CLASSIFICATION

Introducción a la Ciencia de Datos

Some of the figures in this presentation are taken from: An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Some slides are based on Abbass Al Sharif's slides for his course DSO 530: Applied Modern Statistical Learning Techniques.

LOGISTIC REGRESSION

Classification Methods

Credit Card Default Data

- Data for 10,000 credit card users.
- We would like to be able to predict customers that are likely to default.
- The goal is to fit a model such that the relevant predictors of credit card default are elucidated given these variables.

default.qmd

default	student	balance	income
<fct>	<fct>	<dbl>	<dbl>
No	No	729.5264952	44361.6251
No	Yes	817.1804066	12106.1347
No	No	1073.5491640	31767.1389
No	No	529.2506047	35704.4939
No	No	785.6558829	38463.4959
No	Yes	919.5885305	7491.5586
No	No	825.5133305	24905.2266
No	Yes	808.6675043	17600.4513
No	No	1161.0578540	37468.5293
No	No	0.0000000	29275.2683

1-10 of 10,000 rows Previous 1 2 3 4 5 6 ... 1000 Next

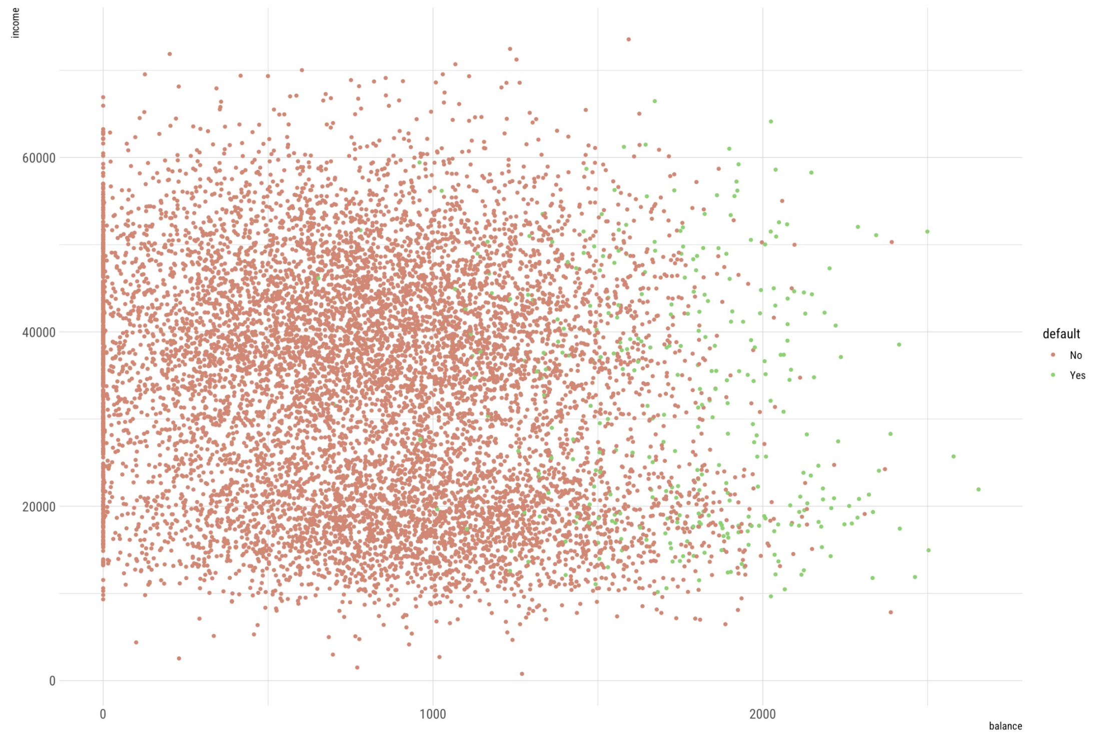
Credit Card Default Data

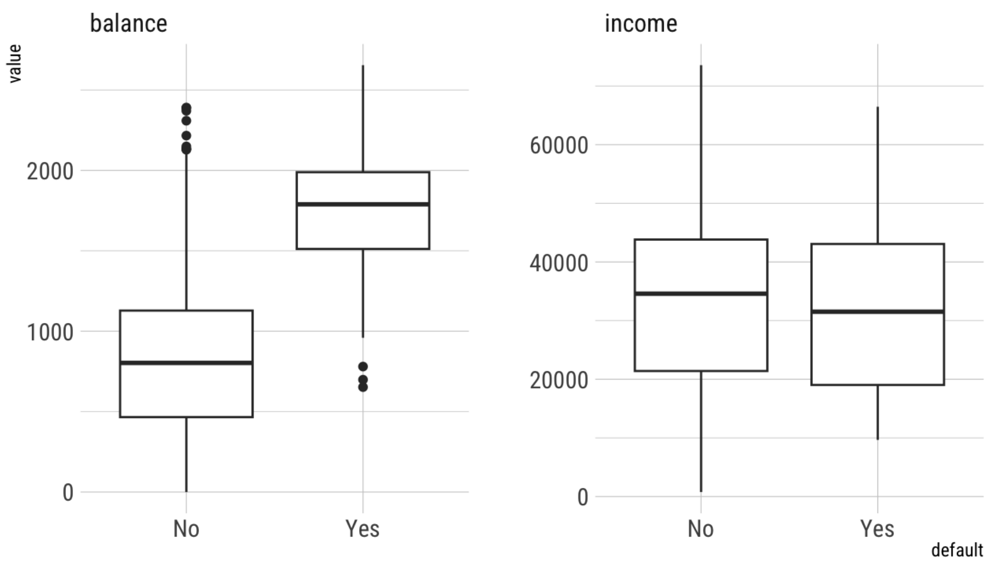
- Predictor variables (X) present in the default data set are:
  - student: A binary factor (Yes/No) containing whether or not a given credit card holder is a student.
  - income: The gross annual income for a given credit card holder.
  - balance: The total credit card balance for a given credit card holder.

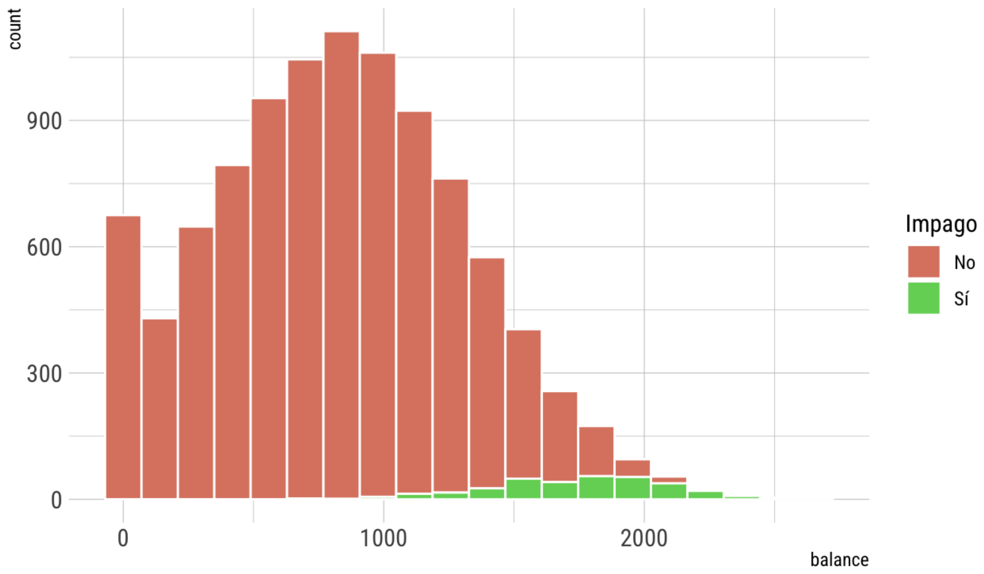
- The target variable (Y), namely default, is categorical: A binary factor (Yes/No) containing whether a given user has defaulted on his/her credit card.

