a1751699_A3

Zhao Ming Soh

19/05/2022

a) Read the data in

```
lung_cancer <- read_csv("lung_cancer.csv")
```

```
## Rows: 24 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): city, age
## dbl (2): pop, cases
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
lung_cancer
```

```
## # A tibble: 24 x 4
##    city       age     pop cases
##    <chr>      <chr> <dbl> <dbl>
##  1 Fredericia 40-54  3059    11
##  2 Horsens    40-54  2879    13
##  3 Kolding    40-54  3142     4
##  4 Vejle      40-54  2520     5
##  5 Fredericia 55-59   800    11
##  6 Horsens    55-59  1083     6
##  7 Kolding    55-59  1050     8
##  8 Vejle      55-59   878     7
##  9 Fredericia 60-64   710    11
## 10 Horsens    60-64   923    15
## # ... with 14 more rows
```

b) Perform EDA

```
skimr::skim(lung_cancer)
```

Table 1: Data summary

Name	lung_cancer
Number of rows	24
Number of columns	4
Column type frequency:	
character	2
numeric	2
Group variables	None

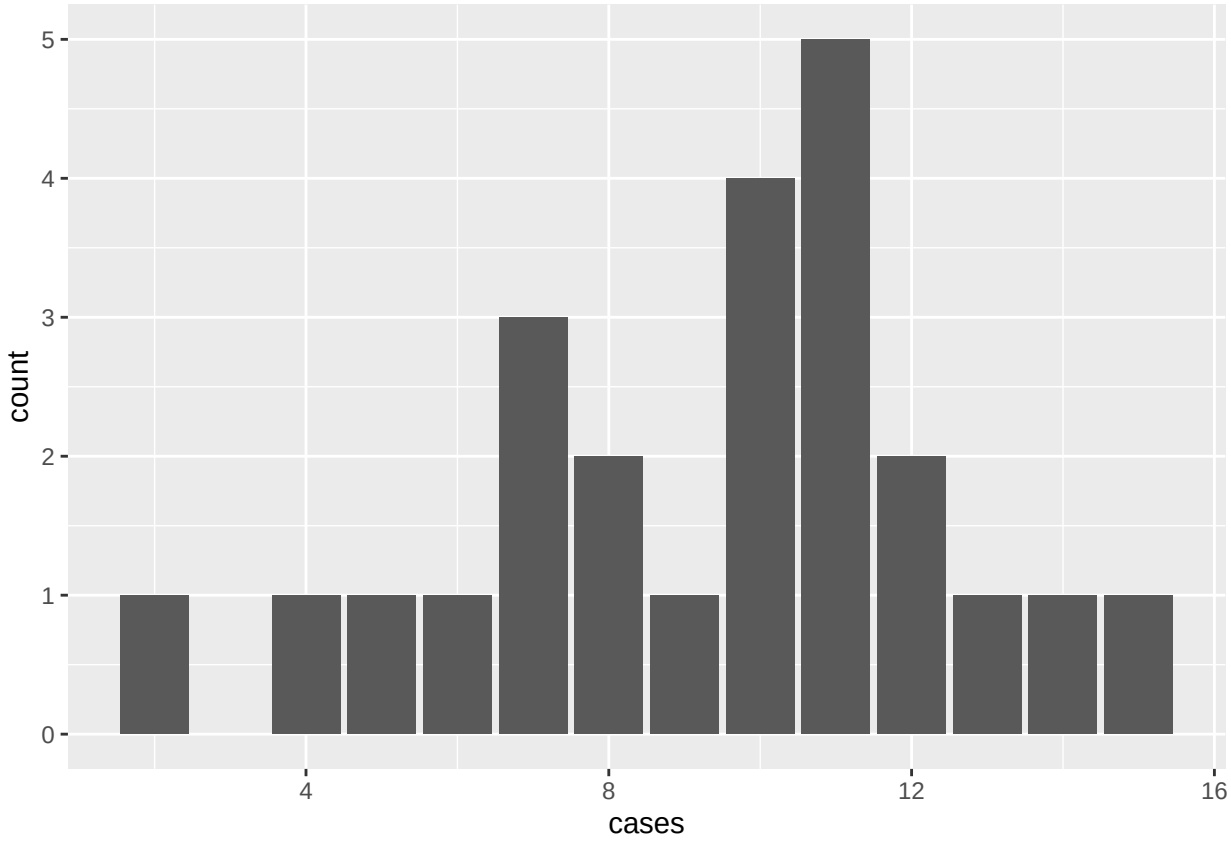
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
city	0	1	5	10	0	4	0
age	0	1	3	5	0	6	0

