Predicting venture capital funding with other characteristics during IPOs of companies

STAT 823: Fall Class Project, 2020

Bhargob Saikia, Chiranjeevi Pippalla, Dinesh Dandamudi

Department of Biostatistics and Data Science
University of Kansas, USA
December 9, 2020

Contents

List of Tables

Abstract

The response of interest is whether or not a company was financed with venture capital funds.Venture capitals comes from well off investors who sees exceptional potential in that certain company which is a very good sign and reflects high growth capacity.Potential predictors include the face value of the company, the number of shares offered, and whether or not the company underwent a leveraged buyout.

Introduction

Private companies often go public by issuing shares of stock referred to as initial public offerings (IPOs). IPOs provide companies with an opportunity to obtain capital by offering shares through the primary market.Companies hire investment banks to market, gauge demand, set the IPO price and date, and more. Prior to an IPO, a company is considered private. As a private company, the business has grown with a relatively small number of shareholders including early investors like the founders, family, and friends along with professional investors such as venture capitalists or angel investors.Overall, the number of shares the company sells and the price for which shares sell are the generating factors for the company's new shareholders' equity value. Shareholders' equity still represents shares owned by investors when it is both private and public, but with an IPO the shareholders' equity increases significantly with cash from the primary issuance. As mentioned in the abstraxt venture capital is one of the bullish signs for the stocks of a company.So it is always interesting to look for factors that attract venture capital and other fundings.

Data Sources

The source of data is online. A study of 482 IPOs was conducted to determine what are the characteristics of companies that attract venture capital funding. Each line of the data set has an identification number and provides information on 4 other variables for a single company. Leveraged buyout

Table 1: IPO data set

Variable	Variable name	Description
1	idnum: Identification number	1 482
2	funding: Venture capital funding	venture capital funding: 1 if yes; 0 otherwise
3	facevalue: Face value company	Estimated facevalue from prospectus (in Million dollars)
4	shares: Number of shares offered	shares offered (in Millions)
5	buyout:	leveraged buyout: 1 if yes; 0 otherwise

Statistical Analysis

The data is available in .xlsx (excel) format. The data analysis is done using the statistical software R and the since the response variable has output only 0 or 1 project focuses mainly on multiple logistic regression. Each of the predictor variable is explored individually and the illustrations used are conducted on the entire dataset for the preliminary investigation. No missing values were found in the dataset. The large sample size and absence of missing value is assumed to ensure better predictibility and less sampling variability. Automatic model selection method has been used to arrive at the final model. The model assumptions are assessed and confirmed while suggesting the final model.

Model Assumptions

All inferences are conducted using $\alpha = 0.05$ unless stated otherwise. No adjustments for multiplicity are made as this is an exploratory analysis. Discrete variables are summarized with proportions and frequencies. The continuous variables, which are summarized using mean, median, standard deviations, coefficient of variation, quantiles, variance, maximum and minimum.

Primary Objective Analysis

Exploring individual predictors and the response variable is very important before starting data analysis. It helps in detecting skewness, presence of outliers or can also suggest if transformations are necessary to fit a better model. Once the dataset variables are explored, the next step is performed to check the relation between the predictors and the outcome individually. This helps us fit a better model which can do a better job at explaining the variation of the response.

Table 2: Analysis of predictor variables

summary	facevalue	shares	buyout
min	1.2	0.30	-
1st quantile	10.2	1.3	..
median	19.5	2.00	..
Mean	26.5	2.23	0.0934
3rd quantile	32.5	2.70	..
max	234.6	11.02	..

Analysis of the predictor variables

Effect of facevalue on funding

We try to get a logistic model between the facevalue and funding. below is the summary table describing the relation.

Table 3: table of the relation between facevalue and funding

Co-efficients	Estimate	Std. Error	z value	Pr(>z)
Intercept	-0.232762	0.130364	-1.79	0.074
facevalue	0.000342	0.003496	-0.10	0.922

So the resultant logistic regression model is

$logit(\pi(x)) = -0.232762 - 0.000342X$ where X is the facevalue of the company (in millions) and $\pi(y)$ is the probability of getting funding.

we check the fit of this model and get $p - value = 7.1 \times 10^{-8}$ which is much lesser than 0.05. so it is not a good fit.

