FinalProject

Eeman, Linda, Sofia

2025-11-06

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
dat <- read_csv("cancer patient data sets 2.csv")
```

```
## Rows: 1000 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient Id, Level
## dbl (24): index, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occu...
##
## 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.
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
dat <- read_csv("cancer patient data sets 2.csv")
```

```
## Rows: 1000 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient Id, Level
## dbl (24): index, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occu...
##
## 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.
```

```r
names(dat) <- make.names(names(dat))

names(dat)
```

```
##  [1] "index"                    "Patient.Id"
##  [3] "Age"                      "Gender"
##  [5] "Air.Pollution"            "Alcohol.use"
##  [7] "Dust.Allergy"             "OccuPational.Hazards"
##  [9] "Genetic.Risk"             "chronic.Lung.Disease"
## [11] "Balanced.Diet"            "Obesity"
## [13] "Smoking"                  "Passive.Smoker"
## [15] "Chest.Pain"               "Coughing.of.Blood"
## [17] "Fatigue"                  "Weight.Loss"
## [19] "Shortness.of.Breath"      "Wheezing"
## [21] "Swallowing.Difficulty"    "Clubbing.of.Finger.Nails"
## [23] "Frequent.Cold"            "Dry.Cough"
## [25] "Snoring"                  "Level"
```

```r
dat$Level <- factor(dat$Level,
                     levels = c("Low","Medium","High"),
                     ordered = TRUE)

dat <- dat |>
  mutate(High = if_else(Level == "High", 1, 0))



m1causes <- glm(High ~ Gender + Genetic.Risk + Age + Smoking + Air.Pollution + Obesity, data = dat, fam

m1symptoms <- glm(High ~  Chest.Pain + Coughing.of.Blood + Shortness.of.Breath, data = dat, family = "b

summary(m1causes)
```

```
##
## Call:
## glm(formula = High ~ Gender + Genetic.Risk + Age + Smoking +
##     Air.Pollution + Obesity, family = "binomial", data = dat)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -16.26976    1.79617  -9.058  < 2e-16 ***
## Gender          1.92822    0.47620   4.049 5.14e-05 ***
## Genetic.Risk   -0.25830    0.14304  -1.806 0.070940 .
## Age            -0.05281    0.01424  -3.708 0.000209 ***
## Smoking         0.79419    0.10285   7.722 1.15e-14 ***
## Air.Pollution   1.22708    0.13725   8.941  < 2e-16 ***
## Obesity         1.59079    0.14919  10.663  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1312.48  on 999  degrees of freedom
## Residual deviance:  236.77  on 993  degrees of freedom
## AIC: 250.77
##
## Number of Fisher Scoring iterations: 8
```

```r
summary(m1symptoms)
```

```
##
## Call:
## glm(formula = High ~ Chest.Pain + Coughing.of.Blood + Shortness.of.Breath,
##     family = "binomial", data = dat)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -9.42731    0.60615 -15.553  < 2e-16 ***
## Chest.Pain           0.12346    0.08084   1.527  0.12673
## Coughing.of.Blood    1.29119    0.11326  11.400  < 2e-16 ***
## Shortness.of.Breath  0.21237    0.06653   3.192  0.00141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1312.48  on 999  degrees of freedom
## Residual deviance:  423.38  on 996  degrees of freedom
## AIC: 431.38
##
## Number of Fisher Scoring iterations: 7
```

```r
exp(7.6069)
```

```
## [1] 2012.031
```

```r
exp(4.1667)
```

```
## [1] 64.50224
```

```r
exp(6.3718)
```

```
## [1] 585.1101
```

```r
exp(-0.7692)
```

```
## [1] 0.4633836
```

```r
exp( 10.4882)
```

```
## [1] 35889.5
```

```r
library(gt)


