Case Study

Question 1 (10 marks)

Generate the square-root and log transformation for MonthlyIncome, namely sqrtInc and logInc respectively. Among the three variables, select the one that is most suitable for analysis of variance. Denote the variable of your choice as tIncome.

A square root transformation and log transformation was done for the MonthlyIncome variable and the result was named as sqrtInc and logInc.

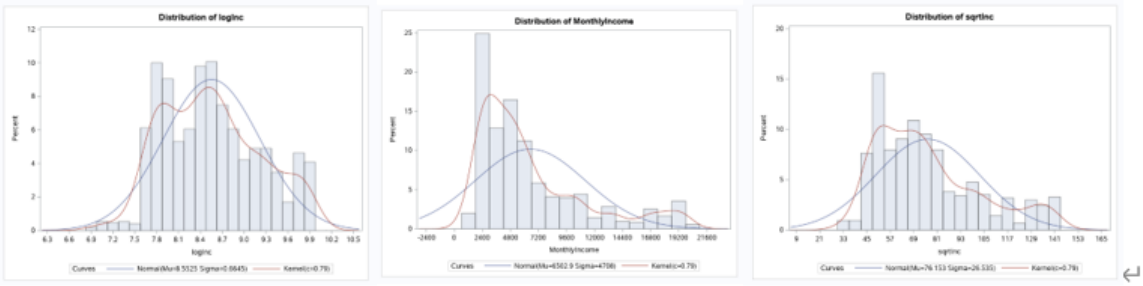
Variable	N	N Miss	Mean	Std Dev	Minimum	Maximum	Median	Lower Quartile	Upper Quartile	Skewness	Kurtosis	Coeff of Variation
MonthlyIncome	1470	0	6502.931	4707.957	1009.000	19999.000	4919.000	2911.000	8380.000	1.370	1.005	72.397
sqrtInc	1470	0	76.153	26.535	31.765	141.418	70.136	53.954	91.542	0.862	-0.113	34.844
logInc	1470	0	8.553	0.664	6.917	9.903	8.501	7.976	9.034	0.286	-0.696	7.769

The above table shows the results of descriptive statistical analysis of the three variables Monthly Income, sqrtInc and logInc.

Since the range of the three variables are not the same, it is not statistically meaningful to compare data such as the mean, standard deviation, and IQR of the three. The point of choosing a variable to be used for research is that the selected variable needs to be as close to a normal distribution as possible.

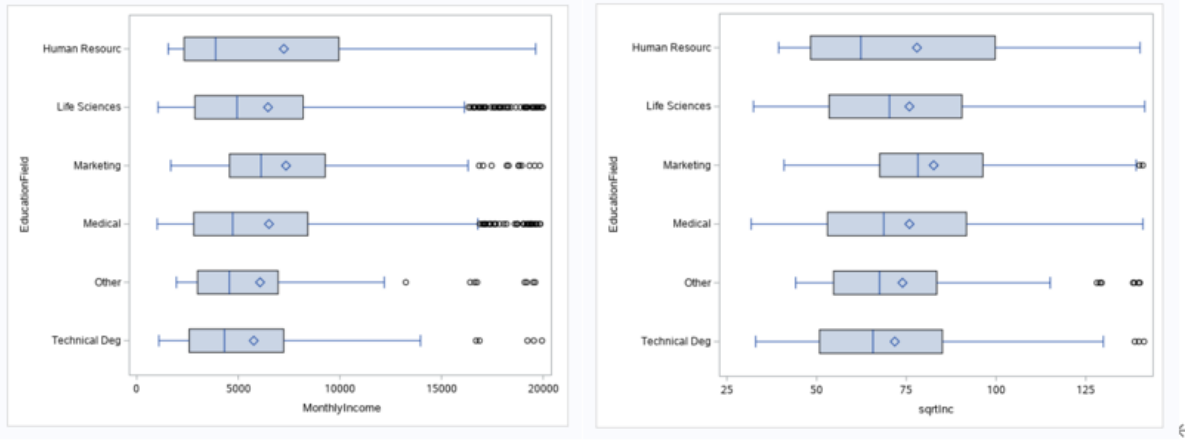
From the data given in the table, it is not difficult to see that all three show a right-skewed distribution, among which the right-skewed distribution of monthly income is the most obvious (skewness = 1.370) and the right-skewed distribution of logInc is the least obvious among the three (skewness = 0.286), while sqrtInc is somewhere in between (skewness = 0.862).

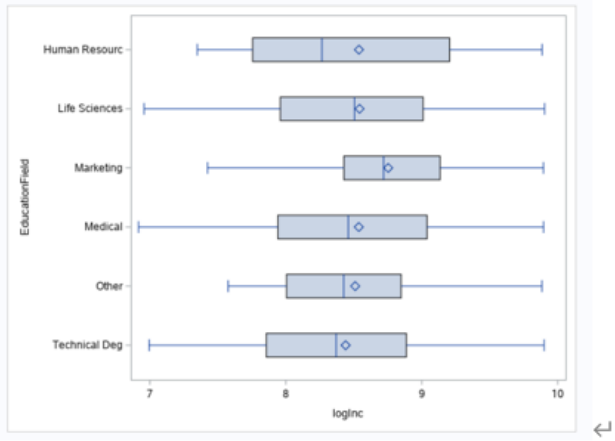
Based on the above analysis, logInc will be preliminarily selected as the most suitable variable for analysis, but more research is needed to verify this judgment.



