Assignment 2
44468423
Gautham Meenakshisundaram

1a) What statistical test has been used?
Two-Samples Independent T-Test with Hospitalstay as outcome variable and Group as predictor variable.

1b) Reproduce this result in SPSS (show the output).

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Hospital stay	Con	104	13.288	8.7230	.8554
	Exp	99	10.687	7.0965	.7132

Independent Samples Test

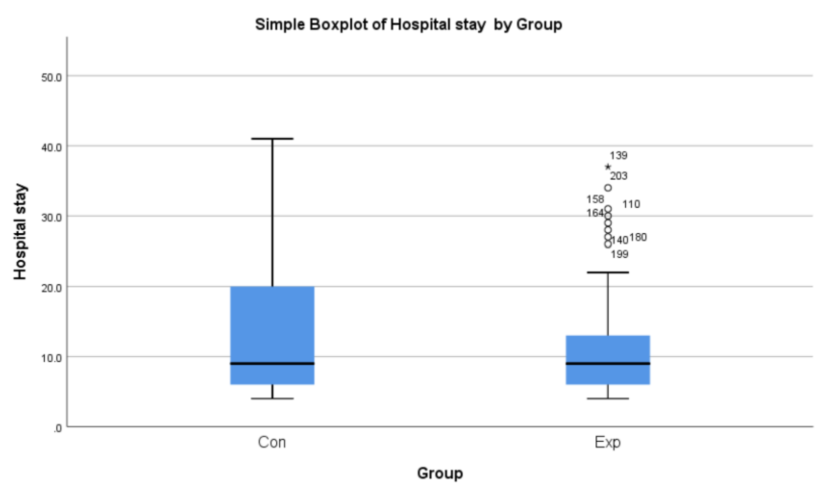
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Hospital stay	Equal variances assumed	12.501	.001	2.324	201	.021	2.6016	1.1193	.3945	4.8087
	Equal variances not assumed			2.336	196.289	.021	2.6016	1.1137	.4052	4.7979

1c) What are the assumption(s) underlying this test? Do you think they are satisfied? (It is not necessary to perform a formal hypothesis test.)
Assumption 1: The outcome variable should be a continuous variable, as in the case of Hospitalstay which measures the length of stay in hospital, in days.

Assumption 2: The predictor variable should be comprised of two categorical and independent groups, as in the case of Group which contains two groups - Con as control group and Exp as treatment group.

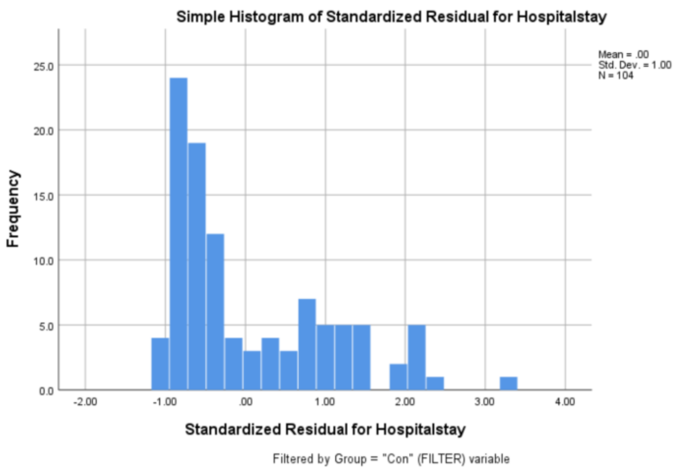
Assumption 3: The observations in each group of the predictor variable should be independent of the other group. The PEP device will be used by the 99 patients in Exp group while the 104 patients in Con group will not have access to it which satisfies assumption 3.

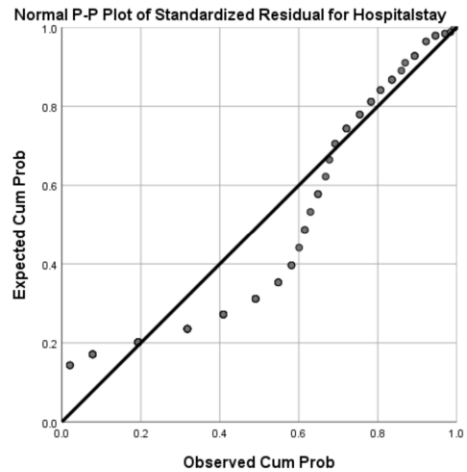
Assumption 4: There should be no significant outliers negatively influencing the results.



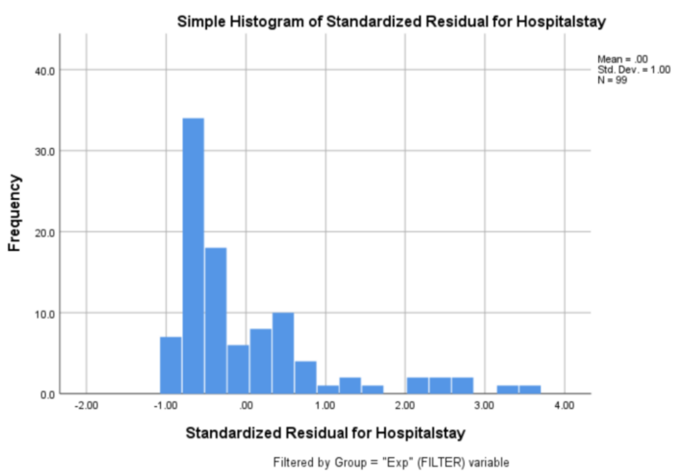
The boxplot of Exp group shows nearly 7 obvious outliers, thereby **not** satisfying assumption 4.

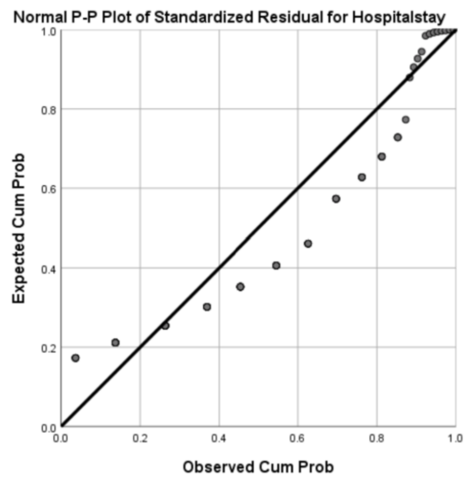
Assumption 5: The outcome variable should at least show hints of normality / linearity for each group in the predictor variable.





The histogram for Con group is right skewed and its P-P plot shows curvature.





The histogram for Exp group is right skewed and its P-P plot does not have residuals near the linear line.

Assumption 5 is **not** satisfied as both the groups does not show normality / linearity against the variable Hospitalstay.

Assumption 6: Check for homogeneity of variances between the groups in the predictor variable.

