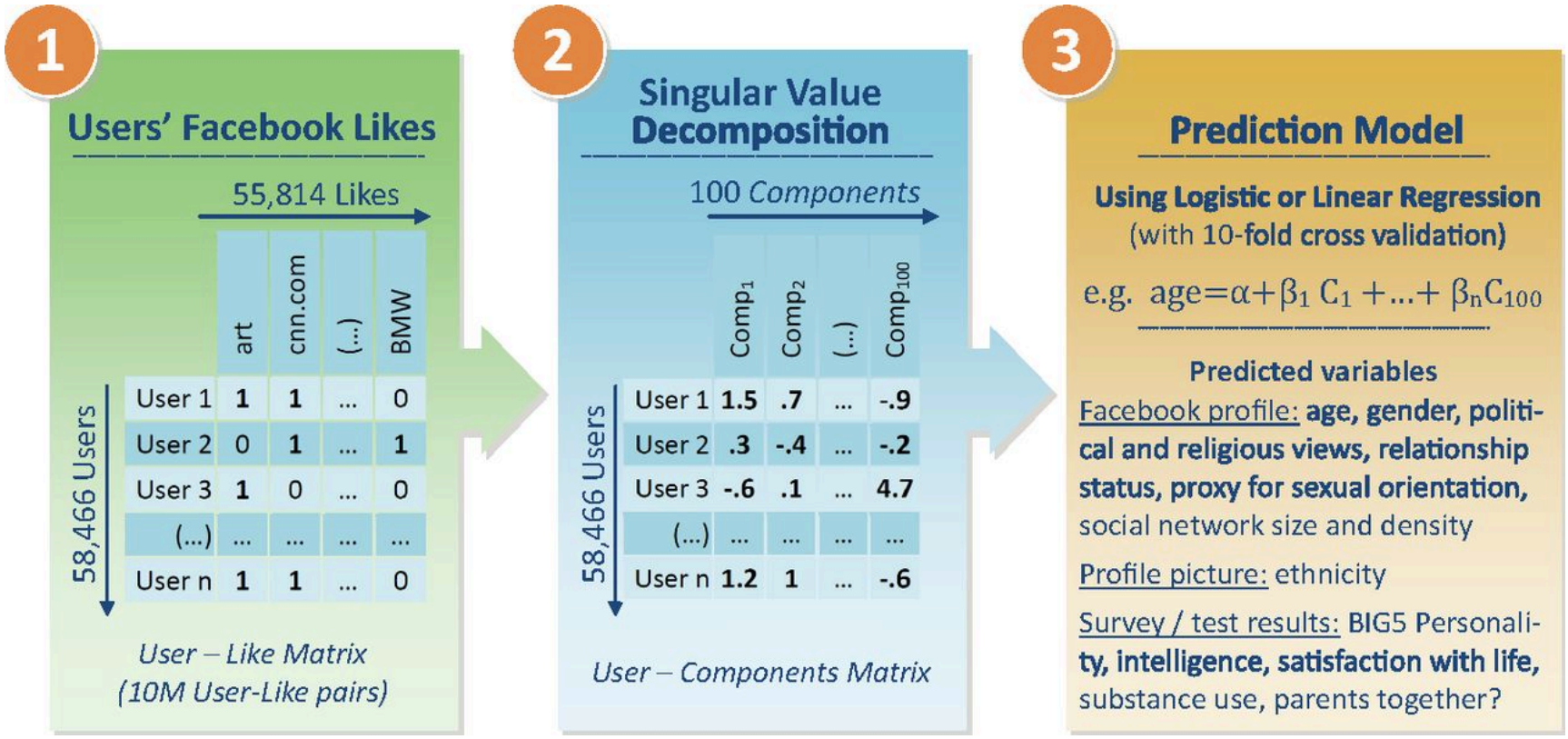
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
alcohol	-0.144	-0.484	-0.207	-0.018	0.266
malic	0.245	-0.225	0.089	0.537	-0.035
ash	0.002	-0.316	0.626	-0.214	0.143
alkalinity	0.239	0.011	0.612	0.061	-0.066
magnesium	-0.142	-0.300	0.131	-0.352	-0.727
phenol	-0.395	-0.065	0.146	0.198	0.149
flavanoid	-0.423	0.003	0.151	0.152	0.109
nonflavphenol	0.299	-0.029	0.170	-0.203	0.501
proanthocyanin	-0.313	-0.039	0.149	0.399	-0.137
colorintense	0.089	-0.530	-0.137	0.066	0.076
hue	-0.297	0.279	0.085	-0.428	0.174
od280	-0.376	0.165	0.166	0.184	0.101
proline	-0.287	-0.365	-0.127	-0.232	0.158


