Homework 2

Denis Ostroushko

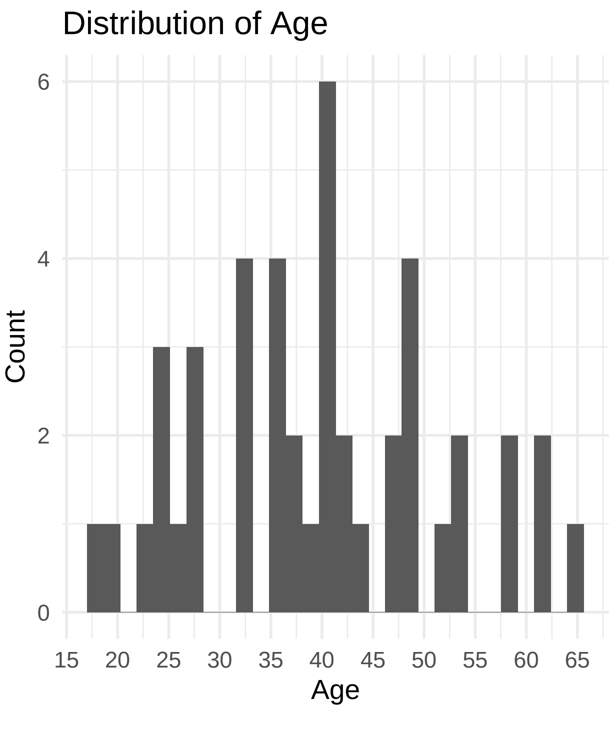
2022-10-29

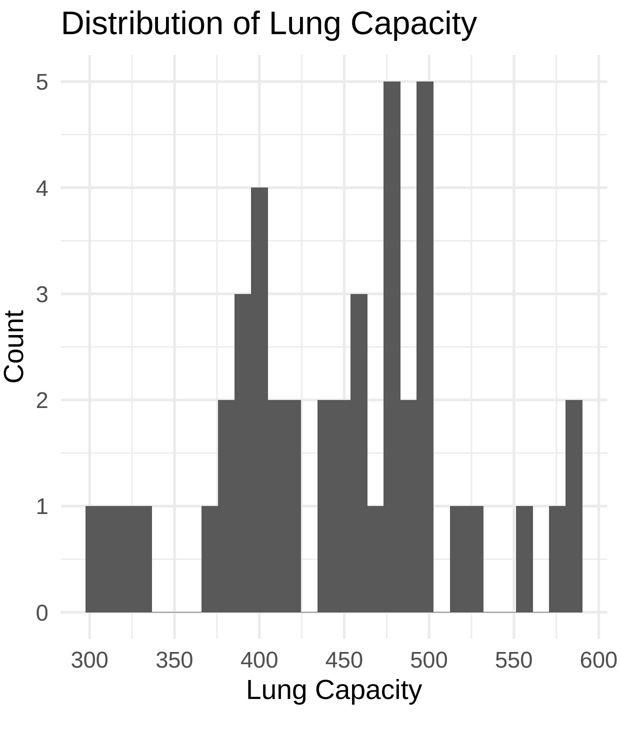
4.2

In this section we need to establish the relationship between Age and Vital Lung Capacity for men working in the cadmium industry, but not exposed to the cadmium fumes. We assign X variable to be Age, and Y variable to be Lung Capacity.

4.2 - A

To establish connection between age and lung capacity we need to develop a regression model with estimates of β_0 and β_1. First, let's look at the distribution of X and Y.





Both X and Y have a distribution with a central tendency. Most values seem to be centered around the mean and median of distributions. So, we will expect confidence and prediction intervals for be narrower near average age and lung capacity values, and wider towards the end of the distribution.

We also want to consider the scope of the model. Minimum age in this sample is 18 while maximum age is 65. Therefore, trying to predict lung capacity outside this range can result in predictions with high margin of error. Moreover, we do not know the relationship between lung capacity and age outside of this range, so we will avoid extrapolation.

Variable	N for Analysis	Mean	Median	Standard Deviation
Age	44	39.84091	41	11.95134
Lung Capacity	44	446.20455	453	69.22615

Finally, we will look at the mean, median, and standard deviation for the two variables. These statistics will give us slightly more insight into the confidence and prediction intervals behavior.