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Hospital stay	Con	104	13.288	8.7230	.8554
	Exp	99	10.687	7.0965	.7132

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Hospital stay	Equal variances assumed	12.501	.001	2.324	201	.021	2.6016	1.1193	.3945	4.8087
	Equal variances not assumed			2.336	196.289	.021	2.6016	1.1137	.4052	4.7979

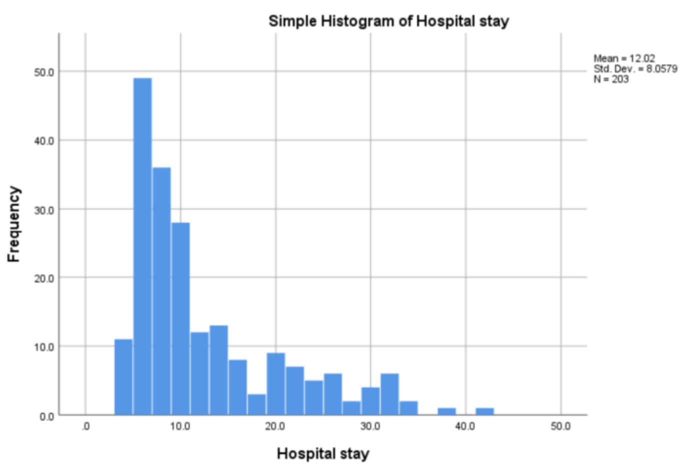
$H_0 : \sigma^2_1 = \sigma^2_2$
H_1 : not H_0

where,
σ^2_1 = variance in hospital stay for patients from group Con who were not facilitated with the PEP device
σ^2_2 = variance in hospital stay for patients from group Exp who were facilitated with the PEP device

The p-value of Levene's Test is $0.001 < 0.05$ which violates the assumption of homogeneity / equal variances between the groups, thereby reject the null hypothesis H_0 and conclude that assumption 6 is **not** satisfied.

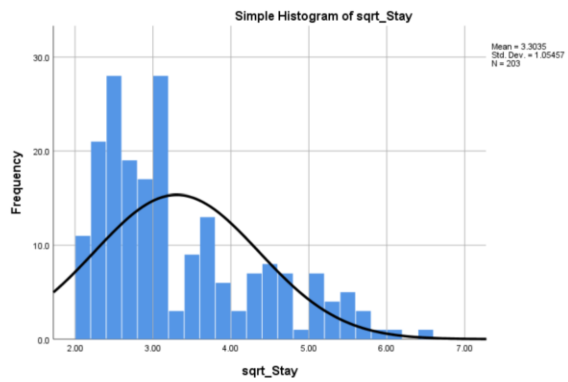
1d) Develop an appropriate linear model for evaluating whether there has been a treatment effect on length of hospital stay. You should only use demographic and baseline variables for prediction (i.e. do not use outcome variables to predict another outcome). Note that in order to evaluate whether there has been a treatment effect, treatment (Group) must be in the model, whether or not it is significant. Perform usual diagnostic model checking. Write down your final model.

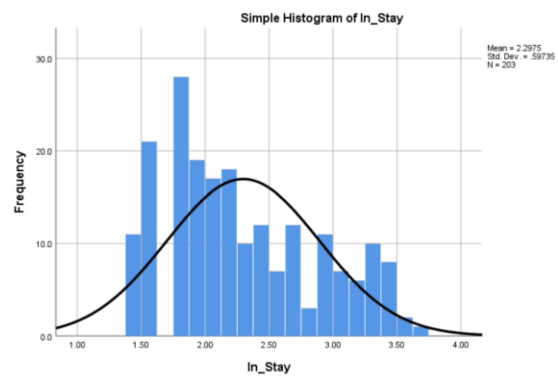
Outcome Variable:

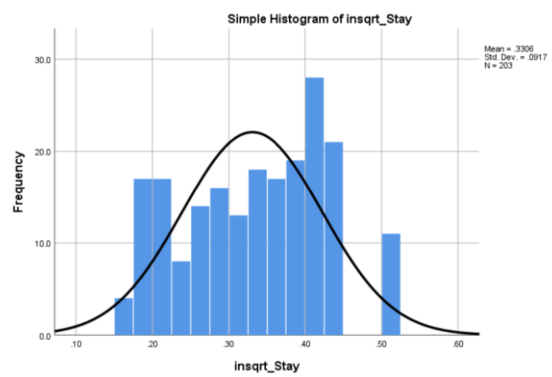


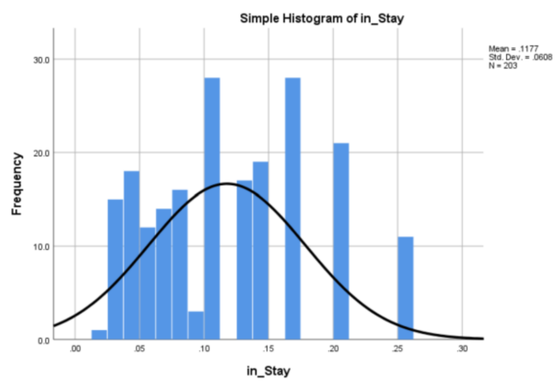
The histogram of the outcome variable Hospitalstay is right skewed and requires transformation.

A comparison of square root, log, inverse square root and inverse transformations on the outcome variable Hospitalstay is shown below.





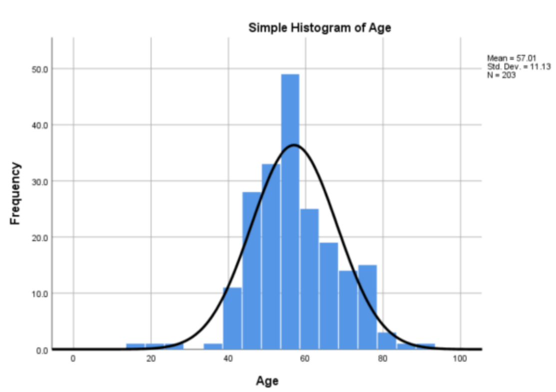


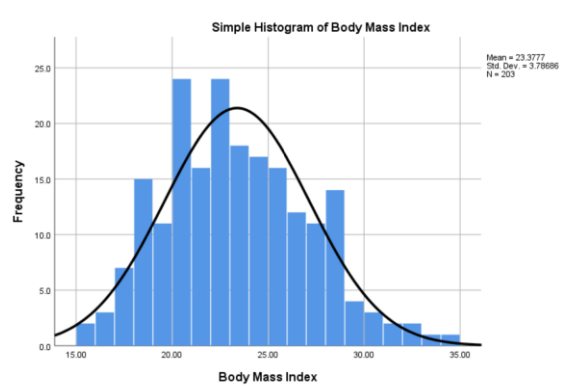


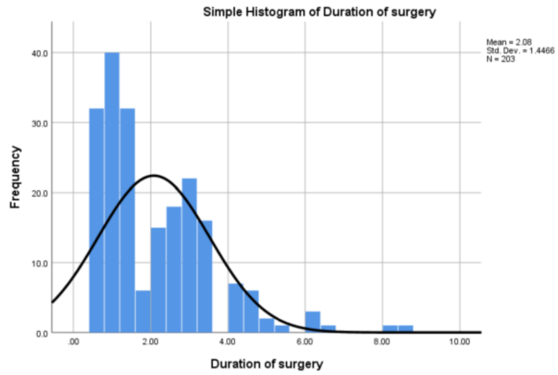
The inverse square root transformation seems to be the most symmetric while the other transformations seems to be right skewed.