Comp.3
Comp.2
Comp.1


Comp.2
Comp.3
Comp.1


Comp.3
Comp.1
Comp.2


Comp.1
Comp.3
Comp.2


1
Users' Facebook Likes
55,814 Likes
58,466 Users
art
cnn.com
(...)
BMW
User 1 1 1 ... 0
User 2 0 1 ... 1
User 3 1 0 ... 0
(...) ... ...
User n 1 1 ... 0
User – Like Matrix
(10M User-Like pairs)
2
Singular Value Decomposition
100 Components
58,466 Users
Comp1
Comp2
(...)
Comp100
User 1 1.5 .7 ... -.9
User 2 .3 -.4 ... -.2
User 3 -.6 .1 ... 4.7
(...) ... ...
User n 1.2 1 ... -.6
User – Components Matrix
3
Prediction Model
Using Logistic or Linear Regression
(with 10-fold cross validation)
e.g. age=α+β1 C1 +...+ βnC100
Predicted variables
Facebook profile: age, gender, political and religious views, relationship status, proxy for sexual orientation, social network size and density
Profile picture: ethnicity
Survey / test results: BIG5 Personality, intelligence, satisfaction with life, substance use, parents together?

Examples of general regression

1. Usual regression (normal or Gaussian response): $Y_i \sim N(\mu, \sigma^2)$ is extended to $Y \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$. That is, $\mu(\mathbf{x}_i) = \mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$.

2. Binary regression (logistic regression as special case): $Y_i \sim$ Bernoulli(π), where $\pi = \Pr(Y_i = 1)$ and $1 - \pi = \Pr(Y_i = 0)$, is extended to $Y_i \sim$ Bernoulli$(\pi(\mathbf{x}_i))$, where $\pi(\mathbf{x}_i) = \exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}/[1 + \exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}$ or

$$\log\left\{\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right\} = \mathbf{x}_i^T\boldsymbol{\beta}.$$

   [Logistic cdf is $F(z) = e^z/(1 + e^z)$, $-\infty < z < \infty$]

3. Count regression (Poisson regression as special case): $Y_i \sim$ Poisson(λ) with mean $\lambda > 0$ and $P(Y_i = y) = \lambda^y e^{-\lambda}/y!$ $(y = 0, 1, \ldots)$ is extended to $Y_i \sim$ Poisson$(\lambda(\mathbf{x}_i))$, where $\lambda(\mathbf{x}_i) = \exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}$ or

$$\log \lambda(\mathbf{x}_i) = \mathbf{x}_i^T\boldsymbol{\beta}.$$

   Poisson regression can be used for insurance claim data to model the number of car accidents (or claims) per year for individuals as a function or demographic and risk factors.

Count Regression

$$Y_i \sim \text{Poisson}(\lambda) \text{ with mean } \lambda > 0$$

We want to model the mean, λ , as a function of covariates X:

$$Y_i \sim \text{Poisson}(\lambda(\mathbf{x}_i)), \text{ where } \lambda(\mathbf{x}_i) = \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \text{ or}$$

$$\log \lambda(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Important property of the Poisson distribution:

E[Y] = λ
Var[Y] = λ

Maximum Likelihood Estimation

Steps to get Maximum Likelihood Estimates:

1. Define the Likelihood as function of the parameters given the data.
2. Define the log-Likelihood.
3. Maximize the log-Likelihood or minimize the negative log-Likelihood.

For Poisson regression, we have:

$$L(\boldsymbol{\beta}; data) = \prod_{i=1}^{n} [\lambda(\mathbf{x}_i)]^{y_i} \frac{e^{-\lambda(\mathbf{x}_i)}}{y_i!}$$

and:

$$\begin{aligned} \log L(\boldsymbol{\beta}; data) &= \sum_{i=1}^{n} \{y_i \log \lambda(\mathbf{x}_i) - \lambda(\mathbf{x}_i) - \log(y_i!)\} \\ &= \sum_{i=1}^{n} \{y_i \mathbf{x}_i^T \boldsymbol{\beta} - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \log(y_i!)\} \end{aligned}$$

Unfortunately…

There is no closed form solution, but statistical software obtain $\hat{\beta}_0, \hat{\beta}_1$ with an iterative method.

Maximum Likelihood Estimation

Poisson: $P(Y = y; \mathbf{x}) = [\lambda(\mathbf{x})]^y e^{-\lambda(\mathbf{x})}/y!$, $\lambda(\mathbf{x}) = \exp\{\mathbf{x}^T\boldsymbol{\beta}\}$ (intercept 1 included in $\mathbf{x}$).
Data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$; $y_i \in \{0, 1, 2, \ldots\}$.
Poisson likelihood in $\boldsymbol{\beta}$ is:

$$L(\boldsymbol{\beta}; data) = \prod_{i=1}^{n} [\lambda(\mathbf{x}_i)]^{y_i} \frac{e^{-\lambda(\mathbf{x}_i)}}{y_i!}$$

