# Assignment 3 quarto

## WEEK 10: Homework Assignments

For this assignment, you'll need to start with a blank le in R, and gure out from there what code you'll need to solve the following problems. As ever, submit your R file through Moodle to complete the assignment.

Figuring out which of the techniques we have used thus far is the core point of this assignment - it's a test of how well you have understood regression techniques, not a test of your programming abilities, and you won't be graded harshly for programming mistakes.

This assignment is worth twice as much as previous homework assignments, and has an extended deadline of January 5th

## Assignment 1: Vocational School Graduate Income

The le grad_income.csv contains a large number of observations of the income of vocational school graduates in a certain country. The data includes the gender and age of each person, along with their annual income.

Your task is to analyse this data and nd the most appropriate linear model to explain the differences in income.

Some questions you should consider as you're looking for the best model:

- Is the relationship between age and income linear, or might it be described better by a curvilinear model?
- Is the relationship between gender and income independent of the relationship with age, or is there an interaction between these terms?

### 1) Find the most appropriate and effective model for this data. Show your workings in your R code (i.e what other models you tried, and any tests you did to compare models.)

```
# Load the data
options(scipen = 999)
rm(list = ls())
library(tidyverse)
library(stargazer)
url <- "https://raw.githubusercontent.com/DanielFabioG/data/refs/heads/main/grad_income_2.csv
grad_income <- read.csv(url)

# Check the data
summary(grad_income)
```

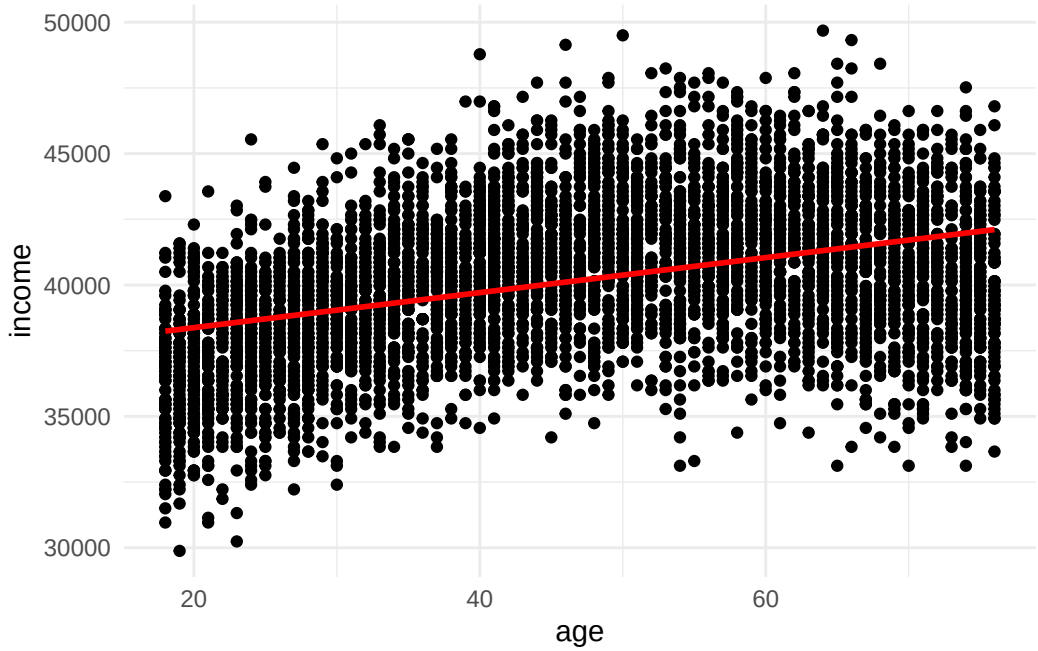
```
    gender                age              income
 Length:5000        Min.   :18.00   Min.   :29880
 Class :character   1st Qu.:32.00   1st Qu.:38160
 Mode  :character   Median :46.00   Median :40140
                    Mean   :46.66   Mean   :40153
                    3rd Qu.:61.00   3rd Qu.:42120
                    Max.   :76.00   Max.   :49680
```

First looking at the linear model

```r
# Fit a linear model
linear_model <- (lm(income ~ age, data = grad_income))

# Check the model with stargazer
stargazer(linear_model, type = "text")
```

```
================================================
                        Dependent variable:
                    ---------------------------
                               income
------------------------------------------------
age                          66.641***
                              (2.216)

Constant                   37,043.430***
                             (110.060)

------------------------------------------------
Observations                   5,000
R2                             0.153
Adjusted R2                    0.153
Residual Std. Error     2,667.344 (df = 4998)
F Statistic          904.369*** (df = 1; 4998)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

