# Logistic Regression

Data Science for Marketing Decision Making
Günter J. Hitsch
Chicago Booth

Winter 2017



## Introduction

- Many applications of marketing analytics involve *discrete outcomes* and *classification problems*

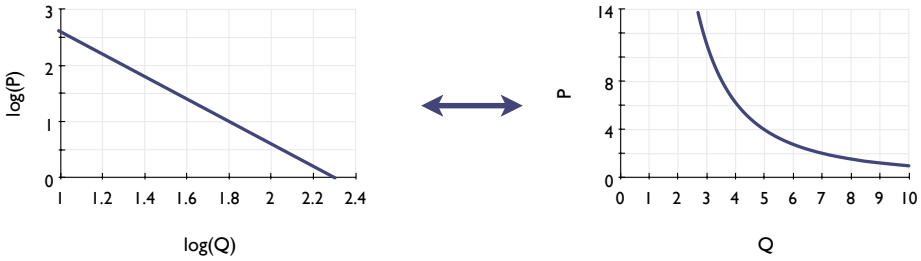
- Examples
  - We send a catalog to a customer in our data base. Will the customer respond?
  - Is a customer in our data base likely to cancel her account in the next three months?
  - Is a household likely to default on a loan?
  - Does online exposure of a consumer to display advertising lead to a sale?

# The regression model we used so far

Linear regression model:

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Predicts an output that varies continuously in the inputs




Limited in usefulness to predict discrete outcomes, such as the examples discussed before:

- 1 = "buy from catalog", 0 = "do not buy from catalog"
- 1 = "cancel account", 0 = "do not cancel account"
- 1 = "default on loan", 0 = "do not default on loan"



# Logistic regression

- Logistic regression is a widely applicable method to predict choices at the individual (consumer) level or to classify outcomes into discrete categories
- The outcome in a *binary logistic regression* model is discrete, and either 0 or 1
  - Extension: Methods for multinomial outcomes (1, 2, 3, . . . )
- Allows us to predict the outcomes in our examples (purchase response, account cancellation, . . . )

Logistic regression predicts the conditional (on the inputs) probability of the two possible outcomes ($Y = 0, 1$) from the independent variables:

$$\Pr\{Y = 0|X\}, \Pr\{Y = 1|X\} \longleftarrow X_1, X_2, \ldots, X_p$$

Why predict the probability of an outcome, not the outcome itself?

- Example: Predict default = 1 if default is more likely than no default
- Will predict default = 1 both if chance of default is 90% and if chance of default is 51%
- Predicting probability is more informative than predicting 0 or 1



## Probability prediction

Is it possible to predict the outcome probability ($Y = 1$) using a linear regression model?

$$\Pr\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

This is called a *linear probability model*

Problems with the linear probability model:

- Predicted probability may be smaller than 0 or larger than 1— may yield poor prediction

Alternative:

- Transform prediction to take values between 0 and 1 using some known function
- Generalized linear model (GLM)

## Probability prediction using logistic function

Use *logistic function* for transformation:

$$\Lambda(z) = \frac{\exp(z)}{1+\exp(z)}, \qquad 0 < \Lambda(z) < 1$$

Logistic regression model:

- Index for level of predicted probability:

$$z = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Transform index to yield probability prediction between 0 and 1

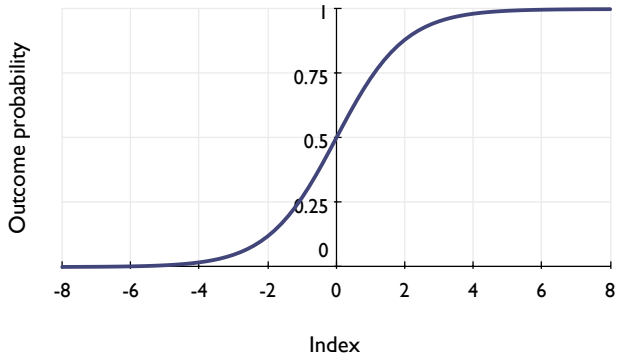
$$\begin{aligned}\Pr\{Y = 1|X) &= \frac{\exp(z)}{1+\exp(z)} \\ &= \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1+\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}\end{aligned}$$



