NBA 4920/6921 Lecture 5

Logistic Regression

Murat Unal

9/14/2021

Agenda

- Quiz 4
- Logistic regression
- Interpretation
- Inference
- Making predictions
- Model performance
- Exercise

```
rm(list=ls())
options(digits = 3, scipen = 999)
library(tidyverse)
library(ISLR)
library(cowplot)
library(ggcorrplot)
library(stargazer)
library(corrr)
library(lmtest)
library(sandwich)
library(MASS)
library(car)
library(jtools)
data <- ISLR::Default
auto <- ISLR::Auto
```

Logistic regression

Logistic regression is suitable for dealing with classification problems

Using logistic regression we model the probability that outcome Y belongs to a specific category

Suppose we want to understand the factors that determine credit card default

We observe the outcome as **Yes/No** in the dataset and recode it as

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

Note: R does this automatically when we call the `glm()` function.

Let's use the `Default` dataset from the ISLR package

```
str(data)
```

```
'data.frame':   10000 obs. of  4 variables:
 $ default: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1
 $ student: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2
 $ balance: num  730 817 1074 529 786 ...
 $ income : num  44362 12106 31767 35704 38463 ...
```

Our outcome of interest is `default`, whether a person failed to pay back their loan.

```
table(data$default)
```

```
  No  Yes
9667  333
```

The rate of default is only 3.3%.

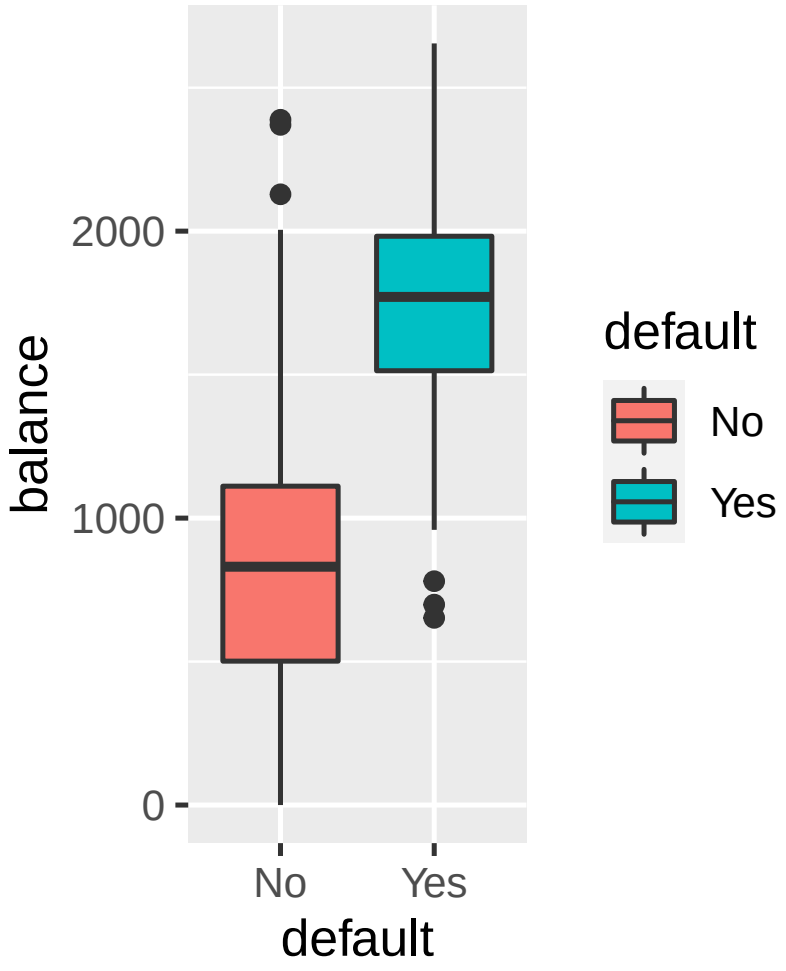
Let's sample from those who did not default and create a new sample data set that we can use to visualize patterns.