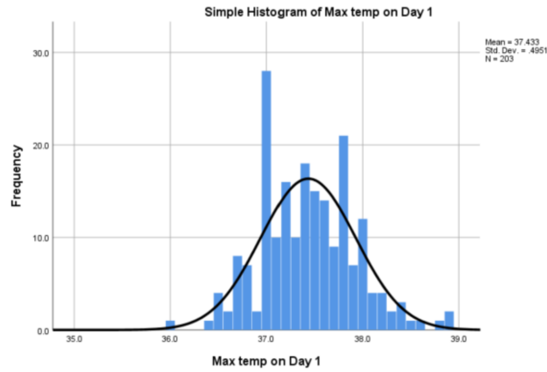
Continuous Predictor Variables:

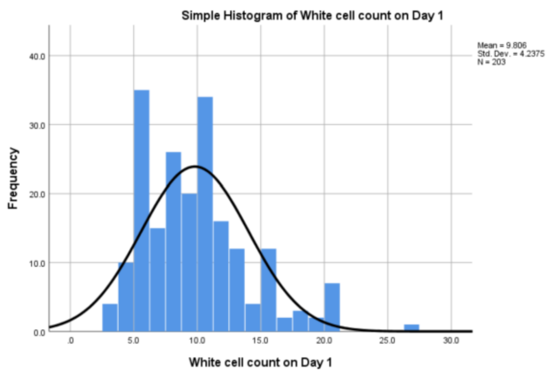
The histograms of all continuous predictor variables are shown below.





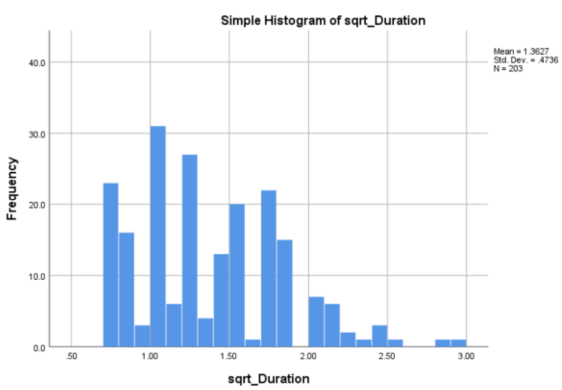


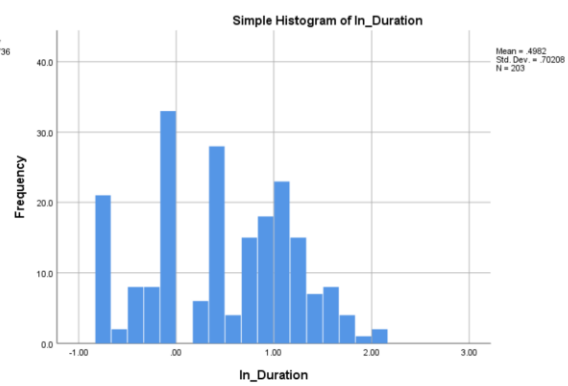




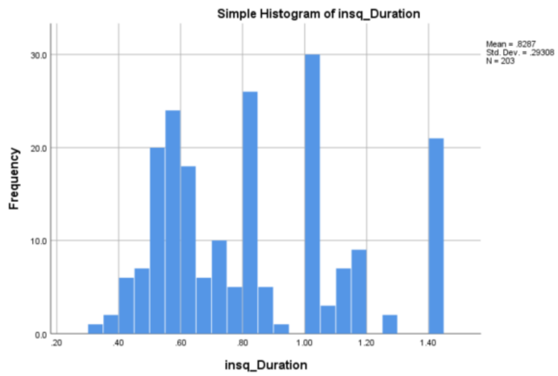
Apart from Duration of Surgery, all the histograms of predictor variables seem to be symmetric. The Duration of Surgery is right skewed and requires transformation.

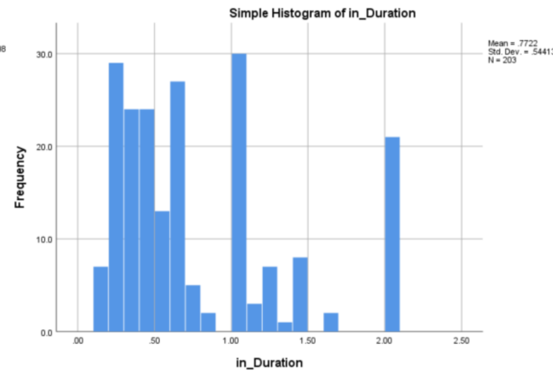
A comparison of square root, log, inverse square root and inverse transformations on the predictor variable Duration of Surgery is shown below.





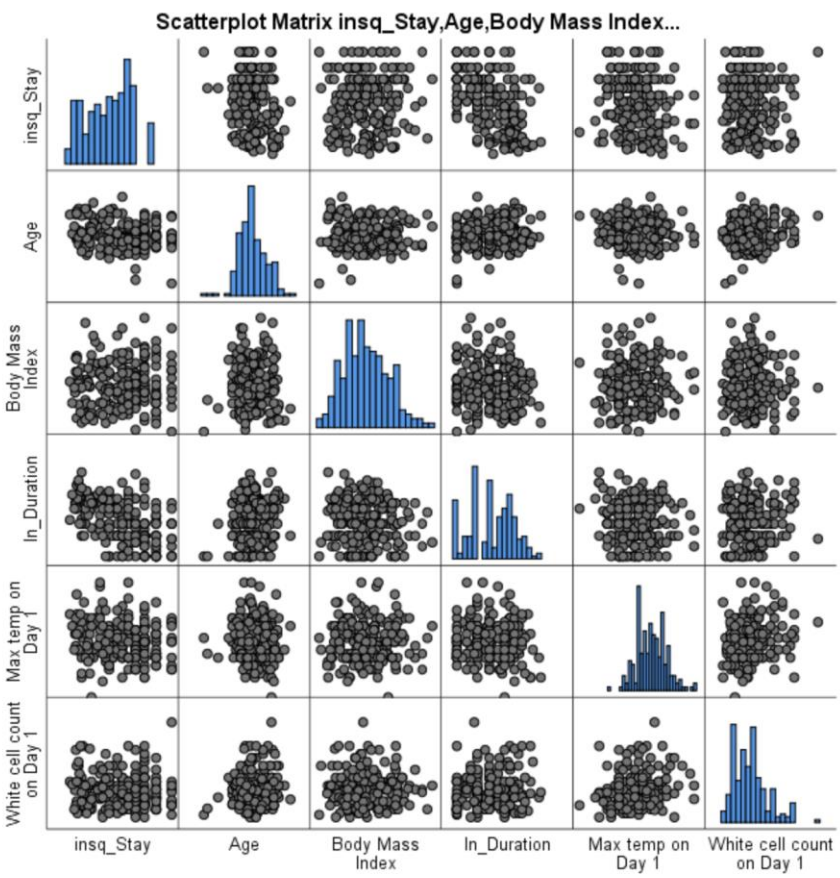
Assignment 2
44468423
Gautham Meenakshisundaram





The log transformation seems to be the most symmetric while the other transformations seems to be right skewed.

