Laboratory Exam

Simple Linear Regression

Charlene Garridos

2023-06-29

Contents

1 The Fractional Distillation Data

The purity of oxygen produced by a fractional distillation process is thought to be related to the percentage of hydrocarbons in the main condensor of the processing unit.

```r
# Specify the file path relative to the working directory
file_path <- "/Users/User/Downloads/fractional_distillation_data.csv"

# Read the CSV file
```

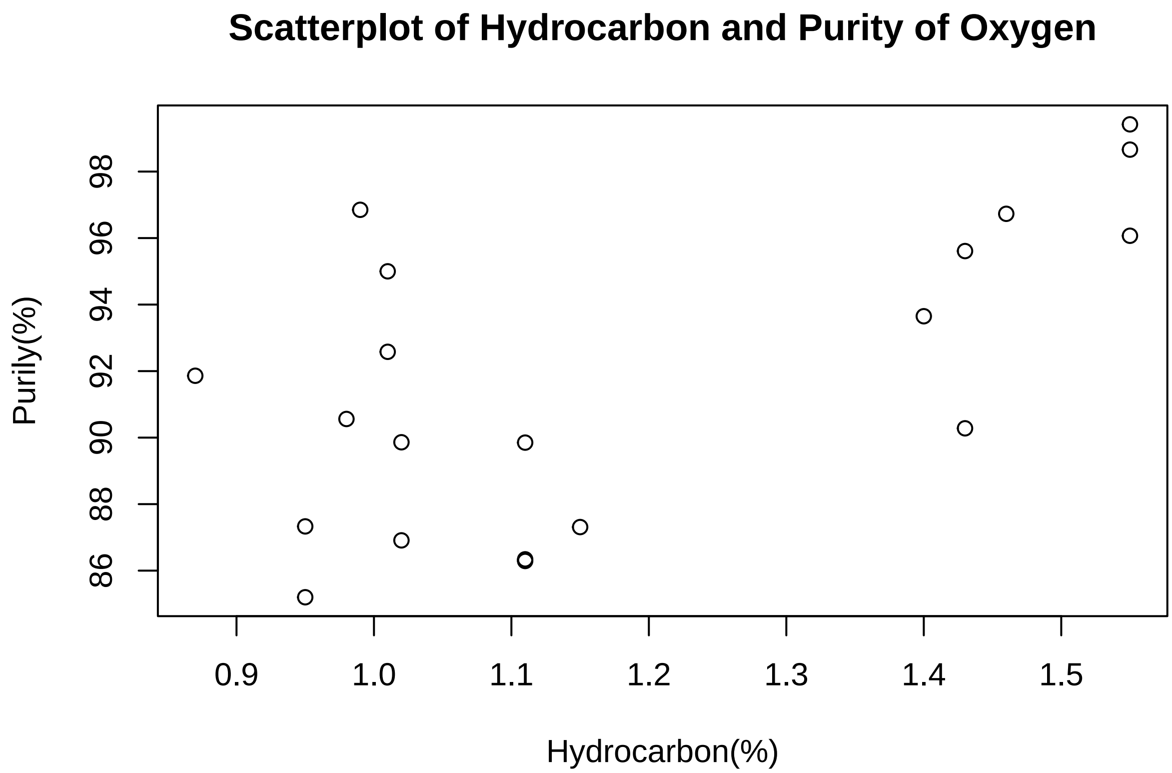
```r
fractional_distillation_data <- read.csv(file_path)

# View the data
print(fractional_distillation_data)
```

```
##    Purily Hydrocarbon
## 1   86.91        1.02
## 2   89.85        1.11
## 3   90.28        1.43
## 4   86.34        1.11
## 5   92.58        1.01
## 6   87.33        0.95
## 7   86.29        1.11
## 8   91.86        0.87
## 9   95.61        1.43
## 10  89.86        1.02
## 11  96.73        1.46
## 12  99.42        1.55
## 13  98.66        1.55
## 14  96.07        1.55
## 15  93.65        1.40
## 16  87.31        1.15
## 17  95.00        1.01
## 18  96.85        0.99
## 19  85.20        0.95
## 20  90.56        0.98
```