Effect of number of shares on funding

Table 4: table of the relation between number of shares and funding

Coefficients:	Estimate	Std. Error	z value	Pr(>z)
(Intercept)	-0.3170	0.1713	-1.85	0.064
shares	0.0337	0.0647	0.52	0.603

So the resultant logistic regression model is

$logit(\pi(y)) = -0.3170 - 0.0337X$ where X is the number of shares of the company (in millions) and $\pi(y)$ is the probability of getting funding.

The p -value is much smaller than 0.05 again. So not a good fit

Table 5: table of the relation between buyout and funding

Coefficient	Estimate	Std. Error	z value	Pr(>z)
(Intercept)	-0.2345	0.0963	-2.43	0.015
buyout	-0.0792	0.3168	-0.25	0.803

Effect of financial buyout on funding

So the resultant logistic regression model is

$logit(\pi(y)) = -0.2345 - 0.0792X$ where X is 1 if there was a buyout else it is zero and $\pi(y)$ is the probability of getting funding.

The p-value of the fit test is much less than 0.05.So not a good fit.

Main model

Automatic variable selection method

We will try to get rid of redundant variables if any.

The final model looks like: $logit[\pi(X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ where

- X_1 = facevalue of the shares (in millions \$)
- X_2=total number of shares in (millions\$)
- X_3= 1 if there was a finacial leverage buyout else it is zero
- β_1 = change in logit when funding increases by 1 unit keeping other covariates fixed.
- β_2 = change in logit when number of shares increases by 1 unit keeping other covariates fixed.
- β_3 = change in logit when there is a financial buyout.

The table 4 gives us the summary of our model.

Table 6: table of summary with all given factors

Coefficients:	Estimate	Std. Error	z value	Pr(>z)
(Intercept)	-0.41997	0.19391	-2.17	0.03
facevalue	-0.01009	0.00809	-1.25	0.21
shares	0.20291	0.14936	1.36	0.17
buyout	-0.07578	0.31860	-0.24	0.81

So it will look like

$$logit[\pi(X)] = -0.41997 - 0.01009X_1 + 0.20291X_2 - 0.07578X_3$$

Goodness of Fit Test

The p-value we get is 6.74×10^{-8}So might not be a good fit. So let us try to see if we can get rid of any redundant variables

Deleting number of shares

Table 7: table of the relation between facevalue, buyout and funding

Coefficients:	Estimate	Std. Error	z value	P(>z)
(Intercept)	-0.22753	0.13211	-1.72	0.085
facevalue -0.00027	0.00351	-0.08	0.939	
buyout -0.07710	0.31796	-0.24	0.808	

Excluding the buyout factor

Table 8: table showing relation between facevalue, number of shares and funding

Coefficients:	Estimate	Std. Error	z value	P(>z)
(Intercept)	-0.42518	0.19271	-2.21	0.027
facevalue -0.01017	0.00808	-1.26	0.027	
shares 0.20306	0.14939	1.36	0.174	

Excluding the facevalue factor

Table 9: Table showing relation between number of shares,buyout and funding

Coefficient:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3117	0.1723	-1.81	0.07
shares	0.0351	0.0649	0.54	0.59
buyout	-0.0922	0.3178	-0.29	0.77

Discussion and Conclusion

We checked by eliminating various factors one by one. But that didnot effect the p-value much. The fit is not very good in almost all of the models.Our main model is the best we have got. So our model will be as below

$$logit[\pi(X)] = -0.41997 - 0.01009X_1 + 0.20291X_2 - 0.07578X_3$$

where the terms have been defined earlier.

