Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 21 - Multiple, non-linear, and logistic regression

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

Simple linear regression model

SIMPLE LINEAR REGRESSION MODEL. In a *simple linear regression model* for a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we assume that $x_1, x_2, \ldots, x_n$ are nonrandom and that $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \ldots, n,$$

where $U_1, \ldots, U_n$ are *independent* random variables with $\mathrm{E}[U_i] = 0$ and $\mathrm{Var}(U_i) = \sigma^2$.

- *Regression line*: $y = \alpha + \beta x$ with *intercept* α and *slope* β
- Least Square Estimators: $\hat{\alpha}$ and $\hat{\beta}$ and $\hat{\sigma}^2$
- Unbiasedness: $E[\hat{\alpha}] = \alpha$ and $E[\hat{\beta}] = \beta$ and $E[\hat{\sigma}^2] = \sigma^2$
- *Standard errors* (estimates of $\sqrt{Var(\hat{\alpha})}$ and $\sqrt{Var(\hat{\beta})}$):

$$se(\hat{\alpha}) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}_n^2}{SXX}\right)} \qquad se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{SXX}}$$

Standard error of fitted values (prediction $\pm$ standard error)

- For a given x_0, the the estimator $\hat{Y} = \hat{\alpha} + \hat{\beta}x_0$ has expectation

$$E[\hat{Y}] = E[\hat{\alpha}] + E[\hat{\beta}]x_0 = \alpha + \beta x_0$$

- Hence, $\hat{Y}$ is unbiased, and $\hat{y} = \hat{\alpha} + \hat{\beta}x_0$ is the best estimate for the fitted value at x_0
- Variance of $\hat{Y}$ is: *[See sdsln.pdf Chpt. 2]*

$$Var(\hat{Y}) = \sigma^2\left(\frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right)$$

- The *standard error* of the fitted value is then the estimate:

$$se(\hat{y}) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right)}$$

where

$$SXX = \sum_1^n (x_i - \bar{x}_n)^2 \qquad \hat{\sigma}^2 = \frac{1}{n-2}\sum_1^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

- Prediction uncertainty at x_0 is reported as $\hat{y} \pm se(\hat{y})$

See R script

Weighted Least Squares and simple polynomial regression

- Weighted Simple Regression

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 w_i$$

  - w_i is the weight (or importance) of observation (x_i, y_i)
  - For natural number weights, it is the same as replicating instances

- Polynomial Simple Regression

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_i - \beta_2 x_i^2 - \ldots - \beta_k x_i^k)^2$$

  - $Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + U_i$ for $i = 1, 2, \ldots, n$
  - May suffer from collinearity (see later in this slides)

See R script

Non-linear simple regression and transformably linear functions

- $Y_i = f(\alpha, \beta, x_i) + U_i$ for $i = 1, 2, \ldots, n$ for a non-linear function $f()$

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - f(\alpha, \beta, x_i))^2$$

- $\mathit{arg}\,\min_{\alpha,\beta} S(\alpha, \beta)$ may be without a closed form
  - use numeric search of the minimum (which may fail to find it!), e.g., gradient descent
- Some $f()$ can be favourably transformed, e.g., $f(\alpha, \beta, x_i) = \alpha x_i^{\beta}$ (recall Power law, Zipf's)
- Solve $\log Y_i = \log \alpha + \beta \log x_i + U_i$ *[Linearization]*
- Let $\log \hat{\alpha}$ and $\hat{\beta}$ be the LSE estimators. By exponentiation:

$$Y_i = \hat{\alpha} x_i^{\hat{\beta}} e^{U_i}$$

where the error term is a multiplicative factor

See R script

Multiple linear regression

- Multivariate dataset of observations:

$$(x_1^1, x_1^2, \dots, x_1^k, y_1), \dots, (x_n^1, x_n^2, \dots, x_n^k, y_n)$$

- $Y_i = \alpha + \beta_1 x_i^1 + \dots + \beta_k x_i^k + U_i$
- In vector terms:
  - $Y_i = \boldsymbol{x}_i \cdot \boldsymbol{\beta}^T + U_i$, where $\boldsymbol{\beta} = (\alpha, \beta_1, \dots, \beta_k)$ and $\boldsymbol{x}_i = (1, x_i^1, \dots, x_i^k)$ the i^{th} observation
  - $\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta}^T + \boldsymbol{U}$, where $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, $\boldsymbol{U} = (U_1, \dots, U_n)$, and $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$
- Ordinary Least Square Estimation (OLS):

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i \cdot \boldsymbol{\beta}^T)^2 = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}^T\|^2 \qquad \hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = (\boldsymbol{X}^T \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^T \cdot \boldsymbol{y}$$

  where $\boldsymbol{y} = (y_1, \dots, y_n)$ and $\|(v_1, \dots, v_n)\| = \sqrt{\sum_{i=1}^n v_i^2}$ is the Euclidian norm
- Meaning of β_i: change of Y due to a unit change in x_i all the x_j with $j \neq i$ unchanged!
- It is a Minimum Variance linear Unbiased Estimator [**Gauss-Markov Thm.**]