df <- data.frame(
  Estimate = c(-16.270, 1.928, -0.258,-0.052,.794,1.227,1.590),
  "Standard Error" = c(1.796, 0.476, 0.142,0.014,0.102,.137,.149),
  pValue = c(2.00e-16,5.14e-5,0.071,2.00e-4,1.15e-14,2.00e-16,2.00e-16),
  row.names = c("Intercept","Gender", "Genetic Risk", "Age","Smoking","Air Pollution", "Obesity")
)
df %>%
  gt(rownames_to_stub = TRUE) %>%
```

Logression Coefficients For Cancer Markers

	Estimate	Standard.Error	pValue
Intercept	-16.270	1.796	2.00e-16
Gender	1.928	0.476	5.14e-05
Genetic Risk	-0.258	0.142	7.10e-02
Age	-0.052	0.014	2.00e-04
Smoking	0.794	0.102	1.15e-14
Air Pollution	1.227	0.137	2.00e-16
Obesity	1.590	0.149	2.00e-16

Logression Coefficients For Cancer Symptoms

	Estimate	Standard.Error	pValue
Intercept	-9.427	0.6060	2.00e-16
Chest Pain	0.123	0.0810	1.27e-01
Coughing of Blood	1.291	0.1130	2.00e-16
Shortness of Breath	0.212	0.0666	1.00e-03

```
  tab_header(
    title = "Logression Coefficients For Cancer Markers"
  )
```

```
dg <- data.frame(
  Estimate = c(-9.427,0.123,1.291,0.212),
  "Standard Error" = c(0.606,0.081,0.113,0.0666),
  pValue = c(2e-16,0.127,2e-16,0.001),
  row.names = c("Intercept","Chest Pain","Coughing of Blood","Shortness of Breath")
)
dg %>%
  gt(rownames_to_stub = TRUE) %>%
  tab_header(
    title = "Logression Coefficients For Cancer Symptoms"
  )
```

Including Plots

You can also embed plots, for example:

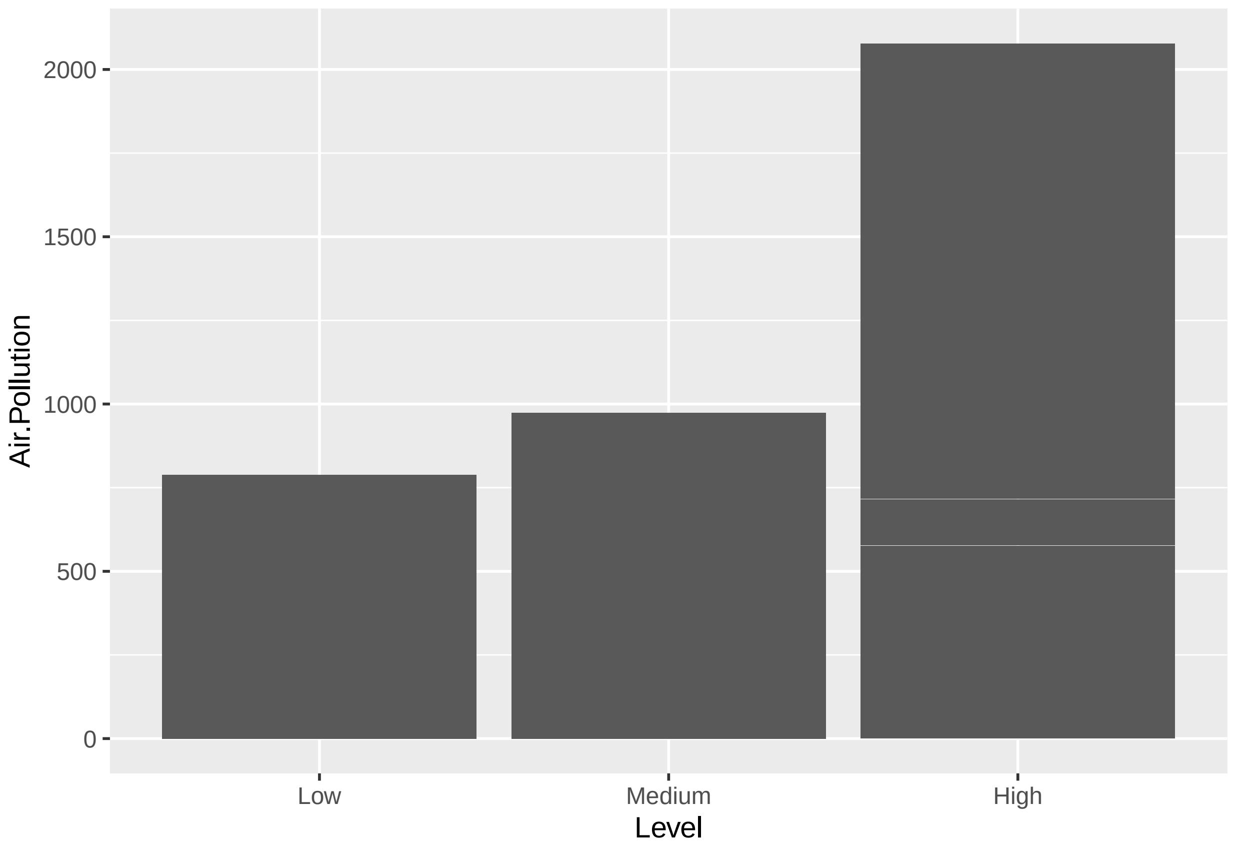
```
library(tidyverse)
data <- read_csv("cancer patient data sets 2.csv")
```

