Data Exploration and Simple Regression Models

Due Date: September 13, 2024 at 11:59 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Problem 1

The dataset teengamb concerns a study of teenage gambling in Britain. You can download this data set by installing faraway library. To get the data set, copy and paste the r command: install.packages("faraway");library(faraway); data(teengamb, package="faraway"). (40 points)

The list variables are described below:

sex:0=male, 1=female

status: Socioeconomic status score based on parents' occupation

income: in pounds per week

verbal: verbal score in words out of 12 correctly defined

gamble: expenditure on gambling in pounds per year

a-) We are interested in predicting the expenditure on gambling. What is the depdent variable? and What are the independent variables? (10 points)

The dependent variable is `gamble` because we are predicting gambling expenditure, while `sex`, `status`, `income`, and `verbal` are independent variables as they influence the prediction.

b-) Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data. (30 points)

```
# load libraries
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.4.1
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.4.1
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.4.1
```

```
# load data
data(teengamb)

# rename sex from 0=male, 1=female
teengamb$sex <- factor(teengamb$sex, levels = c(0, 1), labels = c("male", "female"))

# numerical summary using psych package
summary_data <- describe(teengamb)

# summary using knitr
kable(summary_data, digits = 2, caption = "Summary of the teengamb dataset")
```

Summary of the teengamb dataset

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
sex*	1	47	1.40	0.50	1.00	1.38	0.00	1.0	2	1.0	0.38	-1.90	0.07
status	2	47	45.23	17.26	43.00	45.28	22.24	18.0	75	57.0	0.10	-1.31	2.52
income	3	47	4.64	3.55	3.25	4.14	2.45	0.6	15	14.4	1.33	1.10	0.52
verbal	4	47	6.66	1.86	7.00	6.79	1.48	1.0	10	9.0	-0.79	0.69	0.27
gamble	5	47	19.30	31.52	6.00	12.90	8.75	0.0	156	156.0	2.35	5.97	4.60

Comments on Numerical Summary:

Sex: The mean of 1.40 indicates a higher representation of males, with a slight skew (0.38) toward male. The negative kurtosis (-1.90) suggests a flat distribution with fewer extremes, meaning sex distribution is roughly balanced and won't strongly affect gambling trends by gender.

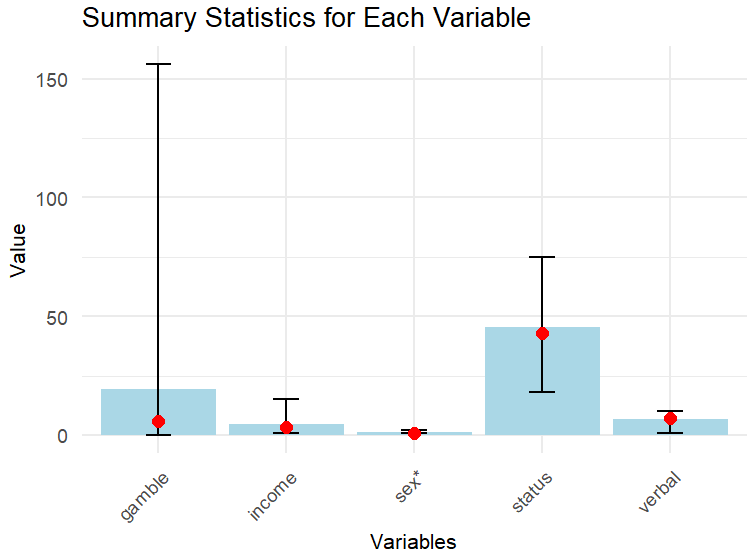
Status: The mean (45.23) and median (43.00) suggest a relatively symmetric distribution. The light-tailed distribution (kurtosis = -1.31) reflects moderate variability in socioeconomic status, implying no extreme concentration.

