HW12 - Poisson Regression

Madhu Peduri

July 18, 2021

0.0.1 Problem 1

```
# Convert the doctors dataframe to appropriate format
doc <- data.frame(age = as.numeric(unclass(as.factor(doctors$age))),
    agesq = (as.numeric(unclass(as.factor(doctors$age))))^2,
    agecat = as.character(doctors$age), smoke = as.numeric(unclass(as.factor(doctors$smoking))),
    deaths = as.numeric(doctors$deaths), personyrs = as.numeric(doctors$"person-years"))
doc
```

```
##    age agesq   agecat smoke deaths personyrs
## 1    1     1 35 to 44     2     32     52407
## 2    2     4 45 to 54     2    104     43248
## 3    3     9 55 to 64     2    206     28612
## 4    4    16 65 to 74     2    186     12663
## 5    5    25 75 to 84     2    102      5317
## 6    1     1 35 to 44     1      2     18790
## 7    2     4 45 to 54     1     12     10673
## 8    3     9 55 to 64     1     28      5710
## 9    4    16 65 to 74     1     28      2585
## 10   5    25 75 to 84     1     31      1462
```

```
des(doc)
```

```
##
##  No. of observations =  10
##   Variable      Class           Description
## 1 age           numeric
## 2 agesq         numeric
## 3 agecat        character
## 4 smoke         numeric
## 5 deaths        numeric
## 6 personyrs     numeric
```

0.0.2 Problem 1(a)

```
# GLM model
poismod <- glm(deaths ~ age + agesq + smoke + smoke:age, offset = log(personyrs),
    family = poisson(link = "log"), data = doc)

sumpois <- summary(poismod)
sumpois
```

```
##
```

```
## Call:
## glm(formula = deaths ~ age + agesq + smoke + smoke:age, family = poisson(link = "log"),
##     data = doc, offset = log(personyrs))
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
##  0.4382  -0.2733  -0.1526   0.2339  -0.0570  -0.8305   0.1340   0.6411
##       9       10
## -0.4106  -0.0127
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.2327     0.7743  -15.80  < 2e-16 ***
## age           2.6840     0.2689    9.98  < 2e-16 ***
## agesq        -0.1977     0.0274   -7.22  5.1e-13 ***
## smoke         1.4410     0.3722    3.87  0.00011 ***
## age:smoke    -0.3075     0.0970   -3.17  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 935.0673  on 9  degrees of freedom
## Residual deviance:   1.6354  on 5  degrees of freedom
## AIC: 66.7
##
## Number of Fisher Scoring iterations: 4
```

```
summpois <- summ(poismod, confint = TRUE, digits = 3, ci.width = 0.95)

sprintf("Residual deviance of the poisson model : %.3f", sumpois$deviance)
```

```
## [1] "Residual deviance of the poisson model : 1.635"
```

```
sprintf("Residual degrees of freedom of the poisson model : %.3f",
    sumpois$df.residual)
```

```
## [1] "Residual degrees of freedom of the poisson model : 5.000"
```

```
print("Pseudo R2 (Mcfadden) of the model : 0.943")
```

```
## [1] "Pseudo R2 (Mcfadden) of the model : 0.943"
```

0.0.3 Problem 1(b)

```
sprintf("Null deviance of the model : %.3f", poismod$null.deviance)
```

```
## [1] "Null deviance of the model : 935.067"
```

```
sprintf("Deviance of the model residuals : %.3f", poismod$deviance)
```

```
## [1] "Deviance of the model residuals : 1.635"
```

```
devresd <- round(residuals(poismod, type = "deviance"), 3)
devresd
```

```
##      1      2      3      4      5      6      7      8      9     10
##  0.438 -0.273 -0.153  0.234 -0.057 -0.830  0.134  0.641 -0.411 -0.013
```

