Solutions:

1. **Summarize the problem they studied, the methods they used, the results they obtained, their recommendations for statistical practice, and how you will apply the recommendations in the future if conducting a two-sample t-test.**

This paper addresses primarily 2 questions i.e. the need if preliminary variance test and emphasizing to equal variances assumption in a two sample mean tests. Previously to measure the quality of means from two independent normally distributed populations has achieved under the assumption that "the two population variances are equal"
They have considered 2 independent random samples which are normally distributed. So based on initial assumption preliminary test

The preliminary test for H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ is to calculate

$$F' = s_2^2/s_1^2 \qquad (1)$$

For the emphasis on variance homogeneity, they have concluded that variance ratios influences for making choice of test procedures with the condition if variance ratio is already known.
Their SWS test can also be appropriate even if the variance ratio is unknown.

$H_a: \mu_1 - \mu_2 \neq Hypothesized\ Difference$

$H_a: \mu_1 - \mu_2 < Hypothesized\ Difference$

$H_a: \mu_1 - \mu_2 > Hypothesized\ Difference$

t = (statistic - parameter | H_0) / $SE_{statistic}$
So while doing two-sample means test, effort should be given on learning different qualities of SWS test and homogeneity of variance should be less emphasized.

Future Recommendations to conduct two sample t test will be-

- Less importance to variance homogeneity
- Learn qualities of SWS test

2a

```
model <- summary( lm(chol ~ wtkg, data=hw7))

model
Call:
lm(formula = chol ~ wtkg, data = hw7)

Residuals:
    1     2     3     4     5     6     7
-55.86  10.14 -44.22 -56.50  10.78  70.32  65.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.395     89.782   1.085   0.3275
wtkg           3.727      1.340   2.781   0.0388 *
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.5 on 5 degrees of freedom
Multiple R-squared: 0.6074, Adjusted R-squared: 0.5289
F-statistic: 7.735 on 1 and 5 DF, p-value: 0.03884

Variable	Estimate	SE	T value	Pr >|t|	Lower 95% CI	Upper 95% CI
Intercept	97.395	89.782	1.085	0.3275	-133.395	328.186
Weight (kg)	3.727	1.340	2.781	0.0388	0.282	7.172

```
1 confint(model)
```

	2.5 %	97.5 %
(Intercept)	-133.3957971	328.186461
hw7$wtkg	0.2823573	7.172412

```
1 confint(model, level=0.9)
```

	5 %	95 %
(Intercept)	-83.519075	278.309739
hw7$wtkg	1.026869	6.427901

2B

y = b0 + b1(X) + e

Y-hat = (the estimate for the intercept was) + (the estimate for the slope was)*X

y = 97.395 + 3.727X

2C

The regression equation can be used to estimate a 7 participant's total cholesterol as a function of his/her BMI.

Here BMI can be calculated as Weight/ (height)^2

Example, suppose a participant has a BMI 30

Estimated intercept is 97.395

We would estimate their total cholesterol to be 97.395 + 3.727(30) = 209.205

Average cholesterol if weight is zero kilograms (probably not at all interpretable)

2D

```
1 confint(model)
```

	2.5 %	97.5 %
(Intercept)	-133.3957971	328.186461
hw7$wtkg	0.2823573	7.172412

[-133.3957971, 328.186461]

we are 95% confident that cholesterol levels for persons with weight of zero

kilograms is in the confidence interval [-133.3957971, 328.186461].

2E

p-value 0.03884 i.e. $p<0.05$ Assuming the null hypothesis that weight has no effect of cholesterol, the probability of observing our effect or something more extreme is 3.884%.

But its strange that is it biologically possible that someone can have weight as 0

Fail to reject the hypothesis that the true intercept is 0, based on our

confidence interval [-133.3957971, 328.186461], which contains 0.

2F

Estimated slope: 3.727

The slope tells us how much increase there will be in the y direction for a one unit increase in the x direction.

We are 95% confident that cholesterol increases, on average, between 0.282 and 7.172 mg/100mL for a 1kg increase in weight.

For a unit increase in weight, for persons with weight between 30 and 84 kilograms, average cholesterol increases by 3.727 mg/100ml.

2G