The diagram above shows the distribution of monthly income, sqrtInc and logInc. Through direct observation, it can be found that although the monthly income has an obvious sharp edge (kurtosis = 1.005), but its distribution is an obvious right-skewed distribution, it also means that there should be a lot of outliers on the "right side" (this can be verified by the box plot). The distribution of logInc and sqrtInc is generally closer to the normal distribution than the data distribution of monthly income (this is consistent with the corresponding skewness value in the first table).

Observe the distribution of sqrtInc in diagram 1 and speculate that there may also be outliers on the "right side" (the boxplot of sqrtInc provided by Diagram 2 can also be verified), At the same time, the data distribution of sqrtInc is flatter than that of logInc, which is not conducive to subsequent analysis.





The distribution of logInc is closer to the normal distribution than the monthly income, and it is more concentrated than the sqrtInc data (the cv value of logInc is 7.769), and there are no outliers in the logInc data (refer to the previous diagram), logInc is the most suitable variable of the three to be used for further analysis (set logInc as tIncome).

Question 2

(a) (20 marks) Carry out a one-way analysis of variance (ANOVA) relating tIncome to EducationField. Use contrasts to test at least one a-priori hypothesis of your choice. Examine and comment on residuals. Also carry out appropriate post-hoc comparisons and discuss your results.

Normality tests are used for all educational field variables and the result is shown as below:

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.886531	Pr < W	0.0067
Kolmogorov-Smirnov	D	0.151547	Pr > D	0.1094
Cramer-von Mises	W-Sq	0.154061	Pr > W-Sq	0.0205
Anderson-Darling	A-Sq	1.083209	Pr > A-Sq	0.0067

EducationField = Human Resourc

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.969769	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.071331	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.659117	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.08048	Pr > A-Sq	<0.0050

EducationField = Life Sciences

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.982129	Pr < W	0.0377
Kolmogorov-Smirnov	D	0.06731	Pr > D	0.0779
Cramer-von Mises	W-Sq	0.107889	Pr > W-Sq	0.0904
Anderson-Darling	A-Sq	0.764668	Pr > A-Sq	0.0466

EducationField = Marketing

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.959127	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.070144	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.803302	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.736584	Pr > A-Sq	<0.0050

EducationField = Medical

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.939186	Pr < W	0.0007
Kolmogorov-Smirnov	D	0.129774	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.229285	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.43335	Pr > A-Sq	<0.0050

EducationField = Other

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.956613	Pr < W	0.0003
Kolmogorov-Smirnov	D	0.093574	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.29939	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.967505	Pr > A-Sq	<0.0050

EducationField = Technical Deg

Judging from the p-value values in several tables, it can be considered that nearly all group's tIncome is not normal distribution.

The GLM Procedure

Levene's Test for Homogeneity of tIncome Variance
ANOVA of Squared Deviations from Group Means

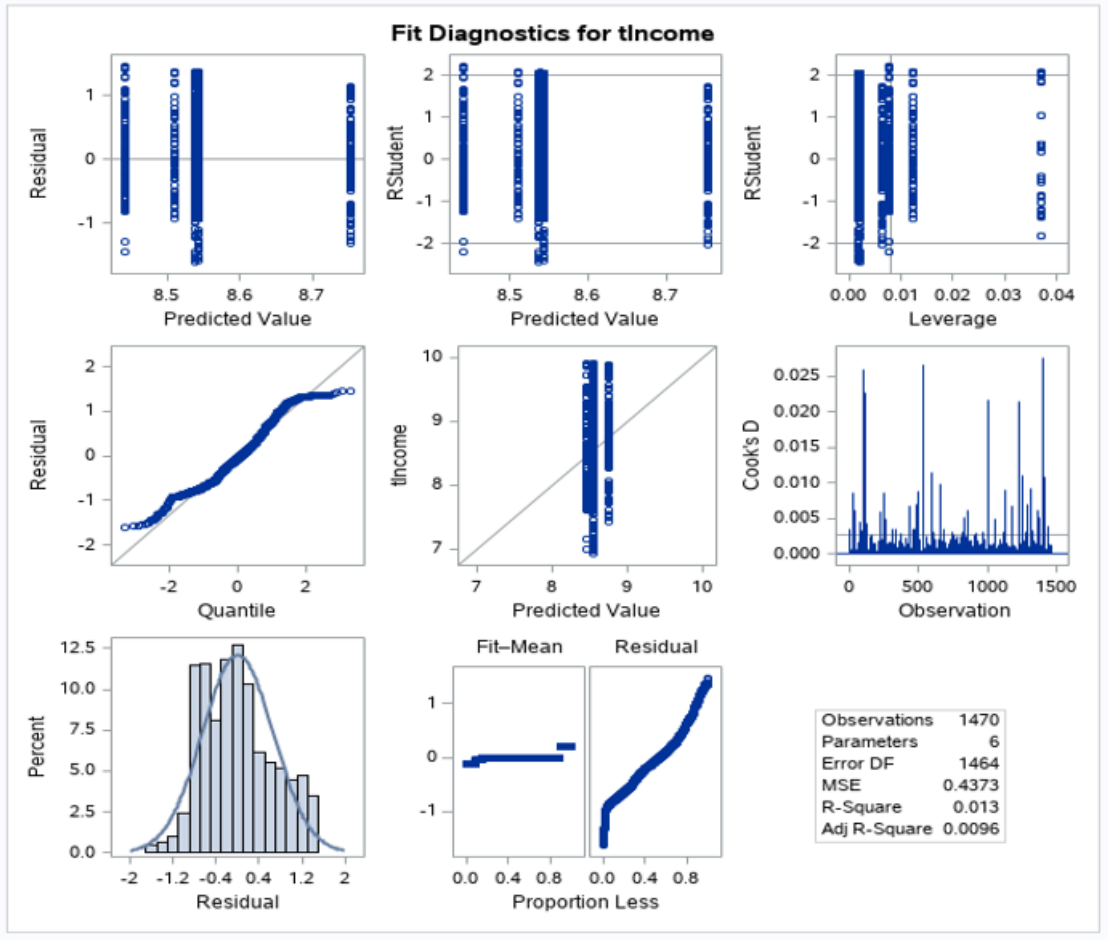
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
EducationField	5	5.4644	1.0929	4.31	0.0007
Error	1464	371.5	0.2537		

