Introduction to Medical Statistics 2026
Exercises Class 8
Logistic regression

Nguyen Lam Vuong and the Biostatistics team

2026-03-27

Exercise 1: Univariable logistic regression

This exercise uses the dataset **cmTbmData.csv**, which contains information on 201 patients with meningitis from 4 different patient groups. For this exercise, we will restrict attention to HIV-positive patients and explore how the CSF white cell count affects the probability of having TBM (compared to having CM). For this exercise, you need to load the packages **ggplot2**, **gtsummary** and **ggeffects**.

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(gtsummary)
```

```
Warning: package 'gtsummary' was built under R version 4.4.3
```

```
library(ggeffects)
```

```
Warning: package 'ggeffects' was built under R version 4.4.3
```

a) Import the dataset (select "stringsAsFactors") and create a new dataset which contains only HIV-positive patients. Because csfwcc has a rather skewed distribution, we create a new variable log2.csfwcc in the dataset which contains the log2-transformed values (more precisely use log2(csfwcc+1) to deal with counts of zero).

Create a boxplot of log2.csfwcc by diagnosis (CM and TBM) to get a first visual impression of the data. Add the individual measurements to the boxplot.

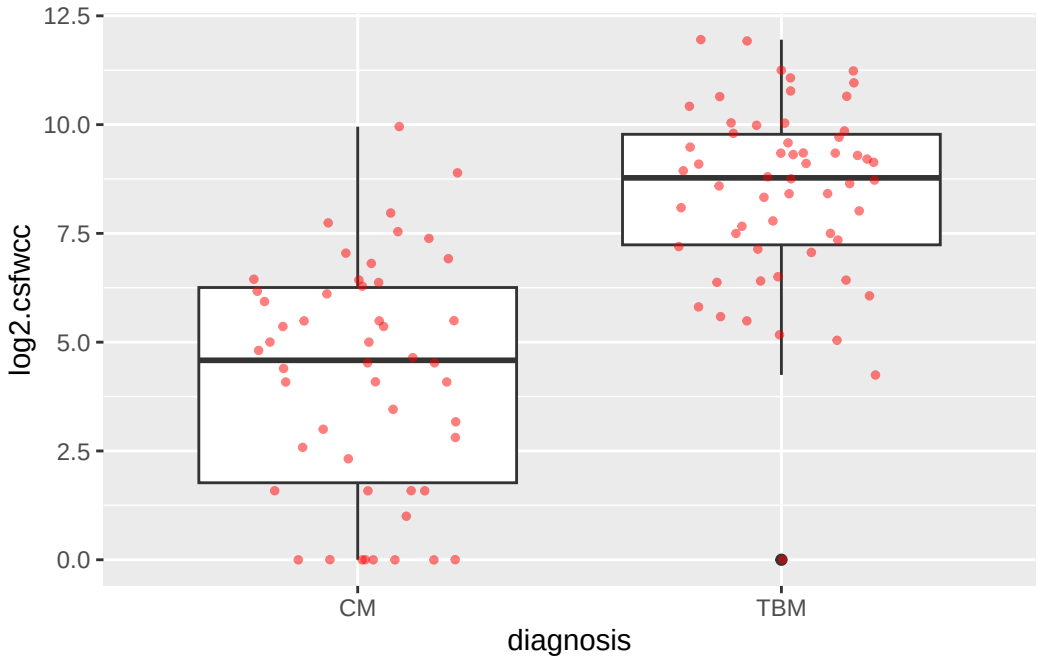
```
# Import data
cm.tbm <- read.csv("https://raw.githubusercontent.com/oucru-biostats/IntroductionToBio

# a)
# create a subset of HIV-positive patients
cm.tbm.hiv <- subset(cm.tbm, (hiv == 1) )
cm.tbm.hiv$log2.csfwcc <- log2(cm.tbm.hiv$csfwcc + 1) # add +1 to cope with 0's
# or use cm.tbm.hiv <- mutate(cm.tbm.hiv, log2.csfwcc=log2(cm.tbm.hiv$csfwcc + 1))