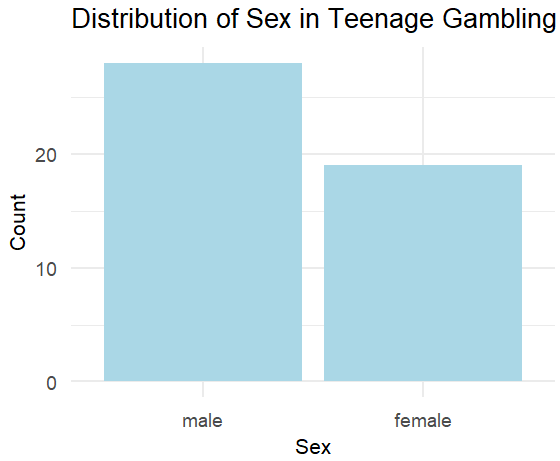
Income: The mean income (£4.64) is higher than the median (£3.25), indicating right skew (1.33), with a few participants earning significantly more. This means income disparity could play a role in predicting gambling expenditure.

Verbal: A near-symmetric distribution is shown by the mean (6.66) and median (7.00), with a slight left skew (-0.79). Verbal ability is generally high among participants and is unlikely to have a strong correlation with gambling behavior.

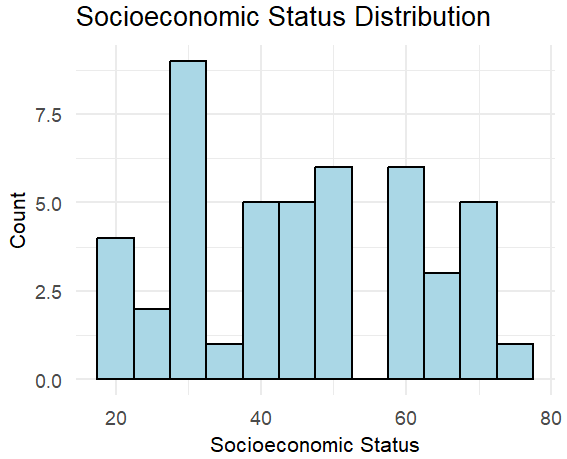
Gamble: Gambling expenditure shows high variability (SD = £31.52) with a strong right skew (2.35), indicating a few heavy gamblers. The high kurtosis (5.97) confirms significant outliers, meaning most gamble little, but a small group spends disproportionately more, heavily influencing the dataset.

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

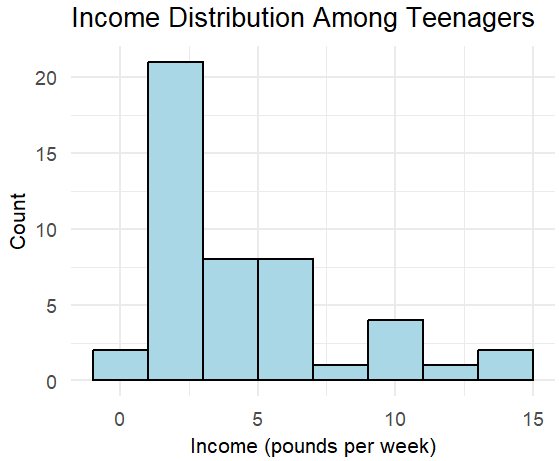




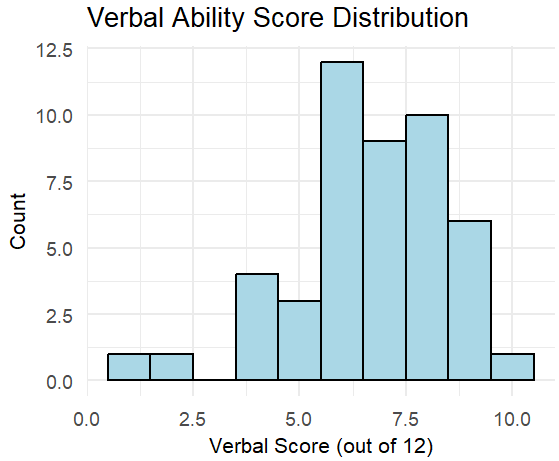
Takeaway: The sex distribution is slightly male-dominated, providing a relatively balanced gender comparison.