```r
# GGplot the model
grad_income %>%
ggplot(aes(x = age, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_minimal()
```



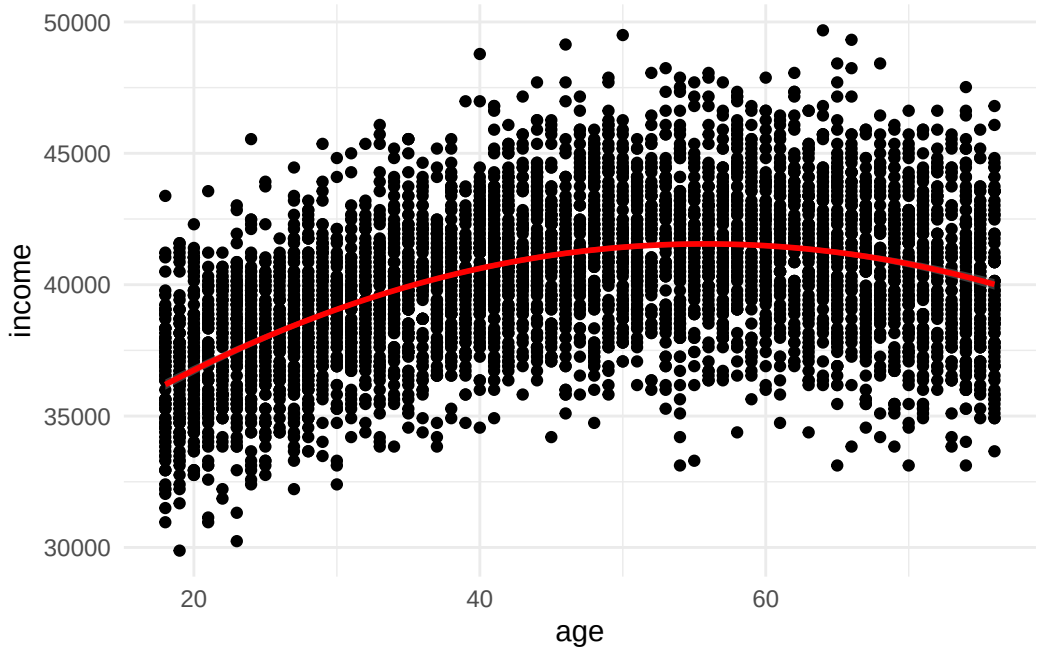

Looking at the quadratic model

```
# Fit a quadratic model
grad_income$gender <- factor(grad_income$gender)


grad_income$age2 <- grad_income$age^2
quad_model <- lm(income ~ age + age2, data = grad_income)


# Check both models with stargazer
stargazer(linear_model, quad_model, type = "text")
```

```
===================================================================
                                Dependent variable:
                    ---------------------------------------------
                                        income
                              (1)                    (2)
-------------------------------------------------------------------
age                        66.641***              418.733***
                           (2.216)                (12.737)
```

```
age2                                                    -3.755***
                                                         (0.134)

Constant                          37,043.430***          29,877.850***
                                    (110.060)              (275.513)

-------------------------------------------------------------------------
Observations                          5,000                  5,000
R2                                    0.153                  0.268
Adjusted R2                           0.153                  0.268
Residual Std. Error          2,667.344 (df = 4998)  2,479.995 (df = 4997)
F Statistic                904.369*** (df = 1; 4998) 915.418*** (df = 2; 4997)
=========================================================================
Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

```r
# GGplot the quadratic model
grad_income %>%
  ggplot(aes(x = age, y = income)) +
  geom_point() +
  geom_smooth(method = "lm",
              formula = y ~ x + I(x^2),
              color = "red")+
  theme_minimal()
```