All Continuous Variables:



All the predictor variables show a random scatter against the transformed outcome variable. Also, there does not seem to be any collinearity between the predictor variables.

Categorical Predictor Variables:

Group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Con	104	51.2	51.2	51.2
	Exp	99	48.8	48.8	100.0
	Total	203	100.0	100.0	

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	79	38.9	38.9	38.9
	M	124	61.1	61.1	100.0
	Total	203	100.0	100.0	

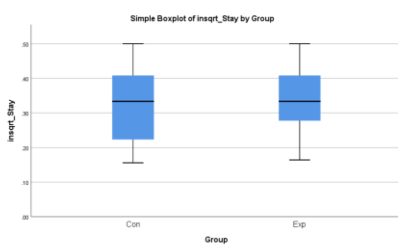
Site of surgery

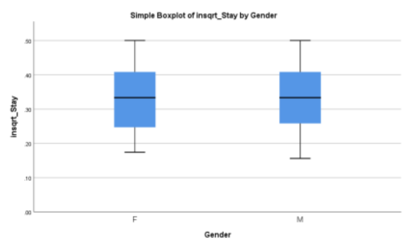
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	abdo	99	48.8	48.8	48.8
	thor	104	51.2	51.2	100.0
	Total	203	100.0	100.0	

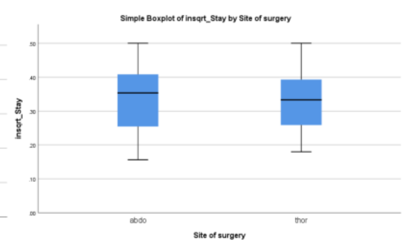
There seems to be a good distribution of observations between the categories in each of these categorical predictor variables.

It looks like Exp might be the reference category in Group which might not be ideal as Con has more observations and can also be considered as a baseline in Group for any treatments. Let's review the reference category of Group while fitting single regression models.

Outcome Variable vs Categorical Predictor Variables:

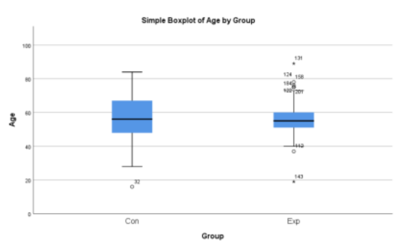


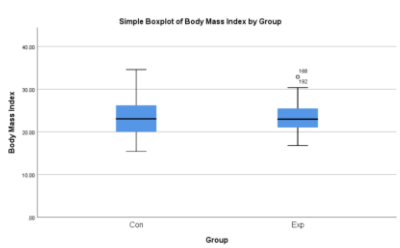


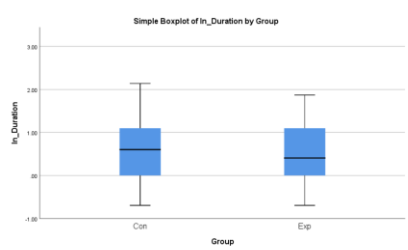


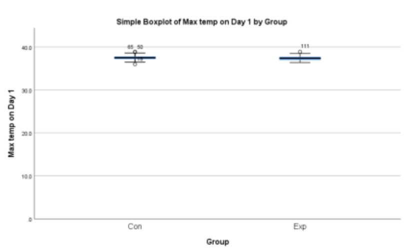
There is no increasing or decreasing trend between any of the categories in these predictor variables, indicating that the categories are not significantly different. The distribution of the transformed outcome variable values is same across the categories in these predictor variables.

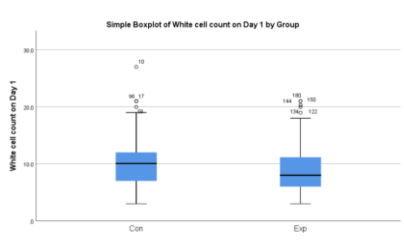
Continuous Predictor Variables vs Categorical Predictor Variables:

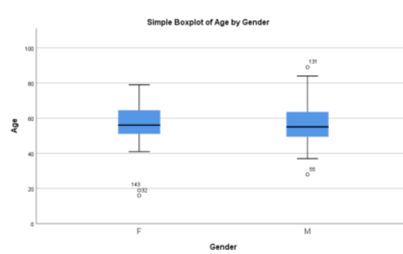


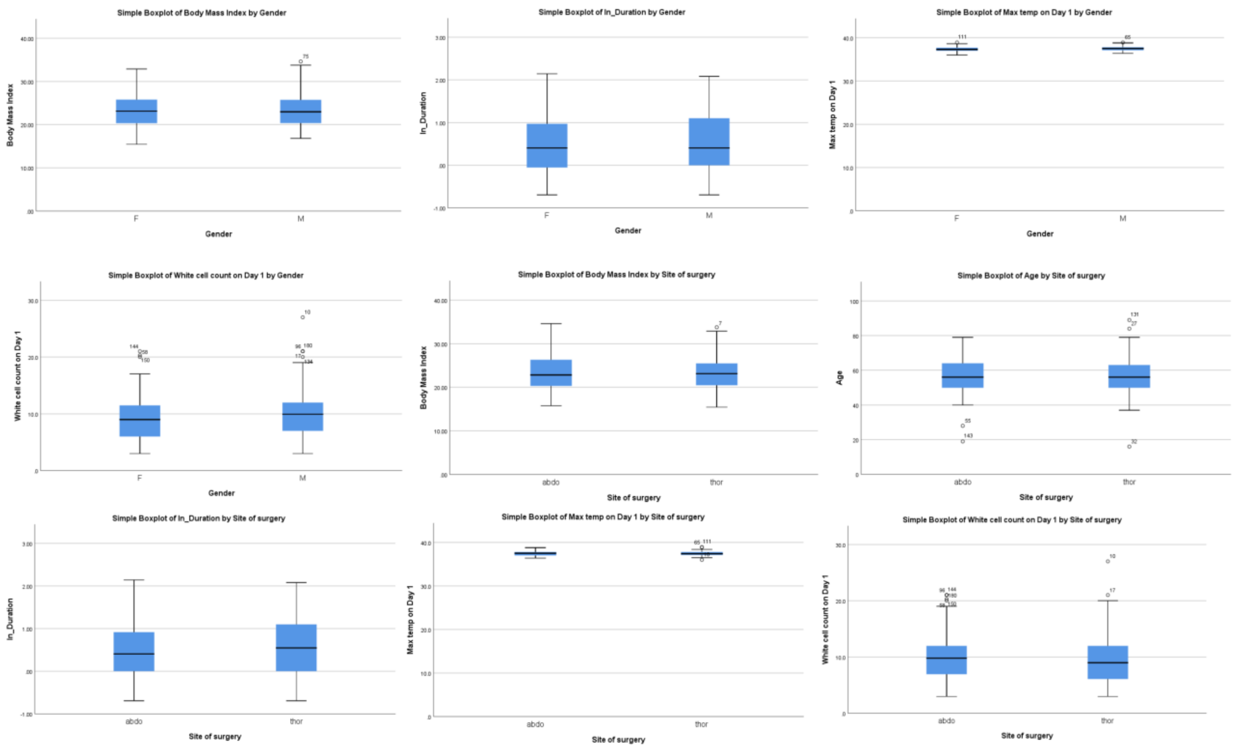












None of the boxplots show a linear trend, thereby indicating that there is no association between these continuous predictor variables and categorical predictor variables.

Single Variable Regressions

Model	**p-value**	**Significance**
InvSqRt(Hospitalstay) ~ Group	0.039	**< 0.05 (significant)**
InvSqRt(Hospitalstay) ~ Gender	0.755	> 0.05 (not significant)
InvSqRt(Hospitalstay) ~ Age	0.014	**< 0.05 (significant)**
InvSqRt(Hospitalstay) ~ BodyMassIndex	0.456	> 0.05 (not significant)
InvSqRt(Hospitalstay) ~ SiteofSurgery	0.349	> 0.05 (not significant)
InvSqRt(Hospitalstay) ~ Log(Durationofsurgery)	p<0.001	**< 0.05 (significant)**
InvSqRt(Hospitalstay) ~ MaxTemp	0.028	**< 0.05 (significant)**
InvSqRt(Hospitalstay) ~ WhiteCellCountonDay1	0.228	> 0.05 (not significant)

While fitting the first model InvSqRt(Hospitalstay) ~ Group, we can see that the reference category for Group is Exp rather than Con which has more observations and can be considered as a baseline in Group for any treatments.

Parameter Estimates

Dependent Variable: insq_Stay

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	.344	.009	37.651	.000	.326	.362
[Group=Con]	-.026	.013	-2.074	.039	-.052	-.001
[Group=Exp]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Let's recode the Exp group as 1 and Con group as 2, for assigning the latter as the reference category in Group.

Parameter Estimates

Dependent Variable: insq_Stay

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Intercept	.318	.009	35.620	.000	.300	.335
[Group2=1.00]	.026	.013	2.074	.039	.001	.052
[Group2=2.00]	0a	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Multiple Variable Regressions

The significant variables from single variable regression models are included together in these multiple variable regression models.

Model	p-values
InvSqRt(Hospitalstay) ~ Group + Age + Log(Durationofsurgery) + MaxTemp	Group=0.187, **Age=0.101**, Log(Durationofsurgery)=p<0.001, MaxTemp=0.004
InvSqRt(Hospitalstay) ~ Group + Log(Durationofsurgery) + MaxTemp	Group=0.162, Log(Durationofsurgery)= p<0.001, MaxTemp=0.005

As advised, Group is included in the model for determining the treatment effect, despite its non-significant p-values.

Final Model

Tests of Between-Subjects Effects

Dependent Variable: insq_Stay

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.511a	3	.170	28.572	.000
Intercept	.082	1	.082	13.684	.000
Group2	.012	1	.012	1.970	.162
ln_Duration	.443	1	.443	74.213	.000
MaxtemponDay1	.048	1	.048	7.962	.005
Error	1.187	199	.006		
Total	23.891	203			
Corrected Total	1.699	202			

a. R Squared = .301 (Adjusted R Squared = .291)

Parameter Estimates

Dependent Variable: insq_Stay

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Intercept	1.525	.415	3.675	.000	.706	2.343
[Group2=1.00]	.015	.011	1.404	.162	-.006	.037
[Group2=2.00]	0a	.	.	.	.	.
ln_Duration	-.067	.008	-8.615	.000	-.082	-.052
MaxtemponDay1	-.031	.011	-2.822	.005	-.053	-.009

a. This parameter is set to zero because it is redundant.

$1/\text{sqrt}(\hat{y}_i) = 1.525 + 0.015\, x_{i1} - 0.067\, \ln(x_{i2}) - 0.031\, x_{i3}$

where, for each i,

x_{i1} = {1 if Group = 1 (Exp), 0 if Group = 2 (Con)}
x_{i2} = Duration of surgery
x_{i3} = Maximum temperature
$\hat{y}_i$ = Length of hospital stay

The reference category in Group is 2 (Con) for this model.

For every hour of increase in the duration of surgery for a patient who is part of Group 1 (Exp) and is facilitated with PEP device, the length of hospital stay decreases by 0.067 days, when compared to patients in Group 2 (Con).

For every degree Celsius of increase in the maximum body temperature for a patient who is part of Group 1 (Exp) and is facilitated with PEP device, the length of hospital stay decreases by 0.031 days, when compared to patients in Group 2 (Con).

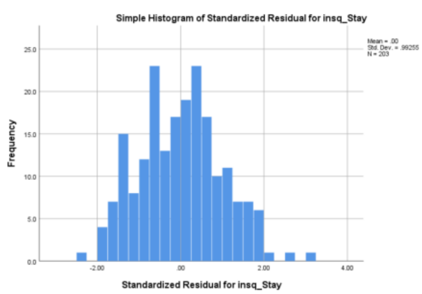
Influential Observations:

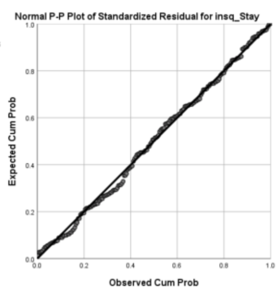
There are no leverage values more than 0.04 which does not satisfy the required moderate leverage values of 0.2 – 0.5 and high leverage values of > 0.5.

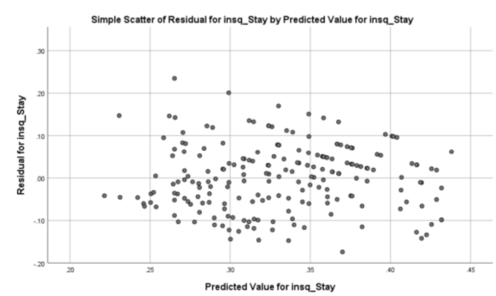
There are no Cook's distance values greater than 0.06 which satisfies the Cook's benchmark value of 1.

To conclude, there doesn't seem to be any influential observations.

