Yuelin Liu
101154473

Final Report

1. Introduction to the problem

It is believed that a person may have more work experience if he or she is healthy, educated, good-looking, and has a well-paid job. But in fact, what factors will influence people's work experience? By analyzing 1260 people's information, I'd like to find out the most important factors related to people's work experience by using regression analysis.

I will also look at which gender is likely to have more work experience. Does marital status influence people's work experience? Does it have large impact on male or female? What's the most popular score people give themselves for their attractiveness. Is there any difference between male and female by giving themselves appearance score? Is it true that people who have the highest wage also have highest education level? It is well known that if a person is not in a good health condition, he or she might not have the high salary. Does the data tell the same story?

2. Introduction to the data

The data I used is from kaggle, it describes the male and female's work experience and wages related with their beauty, which has value from level 1 to 5. And their race, marriage status, education year related to this.

link to original data:

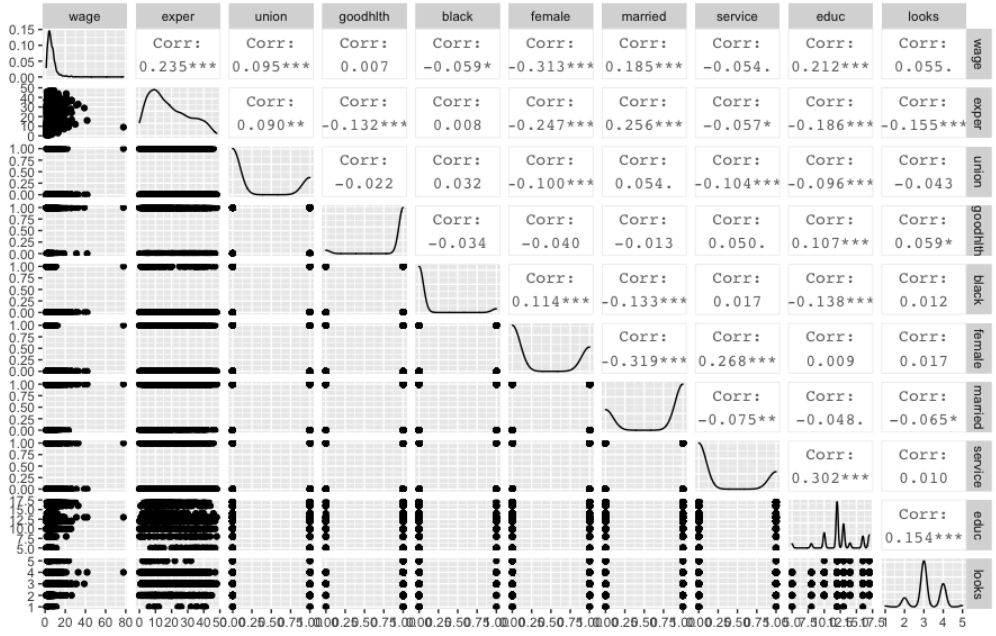
https://www.kaggle.com/aungpyaeap/beauty?select=beauty.csv

variable descriptions:

wage: salary of people
exper: people's year of experience
union: whether people are in the union, an organization to protect their rights and interests
goodhlth: if people are in the good health condition
black: people are black or not
female: people are female or not
married: people's marriage status
service: if people's work is service related
educ: people's year of education
looks: people's appearance score

Exploratory analysis:

A. The graph of all variables against each other:



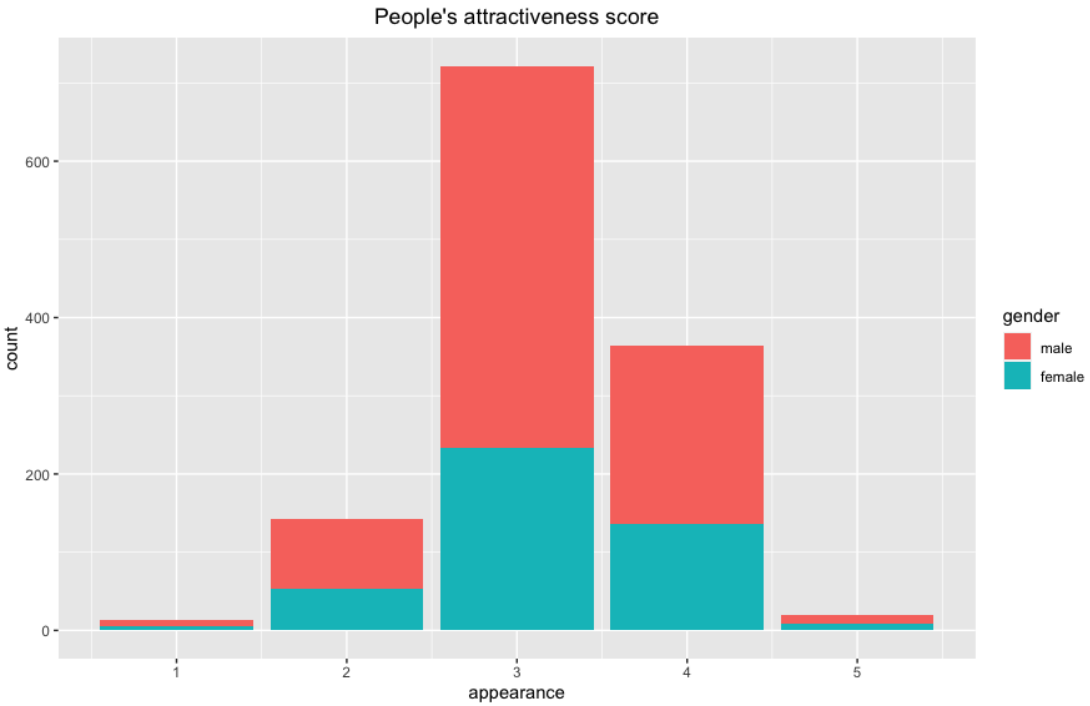
B. People' experience vs wages

Predictor: wage

Response variables: people's year of experience

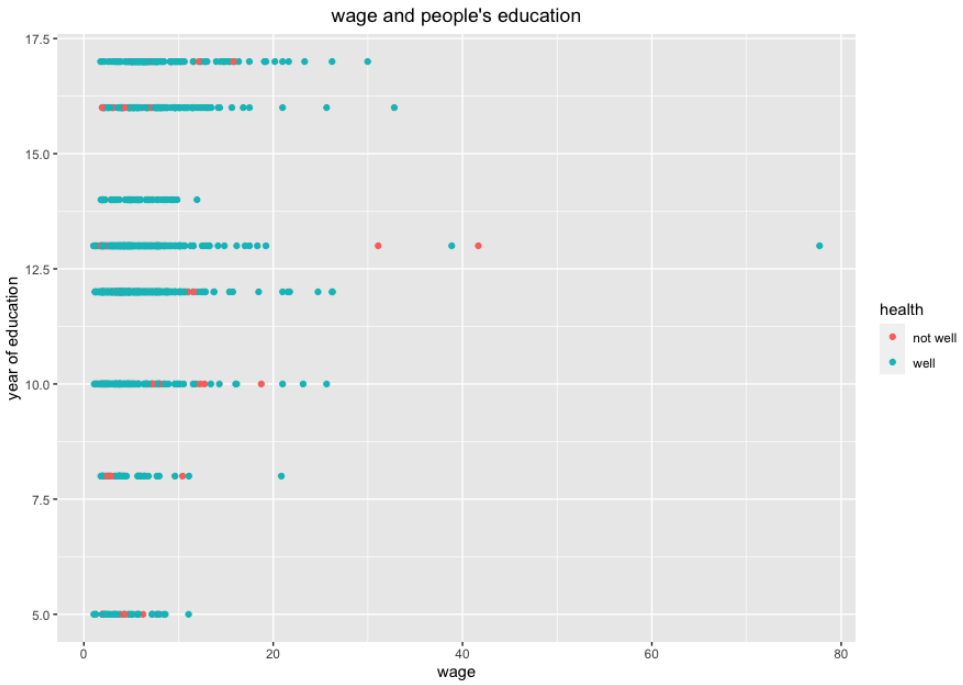
C. Count of people's appearance score

Predictor: appearance score
Response variables: number of people who have this score



