1. Background

(i) Recovery time (in mins) is the response variable, y which is measured between the time at which the drug was discontinued and the time at which the systolic blood pressure had returned to 100 mm Hg. We want to investigate whether the recovery time has a relation between the log of quantity of drug used and the mean level of systolic blood pressure was lowered during hypotension.[1] The two explanatory variables are

x_1 : log(quantity of drug used in mg)

x_2 : mean level of systolic blood pressure during hypotension, in mm Hg

2. Data

log_drug	*blood_pressure*	*recovery_time*	*log_drug*	*blood_pressure*	*recovery_time*
2.26	66	7	2.70	73	39
1.81	52	10	1.90	56	28
1.78	72	18	2.78	83	12
1.54	67	4	2.27	67	60
2.06	69	10	1.74	84	10
1.74	71	13	2.62	68	60
2.56	88	21	1.80	64	22
2.29	68	12	1.81	60	21
1.80	59	9	1.58	62	14
2.32	73	65	2.41	76	4
2.04	68	20	1.65	60	27
1.88	58	31	2.24	60	26
1.18	61	23	1.70	59	28
2.08	68	22	2.45	84	15
1.70	69	13	1.72	66	8
1.74	55	9	2.37	68	46
1.90	67	50	2.23	65	24
1.79	67	12	1.92	69	12
2.11	68	11	1.99	72	25
1.72	59	8	1.99	63	45
1.74	68	26	2.35	56	72
1.60	63	16	1.80	70	25
2.15	65	23	2.36	69	28
2.26	72	7	1.59	60	10
1.65	58	11	2.10	51	25
1.63	69	8	1.80	61	44
2.40	70	14			

Table 1: Blood pressure observations on the 53 patients

(i) **Data Observation**

$n = 53$	x_1	x_2	y
mean	1.992453	66.33962	22.69811
s.d.	0.3401804	7.728428	16.28542

Table 2: Mean and Standard Deviation of Data Columns

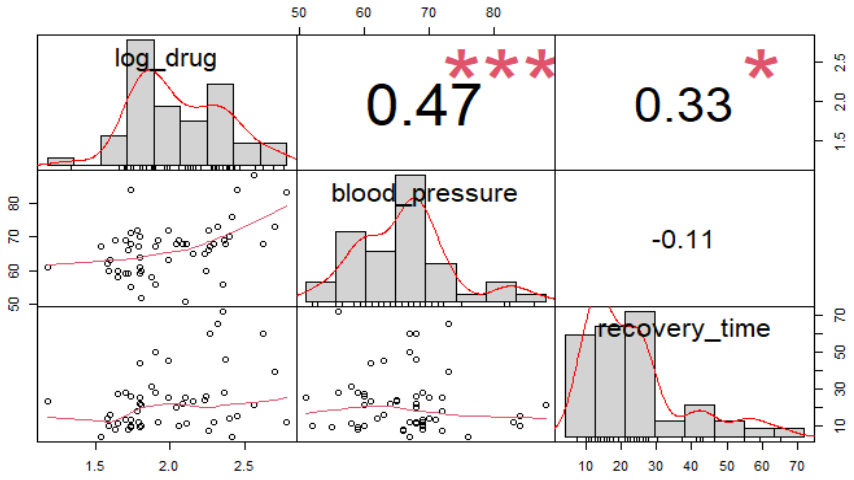


Figure 0: Scatter Plot, Histogram and Correlation Matrix.

(ii) **Data Explaination**

We computed the mean and standard deviations of each explanatory variable and response variable as seen in Table 2. The given data consist of 53 patient samples and consists of 1 response variable and 2 explanatory variables. `recovery_time` ranges between $[4, 72]$ with mean of 22.7 and standard deviation of 16.3. `log_drug` ranges between $[1.18, 2.78]$ with mean of 1.99 and standard deviation of 0.34. `blood_pressure` ranges between $[51, 88]$ with mean of 66.3 and standard deviation of 7.73. In Figure 0, the histogram shows the distribution of each column where the shape of the explanatory variables tend to follow normal distribution and there exist a skewness to the left for the response variable.

To further investigate, we computed the correlation matrix of each variables using `chart.Correlation()` function in `PerformanceAnalytics` package. There exist a positive correlation between `log_drug` and `blood_pressure` of the value of 0.47 such that the change in amount of `log_drug` and the changes `blood_pressure` during hypotension is positively related. Between the response variable and each explanatory variable, the correlation value is 0.33 and -0.11 for `log_drug` and `blood_pressure` respectively. Although there exist a negative correlation between `recovery_time` and `blood_pressure`, the value is small and we can infer that using the variable `blood_pressure` solely cannot represent the entire model. This argument also holds for `log_drug`.