1.1 Scatter Diagram

a. Create a scatter diagram for the data.

```r
# scatter plot of propellant_age and shear_strength
plot(x = fractional_distillation_data$Hydrocarbon, y = fractional_distillation_data$Purily,
     xlab = "Hydrocarbon(%)", ylab = "Purily(%)",
     main = "Scatterplot of Hydrocarbon and Purity of Oxygen")
```



Least-Squares Estimation of the Parameters

Use the lm() function to calculate the linear model based on the data set.

```
# calculate model
model <- lm(data = fractional_distillation_data,
formula = Purily ~ Hydrocarbon)
```

The model object is a list of a number of different pieces of information, which can be seen by looking at the names of the objects in the list.

```
# view the names of the objects in the model
names(model)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
model$coefficients
```

```
## (Intercept) Hydrocarbon
##    77.86328    11.80103
```

1.2 The Least-Squares Fit

The summary() function is a useful way to gather critical information in your model. b. The Least-sqaures fit is

```
model_summary <- summary(model)
model_summary$sigma
```

```
## [1] 3.59656
```

```
model_summary$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 77.86328   4.198888 18.543786 3.537382e-13
## Hydrocarbon 11.80103   3.485119  3.386119 3.291122e-03
```

1.3 The Estimate

c.The estimate of σ^2 is

```
sigma_squared<- (model_summary$sigma)^2
sigma_squared
```

```
## [1] 12.93524
```

Hypothesis Testing on the Slope and Intercept

```
model_summary
```

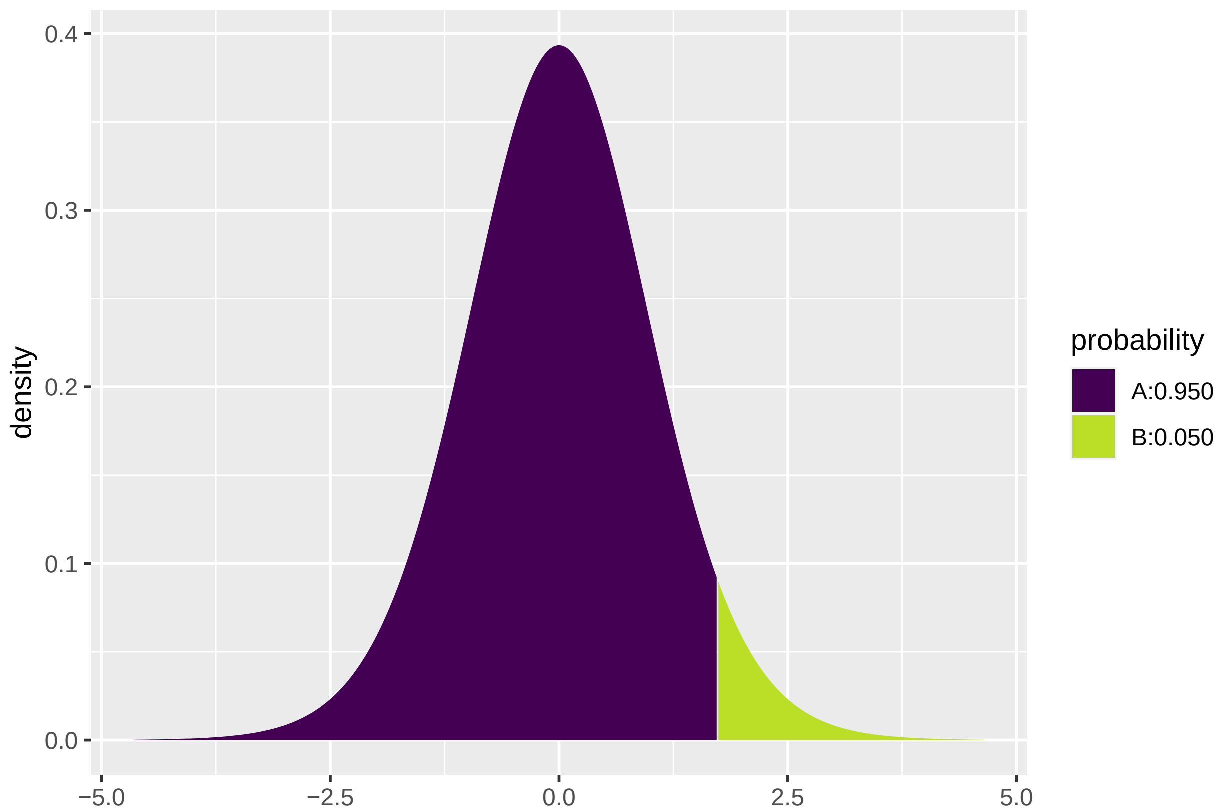
```
##
## Call:
## lm(formula = Purily ~ Hydrocarbon, data = fractional_distillation_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6724 -3.2113 -0.0626  2.5783  7.3037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   77.863      4.199  18.544 3.54e-13 ***
## Hydrocarbon   11.801      3.485   3.386  0.00329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.597 on 18 degrees of freedom
## Multiple R-squared:  0.3891, Adjusted R-squared:  0.3552
## F-statistic: 11.47 on 1 and 18 DF,  p-value: 0.003291
```

```
model_summary$coefficients["Hydrocarbon",]
```

```
##    Estimate   Std. Error      t value     Pr(>|t|)
## 11.801028193  3.485118700  3.386119444  0.003291122
```

```r
mosaic::xqt(0.95, 18)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                            from
##   fortify.SpatialPolygonsDataFrame ggplot2
```




```
## [1] 1.734064
```