See R script

Multivariate linear regression

- The multivariate linear model accommodates two or more dependent variables

$$\boldsymbol{Y} = \boldsymbol{X}\beta^T + \boldsymbol{U}$$

  where
  - $\boldsymbol{Y}$ is $n \times m$: n observations, m dependent variables
  - $\boldsymbol{X}$ is $n \times (k+1)$: n observations, k independent variables $+1$ constants
  - β^T is $(k+1) \times m$: parameters for each of the m dependent variables
  - $\boldsymbol{U}$ is $n \times m$: n observations, m error terms
- It is **not** just a collection of m multiple linear regressions
- Errors in rows (observations) of $\boldsymbol{U}$ are independent, as in a single multiple linear regression
- Errors in columns (dependent variables) are allowed to be correlated.
  - E.g., errors of plasma level and amitriptyline due to usage of drugs
  - Hence, coefficients from the models covary!

See R script

Other variants and generalizations

- Heteroscedastic linear models
  - Relax the assumption of equal variances $Var(U_i) = \sigma^2$
- **Generalized least squares**
  - $U_1, \ldots, U_n$ not necessarily independent
- **Hierarchical linear models**
  - Nested or cluster organization (e.g., Children within classrooms within schools)
  - See **this intro in R**
- Generalized linear models
  - We'll see next at Logistic Regression
- **Tobit regression**
  - Censored dependent variable, e.g., income cannot be negative
- **Truncated regression model**
  - Dependent variable not available/sampled, e.g., income above a poverty threshold
- **Quantile regression**
  - Estimate of the median (or other quantiles) instead of the mean, as in regression

Issues: Omitted variable bias

- Suppose we omit a variable z_i that belongs to the true model

$$Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + U_i$$

  with $\beta_2 \neq 0$ (i.e., Y is determined by Z)
  - Under-specification of the model, due to lack of data
- Fitted model $Y_i = \alpha + \beta_1 x_i + U_i'$
  - Hence, $E[U_i'] = E[\beta_2 z_i + U_i] = \beta_2 z_i + E[U_i] = \beta_2 z_i \neq 0$
- Let $\hat{\alpha}$ and $\hat{\beta}_1$ be the LSE estimators of the fitted model:

$$E[\hat{\beta}_1] = \beta_1 + \beta_2 \delta \qquad Bias(\hat{\beta}_1) = \beta_2 \delta$$

  where δ is the slope of the regression of $z_i = \gamma + \delta x_i + U_i''$, i.e.:

$$\delta = r_{xz} \frac{s_z}{s_x}$$

- $Bias(\hat{\beta}_1) \neq 0$ if X and Z correlated

See R script

Issues: Multi-collinearity and variance inflation factors

- *Multicollinearity*: two or more independent variables (regressors) are strongly correlated.
- $Y_i = \alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + U_i$
- It can be shown that for $j \in \{1, 2\}$:

$$Var(\hat{\beta}_j) = \frac{1}{(1 - r^2)} \cdot \frac{\sigma^2}{SXX_j}$$

  where $r = cor(x^1, x^2)$, $\sigma^2 = Var(U_i)$ and $SXX_j = \sum_1^n (x_i^j - \bar{x}_n^j)^2$
- Correlation between regressors increases the variance of the estimators
- In general, for more than 2 variables:

$$Var(\hat{\beta}_j) = \frac{1}{(1 - R_j^2)} \cdot \frac{\sigma^2}{SXX_j}$$

  where R_j^2 is the coefficient of determination (R^2) in the regression of x_j from all other x_i's.
- The term $^1/_{(1-R_j^2)}$ is called *variance inflation factor*

See R script

Variable selection

- Recall: when $U_i \sim N(0, \sigma^2)$, we have $Y_i \sim N(\boldsymbol{x}_i \cdot \boldsymbol{\beta}, \sigma^2)$, hence we can apply MLE
- Log-likelihood is $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \boldsymbol{x}_i \cdot \boldsymbol{\beta}}{\sigma^2}\right)^2}\right)$
- Akaike information criterion (AIC), balances model fit against model simplicity

$$AIC(\boldsymbol{\beta}) = 2|\boldsymbol{\beta}| - 2\ell(\boldsymbol{\beta})$$

- stepAIC(model, direction="backward") algorithm
  1. $S = \{x^1, \dots, x^k\}$
  2. $b = AIC(S)$
  3. repeat
     - 3.1 $x = arg\, \min_{x \in S} AIC(S \setminus \{x\})$
     - 3.2 $v = AIC(S \setminus \{x\})$
     - 3.3 if $v < b$ then $S, b = S \setminus \{x\}, v$
  4. until no change in S
  5. return S

See R script

Regularization methods: Ridge/Tikhonov

$$\hat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

- Ordinary Least Square Estimation (OLS):

$$S(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2$$

where $\|(v_1, \ldots, v_n)\| = \sqrt{\sum_{i=1}^{n} v_i^2}$ is the Euclidian norm

  - Performs poorly as for prediction (overfitting) and interpretability (number of variables)

- Ridge regression:

$$S(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2$$

where $\|\boldsymbol{\beta}\| = \sqrt{\alpha^2 + \sum_{i=1}^{k} \beta_i^2}$.