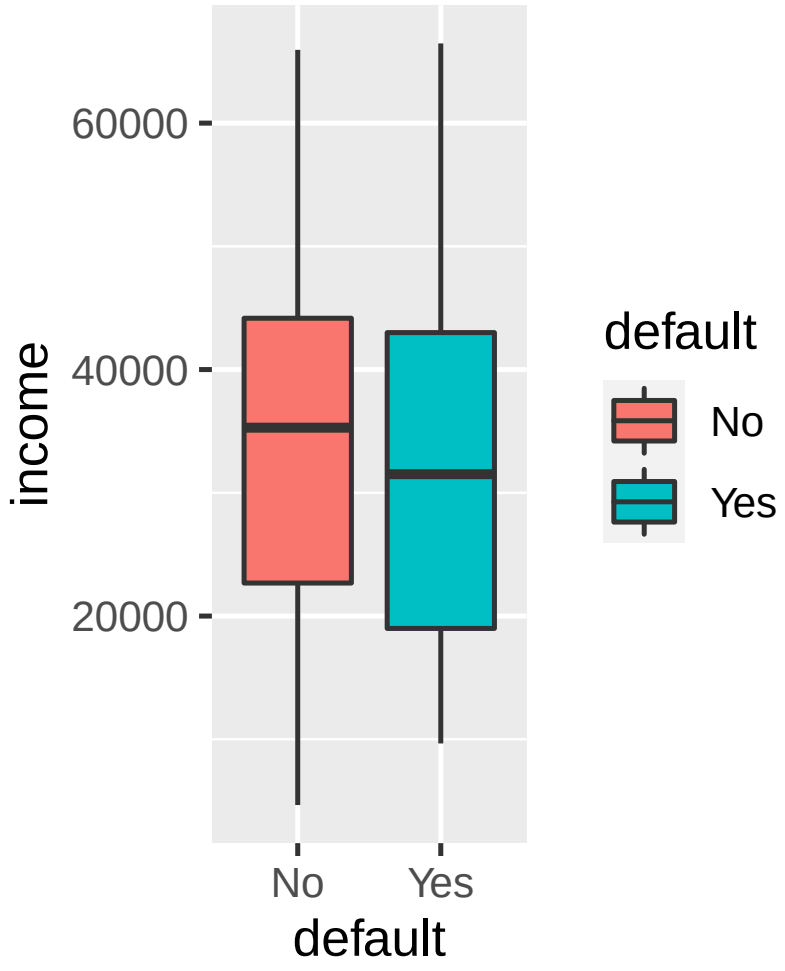
```
# Extract the observations that did not default
default.no.rows<-rownames(data[data$default=="No",])

# Sample 5% from them
no.sample<-sample(default.no.rows,0.05*nrow(data))

# Create new data frame by combining the
# 5% non-defaulters and all that did default.
default.sample<-rbind(data[no.sample,],
                      filter(data,default=="Yes"))
# New default rate.
table(default.sample$default)
```

```
 No Yes
500 333
```

The new default rate is now 0.4

As a first step, what do the following graphs tell us about the relationship between `default` and the observed factors `balance` and `income`.


balance
2000
1000
0
No
Yes
default
default
No
Yes


income
60000
40000
20000
No
Yes
default
default
No
Yes

The plots suggest `balance` is an important factor for `default`

Let us model the probability of `default` as a function of `balance`

$$p(default = 1|balance) = p(balance)$$

The plots suggest `balance` is an important factor for `default`

Let us model the probability of `default` as a function of `balance`

$$p(default = 1|balance) = p(balance)$$

We could use linear regression to estimate this model

$$p(X) = \beta_0 + \beta_1 balance$$

The plots suggest `balance` is an important factor for `default`

Let us model the probability of `default` as a function of `balance`

$$p(default = 1|balance) = p(balance)$$

We could use linear regression to estimate this model

$$p(X) = \beta_0 + \beta_1 balance$$

Q: Do you see any problems with this?