1.4 Test for Significance of Regression in the Fractional Distillation Regression Model.

d. Test for Significance of Regression in the Fractional Distillation Regression Model.

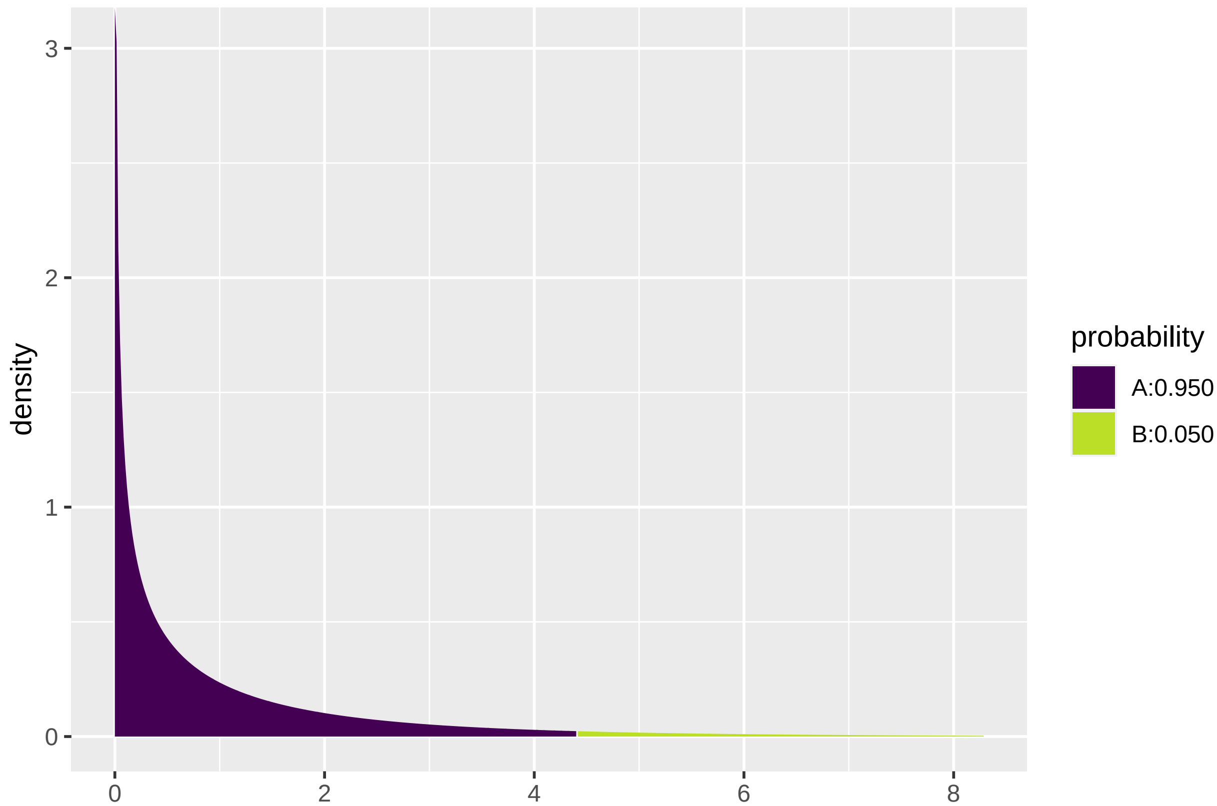
Based on the statistical analysis conducted, the t-statistic for the coefficient of Hydrocarbon, with the 95% confidence interval, is 3.386 ($p < 0.003$), indicating a significant relationship between hydrocarbon and the purity of oxygen. This positive t-value suggests that as the hydrocarbon increases, the purity of oxygen tends to increase. Therefore, we can infer that hydrocarbon has a substantial impact on the purity of oxygen in the fractional distillation data.