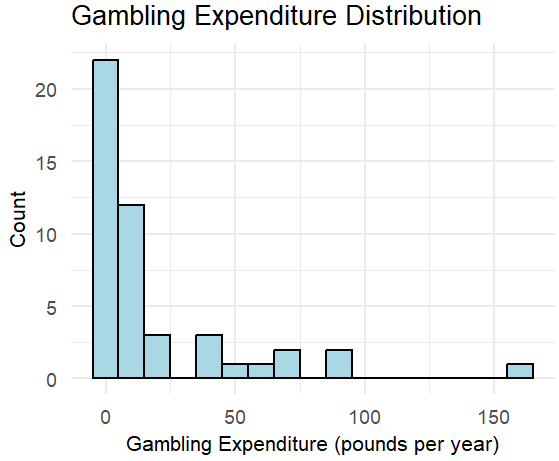
Takeaway: The distribution of socioeconomic status is relatively symmetrical, with most participants clustered around mid-level status.



Takeaway: Most participants earn between £0 and £5 per week, with a few higher-income outliers pulling the mean upward.



Takeaway: The verbal scores are skewed slightly to the higher end, suggesting strong verbal skills in most participants.



Takeaway: Most participants spend very little on gambling, but a few outliers spend disproportionately large amounts.

Problem 2

The dataset uswages is drawn as a sample from the Current Population Survey in 1988. You can download this data set by installing faraway library. To get the data set, copy and paste the r command: install.packages("faraway"); data(uswages, package="faraway"). (60 points, 10 points each)

The wage is the response variable. Please see below for the full list of variables.

wage: Real weekly wages in dollars (deflated by personal consumption expenditures - 1992 base year)

educ:Years of education

exper:Years of experience

race:1 if Black, 0 if White (other races not in sample)

smsa:1 if living in Standard Metropolitan Statistical Area, 0 if not

ne:1 if living in the North East

mw:1 if living in the Midwest

we:1 if living in the West

so:1 if living in the South

pt:1 if working part time, 0 if not

```
library(faraway)

# load 'uswages' dataset
data(uswages, package = "faraway")

# display first few rows
knitr::kable(head(uswages), caption = "First Few Rows of the USWages Dataset")
```

First Few Rows of the USWages Dataset

	wage	educ	exper	race	smsa	ne	mw	so	we	pt
6085	771.60	18	18	0	1	1	0	0	0	0
23701	617.28	15	20	0	1	0	0	0	1	0
16208	957.83	16	9	0	1	0	0	1	0	0
2720	617.28	12	24	0	1	1	0	0	0	0
9723	902.18	14	12	0	1	0	1	0	0	0
22239	299.15	12	33	0	1	0	0	0	1	0

a-) How many observations are in the data set?
There are 2000 observations in the dataset.