Appendix:R code

```
install.packages("readxl")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
install.packages("faraway")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library("readxl")
actual_data <- read_excel("ipo.xlsx")
actual_data
```

```
## # A tibble: 482 x 5
##    idnum funding facevalue shares buyout
##    <dbl>   <dbl>     <dbl>  <dbl>  <dbl>
##  1     1       0      1.2    3         0
##  2     2       0      1.45   1.45      1
##  3     3       0      1.5    0.3       0
##  4     4       0      1.53   0.51      0
##  5     5       0      2      0.8       0
##  6     6       0      2.1    0.7       0
##  7     7       0      2.1    0.4       0
##  8     8       1      2.5    0.5       0
##  9     9       0      2.5    0.5       0
## 10    10       0      2.75   0.55      0
## # ... with 472 more rows
```

```
summary(actual_data$facevalue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.2    10.2    19.5    26.5    32.5   234.6
```

```
summary(actual_data$facevalue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.2    10.2    19.5    26.5    32.5   234.6
```

```
summary(actual_data$shares)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.30    1.30    2.00    2.23    2.70   11.02
```

```
summary(actual_data$buyout)
```

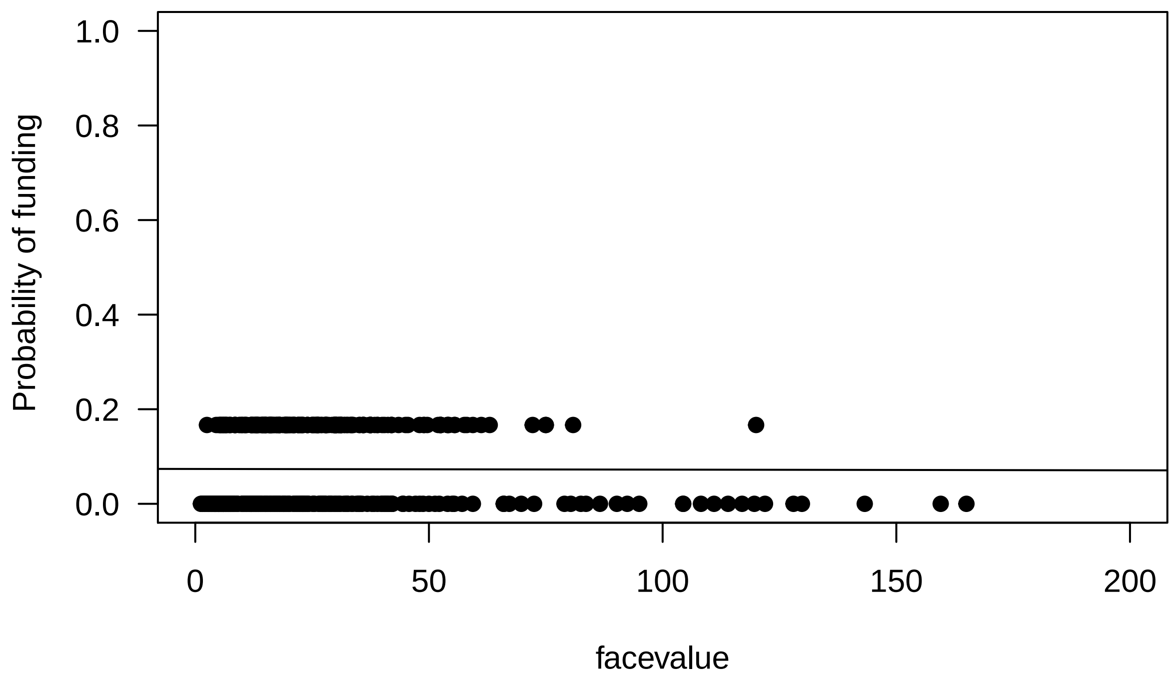
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   0.0000   0.0000   0.0000   0.0934   0.0000   1.0000
```