The plots suggest `balance` is an important factor for `default`

Let us model the probability of `default` as a function of `balance`

$$p(default = 1|balance) = p(balance)$$

We could use linear regression to estimate this model

$$p(X) = \beta_0 + \beta_1 balance$$

Q: Do you see any problems with this?

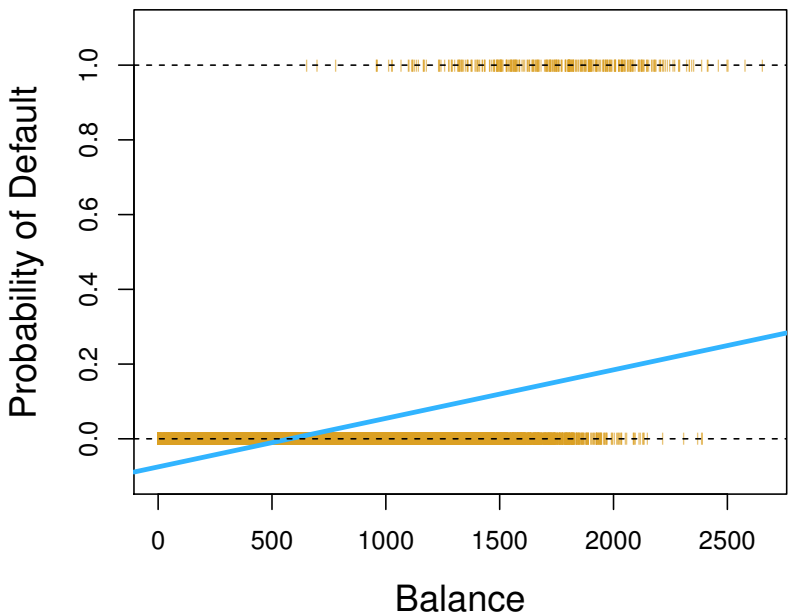
A: We can have probability estimates $[-\infty, +\infty]$

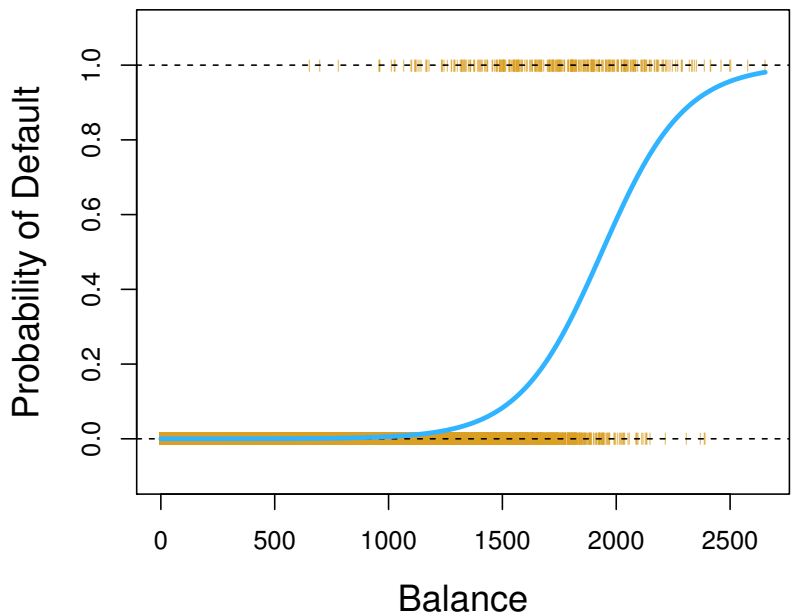
To prevent nonsensical estimates we apply the logistic transformation to the predictors

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

The logistic function ($\frac{e^x}{1+e^x}$) ensures probability estimates are between 0 and 1 no matter what values β_0, β_1 and X take