We do not have missing values, standard deviation is reasonable for both variables in relation to the average values. Mean and median are also pretty close to each other for both variables.

Finally, we can estimate β_0 and β_1, we call estimates b_0 and b_1 respectively.

First need b_1, because the value of b_0 depends of b_1, and we will use this formula Formula:

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

We obtain values from the data set and calculate the value of b_1 below:

Estimate for $\beta_1 = \hat{\beta}_1 = b_1 =$ -3.0716244

Now we can obtain b_0

Formula:

$$\frac{1}{n}(\Sigma Y_i - b_1 \times \Sigma X_i) = \bar{Y} - b_1 \times \bar{X}$$

Estimation code is given below

Estimate for $\beta_0 = \hat{\beta}_0 = b_0 =$ 568.580855

To be sure that our calcualtion went right, we can create a linear model in R, and extract estimates from it. Model output is given below:

```
##
## Call:
## lm(formula = xvc100 ~ age, data = vc_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124.79  -42.75   -1.25   45.20  114.08
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 568.5809    31.4892  18.056 < 0.0000000000000002 ***
## age          -3.0716     0.7578  -4.054             0.000214 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.39 on 42 degrees of freedom
## Multiple R-squared:  0.2812, Adjusted R-squared:  0.2641
## F-statistic: 16.43 on 1 and 42 DF,  p-value: 0.0002135
```

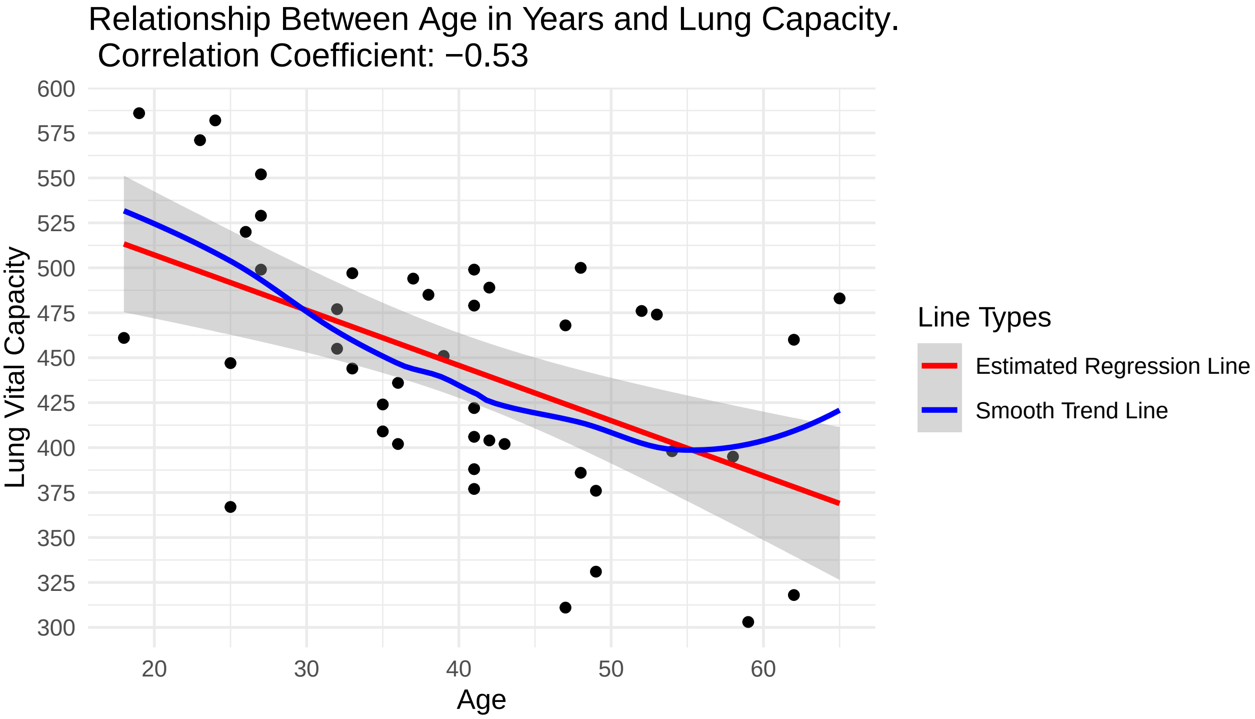
From the summary of the model we can see that $b_0 =$ 568.580855 and $b_1 =$ -3.0716244

Difference in two estimates for b_1 is 0, rounded to 4 decimal points. Difference in two estimates for b_0 is 0, rounded to 4 decimal points.