Welch's ANOVA for tIncome

Source	DF	F Value	Pr > F
EducationField	5.0000	5.05	0.0002
Error	191.5		

Although neither assumption is met, the Anova test will continue as required.

The following diagram shows the analysis of residual:



From the q-q plot of the residual in the above figure, we can find that when our model has a quantile between -2 and 2, the residual of the model is relatively close to the normal distribution, but in general, the residual of the model has light tails.

It is not difficult to see from Cook's D that there are some points whose values need to be further studied according to the Rule of thumb; however, the absence of value in the model leads to a Cause for concern.

The residuals versus fit plots do not show a clear pattern, so linearity and independence can be considered ok. There is no evidence that the vertical distributions are unequal, so a constant error

variance would also work.

An assumption is used: Employees with a marketing education have higher tIncome than other education backgrounds.

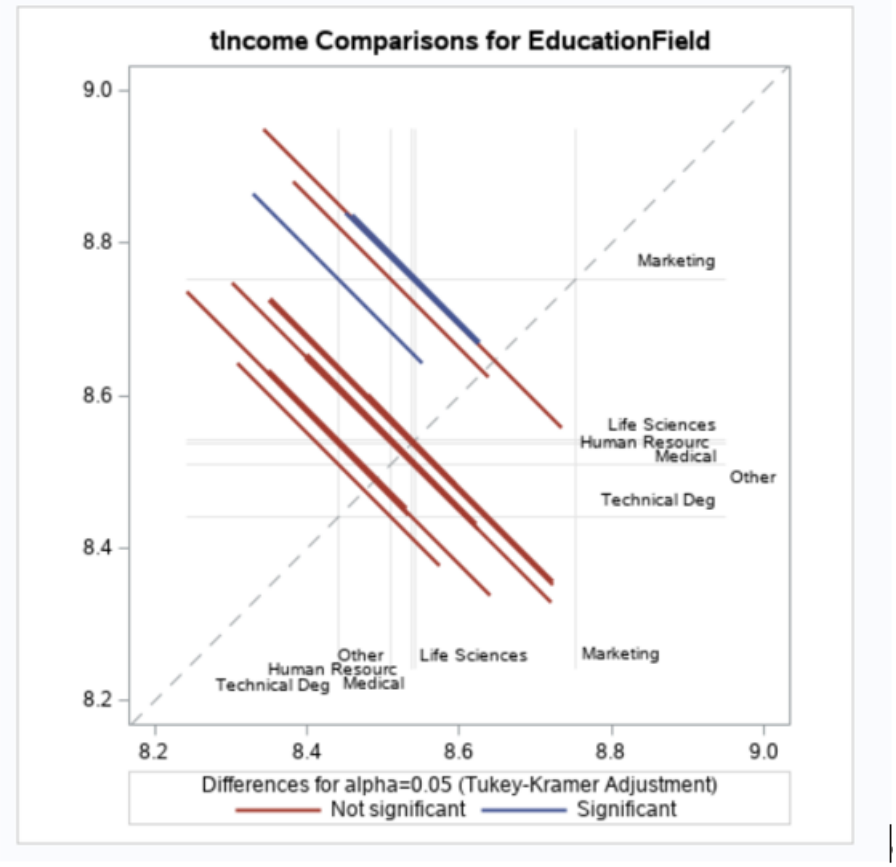
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Marketing vs other education field	1.19855211	0.30863572	3.88	0.0001

Based on the above table, marketing educational fields have significant different from the tIncome of all other educational fields (p-value < 0.05).