```
## Rows: 1000 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient Id, Level
## dbl (24): index, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occu...
##
## 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.
```

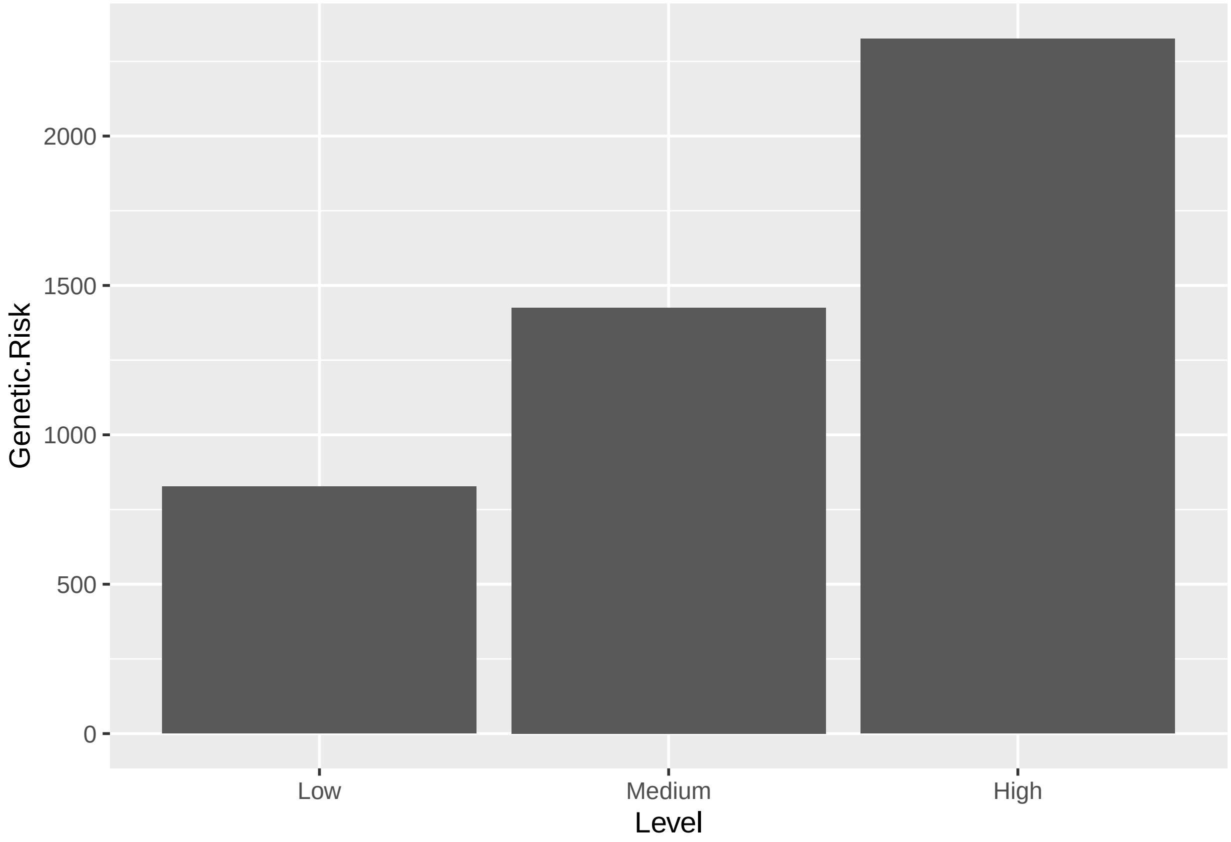
```
m2 <- glm(High ~ Gender + Genetic.Risk + Age + Smoking + Air.Pollution + Obesity, data = dat, family =