3. Regression Models

Each regression model is conducted using `lm()` function in **R** and we obtained its estimate, p-value of conducted t-test and R^2 value from `summary()` function.[2] In addition, we acquired the Analysis of Variance table from `anova()` function. We did not conduct stepwise variable selection,`step()` function, as we only have two explanatory variables and as mentioned above, the correlation between response and each explanatory variable is insignificant. (i.e. We will not conduct univariate linear regression). To begin with, we will make assumptions on each linear regression model,

- Linearity of the data
- Normality of residuals
- Homogeneity of residuals variance (Homoscedasticity)
- Independence of residuals error terms

Now, let us first fit the linear regression model on the given data. The fitted model's equation is shown below,

$$\mathbb{E}(y) = 23.0107 + 23.6386x_1 - 0.7147x_2$$

the `recovery_time` increases on average by about 24 minutes for each increase of 1 in the `log_dose` and decrease by 0.71 minute for every increase of $1mmHg$ in the `blood_pressure` during hypotension.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	23.0107	18.2849	1.258	0.21407
log_drug	23.6386	6.8479	3.452	0.00114 **
blood_pressure	-0.7147	0.3014	-2.371	0.02163 *

Table 3: Summary of Model (1)

	Df	Sum Sq	Mean Sq	F-value	Pr($>F$)
log_drug	1	1545.6	1545.56	7.0202	0.01076 *
blood_pressure	1	1237.7	1237.66	5.6217	0.02163 *
Residuals	50	11007.9	220.16		

Table 4: Analysis of Variance of Model (1)

From the table above, both partial regression coefficients are significant as their `p-value`s are less than 0.05. However, the `p-value` of the intercept(β_0) is greater than 0.05 where we can reject the null hypothesis such that the estimate of the intercept may is true. Moreover, the R^2 value was recorded as 0.2018 which means that of the total sum squares of y, about 80% still present after prediction of y from x_1 and x_2.

Now, we will conduct different linear regression models through transformation of the response variable, y, `recovery_time`. The following models, Log transformation, Square root transformation, Reciprocal transformation and Exponential transformation will be discussed. In addition, due to small data set, we will not divide into training and testing set nor perform cross-validation to measure the performance.

(i) **Model 1. (Log Transformation)**
Firstly, natural logarithmic transformation was undergone on the response variable and fitted with the linear regression model. As we know the distribution of `recovery_time` is skewed, making the log transformation is reasonable. The fitted model's equation is shown below,

$$\mathbb{E}(\log(y)) = 3.09063 + 0.85789x_1 - 0.02876x_2$$

the log(`recovery_time`) increases on the average by about 0.86 for each increase of 1 in the `log_dose` and decrease by 0.03 for every increase of $1mmHg$ in the `blood_pressure` during hypotension.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	3.09063	0.79728	3.876	0.00031 ***
log_drug	0.85789	0.29859	2.873	0.00595 **
blood_pressure	-0.02876	0.01314	-2.188	0.03335 *

Table 5: Summary of Model (1)

	Df	Sum Sq	Mean Sq	F-value	Pr($>F$)
log_drug	1	0.3453	0.34526	4.3732	0.04161 *
blood_pressure	1	0.3781	0.37807	4.7887	0.03335 *
Residuals	50	3.9474	0.07895		

Table 6: Analysis of Variance of Model (1)

From the table above, both partial regression coefficients are significant as their `p-value`s are less than 0.05. Moreover, the R^2 value was recorded as 0.1549 which means that of the total sum squares of y, about 85% still present after prediction of y from x_1 and x_2.

