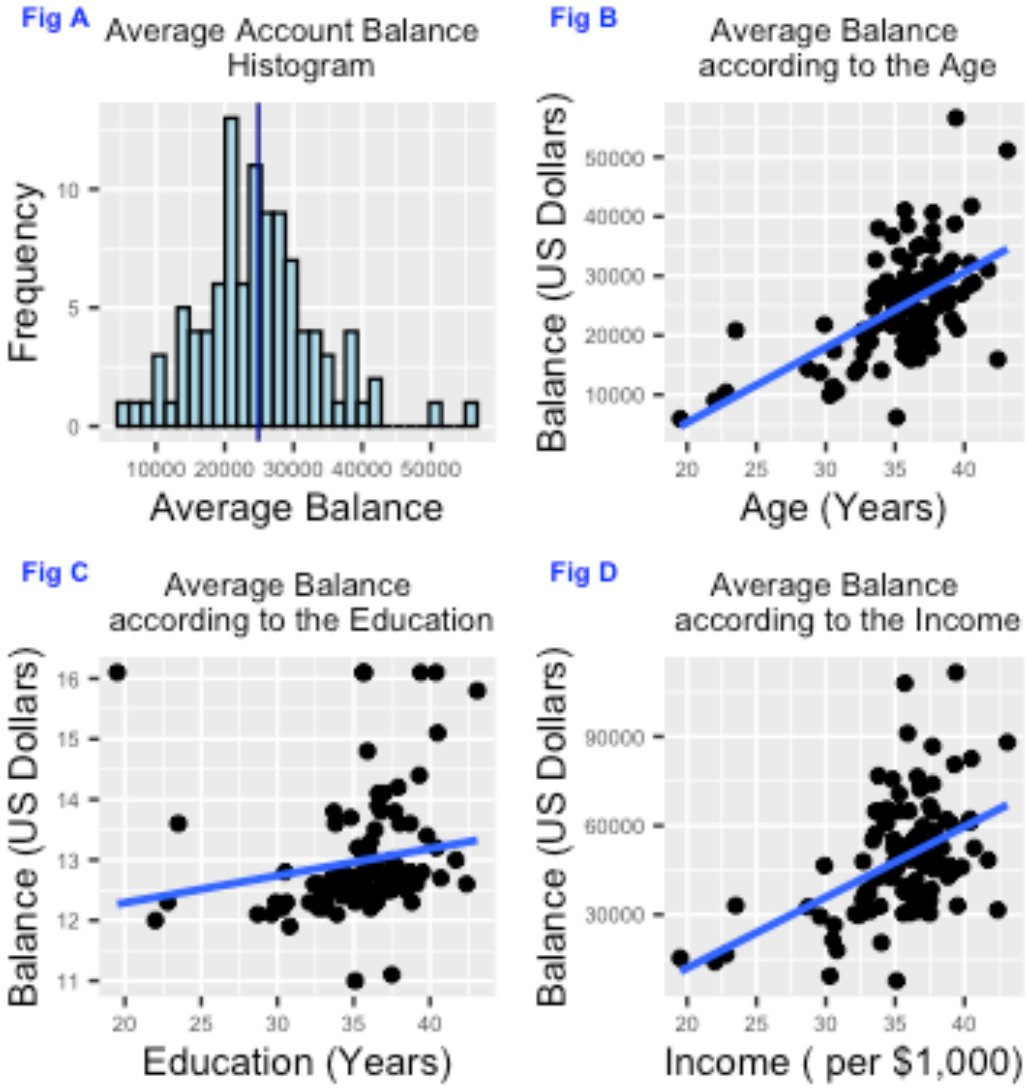
Erik Pak
DSC 423
Assignment 01

Problem 01

A: Analyze the distribution of average account balance using histogram, and compute appropriate descriptive statistics. Write a paragraph describing distribution of Balance and use appropriate descriptive statistics to describe center and spread of the distribution.