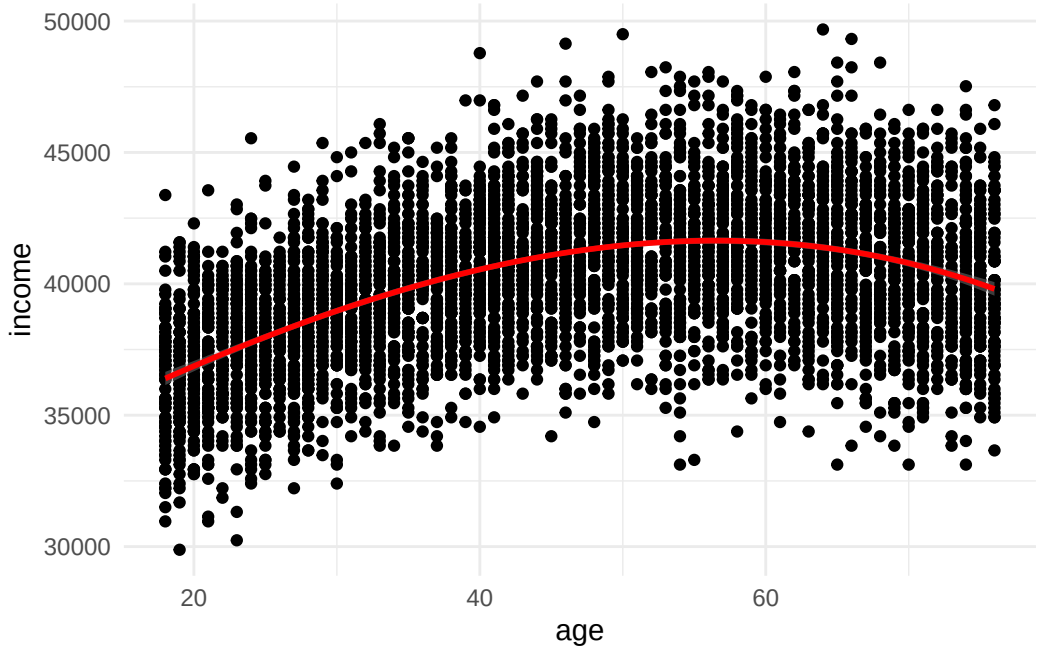


Continuing with looking at cubic model

```
# Fit a polynomial model

poly_model <- lm(income ~ poly(age, 3), data = grad_income)

# Check all models with stargazer
stargazer(linear_model, quad_model, poly_model, type = "text")
```

```
=================================================================================================
                                            Dependent variable:
                        ---------------------------------------------------------------------
                                                    income
                                (1)                   (2)                        (3)
-------------------------------------------------------------------------------------------------
age                          66.641***             418.733***
                             (2.216)               (12.737)

age2                                               -3.755***
                                                   (0.134)

poly(age, 3)1                                                                 80,214.310***
                                                                              (2,478.723)

poly(age, 3)2                                                                -69,469.320***
                                                                              (2,478.723)

poly(age, 3)3                                                                 -6,135.435**
                                                                              (2,478.723)

Constant                   37,043.430***         29,877.850***              40,152.740***
                             (110.060)             (275.513)                  (35.054)

-------------------------------------------------------------------------------------------------
Observations                   5,000                 5,000                      5,000
R2                             0.153                 0.268                      0.269
Adjusted R2                    0.153                 0.268                      0.269
Residual Std. Error   2,667.344 (df = 4998)  2,479.995 (df = 4997)      2,478.723 (df = 499
F Statistic         904.369*** (df = 1; 4998) 915.418*** (df = 2; 4997) 612.947*** (df = 3; 4
=================================================================================================
Note:                                                                  *p<0.1; **p<0.05; ***p<
```

```r
# Ggplot the cubic model

grad_income %>%
  ggplot(aes(x = age, y = income)) +
  geom_point() +
  geom_smooth(method = "lm",
              formula = y ~ poly(x, 3),
              color = "red")+
  theme_minimal()
```




```r
# Code copied from week 09 R assignment to look for the best polynomial model
poly_results <- tibble(
  Level = numeric(),
  AdjR2 = numeric(),
  Significance = numeric(),
  Stars = character()
)
for (polycount in 10:3) {
  fit <- lm(income ~ poly(age, polycount), data = grad_income)
  signif <- summary(fit)$coefficients[polycount + 1, 4]
  poly_results <- add_row(poly_results,
                          Level = polycount,
```

```
                          AdjR2 = summary(fit)$adj.r.squared,
                          Significance = signif,
                          Stars = case_when(
                            signif < 0.01 ~ "***",
                            signif < 0.05 ~ " **",
                            signif < 0.1 ~ "  *",
                            TRUE ~ "   "
                          ))
}
poly_results
```

```
# A tibble: 8 x 4
  Level AdjR2 Significance Stars
  <dbl> <dbl>        <dbl> <chr>
1    10 0.268       0.211  "   "
2     9 0.268       0.259  "   "
3     8 0.268       0.835  "   "
4     7 0.268       0.352  "   "
5     6 0.268       0.721  "   "
6     5 0.268       0.829  "   "
7     4 0.269       0.501  "   "
8     3 0.269       0.0133 " **"
```

The linear model is the worst, while the best model seems to be the cubic model with a slightly higher adjusted R-squared value and a lower p-value than the quadratic model.