```
library(faraway)
library(formatR)
```

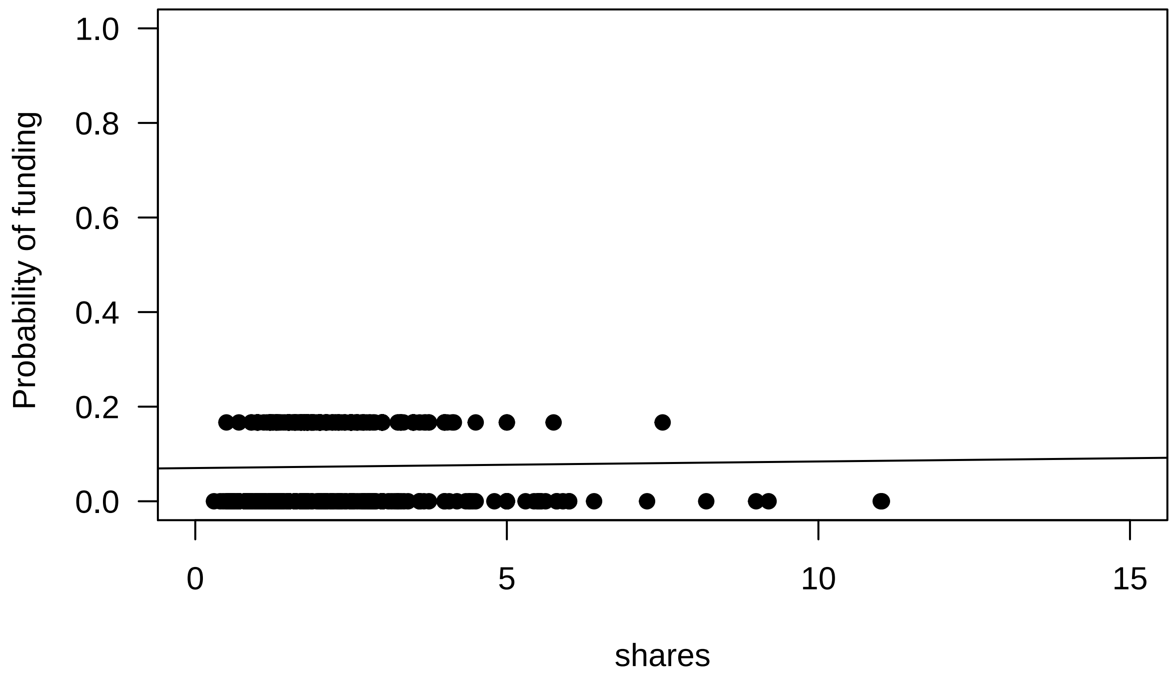
#Plotting relation between facevalue and funding

```
plot(funding/6 ~ facevalue, data = actual_data, xlim = c(0, 200),
    ylim = c(0, 1), xlab = "facevalue", las = 1, ylab = " Probability of funding",
    pch = 19)
abline(lm(funding/6 ~ facevalue, data = actual_data))
```



#plotting relation between number of shares and funding

```
plot(funding/6 ~ shares, data = actual_data, xlim = c(0, 15),
    ylim = c(0, 1), xlab = "shares", las = 1, ylab = " Probability of funding",
    pch = 19)
abline(lm(funding/6 ~ shares, data = actual_data))
```



#logistic regression model between funding and facevalue

```
mod1 <- glm(actual_data$funding ~ actual_data$facevalue, family = binomial(link = logit)

summary(mod1)
```

```
##
## Call:
## glm(formula = actual_data$funding ~ actual_data$facevalue, family = binomial(link = l
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.08   -1.08   -1.07    1.28    1.30
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.232762   0.130364   -1.79    0.074
## actual_data$facevalue -0.000342   0.003496   -0.10    0.922
##
## (Intercept)           .
## actual_data$facevalue
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