Comparisons significant at the 0.05 level are indicated by ***.				
EducationField Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Marketing - Life Sciences	0.21070	0.04257	0.37883	***
Marketing - Human Resourc	0.21520	-0.17756	0.60795	
Marketing - Medical	0.21645	0.04305	0.38984	**
Marketing - Other	0.24307	-0.01347	0.49961	
Marketing - Technical Deg	0.31314	0.09095	0.53532	**
Life Sciences - Marketing	-0.21070	-0.37883	-0.04257	**
Life Sciences - Human Resourc	0.00449	-0.36664	0.37563	
Life Sciences - Medical	0.00575	-0.11065	0.12214	
Life Sciences - Other	0.03237	-0.18966	0.25439	
Life Sciences - Technical Deg	0.10243	-0.07881	0.28367	
Human Resourc - Marketing	-0.21520	-0.60795	0.17756	
Human Resourc - Life Sciences	-0.00449	-0.37563	0.36664	
Human Resourc - Medical	0.00125	-0.37230	0.37480	
Human Resourc - Other	0.02787	-0.39080	0.44655	
Human Resourc - Technical Deg	0.09794	-0.30060	0.49649	
Medical - Marketing	-0.21645	-0.38984	-0.04305	***
Medical - Life Sciences	-0.00575	-0.12214	0.11065	
Medical - Human Resourc	-0.00125	-0.37480	0.37230	
Medical - Other	0.02662	-0.19942	0.25266	
Medical - Technical Deg	0.09669	-0.08945	0.28282	
Other - Marketing	-0.24307	-0.49961	0.01347	
Other - Life Sciences	-0.03237	-0.25439	0.18966	
Other - Human Resourc	-0.02787	-0.44655	0.39080	
Other - Medical	-0.02662	-0.25266	0.19942	
Other - Technical Deg	0.07007	-0.19525	0.33538	
Technical Deg - Marketing	-0.31314	-0.53532	-0.09095	***
Technical Deg - Life Sciences	-0.10243	-0.28367	0.07881	
Technical Deg - Human Resourc	-0.09794	-0.49649	0.30060	
Technical Deg - Medical	-0.09669	-0.28282	0.08945	
Technical Deg - Other	-0.07007	-0.33538	0.19525	

From the results of Tukey's post-hoc procedure shown in Table 3, there are 6 situations that means are significantly different and all six cases are marketing and other variables. This is confirmed by the

diffogram shown below.



(b) (10 marks) If the assumptions for ANOVA is not satisfied, use a nonparametric method to validate the results in question (a).

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable tIncome Classified by Variable EducationField

EducationField	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Life Sciences	606	442893.50	445713.00	8011.40775	730.847360
Other	82	57647.00	60311.00	3735.22968	703.012195
Medical	464	333783.50	341272.00	7564.38515	719.360991
Marketing	159	140557.50	116944.50	5054.93483	884.009434
Technical Deg	132	87460.50	97086.00	4652.97453	662.579545
Human Resourc	27	18843.00	19858.50	2185.39880	697.888889

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
24.7915	5	0.0002

The results of the Kruskal-Wallis's test in above table indicate that there is a significant difference in the means (H = 24.7915 has chi-square distribution with 5 DF, P-value = 0.0002).

The results of further research are shown in the table below:

The NPAR1WAY Procedure

Pairwise Two-Sided Multiple Comparison Analysis

Dwass, Steel, Critchlow-Fligner Method

Variable: tIncome

EducationField	Wilcoxon Z	DSCF Value	Pr > DSCF
Life Sciences vs. Other	0.5216	0.7376	0.9953
Life Sciences vs. Medical	0.4456	0.6302	0.9978
Life Sciences vs. Marketing	-4.0044	5.6630	0.0009
Life Sciences vs. Technical Deg	1.6328	2.3091	0.5766
Life Sciences vs. Human Resourc	0.4023	0.5689	0.9987
Other vs. Medical	-0.2718	0.3844	0.9998
Other vs. Marketing	-3.6449	5.1547	0.0036
Other vs. Technical Deg	0.8970	1.2685	0.9474
Other vs. Human Resourc	0.3440	0.4864	0.9994
Medical vs. Marketing	-4.2297	5.9817	0.0003
Medical vs. Technical Deg	1.4190	2.0068	0.7155
Medical vs. Human Resourc	0.2700	0.3818	0.9998
Marketing vs. Technical Deg	4.3183	6.1070	0.0002
Marketing vs. Human Resourc	1.7108	2.4195	0.5245
Technical Deg vs. Human Resourc	-0.1995	0.2822	1.0000

There are four cases in the above table where the p-value is less than 0.05. From the values provided in the above table, it can be proved that in the above four cases, there is a significant difference in tIncome between the marketing and the object being compared, which is consistent with the results obtained in the previous table.

(c) (25 marks) Use SAS to perform a one-way ANCOVA relating tIncome to EducationField and TotalWorkingYears with TotalWorkingYears as a covariate, including appropriate post-hoc comparisons:

– Confirm that there is a linear relationship between the response variable and the covariate (a scatterplot and correlation coefficient plus a comment will suffice)