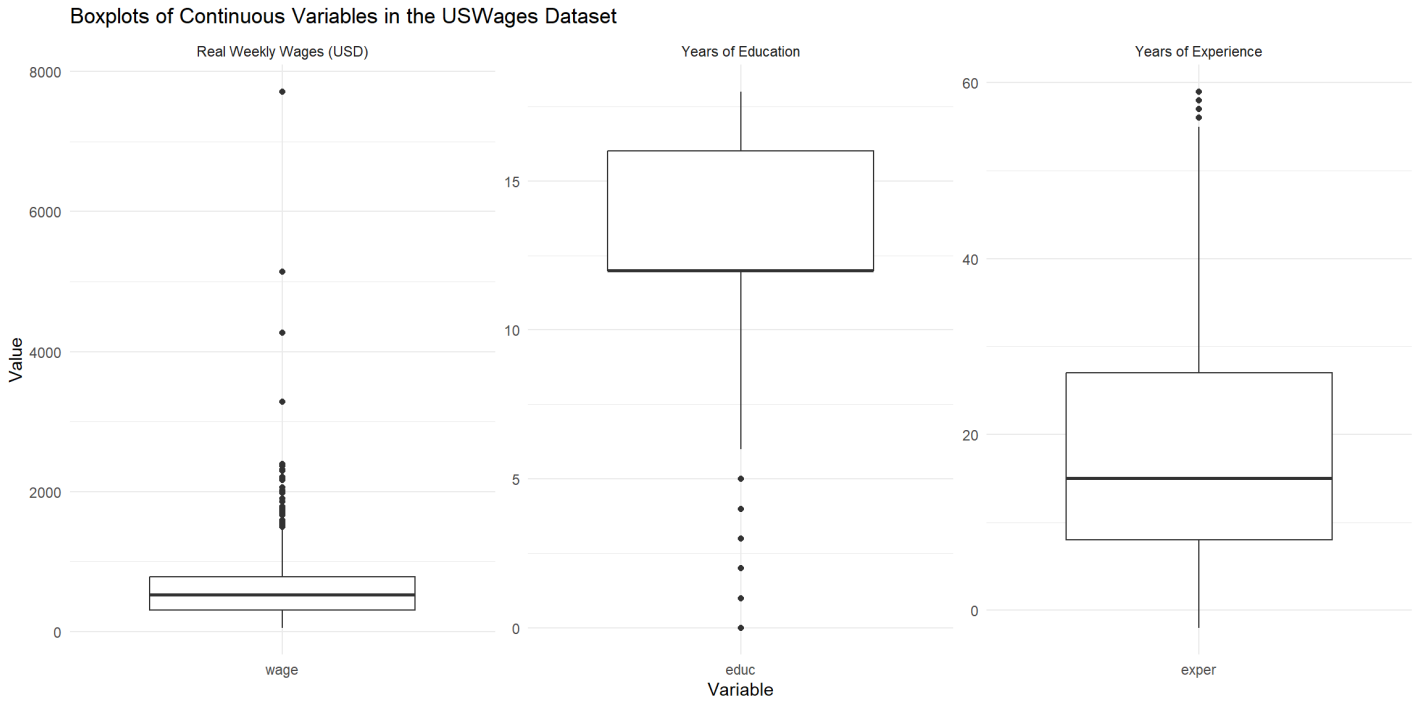
b-) Calculate the mean and median of each variable. Are there any outliers in the data set?
The mean and median of each variable are:

Mean and Median of Each Variable in the USWages Dataset

Variable	Mean	Median
wage	608.1179	522.32
educ	13.1110	12.00
exper	18.4105	15.00
race	0.0780	0.00
smsa	0.7560	1.00
ne	0.2290	0.00
mw	0.2485	0.00
so	0.3125	0.00
we	0.2100	0.00
pt	0.0925	0.00

To understand if variables have outliers, use boxplots.

```
## Warning: package 'reshape2' was built under R version 4.4.1
```



1. Real Weekly Wages (USD) There are several significant outliers in the wage variable, with some earning much higher wages than the rest of the sample.
The majority of the data is concentrated below $2000, indicating that most individuals have relatively moderate weekly wages compared to the outliers.

2. Years of Education There are a few outliers with lower education levels. The distribution of education shows most individuals having between 10 to 16 years of education. These outliers do not significantly skew the dataset.

3. Years of Experience The boxplot for experience shows some outliers on the upper end, representing individuals with exceptionally high experience, possibly indicating older or highly experienced workers. The majority of workers have 10 to 30 years of experience.

c-) Calculate the correlation among wage,education and experience. Plot each of the predictors against the response variable. Identify the variables that are strongly correlated with the response variable.

```
library(faraway)

# load 'uswages' dataset
data(uswages, package = "faraway")

# wage, education, experience
uswages_selected <- uswages[, c("wage", "educ", "exper")]

# correlation matrix
cor_matrix <- cor(uswages_selected, use = "complete.obs")

# display
knitr::kable(cor_matrix, caption = "Correlation Matrix between Wage, Education, and Experience")
```