## Exploring the logistic regression model

Index: $z = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

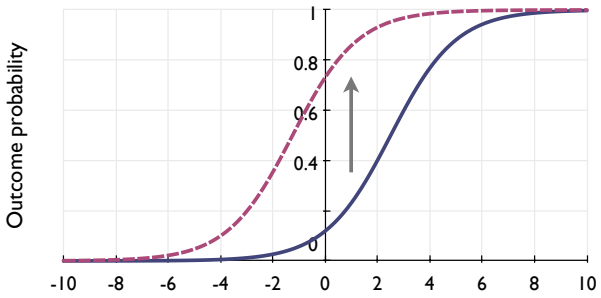
Logistic curve:



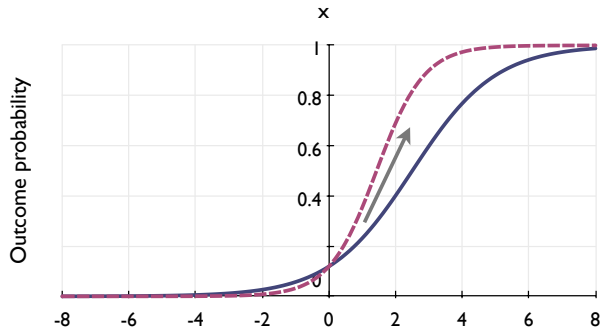

- Predicted outcome probability close to 0 for small (negative) index values
- Predicted outcome probability close to 1 for large index values
- $\Lambda(0) = 0.5$

Note: Variable $X_K$ (not index $z$) on x-axis

- Effect of increasing $\beta_0$ ("intercept")




- Effect of increasing $\beta_k$ ("slope parameter")






## Remarks

- Terminology
  - In marketing (and economics) the logistic regression model for binary outcomes is often called the *logit model* or *binary logit model*
- If we replace the logistic function $\Lambda(z)$ with $F(z)$, the cumulative distribution function for a standard normal distribution, we obtain the *probit model* or *probit regression*
  - The probit model is similar to logistic regression, but does not have a closed-form solution for the outcome probabilities

## Choice-theoretic foundation for the logit model

- Behavioral model of choice with two options, 0 and 1
  - Can be applied to product or brand purchase, voting behavior, or choice of a marriage partner
- We denote the index $z$ by $u$, the *indirect utility* from a choice:

$$u(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - The inputs could be price and promotions, political positions of a candidate, or attributes of a potential partner
- Choice of $Y = 1$ versus option $Y = 0$ is based on utility maximization: Choose option 1 if (and only if)

$$U_1 = u(X) + \epsilon > 0 = U_0$$

  - $\epsilon$ is an unobserved (to the data scientist) component of indirect utility
  - $U_0 = 0$ is simply a convenient normalization

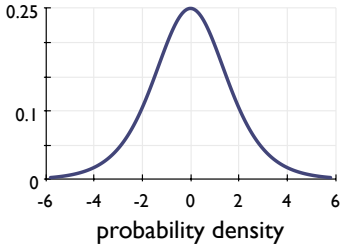


## Choice model prediction

- Because $\epsilon$ is unobserved, we do not know the exact value of $U_1$
- We assume $\epsilon$ has a standard logistic distribution. The cumulative distribution function of the standard logistic distribution is given by

$$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)}$$

- The probability density function of the standard logistic distribution:



## Probability of choice

- If $\epsilon$ has a standard logistic distribution, then the probability that the decision maker choose option 1 over 0 is

$$\begin{aligned}\Pr\{u(X)+\epsilon>0|X\} &= \frac{\exp(u(X))}{1+\exp(u(X))} \\ &= \frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}\end{aligned}$$

- Note the role of the unobserved component in indirect utility: It allows us to rationalize all observed choices in the data.
  - Example: $u(X)$ may be very large, but we many nonetheless observe that option 0 is chosen. This is due to some unmeasured utility component captured by $\epsilon$