We have successfully calculated the two estimates we need.

4.2 - B

In order to plot regression line we save predicted values from the model to the data frame.



Overall, regression line follows smooth trend line, so the relationship between the two variables must be linear. At higher age levels the smooth trend line starts to curve, however, there are less data available in that region, so we should be careful with the interpretation of what we see.

We also can see that the variance around fitted regression line is quite large. This is also supported by a wide regression bound around regression line.

4.2 - C

To obtain this estimate we simply need to plug in the value of $X = 35$ into our regression equation. Note that this age is quite close to the average value of age in the sample, so we should be have a pretty good estimate for the condifence interval of the average lung capacity for a male who is 45 years old.

Estimated value is 461.0740001

We can also obtain a confidence interval for the mean response level when $X_h = 35$

First, we want a standard error, so we need MSE (Mean Squared Error, obtain from residuals), deviation of X_h from the mean, $\bar{X}$, and total variance of X

We will also obtain a coefficient from the t distribution, at 95% confidence level and 42 degrees of freedom

This is a new formula in this assignment, so we will state it below, before estimating standard error using data.

$$se(b_1)^2 = MSE \times [\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}]$$

We estimate that the average lung capacity for men who are 35 years old will be 461.0740001, with a confidence interval given by (441.9978249, 480.1501753).

Again, we can check our work using existing R functions. We will estimate average lung capacity of 35 year old males using code below. It also conveniently provides a confidence interval.

```
##       fit      lwr      upr
## 1 461.074 441.5489 480.5991
```

The two sets of estiamtes align very closely.

4.2 - D

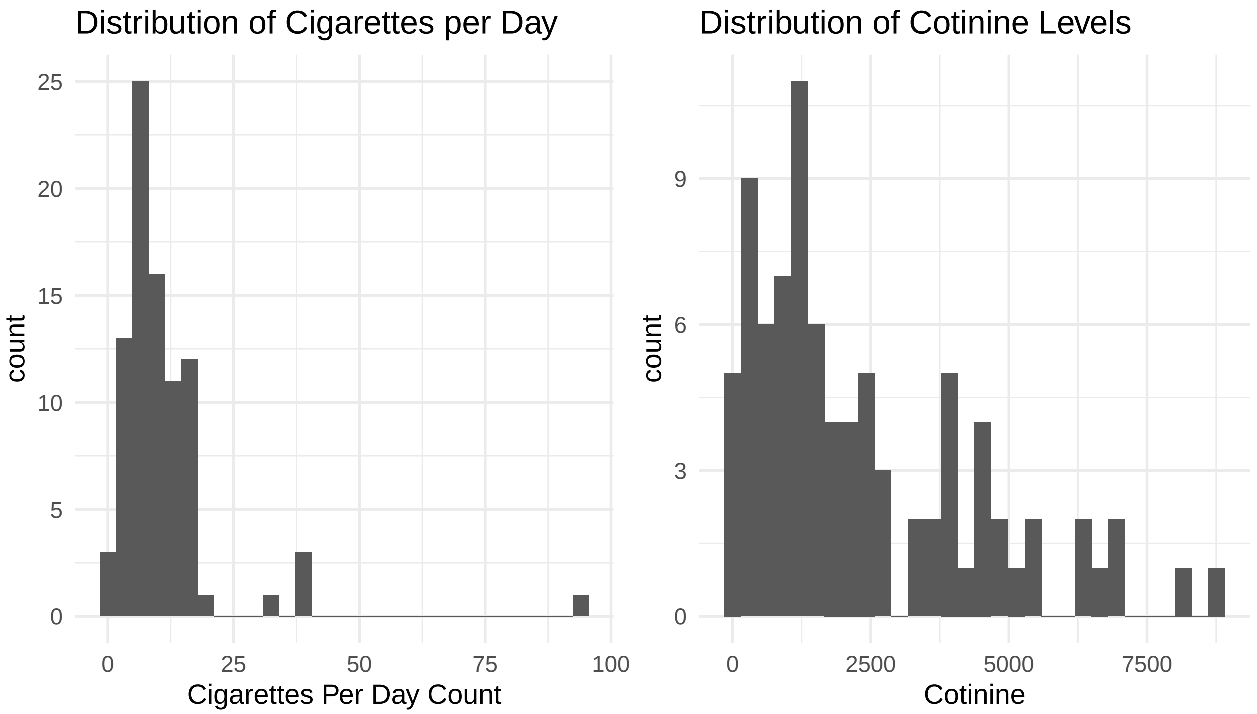
Mean response is the average change in response variable ,lung capacity function, when X , age, is increased by 1 unit, i.e. when person gets one year odler. When age increases by 1 year then VC changes by -3.0716244 units. However, it makes more sense to say that one additional year of age decreases the lung capacity by an overage of 3.0716244 units.