# describe log2.csfwcc by diagnosis
ggplot(cm.tbm.hiv, aes(diagnosis, log2.csfwcc)) + geom_boxplot() +
  geom_jitter(size = 1, alpha = 0.5, width = 0.25, colour = 'red')
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



Answer: *The log-transformed data are fairly symmetric, although there are quite a few individuals with log-value zero in the CM group as well as one outlier of zero in the TBM group. Note that csfwcc is the independent variable. Therefore, what matters is the relation with the outcome (e.g. is it linear on the logit scale), not whether it has a normal distribution. So the log transformation may very well be fine.*

b) How does log2.csfwcc affect the probability of having TBM compared to CM? Perform a univariable logistic regression, summarize the model fit and interpret the resulting odds ratio.

The logistic regression model as implemented in the *glm* function requires the outcome to be a variable of the R type *factor* or a variable with values 0 or 1. If you didn't specify *stringsAsFactors = TRUE* you will get an error message. It may not be clear to you which diagnosis is interpreted as "0" (the reference value "no event") and which as "1" (the "event" value). By default, this is determined by alphabetical order of the levels: the first level acts as reference or "no event", the second level is the "event". Hence, CM is the reference, and we model the probability to have TBM as event. Another approach is to create a variable 0 for CM patients and 1 for TBM patients and then use this as the outcome. Try both approaches.

```
# cm.tbm.hiv$diagnosis <- factor(cm.tbm.hiv$diagnosis)
fit1 <- glm(diagnosis ~ log2.csfwcc, data = cm.tbm.hiv, family = binomial)
# summarize model
summary(fit1)
```

```
Call:
glm(formula = diagnosis ~ log2.csfwcc, family = binomial, data = cm.tbm.hiv)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.6370     0.9739  -4.761 1.92e-06 ***
log2.csfwcc   0.7262     0.1397   5.198 2.01e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.127  on 107  degrees of freedom
Residual deviance:  89.621  on 106  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 93.621

Number of Fisher Scoring iterations: 5
```

```
confint(fit1)
```

```
Waiting for profiling to be done...

                 2.5 %    97.5 %
(Intercept) -6.7791280 -2.934115
log2.csfwcc  0.4828391  1.035422
```

```
# the summary and confint present results on the logit scale
# if we want the OR and its confidence interval, we need to exponentiate these numbers
# using standard R code we write
exp(c(coef(fit1)[2],confint(fit1)[2,]))
```

```
Waiting for profiling to be done...

log2.csfwcc       2.5 %      97.5 %
   2.067237    1.620669    2.816296
```

Characteristic	OR	95% CI	p-value
log2.csfwcc	2.07	1.62, 2.82	<0.001

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

```
# or in a formatted table
tbl_regression(fit1, exponentiate=TRUE)
```

```
# alternative approach for outcome: create new variable tbm which 0 for CM and 1 for T
cm.tbm.hiv$tbm <- ifelse(cm.tbm.hiv$diagnosis == "TBM", 1, 0)
fit2 <- glm(tbm ~ log2.csfwcc, data = cm.tbm.hiv, family = binomial)
```

Answer: *Logistic regression model says that the odds of being a TBM patient increases by a factor of 2.07 (95% CI 1.62-2.82) for each doubling of CSF white cell count (note that we used the 2-log of CSF white cell count, not the 10-log). The p<0.001 strongly suggests there is an association.*

c) Based on the model from b), what is the predicted probability that a subject with log2.csfwcc=6 has TBM? Calculate the answer to this question in 3 ways:

1) "By hand" based on the regression coefficients and the logistic regression model (slide 16).

```
# prediction for new patients with log2.csfwcc = 6
lp <- -4.6370 + 0.7262 * 6 # a + b * x based on the regression coefficients
# or better without rounding of intermediate values
lp <- coef(fit1)[1] + coef(fit1)[2]*6
exp(lp)/(1+exp(lp))
```

```
(Intercept)
  0.4305184
```

Answer: *Based on the logistic regression output, the intercept is -4.6370 and the slope is 0.7262. For the patient, this gives a predicted risk of exp(-4.6370+ 0.7262*6*)/(1+ exp(-4.6370+ 0.7262*6*))=0.43. This corresponds with the result obtained via the other two approaches.*

2) Using the *predict* function

```
predict(fit1, newdata=data.frame(log2.csfwcc=6), type="response")
```

```
        1
0.4305184
```

3) With the help of the *ggpredict* function from the *ggeffects* package, which automatically adds confidence intervals

Make a figure that plots the probability to have TBM for all values of csfwcc. Apply the *plot* function to the output from the *ggpredict* function.

```
pred <- ggpredict(fit2)
```

```
Data were 'prettified'. Consider using `terms="log2.csfwcc [all]"` to
  get smooth plots.
```

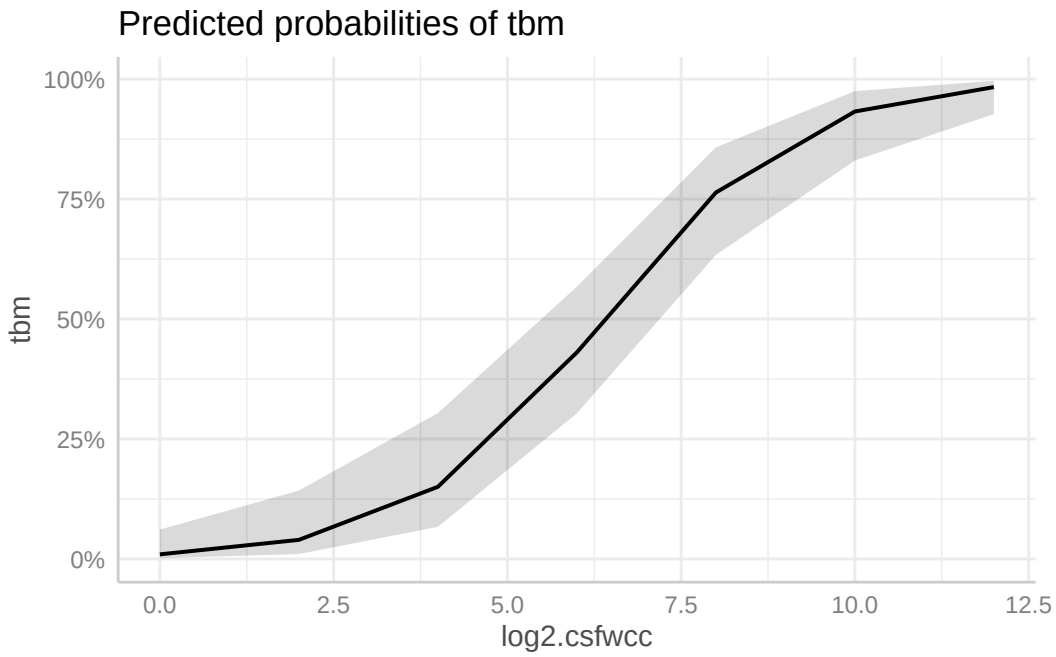
```
pred
```

```
$log2.csfwcc
# Predicted probabilities of tbm

log2.csfwcc | Predicted |     95% CI
-----------------------------------
          0 |      0.01 | 0.00, 0.06
          2 |      0.04 | 0.01, 0.14
          4 |      0.15 | 0.07, 0.30
          6 |      0.43 | 0.30, 0.57
          8 |      0.76 | 0.63, 0.86
         10 |      0.93 | 0.83, 0.97
         12 |      0.98 | 0.93, 1.00


attr(,"class")
[1] "ggalleffects" "list"
attr(,"model.name")
[1] "fit2"
```

```r
plot(pred) # plots at values 0,2,4,6,8,10,12
```