summary(m2)
```

```
##
## Call:
## glm(formula = High ~ Gender + Genetic.Risk + Age + Smoking +
##     Air.Pollution + Obesity, family = "binomial", data = dat)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -16.26976    1.79617  -9.058  < 2e-16 ***
## Gender          1.92822    0.47620   4.049 5.14e-05 ***
## Genetic.Risk   -0.25830    0.14304  -1.806 0.070940 .
## Age            -0.05281    0.01424  -3.708 0.000209 ***
## Smoking         0.79419    0.10285   7.722 1.15e-14 ***
## Air.Pollution   1.22708    0.13725   8.941  < 2e-16 ***
## Obesity         1.59079    0.14919  10.663  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1312.48  on 999  degrees of freedom
## Residual deviance:  236.77  on 993  degrees of freedom
## AIC: 250.77
##
## Number of Fisher Scoring iterations: 8
```

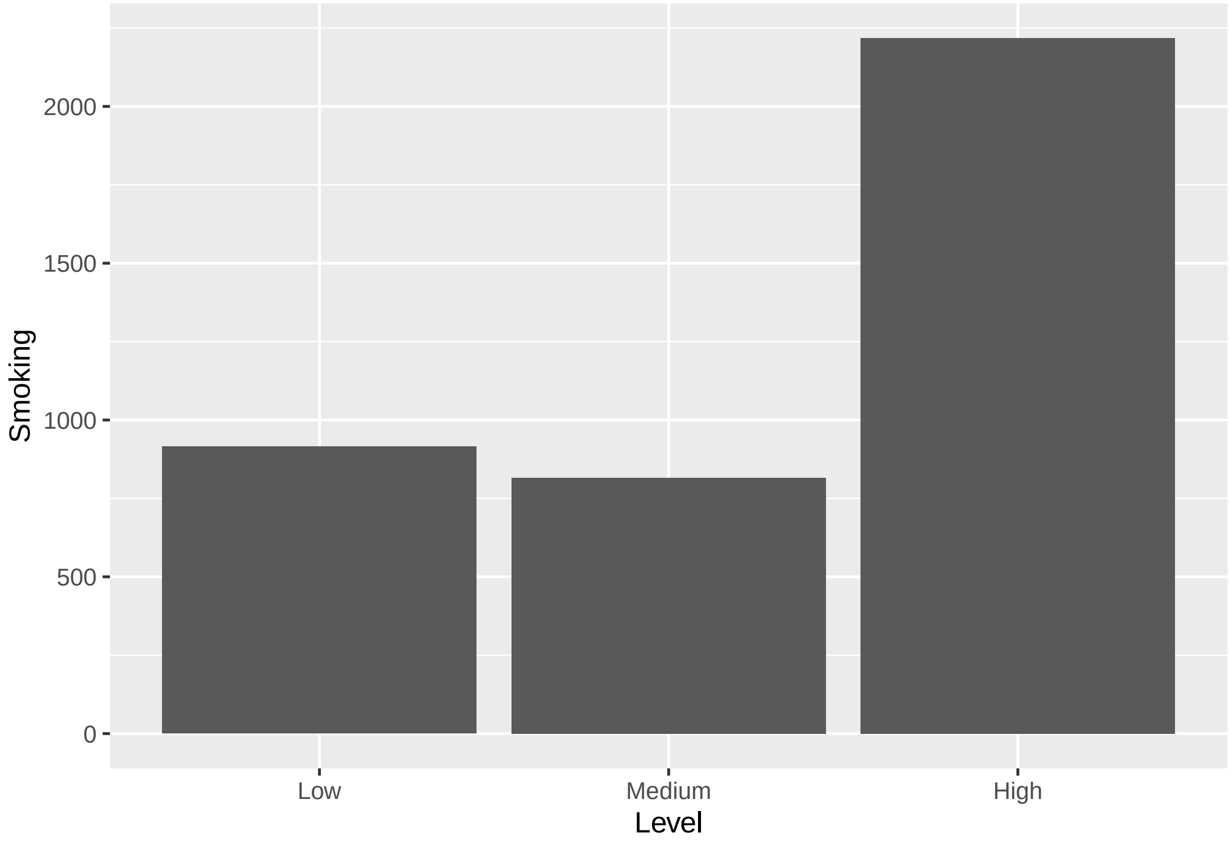
```
library(ggplot2)
ggplot(dat, aes(x = Level, y = `Air.Pollution`)) +
  geom_bar(stat = "identity")
```