Since the relationship is linear, when a person gets 10 years older, VC decreases by 30.7162443

5.3

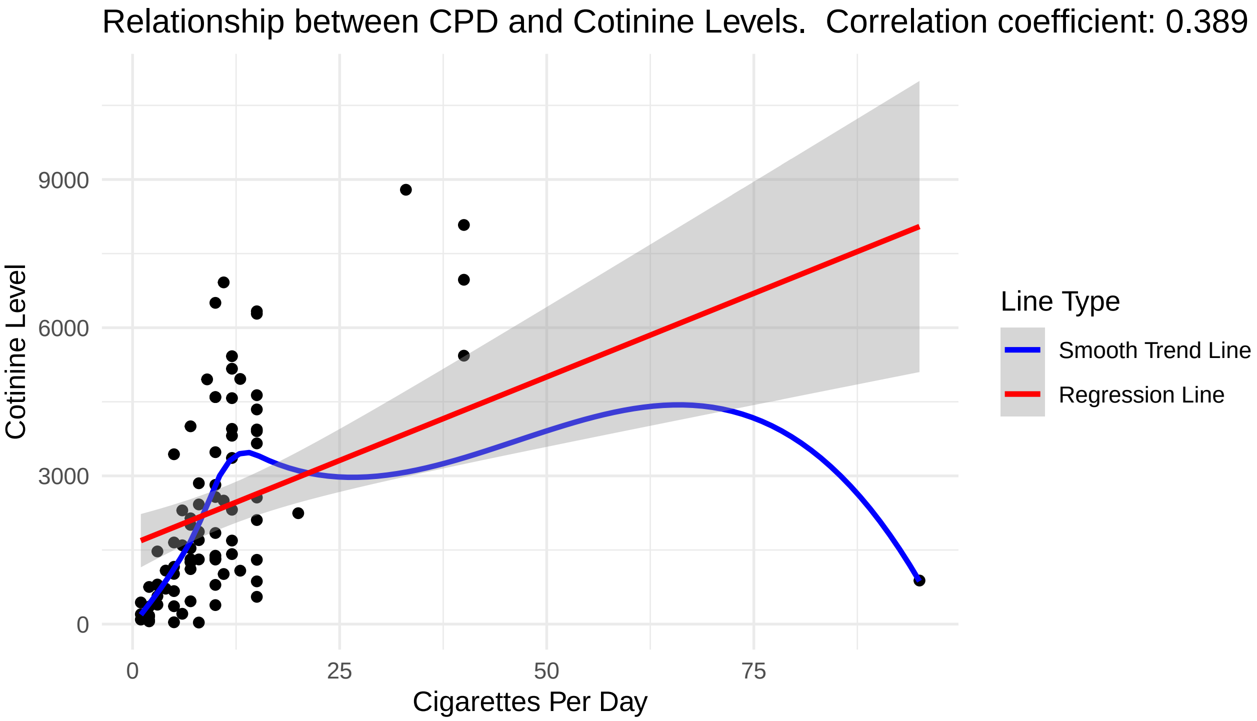
Before we can get an estimate and interpret the meaning of the confidence interval we need to take a look at the summary statistics and the distribution of two variables.