Thus, the statistically significant t-value provides strong evidence to reject $H_0 : \beta_1 = 0$ and support the conclusion that hydrocarbon is an influential factor in determining the purity of an oxygen.

```r
model_summary$fstatistic
```

```
##   value   numdf   dendf 
## 11.4658  1.0000 18.0000
```

```r
mosaic::xqf(0.95,1,18)
```




```
## [1] 4.413873
```

1.5 An analysis-of-variance approach to test significance of regression.

e. An analysis-of-variance approach to test significance of regression. At 95% confidence interval, the F-statistic of 11.4658 ($p < 0.003$) obtained from the regression analysis indicates that the overall model, including the intercept and the predictor variable hydrocarbon, is statistically significant. It suggests the inclusion of hydrocarbon as a predictor in the model significantly improves the ability to predict the purity of the oxygen compared to a model without this variable. Along with the low p-value, it indicates that there is a low probability of obtaining such a strong relationship between the predictors and the response variable by chance alone. This strengthens our confidence in the conclusion that hydrocarbon has a substantial impact on the purity of the oxygen. Therefore, the F-statistic provides strong evidence to reject $H_0 : \beta_1 = 0$.

1.6 95% CI on the slope

```
confint(model_summary)
```

```
##                 2.5 %   97.5 %
## (Intercept) 69.041747 86.68482
## Hydrocarbon  4.479066 19.12299
```

1.7 95% CI on the mean purily when the hydrocarbon percentage is 1.00

```
confint(model_summary$coefficients)
```

```
## Confidence Interval from Bootstrap Distribution (8 replicates)
```

```
##                    2.5%    97.5%
## percentile 0.0005759464 67.48237
```

2 The Steam Consumption Data

The number of pounds of steam used per month at a plant is thought to be related to the average monthly ambient temperature.

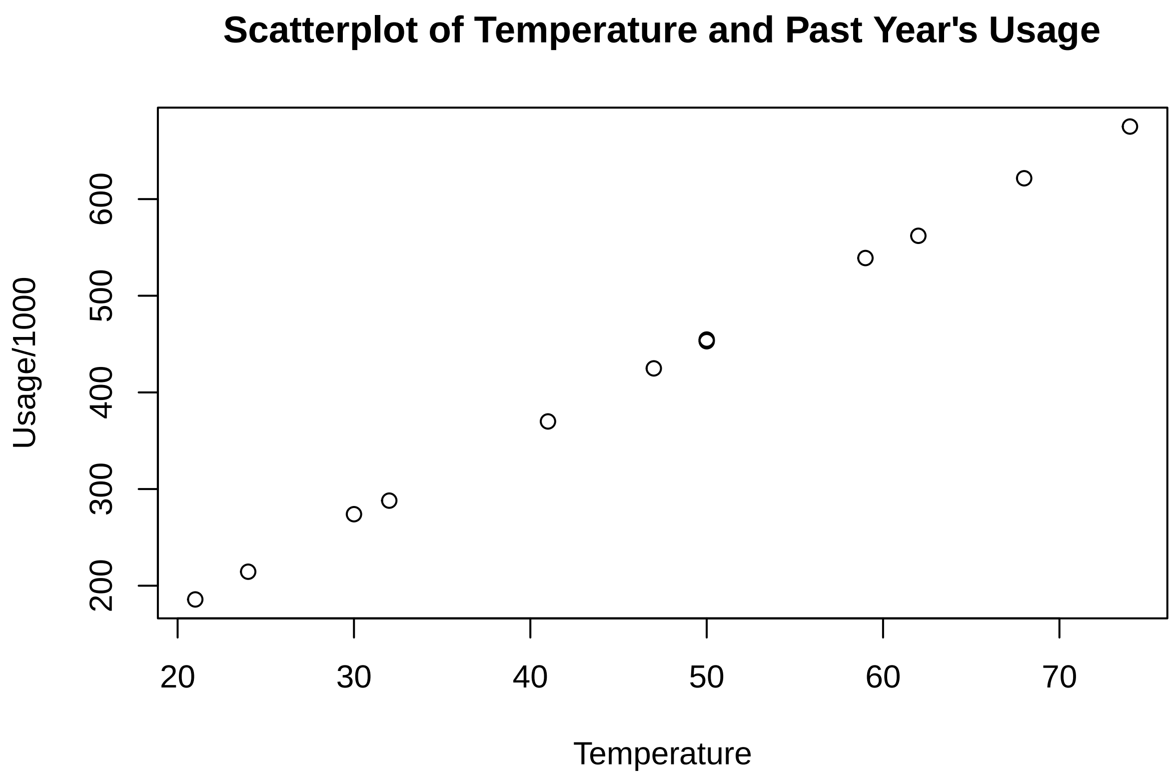
```
# Specify the file path relative to the working directory
file_path <- "/Users/User/Downloads/steam_consumption_data.csv"

# Read the CSV file
steam_consumption_data <- read.csv(file_path)

# View the data
print(steam_consumption_data)
```

```
##    Month Temperature  Usage
## 1    Jan          21 185.79
## 2    Feb          24 214.47
## 3    Mar          32 288.03
## 4    Apr          47 424.84
## 5    May          50 454.68
## 6    Jun          59 539.03
## 7    Jul          68 621.55
## 8    Aug          74 675.06
## 9    Sep          62 562.03
## 10   Oct          50 452.93
## 11   Nov          41 369.95
## 12   Dec          30 273.98
```