Lets continue and look at the log and exp model

```
# Fit a exp
exp_model <- lm(log(income) ~ age, data = grad_income)

# Fit a log model
log_model <- lm(income ~ log(age), data = grad_income)

# log log model
log_log_model <- lm(log(income) ~ log(age), data = grad_income)

# Check poly model with stargazer
stargazer(exp_model, log_model,log_log_model, type = "text")
```

```
=================================================================
                                        Dependent variable:
                                ---------------------------------
                                log(income)    income     log(income)
                                    (1)          (2)          (3)
-----------------------------------------------------------------
age                              0.002***
                                 (0.0001)

log(age)                                     3,175.290***    0.080***
                                               (90.617)      (0.002)

Constant                         10.519***  28,191.940***  10.295***
                                  (0.003)     (343.310)     (0.009)

-----------------------------------------------------------------
Observations                       5,000        5,000        5,000
R2                                 0.156        0.197        0.201
Adjusted R2                        0.155        0.197        0.201
Residual Std. Error (df = 4998)    0.067      2,597.124      0.065
F Statistic (df = 1; 4998)      921.217***  1,227.858***  1,259.749***
=================================================================
Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

Log and exp models are worse as I expected.

Lastly looking at reciprocal model

```r
# Fit a reciprocal model
reciprocal_model <- lm(income ~ I(1/age), data = grad_income)

# Check poly model vs reciprocal with stargazer
stargazer(poly_model,reciprocal_model, type = "text")
```

```
=============================================================================
                                      Dependent variable:
                          -----------------------------------------------
                                             income
                                 (1)                          (2)
-----------------------------------------------------------------------------
poly(age, 3)1              80,214.310***
```

```
                            (2,478.723)

poly(age, 3)2            -69,469.320***
                            (2,478.723)

poly(age, 3)3             -6,135.435**
                            (2,478.723)

I(1/age)                                              -123,991.700***
                                                        (3,235.463)

Constant                  40,152.740***               43,278.820***
                            (35.054)                    (89.179)

------------------------------------------------------------------------
Observations                  5,000                       5,000
R2                            0.269                       0.227
Adjusted R2                   0.269                       0.227
Residual Std. Error   2,478.723 (df = 4996)      2,548.316 (df = 4998)
F Statistic         612.947*** (df = 3; 4996) 1,468.629*** (df = 1; 4998)
========================================================================
Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Cubic model is conslusively the best model.

Now looking if there is a relationship in the model between age and gender

```r
# Fit a model with interaction
interaction_model <- lm(income ~ age + gender, data = grad_income)

# cubic model with interaction
interaction_cubic_model <- lm(income ~ poly(age, 3) + gender, data = grad_income)

# Check the model with stargazer
stargazer(interaction_model, interaction_cubic_model, type = "text")
```

```
========================================================================
                                Dependent variable:
                    ---------------------------------------------------
                                        income
                            (1)                         (2)
```

```
------------------------------------------------------------------------
age                          66.103***
                              (2.027)

poly(age, 3)1                                          79,573.210***
                                                        (2,237.388)

poly(age, 3)2                                         -68,548.260***
                                                        (2,237.474)

poly(age, 3)3                                          -5,418.669**
                                                        (2,237.408)

gendermale                 2,159.574***                 2,134.453***
                             (68.993)                     (63.291)

Constant                   35,987.490***                39,084.240***
                             (106.148)                    (44.777)

------------------------------------------------------------------------
Observations                    5,000                        5,000
R2                              0.292                        0.405
Adjusted R2                     0.292                        0.404
Residual Std. Error    2,439.179 (df = 4997)      2,237.307 (df = 4995)
F Statistic         1,030.625*** (df = 2; 4997) 848.611*** (df = 4; 4995)
========================================================================
Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

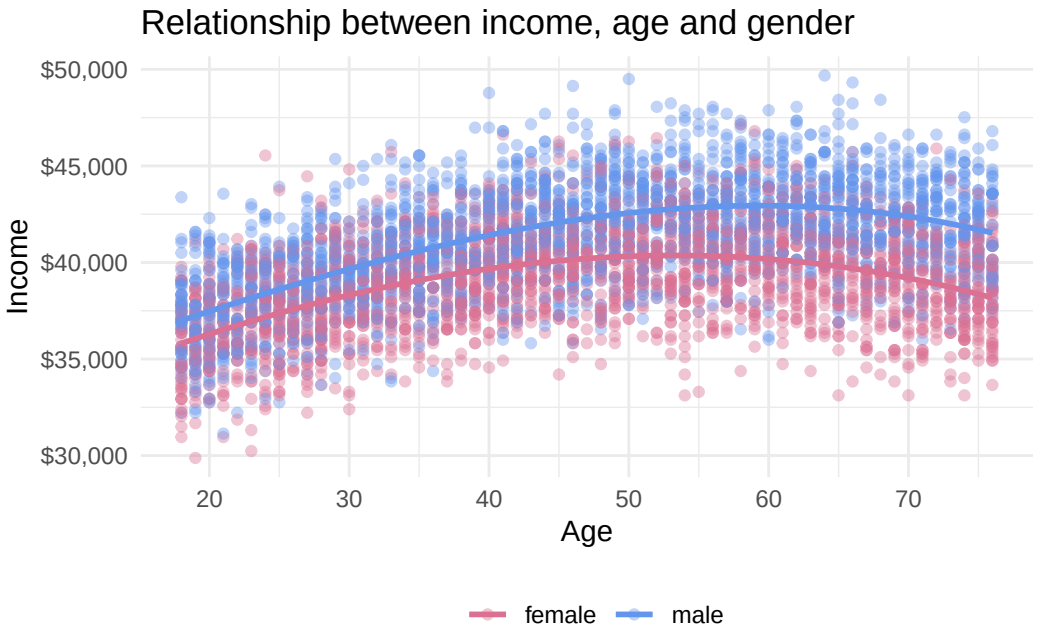
### 2) Write a few lines justifying your choice of model.

The interaction cubic model is the best model as it has the highest adjusted R-squared value and the lowest p-value, it also includes the interaction between age and gender which is important as the relationship between age and income is different between what type of gender the person is.

### 3) Create a graph which illustrates the relationship between age, gender, and income.