Variable	N	Mean	Median	Standard Deviation
CPD	86	11.000	10.0	11.93019
Cotinine	86	2365.709	1623.5	2073.15867



These two variables are heavily skewed, with long tails that have heavy outliers. This was expected when we look at the summary statistics. The mean is greater than the median, which usually means that there are a few outliers on the far positive side, which increase the value of the mean. Visual summary confirms this. Additionally, when we see a large value of standard deviation, this is another sign that there are huge positive outliers.

We also will take a look at the relationship between CPD and Cotinine levels using a scatter plot.



Overall,we should expect a very poor fit of regression model to this data. There are several outliers that skew the fitted line. Especially a data point for someone who smokes over 75 cigarettes per day, but has cotinine levels that are more common for people who smoke between zero and ten cigarettes per day. However, we are not tasked with diagnostics and data tuning in this assignment, so we leave the data point here.

5.3 - A

We need to estimate average response when $X_h = 30$, 30 cigarettes per day, and obtain a confidence interval for it.

We showed how to estimate model parameters, average response level and confidence interval by hand in 4.2, so we will use R functions to get estimates for interpretations.

We provide a summary of the model that we use to obtain an estimate.

```
##
## Call:
## lm(formula = cotinine ~ cpd, data = cig)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7166.0 -1284.0  -545.2  1247.0  4938.1
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  1621.60     282.52   5.740 0.000000147 ***
## cpd            67.65      17.46   3.873    0.000212 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1921 on 84 degrees of freedom
## Multiple R-squared:  0.1515, Adjusted R-squared:  0.1414
## F-statistic:    15 on 1 and 84 DF,  p-value: 0.0002119
```

```
##        fit      lwr     upr
## 1 3650.992 2873.095 4428.89
```

We estimate that people who smoke 30 cigarettes per day have cotinite level of 3650.9922417, bounded by (2873.0946136, 4428.8898697).

There confidence interval is very wide, because we estimate the response level for values that are quite far from the average value of cigarettes per day, which is 11. This high distance from the center of the distribution contributes to the standard error a lot. Moreover, we do not have enough data in that region of the distribution of CPD, and we can even make an argument that 30 cigarettes per day may be considered outside of the model scope, if we change the way these data were collected.

5.3 - B

Obtaining a prediction interval for a single new observed value is a new exercise in this homework, so we will state the formula below. We will also use a built in function to validate our results.

$$se(Y_{h(new)})^2 = MSE * [1 + \frac{1}{n} + \frac{(X_{h(new)} - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}]$$

The prediction interval for cotinine levels of a person who smokes 30 cigarettes per day bounded by (-3422.0104, 4374.8746).

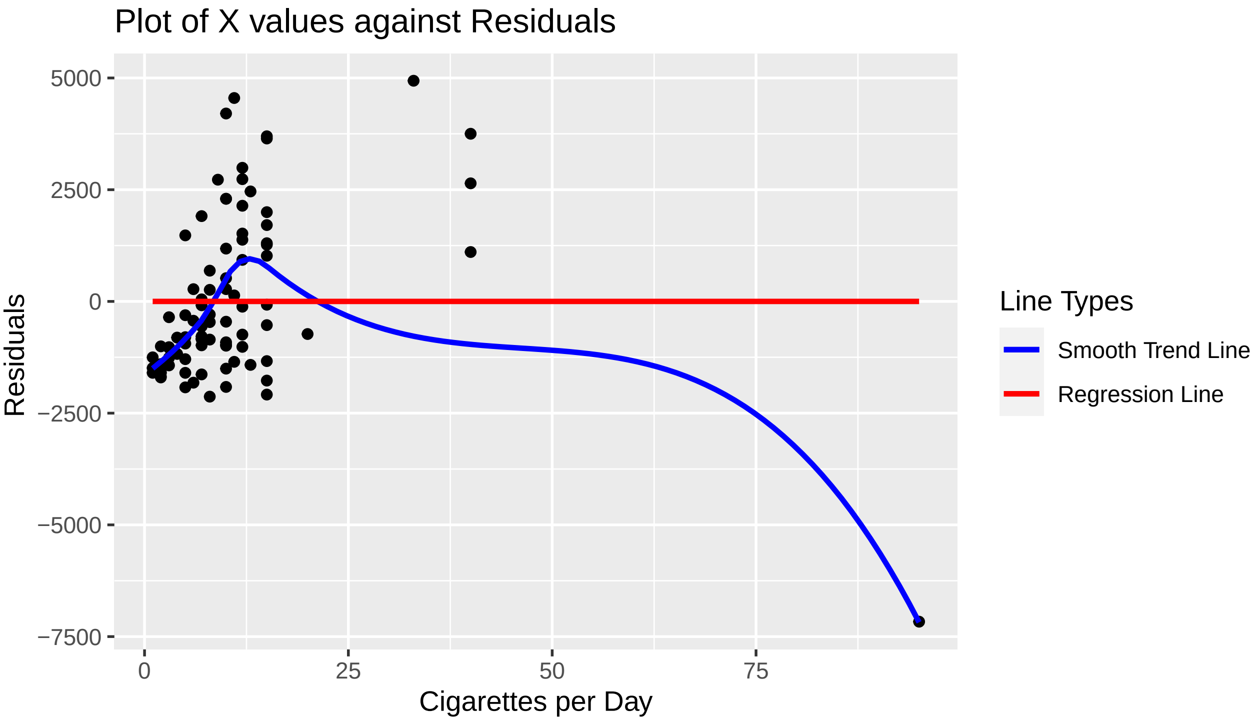
we can check the result using the function below:

```
##        fit       lwr      upr
## 1 3650.992 -247.4502 7549.435
```

