# ROCRegression: un paquete en R para la incorporación de covariables en el análisis ROC

María Xosé Rodríguez-Álvarez[1] Ignacio López de Ullibarri[2] Carmen Cadarso-Suárez[3]

[1]Unit of Clinical Epidemiology and Biostatistics. Complexo Hospitalario Universitario de Santiago de Compostela, Spain

[2]Department of Mathematics. University of A Coruña, Spain

[3]Unit of Biotatistics. Department of Statistics and OR. University of Santiago de Compostela, Spain.

III Jornadas de Usuarios de R. 17-18 de Noviembre de 2011, Madrid


# Outline

- Receiver Operating Characteristic (ROC) curves and covariates.
- ROC regression methodologies.
- Computer-Aided Diagnostic (CAD) system to early detection of breast cancer.
- The `ROCRegression` package.

# Outline

- Receiver Operating Characteristic (ROC) curves and covariates.
- ROC regression methodologies.
- Computer-Aided Diagnostic (CAD) system to early detection of breast cancer.
- The `ROCRegression` package.

# Outline

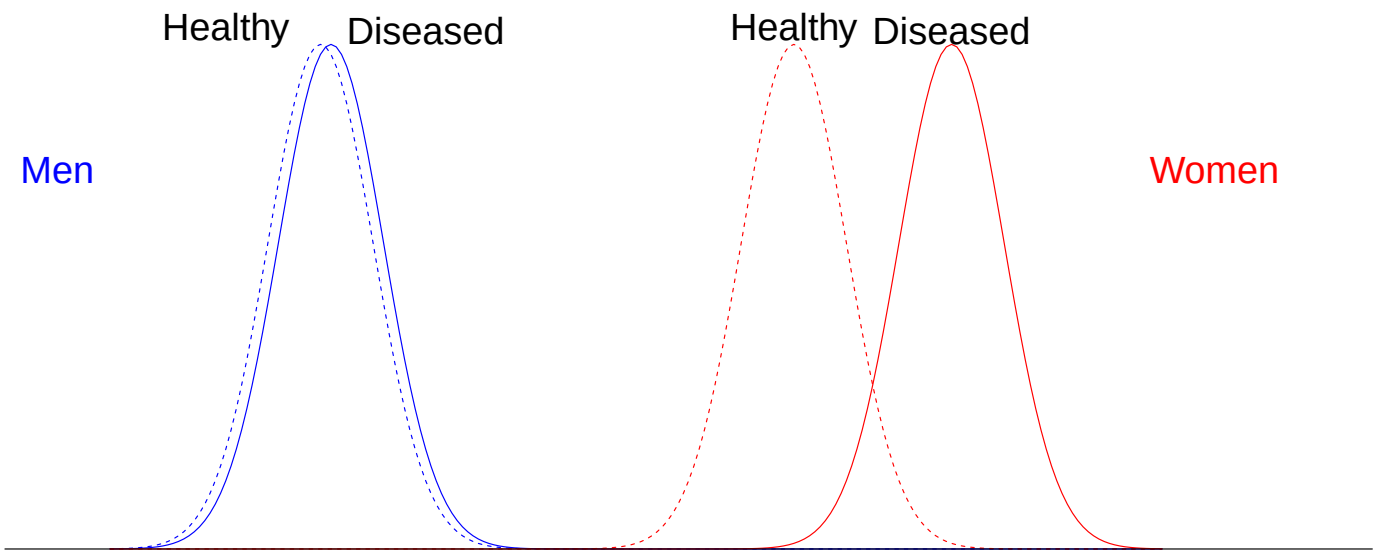
- Receiver Operating Characteristic (ROC) curves and covariates.
- ROC regression methodologies.
- Computer-Aided Diagnostic (CAD) system to early detection of breast cancer.
- The `ROCRegression` package.

# Outline

- Receiver Operating Characteristic (ROC) curves and covariates.
- ROC regression methodologies.
- Computer-Aided Diagnostic (CAD) system to early detection of breast cancer.
- The `ROCRegression` package.