All Plots

A. The average balance histogram is bell-shape normal distribution with slight positive skewness. The median represents the exact middle of the dataset, and the arithmetic average represents the mean. These two values are used to measure the central tendency of the data set. Since the mean is slightly greater than the median, this also confirms positive skewness, and the left side of the mean would have a larger amount of data. The standard deviation ($8,697.81) represents the spread or desperation from the mean, such that, on average, each point would be within $8,697.81 from the mean. Also the range (50,613) represents the spread from the lowest to the highest in the dataset. The 1^{st} and 3^{rd} quartiles represent 25% and 75% of our dataset's data, and the median is right at 50% of the dataset. (Fig A)

B. Age appears to have a linear relationship with the average bank balance, which makes sense because people generally earn more, and their average balance should increase as they age. Nevertheless, I consider three or four outliers far beyond the rest of the data; two or one on the right and the other two lower right, away from the rest. (Fig B)

C. Education level has a linear relationship with the average bank balance, which makes sense since you would likely have more significant earning potential. Therefore, Two or three outliers are all located at the lower right. I would verify the lowest two points on the right to make sure the data is valid. (Fig C)

D. The income level has a great linear relationship with the average balance and more money you make you mostly like would minion a higher back balance. There are couple of points that stray bit aways from the plot and we could conceivably state those point are a outliers.(Fig D)

Descriptive Statistics:

Number of Oversevations	Mean	Standard Deviation	Median	Min	Max	Range
102.00	24,887.88	8,697.81	24,660.50	5,956	56,569	50,613

The Mean is the center of the distribution, the standard deviation is a measure of how dispersed the data is about the Mean, the Min is the smallest value, and the Max is the largest value in the data. The range represents Max - Min. Since the Mean is slightly greater than the Median histogram skews to the right, you can see it in Fig A.

C: Correlation

	Age	Education	Income	Balance
Age	1.0000000	0.1734071	0.4771474	0.5654668
Education	0.1734071	1.0000000	0.5753940	0.5548807
Income	0.4771474	0.5753940	1.0000000	0.9516845
Balance	0.5654668	0.5548807	0.9516845	1.0000000

Age vs. Education:

- there is a positive weak linear relationship

Age vs. Income

- there is a positive moderate linear relationship

Age vs. Balance

- there is a positive moderate linear relationship

Education vs Income

- there is a positive moderate linear relationship

Education vs Balance

- there is a positive moderate linear relationship

Income vs Balance

- strong positive linear relationship

The highest correlation (0.9516845) is highly related to one's income. The higher the income, one is most likely to have the larger the average account balance therefore there is a very strong relationship between income and balance.

D: In this dataset, we have three independent variables (Age, Education, Income) with one dependent variable (Balance). Right side equation containing β's are all independent variables and on the left side is the dependent variable which is the balance.

$$y_{balance} = \beta_0 + \beta_{age} + \beta_{education} + \beta_{income} + \epsilon$$

E: After running the model, the equation below shows that income significantly impacts the balance because a smaller positive coefficient is added to the negative intercept. Also, when standardization of coefficients using lm.beta() function of QuantPsyc

package for consistency measurement of units, it is clear that income significantly impacts the average balance due to the higher coefficient value.

$$\hat{y}_{Balance} = -9540 + 332.5\ (Age) + 288.7\ (Education) + 0.3871\ (Income)$$

Standardized coefficients:

Age	Education	Income
0.14857228	0.03335784	0.86159971

According to the summary from the model, the t value of income is 22.14, and the corresponding p-vale is 2e-16 which is significantly less than $\alpha = 0.05$.Also, the higher the t-value, the greater the confidence we have in the coefficient as a predictor. In conclusion, income has most significant effect on balance. All the t-values from the summary except for education are less than the $\alpha = 0.05$. Therefore education is only value greaten than α which indicate of low reliability of the predictive power of that coefficient.

```
Call:
lm(formula = Balance ~ Age + Education + Income, data = bank)

Residuals:
    Min      1Q  Median      3Q     Max
-7722.0 -1547.4   -56.1  1167.9  8480.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.540e+03  4.423e+03  -2.157   0.0335 *
Age          3.325e+02  7.234e+01   4.597 1.28e-05 ***
Education    2.887e+02  3.005e+02   0.960   0.3392
Income       3.871e-01  1.748e-02  22.137  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2458 on 98 degrees of freedom
Multiple R-squared:  0.9225,    Adjusted R-squared:  0.9201
F-statistic: 388.8 on 3 and 98 DF,  p-value: < 2.2e-16
```

Regression Model

F: I am choosing to remove Education because the t-value from the summary of the model the Pr(> |t|) is 0.3392, which is significantly greater than the p-value of 0.05 and education value has low reliability of the predictive power of that coefficient.