(ii) **Model 2. (Square Root Transformation)**
On the next model, we took square root on the `recovery_time` to fit the linear regression model on the drug use and mean level of systolic blood pressure during hypotension. The fitted equation is shown below,

$$\mathbb{E}(\sqrt{y}) = 4.73363 + 2.17x_1 - 0.06868x_2$$

the square root of (`recovery_time`) increases on the average by about 2.17 for each exponential increase of 1 in the `log_dose` and decrease by 0.069 for every increase of $1mmHg$ in the `blood_pressure` during hypotension.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	4.73363	1.78935	2.645	0.01088 *
log_drug	2.17	0.67013	3.238	0.00214 **
blood_pressure	-0.06868	0.02950	-2.328	0.02397 *

Table 7: Summary of Model (2)

	Df	Sum Sq	Mean Sq	F-value	Pr(> F)
log_drug	1	12.451	12.4511	5.9057	0.01872 *
blood_pressure	1	11.431	11.4305	5.4216	0.02397 *
Residuals	50	105.417	2.1083		

Table 8: Analysis of Variance of Model (2)

From the table above, all regression coefficients are significant as their `p-value`s are less than 0.05. However, the R^2 value obtained is 0.1847 which also means that of the total sun squares of y, about 82% still present after prediction of y from x_1 and x_2.

(iii) **Model 3. (Reciprocal Transformation)**
Continuing from `Model(2)`, we applied reciprocal on the response variable `recovery_time`. Then the fitted equation is as below,

$$\mathbb{E}(\frac{1}{y}) = 0.0414879 - 0.0436053x_1 + 0.0017345x_2$$

the reciprocal of `recovery_time` increases on the average by about 0.04 for each decrease of 1 in the `log_dose` and increase by 0.0017 minute for every increase of $1mmHg$ in the `blood_pressure` during hypotension. Thus the sign of the coefficient of the explanatory variables are different in this model.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.0414879	0606529	0.684	0.4969
log_drug	-0.0436053	0.0227051	-1.921	0.0605
blood_pressure	0.0017345	0.0009994	1.736	0.0888

Table 9: Summary of Model (3)

	Df	Sum Sq	Mean Sq	F-value	Pr(> F)
log_drug	1	0.00380	0.0038004	1.5702	0.21601
blood_pressure	1	0.00729	0.0072899	3.0120	0.08881
Residuals	50	21.0152	0.4203		

Table 10: Analysis of Variance of Model (3)

From Table 9, none of the regression coefficients are significant as their `p-value`s are less than 0.05. This is critical as none of the explanatory variables can explain the fitted model well. Furthermore, the R^2 value obtained was recorded as 0.08395 and this shows that about 92% of the total sum squares of y still present after prediction of y from x_1 and x_2. Thus, the model's performance is expected to be poor in our context.

(iv) **Model 4. (Exponential Transformation)**
Lastly, we conducted a exponential transformation on the regression by taking the exponential on the `recovery_time` The fitted equation is shown below,

$$\mathbb{E}(\exp(y)) = 3.044 * 10^{30} + 2.261 * 10^{30} x_1 - 1.085 * 10^{29} x_2$$

the exponential of `recovery_time` increases on the average by about $2.261 * 10^{30}$ for each increases of 1 in the `log_dose` and decrease by $1.085 * 10^{29}$ for every increase of $1mmHg$ in the `blood_pressure` during hypotension.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	3.044e+30	3.034e+30	1.003	0.3205
log_drug	2.261e+30	1.136e+30	1.990	0.0521
blood_pressure	-1.085e+29	0.002241	-2.169	0.0348 *

Table 11: Summary of Model (4)

	Df	Sum Sq	Mean Sq	F-value	Pr($>F$)
log_drug	1	7.3517e+60	7.3517e+60	1.2129	0.27603
blood_pressure	1	2.8526e+61	2.8526e+61	4.7063	0.03483 *
Residuals	50	3.0306e+62	6.0612e+60		

Table 12: Analysis of Variance of Model (4)

Lastly, for the case of `Model(4)`, the `p-value`s for `blood_pressure` regression coefficient is less than 0.05 and we can conclude that the coefficients is significant. However, the `p-value` of `log_drug`, there is not significant evidence that the estimate is correct and represents the model. Looking at the R^2 values, it is 0.1059 and we see that 90% of the variation in the outcome parameter assessed is unexplained by the model. Referring back to equation, the interpretability of the model is considerably low as it is measured by each increase of the exponential and the coefficients consists of huge decimals.

