# Midterm

Jingwen GAO

2025-10-08

```
library(ggplot2)
library(survival)
library(tidyverse)
library(dplyr)
library(GGally)
library(gridExtra)
library(grid)
library(car)
library(rstatix)
library(moments)
library(confintr)
library(purrr)
library(ggcorrplot)
library(corrplot)
library(MASS)
library(NHANES)
library(lattice)
library(scales)
library(janitor)
library(knitr)
library(ggmosaic)
library(vcd)
```

STAT276 Midterm (20 pts, 5 pts per question) Do work in .Rmd, organize by question number, knit to .doc or .pdf, and submit by the end of class (11:40 am) on Blackboard.

1. In the birthwt (MASS) data, create a new variable (bwt_lbs) from the bwt variable (conversion factor: 1 gram = .0022 lbs). Then, generate means and SDs for the new variable by mom's smoking during pregnancy status. Tidy functions and one piped operation required.

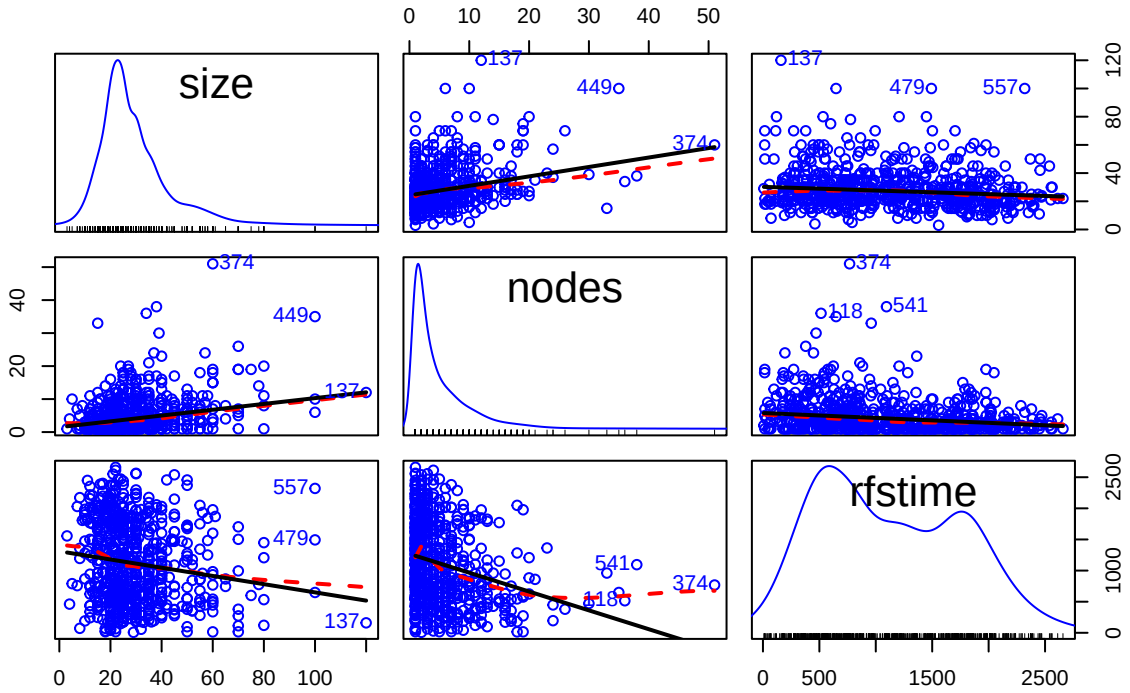
```
birthwt |>
    mutate(bwt_lbs = bwt * 0.0022) |>
    group_by(smoke) |>
    summarise(means = mean(bwt_lbs, na.rm = T), SDs = sd(bwt_lbs, na.rm = T))
```

```
## # A tibble: 2 x 3
##   smoke means   SDs
##   <int> <dbl> <dbl>
## 1     0  6.72  1.66
## 2     1  6.10  1.45
```

2. In the gbsg data (survival), create a scatterplot matrix to explore the relationship between tumor size and number of nodes and survival time (rfstime, the outcome variable). The matrix should have those 3 variables only. Smooth scatterplots with linear and nonlinear smoothers. In a few sentences, explain what you learn about the how tumor size and number of nodes are related (if at all) to survival time.

```
g<-gbsg|>
  relocate(size,.before = 2)|>
  relocate(nodes,.before = 3)|>
  relocate(rfstime,.before = 4)
dat<-g[,2:4]
scatterplotMatrix(dat,


                  regLine=list(method=MASS::rlm,col="black"),
                  smooth=list(spread=F,span=.5,col.smooth="red"),
                  id=list(n=3),
                  main="Tumor size, number of nodes, with survival time"
)
```