New expression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\hat{y}_{balance} = -5912 + 322.7\ (Age) + 0.3966\ (Income)$$

```
Call:
lm(formula = Balance ~ Age + Income, data = bank)

Residuals:
    Min      1Q  Median      3Q     Max
-7385.1 -1577.9  -119.2  1200.6  8362.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.912e+03  2.301e+03  -2.570   0.0117 *
Age          3.227e+02  7.159e+01   4.508  1.8e-05 ***
Income       3.966e-01  1.437e-02  27.600  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2457 on 99 degrees of freedom
Multiple R-squared:  0.9218,    Adjusted R-squared:  0.9202
F-statistic: 583.2 on 2 and 99 DF,  p-value: < 2.2e-16
```

New Linear Regression Model

G:

New model:

$$y_{balance} = \beta_0 + \beta_1 x_{age} + \beta_2 x_{income} + \epsilon$$

$$\hat{y}_{balance} = -5912 + 322.7\ (age) + 0.3966\ (Income)$$

- β_{age} : we estimate the mean balance E(y) to increase $322.70 for every one-year increase in age if β_{income} is held constant
- β_{income} : we estimate the mean balance E(y) to increase 0.40 cents for every dollar in income if β_{age} is held constant

H. The R^2 coefficient is how well it explains the variation in Y in our linear regression model. For example, our second model is 0.9218, which means that the model can explain 92.18% of the variability in our data.

I: Given:

- Median Age = 34.8 & Median Education = 12.5 & Median Income = $42,401 & Average balance = $21,572
  - First Model:
    - $\hat{y}_{Balance} = -9540 + 332.5\ (Age) + 288.7\ (Education) + 0.3871\ (Income)$
      - = -9540 + 332.5*(34.8) + 288.7*(12.5) + 0.3871*(42401) = **22,050.60**
  - Models predicted error: $e = y - \hat{y}$
    - e = $21,572 - $22,050.60 = **-478.60**
  - Final Model:
    - $\hat{y}_{balance} = -5912 + 322.7\ (age) + 0.3966\ (Income)$
    - = -5912 + 322.7*(34.8) + 0.3966*(42401) = **22,135.22**
    - The models prediction error would be: $e = y - \hat{y}$
      - ***e*** = $21,572 - $22,134.20 = **-563.22**

J: Second Model

- $H_0 : \beta_1 = \beta_2 = 0$
- H_a : at least one $\beta_i \neq 0$

F-statistic: 583.2 on 2 and 99 DF, p-value: < 2.2e-16

Since $\alpha = 0.05$ exceeds the observed significance level $\rho = 2.2e - 16$, the data provide very strong evidence that at least one of the model coefficients is non zero. Therefore, there is enough evidence to claim that there is a strong evidence Null Hypothesis is rejected.

Problem 02

A university career center collects information on the job status and starting salary of graduating seniors. Data recently collected over a two-year period included over 900 seniors who had found employment at the time of graduation. The information was used to model starting salary Y as a function of two qualitative independent variables: COLLEGE at four levels {Business, Engineering, Liberal Arts, Nursing} and SEX (male and female).