The following graphs show $p(X)$ modeled using linear regression, versus logistic regression





Interpretation

Rearranging gives us

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

We have the *log* odds/logit on the LHS and linear predictors on the RHS

Interpretation

Rearranging gives us

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

We have the *log* odds/logit on the LHS and linear predictors on the RHS

Increasing X by one unit is associated with changes in the *log* odds by β_1

Odds

Odds are defined as the ratio of the probability of success and the probability of failure

If the probability of success is 80% then the odds of success are $.8/.2 = 4$ to 1

Contrary to linear regression, the relationship between $p(X)$ and X is not a straight line,

so β_1 does **not** correspond to the change in $p(X)$ with a one unit increase in X

Changes in probability due to X are **non-linear** and depend on the value of X

Estimation

We obtain estimates for the coefficients β_0, β_1 by maximizing the **likelihood function**:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

The values $\hat{\beta}_0, \hat{\beta}_1$ are the **maximum likelihood estimates**

Inference

We can apply the same principles from linear regression for inference purposes, i.e. test the **null hypothesis** of

H_0 : The coefficient $\hat{\beta}_j$ has no effect on *log* odds,

i.e $\hat{\beta}_j = 0$ versus

the **alternative hypothesis**

H_A : The coefficient $\hat{\beta}_j$ has some effect on *log* odds i.e $\hat{\beta}_j \neq 0$

A positive (negative) and significant coefficient means that an increase (decrease) in a predictor is associated with an increase (decrease) in the *log* odds as well as $p(X)$

```
logit1 <- glm(default~balance,family = "binomial",
                              data = data)
summ(logit1, model.info = FALSE, model.fit = FALSE,
                               robust="HC0",digits = 3)
```

```
Standard errors: Robust, type = HC0
-------------------------------------------------
                      Est.    S.E.    z val.      p
---------------- --------- ------- --------- -------
(Intercept)         -10.651   0.359   -29.673   0.000
balance               0.005   0.000    25.087   0.000
-------------------------------------------------
```

A one unit increase in `balance` is associated with an increase in the *log* odds of `default` by 0.005 units

Or we can use odds by exponentiating the coefficient estimate. Set `exp=TRUE` inside the `summ()` function.

```
summ(logit1,model.info=FALSE,model.fit = FALSE,
      confint=FALSE,exp=TRUE,robust="HC0",digits = 3)
```

```
Standard errors: Robust, type = HC0
-------------------------------------------------
                    exp(Est.)    S.E.    z val.       p
---------------- ------------ ------- --------- -------
(Intercept)              0.000   0.359   -29.673   0.000
balance                  1.006   0.000    25.087   0.000
-------------------------------------------------
```

A one unit increase in `balance` multiplies the odds that `default` = 1 by a factor of $e^{0.0055} = 1.0055$

A 100 unit increase in `balance` multiplies the odds that `default` = 1 by a factor of $e^{0.0055*100} = 1.73$

Making predictions

The estimated probability of `default` $= 1$ for someone with a `balance` of 1000 is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

with a `balance` of 2000 is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

We can use `predict()` to get predictions out of the fitted model.

`predict()` produces multiple types of predictions.

1. type = `response` predicts on the scale of the response variable for logistic regression, this means predicted probabilities (0 to 1)
2. type = `link` predicts on the scale of the linear predictors for logistic regression, this means predicted log odds $(-\infty, +\infty)$

The default is type = `link`, which you may not want.

```r
# Predictions on scale of response (outcome) variable
p_hat = predict(logit1, type = "response")

# Predict '1' if p_hat is greater or equal to
# some threshold
threshold = 0.5
y_hat = as.numeric(p_hat >= threshold)
```