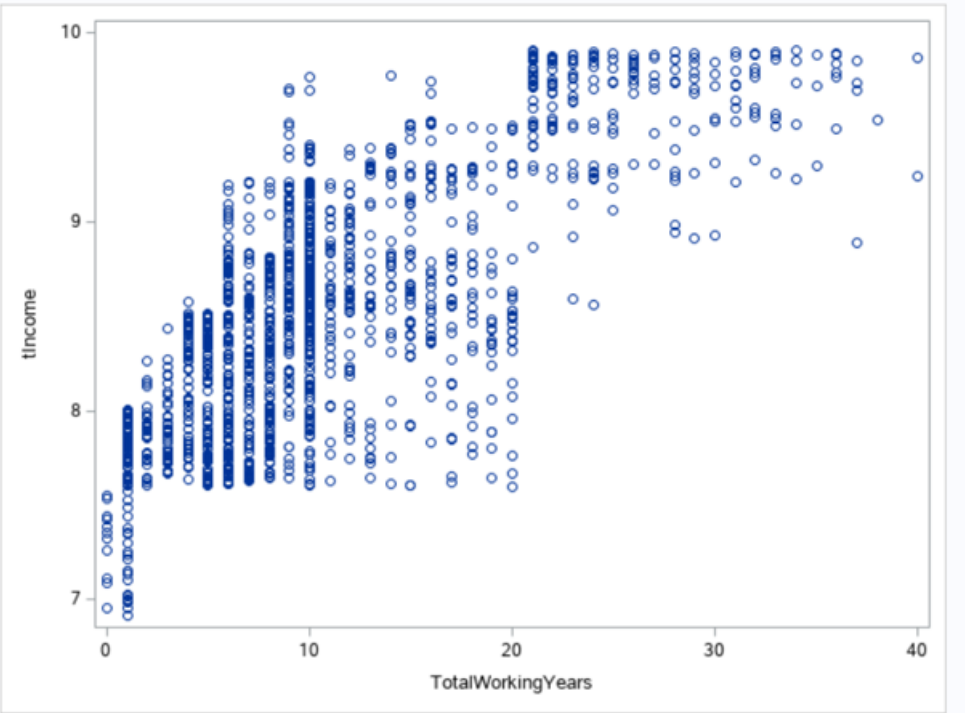


Diagram shows a scatter plot of the relationship between tIncome and TotalWorkingYears. From the above figure, tIncome and TotalWorkingYears can be considered to have a positive linear relationship.

– Check the two additional ANCOVA assumptions (report and comments only on the parts of the output most directly relevant to condition checking):

The GLM Procedure

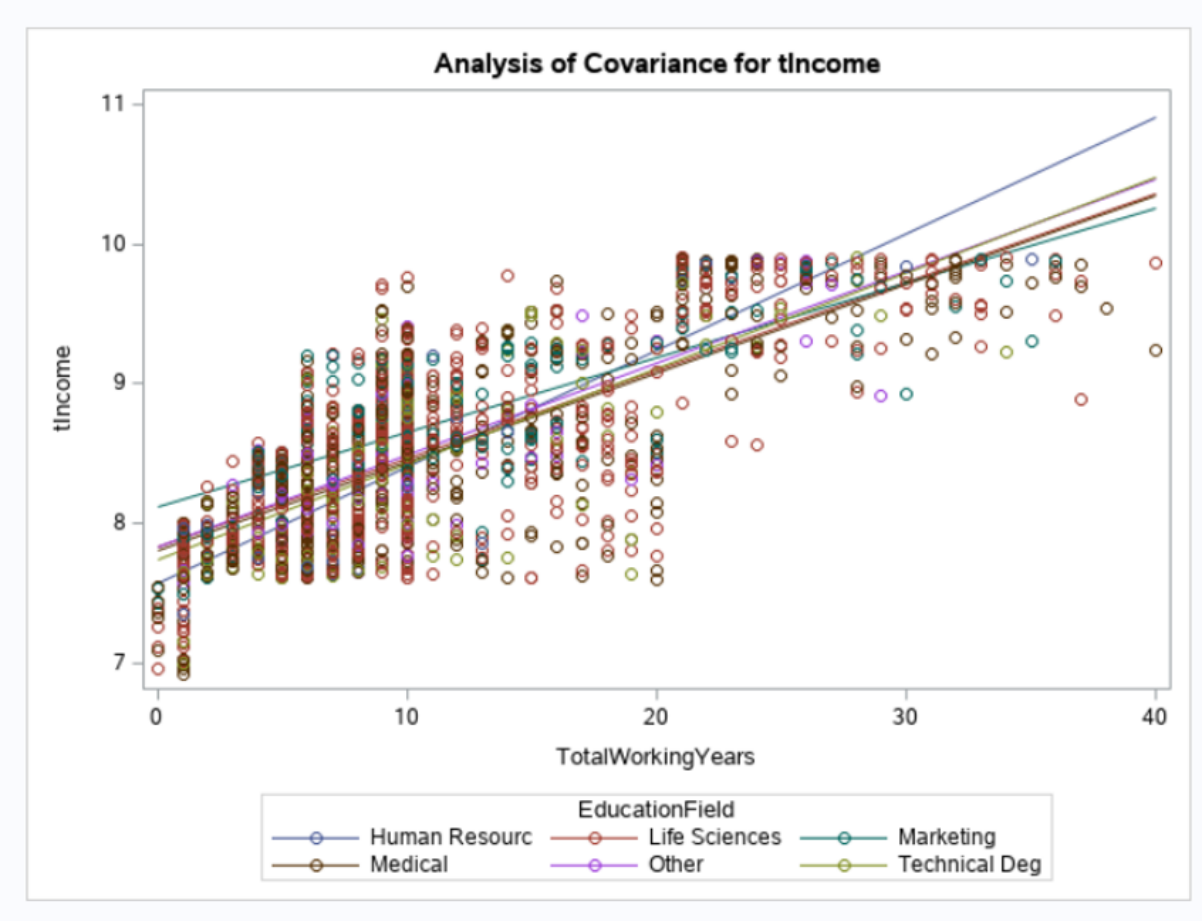
Levene's Test for Homogeneity of TotalWorkingYears Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
EducationField	5	57690.8	11538.2	1.10	0.3581
Error	1464	15347462	10483.2		

Welch's ANOVA for TotalWorkingYears

Source	DF	F Value	Pr > F
EducationField	5.0000	1.17	0.3276
Error	192.1		

Based on the data in table 6, different education fields have no effect on total working years (p-value > 0.05), so it can be considered that education fields and total working years are independent.