Diagnostic Checking







The histogram of standardized residuals seems to be symmetric.
The normal probability plot has the standardized residuals distributed closely on the linear line, indicating linearity.
The unstandardized predicted values vs unstandardized residuals plot shows residuals in a random scatter around the zero line.

1e) In the clinical trials world, there is debate on whether treatment effects should be presented as unadjusted or adjusted. Unadjusted treatment effects do not include any covariates besides treatment in the model, whereas adjusted treatment effects include baseline covariates that are found to be significant. Compute a similar regression model as in (d) for the unadjusted evaluation of treatment effect. (Just present the SPSS output.)

Tests of Between-Subjects Effects

Dependent Variable: insq_Stay

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.036[a]	1	.036	4.300	.039
Intercept	22.222	1	22.222	2685.740	.000
Group2	.036	1	.036	4.300	.039
Error	1.663	201	.008		
Total	23.891	203			
Corrected Total	1.699	202			

a. R Squared = .021 (Adjusted R Squared = .016)

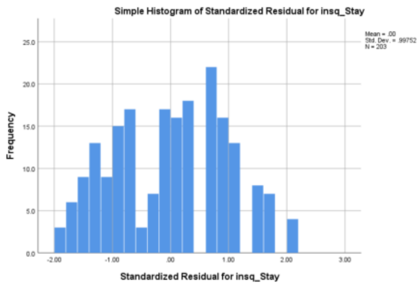
Parameter Estimates

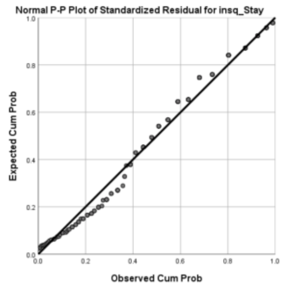
Dependent Variable: insq_Stay

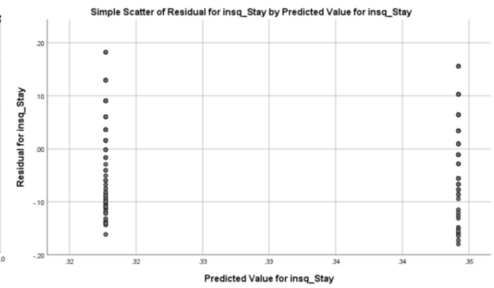
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	.318	.009	35.620	.000	.300	.335
[Group2=1.00]	.026	.013	2.074	.039	.001	.052
[Group2=2.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Assignment 2
44468423
Gautham Meenakshisundaram







The histogram of standardized residuals in the unadjusted model is not as symmetric as adjusted model. The normal probability plot has the standardized residuals of the unadjusted model distributed closely around the linear line but not on the line as adjusted model.
The unstandardized predicted values vs unstandardized residuals of the unadjusted model plot does not show a random scatter, as in the case of adjusted model.

The adjusted model seems to be better in determining the treatment effect on the length of hospital stay, when compared to unadjusted model.

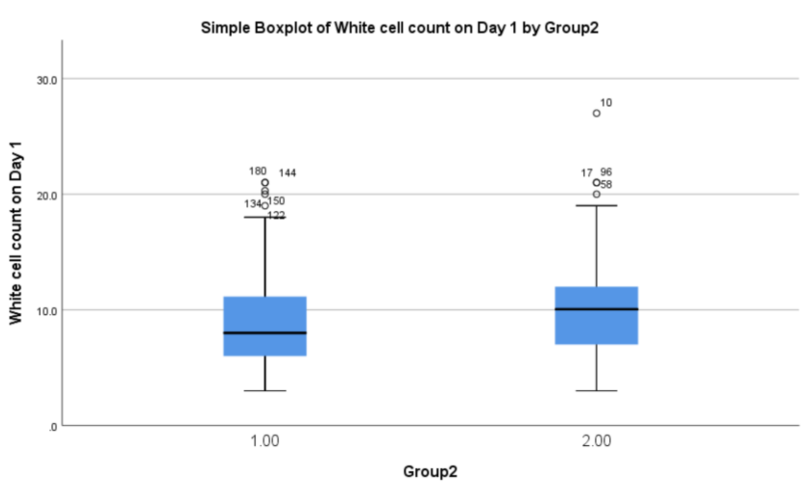
1f) Does the PEP treatment have an effect on hospital stay? Compare your conclusions with that obtained by Zhang et al (2015), and discuss.
The mean days of stay in hospital for the PEP treated group Exp is 10.69 compared to 13.29 mean days for the untreated Con group which confirms the positive effect that PEP device has on the length of hospital stay.

The above observations are in accordance with the conclusions obtained by Zhang et al (2015).

2a) Is there evidence of baseline imbalance? (Examine graphical as well as numerical evidence.)

Baseline variable = WhitecellcountonDay1
Outcome variable = Whitecellcount

Graphical Evidence



There is no increasing or decreasing trend between group Exp (1) and group Con (1), indicating that they are not significantly different. The distribution of baseline variable values is same across these groups.

Numerical Evidence

Group Statistics

	Group2	N	Mean	Std. Deviation	Std. Error Mean
White cell count on Day 1	1.00	99	9.255	4.1839	.4205
	2.00	104	10.332	4.2413	.4159

$\hat{B}_1 = 9.26$
$\hat{B}_2 = 10.33$

where,
$\hat{B}_1$ = mean of baseline white cell count from group Exp (1) who are facilitated with PEP device
$\hat{B}_2$ = mean of baseline white cell count from group Con (2) who are not facilitated with PEP device

The mean baseline white cell count from both groups appears to be not significantly different $\hat{B}_1 = \hat{B}_2$.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
White cell count on Day 1	Equal variances assumed	.005	.946	-1.821	201	.070	-1.0772	.5916	-2.2438	.0894
	Equal variances not assumed			-1.821	200.741	.070	-1.0772	.5914	-2.2434	.0890

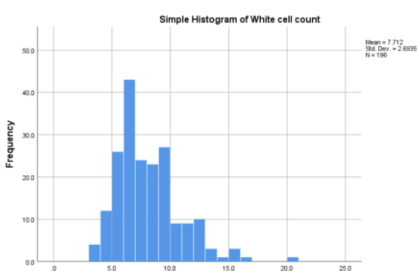
$H_0 : \hat{B}_1 = \hat{B}_2$
H_1 : not H_0

$t_{201} = -1.821$

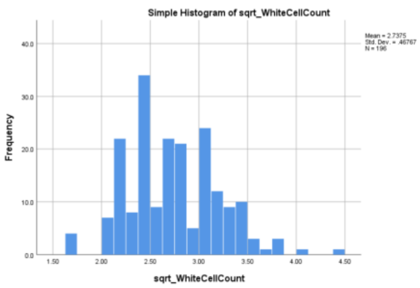
Assuming equal variances between both the groups, the two-samples independent t-test provides a p-value of 0.070 which is greater than 0.05, thereby do not reject null hypothesis H_0 and conclude that they are not significantly different $\hat{B}_1 = \hat{B}_2$.