```
1 confint(mod1lm,'wtkg',level=0.95)
```

	2.5 %	97.5 %
wtkg	0.2823573	7.172412

95 % confidence interval: [0.2823573, 7.172412]

Reject the hypothesis that the average slope is 0. Confidence interval does not contain zero.

2h. Test the hypothesis that the true slope is zero.

We reject the hypothesis that the true slope is 0, based on our
confidence interval[.2823573, 7.172412], which does not contain 0.

2i. Write a brief, but complete (i.e., include the point estimate, p-value, 95% CI, and summary/decision),summary of the effect of weight on cholesterol.

CI = Point estimate ± margin of error
[-133.3957971, 328.186461] → midpoint formula
i.e. [-133.3957971 + 328.186461]/2 = 194.7906639
here point estimate is 194.7 i.e. 195

We got p-value such as 0.03884. **Assuming the weight has zero effect in the population of cholesterol, we obtained the sample effect, or larger, in 3% of studies because of random sample error.**
Conventionally, the P-value for statistical significance is defined as $P < 0.05$. our use case, threshold is breached and the null hypothesis is rejected, that the population means are equal.

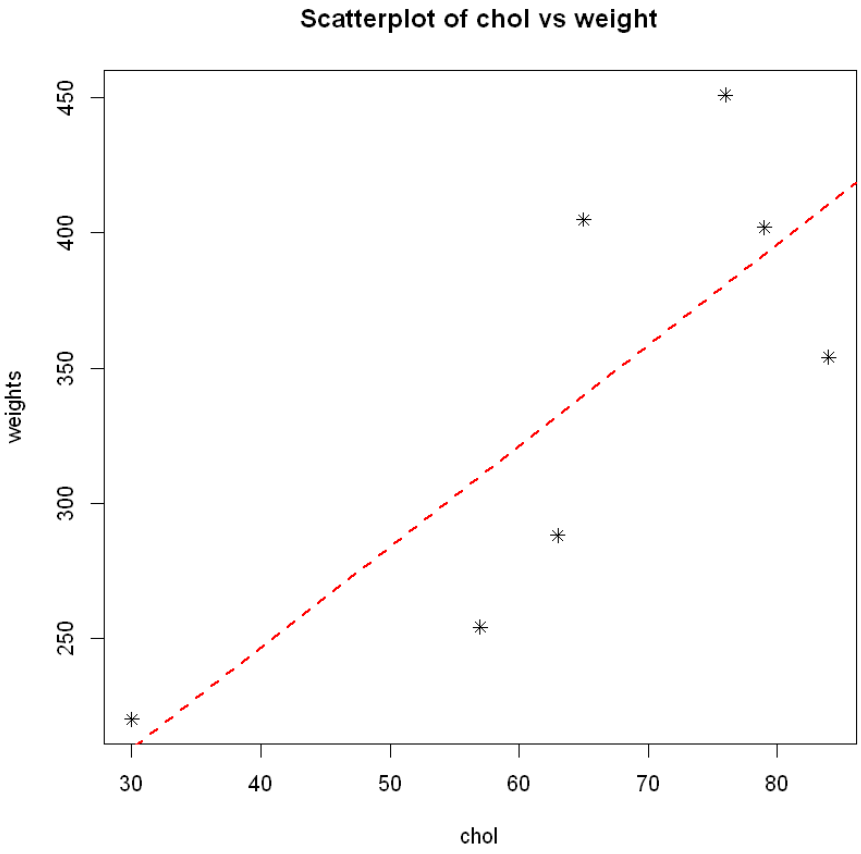
In general 95% CI means that if we have to take 100 different samples and compute a 95% confidence interval for each sample, then approximately 95 of the 100 confidence intervals will contain the true mean value (μ).

We are 95% confident that the true cholesterol for an individual with a weight of 0 kg is between -133.396 mg/100mL and 328.186 mg/100mL.
But in this context we created CI which interval may over or underestimate the mean value. CI does not show the variability of any unknown parameters.
P([sample mean] - margin of error < μ < [sample mean] + margin of error) = 0.95

Here,
we are 95% confident that the true mean is between **-133.3957971** and **328.186461.**
We are 95% percent confident that the effect weight on cholesterol levels are in between interval [-133.3957971, 328.186461].

2j.