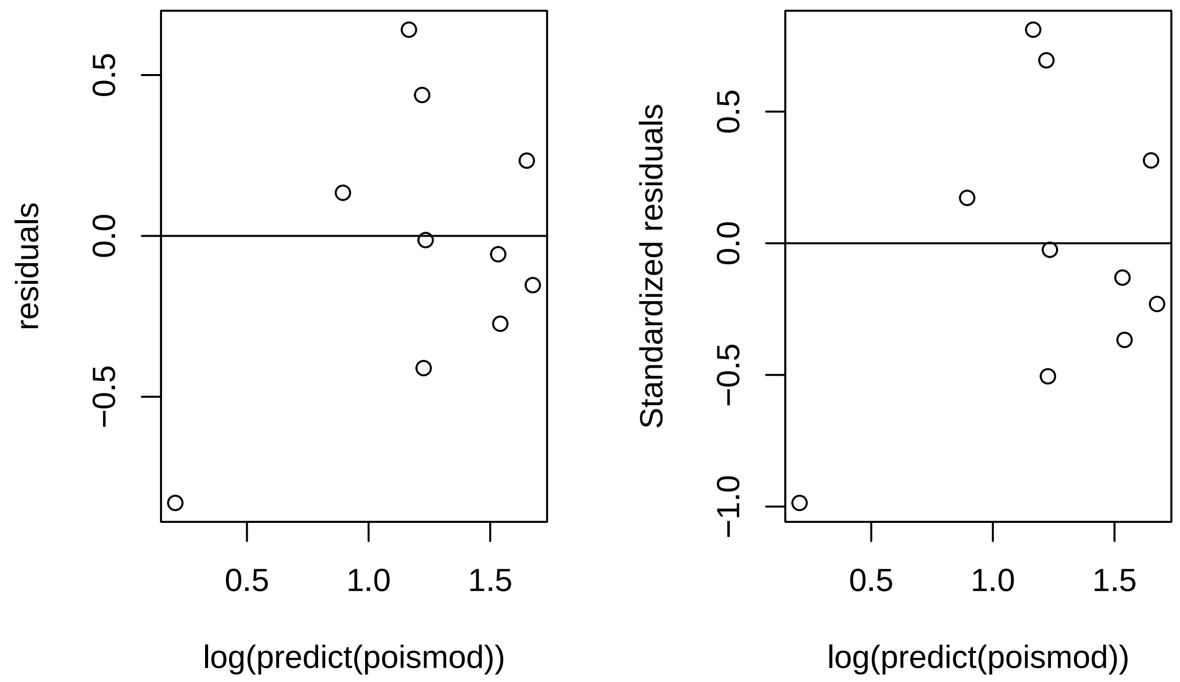
```r
stddevres <- rstandard(poismod)
stddevres
```

```
##       1       2       3       4       5       6       7       8       9      10
##  0.6947 -0.3669 -0.2306  0.3146 -0.1304 -0.9860  0.1723  0.8110 -0.5052 -0.0248
```

```r
sprintf("Chi-square goodness-of-fit value : %.3f", sum(devresd^2))
```

```
## [1] "Chi-square goodness-of-fit value : 1.635"
```

```r
par(mfrow = c(1, 2))
plot(log(predict(poismod)), devresd, ylab = "residuals")
abline(h = 0)
plot(log(predict(poismod)), stddevres, ylab = "Standardized residuals")
abline(h = 0)
```



```r
# standardized Pearson's residuals
pearson.resid <- round(resid(poismod, type = "pearson"), 3)
pearson.resid
```

```
##      1      2      3      4      5      6      7      8      9     10
##  0.444 -0.272 -0.152  0.235 -0.057 -0.766  0.135  0.655 -0.405 -0.013
```