Loglikelihood in $\boldsymbol{\beta}$ is:

$$\begin{aligned} \log L(\boldsymbol{\beta}; data) &= \sum_{i=1}^{n} \{y_i \log \lambda(\mathbf{x}_i) - \lambda(\mathbf{x}_i) - \log(y_i!)\} \\ &= \sum_{i=1}^{n} \{y_i \mathbf{x}_i^T \boldsymbol{\beta} - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \log(y_i!)\} \end{aligned}$$

For the null model of no effect for explanatory variables, $\lambda = e^{\beta_0}$ for all i, log-likelihood is:

$$\sum_{i=1}^{n} \{y_i \log \lambda - \lambda - \log(y_i!)\} = y_+ \log \lambda - n\lambda - \sum_{i=1}^{n} \log(y_i!),$$

with maximum likelihood estimate (MLE) $\hat{\lambda} = \overline{y}$.

Maximum Likelihood Estimation

Steps to get standard error for the Maximum Likelihood Estimates (MLE):

1. Take the second derivative of the log-Likelihood function. This is the Hessian Matrix. The negative of the Hessian is the Fisher information matrix.

2. Evaluate the Fisher Information matrix at the MLE. This is known as the "observed Fisher Information matrix".

3. Take the inverse of the "observed Fisher Information matrix". This is your estimate for the Variance-Covariance matrix of your parameter. The diagonal elements are the estimated Variances.

4. Take the square root of the diagonal elements to obtain the standard errors.

Maximum Likelihood Estimation

Gradient and Hessian.

$$\log L(\boldsymbol{\beta}; data) = \sum_{i=1}^{n} \{y_i \mathbf{x}_i^T \boldsymbol{\beta} - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \log(y_i!)\}$$

$$\frac{\partial \log L(\boldsymbol{\beta}; data)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \{\mathbf{x}_i y_i - \mathbf{x}_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}$$

$$\frac{-\partial^2 \log L(\boldsymbol{\beta}; data)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

SEs for $\hat{\beta}$s come from the square root of the diagonal elements of

$$\left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \Big|_{\hat{\boldsymbol{\beta}}} \right]^{-1}$$

Diagnostics for Poisson regression: **The Deviance.**

The residual deviance becomes ($\sum_{i=1}^{n} \log(y_i!)$ cancels and $\hat{\lambda}_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$)

$$2[\ell(\lambda_1^{(S)}, \ldots, \lambda_n^{(S)}) - \ell(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)] = 2\left[\sum_{i=1}^{n}\{y_i \log y_i - y_i\} - \sum_{i=1}^{n}\{y_i \mathbf{x}_i^T \boldsymbol{\beta} - \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}\right].$$

The null deviance becomes ($\hat{\lambda}_i = \overline{y}$)

$$2[\ell(\lambda_1^{(S)}, \ldots, \lambda_n^{(S)}) - \ell(\overline{y}, \ldots, \overline{y})] = 2\left[\sum_{i=1}^{n}\{y_i \log y_i - y_i\} - \{n\overline{y} \log \overline{y} - n\overline{y}\}\right] = 2\left[\sum_{i=1}^{n} y_i \log y_i - n\overline{y} \log \overline{y}\right].$$

Here $0 \log 0 = 0$ as limit of $z \log z$ as $z \to 0^+$.

$0 \leq$ residual deviance $\leq$ null deviance because ℓ(saturated) $\geq \ell$(explanatory) $\geq \ell$(no explanatory).

Diagnostics for Poisson regression: **The AIC**

1. The AIC seeks to identify "good models" by considering both the likelihood and the number of parameters in the model:

$$AIC = -2 \cdot log(Likelihood) + 2 \cdot (\#\text{ of parameters})$$

2. As such it is similar to the adjusted R^2.

3. Lower **AIC** suggests that the model is better.

Interpretation for Poisson regression:

Suppose model is $P(Y_i = y; \mathbf{x}_i) = e^{-\lambda_i}\lambda_i^y/y!$, where λ_i depends on $\mathbf{x}_i$.
The model-based expected number of occurrences of the k claims is

$$E_k = \sum_{i=1}^{n} \mathrm{E}\left[I(Y_i = k)\right] = \sum_{i=1}^{n} P(Y_i = k; \mathbf{x}_i) = \sum_{i=1}^{n} e^{-\lambda_i}\lambda_i^k/k!.$$

For a fitted model, replace λ_i by $\hat{\lambda}_i = \exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}$ from the regression model.

Expected number of 0s from the model is $\sum_{i=1}^{n} e^{-\hat{\lambda}_i}$.

Expected number of 1s from the model is $\sum_{i=1}^{n} e^{-\hat{\lambda}_i}\hat{\lambda}_i$.

Expected number of 2s from the model is $\sum_{i=1}^{n} e^{-\hat{\lambda}_i}\hat{\lambda}_i^2/2$.

Excellent information on Poisson Regression:

https://freakonometrics.hypotheses.org/9593

https://freakonometrics.hypotheses.org/2289