1. Define the dummy variables to include college (use Business as your baseline) in a regression model for starting salary Y

- Character variables:
- COLLEGE - college major
  - value - Business, Engineering, Liberal Arts, Nursing
  - k - 1 = 3 variables $z_1, z_2,$ and z_3 as follows:
  - $Z_1 = 1$ if X = Engineering; $Z_1 = 0$ Otherwise;
  - $Z_2 = 1$ if X = Liberal Arts; $Z_2 = 0$ Otherwise;
  - $Z_3 = 1$ if X = Nursing; $Z_3 = 0$ Otherwise;

  - The regression model of Y on X is
    - $y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_2 Z_3 + \epsilon$

Table below:

	Engineering	Liberal Arts	Nursing
Engineering	1	0	0
Liberal Arts	0	1	0
Nursing	0	0	1
Business	0	0	0

	Male
Male	1
Female	0

2. Define the dummy variables to include college (use Business as your baseline) in a regression model for starting salary Y

- Major
  - value - Business, Engineering, Liberal Arts, Nursing
  - k - 1 = 3 variables $z_1, z_2,$ and z_3 as follows:
  - $Z_1 = 1$ if X = Engineering; $Z_1 = 0$ Otherwise;
  - $Z_2 = 1$ if X = Liberal Arts; $Z_2 = 0$ Otherwise;
  - $Z_3 = 1$ if X = Nursing; $Z_3 = 0$ Otherwise;
- Sex
  - $Z_4 = 1$ if X = Male; $Z_4 = 0$ Otherwise;

$$y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \epsilon$$

3. The new regression model with Engineering having the save starting salary as students in Business:

$$y = \beta_0 + \beta_1 Z_2 + \beta_2 Z_3 + \epsilon$$

Problem 03

1) Analyze the interrelationships between variables using scatterplots and correlation values. Is salary linearly related to the three predictors? Which variables are more strongly related?

Fig A

Fig B

Fig C

All Scatter Plot

Fig **A.** There is a slight linear relationship with one noticeable outlier lying lower right of the graph and a very insignificant positive correlation between these two variables.

Fig **B.** There is a much better linear relationship than the number of employees. There could be multiple outliers with this scatter plot. I notice two noticeable outliers in the lower and upper parts of the graph and have a low correlation between these two variables.

Fig C. This scatter plot has a similar linear relationship as the firm's gross profit margin. However, this plot may show one outlier lower part of the plot. This plot's correlation is slightly better than the above plot but still has a low correlation between the variables.

The salary is somewhat linearly related to the three predictors. Consequently, I conclude that total admin cost and the firm's gross profit margin are more linearly significant than the number of employees.

Correlation Table:

	salary	numempl	margin	ipcost
salary	1.00000000	0.04267267	0.49884432	0.52975765
numempl	0.04267267	1.00000000	0.12577542	-0.09667573
margin	0.49884432	0.12577542	1.00000000	0.55409931
ipcost	0.52975765	-0.09667573	0.55409931	1.00000000

Salary vs. NumEmpl:

- there is a weak linear relationship

Salary vs. Margin

- there is a weak linear relationship

Salary vs. IPCost

- there is a weak linear relationship

NumEmpl vs Margin

- there is a weak linear relationship

NumEmpl vs IPCost

- there is a negative weak linear relationship

IPCost vs Margin

- There is a weak linear relationship

2) Write down the regression model to predict salary using the other three variables as predictors

$$y = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$\hat{y}_{salary} = 25.900686 + 0.001356 * (numempl) + 0.087484 * (margin) + 0.208865 * (ipcost)$$

3) Examine the $\bar{y}$ equation, Standardized Regression Coefficients, and especially the model summary, and it's evident that ipcost has the most significant effect on the model. When examining the summary of the model, you can see ipcost ρ-value is 0.00598, which is less than $\alpha = 0.05$, which states that it is significant including being the lowest value.

