ASSIGNMENT 4

```
> mymodelsummary

Call:
lm(formula = Minutes ~ Number.of.Copiers, data = CopierData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98570 -0.36780 -0.03733  0.40328  1.65802

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.254192   0.178413   1.425    0.161
Number.of.Copiers 0.063683   0.002046  31.123   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5801 on 43 degrees of freedom
Multiple R-squared:  0.9575,    Adjusted R-squared:  0.9565
F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16

>
```

1. What are the estimated regression coefficients?

```
> coef <- coefficients(mymodel)
> coef
    (Intercept) Number.of.Copiers
     0.25419165       0.06368338
```

Intercept estimate – is the avg value of Minutes (Y) when Number of Copiers (X) = 0.

Slope estimate - shows how much the dependent variable (Minutes) is expected to increase (since the coefficient is positive) when the independent variable (Number of Copiers) increases by one.

2. Are the coefficients significant?

Yes, the coefficients are significant.

3. Interpret the coefficients.

- We can consider a linear model to be statistically significant only when both p-values are less than the pre-determined statistical significance level, which is ideally 0.05. We can also observe this significance level from the stars at the end of the row. The more the stars beside the variable's p value, the more significant the variable.

  So after observing both the p-Values in this case (as we can see the number of stars at the end is 3 which shows that the relationship is significant, and also p value is very less than 0.05), we

can say coefficient is significant -
<2e-16 ***, p-value: < 2.2e-16

- Also, in this case, equation of given linear model is -
  Minutes = $\beta_0 + \beta_1 *$ Number.of.Copiers + ε

  We can decide whether there is any significant relationship between x (Number.of.Copiers) and y (Minutes) by testing the null hypothesis that β = 0.

  Null hypothesis - coefficients associated with the variables is equal to zero i.e. β = 0.

  Alternate hypothesis - coefficients are not equal to zero (i.e. there exists a relationship between the independent variable - Number.of.Copiers and the dependent variable - Minutes).

  Pr(>|t|) or p-value is the probability that we get a t-value as high or higher than the observed value when the Null Hypothesis (the β coefficient is equal to zero or that there is no relationship) is true.

  So, if the Pr(>|t|) is low, the coefficients are significant and
  If the Pr(>|t|) is high, the coefficients are not significant.

  So, when p value is less than significance level (< 0.05), we can safely reject the null hypothesis that the co-efficient β of the predictor is zero. In our case, both these p-Values are well below the 0.05, so we can conclude our model is statistically significant.

- The coefficient **t-value** is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between Lifetime gift amount and Number of Promotions exist.

  In our example, the t-statistic values (31.123) is relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists.

- The **R-squared** statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. The R square is a measure of the linear relationship between our predictor variable (Number of Copiers) and our response variable (Minutes). It always lies between 0 and 1 and the higher R square is, the better is fit.

In our case its 0.9575, so roughly 95.75% of the variance found in the response variable (Minutes) can be explained by the predictor variable (Number of Copiers).