From the above figure, in different educational fields, the correlation between total working years and tIncome are not quite similar.

At the same time, further research was done:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EducationField	5	5.1865909	1.0373182	5.30	<.0001
TotalWorkingYears	1	144.5527760	144.5527760	737.89	<.0001
TotalWork*EducationF	5	1.9816882	0.3963376	2.02	0.0727

Based on the above table, we can consider this assumption to hold because p-value > 0.05. Therefore, "TotalWork*EducationF" will not be put into the model while the model is building.

— **Report and briefly discuss your results.**

Re-run the Ancova test after removing "TotalWork*EducationF", and the results are as follows:

The GLM Procedure

Dependent Variable: tincome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	360.9528510	60.1588085	306.02	<.0001
Error	1463	287.6022826	0.1965839		
Corrected Total	1469	648.5551336			

R-Square	Coeff Var	Root MSE	tincome Mean
0.556549	5.184181	0.443378	8.552515

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EducationField	5	5.2098521	1.0419704	5.30	<.0001
TotalWorkingYears	1	352.5509634	352.5509634	1793.39	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	7.793728209	B	0.04150065	187.80	<.0001
EducationField Human Resourc	0.013263628	B	0.09367047	0.14	0.8874
EducationField Life Sciences	0.039623187	B	0.04261300	0.93	0.3526
EducationField Marketing	0.211886953	B	0.05226246	4.05	<.0001
EducationField Medical	0.013771983	B	0.04378102	0.31	0.7531
EducationField Other	0.063527955	B	0.06234309	1.02	0.3084
EducationField Technical Deg	0.000000000	B	.	.	.
TotalWorkingYears	0.063069634		0.00148930	42.35	<.0001

The model can be considered as significant (F-value = 306.02 with p-value < 0.001).

Using partial sums of squares (Type III SS), EducationField (F-value = 5.30 with p-value < 0.001) is significant, while totalworkingyears (F-value = 1793.39 with p-value < 0.001) is also significant.

EducationField	tIncome LSMEAN	LSMEAN Number
Human Resourc	8.51839156	1
Life Sciences	8.54475112	2
Marketing	8.71701489	3
Medical	8.51889992	4
Other	8.56865589	5
Technical Deg	8.50512793	6

Least Squares Means for effect EducationField
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: tIncome

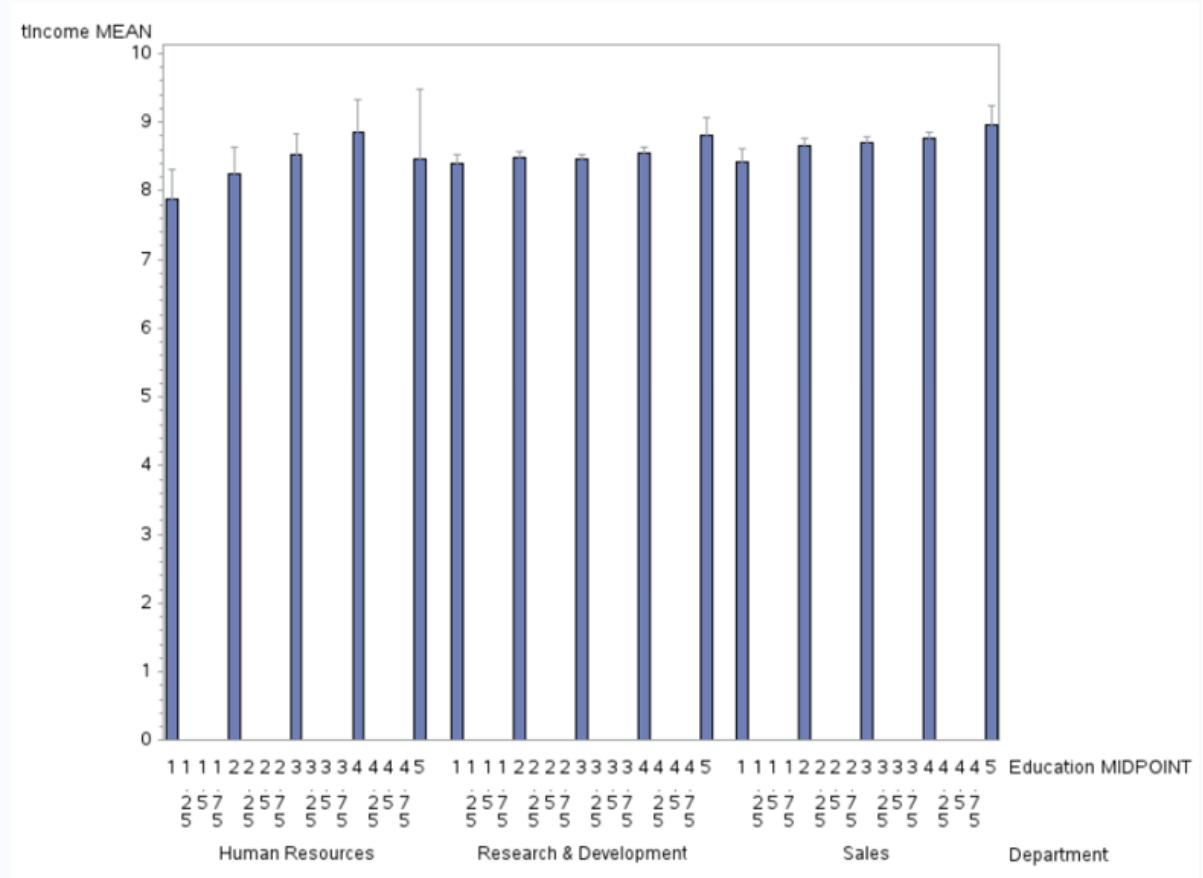
i/j	1	2	3	4	5	6
1		0.9997	0.2612	1.0000	0.9958	1.0000
2	0.9997		0.0002	0.9347	0.9975	0.9388
3	0.2612	0.0002		<.0001	0.1370	0.0008
4	1.0000	0.9347	<.0001		0.9371	0.9996
5	0.9958	0.9975	0.1370	0.9371		0.9117
6	1.0000	0.9388	0.0008	0.9996	0.9117	