## Proof

$$\begin{aligned}\Pr\{Y=0|X\} &= \Pr\{u(X)+\epsilon\leq 0|X\} \\ &= \Pr\{\epsilon\leq -u(X)\} \\ &= \Lambda(-u(X)) \\ &= \frac{\exp(-u(X))}{1+\exp(-u(X))} \\ &= \frac{\exp(-u(X))}{1+\exp(-u(X))}\cdot\frac{\exp(u(X))}{\exp(u(X))} \\ &= \frac{1}{1+\exp(u(X))} \\ &= \frac{1}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}\end{aligned}$$

Hence

$$\begin{aligned}\Pr\{Y=1|X\} &= 1-\Pr\{Y=0|X\} \\ &= \frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}\end{aligned}$$

# Estimation

- Example: Demand for a potential partner when using an online dating site
- Online dating data:
  - Observe various user attributes, and if a site user sends a first-contact (unsolicited) e-mail to another user after viewing his or her profile
  - The decision to send an e-mail is a binary choice
    - 1 = send e-mail, 0 = ignore the other user
  - Because of sample size we only look at choices among opposite-sex partners, e.g. decisions of women to send first-contact e-mails to men
- Data in `Online-Dating.RData`



| Variable | Description |
|---|---|
| `profile_gender` | Gender of person in profile, `male` or `female` |
| `first_contact` | 1 = first-contact e-mail sent, 0 = otherwise |
| `age` | Age of the person in the profile, in years |
| `age_older` | 1 = potential mate in profile is at least 5 years older |
| `age_younger` | 1 = potential mate in profile is at least 5 years younger |
| `looks` | Numerical looks rating |
| `height` | Inches |
| `height_taller` | 1 = potential mate at least 2 inches taller |
| `height_shorter` | 1 = potential mate at least 2 inches shorter |
| `bmi` | Body mass index |
| `yrs_education` | Years of education |
| `educ_more` | 1 = potential mate has at least 2 more years of education |
| `educ_less` | 1 = potential mate has at least 2 years less of education |
| `income` | $1,000 annual income |
| `diff_ethnicity` | 1 = potential mate has different ethnicity than browser |

# Estimation using R

Model:

- Utility of sending a first contact-email ($Y = 1$) depends on utility from a potential match
- Utility of potential match depends on attributes of the mate, such as age, looks, etc. $(X_1, X_2, \dots)$

Estimation using the glm function:

```
fit = glm(y ~ x1 + x2 + ..., family = binomial(), data = DT)
summary(fit)
```

- Binary 0/1 outcome variable `y`
- GLM = generalized linear model
- Use `binomial(link = "probit")` to estimate a probit model