```r
ggplot(dat, aes(x = Level, y = `Genetic.Risk`)) +
  geom_bar(stat = "identity")
```



```r
ggplot(dat, aes(x = Level, y = `Smoking`)) +
  geom_bar(stat = "identity")
```



```
table(dat$"Genetic.Risk", dat$Level)
```

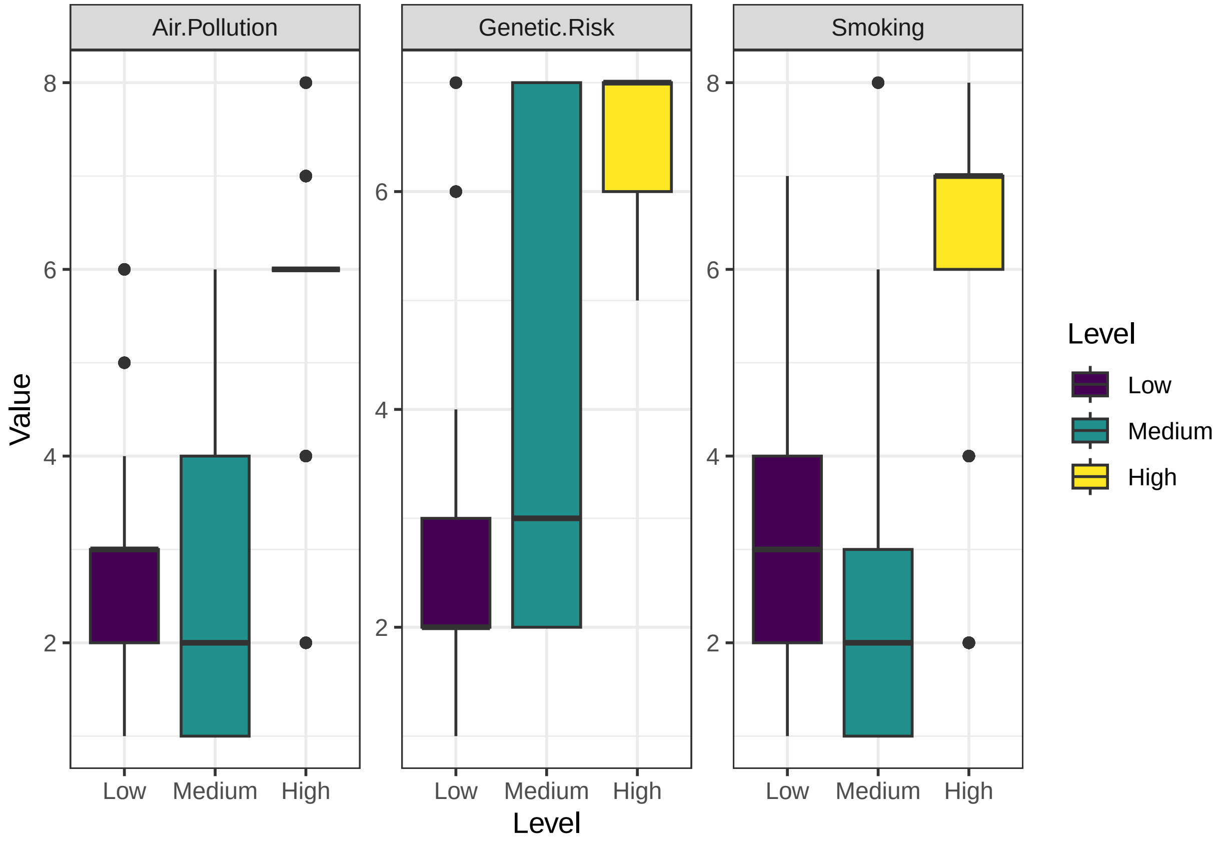
```
## 
##      Low Medium High
##   1   40      0    0
##   2  121     91    0
##   3   92     81    0
##   4   20     20    0
##   5    0     20   80
##   6   20     20   68
##   7   10    100  217
```

