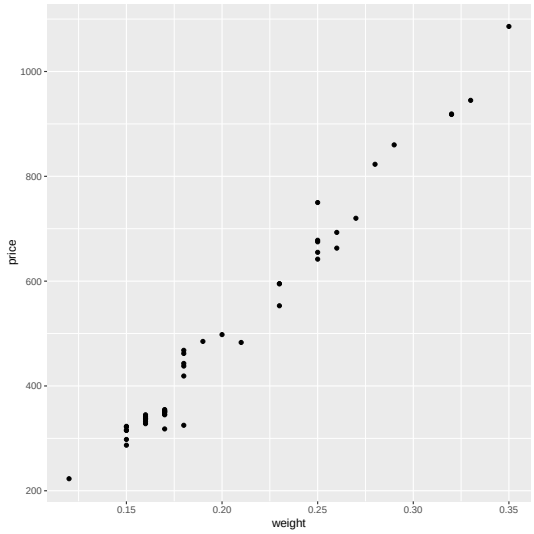
Simple Linear Regression Example

A first plot that shows the linear relationship between price and weight (data set diam.txt),



To define the linear model, use function *lm()*. It's basic summary is given by:

```
> lm1=lm(price~weight, data=diam)
> summary(lm1)

Call:
lm(formula = price ~ weight, data = diam)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

The parameters for the fitted model are $\hat{\beta}_0 = -259.63$ and $\hat{\beta}_1 = 3721.02$. The fitted line (model) is:

$$\hat{y}_i = -259.63 + 3721.02x_i$$

The mean price for a diamond increases 3721.02 for 1 carat increase in weight. Since carats don't increase that way, for an increase in carats of 0.01, the mean price of a diamond increases 37.21.

The output also shows their standard errors, t-values and p-values for the respective test with null hypothesis $H_0 : \beta_0 = 0$ and $\beta_1 = 0$. Observing these p-values, we reject H_0 so we have evidence for $\beta_0 \neq 0$ and $\beta_1 \neq 1$ and they are significantly different from 0. The p-value in the last line is for the overall test for the significance of the regression model:

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_1 : \beta_1 \neq 0$$

and, only in the simpler linear regression model (with only one predictor x), it is equivalent to the one for β_1 above (t-test).

The residual standard error (31.84) is an estimate of σ, being σ^2 the unknown population variance of the residuals (or the response y, which is assumed to be constant around the unkown mean population line).

The Multiple R-squared (or coefficient of determination) is a proportion. Multiplied by 100 measures the percentage of the variation in y that is *explained* by the variation in predictor x, so it is the proportion of variance in y explained by the regression model. The closer to 1 or 100 % the better. It measures the ability to predict Y using X. Only in the simple linear regression model (with only one predictor x), holds this equality:

$$R^2 = r^2, \quad \text{being } r \text{ the Pearson correlation coefficient}$$

Its squared root, R, is known as the multiple correlation coefficient, and is the correlation between the observed variable Y and the predicted values by the model $\hat{Y}$.

The command

```
attributes(lm1)
```

tells you which variables are stored in your lm1 object. You can access the residuals as *lm1$residuals* and the fitted values as *lm1$fitted*. Many other R-functions are applied directly to an *lm* object.

Find the correlation between the observed values and fitted values.

```
cor(lm1$fitted,diam$price)^2
```

and check that it is equal to the Multiple R^2 reported in the output.

Checking model assumptions

Firstly, let's recall what are the assumptions that comprise the linear reggression model:

- The mean of the response $E(Y_i)$ at each value of the predictor, x_i, is a Linear function of the x_i. We are going to check this assumption by visual inspection of the plot Y vs. X altough we will see other type of plots in multiple linear regression.
- The residuals e_i, are Independent. Apart from visual inspection of the residuals plot, we can do a test about independence. For instance, Durbin Watson test. The null hypothesis is that your residuals are independent (no correlation among the residuals). Fubction *dwtest()*, package lmtest.

- The residuals e_i, at each value of the predictor, x_i, are Normally distributed. We will perform a normality test to check that assumption, Jarque-Bera test will be fine or other similar. Normality is the more robust assumption of the linear model, mild departures can still be permitted.
- The residuals e_i, at each value of the predictor, x_i, have Equal variances. This can be checked with the Breusch-Pagan test, *bptest(), package lmtest*, with the null that your residuals are homocedastic (constant variance).

Secondly, let's see why we have to evaluate any regression model that we formulate and subsequently estimate. In short, it's because:

- All of the estimates, intervals, and hypothesis tests arising in a regression analysis have been developed assuming that the assumptions are met.
- If the hypothesis are not met, then the formulas and methods we use are at risk of being incorrect.

The good news is that some of the model conditions are more forgiving than others. So, we really need to learn when we should worry the most and when it's okay to be more carefree about model violations. Here's a pretty good summary of the situation:

- All tests and intervals are very sensitive to departures from independence.
- All tests and intervals are sensitive to moderate departures from equal variance.
- The hypothesis tests and confidence intervals for β_0 and β_1 are fairly "robust"(that is, forgiving) against departures from normality.
- Prediction intervals are quite sensitive to departures from normality.

The important thing to remember is that the severity of the consequences is always related to the severity of the violation. And, how much you should worry about a model violation depends on how you plan to use your regression model. For example, if all you want to do with your model is test for a relationship between x and y, i.e. test that the slope β_1 is 0, you should be okay even if it appears that the normality condition is violated. On the other hand, if you want to use your model to predict a future response y_{new} , then you are likely to get inaccurate results if the error terms are not normally distributed.

To get diagnostics plots for a simple regression model, type out:

```
par(mfrow=c(2,2))
plot(lm1, which=c(1:4), ask=F)
```

Clockwise, starting from top-left, the plots shown are the following:

- Residuals against fitted values. Here is where we have to find those patterns showing lack of independence and non-linear relationships between outcome and predictor.
- Normal Q-Q plot: a graphical tool to help us if the residuals come from a Normal distribution.

- A Scale-location plot (also called Spread-Location plot) of $\sqrt{|e_i|}$ against fitted values. It shows if the residuals are spread equally along the range of the predictor. It is good a roughly horizontal line with equally (randomly) spread points is displayed.
- A plot of Cook's distances in the same order as the observations. It is useful for finding influential observation and we talk more about it on multilple linear regression.

Some tests on the residuals

```
dwtest(lm1, alternative="two.sided") #for independence
Box.test(residuals(lm1)) # also for independence
jarque.bera.test(residuals(lm1)) # for normality
bptest(lm1)  #for constant variance
```

Let's find out what the model predicts (and confidence intervals) for diamons with carats of $0.13, 0.22, 0.24, 0.31, 0.34$, which we haven't observed. The prediction is for the mean price of diamonds with such carats and confidence intervals for these mean prices.

```
new.data=data.frame(weight=c(0.13, 0.22, 0.24, 0.31, 0.34))
predict(lm1, new.data, interval="confidence")
```