```
plot(hw7$wtkg, hw7$chol,pch=8,
        xlab="chol",
        ylab="weights",
        main="Scatterplot of chol vs weight")
abline(coefficients(model), lwd=2, lty=2, col="red")
```



y-intercept: the value of the response variable (y) when the explanatory variable (x) is 0. It's where the least-squares regression line crosses the y-axis.

95% prediction interval:

a) for a data point

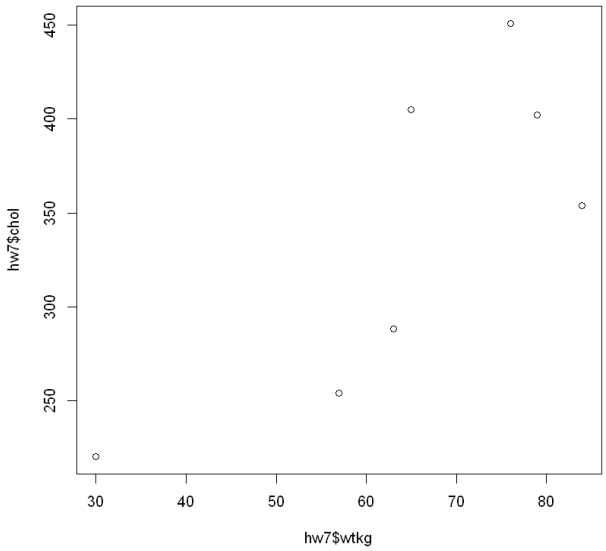
```
mod1lm <- lm(chol ~ wtkg, data=hw7) # need to use lm instead of glm
predict(mod1lm, newdata = data.frame(wtkg=50),  interval="prediction",
                          level = 0.95)

```

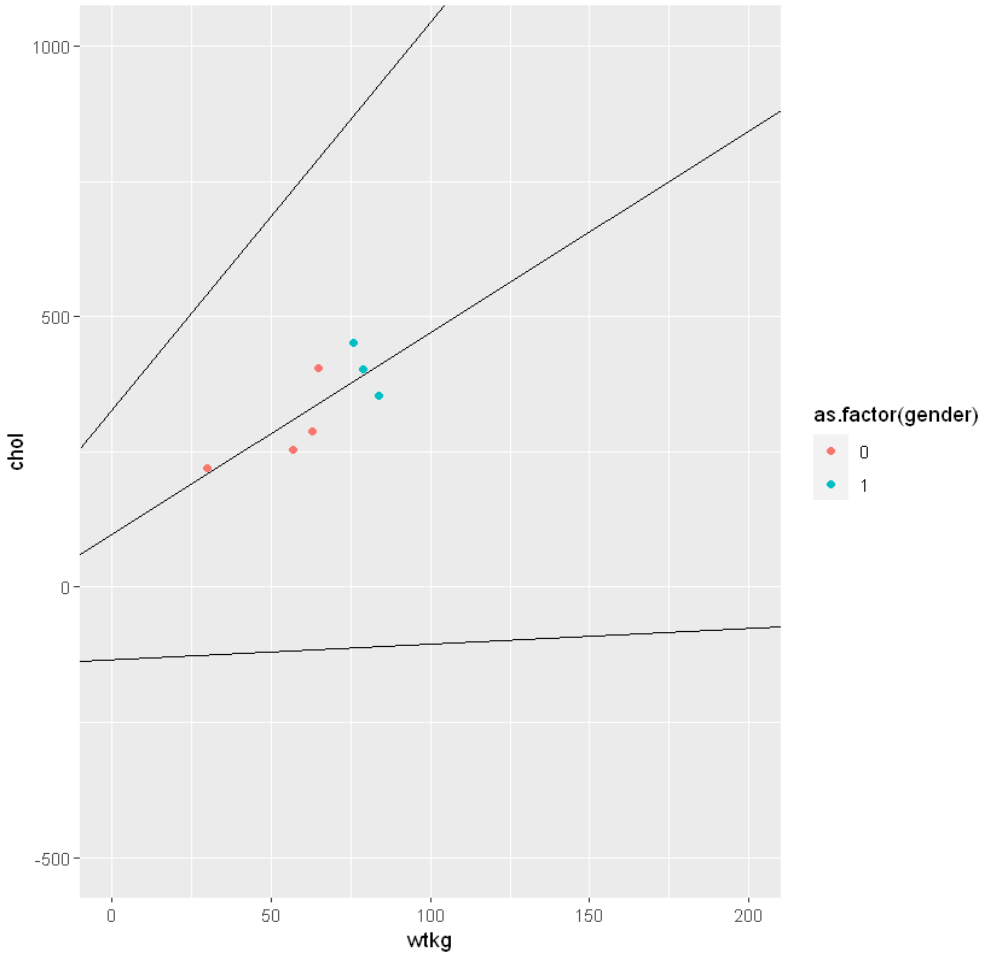
fit	lwr	upr
283.7646	112.4409	455.0882