```
# Ggplot the interaction cubic model
grad_income %>%
```

```r
ggplot(aes(x = age, y = income, color = gender)) +
geom_point(alpha = 0.4) +
geom_smooth(method = "lm",
            formula = y ~ poly(x, 3),
            se = FALSE) +
scale_color_manual(values = c("palevioletred", "cornflowerblue")) +
# fixing the y scale values to dollars
scale_y_continuous(labels = scales::dollar_format()) +
# fixing the x scale for age
scale_x_continuous(breaks = seq(0, 100, 10)) +
labs(title = "Relationship between income, age and gender",
     x = "Age",
     y = "Income",
     color = "") +
theme_minimal()+
theme(legend.position = "bottom")
```



## Assignment 2: Titanic Survivors

The file titanic.csv contains the passenger data from the Titanic, including their survival or nonsurvival of the ship's sinking. We saw in class that paying more for a ticket increased the chances of survival, as did being female.

Your task is to look at a few additional factors and see how they may have impacted the chances of surviving the Titanic. Bear in mind that many of these columns will need to be turned into factors before they can be effectively used in a regression. Also take careful note of the data in your outcome variable, and make sure you're using the right regression model.

### 1) The column `pclass` tells us which class the passenger was in - the highest class is rst, with third-class being the cheapest accommodation on the ship. We know `fare` is correlated with survival; we want to know if this is just because the more expensive first-class accommodation was better positioned for survival.

Test whether `pclass` is also correlated with survival, and see if `fare` is still significant even once `pclass` is included.

```
# Load the data
url <- "https://raw.githubusercontent.com/DanielFabioG/data/refs/heads/main/titanic.csv"

titanic <- read_csv(url)

# Check the data
glimpse(titanic)
```

```
Rows: 1,037
Columns: 10
$ pclass   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ survived <dbl> 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1~
$ name     <chr> "Allen, Miss. Elisabeth Walton", "Allison, Master. Hudson Tre~
$ sex      <chr> "female", "male", "female", "male", "female", "male", "female~
$ age      <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, 63.0000, ~
$ sibsp    <dbl> 0, 1, 1, 1, 1, 0, 1, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0~
$ parch    <dbl> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0~
$ ticket   <chr> "24160", "113781", "113781", "113781", "113781", "19952", "13~
$ fare     <dbl> 52.83438, 25.25833, 25.25833, 25.25833, 25.25833, 26.55000, 2~
$ cabin    <chr> "B5", "C22 C26", "C22 C26", "C22 C26", "C22 C26", "E12", "D7"~
```

```r
# survived and pclass are not factors but R understands
# them as factors since they are 1 and 0 and 1,2,3 respectively

# Fit a model with pclass and fare
pclass_model <- glm(survived ~ pclass + fare, family = binomial(), data = titanic)
fare_model <- glm(survived ~ fare, family = binomial(), data = titanic)

# Check the model with stargazer
stargazer(fare_model,pclass_model, type = "text")
```

```
================================================
                       Dependent variable:
                   ---------------------------
                              survived
                          (1)             (2)
------------------------------------------------
pclass                                  -0.839***
                                         (0.118)

fare                    0.036***         -0.002
                        (0.005)          (0.006)

Constant               -0.948***        1.489***
                        (0.103)          (0.343)

------------------------------------------------
Observations             1,037            1,037
Log Likelihood         -671.195         -646.826
Akaike Inf. Crit.     1,346.390        1,299.653
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Fare is correlated with survival looking at the model with only fare, but when pclass is included fare is no longer significant, but pclass is.

### 2) "Women and children first" was the famous policy for boarding lifeboats in shipwrecks at this time. We found that women were much more likely to survive than men, but what about children?

To test this, you will need create a new factor variable that separates out women, men, and children (we probably don't care about the gender of children, so we can just treat them as a third separate category).