Qualitative predictors

```
logit2 <- glm(default~factor(student),
              family = "binomial",data = data)
summ(logit2, model.info = FALSE, model.fit = FALSE,
                              robust="HC0",digits = 2)
```

```
Standard errors: Robust, type = HC0
-------------------------------------------------
                              Est.   S.E.   z val.      p
------------------------ ------- ------ -------- ------
(Intercept)                -3.50   0.07   -49.55   0.00
factor(student)Yes          0.40   0.12     3.52   0.00
-------------------------------------------------
```

Being a student is associated with an increase in the *log* odds of default by 0.405

In terms of odds, being a student multiplies the odds of `default` $=$ 1 by a factor of $e^{0.405} = 1.5$

Let's check this is indeed true in two ways:

1. The model equation:

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

Let's check this is indeed true in two ways:

1. The model equation:

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

$$log(\frac{\hat{p}(default = 1|student = 1)}{1-\hat{p}(default = 1|student = 1)}) = log(odds((student = 1))$$

$$= -3.5041 + 0.4049 \times 1$$

Let's check this is indeed true in two ways:

1. The model equation:

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

$$log(\frac{\hat{p}(default = 1|student = 1)}{1-\hat{p}(default = 1|student = 1)}) = log(odds((student = 1))$$
$$= -3.5041 + 0.4049 \times 1$$

$$log(\frac{\hat{p}(default = 1|student = 0)}{1-\hat{p}(default = 1|student = 0)}) = log(odds((student = 0))$$
$$= -3.5041 + 0.4049 \times 0$$

Let's check this is indeed true in two ways:

1. The model equation:

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

$$log(\frac{\hat{p}(default = 1|student = 1)}{1-\hat{p}(default = 1|student = 1)}) = log(odds((student = 1))$$
$$= -3.5041 + 0.4049 \times 1$$

$$log(\frac{\hat{p}(default = 1|student = 0)}{1-\hat{p}(default = 1|student = 0)}) = log(odds((student = 0))$$
$$= -3.5041 + 0.4049 \times 0$$

$$\rightsquigarrow log(odds((student = 1)) - log(odds(student = 0)) = 0.4049$$

$$\rightsquigarrow log(\frac{odds(student = 1)}{odds(student = 0)}) = 0.4049$$

$$\rightsquigarrow \frac{odds(student = 1)}{odds(student = 0)} = e^{0.4049} = 1.5 \quad \checkmark$$

2. Using the predicted values:

$$\hat{p}(default = 1|student = 1) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$
$$= \frac{e^{-3.5041+0.4049\times 1}}{1 + e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\hat{p}(default = 1|student = 0) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$
$$= \frac{e^{-3.5041+0.4049\times 0}}{1 + e^{-3.5041+0.4049\times 0}} = 0.0292$$

$$odds(student = 1) = \frac{0.04431}{1 - 0.0431} = 0.045$$
$$odds(student = 0) = \frac{0.0292}{1 - 0.0292} = 0.030$$

$$odds(student = 1) = \frac{0.04431}{1 - 0.0431} = 0.045$$

$$odds(student = 0) = \frac{0.0292}{1 - 0.0292} = 0.030$$

$$\rightsquigarrow \frac{odds(student = 1)}{odds(student = 0)} = 0.45/0.30 = 1.5 \checkmark$$

Multiple predictors

Let us estimate the model with the full predictors: balance,income,student.

```
logit3 <- glm(default~balance + income +
                factor(student),family = "binomial",
                                        data = data)
summ(logit3, model.info = FALSE, model.fit = FALSE,
                                  robust="HC0",digits = 2)
```

