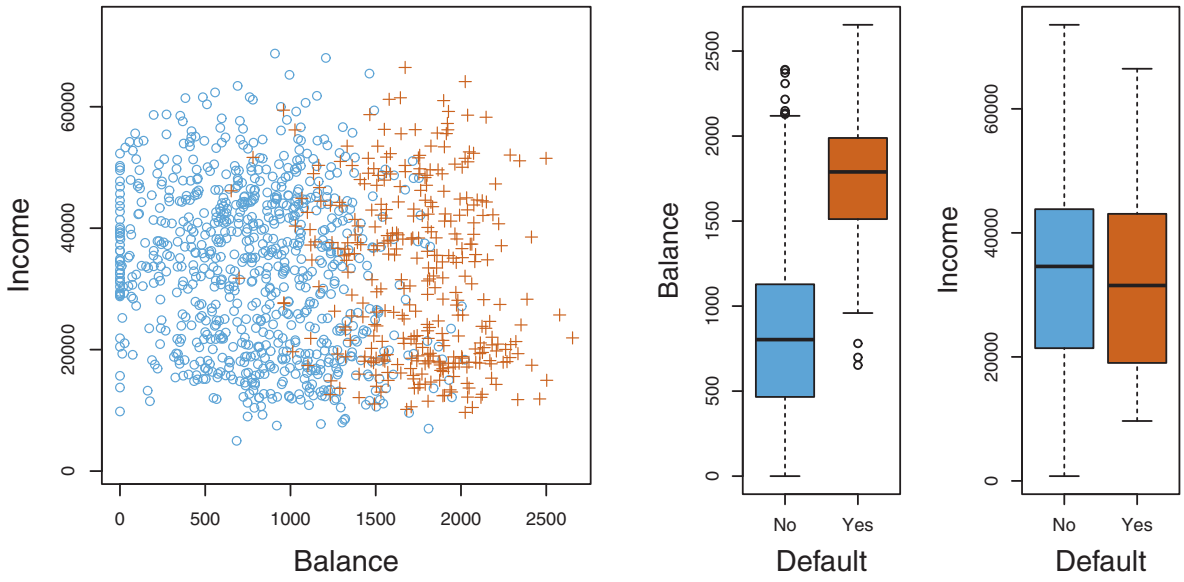
Classification

The linear regression model assumes that the response variable Y is quantitative. But in many situations, the response variable is instead qualitative (categorical).

- Qualitative variables take values in an unordered set $\mathcal{C}$, such as:
  eye color $\in$ {brown, blue, green}
  email $\in$ {spam, non-spam}.
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the probabilities that X belongs to each category in $\mathcal{C}$.
  For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.
- Just as in the regression setting, in the classification setting we have a set of training observations $(x_1, y_1), ..., (x_n, y_n)$ that we can use to build a classifier.

Example: We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.



The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: Boxplots of income as a function of default status.

Can we use Linear Regression?

Suppose for the Default classification task that we code

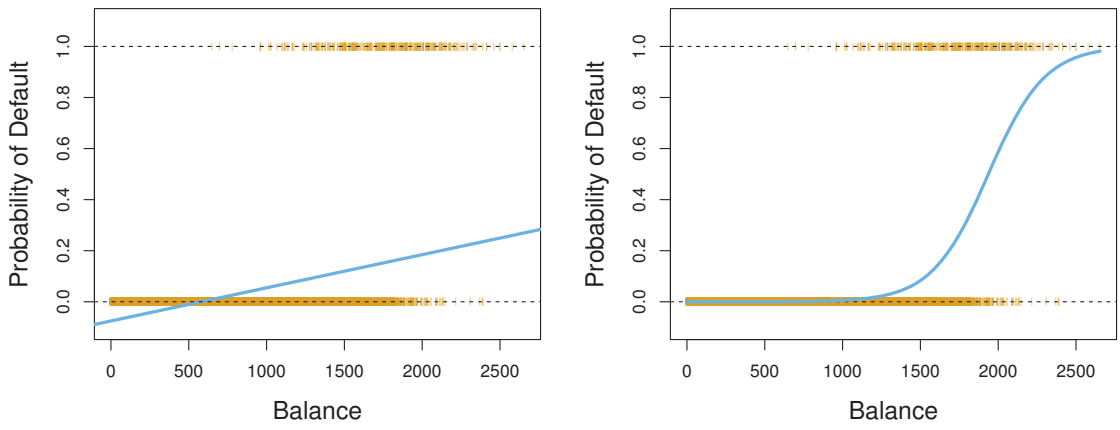
$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?

- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to linear discriminant analysis which we discuss later.
- Since in the population $E(Y|X = x) = \Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.

- However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

If we use $p(X) = pr(Y=1|X) = \beta_0 + \beta_1 X$ to predict default = Yes using balance, then for balances close to zero we predict negative probability of default; if we were to predict for very large balances, we would get values bigger than 1.



Classification using the Default data. Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for default(No or Yes). Right: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

This coding suggests an ordering, and in fact implies that the difference between stroke and drug overdose is the same as between drug overdose and epileptic seizure.

Linear regression is not appropriate here.

Logistic Regression

- Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using balance to predict default. Logistic regression has the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \qquad (*)$$

$0 < p(X) < 1$ for all values of β_0, β_1, X

we can rewrite $(*)$:

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

The quantity $\frac{p(X)}{1-p(X)}$ is called the odds,

and can take on any value between 0 and ∞.

- A bit of rearrangement gives

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

The left-hand side is called the *log odds* or *logit*. (log : ln)

The amount that p(X) changes due to a one-unit change in X depends on the current value of X. But regardless of the value of X, if β_1 is positive then increasing X will be associated with increasing p(X), and if β_1 is negative then increasing X will be associated with decreasing p(X).

Estimating the Regression Coefficients

We use maximum likelihood to estimate the parameters.

$$\text{lik}(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

This likelihood gives the probability of the observed zeros and ones in the data. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

Most statistical packages can fit linear logistic regression models by maximum likelihood.

For the Default data, we use glm() function to find the estimated coefficients of logistic regression model that predicts the probability of default using balance.

```
> library(ISLR)
> names(Default)
[1] "default" "student" "balance" "income"
> glm_fit_def_1 <- glm(default ~ balance, data = Default, family = binomial)
> summary(glm_fit_def_1)

Call:
glm(formula = default ~ balance, family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

$\hat{\beta}_0$ → Estimate (Intercept)

$\hat{\beta}_1$ → Estimate balance

Interpretation:
A one-unit increase in balance is associated with an increase in log odds of default by 0.0055.

since the p-value associated with balance is statistically significant, we can reject $H_0: \beta_1 = 0$. In other words, we conclude that there is indeed an association between balance and the probability of default.

Making Predictions

What is the estimated probability of default for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

$$\hat{\Pr}(Y = 1 \mid X = 1000) = \frac{e^{-10.65 + 0.0055 \times 1000}}{1 + e^{-10.65 + 0.0055 \times 1000}}$$

$$= 0.00576$$

which is below 1%.

The predict() function can be used to predict the probability that the individual will default, given the value of the predictor (in this case balance). The type = "response" option tells R to output probabilities of the form $P(Y = 1|X)$, as opposed to other information such as the logit.

```
> glm_probs_def_1 <- predict(
    glm_fit_def_1,data.frame(balance=c(1000,2000,3000)), type = "response")
> glm_probs_def_1
          1           2           3
0.005752145 0.585769370 0.997115227
```

Note: One can use qualitative predictors with the logistic regression model using the dummy variable approach as discussed before.

As an example, the Default data set contains the qualitative variable student. To fit a model that uses student status as a predictor variable, we simply create a dummy variable that takes on a value of 1 for students and 0 for non-students.

```
> glm_fit_def_2 <- glm(default ~ student, data = Default, family = binomial)
> summary(glm_fit_def_2)

Call:
glm(formula = default ~ student, family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2970  -0.2970  -0.2434  -0.2434   2.6585

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 2908.7  on 9998  degrees of freedom
AIC: 2912.7

Number of Fisher Scoring iterations: 6

> contrasts(Default$student)
    Yes
No    0
Yes   1
```

The coefficient associated with the dummy variable is positive, and the associated p-value is statistically significant. This indicates that

students tend to have higher default probabilities than non-students:

$$\hat{pr}(default = Yes \mid student = Yes) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\hat{pr}(default = Yes \mid student = No) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

```
> glm_probs_def_2 <- predict(
   glm_fit_def_2,data.frame(student=c('Yes','No')), type = "response")
> glm_probs_def_2
         1          2
0.04313859 0.02919501
```

Multiple Logistic Regression

We now consider the problem of predicting a binary response using multiple predictors.
We can generalize $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$ as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

where $X = (X_1, ..., X_p)$ are p predictors.