1. a. Consider the following linear model,

$$\mathbb{E}(y_i|\mathbf{x}_i) = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_i + \beta_3(\text{sex}_i \times \text{age}_i),$$

where sex_i is a binary variable (0 = male, 1 = female) and age_i is a continuous variable. Write down the model separately for males and females and using the two models give interpretations for the four parameters.

Model for males: sex = 0
=> $E(y_i \mid x_i) = B_0 + B_2 \text{age}_i$
Model for females: sex = 1
=> => $E(y_i \mid x_i) = B_0 + B_1 + B_2 \text{age}_i + B_3 \text{age}_i$
Interpretations of the four parameters:
B_0: when males are 0 years old, $y_i = B_0$ and for females, $y_i = B_0 + B_1$ => B_o is the common parameter to be added to the model for both males and females, no matter at what age they are
B_1: this is the constant difference between males and females y_i that is independent of age
B_2: this is a common parameter that adds to the model for both genders, scaled by their age. It denotes the similarity between the 2 sexes in proportion to age
B_3: this parameter is scaled with age for females. This is an additional effect of age that affects only females but not males

b. The data set `galaxy` from the package `ElemStatLearn` contains measurements on the position and radial velocity of the galaxy NGC7531. Fitting a model with the latter as a response, we get the following model summary and residual plot. Does the model fit well? If not, what could be tried next?

```
library(ElemStatLearn)
library(car)
```

```
## Loading required package: carData
```

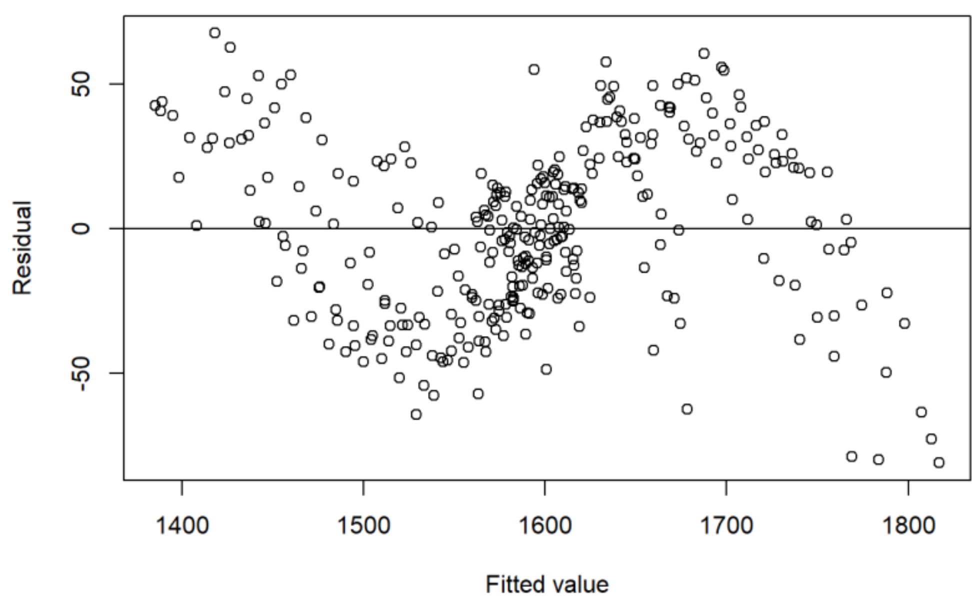
```
lm_galaxy <- lm(velocity ~ ., data = galaxy)
summary(lm_galaxy)
```

```
##
## Call:
## lm(formula = velocity ~ ., data = galaxy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.988 -23.673   0.442  22.770  67.527
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1589.42295    3.92939 404.496  < 2e-16 ***
## east.west          0.77410    0.31202   2.481  0.01362 *
## north.south       -3.19179    0.09697 -32.914  < 2e-16 ***
## angle              0.12454    0.04396   2.833  0.00491 **
## radial.position    0.90118    0.16042   5.618 4.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.13 on 318 degrees of freedom
## Multiple R-squared:  0.8991, Adjusted R-squared:  0.8979
## F-statistic: 708.6 on 4 and 318 DF,  p-value: < 2.2e-16
```

```
vif(lm_galaxy)
```

```
##       east.west     north.south           angle radial.position
##        4.996114        1.747546        1.002817        6.118775
```

```
plot(fitted(lm_galaxy), resid(lm_galaxy), xlab = "Fitted value", ylab = "Residual")
abline(h = 0)
```



- The line lm_galaxy <- lm(velocity ~ ., data = galaxy) means that the a linear regression model is done with response variable velocity and explanatory variables are other attributes of the galaxy
- Ideally, residual values should be **equally and randomly spaced around the horizontal axis** y = 0 in the residual-fitted value graph. From the scatterplot above, it indicates that the linear regression model is not good, because the data are distributed non-linearly but rather in a third degree polynomial shape.
- A high R squared doesn't necessarily mean a good fit. The regression line consistently under and over-predicts the data along the curve, which is bias. The Residuals versus Fits plot emphasizes this unwanted pattern. An unbiased model has residuals that are randomly scattered around zero. Non-random residual patterns indicate a bad fit despite a high R2

Other solutions: Since the linear model regression doesnt fit well, we could have several different options:

- Use Polynomial regressions: It will fit the data better
- Reduce number of predicting variables by backward selection method and continue to use the linear model
- Or start using forward selection from begin with no explaining variables until the model is well explained by the chosen variables from the method