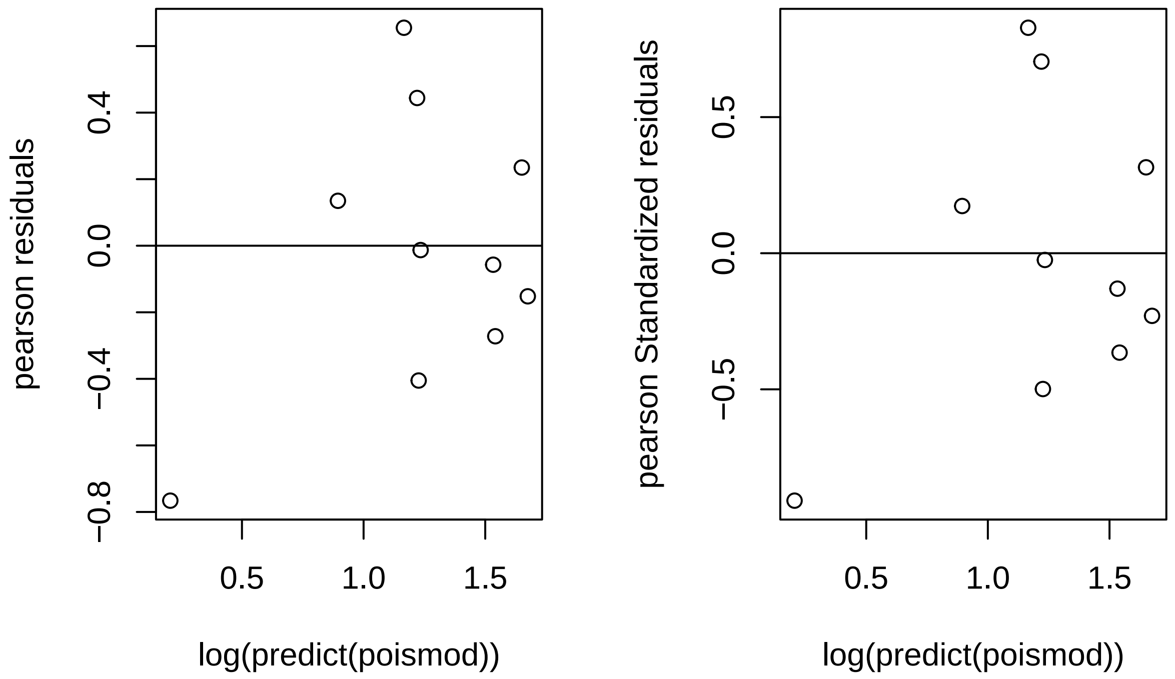
```r
stdpearsons <- rstandard(poismod, type = "pearson")
stdpearsons
```

```
##       1       2       3       4       5       6       7       8       9      10
##  0.7040 -0.3652 -0.2302  0.3155 -0.1303 -0.9090  0.1734  0.8283 -0.4989 -0.0248
```

```
sprintf("Pearson goodness-of-fit value : %.3f", sum(pearson.resid^2))
```

```
## [1] "Pearson goodness-of-fit value : 1.551"
```

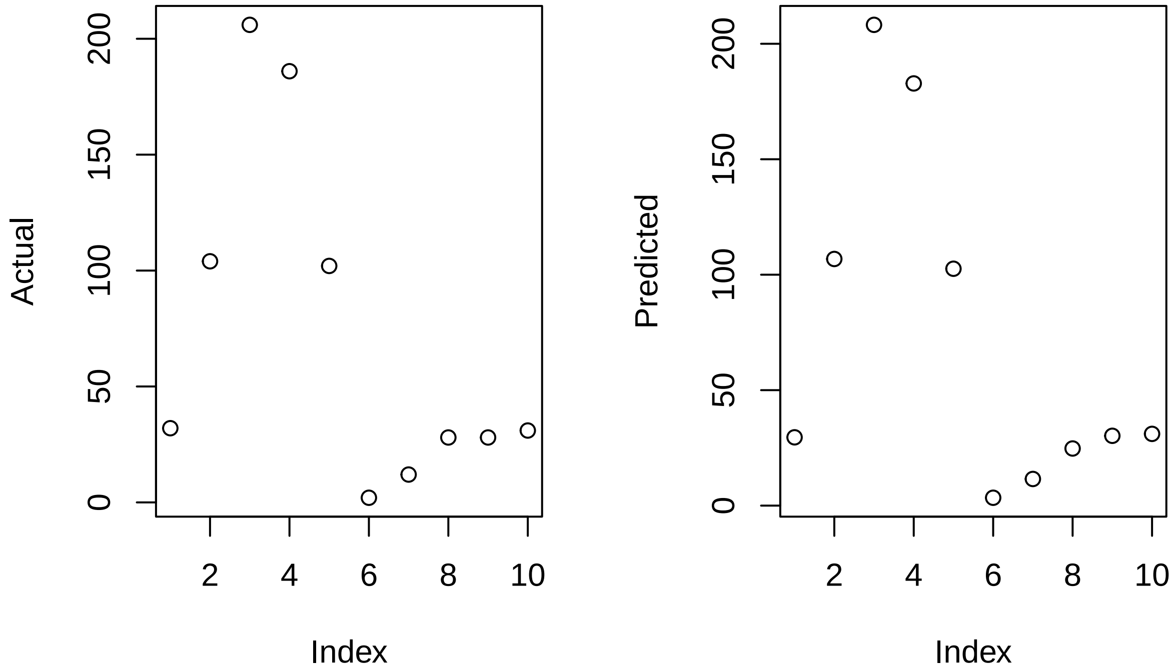
```
par(mfrow = c(1, 2))
plot(log(predict(poismod)), pearson.resid, ylab = "pearson residuals")
abline(h = 0)
plot(log(predict(poismod)), stdpearsons, ylab = "pearson Standardized residuals")
abline(h = 0)
```



```
# expected <- fitted.values(docm1, type='response')
expected <- round(predict(poismod, type = "response"), 2)
expected
```

```
##      1      2      3      4      5      6      7      8      9     10
##  29.58 106.81 208.20 182.83 102.58   3.41  11.54  24.74  30.23  31.07
```

```
par(mfrow = c(1, 2))
plot(doc$deaths, ylab = "Actual")
plot(expected, ylab = "Predicted")
```