4. Evaluation

In this report, we conducted fitting on 4 different models with different transformation. Due to the nature of the clinical data, we may not have a strong model with high accuracy that can represent the entire dataset. However, we want to evaluate the performances and validate our assumptions stated above on each model by comparing the values and plots obtained. [3]

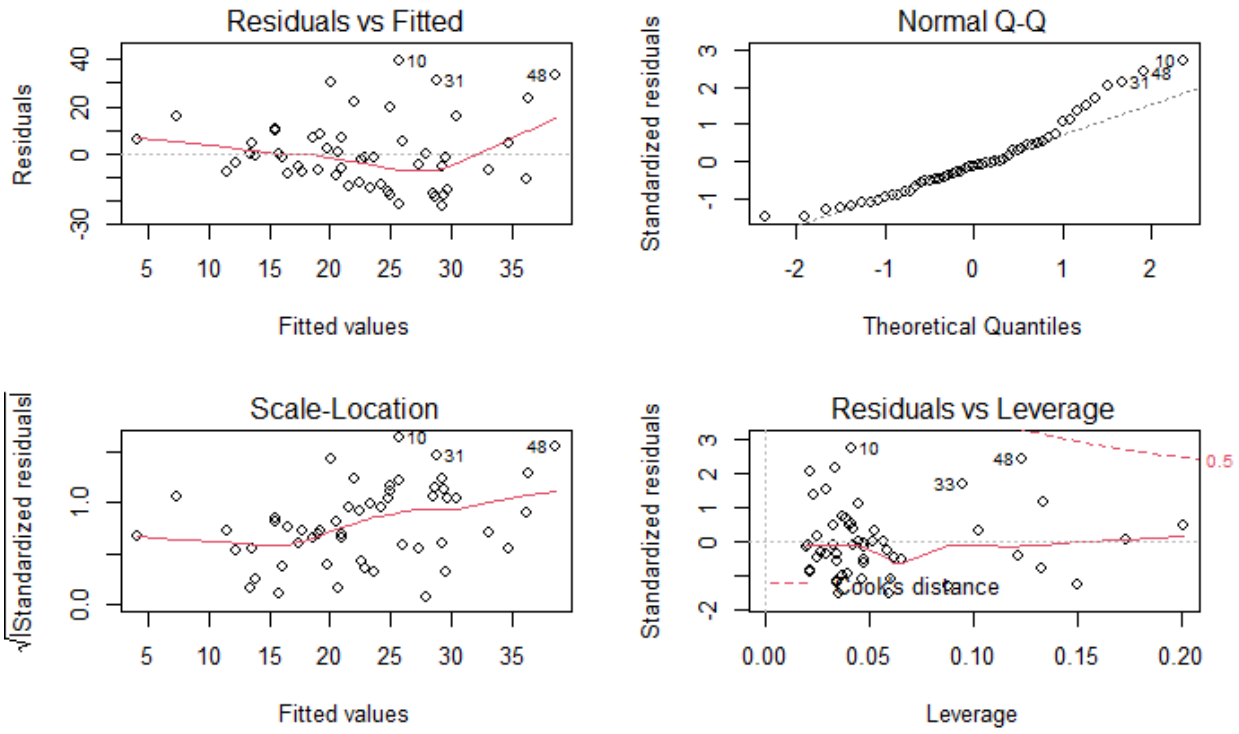


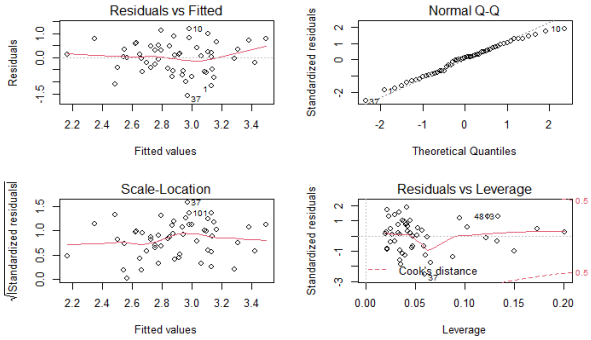
Figure 1: Residual, Normal Q-Q, Scale-Location and Residual vs Leverage plots of a model before transformation

Looking at the residuals plot, the residuals appear to be randomly scattered around the zero line without a distinctive pattern. However, the mean of the residuals (red line) seem to have a increasing trend after 30. Referring to Figure 0, 2 scatter plots at the last row show that there is no distinctive linear relationship between `recover_time` and the explanatory variables, `log_drug` and `blood_pressure` and the residual plot also shows that there is no noticeable linear relationship.