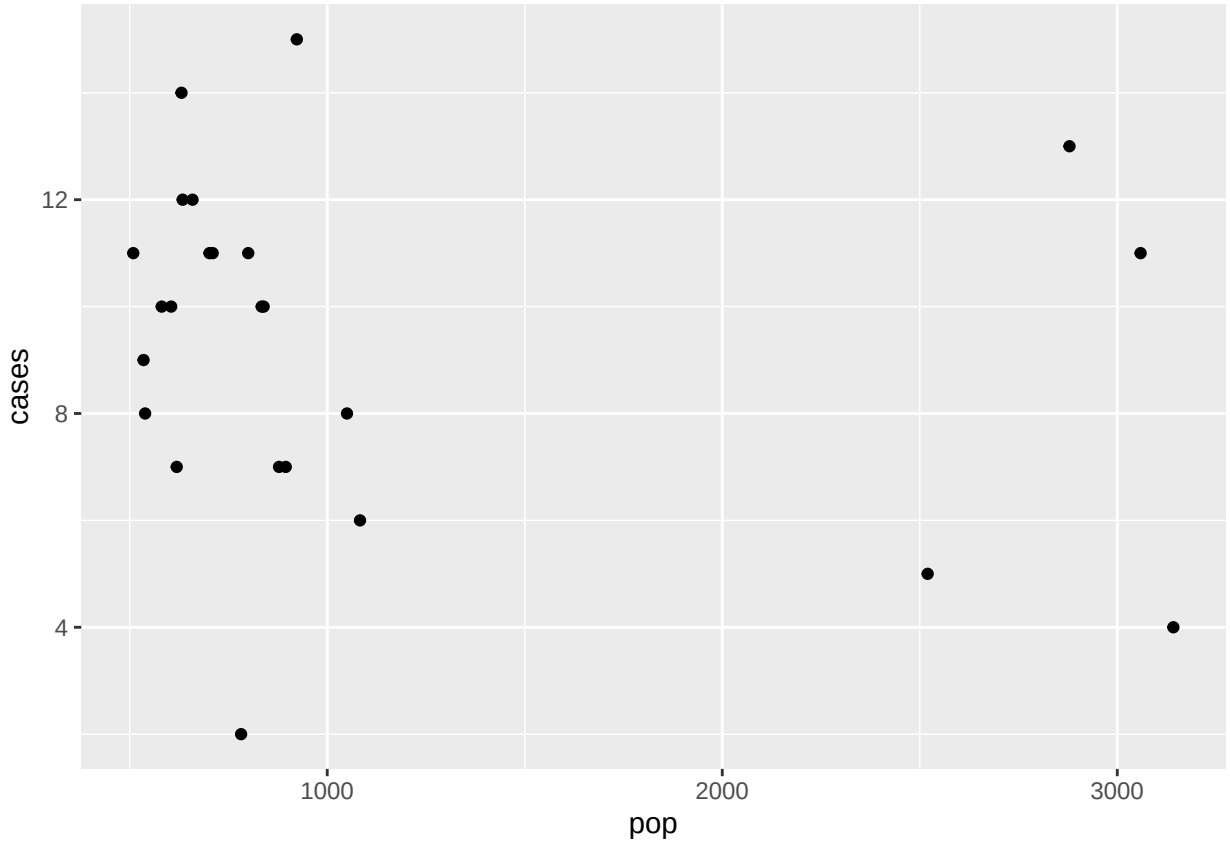
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pop	0	1	1100.33	842.23	509	628	791	954.75	3142	
cases	0	1	9.33	3.16	2	7	10	11.00	15	

```
lung_cancer %>%
  ggplot(aes(cases)) + geom_bar()
```