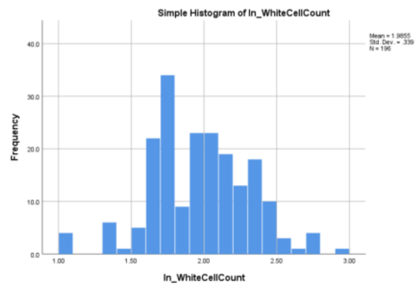
2b) Develop an appropriate statistical model for determining the treatment effect on white cell count.

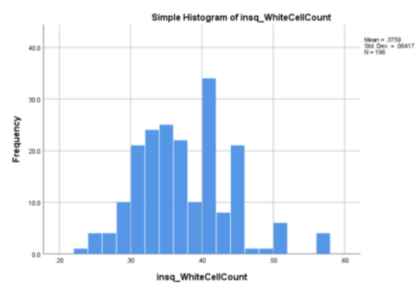
Outcome Variable:

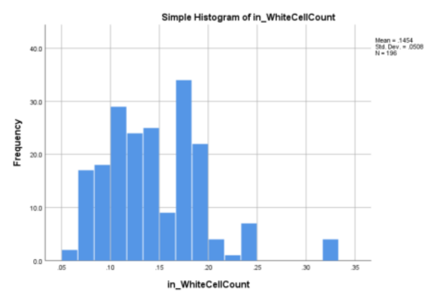


The distribution of outcome variable White Cell Count seems to be right skewed and requires transformation.





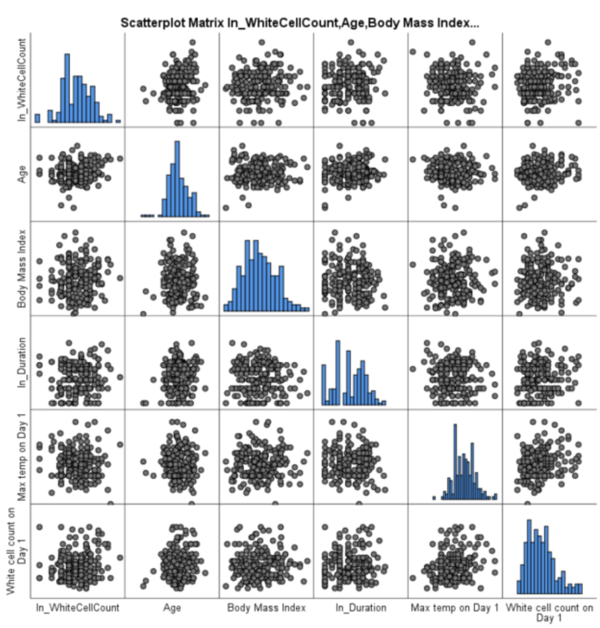




The distribution of log transformation seems to be the most symmetric while the other transformations seems to be right skewed.

Continuous Predictor Variables:
The exploratory data analysis for all the continuous predictor variables have been performed already in 1(d).

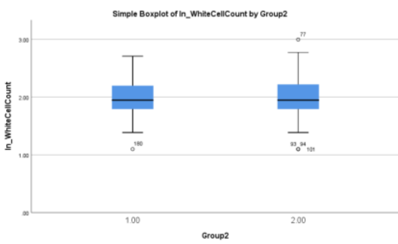
All Continuous Variables:

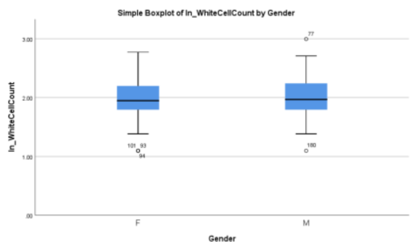


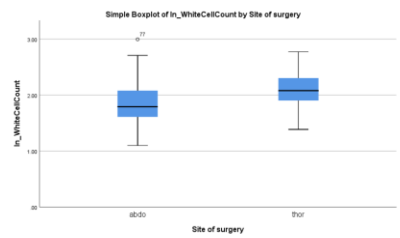
All the predictor variables show a random scatter against the transformed outcome variable. Also, there does not seem to be any collinearity between the predictor variables.

Categorical Predictor Variables:
The exploratory data analysis for all the categorical predictor variables have been performed already in 1(d).

Outcome Variable vs Categorical Predictor Variables:







There is no increasing or decreasing trend between any of the categories in Group and Gender against the transformed outcome variable, indicating that the categories are not significantly different.

However, the categories in Site of Surgery show an increasing linear trend, indicating a slight association with the transformed outcome variable. The distribution of the transformed outcome variable values is different across the categories of Site of Surgery.

Continuous Predictor Variables vs Categorical Predictor Variables:
The exploratory data analysis for continuous vs categorical predictor variables have been performed already in 1(d).

Single Variable Regressions:

Model	**p-value**	**Significance**
In(WhiteCellCount) ~ Group	0.990	> 0.05 (not significant)
In(WhiteCellCount) ~ Gender	0.137	> 0.05 (not significant)
In(WhiteCellCount) ~ Age	0.050	**= 0.05 (significant)**
In(WhiteCellCount) ~ BodyMassIndex	0.199	> 0.05 (not significant)
In(WhiteCellCount) ~ SiteofSurgery	p<0.001	**< 0.05 (significant)**
In(WhiteCellCount) ~ Log(Durationofsurgery)	0.287	> 0.05 (not significant)
In(WhiteCellCount) ~ MaxTemp	0.245	> 0.05 (not significant)
In(WhiteCellCount) ~ WhiteCellCountonDay1	0.033	**< 0.05 (significant)**

Multiple Variable Regressions:
The significant variables from single variable regression models are included together in these multiple variable regression models.

Model	**p-values**
In(WhiteCellCount) ~ Group + Age + SiteofSurgery + WhiteCellCountonDay1	Group=0.692, **Age=0.076**, Siteofsurgery=p<0.001, WhiteCellCountonDay1=0.023
In(WhiteCellCount) ~ Group + Siteofsurgery + WhiteCellCountonDay1	Group=0.721, Siteofsurgery=p<0.001, WhiteCellCountonDay1=0.013

As advised, Group is included in the model for determining the treatment effect, despite its non-significant p-values.

Final Model

Tests of Between-Subjects Effects

Dependent Variable: ln_WhiteCellCount

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3.402[a]	3	1.134	11.457	.000
Intercept	96.293	1	96.293	972.700	.000
Group2	.013	1	.013	.128	.721
Siteofsurgery	2.874	1	2.874	29.034	.000
WhitecellcountonDay1	.629	1	.629	6.356	.013
Error	19.007	192	.099		
Total	795.103	196			
Corrected Total	22.410	195			

a. R Squared = .152 (Adjusted R Squared = .139)

Parameter Estimates

Dependent Variable: ln_WhiteCellCount

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	1.961	.068	28.775	.000	1.826	2.095
[Group2=1.00]	.016	.045	.358	.721	-.073	.105
[Group2=2.00]	0[a]	.	.	.	.	.
[Siteofsurgery=abdo]	-.242	.045	-5.388	.000	-.331	-.154
[Siteofsurgery=thor]	0[a]	.	.	.	.	.
WhitecellcountonDay1	.014	.006	2.521	.013	.003	.025

a. This parameter is set to zero because it is redundant.