The cases circled in red in the above figure are all significantly different, because their p-values are all less than 0.05, And it can be found that these numbers are marketing corresponding to other variables.

According to the results of the above analysis, if we consider totalworkingyears as covariate, educationfield is significant affect on tIncome, especially in the case of marketing.

Question 3 (25 marks)

(a) Perform and analyze a factorial ANOVA model to determine whether there is statistically significant difference in tIncome by Department and Education. Carry out to test whether there is evidence of interaction between Department and Education. Examine and comment on residuals. Carry out appropriate follow-up analysis and discuss your results.



From the above figure, in the time of human resources, the education midpoint showed a trend of rising first and then falling last, while Research&Development showed the situation that it did not change at first and then rose. At Sales, it showed a steady rise.

The changing trend of human resources is different from the other two, so there is an interaction between department and education.

Further research is required before checking assumptions.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.951176	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.120755	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.586837	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	3.258803	**Pr > A-Sq**	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.96611	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.069358	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.323112	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.626018	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.964515	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.082749	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.830349	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.884714	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.967224	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.070388	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.444993	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.31614	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.955015	Pr < W	0.0635
Kolmogorov-Smirnov	D	0.104045	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.094855	Pr > W-Sq	0.1314
Anderson-Darling	A-Sq	0.66221	Pr > A-Sq	0.0826

It can be found that Education basically presents a non-normal distribution at five levels, except that normality can be assumed when education=5.

But we can check the sample size:

The MEANS Procedure

Analysis Variable : tIncome		
Education	N Obs	N
1	170	170
2	282	282
3	572	572
4	398	398
5	48	48

It can be found that the value of N Obs of the five is greater than 30. According to the IC theorem, the Anova test can continue to be used.

Department was also subjected to Normality test:

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.893018	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.150309	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.331682	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.245558	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.947185	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.077014	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.434635	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	16.42991	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.984372	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.073071	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.317251	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.9721	Pr > A-Sq	<0.0050

The results of the test show that there is no situation that can assume normality (all p-values < 0.05).

But based on the same logic as before:

The MEANS Procedure

Analysis Variable : tIncome		
Department	N Obs	N
Human Resources	63	63
Research & Development	961	961
Sales	446	446

Since the size of the three is greater than 30, according to the CI theorem, the Anova test will continue to be used.

The next two tables show the results after testing Education and Department:

The GLM Procedure

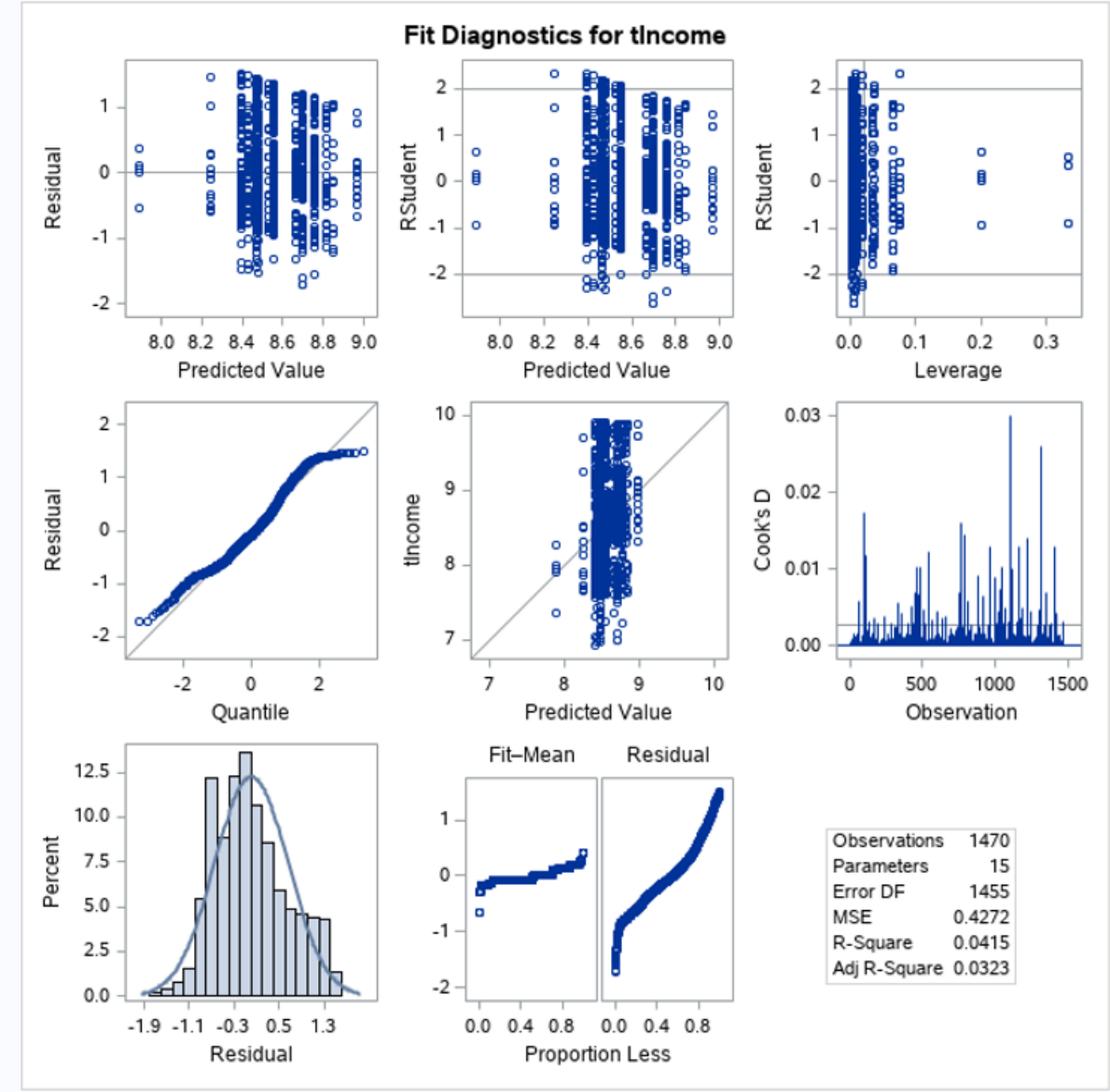
Levene's Test for Homogeneity of tIncome Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Education	4	2.9586	0.7396	2.97	0.0186
Error	1465	364.7	0.2490		

Welch's ANOVA for tIncome			
Source	DF	F Value	Pr > F
Education	4.0000	6.71	<.0001
Error	279.3		

The GLM Procedure

Levene's Test for Homogeneity of tIncome Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Department	2	7.2204	3.6102	13.34	<.0001
Error	1467	397.0	0.2706		

The p-values of both are less than 0.05, so there is not enough reason to assume equal variance.



From the Q-Q plot and histogram, the distribution of the residuals of the data is somewhat close to the normal distribution, but strictly speaking, the distribution presents a left-skewed distribution.

From the graph of Residual vs. Predicted value, it can be found that no pattern is displayed, which is very ideal, but there are some points whose value exceeds 2 or is less than -2, so it can be considered that there are outliers.

The GLM Procedure

Dependent Variable: tIncome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	26.9387259	1.9241947	4.50	<.0001
Error	1455	621.6164077	0.4272278		
Corrected Total	1469	648.5551336			

R-Square	Coeff Var	Root MSE	tIncome Mean
0.041537	7.642508	0.653627	8.552515

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Education	4	8.63622442	2.15905611	5.05	0.0005
Department	2	5.32046916	2.66023458	6.23	0.0020
Education*Department	8	4.38603540	0.54825443	1.28	0.2478

It can be found that only the p-value of Education*Department is greater than 0.05, so this variable does not significantly affect tIncome, and the amount of change will be excluded.

Question 4 (10 marks)

Write a summary of your findings from Questions 1–3. Keep the technical details of the analyses that led you to these conclusions to the absolute minimum. Rather, focus on practical significance and present your findings in non-specialist terms. One to two paragraphs (up to a page) will be sufficient.

The data provided Monthly Income as raw data, but through descriptive statistical analysis and visualization of it, this data can be judged not to be used as a variable for further research, because monthly income is not close to normal distribution, and there are still many outliers. The data is transformed by square root and log. After performing the same research analysis on the transformed two data, the log-transformed data was finally used as the research variable, because this log transformation can not only eliminate outliers in the data (square root does not), but also make the cv value of the data drop and the distribution closer to a normal distribution.

When analyzing log income, it can be found from the results that when there is marketing, the result value will be higher than that of other centralized cases, so Anova test is used to verify the hypothesis. The results show significant results which confirm the assumption. Since the distribution of the data is basically not normal, the non-parameter test is used for research, and the results obtained are more rigorous and accurate, but are consistent with the results of the Anova test. At the same time, if the total working years are used as covariate for research, and the Ancova test is used for research, the results show that educationfield is a significant affect on tIncome when total working years is used as covariate, especially in the case of marketing.

Taking Education and Department as two variables to conduct Anova test research, it can be found that these two variables are sifinicant affect on tIncome, but the interaction between the two is not significant.