0.0.3.1 Observations

- Outliers will be 2 standard deviations away from mean. We do not have standard residuals more than +2 or -2 suggesting no outliers.
- We do not see much difference between actual and predicted values.

0.0.4 Problem 1(c)

```
kable(cbind(doc[, c(1, 3, 4, 5)], expected, pearson.resid, devresd))
```

age	agecat	smoke	deaths	expected	pearson.resid	devresd
1	35 to 44	2	32	29.58	0.444	0.438
2	45 to 54	2	104	106.81	-0.272	-0.273
3	55 to 64	2	206	208.20	-0.152	-0.153
4	65 to 74	2	186	182.83	0.235	0.234
5	75 to 84	2	102	102.58	-0.057	-0.057
1	35 to 44	1	2	3.41	-0.766	-0.830
2	45 to 54	1	12	11.54	0.135	0.134
3	55 to 64	1	28	24.74	0.655	0.641
4	65 to 74	1	28	30.23	-0.405	-0.411

age	agecat	smoke	deaths	expected	pearson.resid	devresd
5	75 to 84	1	31	31.07	-0.013	-0.013

0.0.5 Problem 1(d)

```
# Chi square
pchisq(deviance(poismod), df.residual(poismod), lower = F)
```

```
## [1] 0.897
```

```
deviance(poismod)
```

```
## [1] 1.64
```