Predict if women send a first-contact e-mail after viewing the profile of a potential male partner

```
library(data.table)
load("./Data/Dating-Data.RData")

women_DT = dating_DT[profile_gender == "male",
                     -"profile_gender", with = FALSE]

womens_choices = glm(first_contact ~ .,
                     family = binomial(),
                     data = women_DT)
```

```
Call:
glm(formula = first_contact ~ ., family = binomial(), data = women_DT)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7989  -0.4184  -0.3644  -0.3115   3.0393

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.5251886  0.4706216 -15.990  < 2e-16 ***
age             0.0150776  0.0020255   7.444 9.77e-14 ***
age_older      -0.2693740  0.0324052  -8.313  < 2e-16 ***
age_younger    -0.3078933  0.0364871  -8.438  < 2e-16 ***
looks           0.5236808  0.0285258  18.358  < 2e-16 ***
height          0.0417912  0.0058863   7.100 1.25e-12 ***
height_taller   0.3470403  0.0690557   5.026 5.02e-07 ***
height_shorter -0.3188366  0.1272128  -2.506   0.0122 *
bmi             0.0360348  0.0058656   6.143 8.07e-10 ***
yrs_education   0.0133780  0.0075301   1.777   0.0756 .
educ_more      -0.1812976  0.0338545  -5.355 8.55e-08 ***
educ_less      -0.2295514  0.0395019  -5.811 6.20e-09 ***
income          0.0025459  0.0002513  10.130  < 2e-16 ***
diff_ethnicity -0.4960341  0.0761361  -6.515 7.26e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41036  on 79999  degrees of freedom
Residual deviance: 40129  on 79986  degrees of freedom
AIC: 40157

Number of Fisher Scoring iterations: 6
```

```
Call:
glm(formula = first_contact ~ ., family = binomial(), data = men_DT)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8318  -0.4696  -0.4157  -0.3480   3.0594

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.0517899  0.4060820   0.128 0.898517
age             0.0061920  0.0018682   3.314 0.000918 ***
age_older      -0.1792139  0.0384492  -4.661 3.15e-06 ***
age_younger    -0.0734181  0.0304971  -2.407 0.016067 *
looks           0.4837548  0.0214545  22.548  < 2e-16 ***
height         -0.0153335  0.0052140  -2.941 0.003273 **
height_taller  -0.5151478  0.1042932  -4.939 7.84e-07 ***
height_shorter  0.1107209  0.0541539   2.045 0.040898 *
bmi            -0.0620012  0.0044635 -13.891  < 2e-16 ***
yrs_education  -0.0204628  0.0065302  -3.134 0.001727 **
educ_more      -0.0611685  0.0318745  -1.919 0.054979 .
educ_less      -0.1431284  0.0329015  -4.350 1.36e-05 ***
income          0.0002289  0.0003797   0.603 0.546576
diff_ethnicity -0.2465653  0.0421174  -5.854 4.79e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# Logistic regression output

- Coefficient estimate: Shows how each variable enters $u(X)$, and thus the probability of choosing option 1
- $z$-statistic: Standard normal test statistic used to test for statistical significance of coefficient estimate
  - For all practical purposes use like a $t$-statistic
- $p$-value: As in standard regression model

- AIC: Akaike information criterion
  - AIC value not directly interpretable
  - Useful to select a model among many candidate models, and also allows us to compare models with a different number of independent variables
  - Approach:
    1. Estimate each candidate model and record the AIC
    2. Select model with lowest AIC