4)

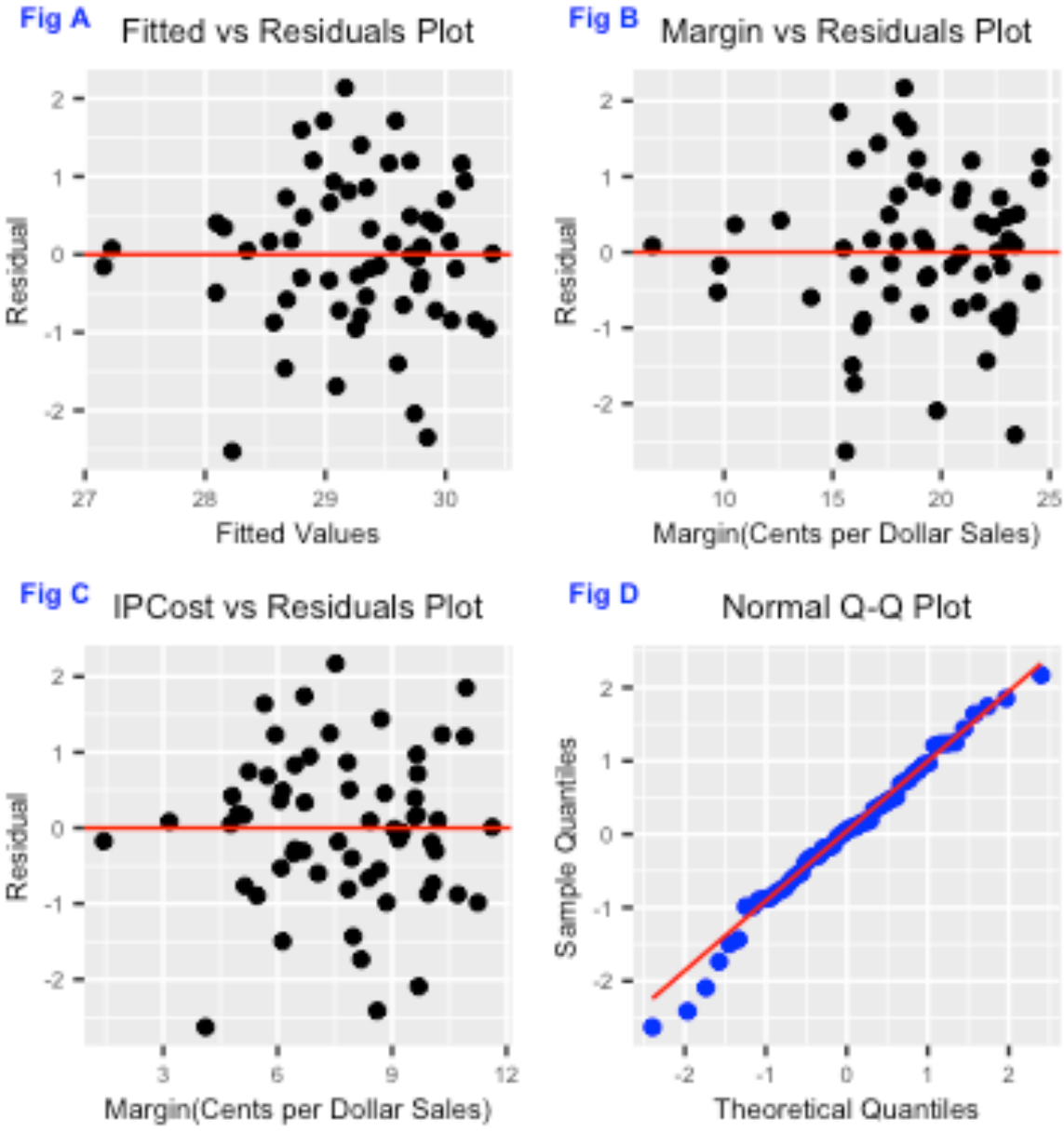
$$y = \beta_0 + \beta_1 x_{margin} + \beta_2 x_{ipcost} + \epsilon$$

$$\hat{y} = 25.97304 + 0.09092 x_{margin} + 0.20311 x_{ipcost}$$

- β_{margin} : we estimate the mean salary E(y) to increase $90.92 per dollar of sales if β_{ipcost} is held constant
- β_{ipcost} : we estimate the mean salary E(y) to increase $203.11 per administrative costs if β_{margin} is held constant

5) The R^2 coefficient of determination is how well it explains the variation in Y in our linear regression model. For example, our second model is 0.3415, which means that the model can explain 34.15% of the variability in our data.