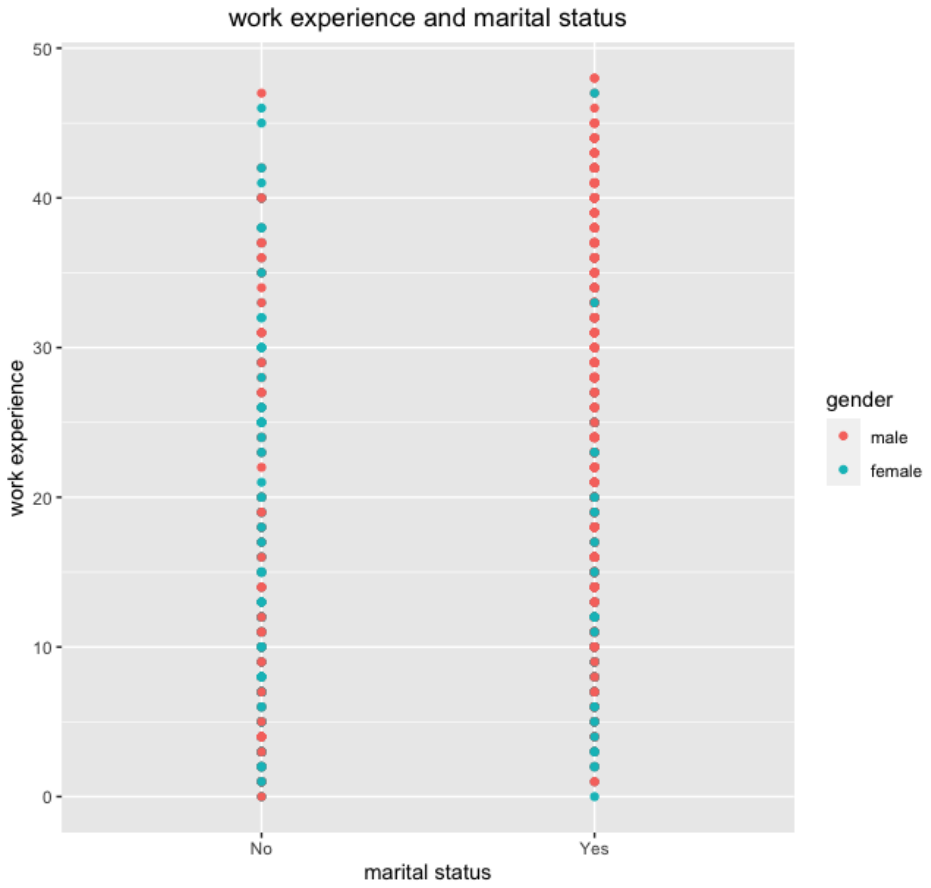
D. People's wage and their education by their health condition

Predictor: wage
Response variable: year of education



D. People's work experience and their marital status

Predictor: people's marital status
Response variable: year of work experience



3. Regression analysis

Regression model:

Full model: $Y_{experience} = \beta 0 + X\text{wage}\beta\text{wage} + X\text{union}\beta\text{union} + X\text{good health}\beta\text{good health} +$

$+ X_{black}\,\beta_{black} + X_{female}\,\beta_{female} + X_{married}\,\beta_{married} + X_{service}\,\beta_{service} + X_{education}\beta_{education} + X_{looks}\,\beta_{looks}$

I split my data into half, so I can use half for model selection, and the other half for model inference.

This is the summary of my full model:

```
Call:
lm(formula = exper ~ ., data = beautyTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-38.785  -7.674  -2.212   6.906  30.538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.49202    3.12203  12.649  < 2e-16 ***
wage         0.44728    0.09175   4.875 1.38e-06 ***
union        1.05888    0.98605   1.074 0.283305
goodhlth    -8.31711    1.59698  -5.208 2.60e-07 ***
black       -0.39629    1.86918  -0.212 0.832168
female      -4.44869    1.00180  -4.441 1.06e-05 ***
married      2.45505    0.98446   2.494 0.012896 *
service      2.37009    1.04877   2.260 0.024174 *
educ        -0.79326    0.17716  -4.478 8.98e-06 ***
looks       -2.21973    0.62398  -3.557 0.000403 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.66 on 622 degrees of freedom
Multiple R-squared:  0.1974,    Adjusted R-squared:  0.1858
F-statistic:    17 on 9 and 622 DF,  p-value: < 2.2e-16
```

Model selection: I’m going to use both forward and backward selection to select the best model.