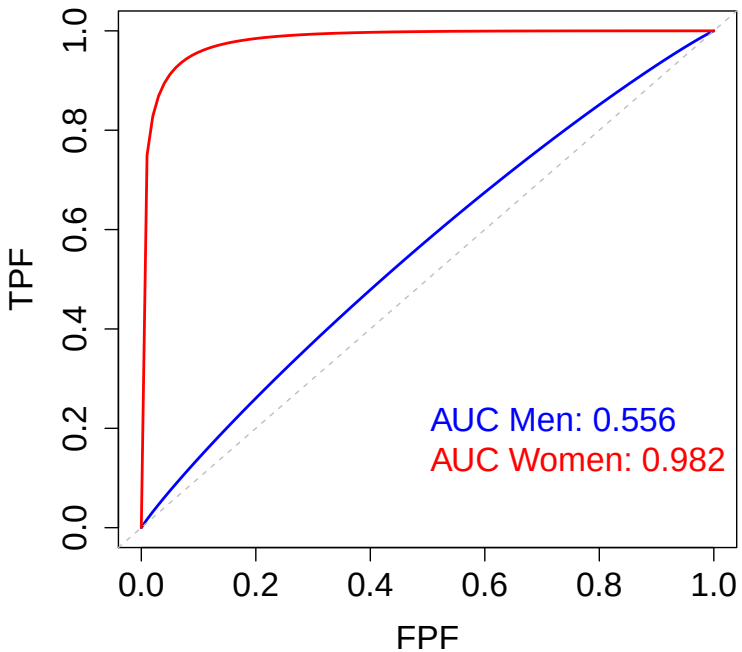
# ROC curve and covariates

- What is meant by "ROC curve and covariates"?




Density of the diagnostic test in diseased (solid line) and healthy (dashed line) in men (blue) and women (red).

# ROC curve and covariates




ROC curve in men (blue) and women (red).

- Among men the diagnostic test is almost uninformative about true disease status.
- In the case of women, however, the diagnostic test displays a very high discriminatory capacity.
- What then are the consequences of this from a practical point of view? Simply, that this diagnostic test should not be used for men.

# ROC curve and covariates

- In many practical situations the discriminatory capacity of a diagnostic test can be affected by a set of covariates (characteristics of the subject, variations on how the test is performed, ...).
- In such cases, interest should be focused on assessing the accuracy of the test according to the values of the covariates $\mathbf{X} = (X_1, \ldots, X_p)$. The **covariate-specific ROC** curve is defined as

$$ROC_{\mathbf{X}}(t) = S_{D\mathbf{X}}\left(S^{-1}_{\bar{D}\mathbf{X}}(t)\right), t \in (0,1),$$

where

$$S_{D\mathbf{X}}(y) = P[Y \geq y | D = 1, \mathbf{X}],$$
$$S_{\bar{D}\mathbf{X}}(y) = P[Y \geq y | D = 0, \mathbf{X}].$$

# ROC regression methodologies

How can we incorporate the information of the covariates in the ROC analysis?

Within the general regression framework:

- **Induced ROC methodology** (Pepe, 1998; Faraggi, 2003; Zheng and Heagerty, 2004)
- **Direct ROC methodology** (Pepe, 2000; Cai and Pepe, 2002; Alonzo and Pepe, 2002; Cai, 2004)

# Induced ROC methodology

The **induced ROC methodology** is based on specifying a regression model for the test result as a function of covariates, in both healthy and diseased populations

$$Y_{\bar{D}} = \mathbf{X}\boldsymbol{\beta}_{\bar{D}} + \sigma_{\bar{D}}\varepsilon_{\bar{D}},$$
$$Y_D = \mathbf{X}\boldsymbol{\beta}_D + \sigma_D\varepsilon_D.$$

From these regression models, the **induced covariate-specific ROC curve** is then computed

$$ROC_{\mathbf{X}}(t) = S_D\left(\mathbf{X}\left(\frac{\boldsymbol{\beta}_{\bar{D}} - \boldsymbol{\beta}_D}{\sigma_D}\right) + \frac{\sigma_{\bar{D}}}{\sigma_D}S_{\bar{D}}^{-1}(t)\right),$$

where $S_D$ and $S_{\bar{D}}$ are the survival functions of $\varepsilon_D$ and $\varepsilon_{\bar{D}}$ respectively.

# Approximations in the induced ROC methodology

We have implemented in the `ROCRegression` package different parametric/semiparametric proposals which **differ** in the assumptions made about the **distribution of errors** $\varepsilon_D$ and $\varepsilon_{\bar{D}}$.

- **Induced normal method (NM)** (Faraggi, 2003): the errors $\varepsilon_{\bar{D}}$ and $\varepsilon_D$ are considered to be **normally distributed**: $\varepsilon_{\bar{D}}, \varepsilon_D \sim N(0, 1)$.