## Estimation: Mathematical background

How does the computer estimate the logistic regression model coefficients?

- In regression analysis, the idea is to make the predicted outcome as close as possible to the actual outcome
- The same idea is at work here: the software selects parameter values to match the predicted probability of an outcome (between 0 and 1) to the actual outcome (0 or 1)
- The estimation method used is called *maximum likelihood*

$$\max_{\boldsymbol{\beta}} \log(l(\textit{data} = (\boldsymbol{y}, \boldsymbol{X})|\boldsymbol{\beta})) = \sum_{i=1}^{n} \log(\Pr\{Y = y_i|\boldsymbol{x}_i; \boldsymbol{\beta}\})$$



## Interpretation of the estimates

- The estimated coefficients cannot be interpreted in the same way as the coefficient estimates in the linear regression model!
  - In the example before, the estimate on the dummy variable `height_taller` (men making choices) was -0.515
  - This does not mean that the probability of a first contact decreases by 51.5 percent if men choose between taller women and women of similar height
- Even if we take the log of a non-categorical independent variable the coefficient estimate is not an elasticity
  - Note: Generally, there is no benefit of taking the log of the independent variables in the logistic regression model

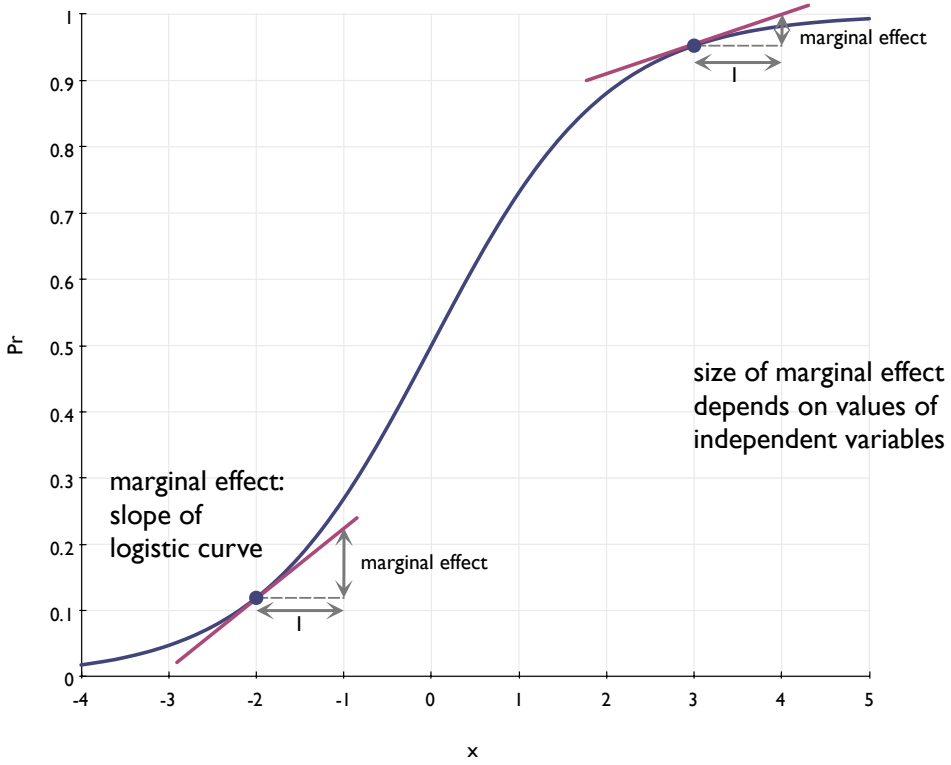
# Interpretation of the estimates:Marginal effects

- The marginal effect of the independent variable $X_k$ is the change in the outcome probability $\Pr\{Y = 1|X\}$ relative to an increase in $X_k$ by a "small" unit

$$\text{marginal effect} = \frac{d\Pr\{Y = 1|X\}}{dX_k} \approx \frac{\Delta\Pr\{Y = 1|X\}}{\Delta X_k}$$

- $\Delta X_k$ is the "small" increase and $\Delta\Pr\{Y = 1|X\}$ is the corresponding change in the outcome probability
- Mathematically, the marginal effect is the derivative of the outcome probability with respect to $X_k$

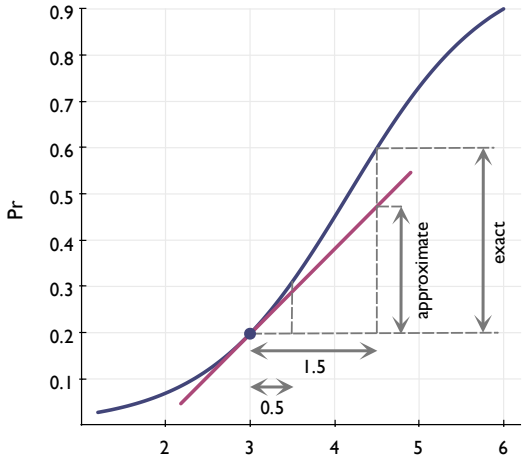




# Using the marginal effects

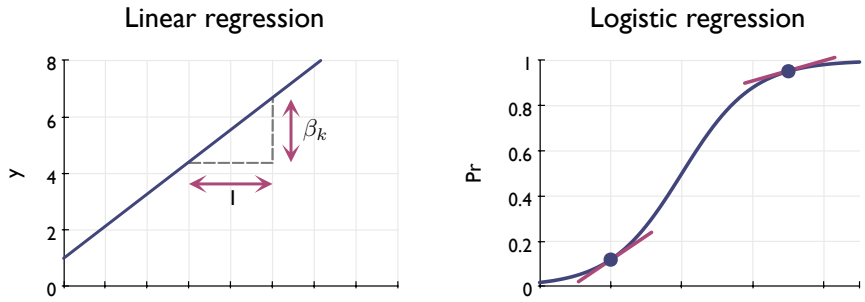
- We would like to predict by how much the outcome probability $\Pr\{Y = 1|X\}$ changes for a change in $X_k$ by $\Delta X_k$ units

$$\Delta \Pr\{Y = 1|X\} \approx \Delta X_k \cdot \text{marginal effect}$$






# Marginal effects: Linear vs. logistic regression model




- In linear regression model, marginal effect = regression coefficient $= \beta_k$. Marginal effect does not depend on values of inputs
- In logistic regression, marginal effect = slope of the logistic curve. Depends on values of inputs

Marginal effect concept is simple in linear regression, hence we did not previously discuss it

## Calculation of marginal effects

1. Decide for what values of the independent variables you want to calculate the marginal effect
   - Observed or mean of each input
2. Calculate the outcome probability at the chosen values of the independent variables:

$$p = \Pr\{Y = 1|X\} = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

3. The marginal effect of increasing $X_k$ is

$$\text{marginal effect} = \beta_k \cdot p \cdot (1 - p)$$



## Exact effect of increasing $X_k$

Compare predictions at $X = X_0$ and $X = X_1$:

$$\Delta \Pr\{Y = 1\} = \Pr\{Y = 1|X_1\} - \Pr\{Y = 1|X_0\}$$

Best done using the `predict` command in R

# Calculating marginal effects in R

- Install and load the package `erer`
- Estimate the model and then calculate the marginal effects:

```
fit = glm(y ~ x1 + x2 + ..., family = binomial(), x = TRUE, data = DT)
maBina(fit, digits = 6)
```

- Need to set the `x = TRUE` option in the `glm` function!
- Use the `digits` option to choose the number of digits in the output
- By default, `maBina` calculates the marginal effects at the means of all independent variables
- To calculate the average marginal effects across all observations in the data, use

```
maBina(fit, x.mean = FALSE, digits = 6)
```

- For dummy variables, `maBina` predicts the exact effect of increasing the variable from $X_k = 0$ to $X_k = 1$



```
womens_choices = glm(first_contact ~ ., family = binomial(),
                     x = TRUE, data = women_DT)
library(erer)
maBina(womens_choices, x.mean = FALSE, digits = 6)

                  effect    error    t.value  p.value
(Intercept)    -0.491205 0.030399 -16.158449 0.000000
age             0.000984 0.000132   7.475458 0.000000
age_older      -0.016271 0.001910  -8.516871 0.000000
age_younger    -0.017896 0.001989  -8.998352 0.000000
looks           0.034183 0.001825  18.734088 0.000000
height          0.002728 0.000383   7.121283 0.000000
height_taller   0.019068 0.003350   5.691873 0.000000
height_shorter -0.017346 0.006045  -2.869600 0.004111
bmi             0.002352 0.000382   6.161249 0.000000
yrs_education   0.000873 0.000491   1.777149 0.075548
educ_more      -0.010938 0.001993  -5.488605 0.000000
educ_less      -0.013511 0.002211  -6.111024 0.000000
income          0.000166 0.000016  10.175925 0.000000
diff_ethnicity -0.025411 0.003176  -8.002155 0.000000

mean(women_DT$first_contact)     # For comparison

[1] 0.0711
```

# Calculating exact effects of increasing $X_k$ in R

- Predict outcome probabilities:

```
Pr = predict(fit, type = "response")
Pr = predict(fit, newdata = new_DT, type = "response")
```

- If the `newdata` option is not specified probabilities will be predicted for the data set used to estimate the model
- Make sure the `data.table` or data frame in `newdata` includes all variables used in the original logistic regression model



Example: Change (increase) in first-contact probabilities if all men become millionaires

```
Pr_0 = predict(womens_choices, type = "response")

# Copy data and set income to 1000 (= $1,000,000)
millionaires_DT = women_DT
millionaires_DT[, income := 1000]

# Predict first-contact probabilities at high income data
Pr_1 = predict(womens_choices, newdata = millionaires_DT,
               type = "response")

Delta_Pr = Pr_1 - Pr_0
summary(Delta_Pr)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.07189 0.30870 0.35270 0.34880 0.39580 0.53020
```

## Historical background

- Logistic regression—the logit model—is a specific instance of a *discrete choice model*
- Discrete choice models were first developed and used to explain choice behavior in the 1970s
- Discrete choice models have revolutionized marketing — much of the work conducted at Booth and other schools during the last thirty years would not have been possible without these models
- Jim Heckman (University of Chicago) and Dan McFadden (UC Berkeley) received the Nobel Prize in Economics in 2000 for their work on statistical methods to capture individual choices



## Summary

- Logistic regression predicts the probability of the two possible outcomes $Y = 0, 1$ conditional on the values of the independent variables $X_1, \ldots, X_p$
- The logistic regression model can be interpreted as a choice model, where the choice of 1 versus 0 is based on a comparison of the corresponding indirect utilities
- Estimating the logistic regression model is as simple as estimating the linear regression model, but interpreting the coefficient estimates is more difficult
  - Marginal effects: Change in outcome probability relative to a "small" change in one of the independent variables