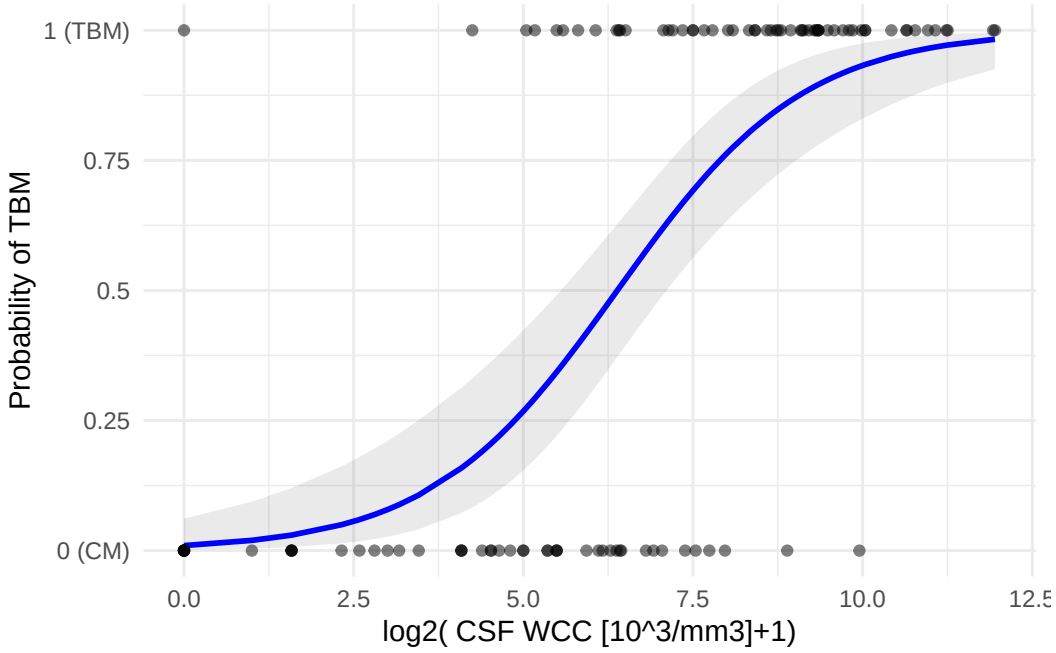



```r
# we can make a slightly more beautiful figure via

pred_data <- ggpredict(fit2, terms = "log2.csfwcc [all]")

ggplot(pred_data, aes(x = x, y = predicted)) +
  # Add the confidence interval ribbon
  geom_ribbon(aes(ymin = conf.low, ymax = conf.high), alpha = 0.1) +
  # Add the prediction line
  geom_line(color = "blue", size = 1) +
  # Add the raw data points (this replaces add.data = TRUE)
  geom_point(data = attr(pred_data, "rawdata"), aes(x = x, y = response), alpha = 0.5)
  # Apply your custom scales
  scale_y_continuous(
    name = "Probability of TBM",
    breaks = c(0, 0.25, 0.5, 0.75, 1),
    labels = c("0 (CM)", "0.25", "0.5", "0.75", "1 (TBM)")
  ) +
  xlab("log2( CSF WCC [10^3/mm3]+1)") +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```