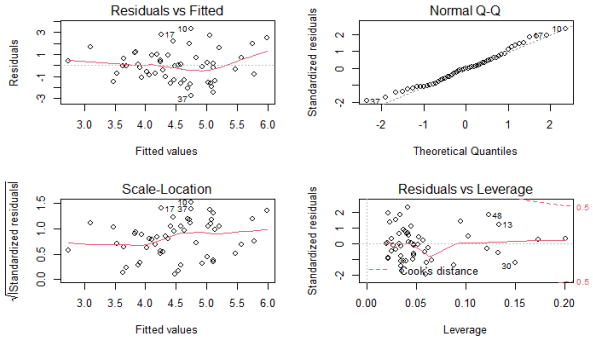
Assessing the normality assumption, the values on the right end of the plot tend to deviate away from the straight line. However, we can conclude that this generally follow the straight line. Looking at the scale-location plot, the spread around the red line does not vary greatly with the fitted values. However, the red line seems to not have a horizontal shape as it increases after 15. Thus, we cannot determine the homoscedasticity. All the points in the

Residuals vs Leverage plot seems to be under the Cook's distance and there exist no influential case.

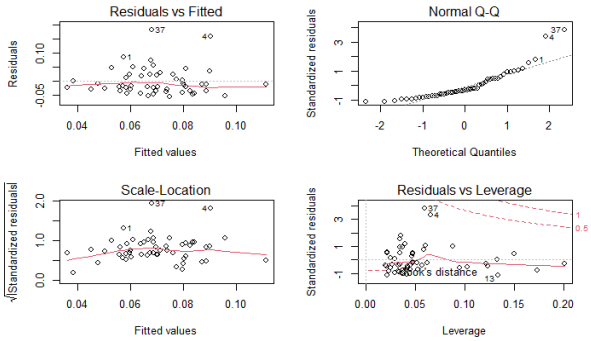
Overall, the linear model before transformation seems to not satisfy all our assumptions and we want to choose a model that best satisfies the assumptions we stated above. Now we will discuss the plots of the transformed models.



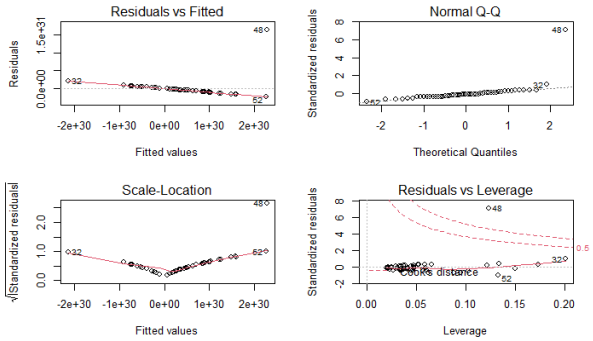
(a) `Model(1)`



(b) `Model(2)`



(c) `Model(3)`



(d) `Model(4)`

Figure 2: Residual, Normal Q-Q, Scale-Location and Residual vs Leverage plots of each model

(i) **Residual Plot**
Referring to Figure 2 above, the general shape of the zero line in the residual plots of `Model(1)` and `Model(2)` are similar. On the other hand, `Model(4)` shows a distinctive decreasing trend with a critical outlier. Thus, we can say that the model does not has a linear relationship between the response variable and explanatory variables. For the rest of the three models we can then conclude that the residuals appear to be randomly scattered around the zero line with not distinctive pattern and assume the linearly relationship holds. To be precise, we can observe that `Model(3)` has the most straight red line that lies above zero line and there is no fitted pattern and `Model(2)` has a slight increasing trend at the end the red line.

(ii) **Normal Q-Q Plot**
To check normality assumption on the residuals, we look at Normal Q-Q plot. The

residuals generally follow the straight line on `Model(1)` and `Model(2)`. These two models show best fit to the straight line where we can say that residuals are normally distributed and the normality assumption holds. Both `Model(3)` and `Model(4)` shows a sudden spike in the values at the tail. Hence, the normality assumption is violated in `Model(3)` and `Model(4)`

(iii) **Scale-Location Plot**
To check if the residuals are equally spread along the ranges of predictors, we look at the Scale-Location plot. We can see sufficient variation among the plots except for `Model(4)`. In `Model(4)` we see that there is a noticeable trend and this may violate the constant variance. On the other hand, we can see sufficient variation among the plots in `Model(1)` as the red line has the most straight fitting.