2.1 Scatter Diagram

a. Create a scatter diagram for the data.

```
# scatter plot of propellant_age and shear_strength
plot(x = steam_consumption_data$Temperature, y = steam_consumption_data$Usage,
     xlab = "Temperature", ylab = "Usage/1000",
     main = "Scatterplot of Temperature and Past Year's Usage")
```



Least-Squares Estimation of the Parameters

Use the lm() function to calculate the linear model based on the data set.

```
# calculate model
model_1 <- lm(data = steam_consumption_data,
formula = Usage ~ Temperature)
```

The model object is a list of a number of different pieces of information, which can be seen by looking at the names of the objects in the list.

```
# view the names of the objects in the model
names(model_1)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```r
model_1$coefficients
```

```
## (Intercept) Temperature
##   -6.332087    9.208468
```

2.2 The Least-Squares Fit

The summary() function is a useful way to gather critical information in your model. b. The Least-sqaures fit is

```r
model_1_summary <- summary(model_1)
model_1_summary$sigma
```

```
## [1] 1.945628
```

```r
model_1_summary$coefficients
```

```
##              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -6.332087 1.67004573  -3.791565 3.534310e-03
## Temperature  9.208468 0.03382295 272.254999 1.099192e-20
```

2.3 The Estimate

c.The estimate of σ^2 is

```r
sigma_squared<- (model_1_summary$sigma)^2
sigma_squared
```

```
## [1] 3.78547
```