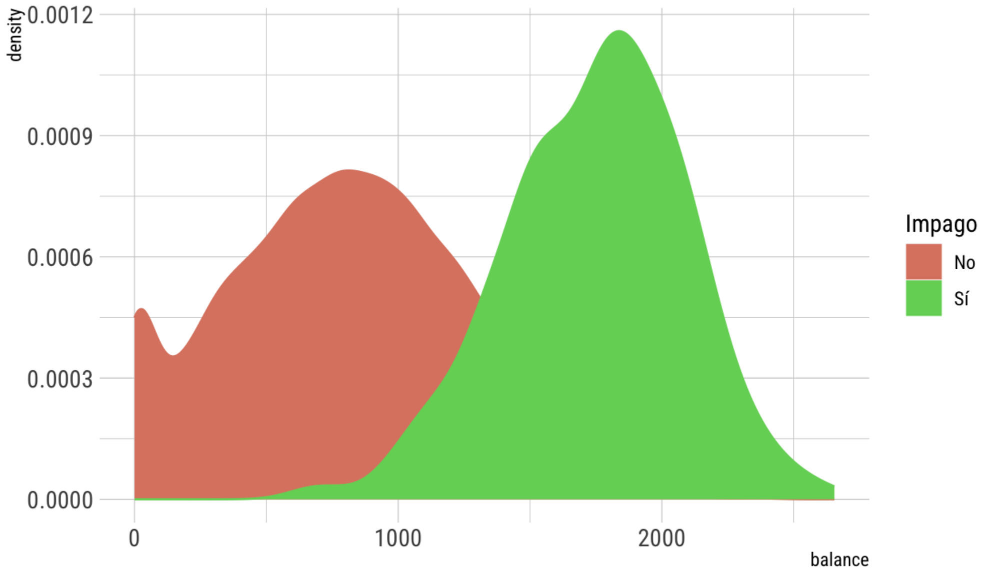
- How do we model the relationship between Y and X?