```
confint(mod1lm,'wtkg',level=0.95)
```

	2.5 %	97.5 %
wtkg	0.2823573	7.172412

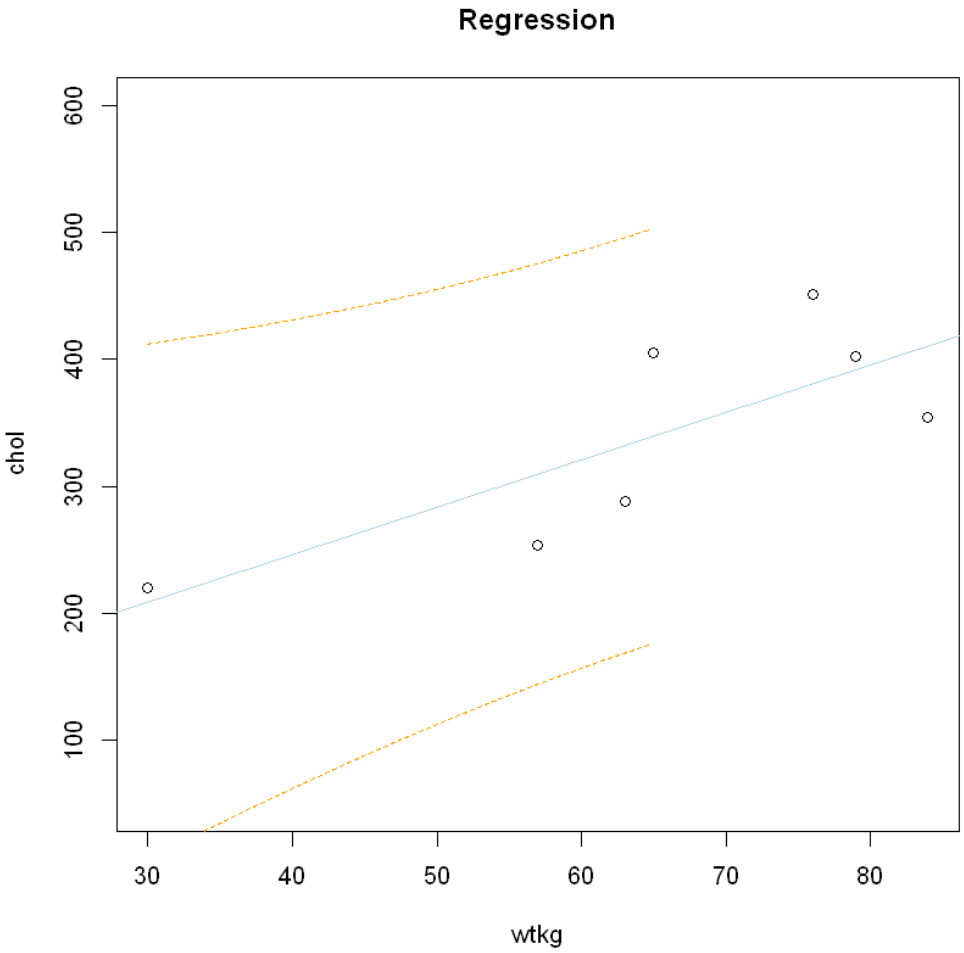
```
ggplot(data = hw7, mapping = aes(x = wtkg, y = chol, color = as.factor(gender))) +
  geom_point() +
  geom_abline(intercept = 97.395, slope = 3.727) +
  geom_abline(intercept = -133.43957971, slope = 0.2823573) +
  geom_abline(intercept = 328.186461, slope = 7.172412) +
  scale_x_continuous(limits = c(0, 200)) +
  scale_y_continuous(limits = c(-500, 1000))
```





95% prediction interval for regression