6)



Goodness of Fit

$H_0 : \beta_{margin} = \beta_{ipcost} = 0$

H_a : At least one coefficient $\beta_j \neq 0$

According to the Analysis of Variance F-value of 15.04 with $\rho = 5.479e - 9$ being significantly less than $\alpha = 0.05$, age and income variables are statistically significant in the model according to ANOVA analysis. The plots displaying residuals with fitted, margin, and ipcost values are displayed with random plots are randomly displayed. The normal Q-Q plot is closely following a straight line and approximately at 45^o angle upwards, which tells me that my residuals are normally distributed. All these validations lead me to believe this model is viable according to the goodness of fit. Overall, this is poor performing model due to $R^2 = 0.3415$, and transformation can improve the performance of this model. For comparison purposes, R^2_{adj} should be used to compare the OLD & NEW model.

```
Call:
lm(formula = salary ~ margin + ipcost, data = sal)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5260 -0.5797  0.0490  0.6611  2.1359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.97304    0.65063  39.920  < 2e-16 ***
margin       0.09091    0.03928   2.315  0.02420 *
ipcost       0.20311    0.07111   2.856  0.00594 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9923 on 58 degrees of freedom
Multiple R-squared:  0.3415,    Adjusted R-squared:  0.3188
F-statistic: 15.04 on 2 and 58 DF,  p-value: 5.479e-06
```

New Salary Model

```
Analysis of Variance Table

Response: Balance
          Df     Sum Sq    Mean Sq F value    Pr(>F)
Age        1 2443180856 2443180856  404.61 < 2.2e-16 ***
Income     1 4599872720 4599872720  761.78 < 2.2e-16 ***
Residuals 99  597790568    6038289
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA

7) The IPCost has the more significant effect on the latest model according to the standardized coefficients of the newest model

Standardized Regression Coefficients (New Model)

margin	ipcost
0.2962680	0.3655958

Problem 04

Laughter is often called “the best medicine,” since studies have shown that laughter can reduce muscle tension and increase oxygenation of the blood. In the International Journal of Obesity (January 2007), researchers at Vanderbilt University investigated the physiological changes that accompany laughter. Ninety subjects (18–34 years old) watched film clips designed to evoke laughter. During the laughing period, the researchers measured the heart rate (beats per minute) of each subject with the following summary results: y_bar = 73.5, s = 6. It is well known that the mean resting heart rate of adults is 71 beats/minute. At α = .05, is there sufficient evidence to indicate that the true mean heart rate during laughter exceeds 71 beats/minute? Show your work.

*** Are there two possible solutions to this problem?**

A)

What problem is this?

Large sample $(n \geq 30)$ Test Hypothesis about μ

What data were we given?

$n = 90\ , \bar{y} = 73.5\ , s = 6\ , \mu_0 = 71$

What formula do we need to apply?

Test statistic: $z = (\bar{y} - \mu_0)/\sigma_{\bar{y}} \approx (\bar{y} - \mu_0)/s/\sqrt{n})$

$H_0 : \mu = 71$

$H_a : \mu > 71$

Rejection region: $z > z_\alpha = (1.96 > z_c)$

$\rho - value : P(z > z_c)$

= (73.5 - 71) / (6 / $\sqrt{90}$)

z_c = 3.9528

ρ_c = 1 - pnorm(z_c) = 3.8621e-05

Conclusion: The p-value is ≈ 0.000, therefore we conclude that the null hypothesis is rejected at 5% level. There is enough evidence to claim that that the true mean heart rate during laughter exceeds 71 beats/minute.

B)

What problem is this?

Large sample 100(1 - α)% Confidence Interval for μ because $n > 30$

What data were we given?

$n = 90$, $\bar{y} = 73.5$, $s = 6$, $\mu_0 = 71$

What formula do we need to apply?

$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}(s/\sqrt{n})$

The 95% C.I for the beats per minute computation

= 73.5 ± 1.96 (6 / $\sqrt{}$ 90)

= (72.26, 74.74)

Based one the 95% C.I., there is enough evidence to claim that that the true mean heart rate during laughter exceeds 71 beats/minute.