The default Dataset

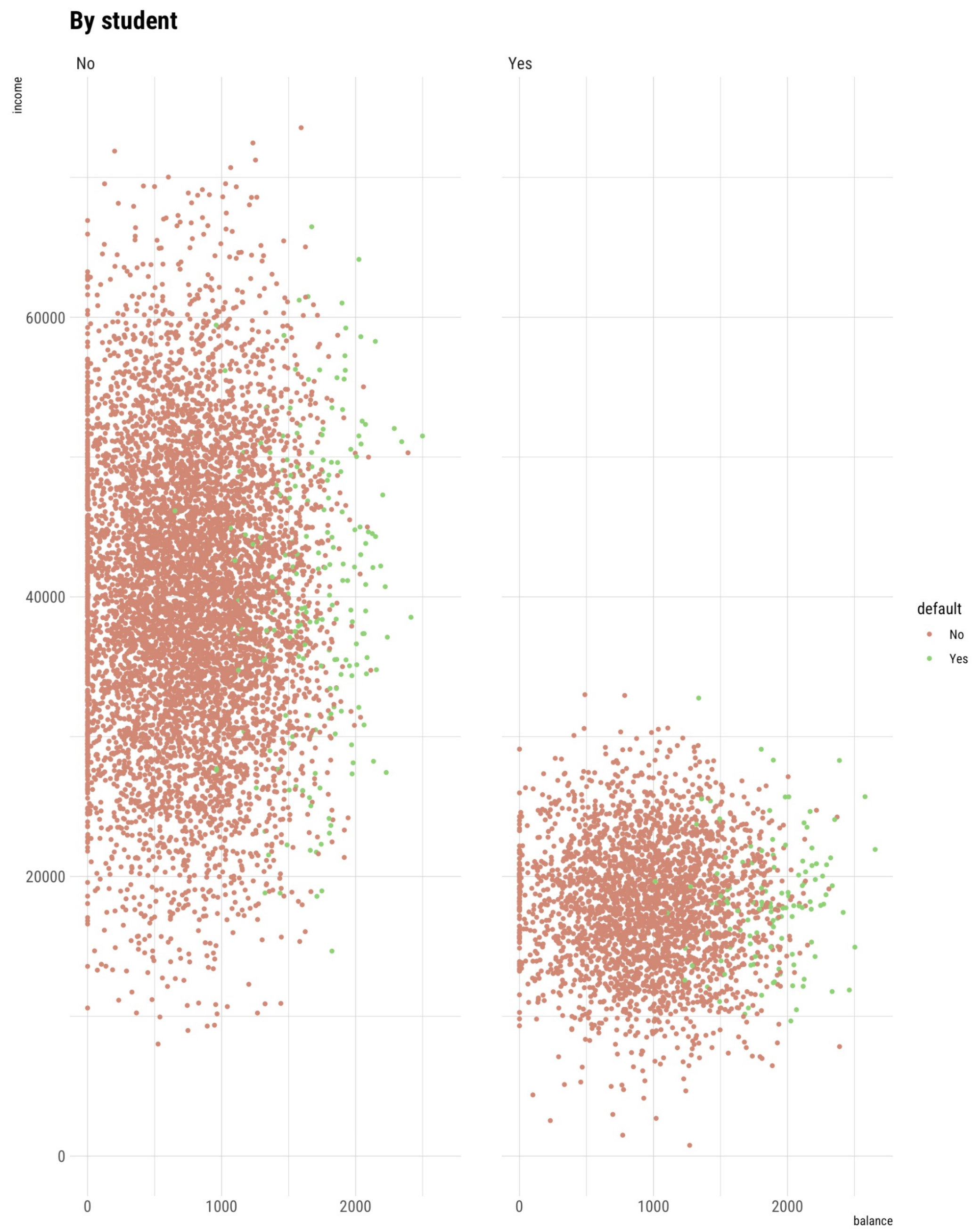


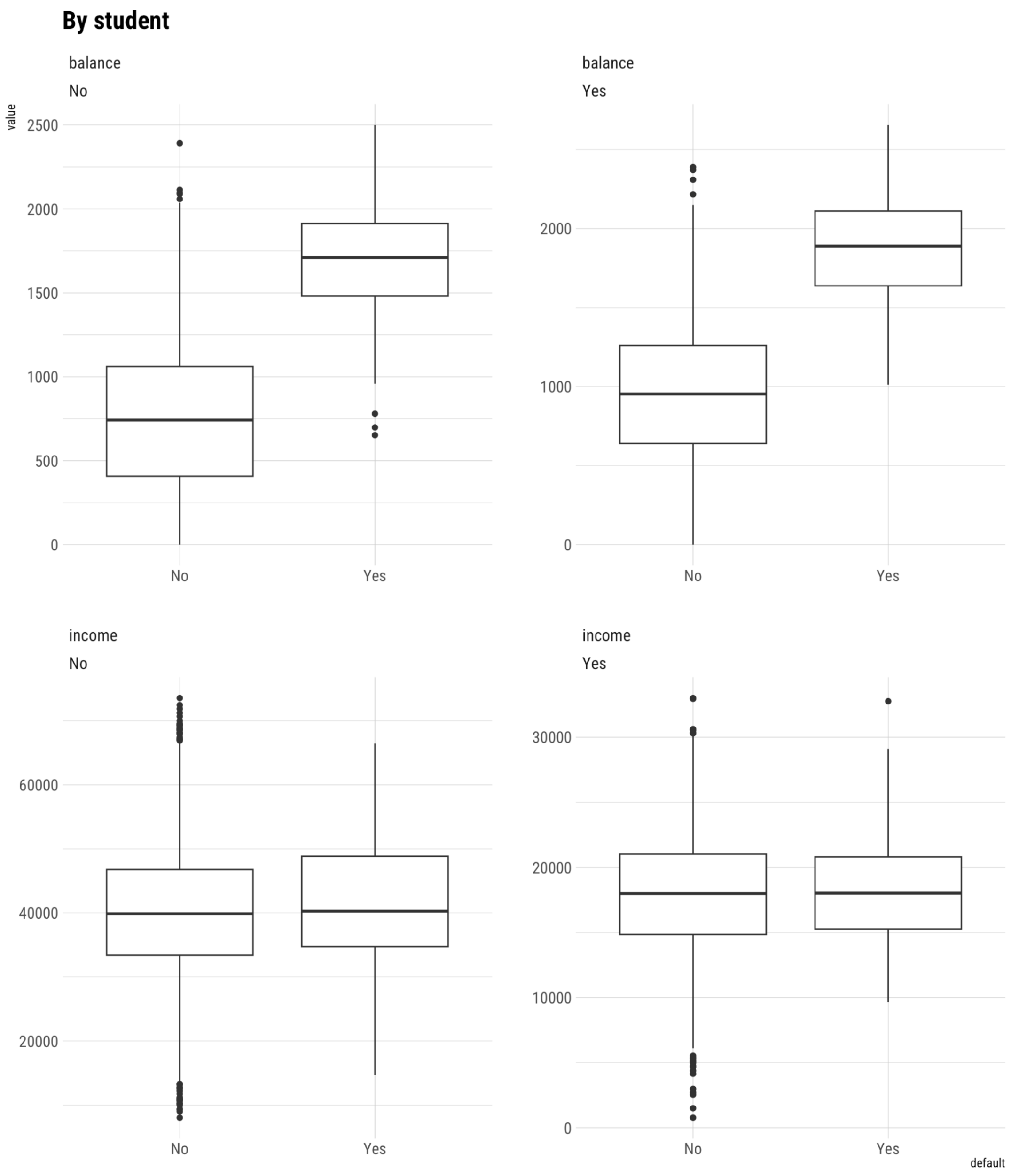






The default Dataset





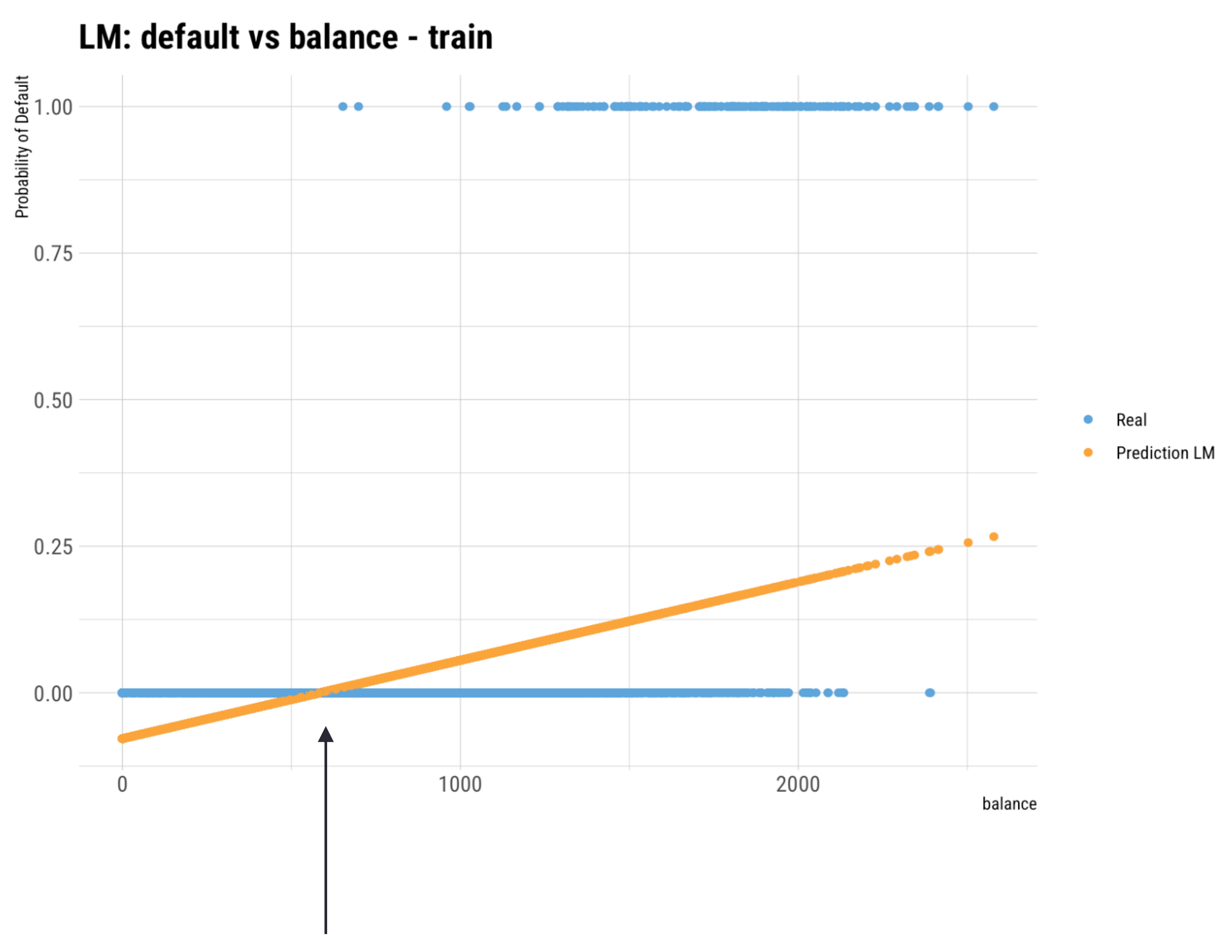
Logistic Regression (logit model)

- Rather than modeling the response *Y* (Yes or No) directly, logistic regression models the **probability** that *Y* belongs to a particular category.

- For any given value of balance, a prediction can be made for the probability of default.

Why not Linear Regression?

- If we fit a linear regression to the Default data, then for very low balances we predict a negative probability, and for high balances we predict a probability above 1!