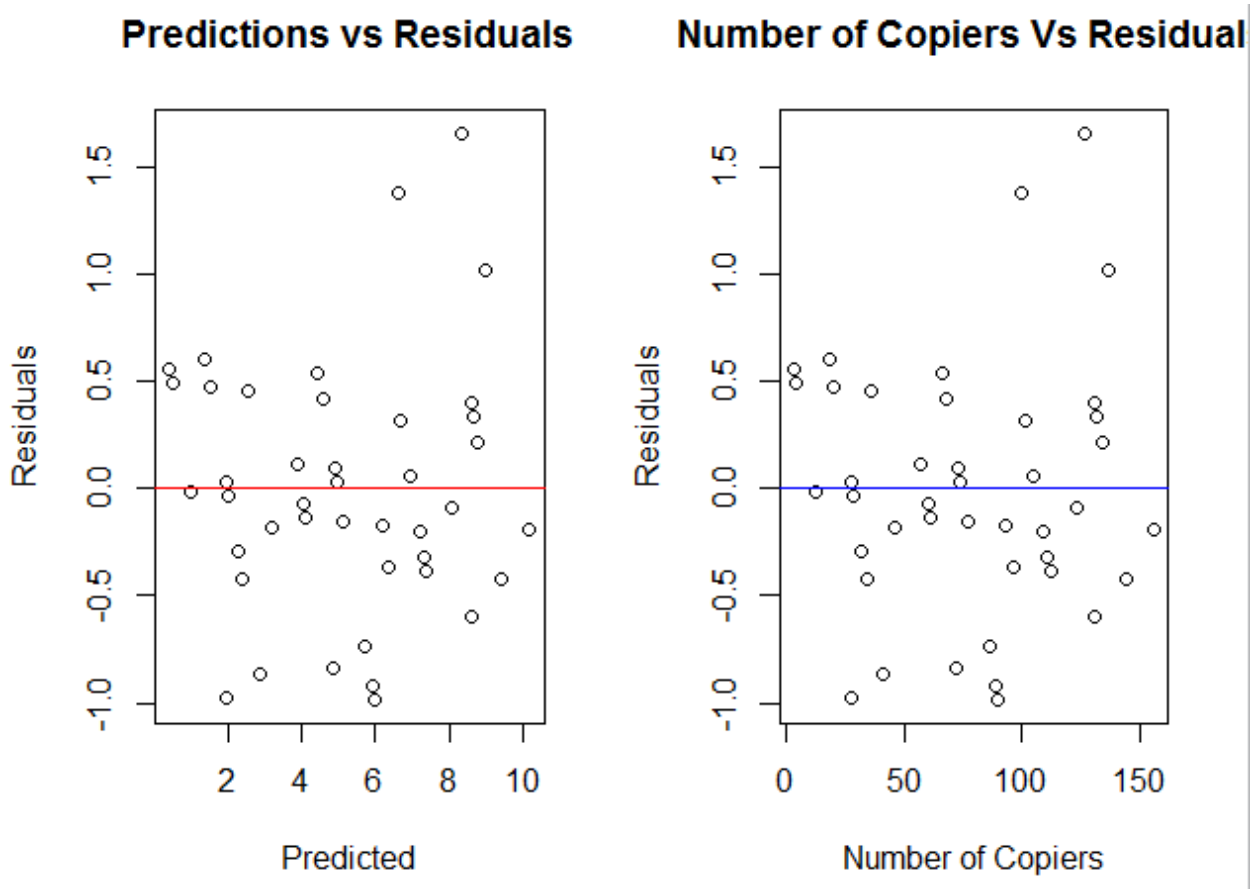
4. What is the r-square for the model?

```
> rsq <- mymodelsummary$r.squared
> rsq
[1] 0.9574955
```

5. From the residual plot, does the model appear to fit the data? Explain.

The prediction made by the model is on the x-axis, and the accuracy of the prediction is on the y-axis. The distance from the line at 0 is how bad the prediction was for that value.
Since, Residual = Observed – Predicted

Positive values for the residual (on the y-axis) mean the prediction was too low, and negative values mean the prediction was too high; 0 means the guess was correct.



As we can see residuals are randomly scattered above and below the horizontal axis, almost similar distances from horizontal axis and no clustering hence we can conclude that model fit to the data.

6. Provide the estimates using the new data along with the 99% confidence intervals.

```
> data1 <- data.frame(Number.of.Copiers = c(2, 4, 6))
> data1
  Number.of.Copiers
1                 2
2                 4
3                 6
>
> predict(mymodel, newdata = data1, interval = "prediction", level = .99)
        fit        lwr      upr
1 0.3815584 -1.2513538 2.014471
2 0.5089252 -1.1212557 2.139106
3 0.6362920 -0.9912278 2.263812
```