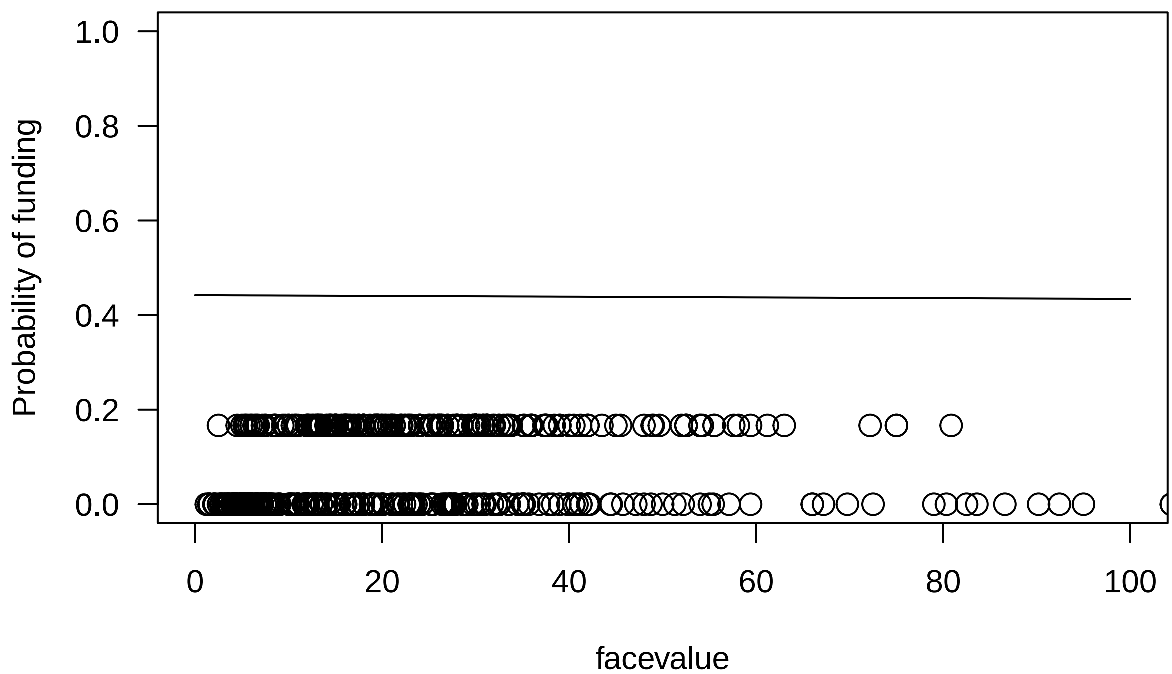
```
##     Null deviance: 661.20  on 481  degrees of freedom
## Residual deviance: 661.19  on 480  degrees of freedom
## AIC: 665.2
##
## Number of Fisher Scoring iterations: 3
```

```
pchisq(deviance(mod1), df.residual(mod1), lower.tail = FALSE)
```

```
## [1] 7.1323e-08
```

#plotting the fit curve

```
plot(funding/6 ~ facevalue, data = actual_data, xlim = c(0, 100),
    las = 1, cex = 1.5, ylim = c(0, 1), xlab = "facevalue", ylab = "Probability of fundi
x <- seq(0, 100, 1)
lines(x, ilogit(-0.232726 - 0.00032 * x))
```