Our result matches, so we can interpret the results now.

Note that the cotinine levels range from 32 and 8792 in the sample of data we have for analysis. This means that the prediction interval spans over almost entire range of the Y variable. We saw in the introduction of this analysis section that the two variables have outliers and extreme values. They contribute meaningfully to the standard error and uncertainty in the estimate. Moreover, estimating a prediction interval for a single observation brings another level of uncertainty and variation. All together these factors result in a prediction interval that is essentially unusable, since it captures alsmost the entire range of cotinine values.

5.3 - C



The plot of residuals and corresponding values of cigarettes per day has the same issues as the scatter plot. If we focus on the cluster of the points for people who smoke less than 25 cigarettes per day, residuals are randomly scattered above and below the average regression line. Extreme values and outliers are harder to interpret due to their nature. overall, there is no notable linear or other trend in residuals against the values of the predictor, so residuals are independent of cpd, which follows the assumption of the model.

We can also see that residuals may vary between 2500 and 5000 values, which is a large error.

Summary of residuals also suggests that their distribution might be approximately normal.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -7166.0 -1284.0  -545.2     0.0  1247.0  4938.1
```

average residual value: 0, which also follows a model assumption.

5.3 - D

We can set up the ANOVA table using a model we created earlier and a simple function available in R.

First of all we can observe that the sum of squares is huge for both regression and error terms. We saw that residuals had a lot of variation when evaluating the residual plot. We also know that regression mean square would be large because of extreme values and outliers in the data.

ANOVA table allows us to test the following hypotheses:

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

```
## Analysis of Variance Table
##
## Response: cotinine
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## cpd        1  55360992 55360992  15.003 0.0002119 ***
## Residuals 84 309967894  3690094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The goal of the ANOVA table is to evaluate the relationship response and predictor when comparing the portion of variation explained by the predictor. If the regression mean square is large, then the coefficient of the fitted line is far from zero, and must be related to the response variable. If the error mean square is small, then the values of response are distributed close to the fitted line, and therefore predictor value is a good proxy for values of Y.

So, we obtain an F statistic to see is regression mean square is far greater than the error mean square.

$$Fstatistic = F^* = \frac{RegressionMS}{ResidualMS}$$

We obtain $F^* = 15.0025968$, NA. In order to know if this ratio is great enough we prepare a cutoff. Cutoff is given by a value of F that depends on the desired confidence level and degrees of freedom. Degrees of freedom of F include 2 numbers. Regression degrees of freedom work out to be number of model terms k minus 1 , which is 1, since we have 1 predictor and 1 estimate for the intercept. Error degrees of freedom is $n - 2$, which is 84.

Hence, the cutoff value for $F = 3.9545684$ for the 95% confidence level.

Clearly, $15.0025968 > 3.9545684$, so we reject the null hypothesis and accept that β_1 is not zero. Therefore, there exists a statistically significant relationship between Cotinine levels and the daily cigarette consumption.