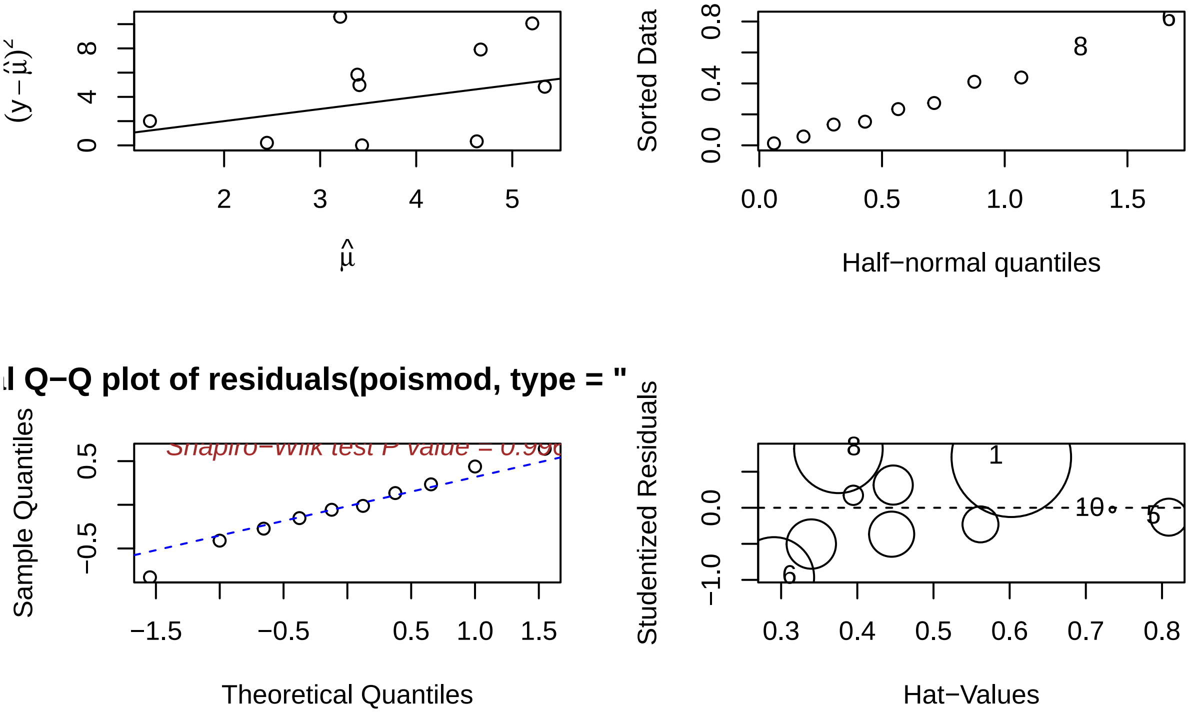
```
pr <- residuals(poismod, "pearson")
sum(pr^2)
```

```
## [1] 1.55
```

```
# poisson goodness of fit values
poisgof(poismod)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 1.64
##
## $df
## [1] 5
##
## $p.value
## [1] 0.897
```

```
# plots
par(mfrow = c(2, 2))
plot(log(fitted(poismod)), (doc$deaths - fitted(poismod))^2,
    xlab = expression(hat(mu)), ylab = expression((y - hat(mu))^2))
abline(0, 1)
halfnorm(residuals(poismod))
shapiro.qqnorm(residuals(poismod, type = "deviance"))
influencePlot(poismod)
```



```
##    StudRes   Hat   CookD
## 1   0.7003 0.602 0.15003
## 5  -0.1303 0.809 0.01436
## 6  -0.9643 0.291 0.06770
## 8   0.8175 0.375 0.08240
## 10 -0.0248 0.735 0.00034
```

```
# Confidence intervals of each predictor
kable(cbind(summary(poismod)$coef, confint(poismod)))
```

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

	Estimate	Std. Error	z value	Pr(>|z|)	2.5 %	97.5 %
(Intercept)	-12.233	0.774	-15.80	0.000	-13.820	-10.782
age	2.684	0.269	9.98	0.000	2.171	3.226
agesq	-0.198	0.027	-7.22	0.000	-0.252	-0.145
smoke	1.441	0.372	3.87	0.000	0.736	2.198
age:smoke	-0.308	0.097	-3.17	0.002	-0.501	-0.120

```
# Significance of each predictor relative to full model
kable(drop1(poismod, test = "F"))
```

```
## Warning in drop1.glm(poismod, test = "F"): F test assumes 'quasipoisson' family
```

	Df	Deviance	AIC	F value	Pr(>F)
	NA	1.64	66.7	NA	NA
agesq	1	59.89	123.0	178.1	0.000
age:smoke	1	12.18	75.2	32.2	0.002

```r
# Goodness of fit measures using anova
kable(anova(poismod, test = "Chisq"))
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	9	935.07	NA
age	1	850.1	8	85.01	0.000
agesq	1	61.0	7	24.03	0.000
smoke	1	11.9	6	12.18	0.001
age:smoke	1	10.5	5	1.64	0.001

```r
# mean and variance
phi <- sum(pr^2)/df.residual(poismod)
round(c(phi, sqrt(phi)), 4)
```