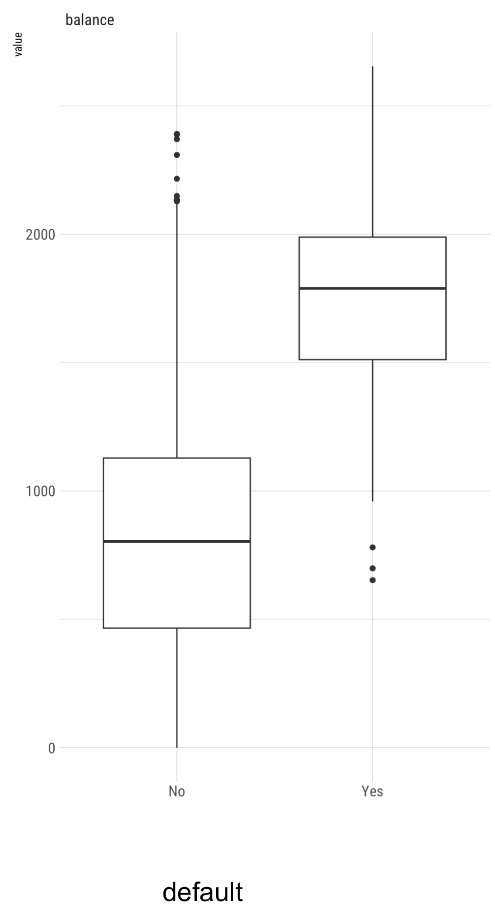


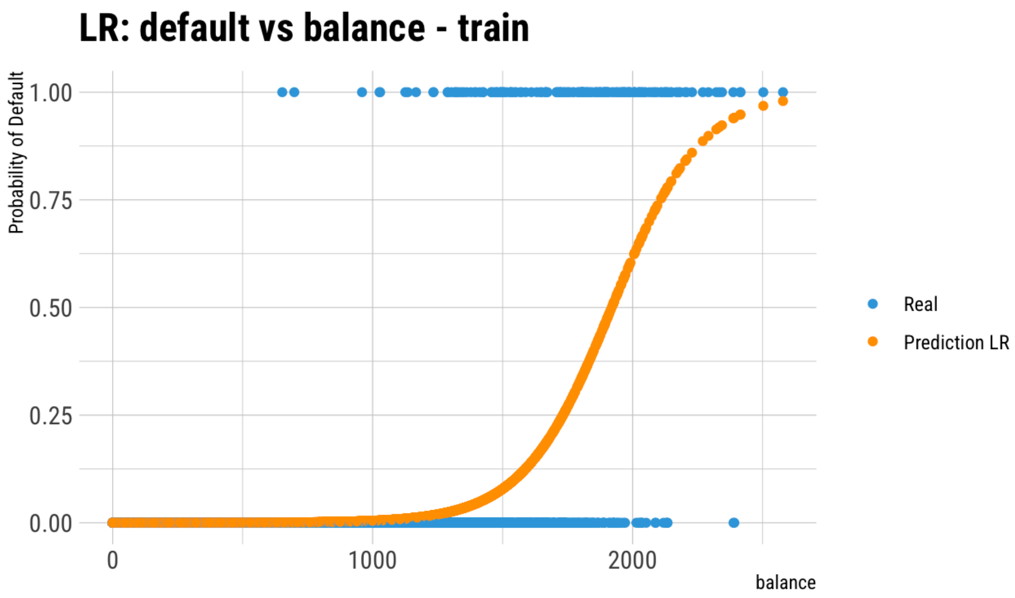
When Balance < ~500,

P(default) is negative!

Logistic Function on Default Data

- The probability of default is close to, but not less than zero for low balances. And close to but not above 1 for high balances.





The Logistic Model

- In logistic regression we use the logistic function:

$$P(Y|X=x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

- We come up with $\hat{\beta}_0$ and $\hat{\beta}_1$ to estimate β_0 and β_1

The logit function

- We assume a linear relationship between the predictor variables and the log-odds ratio (relación logarítmica de probabilidades) of the event (Default).

$$\hat{p} = P(Y|X)$$

$$\beta_o + \beta_1 X = ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$$

The logic model

- Calculate the inverse function:

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 X$$

$$\frac{\hat{p}}{1-\hat{p}} = e^{\beta_0+\beta_1 X} \rightarrow \hat{p} = e^{\beta_0+\beta_1 X}(1-\hat{p}) \rightarrow \hat{p} = e^{\beta_0+\beta_1 X} - e^{\beta_0+\beta_1 X}\hat{p}$$

$$\rightarrow \hat{p} + e^{\beta_0+\beta_1 X}\hat{p} = e^{\beta_0+\beta_1 X} \rightarrow \hat{p}(1+e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X}$$

$$\rightarrow \hat{p} = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Interpreting β_1

- Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting *P(Y)* and not *Y*.
  - If $\beta_1 = 0$, this means that there is no relationship between *Y* and *X*.
  - If $\beta_1 > 0$, this means that when *X* gets larger so does the probability that $Y = 1$.
  - If $\beta_1 < 0$, this means that when *X* gets larger, the probability that $Y = 1$ gets smaller.
- But how much bigger or smaller depends on where we are on the slope.

Are the coefficients significant?

- We still want to perform a **hypothesis test** to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a Z test instead of a T test, but that doesn't change the way we interpret the p-value.
- Here the p-value for balance is very small, and β_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Making Prediction

- Suppose an individual has an average balance of \$1000. What is their probability of default?

$$P(Y=1|X=1000)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}=\frac{e^{-10.6513+0.0055\times1000}}{1+e^{-10.6513+0.0055\times1000}}=0.00576$$

- The predicted probability of default for an individual with a balance of \$1000 is less than 1%

- For a balance of \$2000, the probability is much higher, and equals to 0.586 (58.6%)

Qualitative Predictors in Logistic Regression

- We can predict if an individual default by checking if he/she is a student or not. Thus we can use a qualitative variable "Student" coded as (Student = 1, Non-student = 0)
- $\hat{\beta}_1$ is positive: This indicates students tend to have higher default probabilities than non-students

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

$$P(Y = 1|X = 1) = \frac{e^{-3.5041+0.4049\times 1}}{1 + e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$P(Y = 1|X = 0) = \frac{e^{-3.5041+0.4049\times 0}}{1 + e^{-3.5041+0.4049\times 0}} = 0.0292$$

Multiple Logistic Regression

- We can fit multiple logistic just like regular regression

$$P(Y|X_1 \dots X_p) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Multiple Logistic Regression - Default Data

- Predict Default using:
  - Balance (quantitative)
  - Income (quantitative)
  - Student (qualitative)

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
```

Predictions

- A student with a credit card balance of \$1,500 and an income of \$40,000 has an estimated probability of default:

$$P(X) = \frac{e^{-10.859+0.00574\times 1500+0.003\times 40000-0.6468\times 1}}{1+e^{-10.859+0.00574\times 1500+0.003\times 40000-0.6468\times 1}} = 0.058$$

An Apparent Contradiction!

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

Positive

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
```

Negative

To whom should credit be offered?

- A student is risker than non students if no information about the credit card balance is available

- However, that student is less risky than a non-student with the same credit card balance!

Logistic regression (ordinal logit model)

- The ordered logit model is a regression model for an ordinal response variable:
  - Low, Medium, High
  - Bad, Regular, Good, Very Good

- The model is based on the cumulative probabilities of the response variable

- The logit of each cumulative probability is assumed to be a linear function with regression coefficients constant across response categories

Logistic regression (ordinal logit model)

- Let the response be $Y = 1,2, \dots, J$ where the ordering is natural.
- The associated probabilities are $\pi_1, \pi_2, \dots, \pi_J$, and a cumulative probability of a response less than equal to *j* is:

$$P(Y \leq j) = \pi_1 + \cdots + \pi_j$$

- The cumulative logit is defined as:

$$ln\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = ln\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = ln\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right)$$

Logistic regression (ordinal logit model)

- The sequence of cumulative logits may be defined as:

$$L_1 = ln\left(\frac{\pi_1}{\pi_2 + \cdots + \pi_J}\right) \qquad L_2 = ln\left(\frac{\pi_1 + \pi_2}{\pi_3 + \cdots + \pi_J}\right) \qquad \ldots \qquad L_{J-1} = ln\left(\frac{\pi_1 + \cdots + \pi_{J-1}}{\pi_J}\right)$$

- Then…

$$L_1 = \alpha_1 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$L_2 = \alpha_2 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$L_{J-1} = \alpha_{J-1} + \beta_1 X_1 + \cdots + \beta_p X_p$$

Logistic regression (ordinal logit model)

- Implementation

`glm` → `polr`

MASS (version 7.3-58.3)

polr: Ordered Logistic or Probit Regression

Description

Fits a logistic or probit regression model to an ordered factor response. The default logistic case is *proportional odds logistic regression*, after which the function is named.