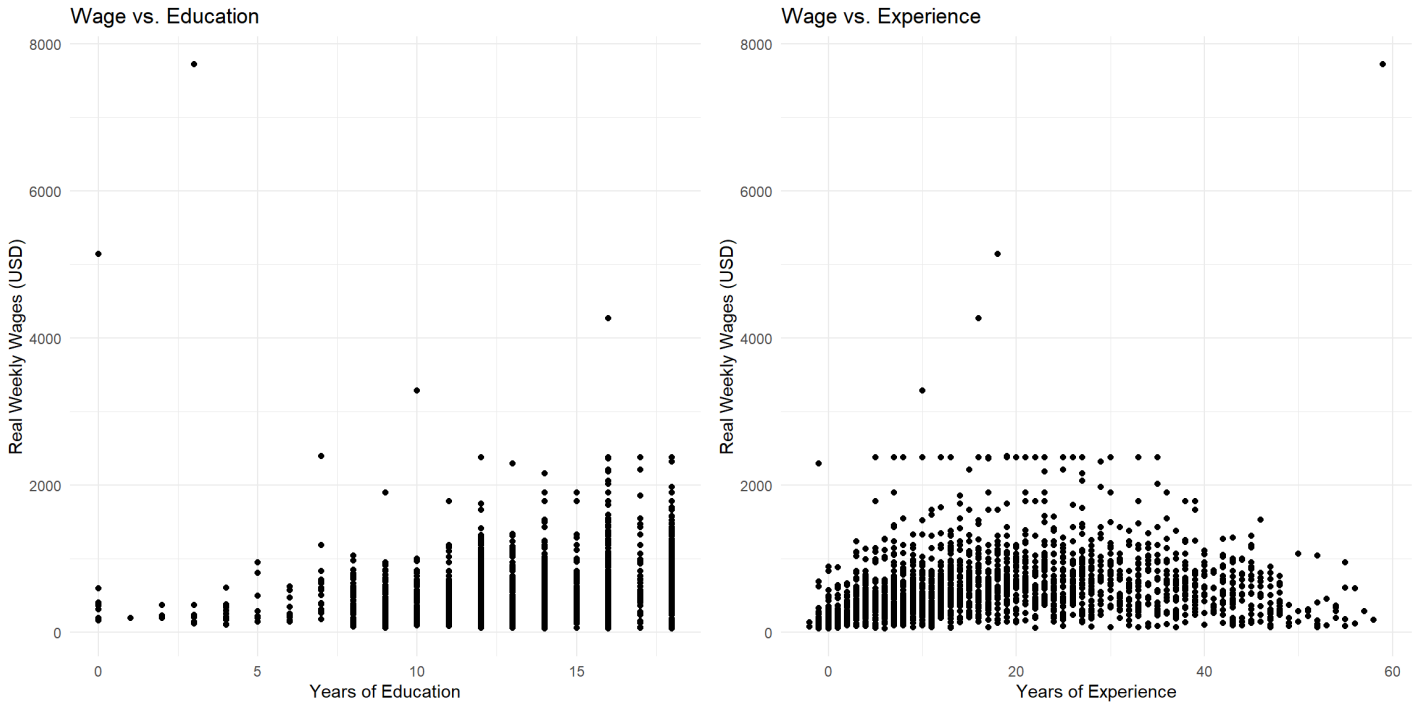
Correlation Matrix between Wage, Education, and Experience

	wage	educ	exper
wage	1.0000000	0.2483358	0.1832012
educ	0.2483358	1.0000000	-0.3024788
exper	0.1832012	-0.3024788	1.0000000

Wage and Education (0.2483): Weak positive correlation. Higher education slightly increases wages.
Wage and Experience (0.1832): Weak positive correlation. More experience has minimal impact on wages.
Education and Experience (-0.3025): Moderate negative correlation. More education generally means less work experience.

Overall, both education and experience have weak effects on wage.