  - Notice that λ_2 is not in the parameters of the minimization problem!
  - Variables with minor contribution have their coefficients **close** to zero
  - It improves prediction error by reducing overfitting through a bias-variance trade-off
  - It is **not** a parsimonious method, i.e., does not reduce features

Regularization methods: Lasso and Penalized

- Lasso (Least Absolute Shrinkage and Selection Operator) regression:

$$S(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

  where $\|\boldsymbol{\beta}\|_1 = |\alpha| + \sum_{i=1}^{k} |\beta_i|$.
  - Notice that λ_1 is not in the parameters of the minimization problem!
  - Variable with minor contribution have their coefficients **equal** to zero
  - It improves prediction error by reducing overfitting through a bias-variance trade-off
  - It **is** a parsimonious method, i.e., it reduces the number of features
- Penalized linear regression:

$$S(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

  - Both Ridge and Lasso regularization parameters
- How to solve the minimization problems? **Lagrange multiplier method** or **reduction to Support Vector Machine** learning
- How to find the best λ_1 and/or λ_2? Cross-validation!

See R script

Towards logistic regression

- Consider a bivariate dataset
$$(x_1, y_1), \dots, (x_n, y_n)$$
where $y_i \in \{0, 1\}$, i.e., Y_i is a binary variable
- Using directly linear regression:
$$Y_i = \alpha + \beta x_i + U_i$$
results in poor performances (R^2)

See R script

Towards logistic regression

- Consider a bivariate dataset
$$(x_1, y_1), \ldots, (x_n, y_n)$$
where $y_i \in \{0, 1\}$, i.e., Y_i i binary variable
- Group by x values:
$$(d_1, f_1), \ldots, (d_m, f_m)$$
where $d_1, \ldots, d_m$ are the distinct values of $x_1, \ldots, x_n$ and f_i is the fraction of 1's:
$$f_i = \frac{|\{j \in [1, n] \mid x_j = d_i \wedge y_j = 1\}|}{|\{j \in [1, n] \mid x_j = d_i\}|}$$
and the linear model (we continue using x_i but it should be d_i):
$$F_i = \alpha + \beta x_i + U_i$$

See R script

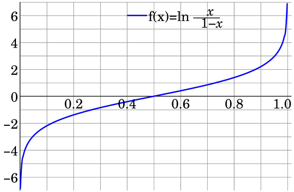
Towards logistic regression

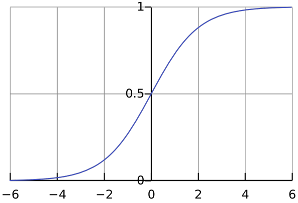
- Rather than f_i, we model the logit of f_i

$$logit(F_i) = \alpha + \beta x_i + U_i$$

where logit and its inverse (**logistic function**) are:

$$logit(p) = \log \frac{p}{1-p} \qquad inv.logit(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$





- Why?
  - $F_i \in [0,1]$ while the RHS is in $\mathbb{R}$
  - Relation between RHS and F_i is typically sigmoidal, not linear

See R script

Logistic regression and generalized linear models

- Since Y_i's are binary, $F_i = P(Y_i = 1|X = x_i) \sim Ber(f_i)$, and U_i is not necessary
$$logit(F_i) = \alpha + \beta x_i$$
and then $F_i = P(Y_i = 1|X = x_i) = inv.logit(\alpha + \beta x_i) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$
- Since $F_i/(1 - F_i) = e^{\alpha+\beta x_i}$, β can be interpreted as:
  - the expected change in log odds of having the outcome per unit change in X
  - e.g., $\beta = 0.38$ in predicting heart disease from smoking: the smoking group has $e^{\beta} = 1.46$ times the odds of the non-smoking group of having heart disease
  - e.g., $\alpha = -1.93$ means the probability a non-smoker has heart disease is $e^{\alpha}/(1 + e^{\alpha}) = 0.13$.
- Generalized linear models: family = distribution + link function
  - E.g., Binomial + logit for logistic regression
  - For $Y_i \in \{0, 1\}$, actually Bernoulli + logit *[Binary logistic regression]*
- Since distribution is known, MLE can be adopted for estimating α and β:
$$\ell(\alpha, \beta) = \sum_{i=1}^{n} [y_i \log(inv.logit(\alpha + \beta x_i)) + (1 - y_i) \log(1 - inv.logit(\alpha + \beta x_i))]$$

See R script

Elastic net logistic regression

- Penalized linear regression minimizes:

$$\|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

  - $\lambda_1 = 0$ is the Ridge penalty
  - $\lambda_2 = 0$ is the Lasso penalty
- Elastic net regularization for logistic regression minimizes:

$$-\ell(\boldsymbol{\beta}) + \lambda \left(\frac{(1-\alpha)}{2} \|\boldsymbol{\beta}\|^2 + \alpha \|\boldsymbol{\beta}\|_1 \right)$$

  - $\alpha = 0$ is the Ridge penalty
  - $\alpha = 1$ is the Lasso penalty
  - λ is to be found, e.g., by cross-validation

See R script