- **Induced semiparametric method (SM)** (Pepe, 1998): **no assumptions on the distributions** of the errors $\varepsilon_{\bar{D}}$ and $\varepsilon_D$ are made.

# Direct ROC methodology

In contrast to the induced ROC methodology, in the **direct ROC methodology** the effect of covariates is evaluated directly on the ROC curve.
These effects are modelled by a Generalized Linear Model (GLM)

$$ROC_{\mathbf{X}}(t) = g\left(\mathbf{X}\boldsymbol{\beta} + h(t)\right),$$

where

- $\boldsymbol{\beta}$ quantify the effects of the covariates on the ROC curve (**unknown**).
- $h(\cdot)$ represents the effect of the FPFs ($t$) on the TPFs ($ROC(t)$) (**unknown**).
- $g^{-1}(\cdot)$ is a **known** link function (e.g. logit or probit).

The above model is denoted as an **ROC-GLM regression model**.

# Approximations in the direct ROC methodology

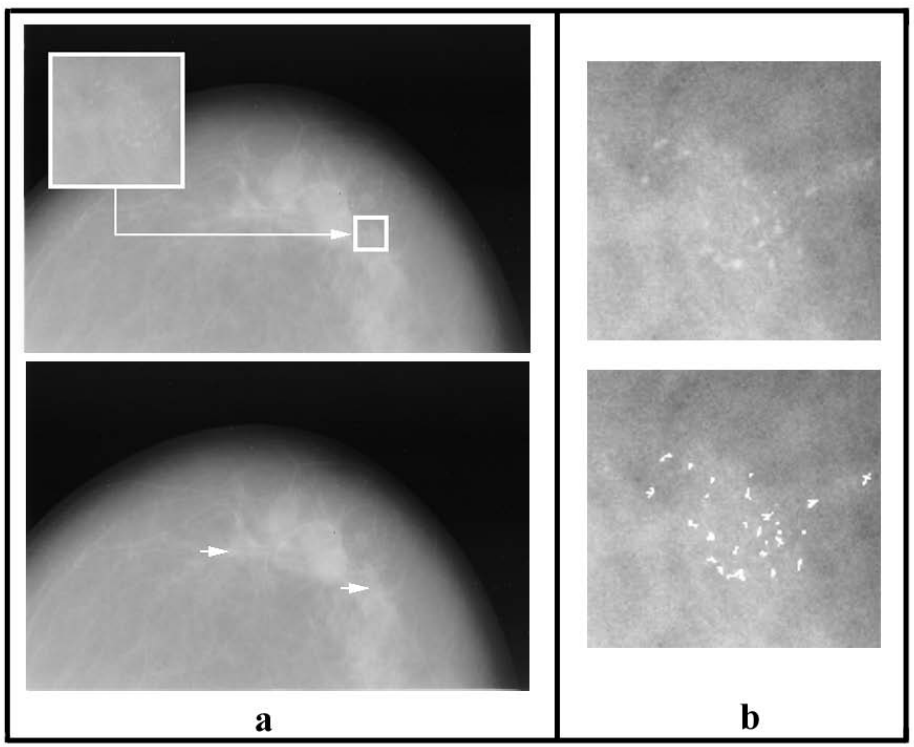
Different proposals have been suggested in the literature, which differ in the assumptions made about funcion of the FPFs, $h(\cdot)$:

- **Parametric ROC-GLM (PROCGLM)** (Pepe,2000; Alonzo and Pepe, 2002). A **parametric form** for $h(\cdot)$ is assumed

$$h(t) = \sum_{k=1}^{K} \alpha_k h_k(t),$$

  where $h_1, \ldots, h_k$ are **known** funtions.
- **Semiparametric ROC-GLM (SROCGLM)** (Cai and Pepe, 2002; Cai 2004). The function $h(\cdot)$ remains **completely unspecified**.

# Automated detection of clustered microcalcifications on digital mammograms




(a) Original mammogram containing a cluster of microcalcifications (zoomed), and the results of the CAD system (white arrows): true cluster and false detection; and (b) original cluster; and detection of the cluster.

- Breast cancer is one of the main causes of death among women, but an early detection can considerably reduce the mortality rates.
- Computer-Aided Diagnostic (CAD) systems, dedicated to the detection of lesions, are usually used to help radiologists in the interpretation of mammograms.
- A CAD system produces, as a result, suspicious areas that can be recognized as true lesions or false detections.