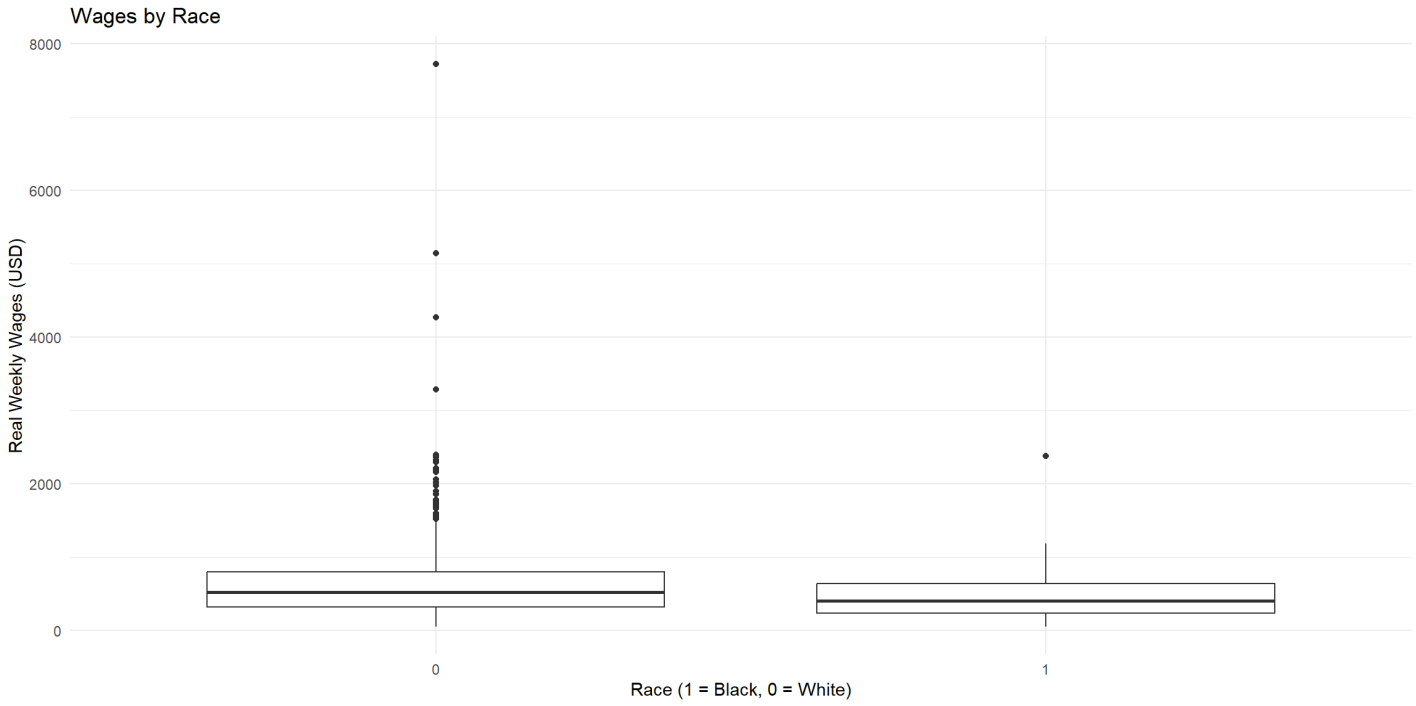
```
## Warning: package 'gridExtra' was built under R version 4.4.1
```



Wage vs. Education: Weak positive trend, wages increase slightly with education. Large variance at all education levels with notable outliers.

Wage vs. Experience: Wages rise with experience until around 20-30 years, then plateau or decline. Significant wage spread and some outliers.

d-) Is there difference in wages based on race?



From the boxplot:

The median wages for both Black (1) and White (0) individuals appear quite similar in the boxplot. The range of wages is larger for White individuals, with many outliers on the higher end (greater than $2000). There is less spread for wages among Black individuals, with fewer outliers compared to the White group. To statistically confirm this, we check the t-test result.

```
t_test_result <- t.test(wage ~ race, data = uswages)

t_test_result$p.value
```

```
## [1] 4.095868e-09
```

The t-test p-value is 4.0958683^{-9}. There is a statistically significant difference in wages between Black (1) and White (0) individuals in this dataset. This suggests that race plays a role in wage disparity, with the observed difference unlikely to be due to chance.

e-) Build a regression model by using only education to predict the response variable. State the regression model.

```
# regression model using education to predict wage
education_model <- lm(wage ~ educ, data = uswages)

# coefficients
intercept <- coef(education_model)[1]
slope <- coef(education_model)[2]
```

The regression model is:

$$\hat{wage} = 109.75 + 38.01 \times education$$

f-) Build a regression model by using only experience to predict the response variable. State the regression model.

```
# regression model using experience to predict wage
experience_model <- lm(wage ~ exper, data = uswages)

# coefficients
intercept_exper <- coef(experience_model)[1]
slope_exper <- coef(experience_model)[2]
```

The regression model is:

$$\hat{wage} = 492.17 + 6.3 \times experience$$