```
plot(hw7$wtkg, hw7$chol, ylim=c(50, 600), xlab="wtkg", ylab="chol", main="Regression")
abline(mod1lm, col="lightblue")

pred_interval <- predict(mod1lm, newdata=data.frame(wtkg=newx), interval="prediction",
                         level = 0.95)
lines(newx, pred_interval[,2], col="orange", lty=2)
lines(newx, pred_interval[,3], col="orange", lty=2)
```

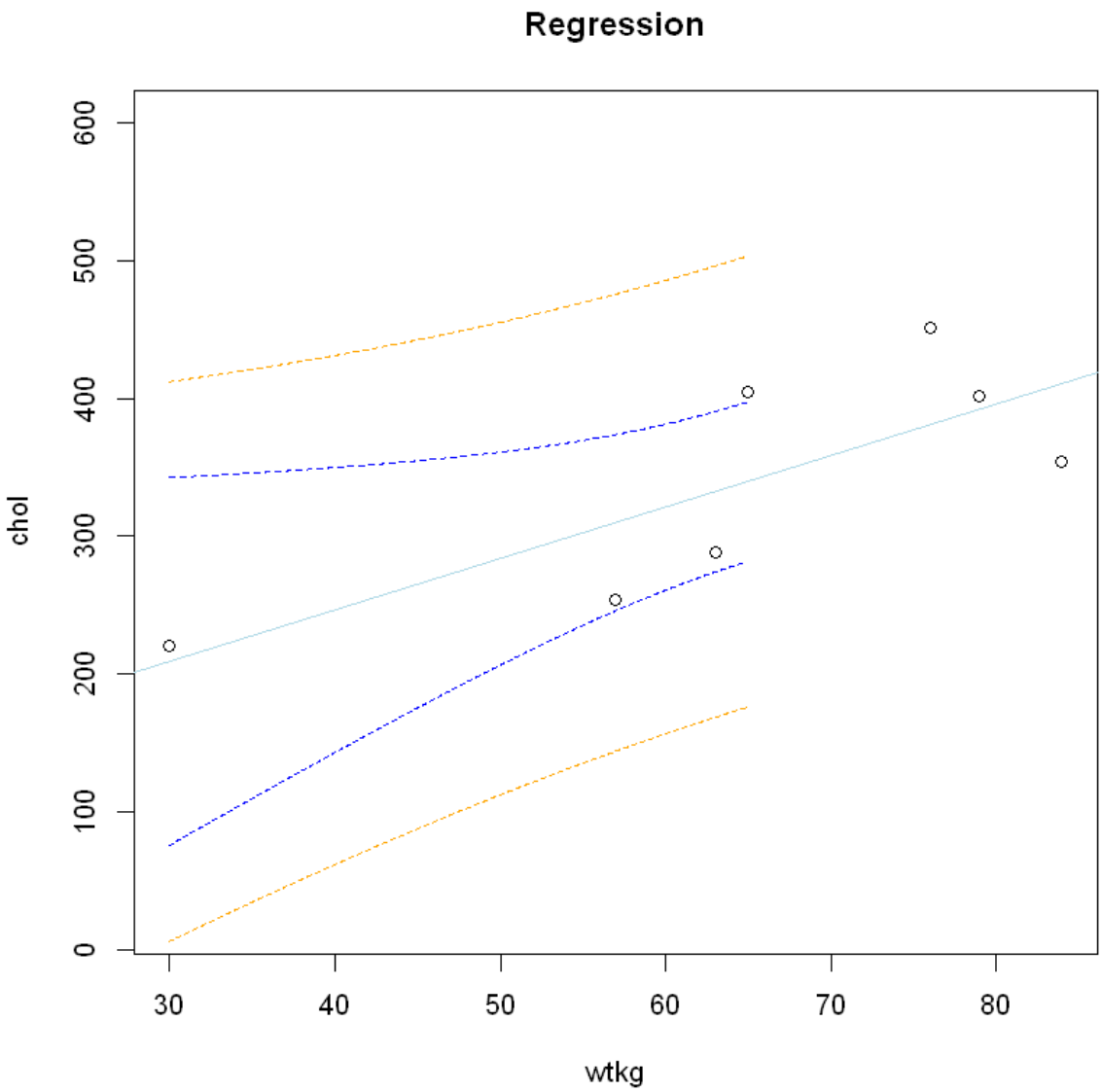


The linearity of these relationships suggests that there is an incremental risk with each additional weight in kg and the additional risk is estimated by the slopes. This perhaps helps us think about the consequences of Higher cholesterol.

The regression suggests that their risk would increase by slope: 3.727 with more than 500 as cholesterol. Females have higher chance of getting higher cholesterol rate.

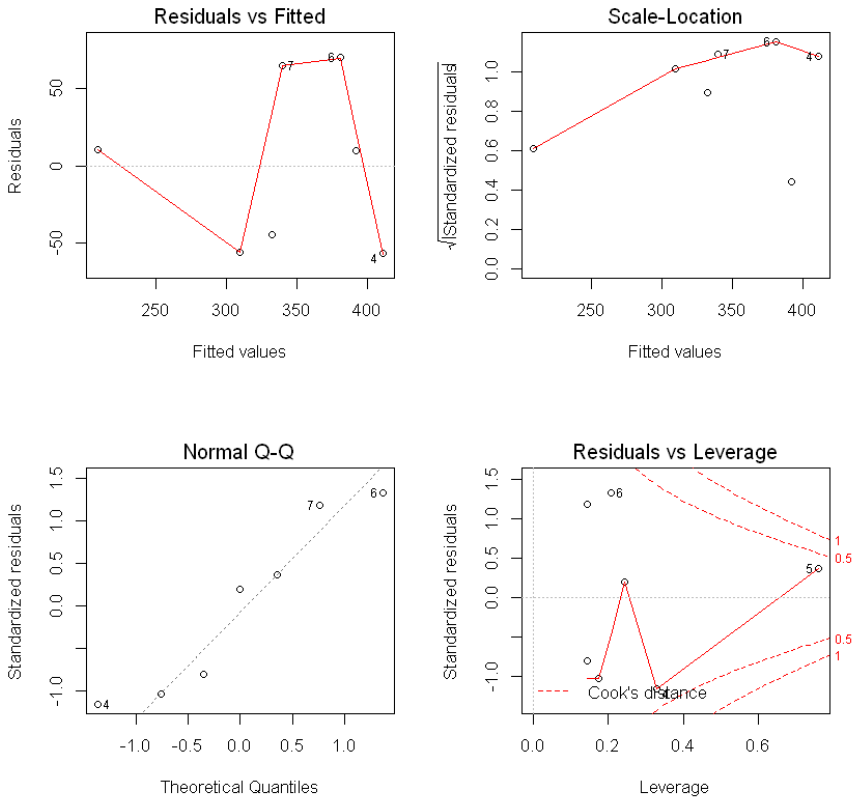
95% confidence and prediction intervals for regression

```
plot(hw7$wtkg, hw7$chol, ylim=c(20, 600), xlab="wtkg", ylab="chol", main="Regression")
abline(mod1lm, col="lightblue")

conf_interval <- predict(mod1lm, newdata=data.frame(wtkg=newx), interval="confidence",
                         level = 0.95)
lines(newx, conf_interval[,2], col="blue", lty=2)
lines(newx, conf_interval[,3], col="blue", lty=2)

pred_interval <- predict(mod1lm, newdata=data.frame(wtkg=newx), interval="prediction",
                         level = 0.95)
lines(newx, pred_interval[,2], col="orange", lty=2)
lines(newx, pred_interval[,3], col="orange", lty=2)
```



.

For my understanding I plotted Q-Q plot.



The residual errors plotted versus their fitted values, here residuals should be randomly distributed around the horizontal line as a residual error of zero. Q-Q plot, which is suggesting the residual errors are normally distributed. Scale location plot where no obvious trend found. Residuals vs Leverage shows each points leverage, which is a measure of its importance while determining any regression result.