# Automated detection of clustered microcalcifications on digital mammograms

- A fundamental aspect of any computerized scheme is the **reduction** of the false-detection rate. To this aim, several statistical methods, like the ROC curve, are used.
- For the analysis that will be presented in this talk, the **diagnostic test** (or marker ($Y$)) considered was the ratio of the cluster **size** to the mean distance between microcalcifications of each cluster detected on the digital mammograms.
- The objective is to evaluate the **discriminatory capacity** of this **size** in distinguishing **true clustered microcalcifications** (diseased population) and **false detections** (healthy population).
- The radiologists suspect that the accuracy of this marker can be affected by:
  - The breast **tissue** type: dense or fatty ($X_1$).
  - The ratio of the cluster average **grey level** to that of the image ($X_2$).

# The ROCRegression package

- So far, the scarcity of implemented ROC regression software is probably responsible for these models' lack of popularity in the medical community.
- We have implemented the methods presented in this talk in an R (R Development Core Team, 2011) package, named ROCRegression
- The implementation of the ROCRegression package has been done in a similar fashion to other regression funtions/packages:
  - ROCreg
  - print.ROCreg and summary.ROCreg
  - predict.ROCreg
  - plot.ROCreg.

# ROCreg function: example

- The `ROCreg()` function fits a **ROC regression model** with a vector of continuous and/or categorical covariates, and their possible interactions.

- For the CAD system, the `ROCRegression` package was used to fit a **Semiparametric ROC-GLM** including the 'grey level-by-tissue type' interaction

$$ROC_{(GL,TT)}(t) = \Phi\left(\beta_1 I_{\{TT=Fatty\}} + \beta_2 GL + \beta_3 GL \times I_{\{TT=Fatty\}} + h(t)\right).$$

- The syntax in this case is as follows:

```
R > fit.CAD <- ROCreg(method="SROCGLM", model=~tissue*greyLevel,
+ marker = "size", group = "lesion", tag.healthy = 0,
+ se.fit = TRUE, data=radio)
```

# ROCreg function: example

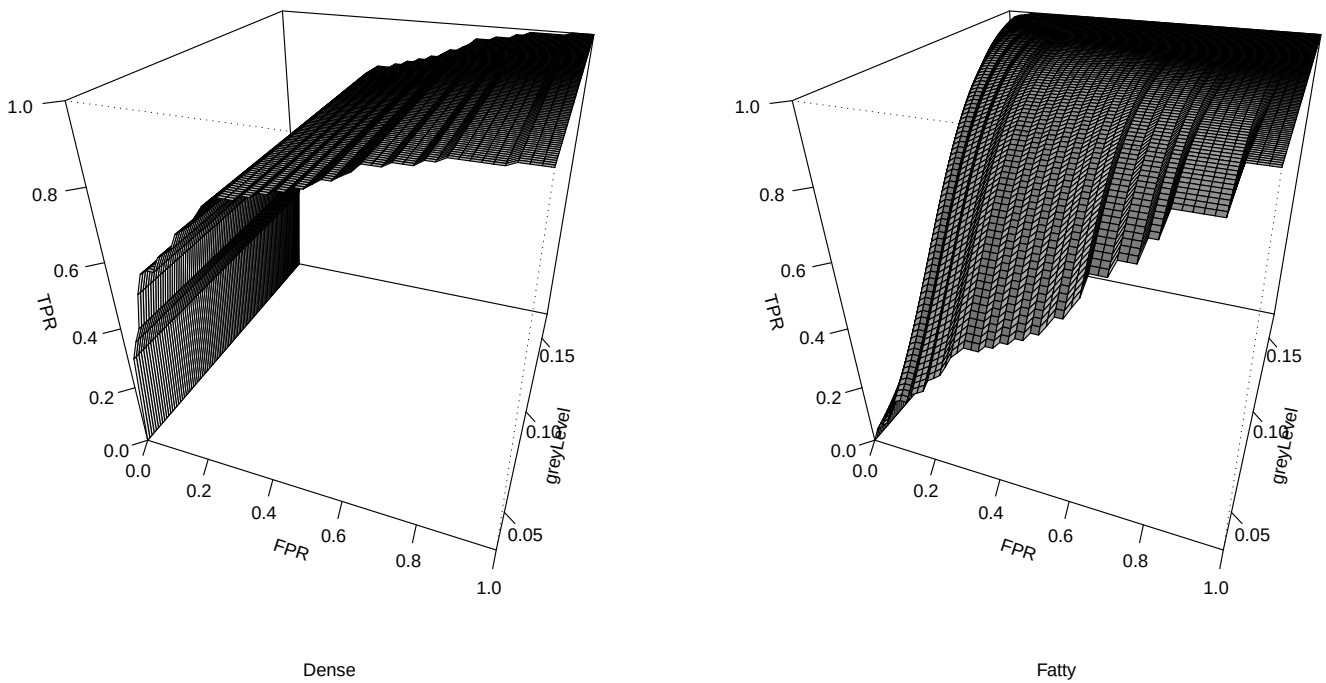
```
R > summary(fit.CAD)

Call:
ROCreg(method = "SROCGLM", model = ~tissue * greyLevel, marker = "size",
group = "lesion", tag.healthy = 0, se.fit = TRUE, data = radio)


************************************************
Semiparametric ROC-GLM Method
************************************************

ROC Coefficients:
---------------

                        Estimate Std. Error 95% Conf. Interval p-value
tissueFatty             -2.0883  0.8974     (-3.8472,-0.3294)  0.0200
greyLevel               0.4620   5.2876     (-9.9014,10.8254)  0.9304
`tissueFatty:greyLevel` 27.4304  13.7866    (0.4091,54.4517)   0.0466
```