```
# using fit1 doesn't work if we want to add the raw data
```

Exercise 2: Multivariable logistic regression

Prediction of re-shock in patients with DSS (dengue shock syndrome)

The dataset **DF.csv** contains data from 2007 children with DSS which were recruited into the DF study between 2003-2009. For this exercise, we aim to predict the occurrence of re-shock based on the 268 subjects recruited into the DF study in 2009.

a) Import the **DF.csv** dataset and create a new dataset *df2009* which contains only the 268 subjects recruited in 2009 and the variables:

- Outcome Y: re-shock (reshock)
- Covariables X (measured at onset of shock): Age (*age*), *sex*, day of illness at shock (*day_ill*), temperature (*temp*), platelet count (*plt*), hematocrit (*hct*).

Look at descriptive statistics for the outcome and the covariables using the *summary* function. How many re-shocks occur in the 268 subjects? Does any of the covariables have missing values? Do you notice something peculiar about the temperature values?

Additionally make a summary of the covariables by outcome value, using the tbl_summary function from the gtsummary package. Do you recommend a log-transformation of the platelet count or hematocrit?

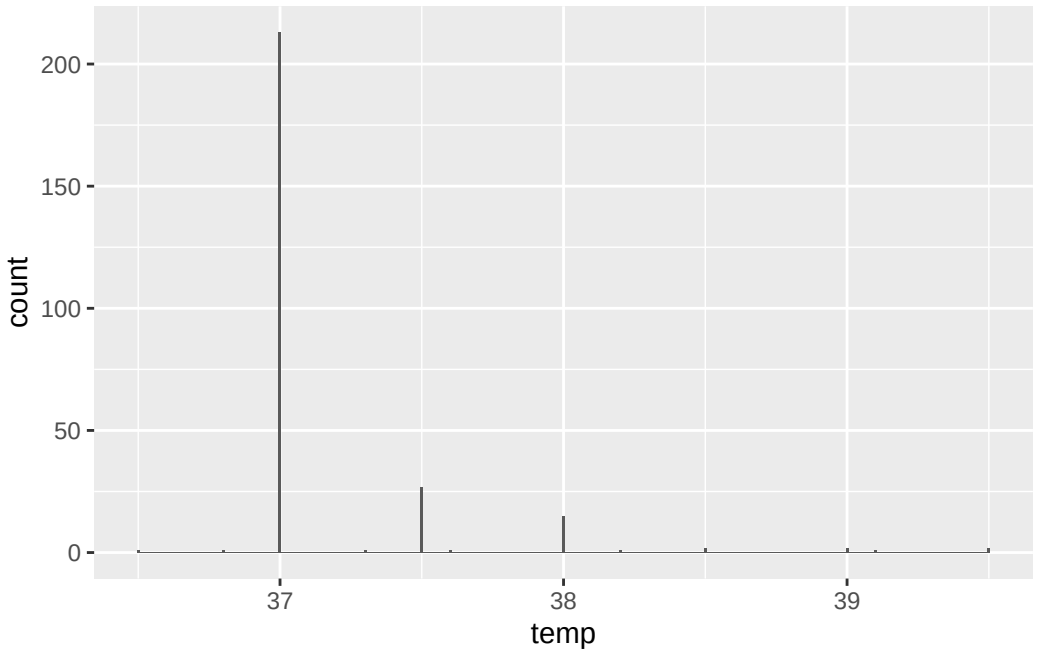
```
# Import data
dfstudy <- read.csv("https://raw.githubusercontent.com/oucru-biostats/IntroductionToBi
df2009 <- subset(dfstudy, year == 2009)
df2009 <- subset(df2009, select=c(age, sex, day_ill, temp, plt, hct, reshock))