To construct this new column, I suggest that you look for information about the ifelse() function (which we discussed a little in class) and the case_when() function. You'll need to use a combination of information from the `age` column and the `sex` column.

Once it's created, test whether being a woman or a child makes a major difference to survival.

```
# Creating new column for men, women and children
titanic <- titanic %>%
  mutate(
    person_type = case_when(
      age < 16                  ~ "child",
      sex == "female"           ~ "woman",
      sex == "male" | is.na(sex) ~ "man"   # Catch any NA for safety
    ),
    # make this a factor for cleaner analyses
    person_type = factor(person_type, levels = c("man", "woman", "child"))
  )
```

```
# Fit a model with person_type
person_type_model <- glm(survived ~ person_type, family = binomial(), data = titanic)

# Check the model with stargazer
stargazer(person_type_model, type = "text")
```

```
================================================
                        Dependent variable:
                    ----------------------------
                               survived
------------------------------------------------
person_typewoman               2.739***
                                (0.169)

person_typechild               1.840***
                                (0.217)
```

```
Constant                    -1.542***
                             (0.108)

-----------------------------------------------
Observations                  1,037
Log Likelihood              -533.019
Akaike Inf. Crit.           1,072.037
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

```
chisq.test(table(titanic$person_type, titanic$survived))
```

```
    Pearson's Chi-squared test

data:  table(titanic$person_type, titanic$survived)
X-squared = 322.03, df = 2, p-value < 0.00000000000000022
```

The chi-squared test shows that there is a significant difference between the survival of male, female and children.

```
xtabs(~person_type+survived, data = titanic)
```

```
           survived
person_type   0   1
      man   486 104
      woman  77 255
      child  49  66
```

The table shows that being a woman or a child makes a major difference to survival than being a man. It also shows that being a woman is more likely to survive than being a child.

```
anova(pclass_model, person_type_model, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: survived ~ pclass + fare
Model 2: survived ~ person_type
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1034     1293.7
2      1034     1066.0  0   227.62
```

Looking at the models we can see that the person_type model is better than the pclass model, as the large drop in residual deviance shows, which means the person_type model is a better fit for the data. We cannot see the p-value for the anova test, because the person_type model is not nested in the pclass model.

**3) Finally, combine these factors and find the most effective model to explain the odds of surviving the Titanic, based on whatever you think is the most effective combination of ticket price, passenger class, gender, and age. You don't have to include all these factors if you don't think they're all necessary.**

```
# Fit a model with person_type, pclass and fare
persontypefare <- glm(survived ~ person_type * pclass + pclass + fare, family = binomial()
                      , data = titanic)

# Check the model with stargazer
stargazer(persontypefare, type = "text")
```

```
================================================
                         Dependent variable:
                     ---------------------------
                                survived
------------------------------------------------
person_typewoman                6.154***
                                 (0.735)

person_typechild                7.940***
                                 (1.873)

pclass                          -0.367**
                                 (0.174)

fare                             0.009
                                (0.008)

person_typewoman:pclass         -1.444***
                                 (0.284)

person_typechild:pclass         -2.109***
                                 (0.652)

Constant                         -0.908*
                                 (0.489)

------------------------------------------------
Observations                      1,037
```

```
Log Likelihood                   -460.295
Akaike Inf. Crit.                 934.591
================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```r
anova(person_type_model, persontypefare, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: survived ~ person_type
Model 2: survived ~ person_type * pclass + pclass + fare
  Resid. Df Resid. Dev Df Deviance              Pr(>Chi)
1      1034    1066.04
2      1030     920.59  4   145.45 < 0.00000000000000022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seems like adding it all together plus the interaction between person_type and pclass is a better fit for the data, as the large drop in residual deviance shows. The p-value is also significant, which means the final model is a better fit for the data than the person_type model.

Lastly I will make some new variables to see if they can improve the model even further.

```r
# Creating a new categorical variable for family size or if the person is alone

titanic <- titanic %>%
  mutate(
    family_size = sibsp + parch,
    is_alone = ifelse(family_size == 0, "alone", "not_alone"),
    is_alone = factor(is_alone)
  )

# Creating a new variable for looking if living closer to the deck had an impact on survival
titanic <- titanic %>%
  mutate(
    deck = substr(cabin, 1, 1), # Adding na to the missing values
    deck = ifelse(is.na(deck), "no_cabin", deck),
    deck = factor(deck)
  )

# Lastly I will extract the title from the name column
```

```r
titanic <- titanic %>%
  mutate(
    title = gsub("(.*, )|(\\..*)", "", name),
    title = factor(title)
  )
```

```r
# Fit a model with person_type, pclass, fare, family_size, is_alone, deck and title
final_model <- glm(survived ~ person_type * pclass + fare + family_size + is_alone + deck + t
                   family = binomial()
                   , data = titanic)

# Check the model with stargazer
stargazer(final_model, type = "text")
```

```
=================================================
                          Dependent variable:
                      ---------------------------
                                survived
-------------------------------------------------
person_typewoman                4.070***
                                 (1.502)

person_typechild                6.127***
                                 (2.192)

pclass                           0.083
                                (0.225)

fare                             0.008
                                (0.009)

family_size                     -0.419***
                                 (0.101)

is_alonenot_alone                0.514*
                                 (0.282)

deckB                            -0.173
                                 (0.758)

deckC                            -0.990
```