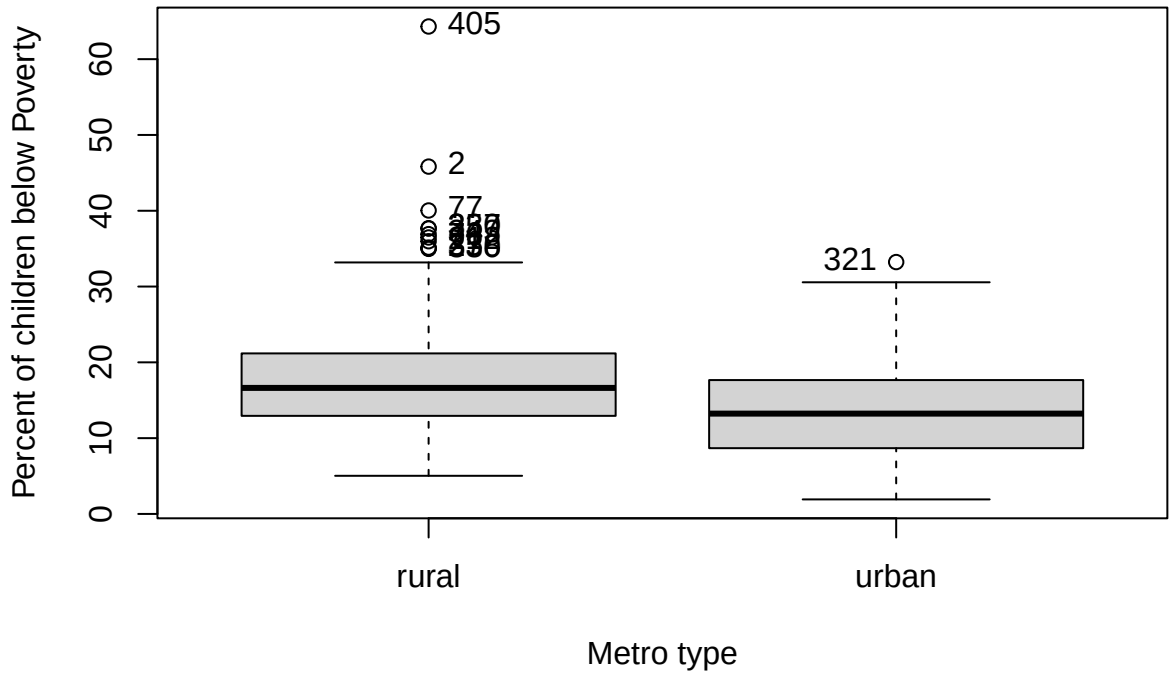



According to the leftest bottom panel, by reading the loess and linear regression, we can tell that as tumor size increases, the corresponding survival time decreases. They have a moderate negative correlation. Their relationship is approximately linear, since the loess regression and linear regression greatly overlap with each other. However, by the scatterplot, we can tell that most tumor sizes cluster at a the interval of [20,40]. In this cluster region, the survival time has a great variance, which might due to other factors that influencing the survival time. Conclusively, there is a negative correlation between tumor size and survival time. But when tumor size is moderately small (within [20,40]), more factors should be taken into account to predict the survival time.

According to the middle bottom panel, the linear regression and loess regression both show a negative correlation between nodes and survival time when node number is less than 30, whereas the loess regression shows that when there are more than 30 nodes, the survival time will increase. Also, by the scatterplot, we can tell that most tumor has less than 10 nodes, and their survival time also has a great variance, meaning that we should consider other factors to predict the survival time when there are few nodes.

3. In the midwest (ggplot2) data, generate boxplots of percent of children below the poverty line (percchildbelowpovert) by metro status (inmetro, 0=rural, 1=urban) for the counties in Illinois. Then find either the case numbers or the values of the outliers shown in the boxplot (if any are shown). What do these plots suggest about the relationship between child poverty in urban vs. rural counties in Illinois?

```
m <- midwest |>
    mutate(inmetro = as.factor(dplyr::recode(inmetro, `0` = "rural", `1` = "urban")))
p <- car::Boxplot(m$percchildbelowpovert ~ m$inmetro, id = list(n = Inf), xlab = "Metro type",
    ylab = "Percent of children below Poverty")
```




```
m |>
    slice(as.integer(p)) |>
    dplyr::select("PID", "percchildbelowpovert")
```

```
## # A tibble: 12 x 2
##      PID percchildbelowpovert
##    <int>                <dbl>
##  1   562                 45.8
```

```
##  2   595                35.1
##  3   636                36.5
##  4   637                40.0
##  5  1214                36.0
##  6  1239                37.7
##  7  2009                35.0
##  8  2061                35.0
##  9  2074                36.9
## 10  2081                37.7
## 11  3020                64.3
## 12  2052                33.2
```