# quick descriptive stats
summary(df2009)
```

```
      age             sex         day_ill          temp            plt
 Min.   : 1.000   Female:119   Min.   :3.000   Min.   :36.50   Min.   :  7010
 1st Qu.: 7.000   Male  :149   1st Qu.:5.000   1st Qu.:37.00   1st Qu.: 26000
 Median : 9.000                Median :5.000   Median :37.00   Median : 39000
 Mean   : 9.306                Mean   :5.172   Mean   :37.16   Mean   : 43650
 3rd Qu.:12.000                3rd Qu.:6.000   3rd Qu.:37.00   3rd Qu.: 53775
 Max.   :14.000                Max.   :7.000   Max.   :39.50   Max.   :196000
                                               NA's   :1       NA's   :2
      hct        reshock
 Min.   :40.00   No :188
 1st Qu.:47.00   Yes: 80
 Median :50.00
 Mean   :49.93
 3rd Qu.:53.00
 Max.   :60.00
```

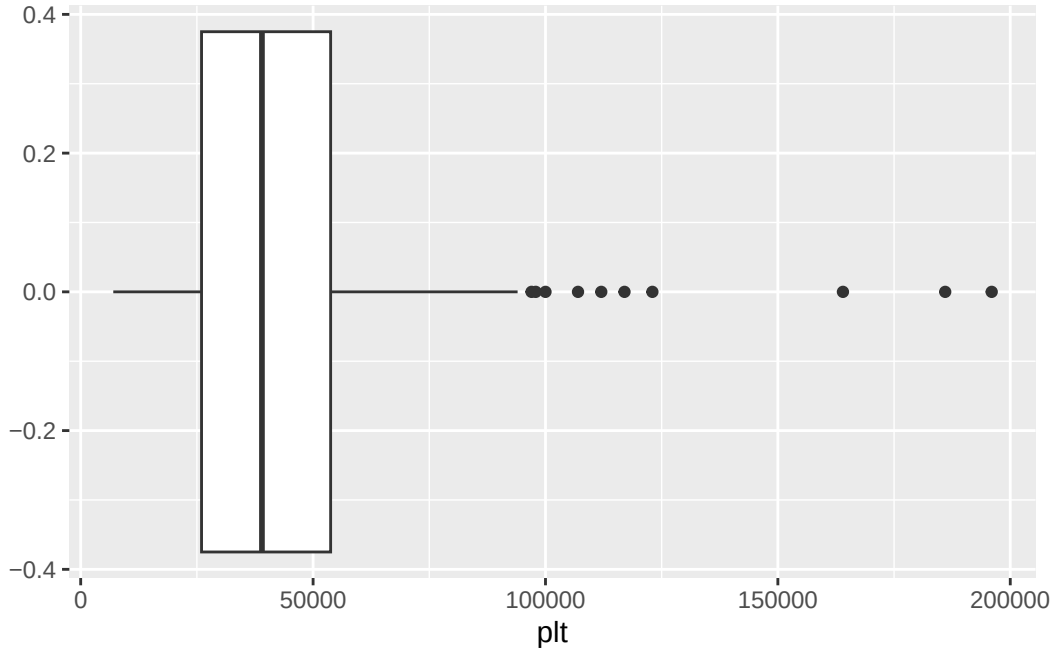
```
## almost all temperatures are 37.00 degrees
ggplot(df2009,aes(temp)) + geom_histogram(binwidth=0.01)
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_bin()`).
```




```
## platelet has a somewhat skewed distribution, or at least some very high values
ggplot(df2009,aes(plt)) + geom_boxplot()
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Characteristic	**No** N = 188[1]	**Yes** N = 80[1]
age	10 (7, 12)	8 (6, 11)
sex		
Female	83 (44%)	36 (45%)
Male	105 (56%)	44 (55%)
day_ill		
3	3 (1.6%)	0 (0%)
4	29 (15%)	27 (34%)
5	76 (40%)	34 (43%)
6	71 (38%)	19 (24%)
7	9 (4.8%)	0 (0%)
temp	37.00 (37.00, 37.00)	37.00 (37.00, 37.50)
Unknown	0	1
plt	38,000 (25,000, 52,000)	40,000 (29,000, 63,000)
Unknown	1	1
hct	50.0 (47.0, 52.5)	50.0 (47.0, 53.0)

[1]Median (Q1, Q3); n (%)