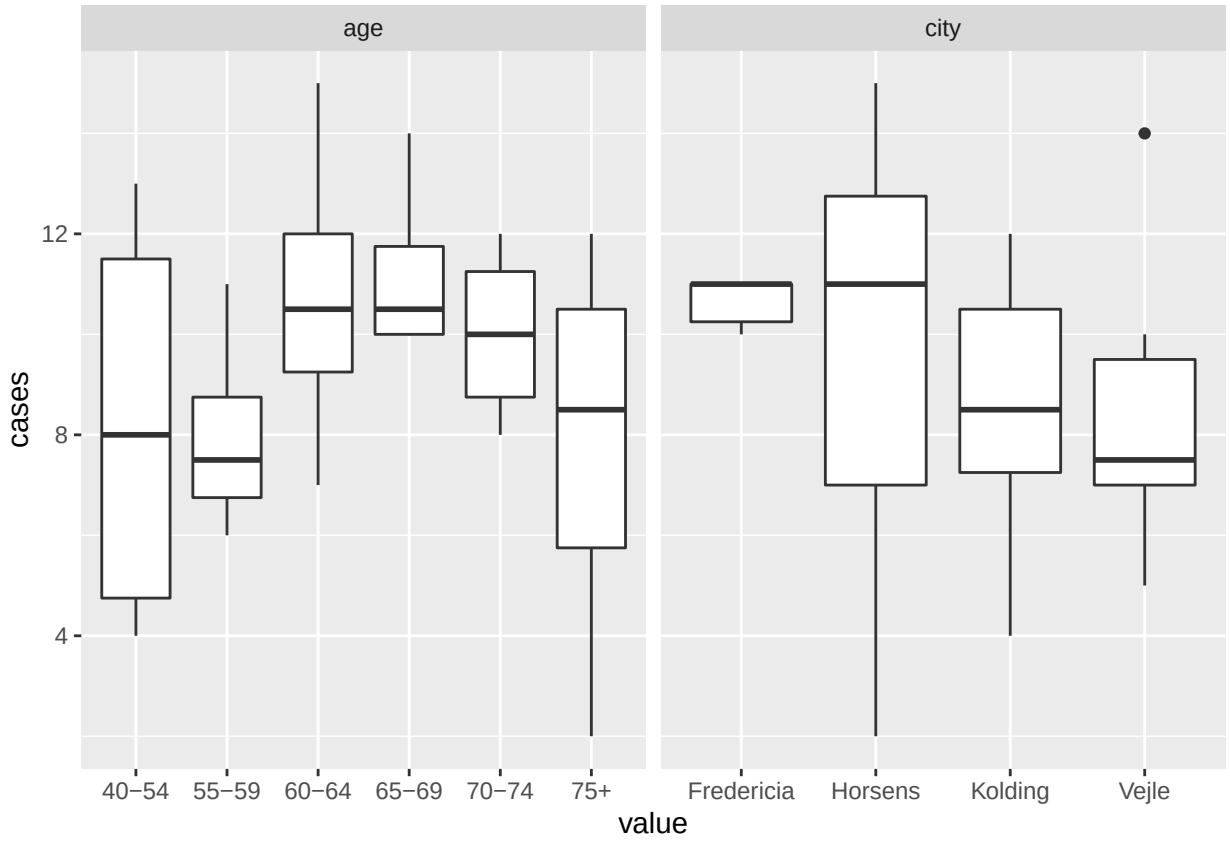



```
lung_cancer %>%
  ggplot(aes(pop, cases)) + geom_point()
```