| | |
|---|---|
| | (0.717) |
| deckD | 0.184 |
| | (0.759) |
| deckE | 0.568 |
| | (0.755) |
| deckF | -0.145 |
| | (0.958) |
| deckG | -1.212 |
| | (1.182) |
| deckno_cabin | -1.274* |
| | (0.690) |
| deckT | -16.052 |
| | (2,399.545) |
| titleCol | 16.853 |
| | (2,399.545) |
| titleDon | 0.839 |
| | (3,393.469) |
| titleDona | 31.160 |
| | (3,393.469) |
| titleDr | 17.013 |
| | (2,399.545) |
| titleLady | 30.048 |
| | (3,393.469) |
| titleMajor | 16.706 |
| | (2,399.545) |
| titleMaster | 17.685 |
| | (2,399.545) |
| titleMiss | 18.024 |
| | (2,399.545) |

```
titleMlle                          30.745
                                  (2,929.544)

titleMme                           30.356
                                  (3,393.469)

titleMr                            15.554
                                  (2,399.545)

titleMrs                           18.570
                                  (2,399.545)

titleMs                            33.158
                                  (3,393.469)

titleRev                            0.884
                                  (2,544.634)

titleSir                           32.596
                                  (3,393.469)

titlethe Countess                  30.403
                                  (3,393.469)

person_typewoman:pclass           -1.611***
                                   (0.307)

person_typechild:pclass           -2.002***
                                   (0.697)

Constant                           -16.437
                                  (2,399.545)

-------------------------------------------------
Observations                         1,037
Log Likelihood                     -429.162
Akaike Inf. Crit.                   924.325
=================================================
Note:                   *p<0.1; **p<0.05; ***p<0.01
```

```r
# looking at the categorical variables to see if the model is overfitting
table(titanic$deck, titanic$survived)
```

```
            0   1
  A         7  11
  B        16  45
  C        34  52
  D        12  30
  E        10  28
  F         6  12
  G         2   3
  no_cabin 524 244
  T         1   0
```

```r
table(titanic$title, titanic$survived)
```

```
                0   1
  Capt          1   0
  Col           2   2
  Don           1   0
  Dona          0   1
  Dr            3   4
  Lady          0   1
  Major         1   1
  Master       25  28
  Miss         64 146
  Mlle          0   2
  Mme           0   1
  Mr          475  98
  Mrs          32 138
  Ms            0   1
  Rev           8   0
  Sir           0   1
  the Countess  0   1
```

```r
table(titanic$is_alone, titanic$survived)
```

```
              0   1
  alone     395 186
  not_alone 217 239
```

```r
anova(persontypefare, final_model, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: survived ~ person_type * pclass + pclass + fare
Model 2: survived ~ person_type * pclass + fare + family_size + is_alone +
    deck + title
  Resid. Df Resid. Dev Df Deviance   Pr(>Chi)
1      1030     920.59
2      1004     858.32 26   62.266 0.00008255 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model is improved but it got worse by overfitting the data as many of the new categories have too few observations to be significant. The p-value is also not significant, which means the final model is not a better fit for the data than the persontypefare model at the moment, I will try to group them better to see if the model improves correctly.

```r
titanic <- titanic %>%
  mutate(deck_grouped = case_when(
    deck %in% c("G","T","F") ~ "rare_deck",
    deck %in% c("A","B","C","D","E") ~ as.character(deck),  # keep them
    deck == "no_cabin" ~ "no_cabin",
    TRUE ~ deck  # just in case
  ))

titanic <- titanic %>%
  mutate(deck_grouped = factor(deck_grouped))

# Combining the rare titles

titanic <- titanic %>%
  mutate(title_grouped = case_when(
    title %in% c("Dr", "Rev", "Major", "Col", "Capt", "Jonkheer", "Don", "Sir") ~ "rare_titl
    title %in% c("Mme", "Ms", "Lady", "Mlle", "the Countess", "Dona") ~ "Miss",
    TRUE ~ title))
```

```r
# looking at the categorical variables to see if the model is overfitting
table(titanic$deck_grouped, titanic$survived)
```

```
             0   1
  A          7  11
  B         16  45
  C         34  52
  D         12  30
  E         10  28
  no_cabin 524 244
  rare_deck   9  15
```