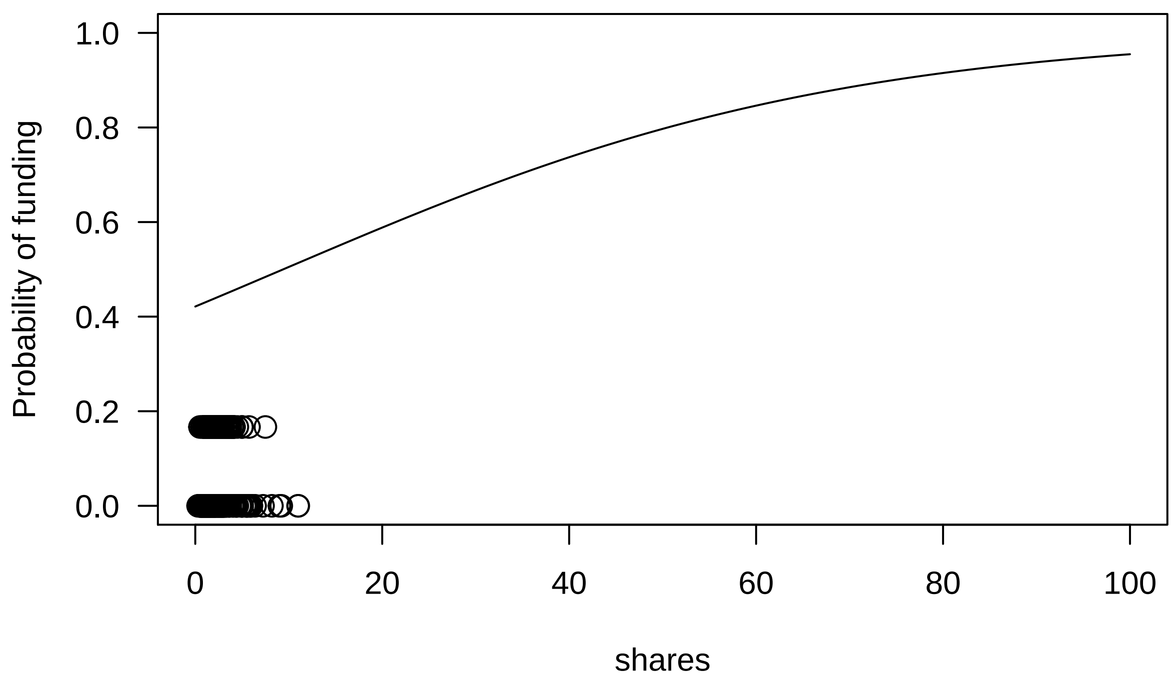
#logistic relation between number of shares and funding

```
mod2 <- glm(actual_data$funding ~ actual_data$shares, family = binomial(link = logit))

summary(mod2)
```

```
##
## Call:
## glm(formula = actual_data$funding ~ actual_data$shares, family = binomial(link = logi
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.20   -1.07   -1.06    1.28    1.31
```

```
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.3170     0.1713   -1.85    0.064 .
## actual_data$shares     0.0337     0.0647    0.52    0.603
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 661.20  on 481  degrees of freedom
## Residual deviance: 660.93  on 480  degrees of freedom
## AIC: 664.9
##
## Number of Fisher Scoring iterations: 3
```

```
plot(funding/6 ~ shares, data = actual_data, xlim = c(0, 100),
    las = 1, cex = 1.5, ylim = c(0, 1), xlab = "shares", ylab = "Probability of funding"
x <- seq(0, 100, 1)
lines(x, ilogit(-0.317 + 0.0337 * x))
```




```
pchisq(deviance(mod2), df.residual(mod2), lower.tail = FALSE)
```

```
## [1] 7.4019e-08
```

