**Explanation**: Here, the data is only from when law was not passed to measure a better affect of Petrol Prices on the number of car drivers killed. Petrol Prices here serve as the independent variable while the Driverskilled serves as the dependent variable. Using the lm command, we generate a linear regression model. Lets interpret the results;

Residuals: The residuals are the difference between the actual values and the predicted values. We want our median to be zero. From the data we have, our distribution is slightly right skewed.

Coefficients — Estimate: In the plot below, the line where it meets the y-axis is the y-intercept (b) written next to Intercept and the slope of the line (m) is written next to Petrol Prices . Because regression model finds the line that fits the points in such a way that it minimizes the distance between each point and the line, we have now the values to formulate an equation.

*y = -634.69 (x) + 190.53*

The negative slope denotes a negative relation between Petrol Prices and DriversKilled, meaning one would go up if other one goes down and vice-versa.

Coefficients – Std. Error: The standard error of the coefficient is an estimate of the standard deviation of the coefficient. In effect, it is telling us how much uncertainty there is with our coefficient. Using standard error, we can create Confidence Intervals. Lets say we create an interval of 95% then it would be as following:

-634.69 +- 1.96 (149.51) = -341.65 , -927.73

Hence we are 95% sure the slope falls in this range.

Coefficients – t value: The t-statistic is simply the coefficient divided by the standard error. In general, we want our coefficients to have large t-statistics, because it indicates that our standard error is small in comparison to our coefficient. The larger our t-statistic is, the more certain we can be that the coefficient is **not** zero. The t-statistic is then used to find the p-value.

Coefficients – Pr(>|t|) [p values] and Signif. Codes: The p-value is calculated using the t-statistic from the T distribution. The p-value, in association with the t-statistic, help us to understand how *significant* our coefficient is to the model. If p-value is below the significance level (usually 0.05), then it is significant meaning the coefficient is **not zero.** As we can see in the data, the p-value for both the intercept and the independent variable Petrol Price is very low denoting that both are significant and **non-zero.** However, the significance of the y-intercept or otherwise does not make much of a difference but excluding it when it is not significant will lead to a model where the dependent variable is zero when the dependent variable is zero hence including it is always vital even though its meaningless for the most part.

This is the same as running t-test on parametres of regression. Let us run it on slope.

Ho = null hypothesis: slope is zero

H1 = alternative hypothesis: slope is not zero

Running t-test on it will yield the t and p value already given in the linear model. Testing it against 95% confidence interval will make us reject null hypothesis in favor of alternative hypothesis as alpha value > p value. This is the same conclusion we reached above.

The coefficient codes give us a quick way to visually see which coefficients are significant to the model. To the right of the p-values you’ll see several asterisks (or none if the coefficient is not significant to the model). The number of asterisks corresponds with the significance of the coefficient as described in the legend just under the coefficients section. The more asterisks, the more significant.

Residual Standard Error: The residual standard error is a measure of how well the model fits the data. If we look at the plot below, we notice that the line doesn’t perfectly flow through each of the points and that there is a “residual” between the Deaths and the line (shown as a blue line). The residual standard error tells us the *average* amount that the actual values of Y (the dots) differ from the predictions (the line) in units of Y. In general, we want the smallest residual standard error possible, because that means our model’s prediction line is very close to the actual values, on average.

For our current model, we can see that on average, the actual values are around 23 deaths away from the predicted values (regression line).

Multiple R-squared and Adjusted R-squared: The Multiple R-squared value is most often used for simple linear regression (one predictor). It tells us what percentage of the variation within our dependent variable that the independent variable is explaining. In other words, it’s another method to determine how well our model is fitting the data. In the example above, Petrol Prices explains **9.74%** of the variation within Monthly Deaths, our dependent variable. This means that Petrol Prices helps to explains some of the variation within Deaths, but not as much as we would like. Ultimately, our model isn’t fitting the data very well.

When running a regression model, either simple or multiple, a hypothesis test is being run on the global model.

Ho = Null hypothesis: no relationship between the dependent variable and the independent variable(s)

H1 = Alternative hypothesis: there is a relationship.

Said another way, the null hypothesis is that the coefficients for all of the variables in your model are zero. The alternative hypothesis is that at least one of them is not zero. The F-statistic and overall p-value help us determine the result of this test. Looking at the F-statistic alone can be a little misleading depending on how many variables are in your test. If you have a lot of independent variables, it’s common for an F-statistic to be close to one and to still produce a p-value where we would reject the null hypothesis. However, for smaller models, a larger F-statistic generally indicates that the null hypothesis should be rejected. A better approach is to utilize the p-value that is associated with the F-statistic.

We can see from the anova test, the F-statistic is large and our p-value is very small. This would lead us to **reject the null hypothesis** and conclude that there is strong evidence that a relationship does exist between *Monthly Deaths* and *Petrol Prices*.

**Explanation**: Here, the data is only from when law was not passed to measure a better affect of Petrol Prices and kms on the number of car drivers killed. Petrol Prices and kms here serve as the independent variables while the Driverskilled serves as the dependent variable. Using the lm command, we generate a linear regression model. Lets interpret the results;

Residuals: The residuals are the difference between the actual values and the predicted values. We want our median to be zero. From the data we have, our distribution is slightly right skewed.