This is the result from each step I drop the variable with the least AIC value:

```
Start:  AIC=3001.66
exper ~ wage + union + goodhlth + black + female + married +
    service + educ + looks

           Df Sum of Sq   RSS    AIC
- black     1      5.11 70741 2999.7
- union     1    131.14 70867 3000.8
<none>                  70736 3001.7
- service   1    580.78 71317 3004.8
- married   1    707.26 71443 3005.9
- looks     1   1439.17 72175 3012.4
- female    1   2242.60 72978 3019.4
- educ      1   2280.06 73016 3019.7
- wage      1   2702.98 73439 3023.4
- goodhlth  1   3084.56 73820 3026.6

Step:  AIC=2999.71
exper ~ wage + union + goodhlth + female + married + service +
    educ + looks

           Df Sum of Sq   RSS    AIC
- union     1    129.51 70870 2998.9
<none>                  70741 2999.7
+ black     1      5.11 70736 3001.7
- service   1    576.29 71317 3002.8
- married   1    736.32 71477 3004.2
- looks     1   1439.54 72180 3010.4
- female    1   2289.75 73031 3017.8
- educ      1   2304.14 73045 3018.0
- wage      1   2704.94 73446 3021.4
- goodhlth  1   3079.83 73821 3024.6
```

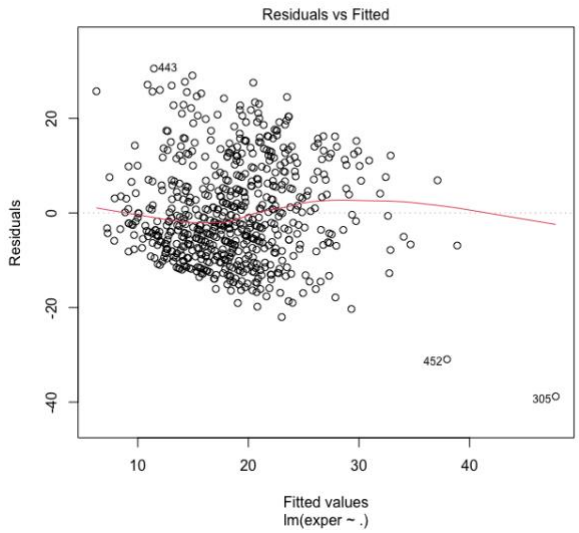
```
Step:  AIC=2998.86
exper ~ wage + goodhlth + female + married + service + educ +
    looks

          Df Sum of Sq   RSS    AIC
<none>                  70870 2998.9
+ union    1    129.51 70741 2999.7
+ black    1      3.48 70867 3000.8
- service  1    551.20 71422 3001.8
- married  1    778.25 71649 3003.8
- looks    1   1449.44 72320 3009.7
- female   1   2272.47 73143 3016.8
- educ     1   2422.84 73293 3018.1
- wage     1   2885.65 73756 3022.1
- goodhlth 1   3059.11 73930 3023.6
```

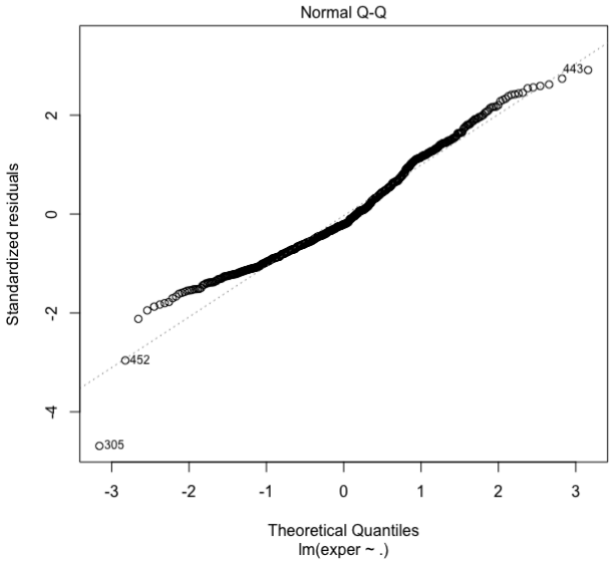
Therefore, the final model I select is:

$Y_{experience} = \beta 0 + X_{\text{wage}}\beta_{\text{wage}} + X_{\text{good health}}\beta_{\text{good health}} + X_{female}\,\beta_{female} + X_{married}\,\beta_{married} + X_{service}\,\beta_{service} + X_{education}\beta_{education} + X_{looks}\,\beta_{looks}$