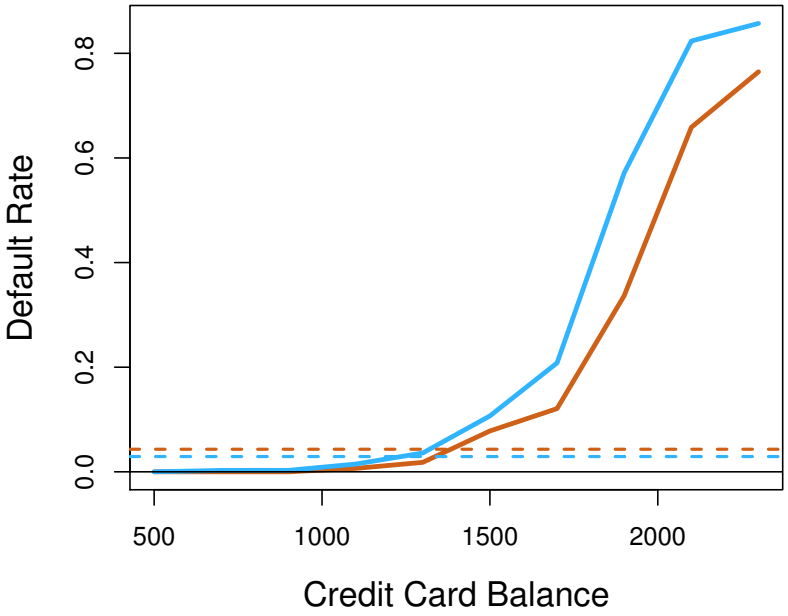
```
Standard errors: Robust, type = HC0
```

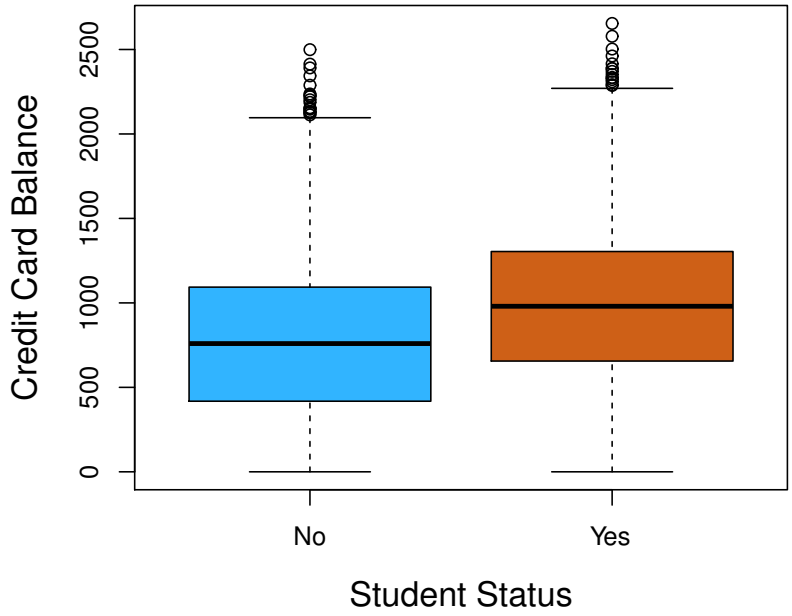
	Est.	S.E.	z val.	p
(Intercept)	-10.87	0.49	-22.07	0.00
balance	0.01	0.00	24.77	0.00
income	0.00	0.00	0.36	0.72
factor(student)Yes	-0.65	0.24	-2.67	0.01

Now, the multiple logistic regression results indicate that for a fixed value of `balance` and `income` a student is less likely to default than a non-student

Now, the multiple logistic regression results indicate that for a fixed value of `balance` and `income` a student is less likely to default than a non-student

Ok. But we just saw the opposite. Let us investigate.





`balance` and `student` are correlated. Students have higher levels of debt, which is associated with higher default rates.

So, overall, students tend to default at a higher rate than non-students (see dashed lines)

However, conditional on having the same balance, a student has a lower probability of default than a non-student (see solid lines)

This is again an example of Omitted Variable Bias (leaving out `balance` from the model results in biased estimates for the effect of `student` on the probability of default) and illustrates why we need to be careful in the conclusions we draw from model outputs using observational data

This is again an example of Omitted Variable Bias (leaving out `balance` from the model results in biased estimates for the effect of `student` on the probability of default) and illustrates why we need to be careful in the conclusions we draw from model outputs using observational data

Q: What else do you think we are missing in the model?