# `plot` function: example

The `plot()` function plots the **ROC curve**, and, optionally, **AUC**, **Youden Index** and **optimal treshold** (cut-off) based on the Youden Index, from a `ROCreg` object. The suitable type of graphic is chosen according to the number and nature of the covariates.

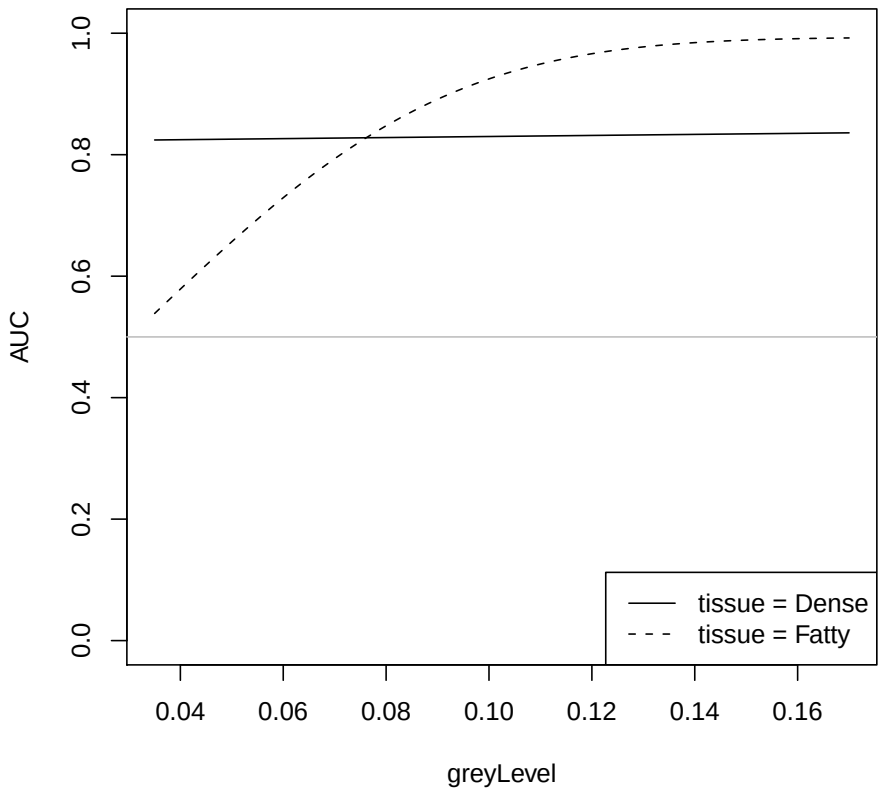
# plot function: example

```
R > plot(fit.CAD, accuracy="AUC")
```




Estimated ROC surfaces according to grey level, for dense (left) and fatty (right) tissue.

# `plot` function: example




Estimated AUC according to grey level, for dense (solid line) and fatty (dashed line) tissue.

**Gracias por vuestra atención!**