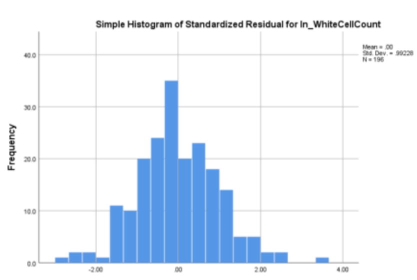
Influential Observations

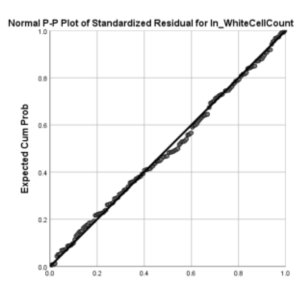
There are no leverage values more than 0.06 which does not satisfy the required moderate leverage values of 0.2 – 0.5 and high leverage values of > 0.5.

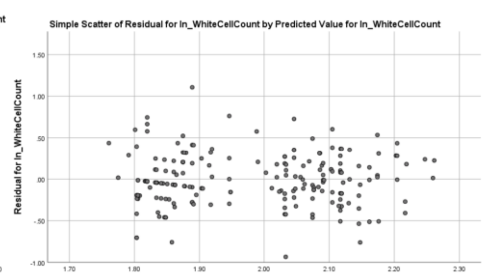
There are no Cook's distance values greater than 0.14 which satisfies the Cook's benchmark value of 1.

To conclude, there doesn't seem to be any influential observations.

2b)(i) Perform diagnostic checking for your final model (examine residuals only);







The histogram of standardized residuals seems to be symmetric.
The normal probability plot has the standardized residuals distributed closely on the linear line, indicating linearity.
The unstandardized predicted values vs unstandardized residuals plot shows residuals in a random scatter around the zero line.

2b)(ii) Write down your final model;

Parameter Estimates

Dependent Variable: ln_WhiteCellCount

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	1.961	.068	28.775	.000	1.826	2.095
[Group2=1.00]	.016	.045	.358	.721	-.073	.105
[Group2=2.00]	0[a]	.	.	.	.	.
[Siteofsurgery=abdo]	-.242	.045	-5.388	.000	-.331	-.154
[Siteofsurgery=thor]	0[a]	.	.	.	.	.
WhitecellcountonDay1	.014	.006	2.521	.013	.003	.025

a. This parameter is set to zero because it is redundant.

$\ln(\hat{y}_i) = 1.961 + 0.016\, x_{i1} - 0.242\, x_{i2} + 0.014\, x_{i3}$

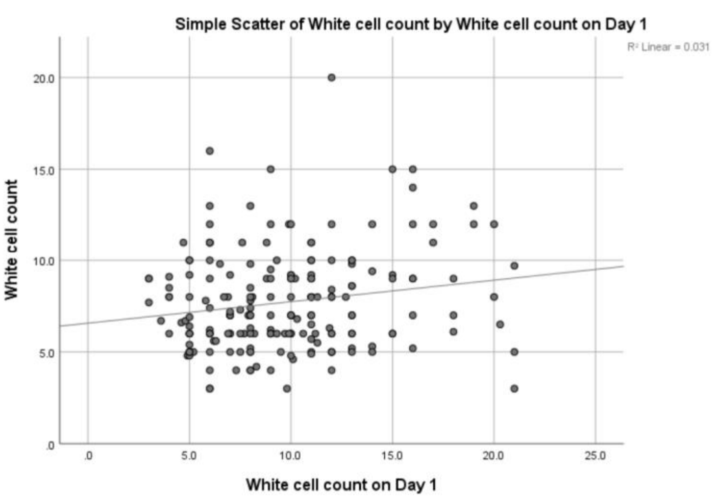
where, for each i,

x_{i1} = {1 if Group = 1 (Exp), 0 if Group = 2 (Con)}
x_{i2} = {1 if Siteofsurgery = abdo, 0 if Siteofsurgery = thor}
x_{i3} = White cell count on Day 1
$\hat{y}_i$ = White cell count on Day 5

For this model, the reference category in Group is 2 (Con) and Siteofsurgery is thor.

For every increase of (x 10^9/L) in the white cell count on day 1 for a patient whose site of surgery is abdomen and is part of Group 1 (Exp), the white cell count on day 5 increases by 0.014 (x 10^9/L) white cells, when compared to patients in Group 2 (Con).

2b)(iii) Is the treatment effective in reducing white cell count?



Regression Line:
y = x
WhiteCellCountonDay5 = WhiteCellCountonDay1

The scatter points are equally and randomly distributed across the regression line on both sides, indicating that there has been no effect on the white cell count from the treatment of PEP device.

2c) From Table 3, is there a treatment effect on White cell count on Day 5? Explain.

Table 3
Comparison of days of antibiotic therapy, white blood cell count, length of hospital stay, total expense of treatment and days of fever between groups.

Characteristics	Exp (n = 99) Mean (SD)	Con (n = 104) Mean (SD)	Exp – Con Mean difference (95% CI)
Length of hospital stay *(d)*	10.7 (7.1)	13.3 (8.7)	−2.6 (−4.8 to −0.4)
Antibiotic therapy *(d)*	7.23 (5.93)	8.85 (6.62)	1.61 (−0.13 to 3.36)
White cell count on Day 5 *($\times 10^9$/L)*	7.66 (2.50) [a]	7.76 (2.87) [b]	0.09 (−0.67 to 0.86)
Total expense of treatment *(RMB × 10 000)*	2.07 (1.67)	2.44 (1.68)	−0.38 (−0.84 to 0.09)

Exp = experimental group, Con = control group.
[a] n = 96,
[b] n = 100.

$\hat{B}_1 = 7.66$
$\hat{B}_2 = 7.76$

where,
$\hat{B}_1$ = mean of baseline white cell count from group Exp (1) who are treated with PEP device
$\hat{B}_2$ = mean of baseline white cell count from group Con (2) who are not treated with PEP device

The mean baseline white cell count from both groups appears to be not significantly different $\hat{B}_1 = \hat{B}_2$. Therefore, it can be concluded that there is no considerable treatment effect on white cell count on the 5th day.

2d) Comment on the difference between your method of analysis and that used by Zhang et al (2015) to arrive at the result given in Table 3.
The method that has been employed here is Analysis of Covariance where general linear models have been employed to check whether the means of an outcome variable are equal across the categories of predictor variables, while taking into account the effect of covariates.

Zhang et al (2015) have used Intention to Treat analysis where the observations from subjects are randomized into groups and analyzed as per the groups and not according to the treatments that the groups receive.