Coefficients — Estimate: In the plot below, the line where it meets the y-axis is the y-intercept (b) written next to Intercept and the slope of first variable (PetrolPrices) (m1) is written next to Petrol Prices while slope of second variable (kms) (m2) is written next to kms. Because regression model finds the line that fits the points in such a way that it minimizes the distance between each point and the line, we have now the values to formulate an equation.

*y = -573.7 (x1) -1113 (x2) + 190.53*

The negative slopes denotes a negative relation between Petrol Prices, kms and DriversKilled, meaning if independents would go up, dependent goes down and vice-versa.

Coefficients – Std. Error: The standard error of the coefficient is an estimate of the standard deviation of the coefficient. In effect, it is telling us how much uncertainty there is with our coefficient. Using standard error, we can create Confidence Intervals. Lets say we create an interval of 95% then it would be as following:

-573.7 +- 1.96 (153.5) = -272.84 , -874.56

-1113 +- 1.96 (0.0006875) = -1112.99, -1113.00

Hence we are 95% sure the slopes fall in these ranges.

Coefficients – t value: The t-statistic is simply the coefficient divided by the standard error. In general, we want our coefficients to have large t-statistics, because it indicates that our standard error is small in comparison to our coefficient. The larger our t-statistic is, the more certain we can be that the coefficient is **not** zero. The t-statistic is then used to find the p-value.

Coefficients – Pr(>|t|) [p values] and Signif. Codes: The p-value is calculated using the t-statistic from the T distribution. The p-value, in association with the t-statistic, help us to understand how *significant* our coefficient is to the model. If p-value is below the significance level (usually 0.05), then it is significant meaning the coefficient is **not zero.** As we can see in the data, the p-value for both the intercept and the independent variable Petrol Price is very low denoting that both are significant and **non-zero.** But the p-value for kms is greater than the significance level (0.1 > 0.05) and is **not significant and zero.** However, the significance of the y-intercept or otherwise does not make much of a difference but excluding it when it is not significant will lead to a model where the dependent variable is zero when the dependent variable is zero hence including it is always vital even though its meaningless for the most part.

This is the same as running t-test on parametres of regression. Let us run it on slopes.

Ho = null hypothesis: slope is zero

H1 = alternative hypothesis: slope is not zero

Running t-test on it will yield the t and p value already given in the linear model. Testing it against 95% confidence interval will make us reject null hypothesis in favor of alternative hypothesis as alpha value > p value. This is the same conclusion we reached above.

The coefficient codes give us a quick way to visually see which coefficients are significant to the model. To the right of the p-values you’ll see several asterisks (or none if the coefficient is not significant to the model). The number of asterisks corresponds with the significance of the coefficient as described in the legend just under the coefficients section. The more asterisks, the more significant. As you can see, there are no asterisks on kms indicating it is not significant.

Residual Standard Error: The residual standard error is a measure of how well the model fits the data. If we look at the plot below, we notice that the line doesn’t perfectly flow through each of the points and that there is a “residual” between the Deaths and the line (shown as a red line). The residual standard error tells us the *average* amount that the actual values of Y (the dots) differ from the predictions (the line) in units of Y. In general, we want the smallest residual standard error possible, because that means our model’s prediction line is very close to the actual values, on average.

For our current model, we can see that on average, the actual values are around 23 deaths away from the predicted values (regression line).

Multiple R-squared and Adjusted R-squared: The Adjusted R-squared value is most often used for multiple linear regression (more than one predictor). It tells us what percentage of the variation within our dependent variable that the independent variables is explaining. In other words, it’s another method to determine how well our model is fitting the data. In the example above, Petrol Prices and kms explains **10.07%** of the variation within Monthly Deaths, our dependent variable. As from the SLR, we know most of it is from Petrol prices and we saw above that kms is not significant. This means that Petrol Prices and a minuscule proportion of kms helps to explains some of the variation within Deaths, but not as much as we would like. Ultimately, our model isn’t fitting the data very well.

When running a regression model, either simple or multiple, a hypothesis test is being run on the global model.

Ho = Null hypothesis: no relationship between the dependent variable and the independent variable(s)

H1 = Alternative hypothesis: there is a relationship.

Said another way, the null hypothesis is that the coefficients for all of the variables in your model are zero. The alternative hypothesis is that at least one of them is not zero. The F-statistic and overall p-value help us determine the result of this test. Looking at the F-statistic alone can be a little misleading depending on how many variables are in your test. If you have a lot of independent variables, it’s common for an F-statistic to be close to one and to still produce a p-value where we would reject the null hypothesis. However, for smaller models, a larger F-statistic generally indicates that the null hypothesis should be rejected. A better approach is to utilize the p-value that is associated with the F-statistic.

We can see from the anova test, the F-statistic , for Petrol Prices, is large and our p-value is very small. For kms, the F-statistic is small and the p-value is also larger than the level of significance.

Independent Variable Petrol Prices:

This would lead us to **reject the null hypothesis** and conclude that there is strong evidence that a relationship does exist between *Monthly Deaths* and *Petrol Prices*.

Independent Variable kms:

This would lead us to **accept the null hypothesis** and reject alternative hypothesis and conclude that there is strong evidence that a relationship does **not** exist between *Monthly Deaths* and *kms*.