Hypothesis Testing on the Slope and Intercept

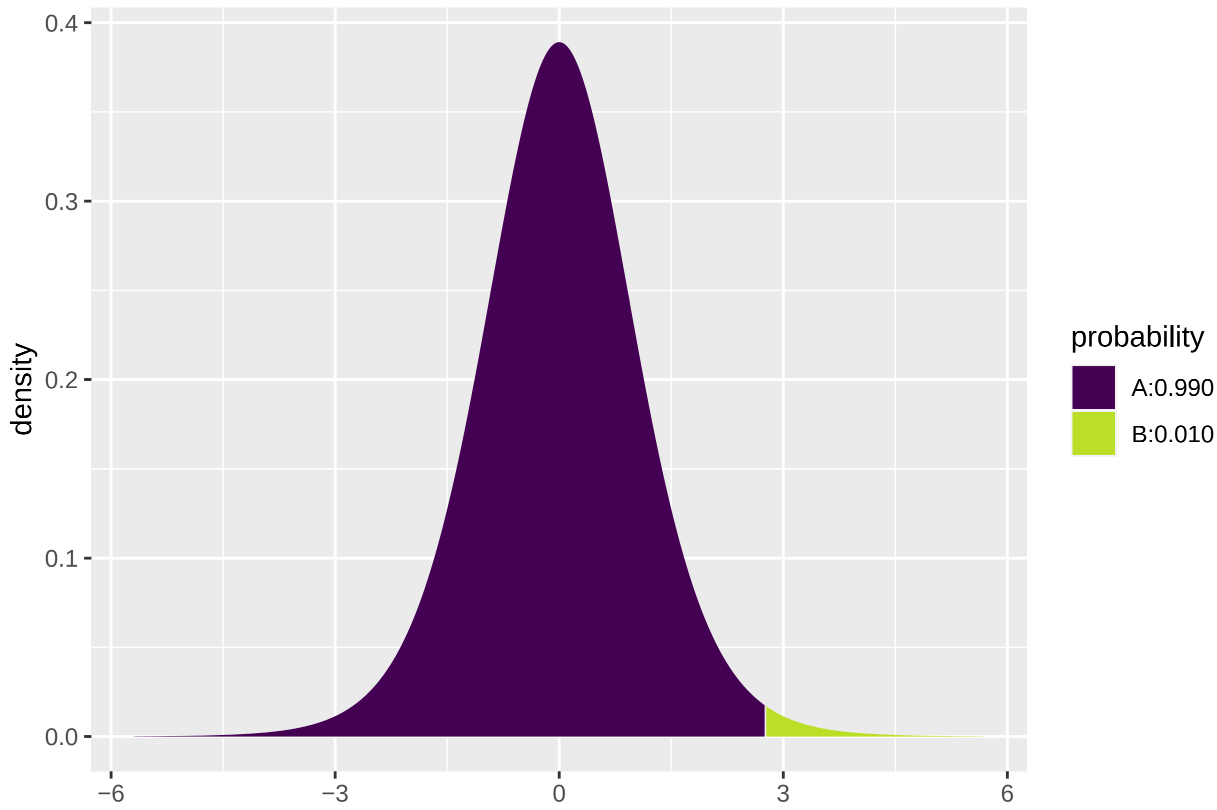
```r
model_1_summary
```

```
##
## Call:
## lm(formula = Usage ~ Temperature, data = steam_consumption_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5629 -1.2581 -0.2550  0.8681  4.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.33209    1.67005  -3.792  0.00353 **
## Temperature  9.20847    0.03382 272.255  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.946 on 10 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 7.412e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

```r
model_1_summary$coefficients["Temperature",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 9.208468e+00 3.382295e-02 2.722550e+02 1.099192e-20
```