(iv) **Residuals vs Leverage Plot**
The Cook's distance lines are not seen in a large portion within the plots except for `Model(4)`. This is expected in `Model(4)` as there is critical outlier that affects the model. From this, we can conclude that there is no influential case and no outliers that affect the model's performance in `Model(1)`,`Model(2)` and `Model(3)`

Continuing from the analysis, we identified an outlier that influences `Model(4)` and removed the outlier(48). We expected the model to perform better but the model had another noticeable outlier(10) which also resulted in a poor performance. If we iterate this by dropping single outlier one by one, the reliability of the outcome of `Model(4)` becomes smaller as we do note have a large dataset. Our focus is not to optimise the performance of `Model(4)` hence, we will leave this for future discussion

Referring to Section 3, Models' partial coefficient `p-value`s vary greatly depending on the choice of model. For `Model(1)` and `Model(2)` we can conclude that each predictor parameter contributes separately to the effectiveness of the regression model as their partial regression coefficients are significant.

	`Model(1)`	`Model(2)`	`Model(3)`	`Model(4)`
R^2	0.1549	0.1847	0.08395	0.1059

Table 13: R^2 Values of each Model

Considering the fact that we are fitting a model in the context of clinical data, R^2 value is not high as each patients have different gene thus different reaction towards drugs. Despite the fact that the model before transformation had R^2 value of 0.2018 and `Model(2)` having the value of 0.1847, comparing among low R^2 values may not lead to critical decision of model selection. Since logarithmic transformation normalises skewed frequency distributions, we believe `Model(1)` is the best model. Using our previous analysis, `Model(1)` with log transformation seems to have satisfied our key assumptions as the residual plots and normal

Q-Q plots shows linearity of data and normality of residuals respectively, It also satisfies the homoscedasticity and independence of residual error terms shown in scale-location plot and residual vs leverage plot.

5. Conclusion

To sum up, we conducted 4 different regression models and compared each model to find the best fit to our given data. Both `Model(1)` and `Model(2)` have met the criteria to keep our assumptions. For our decision, we chose `Model(1)` (Log Transformation)

$$\mathbb{E}(\log(y)) = 3.09063 + 0.85789x_1 - 0.02876x_2$$

and computed its prediction interval when the value of `log_drug` $= 2$ and `blood_pressure` $= 75$ using `predict.lm()` function. Noting that `Model(1)` undergoes log transformation on the response variable, we took the exponential on the prediction interval compute the prediction of `recovery_time`. Therefore, we get the `recovery_time` $= 14.14435$ with a 95% prediction interval of $[3.736709, 53.5398]$. In words we can interpret as approximately 14 minutes of recovery time was taken between the time at the stop of drug administrating and the time at which the blood pressure returned to the desired level.

References

[1] Peter Armitage, Geoffrey Berry, and John Nigel Scott Matthews. *Statistical methods in medical research*. John Wiley & Sons, 2008.

[2] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. CRC press, 2018.

[3] Bommae Kim. *Understanding Diagnostic Plots for Linear Regression Analysis*. 2015. URL: https://data.library.virginia.edu/diagnostic-plots/.

Appendix(Code)

(A) `Correlation Matrix`, (§2)

```
library(PerformanceAnalytics)
bloodpressure <- read.table('data/bloodpressure.txt',head=T)
chart.Correlation(bloodpressure)
```

(B) `Model Fitting`, (§3 & §4)

```
# Model (before transformation)
model <- lm(recovery_time) ~ log_drug +
            blood_pressure, data = bloodpressure)
summary(model)
anova(model)
layout(matrix(c(1,2,3,4),2,2))
plot(model)

# Model. 1
modellog <- lm(log(recovery_time) ~ log_drug +
                blood_pressure, data = bloodpressure)
summary(modellog)
anova(modellog)
layout(matrix(c(1,2,3,4),2,2))
plot(modellog)

# Model. 2
rtreci <- 1/bloodpressure$recovery_time
modelreci <- lm(rtreci ~ log_drug +
                blood_pressure, data = bloodpressure)
summary(modelreci)
anova(modelreci)
layout(matrix(c(1,2,3,4),2,2))
plot(modelreci)

# Model. 3
modelsqrt <- lm(sqrt(recovery_time) ~ log_drug +
                blood_pressure, data = bloodpressure)
summary(modelsqrt)
anova(modelsqrt)
layout(matrix(c(1,2,3,4),2,2))
plot(modelsqrt)
```

```
# Model. 4
modelexp <- lm(exp(recovery_time) ~ log_drug +
                blood_pressure, data =bloodpressure)
summary(modelexp)
anova(modelexp)
layout(matrix(c(1,2,3,4),2,2))
plot(modelexp)
```

(C) `Prediction Interval on Model(1)`, (§5)

```
newdata <- data.frame(log_drug = 2, blood_pressure = 75)
pi <- predict(modellog, newdata = newdata, interval = "prediction")
exp(pi)
```