The probability of default for a student with a credit card balance \$1,500 and income \$40,000 is:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-+0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-+0.6468\times 1}} = 0.058$$

The probability of default for a non-student with the same balance and income is:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-+0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-+0.6468\times 0}} = 0.105$$

Model performance

How does our logistic model fit the data? We want to quantify it.

The **deviance** -negative two times the maximized log-likelihood- plays the role of *RSS* in logistic regression.

Similar to *RSS* in OLS, the **deviance** decreases as the number of variables in the model increase.

```
logLik(logit3)
```

```
'log Lik.' -786 (df=4)
```

```
#Deviance = -2*logLik
logit3$deviance
```

```
[1] 1572
```

```
#Null deviance
logit3$null.deviance
```

```
[1] 2921
```

We define a test statistic based on the differences between the residual deviance -*RSS* for the model with predictors and the null model, null deviance -*TSS*

```
with(logit3, null.deviance - deviance)
```

```
[1] 1349
```

```
with(logit3, df.null - df.residual)
```

```
[1] 3
```

The test statistic is distributed $chi - squared$ with d.o.f equal to the differences in d.o.f between the current and the null model, i.e. the number of predictors in the model.

Using this test we evaluate whether there's statistically meaningful decrease in the residual deviance-*RSS*

```
#p-value of the test:
sprintf("p-value: %f",with(logit3,
        pchisq(null.deviance-deviance,
               df.null-df.residual,lower.tail=FALSE)))
```

```
[1] "p-value: 0.000000"
```

If yes, we conclude that our model is overall significant, thus useful

We could get to the same conclusion by looking into the model fit using the summ() function

```
summ(logit3,model.info = FALSE,robust="HC0",digits = 2)
```

```
MODEL FIT:
<U+03C7>²(3) = 1349.10, p = 0.00
Pseudo-R² (Cragg-Uhler) = 0.50
Pseudo-R² (McFadden) = 0.46
AIC = 1579.54, BIC = 1608.39

Standard errors: Robust, type = HC0
```

	Est.	S.E.	z val.	p
(Intercept)	-10.87	0.49	-22.07	0.00
balance	0.01	0.00	24.77	0.00
income	0.00	0.00	0.36	0.72
factor(student)Yes	-0.65	0.24	-2.67	0.01

Similar to the **anova test** in liner regression, the **likelihood ratio test** is used to compare two nested models.

Use the `lrtest()` in R.

Exercise

1. Test whether `income` is an important variable in the model by comparing two models: with and without `income`. Do it both manually and through the `lrtest()` function.
2. Obtain prediction of default using the better model.
3. Plot the density distributions of the predictions based on default status.

```
logit.no.income <- glm(default~balance +
                factor(student),family = "binomial",
                                        data = data)
logit3 <- glm(default~balance + income +
                factor(student),family = "binomial",
                                        data = data)
```

```r
# Differences in the deviance:
dev.diff <- logit.no.income$deviance - logit3$deviance
# Get the d.o.f
dof <- logit.no.income$df.residual - logit3$df.residual
# Call the test
sprintf("p-value: %f",
        pchisq(dev.diff,dof,lower.tail=FALSE))
```

```
[1] "p-value: 0.711514"
```

