STATS 3001 / STATS 4104 / STATS 7054
Statistical Modelling III
Practical 4 - Logistic regression

Week 7

GOAL

The purpose of this practical is to explore the application of the `glm` function to logistic regression.

OVERVIEW

- To enter and prepare data for logistic regression analysis in R.
- To perform a logistic regression in R and interpret the output.
- To extract estimated probabilities from a logistic regression.

DATA

The incidence of non-melanoma skin cancer among women in Minneapolis-St Paul, Minnesota, and Dallas-Fort Worth, Texas was recorded in a study.

The data are available in the file `skin.xlsx`.

STEPS

1. Load data into R.

```
skin <- readxl::read_excel(here::here("data", "skin.xlsx"))
skin
```

```
## # A tibble: 15 x 4
##    Cases Town                Age   Population
##    <dbl> <chr>               <chr>      <dbl>
##  1     1 St  Paul            15-24     172675
##  2    16 St  Paul            25-34     123065
##  3    30 St  Paul            35-44      96216
##  4    71 St  Paul            45-54      92051
##  5   102 St  Paul            55-64      72159
##  6   130 St  Paul            65-74      54722
##  7   133 St  Paul            75-84      32185
##  8    40 St  Paul            85+         8328
##  9     4 Dallas - Fort Worth 15-24     181343
## 10    38 Dallas - Fort Worth 25-34     146207
## 11   119 Dallas - Fort Worth 35-44     121374
## 12   221 Dallas - Fort Worth 45-54     111353
## 13   259 Dallas - Fort Worth 55-64      83004
## 14   310 Dallas - Fort Worth 65-74      55932
## 15    65 Dallas - Fort Worth 85+         7583
```

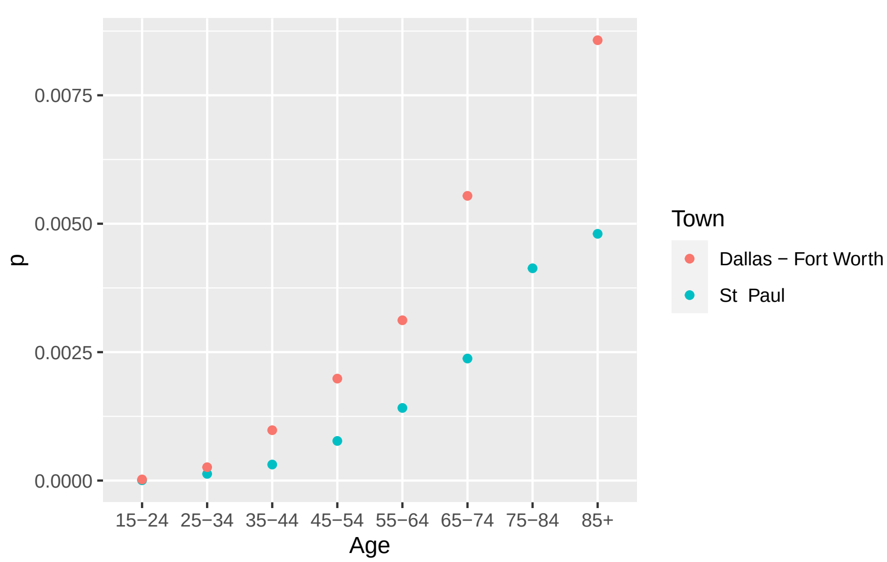
2. Add column that is the proportion of cases for each row.

```
skin <-
  skin %>%
  mutate(
    p = Cases / Population
  )
skin
```

```
## # A tibble: 15 x 5
##    Cases Town                Age   Population          p
##    <dbl> <chr>               <chr>      <dbl>      <dbl>
##  1     1 St  Paul            15-24     172675 0.00000579
##  2    16 St  Paul            25-34     123065 0.000130
##  3    30 St  Paul            35-44      96216 0.000312
##  4    71 St  Paul            45-54      92051 0.000771
##  5   102 St  Paul            55-64      72159 0.00141
##  6   130 St  Paul            65-74      54722 0.00238
##  7   133 St  Paul            75-84      32185 0.00413
##  8    40 St  Paul            85+         8328 0.00480
##  9     4 Dallas - Fort Worth 15-24     181343 0.0000221
## 10    38 Dallas - Fort Worth 25-34     146207 0.000260
## 11   119 Dallas - Fort Worth 35-44     121374 0.000980
## 12   221 Dallas - Fort Worth 45-54     111353 0.00198
## 13   259 Dallas - Fort Worth 55-64      83004 0.00312
## 14   310 Dallas - Fort Worth 65-74      55932 0.00554
## 15    65 Dallas - Fort Worth 85+         7583 0.00857
```