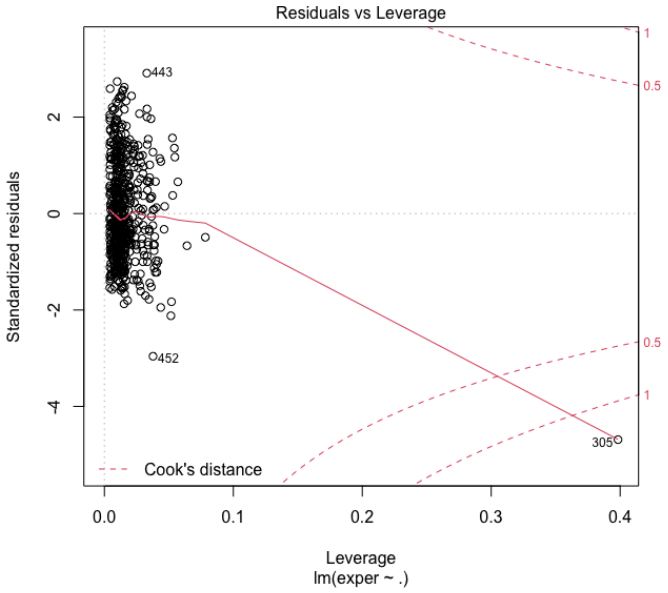
Model diagnostic:

The residual vs fitted plot is flat, there is no relationship, we can assume linear relationship.



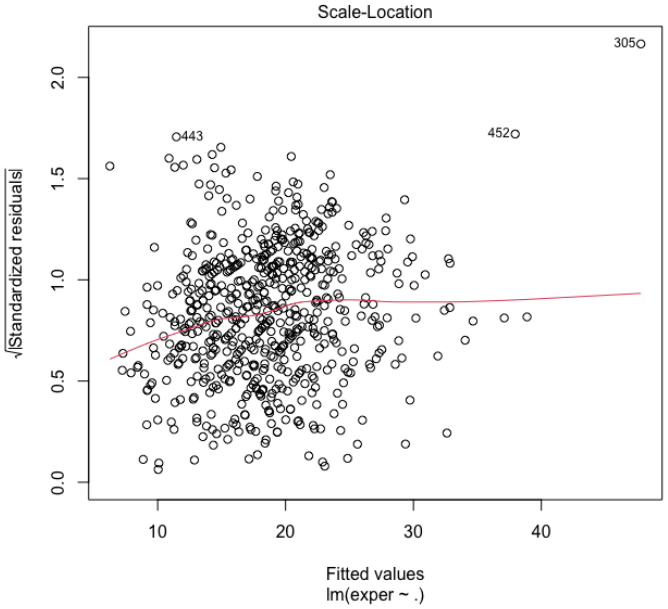


The QQ-plot has most residuals are on the line, but the bottom corner is a little bit higher, which means residuals aren't as low as they should be.



The residuals vs leverage plot shows observations which could have a large impact on the regression slopes, there is a high leverage point that is outside of the dashed lines. So the plot has highly influential outlier.

The scale-location plot shows how residual variance varies with the predicted values. The first half part of the red line is not flat enough, which means that appears to be a non-constant residual variance.



4. Discussion

Result: So, I found that people's work experience is related to their wages, health conditions, gender, marital status, service-related work, education and their appearance.

Prediction: I'm going to predict a person's work experience given his wages of 20.5, good health condition, male, married, not has service-related job, education of 12 years, appearance score of 13.

With 95% prediction interval, I got

fit	lwr	upr
27.06628	5.970739	48.16183

Which means the estimate year of experience is 27.06628 years, with 95% confidence, the prediction interval is from 5.970739 to 48.16183.

Confidence interval:

I use test data to produce the 95% confidence interval:

	2.5 %	97.5 %
(Intercept)	25.08402950	38.4716332
wage	0.48759032	0.9311104
goodhlth	-4.84366416	2.6491047
female	-4.97686803	-0.8479525
married	3.41313908	7.3304008
service	0.06232494	4.2178487
educ	-1.56651786	-0.8381862
looks	-3.03134552	-0.4980337

5. Limitations

There's a person's wage that's too high while has few work experience compared to others , I'd like to delete this person's entry, since it may cause the outlier of the model.