Usage

```
polr(formula, data, weights, start, ..., subset, na.action,
     contrasts = NULL, Hess = FALSE, model = TRUE,
     method = c("logistic", "probit", "loglog", "cloglog", "cauchit"))
```

Logistic regression (multinomial)

- The multinomial logistic regression is a regression model for:
  - multiclass problems (a multiple valued variable that does not have order)
  - ordered class values but with regression coefficients not constant across response categories

- Implementation with `glm`
  - `family = binomial` → `family = multinomial`

Logistic regression (multinomial)

- To arrive at the multinomial logit model, one can imagine, for *K* possible outcomes, running *K*-1 independent binary logistic regression models, in which one outcome is chosen as a "base" and then the other *K*-1 outcomes are separately regressed against the base outcome.

$$ln\left(\frac{P(Y=1)}{P(Y=K)}\right)$$

$$ln\left(\frac{P(Y=2)}{P(Y=K)}\right)$$

$$\ldots$$

$$ln\left(\frac{P(Y=K-1)}{P(Y=K)}\right)$$

Logistic regression assumptions

- Logistic regression <u>does not</u> require:
  - linear relationship between the dependent and independent variables
  - error terms (residuals) normally distributed
  - homoscedasticity (homogeneity of variances)
- Logistic regression <u>does</u> require that:
  - little or no multicollinearity among the independent variables
  - the independent variables are linearly related to the log odds
  - no influential values (extreme values or outliers) in the continuous predictors
  - large sample size

http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/#logistic-regression-assumptions

Calculations

- Based on least squares algorithms:
  - Iteratively Reweighted Least Squares (Fisher scoring)

```
glm_irls = function(X, y, weights=rep(1,nrow(X)), family=poisson(log), maxit=25, tol=1e-16)
    if (!is(family, "family")) family = family()
    variance = family$variance
    linkinv = family$linkinv
    mu.eta = family$mu.eta
    etastart = NULL

    nobs = nrow(X)    # needed by the initialize expression below
    nvars = ncol(X)   # needed by the initialize expression below
    eval(family$initialize) # initializes n and fitted values mustart
    eta = family$linkfun(mustart) # we then initialize eta with this
    dev.resids = family$dev.resids
    dev = sum(dev.resids(y, linkinv(eta), weights))
    devold = 0
    beta_old = rep(1, nvars)

    for(j in 1:maxit)
    {
      mu = linkinv(eta)
      varg = variance(mu)
      gprime = mu.eta(eta)
      z = eta + (y - mu) / gprime # potentially -offset if you would have an offset argumer
      W = weights * as.vector(gprime^2 / varg)
      beta = solve(crossprod(X,W*X), crossprod(X,W*z), tol=2*.Machine$double.eps)
      eta = X %*% beta # potentially +offset if you would have an offset argument as well
      dev = sum(dev.resids(y, mu, weights))
      if (abs(dev - devold) / (0.1 + abs(dev)) < tol) break
      devold = dev
      beta_old = beta
    }
    list(coefficients=t(beta), iterations=j)
}
```

Logistic regression

STRENGTHS

- Interpretable
- Efficient training
- No assumptions about distributions of classes in feature space
- Fast classification
- Performs well when the dataset is linearly separable
- Less inclined to over-fitting

WEAKNESSES

- Underperform when there are multiple or non-linear decision boundaries
- Not flexible enough to naturally capture more complex relationships
- Overfits if the number of observations is lesser than the number of features
- Requires average or no multicollinearity between independent variables

LOGISTIC REGRESSION

R session

Exercise 2

- Using the breast cancer dataset:
  - Divide into training and validaiton (80%, 20%)
  - Perform 10-fold cross validation with logistic regression over the training data.
  - Test final model on validaiton data.

Bibliography

- Machine Learning with R. Brett Lantz. Packt Publishing. 2013.
- DSO 530: Applied Modern Statistical Learning Techniques. Abbass Al Sharif. http://www.alsharif.info/#!iom530/c21o7
- An Introduction to Statistical Learning with Applications in R (2nd Edition). Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. https://www.statlearning.com
- Applied Predictive Modeling. Max Kuhn and Kjell Johnson. 2013th Edition. Springer. http://appliedpredictivemodeling.com