Answer: *80/268 (30%) subjects had re-shock. Only 3 missing values (1 for temp, 2 for plt).*

```
# by group using gtsummary
tbl_summary(data=df2009, by=reshock)
```

b) Perform univariable logistic regressions for each covariable separately on outcome and interpret the results. Which covariables clearly show an association with the occurrence of re-shock?

```
## we log-transform platelet and the further results are based on log-transformed pl
## (though we will see that this does not really affect the results).
## Also, we need to use reshock as a 0-1 variable or a variable of type factor.
df2009$log2.plt <- log2(df2009$plt)

# Univariable regression
summary(glm(reshock ~ age, data = df2009, family = binomial))
```

```
Call:
```

```
glm(formula = reshock ~ age, family = binomial, data = df2009)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4275     0.4176   1.024  0.30596
age          -0.1418     0.0449  -3.157  0.00159 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 326.74  on 267  degrees of freedom
Residual deviance: 316.32  on 266  degrees of freedom
AIC: 320.32

Number of Fisher Scoring iterations: 4
```

```
summary(glm(reshock ~ sex, data = df2009, family = binomial))
```

```
Call:
glm(formula = reshock ~ sex, family = binomial, data = df2009)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.83532    0.19956  -4.186 2.84e-05 ***
sexMale     -0.03445    0.26847  -0.128    0.898
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 326.74  on 267  degrees of freedom
Residual deviance: 326.73  on 266  degrees of freedom
AIC: 330.73

Number of Fisher Scoring iterations: 4
```

```r
summary(glm(reshock ~ day_ill, data = df2009, family = binomial))
```

```
Call:
glm(formula = reshock ~ day_ill, family = binomial, data = df2009)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0780     0.8583   2.421 0.015482 *
day_ill      -0.5756     0.1687  -3.411 0.000646 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 326.74  on 267  degrees of freedom
Residual deviance: 314.45  on 266  degrees of freedom
AIC: 318.45

Number of Fisher Scoring iterations: 4
```

```r
summary(glm(reshock ~ temp, data = df2009, family = binomial))
```

```
Call:
glm(formula = reshock ~ temp, family = binomial, data = df2009)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -27.6017    11.4947  -2.401   0.0163 *
temp          0.7190     0.3091   2.327   0.0200 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 324.32  on 266  degrees of freedom
Residual deviance: 318.79  on 265  degrees of freedom
  (1 observation deleted due to missingness)
```

```
AIC: 322.79

Number of Fisher Scoring iterations: 4
```

```
summary(glm(reshock ~ plt, data = df2009, family = binomial)) # untransformed plt
```

```

Call:
glm(formula = reshock ~ plt, family = binomial, data = df2009)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.073e+00  2.563e-01  -4.188 2.82e-05 ***
plt          4.777e-06  4.863e-06   0.982    0.326
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 323.61  on 265  degrees of freedom
Residual deviance: 322.66  on 264  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 326.66

Number of Fisher Scoring iterations: 4
```

```
summary(glm(reshock ~ log2.plt, data = df2009, family = binomial)) # log-transformed p
```

```

Call:
glm(formula = reshock ~ log2.plt, family = binomial, data = df2009)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2360     2.4729  -1.713   0.0867 .
log2.plt      0.2219     0.1620   1.369   0.1709
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 323.61  on 265  degrees of freedom
Residual deviance: 321.70  on 264  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 325.7

Number of Fisher Scoring iterations: 4
```

```r
summary(glm(reshock ~ hct, data = df2009, family = binomial))
```

```

Call:
glm(formula = reshock ~ hct, family = binomial, data = df2009)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.32268    1.72087  -1.931   0.0535 .
hct          0.04929    0.03416   1.443   0.1490
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 326.74  on 267  degrees of freedom
Residual deviance: 324.64  on 266  degrees of freedom
AIC: 328.64

Number of Fisher Scoring iterations: 4
```

```r
# We can do it all at once via the tbl_uvregression function from the gtsummary packag
tbl_uvregression(data=df2009, method=glm, y=reshock, method.args=list(family=binomial)
```

Note the values of plt on the original scale. What is happening here? Do you have a suggestion to change this.

c) Perform a multivariable logistic regression with all covariables jointly and interpret the results. Which covariables are significant after adjustment for all others and what is their effect size?