```r
mosaic::xqt(0.99, 10)
```




```
## [1] 2.763769
```

2.4 Test for Significance of Regression in the Steam Consumption Regression Model.

d. Test for Significance of Regression in the Steam Consumption Regression Model.

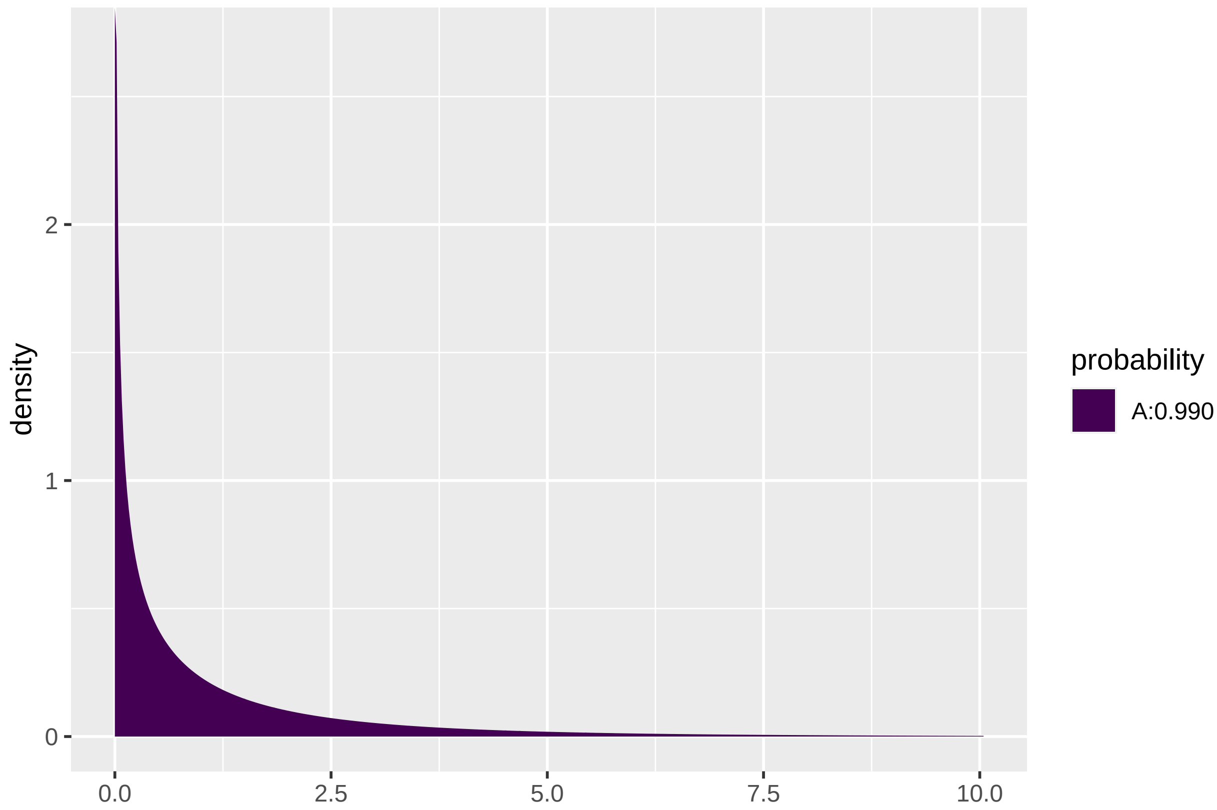
Based on the statistical analysis conducted, the t-statistic for the coefficient of Temperature, with the 99% confidence interval, is 2.722 ($p < 0.002$), indicating a significant relationship between temperature and usage. This positive t-value suggests that as the temperature increases, the usage tends to increase. Therefore, we can infer that temperature has a substantial impact on the usage of the steam consumption.

Thus, the statistically significant t-value provides strong evidence to reject $H_0 : \beta_1 = 0$ and support the conclusion that temperature is an influential factor in determining the usage.

```r
model_1_summary$fstatistic
```

```
##    value    numdf    dendf 
## 74122.78     1.00    10.00
```

```r
mosaic::xqf(0.99,1,10)
```




```
## [1] 10.04429
```

2.5 An analysis-of-variance approach to test significance of regression.

e. An analysis-of-variance approach to test significance of regression. At 99% confidence interval, the F-statistic of 74122.78 ($p < 0.002$) obtained from the regression analysis indicates that the overall model, including the intercept and the predictor variable temperature, is statistically significant. It suggests the inclusion of temperature as a predictor in the model significantly improves the ability to predict the usage compared to a model without this variable. Along with the low p-value, it indicates that there is a low probability of obtaining such a strong relationship between the predictors and the response variable by chance alone. This strengthens our confidence in the conclusion that temperature has a substantial impact on the usage of the steam consumption. Therefore, the F-statistic provides strong evidence to reject $H_0 : \beta_1 = 0$.

2.6 99% CI on the slope

```r
confint(model_1_summary)
```

```
##                   2.5 %    97.5 %
## (Intercept) -10.053181 -2.610993
## Temperature   9.133106  9.283830
```

2.7 99% prediction interval on the steam usage in a month with average ambient tempreature of 58 degrees.

```r
(new_data <- data.frame(
  Temperature = c(58, 65, 65),
  Usage = c(455, 632, 423.09)
))
```

```
##   Temperature  Usage
## 1          58 455.00
## 2          65 632.00
## 3          65 423.09
```

```r
predict(model_1, new_data)
```

```
##        1        2        3
## 527.7590 592.2183 592.2183
```

```r
predict(model_1, new_data, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 527.7590 523.1644 532.3537
## 2 592.2183 587.4957 596.9410
## 3 592.2183 587.4957 596.9410
```

```r
predict(model_1, new_data, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 527.7590 526.2368 529.2813
## 2 592.2183 590.3448 594.0918
## 3 592.2183 590.3448 594.0918
```

```r
confint(model_1)
```

```
##                   2.5 %    97.5 %
## (Intercept) -10.053181 -2.610993
## Temperature   9.133106  9.283830
```