#logistic relation between buyout and funding

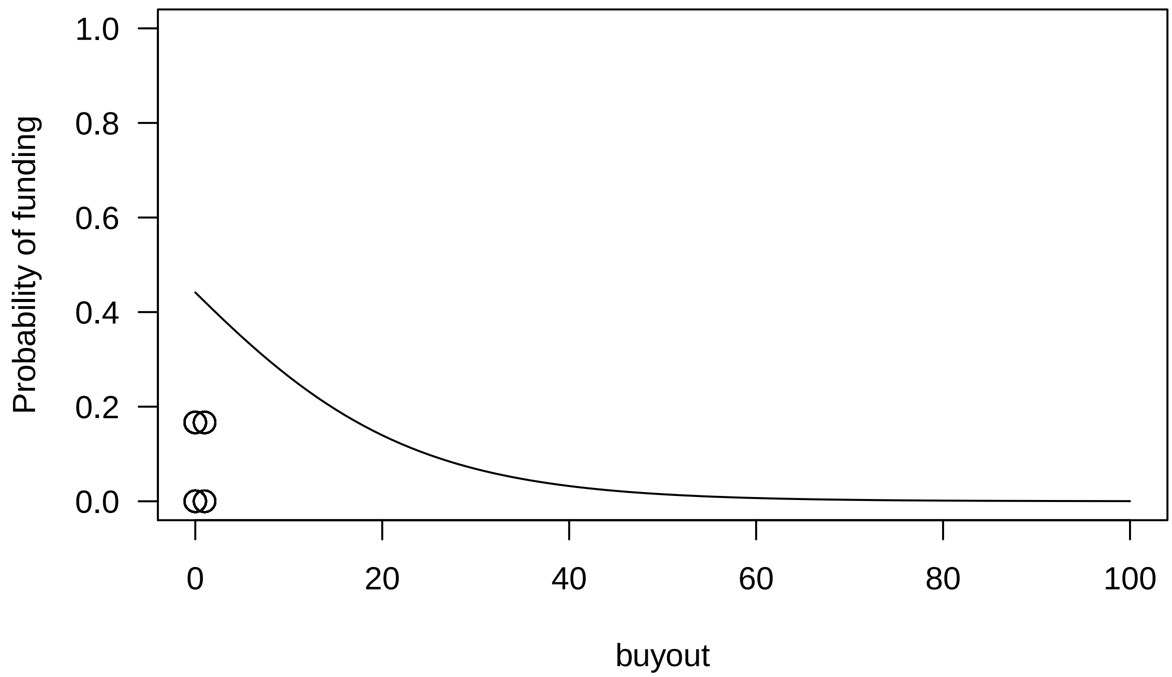
```r
mod3 <- glm(actual_data$funding ~ actual_data$buyout, family = binomial(link = logit))

summary(mod3)
```

```
##
## Call:
## glm(formula = actual_data$funding ~ actual_data$buyout, family = binomial(link = logi
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.08   -1.08   -1.08    1.28    1.31
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.2345     0.0963   -2.43    0.015 *
## actual_data$buyout  -0.0792     0.3168   -0.25    0.803
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 661.20  on 481  degrees of freedom
## Residual deviance: 661.14  on 480  degrees of freedom
## AIC: 665.1
##
## Number of Fisher Scoring iterations: 3
```

```r
pchisq(deviance(mod3), df.residual(mod3), lower.tail = FALSE)
```

```
## [1] 7.1863e-08
```

```r
plot(funding/6 ~ buyout, data = actual_data, xlim = c(0, 100),
    las = 1, cex = 1.5, ylim = c(0, 1), xlab = "buyout", ylab = "Probability of funding"
x <- seq(0, 100, 1)
lines(x, ilogit(-0.2345 - 0.0792 * x))
```




```r
pchisq(deviance(mod3), df.residual(mod3), lower.tail = FALSE)
```

```
## [1] 7.1863e-08
```

#The main multiple regression logistic model

```r
log_mod <- glm(actual_data$funding ~ actual_data$facevalue +
    actual_data$shares + actual_data$buyout, family = binomial(link = logit))

summary(log_mod)
```

```
##
## Call:
## glm(formula = actual_data$funding ~ actual_data$facevalue + actual_data$shares +
##     actual_data$buyout, family = binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.56   -1.07   -1.03    1.28    1.39
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.41997    0.19391   -2.17     0.03
## actual_data$facevalue -0.01009    0.00809   -1.25     0.21
## actual_data$shares     0.20291    0.14936    1.36     0.17
## actual_data$buyout    -0.07578    0.31860   -0.24     0.81
##
## (Intercept)           *
```

```
## actual_data$facevalue
## actual_data$shares
## actual_data$buyout
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 661.20  on 481  degrees of freedom
## Residual deviance: 659.24  on 478  degrees of freedom
## AIC: 667.2
##
## Number of Fisher Scoring iterations: 4
```

#We check the fit

```
pchisq(deviance(log_mod), df.residual(log_mod), lower.tail = FALSE)
```

```
## [1] 6.7457e-08
```

