Assignment 1 - MDA9159

Instructor: Dr. Guowen Huang

Fall 2025

-Student name: Rui Deng

-Student number: 251509628

Question 1

(a) Compute the mean and SD of salary

```
data=read.table("p130.txt",header=TRUE)
mean(data$S)
```

```
## [1] 17270.2
```

```
sd(data$S)
```

```
## [1] 4716.632
```

Therefore, the mean of salary is 17270.2 and the standard deviation is 4716.632.

(b) Compute mean salary by education level

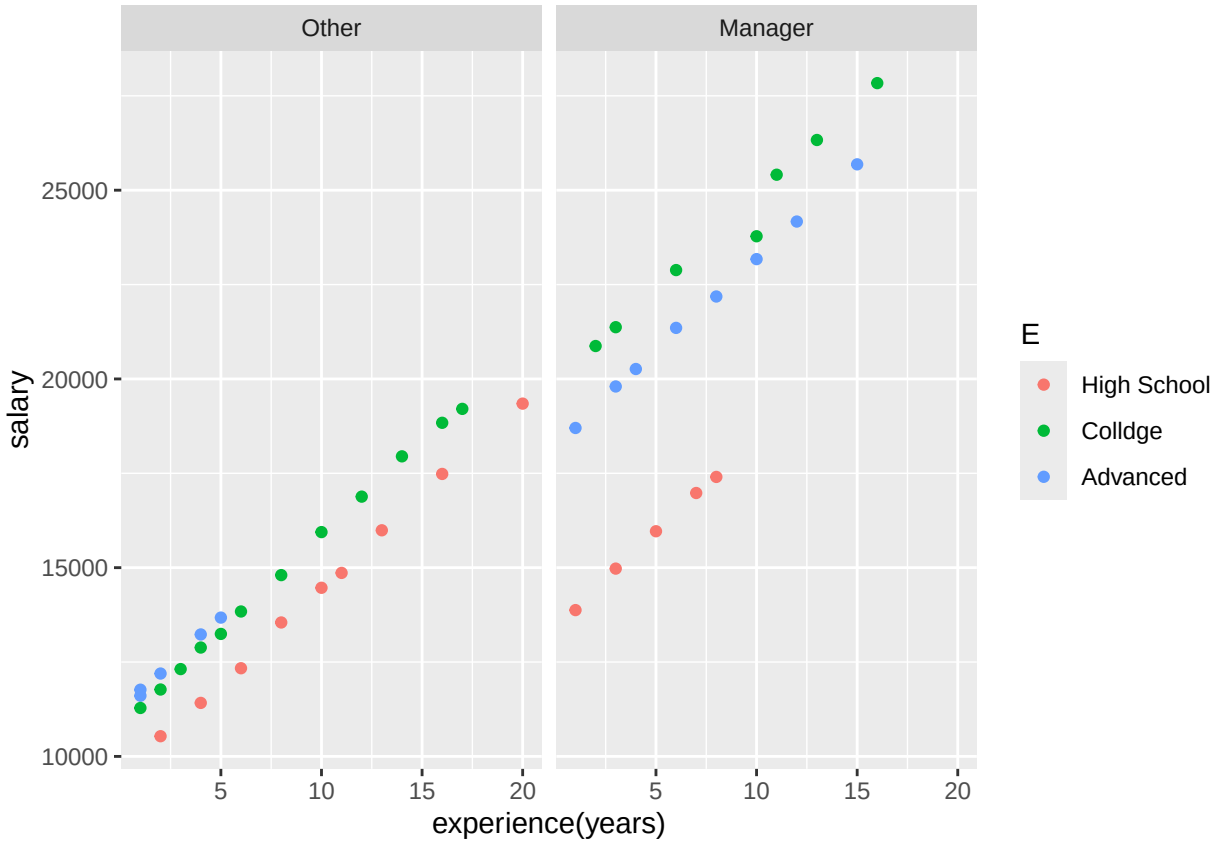
```
aggregate(S~E, data=data, mean)
```

```
##   E        S
## 1 1 14941.50
## 2 2 18286.37
## 3 3 18292.85
```

Therefore, the mean of salary of education level1 is 14941.5; the mean of salary of educaation level2 is 18286.37; the mean of salary of education level3 is 18292.85. From this, we can see that the higher education level is the higher salary might be.