The scores of people's appearance are so subjective, which cannot reflect the real situation, so it might not be reliable. And the number of years of people's experience does not depend on the types of jobs people have. So some types of work, like IT jobs, which needs only a few years of experience to receive relatively high salary compare to other types of jobs. And for the race, the data only has the race of black or not black, which ignores other races like Asian. Therefore, the data needs extra columns and variables to improve its appropriateness. And the data only has 1260 people's information, which is not enough to conclude which factor should be included in the model. Hence, the range and quality of data needs to improve as well.

For the future extensions, I'd like to merge some data so that there will be more variables used to predict people's work experience. I will mutate more columns so that I can study relationships between them better. I would like to try to use different versions of model, such as the square of a specific variable, the logarithm of the response variable, or the interaction between each variable. And I want to make data visualization more attractive by using some tools like scatterplot3d and wesanderson for color palette.

6. Conclusions:

From the plot “wage effects on people’s experience”, I found that most female’s wages are lower than male’s, and more years of work experience does not mean more wages.

From the bar chart “People’s attractiveness score”, I found that most people give them appearance score in the middle range, very few people gave them the lowest score, and the score patterns of men and women are very similar.

I learned that the higher wage is positive related to people’s education level from the graph “wage and people’s education”. The starting salary is a little bit higher for people who are educated for more years. Even though some people don’t have a good health condition, they still make a lot of money.

I also noticed that before people got married, their work experiences are pretty much the same by analyzing the graph “work experience and marital status”, but after they are married, there is a significant change so that men have more work experience than women.

Looking at the regression analysis, I found that people’s work experience is related to their wages, health conditions, gender, marital status, service-related work, education and their appearance. But people’s work experience does not depend on the union they are part of, or the race they belong to.

Code Appendix

```
library(tidyverse)
library(lubridate)

//data is downloaded from
"https://www.kaggle.com/aungpyaeap/beauty?select=beauty.csv"
setwd("Downloads")

//plot of ggpairs
beauty <- read.csv("beauty.csv")
beauty%>%ggpairs()

//mutate a new variable gender
library(dplyr)
beauty <- beauty %>%
   mutate(gender=factor(female, levels = c(0,1), labels=c("male", "female")))

//experience~wage plot
beauty %>% ggplot(aes(y=exper, x=wage, color=gender))+
   geom_point()+
   labs(x="wage", y="year of experience", title="wage effects on people's
experience")+
   theme(plot.title = element_text(hjust = 0.5))

//bar chart of people look
beauty %>% ggplot(aes(x=looks, fill=gender))+
   geom_bar()+
   labs(x="appearance", title="People's attractiveness score")+
   theme(plot.title = element_text(hjust = 0.5))

//mutate a new variable health
beauty <- beauty %>%
   mutate(health=factor(goodhlth, levels = c(0,1), labels=c("not well", "well")))

//education~wage plot
beauty %>% ggplot(aes(y=educ, x=wage, color=health))+
   geom_point()+
   labs(x="wage", y="year of education", title="wage and people's education")+
   theme(plot.title = element_text(hjust = 0.5))

//mutate a new variable married
marry <- beauty %>%
   mutate(got_married=factor(married, levels = c(0,1), labels=c("No", "Yes")))
```

```
//experience~marriage plot
marry %>% ggplot(aes(y=exper, x=got_married, color=gender))+
   geom_point()+
   labs(x="marital status", y="work experience", title="work experience and marital status")+
   theme(plot.title = element_text(hjust = 0.5))

//model selection
beauty <- beauty%>%select(-gender, -health)
N = beauty %>% nrow
beauty = beauty%>%mutate(train=runif(N)<.5)
beautyTrain = beauty%>%filter(train == TRUE)
beautyTest = beauty%>%filter(train == FALSE)

beautyTrain <- beautyTrain%>%select(-train)
m_beauty = lm(exper~., beautyTrain)
summary(m_beauty)
stepMains = step(m_beauty, direction="both")
m_beauty = lm(exper~wage + goodhlth + female + married + service + educ + looks, beautyTrain)

//diagnostic plot
plot(m_beauty)

//predict experience for a person
newPerson <-data.frame(wage = 20.5, goodhlth = 1, female = 0, married   = 1, service = 0, educ = 12, looks = 3)
predict(m_beauty, newPerson, interval = "prediction", level = 0.95)

//confidence interval
beautyTest <- beautyTest%>%select(-train)
m_beautyTest = lm(exper~wage + goodhlth + female + married + service + educ + looks, beautyTest)
confint(m_beautyTest)
```