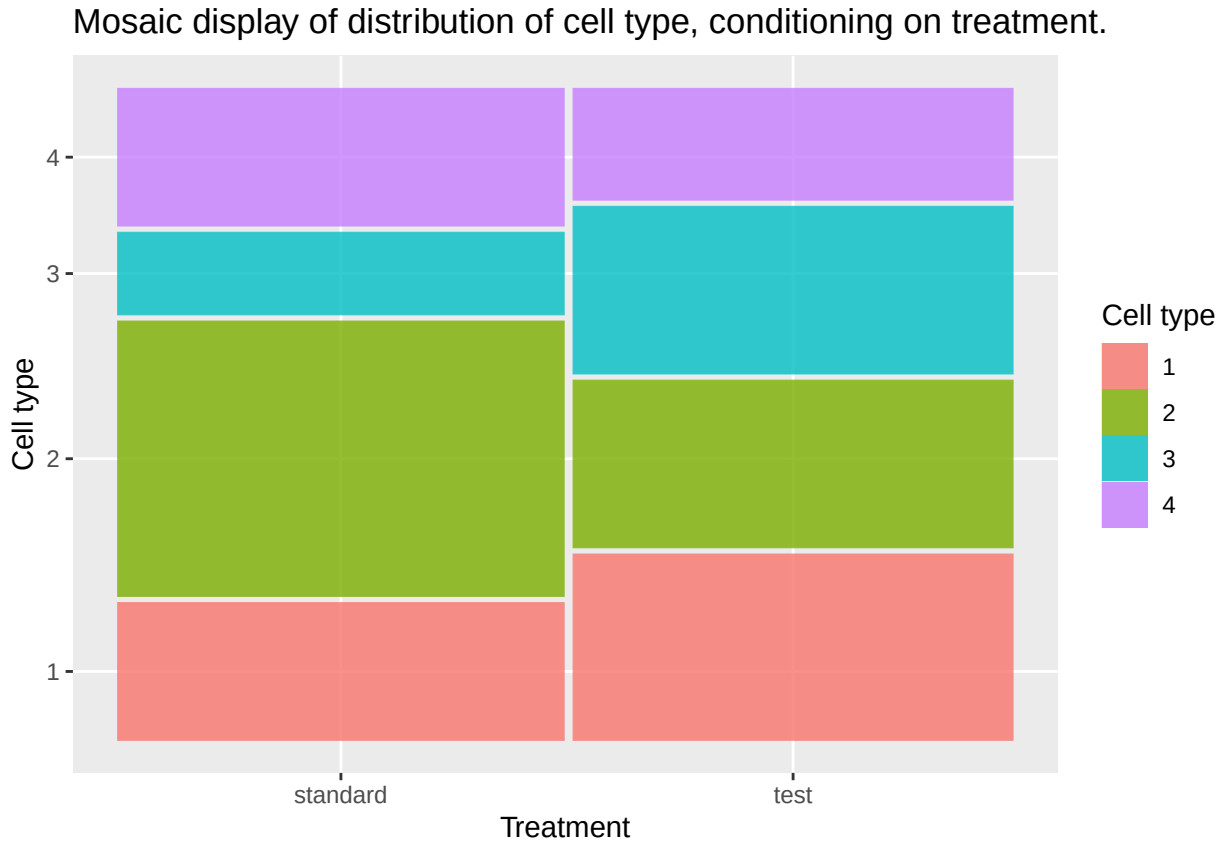
From the boxplot by Metro type, we can see that boxplot for rural counties shows a higher median, more high-value outliers and a relatively more clustered distribution. We can tell that the percent of children under poverty in rural counties has a more right skewed distribution than that in urban counties, meanning that rural kids tend to have a more severe poverty in general. Also, rural counties are more likely to have children with extreme serious poverty. The distribution of percentage for rural counties clustered at a higher level than urban counties, meaning that the poverty level of most rural counties is a bit higher than most urban counties.

4. In the VA (MASS) data, generate a mosaic plot of the relationship between treatment (1=standard, 2=test) and cancer cell type. Make treatment the conditioning variable. Does the plot suggest that lung cancer treatment might be confounded by cell type? To further examine this, generate a table of proportions of cell type by treatment condition. Summarize in a couple sentences.

```
V<-VA|>mutate(treat=dplyr::recode(treat,"1"="standard","2"="test"))
ggplot(data=V|>drop_na(cell)|>drop_na(treat)) +
  geom_mosaic(aes(x=product(cell, treat), fill=cell)) +


  labs(title = "Mosaic display of distribution of cell type, conditioning on treatment.",


       fill="Cell type",
       x="Treatment",
       y="Cell type")
```




From the Mosaic plot, we can tell that the proportions of cell type are different across different treatment. Therefore, we think there is an association between treatment and cell type.

```
library(kableExtra)
tab <- V %>%
    tabyl(treat, cell) %>%
    adorn_totals(c("row", "col")) %>%
    adorn_percentages("row") %>%  adorn_percentages("row") %>%

adorn_pct_formatting(rounding = "half up", digits = 0) %>%
    adorn_ns() %>%
    adorn_title("combined")

tab %>%
    kbl()
```

| treat/cell | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| standard | 22% (15) | 43% (30) | 13% (9) | 22% (15) | 100% (69) |
| test | 29% (20) | 26% (18) | 26% (18) | 18% (12) | 100% (68) |
| Total | 26% (35) | 35% (48) | 20% (27) | 20% (27) | 100% (137) |

From the tab we have shown here, we can tell that - Standard treatment is used more for type 2 cell, least for type 3 cell, approximately the same for type 1 and type 4 cell. - The test treatment is mostly used for type 1 cell, least for type 4 cell, approximately the same for type 2 and type 3 cell.