(c) Draw a scatterplot of salary vs. experience, colored by education, faceted by management. Interpret the plot.

```
data$E=factor(data$E, labels=c("High School","Colldge","Advanced"))
data$M=factor(data$M,labels=c("Other","Manager"))
library(ggplot2)
ggplot(data, aes(x=X, y=S, col=E)) +
  geom_point() +
  facet_grid(~M) +
  xlab("experience(years)") +
  ylab("salary")
```



From the picture, we can see that for the same education level as years of experience increase, salary also increases. With the same year of experience, the amount of salary is related to education level, but the highest is for education level2.

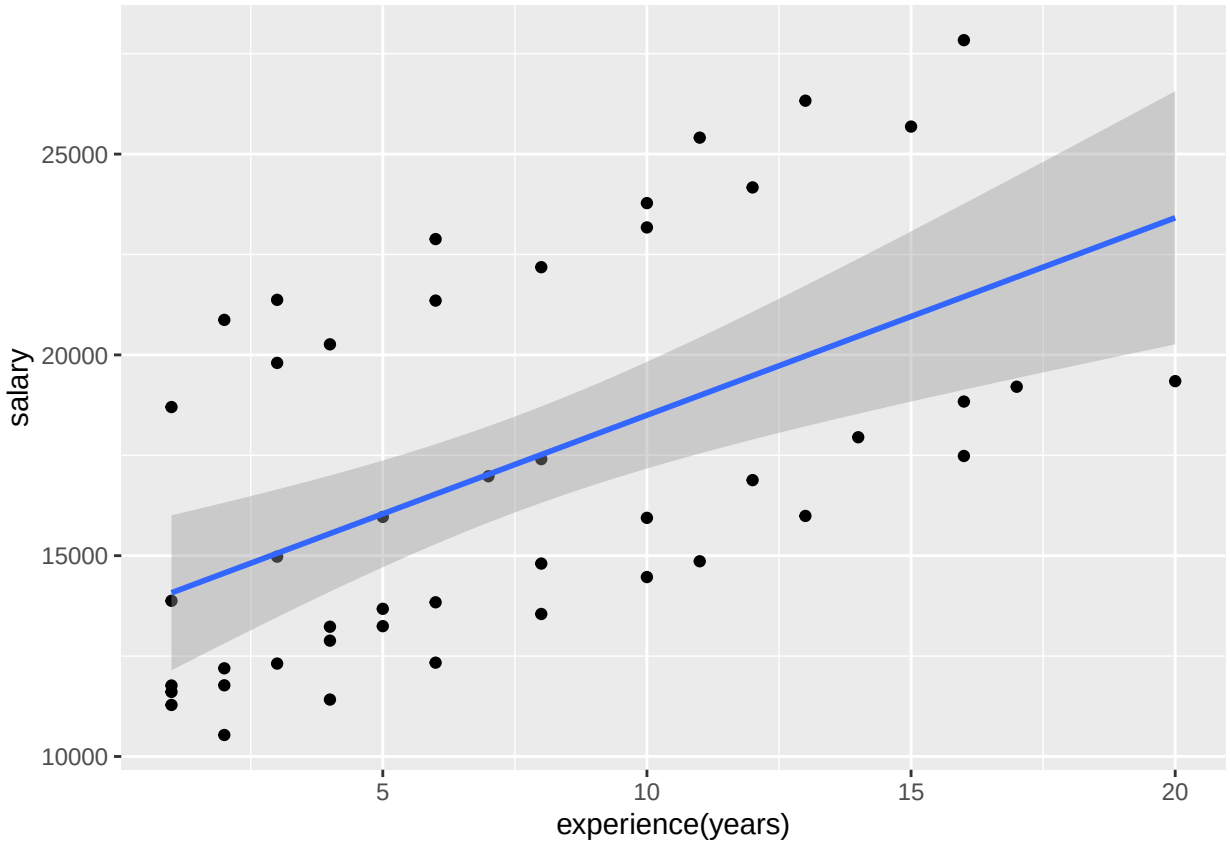
Question 2

Fit a regression of salary on experience: S~X.