#we check by excluding buyout

```
log_mod1 <- glm(actual_data$funding ~ actual_data$facevalue +
    actual_data$shares, family = binomial(link = logit))

summary(log_mod1)
```

```
##
## Call:
## glm(formula = actual_data$funding ~ actual_data$facevalue + actual_data$shares,
##     family = binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.55   -1.07   -1.03    1.28    1.39
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.42518    0.19271   -2.21    0.027
## actual_data$facevalue -0.01017    0.00808   -1.26    0.208
## actual_data$shares     0.20306    0.14939    1.36    0.174
##
## (Intercept)           *
## actual_data$facevalue
## actual_data$shares
## ---
```

```
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 661.2  on 481  degrees of freedom
## Residual deviance: 659.3  on 479  degrees of freedom
## AIC: 665.3
##
## Number of Fisher Scoring iterations: 4
```

#check fit

```
pchisq(deviance(log_mod1), df.residual(log_mod1), lower.tail = FALSE)
```

```
## [1] 7.9032e-08
```

#We check excluding number of shares

```
log_mod2 <- glm(actual_data$funding ~ actual_data$facevalue +
    actual_data$buyout, family = binomial(link = logit))

summary(log_mod2)
```

```
##
## Call:
## glm(formula = actual_data$funding ~ actual_data$facevalue + actual_data$buyout,
##     family = binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.08   -1.08   -1.07    1.28    1.32
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.22753    0.13211   -1.72    0.085
## actual_data$facevalue -0.00027    0.00351   -0.08    0.939
## actual_data$buyout    -0.07710    0.31796   -0.24    0.808
##
## (Intercept)           .
## actual_data$facevalue
## actual_data$buyout
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 661.20  on 481  degrees of freedom
## Residual deviance: 661.13  on 479  degrees of freedom
## AIC: 667.1
##
## Number of Fisher Scoring iterations: 3
```

#check fit

```
pchisq(deviance(log_mod2), df.residual(log_mod2), lower.tail = FALSE)
```

```
## [1] 6.0875e-08
```

#excluding facevalue

```
log_mod3 <- glm(actual_data$funding ~ actual_data$shares + actual_data$buyout,
    family = binomial(link = logit))

summary(log_mod3)
```

```
##
## Call:
## glm(formula = actual_data$funding ~ actual_data$shares + actual_data$buyout,
##     family = binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.21   -1.07   -1.06    1.28    1.33
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.3117     0.1723   -1.81     0.07 .
## actual_data$shares    0.0351     0.0649    0.54     0.59
## actual_data$buyout   -0.0922     0.3178   -0.29     0.77
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 661.20  on 481  degrees of freedom
## Residual deviance: 660.84  on 479  degrees of freedom
## AIC: 666.8
##
## Number of Fisher Scoring iterations: 3
```

#t-test

```r
pchisq(deviance(log_mod3), df.residual(log_mod3), lower.tail = FALSE)
```

```
## [1] 6.3418e-08
```

```r
xtable(summary(log_mod))
```

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & z value & Pr($>$$|$z$|$) \\
##   \hline
## (Intercept) & -0.4200 & 0.1939 & -2.17 & 0.0303 \\
##   actual\_data\$facevalue & -0.0101 & 0.0081 & -1.25 & 0.2121 \\
##   actual\_data\$shares & 0.2029 & 0.1494 & 1.36 & 0.1743 \\
##   actual\_data\$buyout & -0.0758 & 0.3186 & -0.24 & 0.8120 \\
##    \hline
## \end{tabular}
## \end{table}
```

Bibliography

[1]Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2014), Applied Linear Statistical Models, 5th ed., McGraw-Hill Irwin.
[2] Improve coding in latex.https://www.overleaf.com/learn/latex/
[3] Improve coding in R. https://community.rstudio.com/
[4] Check installed packages in R. Available online. https://stackoverflow.com/questions/9341635/check-for-installed-packages-before-running-install-packages
[5]Statistics by Jim. Available online. https://statisticsbyjim.com/