4. Plot proportion against age with colour of plots for each town. Describe the relationships, do the relationships make sense?

```
skin %>%
  ggplot(aes(Age, p, col = Town)) +
  geom_point()
```



5. Fit the following logistic regression models

$$M_0 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{constant}$$

$$M_1 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{Town}_i$$

$$M_2 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{Age}_i$$

$$M_3 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{Age}_i + \text{Town}_i$$

```r
M0 <- glm(
  cbind(Cases, Population - Cases) ~ 1,
  data = skin,
  family = binomial
  )
M1 <- glm(
  cbind(Cases, Population - Cases) ~ Town,
  data = skin,
  family = binomial(logit)
  )
M2 <- glm(
  cbind(Cases, Population - Cases) ~ Age,
  data = skin,
  family = binomial
  )
M3 <- glm(
  cbind(Cases, Population - Cases) ~ Town + Age,
  data = skin,
  family = binomial
  )
```

6. Choose the best model.

```r
anova(M0, M1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Cases, Population - Cases) ~ 1
## Model 2: cbind(Cases, Population - Cases) ~ Town
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        14     2330.5
## 2        13     2207.3  1   123.21 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(M0, M2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Cases, Population - Cases) ~ 1
## Model 2: cbind(Cases, Population - Cases) ~ Age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        14    2330.46
## 2         7     232.28  7   2098.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(M1, M3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Cases, Population - Cases) ~ Town
## Model 2: cbind(Cases, Population - Cases) ~ Town + Age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        13    2207.26
## 2         6       5.15  7   2202.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(M2, M3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Cases, Population - Cases) ~ Age
## Model 2: cbind(Cases, Population - Cases) ~ Town + Age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         7    232.276
## 2         6      5.151  1   227.12 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AIC(M0, M1, M2, M3)
```

```
##    df       AIC
## M0  1 2419.4151
## M1  2 2298.2083
## M2  8  335.2271
## M3  9  110.1024
```

With both LRT and AIC we see that the full model, $M3$ is the best.

7. For you final model, give an interpretation of the coefficient `TownSt Paul`.

```r
summary(M3)
```

```
##
## Call:
## glm(formula = cbind(Cases, Population - Cases) ~ Town + Age,
##     family = binomial, data = skin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2830  -0.3355   0.0000   0.3927   1.0820
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -10.83872    0.44755 -24.218  < 2e-16 ***
## TownSt  Paul  -0.85492    0.05969 -14.322  < 2e-16 ***
## Age25-34       2.62915    0.46747   5.624 1.86e-08 ***
## Age35-44       3.84627    0.45467   8.459  < 2e-16 ***
## Age45-54       4.59538    0.45104  10.188  < 2e-16 ***
## Age55-64       5.08901    0.45031  11.301  < 2e-16 ***
## Age65-74       5.65031    0.44976  12.563  < 2e-16 ***
## Age75-84       6.20887    0.45756  13.570  < 2e-16 ***
## Age85+         6.18346    0.45783  13.506  < 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: 2330.4637  on 14  degrees of freedom
## Residual deviance:    5.1509  on  6  degrees of freedom
## AIC: 110.1
##
## Number of Fisher Scoring iterations: 4
```

If you move from Dallas to St. Paul, we expect the log odds of skin cancer to decrease by 0.85492.