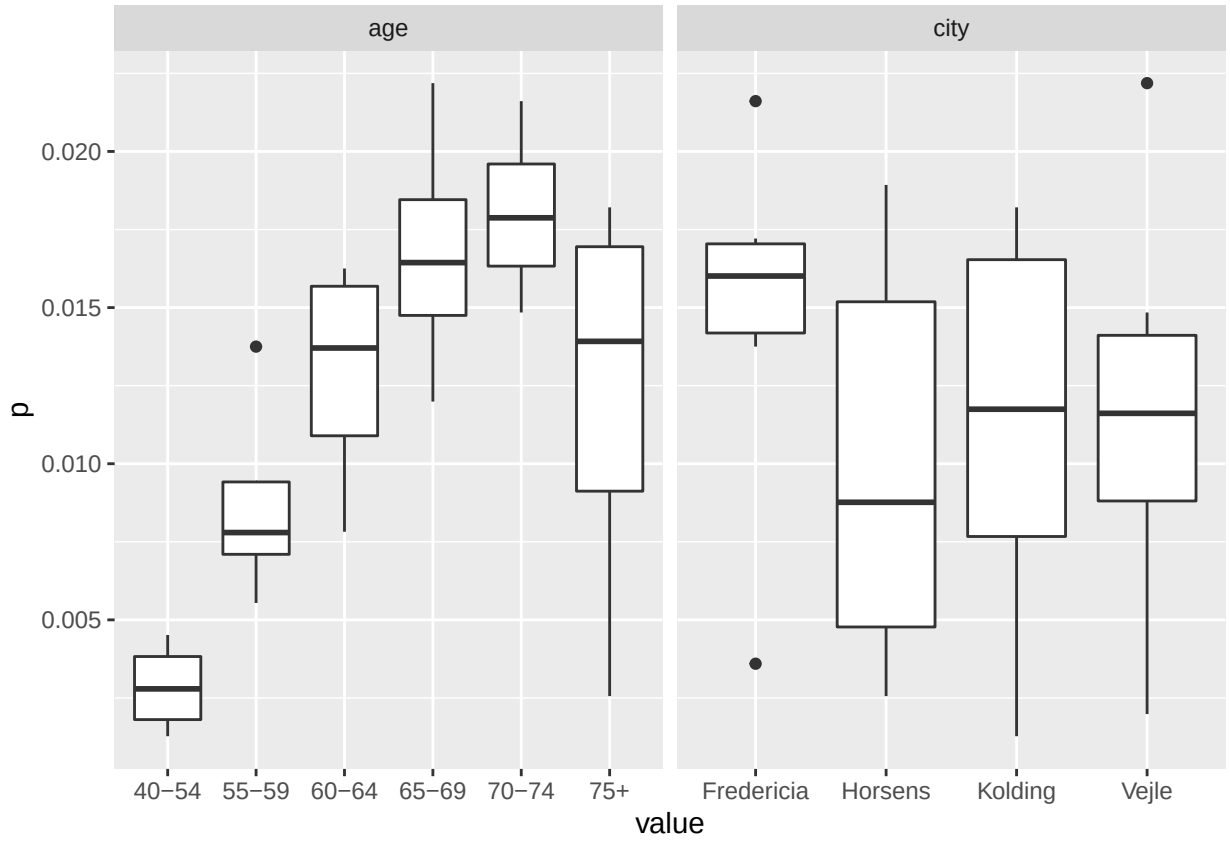



```
lung_cancer <- lung_cancer %>%
  mutate(
    across(where(is.character), factor)
    )

# Boxplot that shows the relationship between cases in the age and city groups
lung_cancer %>%
  select(where(is.factor),cases) %>%
  pivot_longer(-cases) %>%
  ggplot(aes(value,cases)) +
  geom_boxplot() +
  facet_wrap(~name, scales = "free_x")
```




```
# Boxplot that shows the relationship between proportion of cases to populations in the age and city gr
lung_cancer %>%
  mutate(
    p = cases/pop
  ) %>%
  select(p,where(is.factor)) %>%
  pivot_longer(-p) %>%
  ggplot(aes(value,p)) +
  geom_boxplot() +
  facet_wrap(~name, scales = "free_x")
```