(a) Report the fitted regression line.

```
ggplot(data, aes(x=X, y=S)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlab("experience(years)") +
  ylab("salary")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
lmdata=lm(S~X, data=data)
summary(lmdata)
```

```
##
## Call:
## lm(formula = S ~ X, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
##  -4197  -2781  -2368   4685   6420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13584.0     1051.5  12.919  < 2e-16 ***
## X              491.5      115.8   4.243 0.000112 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4018 on 44 degrees of freedom
## Multiple R-squared:  0.2904, Adjusted R-squared:  0.2743
## F-statistic: 18.01 on 1 and 44 DF,  p-value: 0.0001117
```

(b) Interpret the slope.

From the output, we see that $\hat{S}$=491.5*X+1358.4. The intercept 491.5 means there is a positive relationship between salary and experience. When experence increases 1, salary will increase $491.5.

(c) What percent of salary variation is explained by experience?

As adjusted R^2=0.2743, only 27% of salary variation is explained by experience. This is probably because experience alone explains only a small proportion of the variation in salary.This suggests that other factors may also play important roles in determining salary levels.

Question 3

Fit the model S~X + Education (treat Education as a factor).

(a) Write the regression equation using dummy variables (reference = High School).

```
lmmulti=lm(S~X+E, data=data)
summary(lmmulti)
```

```
##
## Call:
## lm(formula = S ~ X + E, data = data)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4320  -3182  -1372   2812   6079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10474.3     1305.4   8.024 5.19e-10 ***
## X              548.6      107.6   5.100 7.69e-06 ***
## EColldge      3221.1     1275.8   2.525  0.01544 *
## EAdvanced     4780.1     1422.7   3.360  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3622 on 42 degrees of freedom
## Multiple R-squared:  0.4498, Adjusted R-squared:  0.4104
## F-statistic: 11.44 on 3 and 42 DF,  p-value: 1.291e-05
```

From the result we can see that adjusted R^2=0.4104, it is better than the model in question 2. But it's still low, the next step to improve the model will be to add in the factor of job position.

(b) Interpret the coefficient for College.

E2 means the group of college. The coefficient of E2 is 3221.1, which means When X remains unchanged, the average salary of those with a College education is $3,221.10 higher than that of the High School group.

(c) Conduct an overall F-test for whether Education matters (i.e., all education-level effects = 0).

Model: S=β_0+β_1D+ β_1EColldge+ β_2EAdvanced

H_0: β_1=β_2=0

H_1: one of β_1, β_2 is not 0

```
anova_result=anova(lmdata, lmmulti)
anova_result
```

```
## Analysis of Variance Table
##
## Model 1: S ~ X
## Model 2: S ~ X + E
##   Res.Df       RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1     44 710380856
## 2     42 550853135  2 159527722 6.0816 0.004791 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result, we can see F=6.0816 and p value=0.00479<0.5, which means we reject the null hypothesis. So education has a statistically significant overall effect on salary after controlling for experience.

Question 4

At 10 years of experience, predict mean salary and compute:

(a) a 95% confidence interval for the mean salary

```
conf_interval=predict(lmmulti, newdata=data[data$X==10,],
                      interval="confidence",level=0.95)
conf_interval
```

```
##         fit      lwr      upr
## 30 15960.34 13965.87 17954.81
## 31 19181.47 17467.76 20895.17
## 32 20740.48 18493.93 22987.03
## 33 19181.47 17467.76 20895.17
```

(b) a 95% prediction interval for a single individual

```
pred_interval=predict(lmmulti,newdata=data[data$X==10,],
                      interval="prediction",level=0.95)
pred_interval
```

```
##         fit       lwr      upr
## 30 15960.34  8384.526 23536.16
## 31 19181.47 11674.676 26688.25
## 32 20740.48 13094.431 28386.53
## 33 19181.47 11674.676 26688.25
```

Question 5

(a) Why do we drop one indicator when coding education with dummies?

1. To avoid perfect multicollinearity. If all dummies are included, they would sum to 1 for every observation. So one dummy is exactly predictable from the others, creating perfect collinearity.
2. By dropping one dummy, the intercept can represent the mean outcome for the dropped group.

(b) Explain (in words) why a prediction interval is wider than a confidence interval

The prediction interval is wider because predicting one new observation includes both the uncertainty of estimating the mean and the natural variability of individual outcomes around that mean.