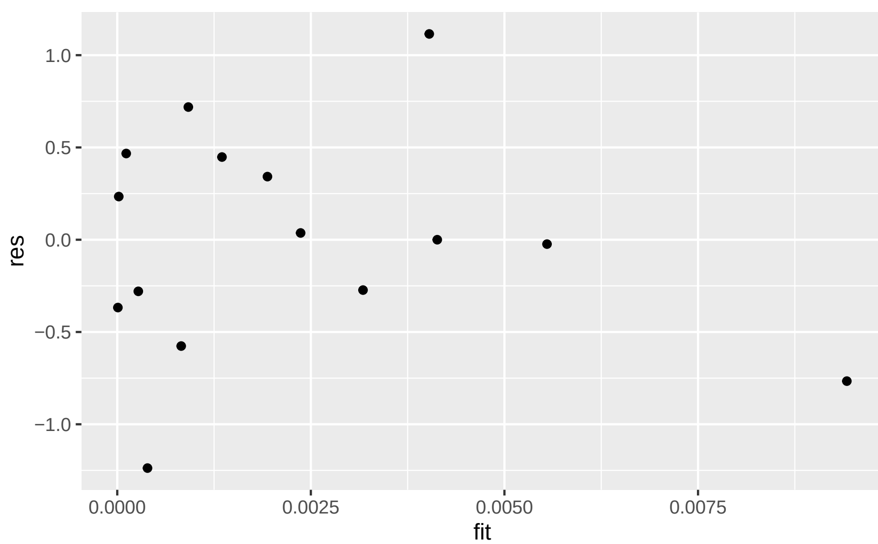
8. Check the assumptions of the model.

```r
skin <-
  skin %>%
  add_column(
    res = residuals(M3, type = "pearson"),
    fit = fitted(M3)
  )
skin
```

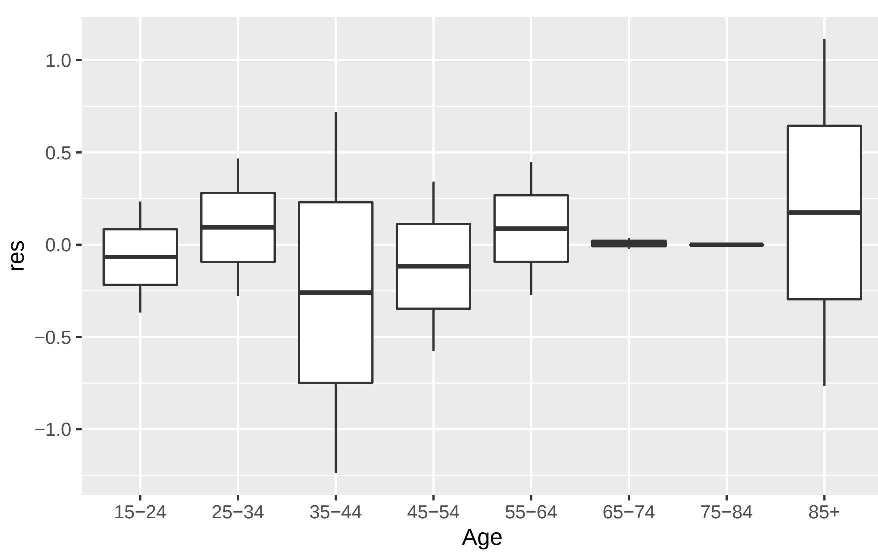
```
## # A tibble: 15 x 7
##    Cases Town                Age   Population          p       res        fit
##    <dbl> <chr>               <chr>      <dbl>      <dbl>     <dbl>      <dbl>
##  1     1 St  Paul            15-24     172675 0.00000579 -3.68e- 1 0.00000835
##  2    16 St  Paul            25-34     123065 0.000130    4.67e- 1 0.000116
##  3    30 St  Paul            35-44      96216 0.000312   -1.24e+ 0 0.000391
##  4    71 St  Paul            45-54      92051 0.000771   -5.76e- 1 0.000826
```

```
## 5   102 St  Paul            55-64      72159 0.00141     4.48e- 1 0.00135
## 6   130 St  Paul            65-74      54722 0.00238     3.67e- 2 0.00237
## 7   133 St  Paul            75-84      32185 0.00413    -4.12e-14 0.00413
## 8    40 St  Paul            85+         8328 0.00480     1.11e+ 0 0.00403
## 9     4 Dallas - Fort Worth 15-24     181343 0.0000221   2.34e- 1 0.0000196
## 10   38 Dallas - Fort Worth 25-34     146207 0.000260   -2.80e- 1 0.000272
## 11  119 Dallas - Fort Worth 35-44     121374 0.000980    7.19e- 1 0.000918
## 12  221 Dallas - Fort Worth 45-54     111353 0.00198     3.42e- 1 0.00194
## 13  259 Dallas - Fort Worth 55-64      83004 0.00312    -2.73e- 1 0.00317
## 14  310 Dallas - Fort Worth 65-74      55932 0.00554    -2.37e- 2 0.00555
## 15   65 Dallas - Fort Worth 85+         7583 0.00857    -7.66e- 1 0.00942
```