Characteristic	N	OR	95% CI	p-value
age	268	0.87	0.79, 0.95	0.002
sex	268			
Female		—	—	
Male		0.97	0.57, 1.64	0.9
day_ill	268	0.56	0.40, 0.78	<0.001
temp	267	2.05	1.13, 3.85	0.020
plt	266	1.00	1.00, 1.00	0.3
hct	268	1.05	0.98, 1.12	0.15
log2.plt	266	1.25	0.91, 1.73	0.2

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Characteristic	OR	95% CI	p-value
age	0.87	0.79, 0.95	0.004
sex			
Female	—	—	
Male	1.05	0.59, 1.89	0.9
day_ill	0.63	0.44, 0.89	0.009
temp	2.11	1.11, 4.19	0.026
log2(plt)	1.11	0.76, 1.61	0.6
hct	1.09	1.01, 1.18	0.031

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

```
# Multivariable regression
fit <- glm(reshock ~ age + sex + day_ill + temp + log2(plt) + hct, data = df2009, fami
tbl_regression(fit, exponentiate=TRUE)
```

Answer: *Univariable and multivariable analyses do not lead to substantially different results. The following covariables are associated with an increased risk of shock: younger age and earlier day of illness (strong association), higher temperature (fairly strong) and higher hematocrit (especially in the multivariable model).*

d) Based on c), age and day of illness are two important predictors of re-shock. Is there any evidence that age or the day of illness at shock affect the outcome non-linearly? Evaluate this by performing a test for the presence of a quadratic effect in each. Hint: compare both models using the anova function.

```r
fit.quad <- glm(reshock ~ poly(age, 2) + sex + poly(day_ill, 2) + temp + log2(plt) + h
anova(fit, fit.quad, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: reshock ~ age + sex + day_ill + temp + log2(plt) + hct
Model 2: reshock ~ poly(age, 2) + sex + poly(day_ill, 2) + temp + log2(plt) +
    hct
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       258     293.03
2       256     290.09  2   2.9395     0.23
```

Answer: *No clear evidence for a quadratic effect (p=0.23).*

e) The effect of hematocrit on the risk of re-shock might be different for males and females. Does the data provide any evidence for an interaction between sex and hematocrit?

```r
# Test for interaction between sex and hct
fit.ia <- glm(reshock ~ age + hct*sex + day_ill + temp + log2(plt) , data = df2009, fa
anova(fit, fit.ia, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: reshock ~ age + sex + day_ill + temp + log2(plt) + hct
Model 2: reshock ~ age + hct * sex + day_ill + temp + log2(plt)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       258     293.03
2       257     291.81  1   1.2189   0.2696
```

```r
summary(fit.ia)
```

```
Call:
glm(formula = reshock ~ age + hct * sex + day_ill + temp + log2(plt),
    family = binomial, data = df2009)

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -33.39045   13.39653  -2.492  0.01269 *
age          -0.14407    0.04943  -2.915  0.00356 **
hct           0.13683    0.06235   2.195  0.02819 *
sexMale       4.35749    3.93759   1.107  0.26845
day_ill      -0.47483    0.17702  -2.682  0.00731 **
temp          0.76289    0.33367   2.286  0.02223 *
log2(plt)     0.06787    0.19093   0.355  0.72222
hct:sexMale  -0.08618    0.07849  -1.098  0.27222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 321.18  on 264  degrees of freedom
Residual deviance: 291.81  on 257  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 307.81

Number of Fisher Scoring iterations: 4
```

Answer: *No clear evidence for an interaction (p=0.27). Because the interaction concerns a single parameter, the summary function gives the same information with a slightly different p-value (using Wald test instead of likelihood ratio test).*