c) Fit a Poisson Rate Regression (M1)

```
M1 <- glm(cases ~ 1, family = poisson, offset = log(pop), data = lung_cancer)
summary(M1)
```

```
##
## Call:
## glm(formula = cases ~ 1, family = poisson, data = lung_cancer,
##     offset = log(pop))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.4891  -0.5126   1.2413   1.9248   3.1028
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.76978    0.06682  -71.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 129.91  on 23  degrees of freedom
```

```
## Residual deviance: 129.91  on 23  degrees of freedom
## AIC: 228.3
##
## Number of Fisher Scoring iterations: 5
```

d) Fit a Poisson Rate Regression (M2)

```
M2 <- glm(cases ~ age + city, family = poisson, offset = log(pop), data = lung_cancer)
summary(M2)
```

```
##
## Call:
## glm(formula = cases ~ age + city, family = poisson, data = lung_cancer,
##     offset = log(pop))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.63573  -0.67296  -0.03436   0.37258   1.85267
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.6321     0.2003 -28.125  < 2e-16 ***
## age55-59      1.1010     0.2483   4.434 9.23e-06 ***
## age60-64      1.5186     0.2316   6.556 5.53e-11 ***
## age65-69      1.7677     0.2294   7.704 1.31e-14 ***
## age70-74      1.8569     0.2353   7.891 3.00e-15 ***
## age75+        1.4197     0.2503   5.672 1.41e-08 ***
## cityHorsens  -0.3301     0.1815  -1.818   0.0690 .
## cityKolding  -0.3715     0.1878  -1.978   0.0479 *
## cityVejle    -0.2723     0.1879  -1.450   0.1472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 129.908  on 23  degrees of freedom
## Residual deviance:  23.447  on 15  degrees of freedom
## AIC: 137.84
##
## Number of Fisher Scoring iterations: 5
```

e) Fit a Poisson Rate Regression (M3)

```
M3 <- glm(cases ~ age + city + log(pop), family = poisson, offset = log(pop), data = lung_cancer)
summary(M3)
```

```
##
## Call:
## glm(formula = cases ~ age + city + log(pop), family = poisson,
```

```
##     data = lung_cancer, offset = log(pop))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.44001  -0.64195  -0.04286   0.50052   1.51893
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.7496     8.8151   1.333   0.1826
## age55-59     -1.3842     1.2729  -1.087   0.2768
## age60-64     -1.2367     1.4049  -0.880   0.3787
## age65-69     -1.4378     1.6310  -0.882   0.3780
## age70-74     -1.8049     1.8608  -0.970   0.3321
## age75+       -1.8383     1.6588  -1.108   0.2678
## cityHorsens   0.1833     0.3193   0.574   0.5660
## cityKolding  -0.0483     0.2520  -0.192   0.8480
## cityVejle    -0.1679     0.1965  -0.855   0.3927
## log(pop)     -2.2096     1.1227  -1.968   0.0491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 129.908  on 23  degrees of freedom
## Residual deviance:  19.498  on 14  degrees of freedom
## AIC: 135.89
##
## Number of Fisher Scoring iterations: 4
```

f) Use Anova to compare M1 and M2

```
anova(M1,M2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cases ~ 1
## Model 2: cases ~ age + city
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        23    129.908
## 2        15     23.447  8   106.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(M2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: cases
##
```

```
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    23    129.908
## age   5  101.601        18     28.307   <2e-16 ***
## city  3    4.859        15     23.447   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant difference between M1 and M2 as shown by the anova test where M2 has a p-value $<$ 0.05. Therefore, both city and age are significant predictors in predicting the number of cases of lung cancer.

g) Find AIC of all 3 models

```
AIC(M1,M2,M3)
```

```
##    df      AIC
## M1  1 228.2960
## M2  9 137.8355
## M3 10 135.8862
```

```
AIC(M2) - AIC(M3)
```

```
## [1] 1.949273
```

M2 is selected as the best model using the rule of thumb that if the AIC value between any 2 models is within 2 of each other, then always select the smaller model.

h) Summary of coefficients for M2