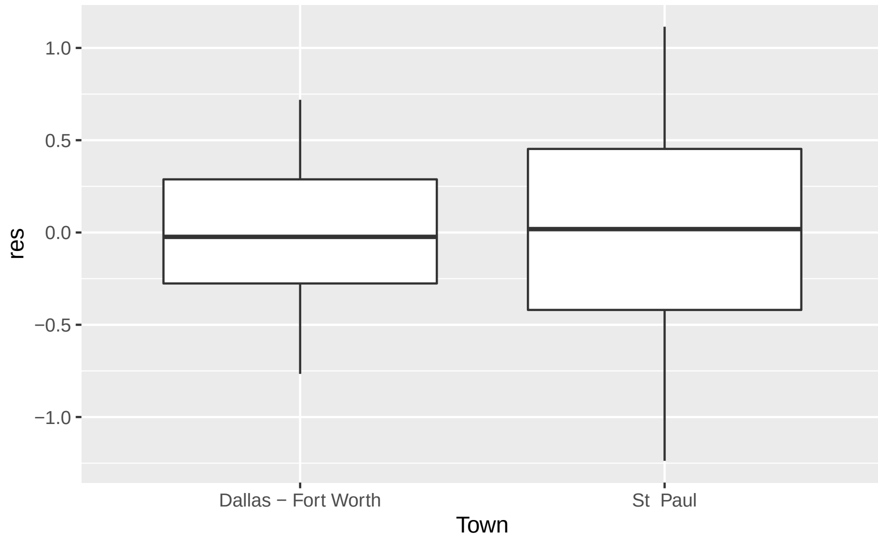
```
skin %>%
  ggplot(aes(fit, res)) + geom_point()
```




```
skin %>%
  ggplot(aes(Age, res)) + geom_boxplot()
```



```
skin %>%
  ggplot(aes(Town, res)) + geom_boxplot()
```



Still a slight sign of increase variance for larger values of fitted values. No patterns in residual versus fitted so the model seems good.

9. Use the model to predict the probability of skin cancer for a 51 year old living in Texas.

```
new_pt <- tibble(Age = "45-54", Town = "Dallas - Fort Worth")
logit_pred <- predict(M3, newdata = new_pt)
logit_pred
```

```
##        1
## -6.24334
```

```
pred <- exp(logit_pred) / (1 + exp(logit_pred))
pred
```

```
##           1
## 0.001939584
```

```
predict(M3, newdata = new_pt, type = "response")
```

```
##           1
## 0.001939584
```