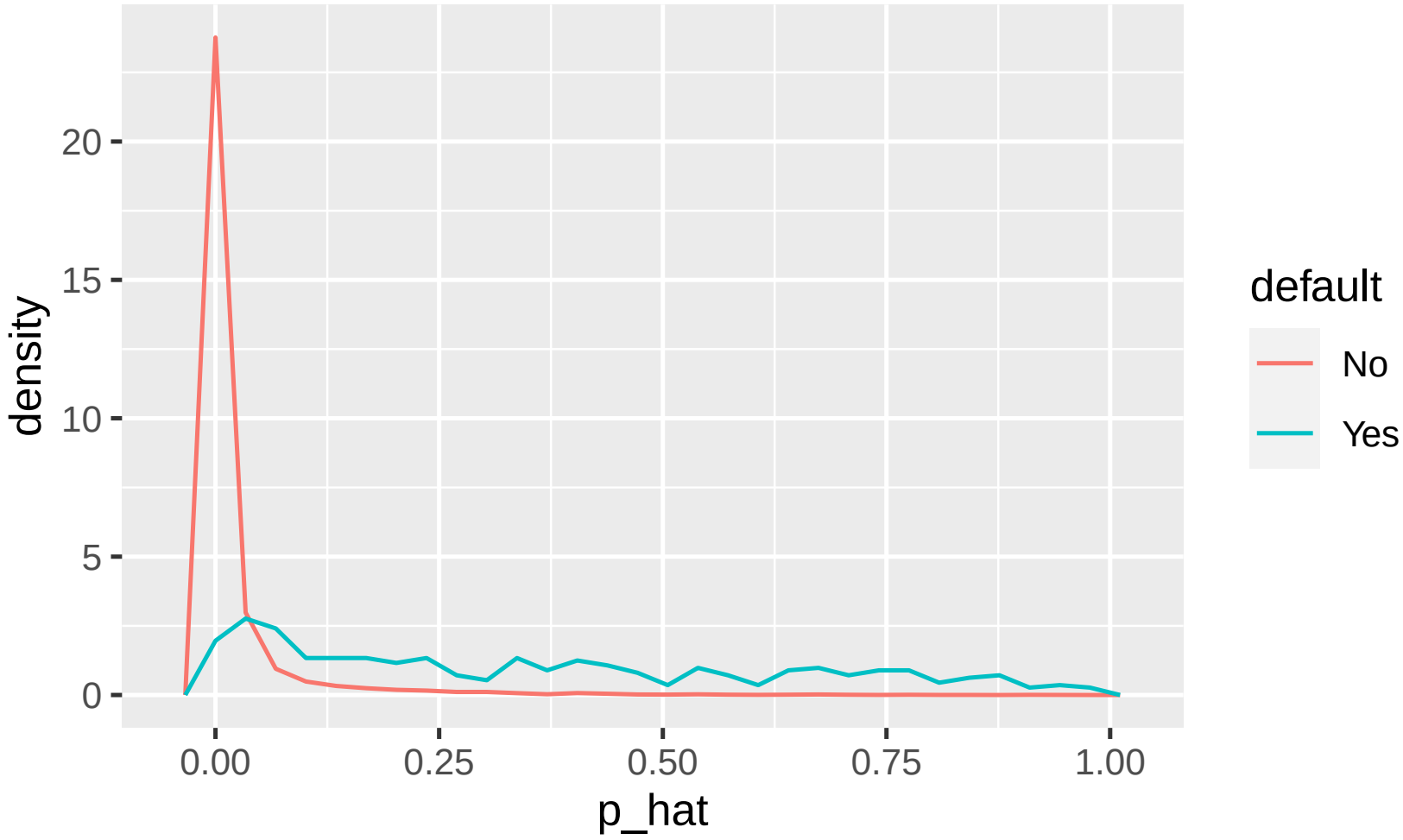
```r
# Likelihood retio test
lrtest(logit.no.income, logit3)
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
3	-786	NA	NA	NA
4	-786	1	0.137	0.712

Since `p-val`$>$ 0.05 we conclude that including `income` in the model does not have a statistically meaningful impact.

```r
# Predictions on scale of response (outcome) variable
p_hat = predict(logit.no.income, type = "response")
# Add the predictions to the data:
data$p_hat <- p_hat
```

```r
ggplot(data,aes(y=..density..,x=p_hat,color=default))+
  geom_freqpoly()
```


density
20
15
10
5
0
0.00
0.25
0.50
0.75
1.00
p_hat
default
No
Yes