```r
table(titanic$title_grouped, titanic$survived)
```

```
              0   1
  Master      25  28
  Miss        64 153
  Mr         475  98
  Mrs         32 138
  rare_title  16   8
```

```r
# Fit a model with person_type, pclass, fare, family_size, is_alone, deck_grouped and title_g
final_model_2 <- glm(survived ~ person_type * pclass + fare + family_size + is_alone + deck_g
                     family = binomial()
                     , data = titanic)

# Check the model with stargazer
stargazer(final_model_2, type = "text")
```

```
================================================
                         Dependent variable:
                     ---------------------------
                                survived
------------------------------------------------
person_typewoman                4.287***
                                 (1.470)
```

| | |
|---|---|
| person_typechild | 6.451*** |
| | (2.186) |
| pclass | 0.132 |
| | (0.226) |
| fare | 0.009 |
| | (0.009) |
| family_size | -0.416*** |
| | (0.101) |
| is_alonenot_alone | 0.498* |
| | (0.277) |
| deck_groupedB | -0.463 |
| | (0.712) |
| deck_groupedC | -1.185* |
| | (0.681) |
| deck_groupedD | -0.074 |
| | (0.727) |
| deck_groupedE | 0.304 |
| | (0.725) |
| deck_groupedno_cabin | -1.596** |
| | (0.653) |
| deck_groupedrare_deck | -0.883 |
| | (0.849) |
| title_groupedMiss | 0.291 |
| | (0.487) |
| title_groupedMr | -2.120* |
| | (1.264) |
| title_groupedMrs | 0.818 |
| | (0.584) |
| title_groupedrare_title | -1.507 |

```
                                  (1.343)

person_typewoman:pclass          -1.663***
                                  (0.308)

person_typechild:pclass          -2.106***
                                  (0.697)

Constant                           1.407
                                  (1.452)

--------------------------------------------------
Observations                       1,037
Log Likelihood                   -434.162
Akaike Inf. Crit.                 906.324
==================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```r
anova(persontypefare, final_model_2, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: survived ~ person_type * pclass + pclass + fare
Model 2: survived ~ person_type * pclass + fare + family_size + is_alone +
    deck_grouped + title_grouped
  Resid. Df Resid. Dev Df Deviance     Pr(>Chi)
1      1030     920.59
2      1018     868.32 12   52.267 0.0000005561 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even though the p-value is significant and the model is improved, the final model is not a good enough improvement to the persontypemodel. Personally I think the model is too complex and gets improved too little for it to be worth to devide the categories even further. Lastly I will remove some of the variables that are not significant and see how much I can remove to get a better model that is not overfitting the data.

```r
simple_model <- glm(
  survived ~ person_type * pclass + pclass + family_size + is_alone,
  family = binomial(link = "logit"),
  data = titanic
```

```
)

stargazer(persontypefare, simple_model, type = "text")
```

```
=================================================
                          Dependent variable:
                      ---------------------------
                                survived
                          (1)            (2)
-------------------------------------------------
person_typewoman         6.154***       6.151***
                         (0.735)        (0.724)

person_typechild         7.940***       7.235***
                         (1.873)        (1.816)

pclass                   -0.367**       -0.475***
                         (0.174)        (0.132)

fare                      0.009
                         (0.008)

family_size                             -0.420***
                                         (0.097)

is_alonenot_alone                        0.659***
                                         (0.255)

person_typewoman:pclass  -1.444***      -1.400***
                         (0.284)        (0.280)

person_typechild:pclass  -2.109***      -1.643***
                         (0.652)        (0.637)

Constant                 -0.908*        -0.523*
                         (0.489)        (0.310)

-------------------------------------------------
Observations              1,037          1,037
Log Likelihood           -460.295       -449.600
Akaike Inf. Crit.        934.591        915.201
```

```
================================================
Note:                    *p<0.1; **p<0.05; ***p<0.01
```

```r
anova(persontypefare, simple_model, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: survived ~ person_type * pclass + pclass + fare
Model 2: survived ~ person_type * pclass + pclass + family_size + is_alone
  Resid. Df Resid. Dev Df Deviance    Pr(>Chi)
1      1030     920.59
2      1029     899.20  1    21.39 0.000003748 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 4) Write a few lines justifying your choice of this model.

I chose the "simple model" because it is the best fit for the data, as it improves the model significantly while removing the term fare which didnt seem to fit the model and adds family size and shows if the passenger were alone, which seems to improve it also.

The p-value is also significant, which means the simple model is a better fit for the data than the persontypefare model. The simple model is also not overfitting the data, as the p-values for the variables are significant and the model is not too complex. The simple model is also relatively easy to interpret and understand, which is important when presenting the results to others. The final_model_2 is too complex and does not improve the model enough to be worth the extra complexity in my opinion.