```
chisq.test(table(dat$"Genetic.Risk", dat$Level))
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table(dat$Genetic.Risk, dat$Level)
## X-squared = 632.14, df = 12, p-value < 2.2e-16
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
library(tidyverse)

dat_long <- dat|>
  pivot_longer(
    cols = c(`Air.Pollution`, `Genetic.Risk`, Smoking),
    names_to = "Variable",
    values_to = "Value")


ggplot(dat_long, aes(x = Level, y = Value, fill = Level)) +
  geom_boxplot() +
  facet_wrap(~ Variable, scales = "free_y") +
  theme_bw()
```



```
library(tidyverse)

dat <- read_csv("cancer patient data sets 2.csv")
```

```
## Rows: 1000 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient Id, Level
## dbl (24): index, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occu...
##
## 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.
```

```
library(tidyverse)
library(MASS)

data <- read_csv("cancer patient data sets 2.csv")
```

```
## Rows: 1000 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient Id, Level
## dbl (24): index, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occu...
##
## 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.
```

```r
names(data) <- make.names(names(data))

names(data)
```

```
##  [1] "index"                    "Patient.Id"
##  [3] "Age"                      "Gender"
##  [5] "Air.Pollution"            "Alcohol.use"
##  [7] "Dust.Allergy"             "OccuPational.Hazards"
##  [9] "Genetic.Risk"             "chronic.Lung.Disease"
## [11] "Balanced.Diet"            "Obesity"
## [13] "Smoking"                  "Passive.Smoker"
## [15] "Chest.Pain"               "Coughing.of.Blood"
## [17] "Fatigue"                  "Weight.Loss"
## [19] "Shortness.of.Breath"      "Wheezing"
## [21] "Swallowing.Difficulty"    "Clubbing.of.Finger.Nails"
## [23] "Frequent.Cold"            "Dry.Cough"
## [25] "Snoring"                  "Level"
```

```r
data$Level <- factor(data$Level,
                     levels = c("Low","Medium","High"),
                     ordered = TRUE)

m1 <- lm(Level ~ Age + Gender + Air.Pollution + Alcohol.use + Dust.Allergy + OccuPational.Hazards + Gene
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a factor
## response will be ignored
```

```
## Warning in Ops.ordered(y, z$residuals): '-' is not meaningful for ordered
## factors
```

```r
print(m1)
```

```
##
## Call:
## lm(formula = Level ~ Age + Gender + Air.Pollution + Alcohol.use +
##     Dust.Allergy + OccuPational.Hazards + Genetic.Risk + chronic.Lung.Disease +
##     Balanced.Diet + Obesity + Smoking + Passive.Smoker + Chest.Pain +
##     Coughing.of.Blood + Fatigue + Weight.Loss + Shortness.of.Breath +
##     Wheezing + Swallowing.Difficulty + Clubbing.of.Finger.Nails +
##     Frequent.Cold + Dry.Cough + Snoring, data = data)
##
## Coefficients:
##              (Intercept)                       Age                    Gender
##                -0.732411                  0.001459                  0.096549
##            Air.Pollution               Alcohol.use              Dust.Allergy
##                 0.056999                 -0.012926                 -0.004656
##     OccuPational.Hazards              Genetic.Risk      chronic.Lung.Disease
##                -0.053501                  0.169682                  0.009071
##            Balanced.Diet                   Obesity                   Smoking
##                 0.027874                  0.050272                 -0.013888
##           Passive.Smoker                Chest.Pain         Coughing.of.Blood
##                 0.026465                 -0.085737                  0.121869
##                  Fatigue               Weight.Loss       Shortness.of.Breath
##                 0.064261                 -0.019988                  0.053201
##                 Wheezing     Swallowing.Difficulty  Clubbing.of.Finger.Nails
##                 0.001651                  0.076921                  0.048068
##            Frequent.Cold                 Dry.Cough                   Snoring
##                -0.008027                  0.014008                  0.132843
```