```
## [1] 0.310 0.557
```

```r
mean(fitted(poismod))
```

```
## [1] 73.1
```

```r
sqrt(var(fitted(poismod)))
```

```
## [1] 73.6
```

0.0.5.1 Observations

- Deviance: This explains the un-explained variance in our data. Lesser the deviance more good the model is. We have a less deviance of 1.64 suggesting good explainability of the variance.
- Chisquare: It gives an idea of how much the fitted values differ from the expected values. Lesser the chisquare more dood the model is. We have a less chisquare of 0.897 suggesting good fitted values.
- poisgof: The null hypothesis of this is a good goodness-of-fit measure for the poisson regression model. We have the p-value 0.897 > 0.05 significant value suggesting good fit of the model.
- Plots: We can see the normality of the residuals.
- drop1: We can see less AIC value for the predictor age*smoke suggesting it as good predictor.
- Anova: We can see the Pr values are small for all the predictors.
- Assumption of the poission model is that mean and variance of the fitted values should be same. We can see they are approximately near suggesting a good model.

0.1 Document Information.

All of the statistical analyses in this document will be performed using R version 4.1.0 (2021-05-18). R packages used will be maintained using the packrat dependency management system.

```r
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] Rcpp_1.0.7         jtools_2.1.3       dobson_0.4         Matrix_1.3-4
##  [5] psych_2.1.6        leaps_3.1          faraway_1.0.7      xtable_1.8-4
##  [9] lmtest_0.9-38      zoo_1.8-9          PairedData_1.1.1   mvtnorm_1.1-2
## [13] gld_2.6.2          ggpubr_0.4.0       car_3.0-11         carData_3.0-4
## [17] mnormt_2.0.2       vcd_1.4-8          epiDisplay_3.5.0.1 nnet_7.3-16
## [21] foreign_0.8-81     Hmisc_4.5-0        Formula_1.2-4      survival_3.2-11
## [25] lattice_0.20-44    MASS_7.3-54        ggplot2_3.3.5      rmarkdown_2.8
## [29] knitr_1.33
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-152        RColorBrewer_1.1-2  tools_4.1.0
##  [4] backports_1.2.1     utf8_1.2.1          R6_2.5.0
##  [7] rpart_4.1-15        colorspace_2.0-1    withr_2.4.2
## [10] tidyselect_1.1.1    gridExtra_2.3       curl_4.3.1
## [13] compiler_4.1.0      formatR_1.11        htmlTable_2.2.1
## [16] scales_1.1.1        checkmate_2.0.0     proxy_0.4-26
## [19] stringr_1.4.0       digest_0.6.27       minqa_1.2.4
## [22] rio_0.5.27          base64enc_0.1-3     jpeg_0.1-8.1
## [25] pkgconfig_2.0.3     htmltools_0.5.1.1   lme4_1.1-27.1
## [28] highr_0.9           htmlwidgets_1.5.3   rlang_0.4.11
## [31] readxl_1.3.1        rstudioapi_0.13     generics_0.1.0
## [34] dplyr_1.0.7         zip_2.2.0           magrittr_2.0.1
## [37] munsell_0.5.0       fansi_0.5.0         abind_1.4-5
## [40] lifecycle_1.0.0     stringi_1.6.1       yaml_2.2.1
## [43] parallel_4.1.0      forcats_0.5.1       crayon_1.4.1
## [46] lmom_2.8            haven_2.4.1         splines_4.1.0
## [49] pander_0.6.4        hms_1.1.0           tmvnsim_1.0-2
## [52] pillar_1.6.1        boot_1.3-28         ggsignif_0.6.2
## [55] glue_1.4.2          evaluate_0.14       latticeExtra_0.6-29
## [58] data.table_1.14.0   nloptr_1.2.2.2      png_0.1-7
## [61] vctrs_0.3.8         cellranger_1.1.0    gtable_0.3.0
## [64] purrr_0.3.4         tidyr_1.1.3         xfun_0.23
## [67] openxlsx_4.2.4      broom_0.7.8         e1071_1.7-7
## [70] rstatix_0.7.0       class_7.3-19        tibble_3.1.2
## [73] cluster_2.1.2       ellipsis_0.3.2
```