```
coef_M2_age55_59 <- 1.1010
exp(1.1010)
```

```
## [1] 3.007172
```

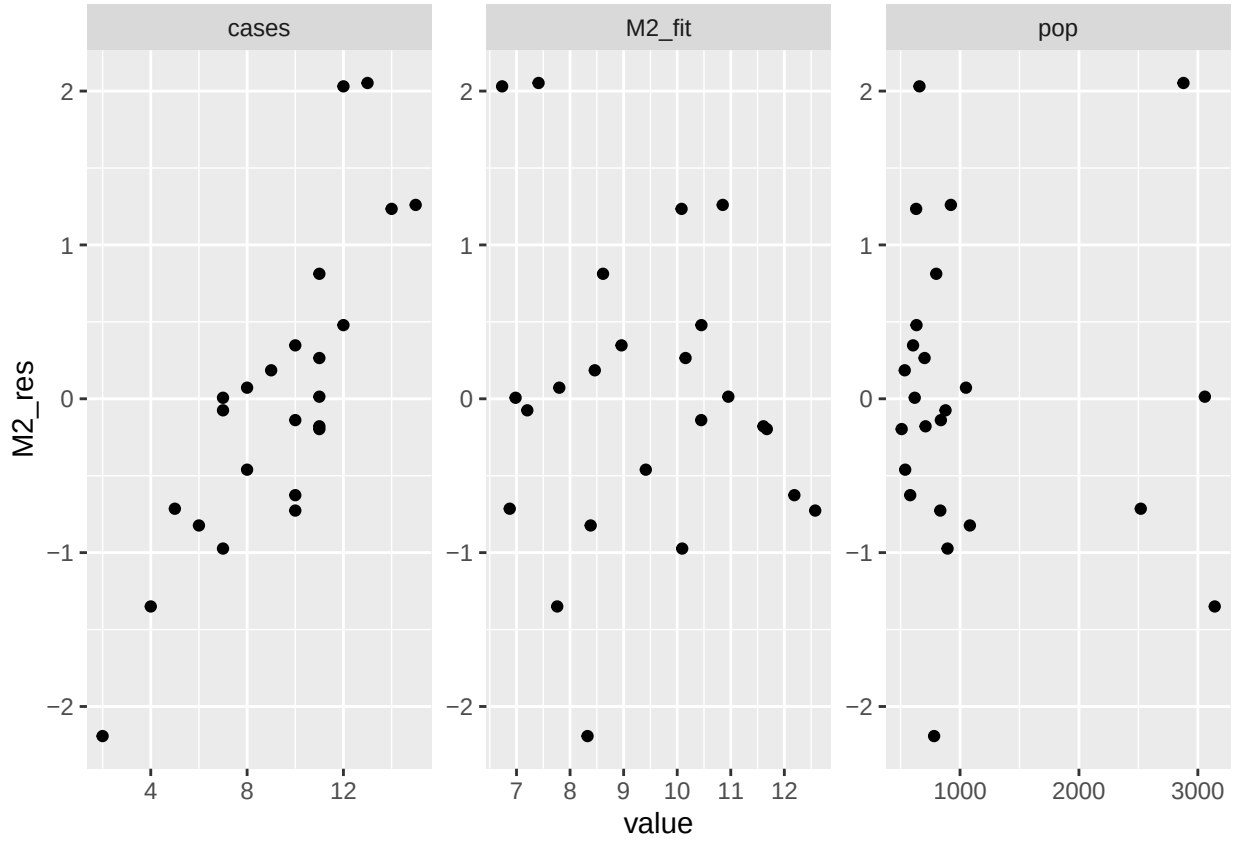
The coefficient of the age group 55-59 is 1.1010 , this is equivalent to an increase of 300% for each increase in the population of the age group 55-59 by 1 person.

i) Obtain the Pearson Residuals for M2

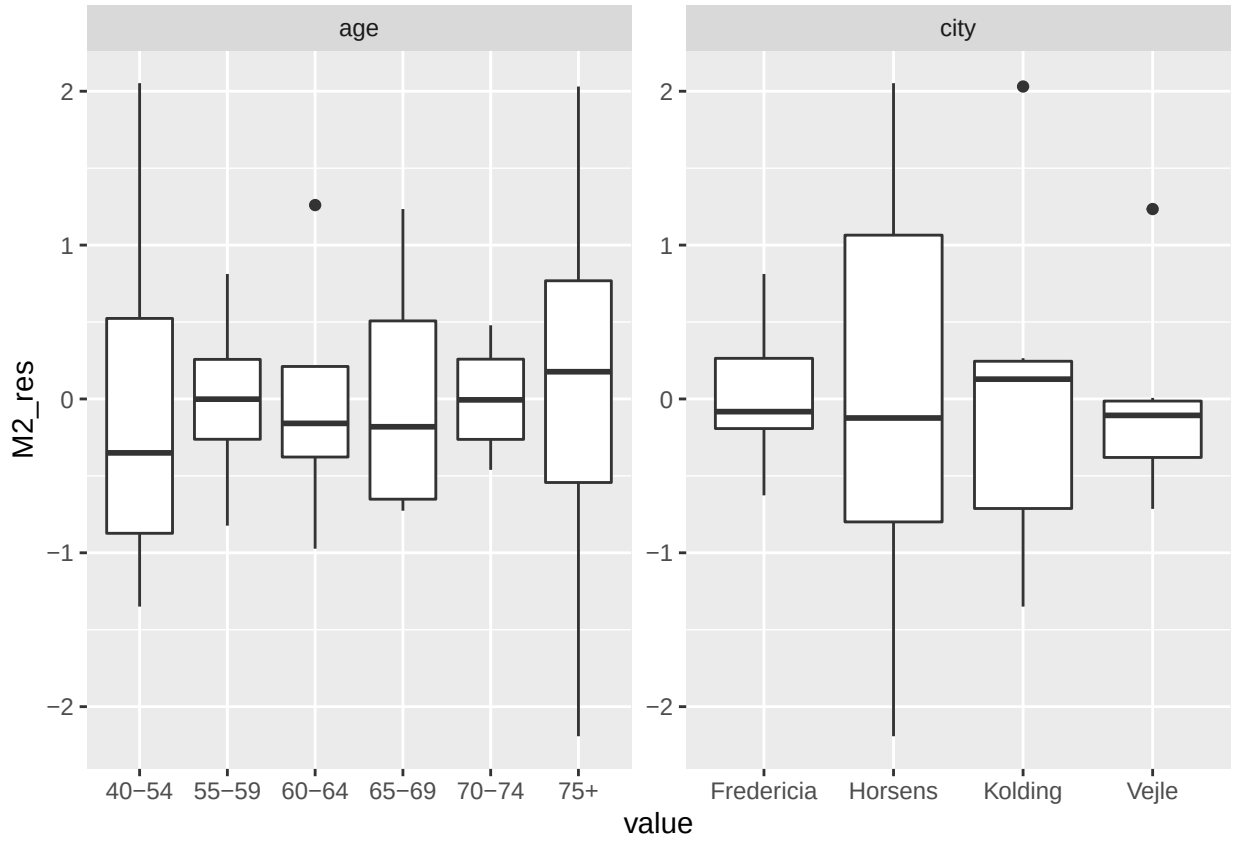
```
lung_cancer <- lung_cancer %>%
  add_column(
    M2_res = residuals(M2, type="pearson"),
    M2_fit = fitted(M2)
  )
```

j) Plot residuals vs

```
lung_cancer %>%
  select(where(is.double)) %>%
  pivot_longer(-M2_res) %>%
  ggplot(aes(value, M2_res)) + geom_point() + facet_wrap(~name, scale = "free")
```




```
lung_cancer %>%
  select(where(is.factor), M2_res) %>%
  pivot_longer(-M2_res) %>%
  ggplot(aes(value, M2_res)) + geom_boxplot() + facet_wrap(~name, scale = "free")
```



k)

```
new_data<-tibble(age = "40-54", pop = 4000, city = "Fredericia")
lambda<-predict(M2,newdata = new_data, type = "response")
ppois(5, lambda)
```

```
## [1] 0.004440372
```

The log additive (multiplicative) model

$$\mu_{ijk} = n_{ijk} e^{\alpha_i + \beta_j + \gamma_k}$$