Danmarks Tekniske Universitet

02323 Introduction to Statistics Fall 22

Project 2: Heating in Sønderborg

November 10, 2022

Written by:
Marcus Wahlers Sand (s215827 / mwasa)

1 Introduction

The following report is handed in for the second mandatory project in the Introduction to Statistics course at DTU (02323).

> In this project, we'll formulate and select a suitable multiple linear regression model which describes the heat consumption during winter for four houses in Sønderborg. [2]

Below you will find the introductory text describing the project:

> In this project, we'll analyse data from the Sønderborg database, which consists of measurements of the heat consumption of buildings together with climatic measurements like outdoor temperature, solar irradiance, and wind speed. The buildings are typical single-family houses built with bricks during the 1950s and 60s. The houses are located in the area around Borgmester Andersens vej in Sønderborg and are of varying sizes. The climatic measurements were recorded by a weather station located at the local district heating plant.
>
> In this project, we'll only analyse the data from the winter of 2009/2010, defined as the period from 15 October 2009 to 15 April 2010. Furthermore, we'll only consider the houses with id numbers 3, 5, 10, and 17, respectively. [2]

We will also only be looking at data from the dataset, which doesn't have missing values for any of the variables.

2 Statistical analysis

2.1 Question a

Present a short descriptive analysis and summary of the data for the variables Q, Ta, and G.

2.1.1 Introduction to the variables

In the dataset, we have dated observations of the heat consumption in houses, identified by an integer, together with the daily averages of centrally observed climatic variables [2]. There are two climatic variables (Ta & G) and one heat consumption variable (Q). These variables are all quantitative. We also get a date variable, t, in order to tell at what time the observation took place, and a house identification variable, $houseId$, in order to differentiate the observations from each other, and link the data to a source.

The original dataset consists of 15.568 observations, ranging from October 2nd, 2008, to June 1st, 2011. The subset of the original dataset we will be analyzing in this project contains 576 observations, with a period ranging from October 15th, 2009, to April 15th, 2010, and where there are no missing values for any of the variables. Below you can find a table showing the variables (Table 1).

Variable	**Explanation**
t	Date
$houseId$	House identification number
Q	Heat consumption (kW)
Ta	Outdoor temperature ($°C$)
G	Solar irradiance (W/m^2)

Table 1: Table showing the variables in the dataset. [2]

2.1.2 Scatter Plots

We start by looking at scatter plots of the heat consumption against the outdoor temperature and the solar irradiance. Below (Figure 1), you can see a scatter plot between the heat consumption and the outdoor temperature.

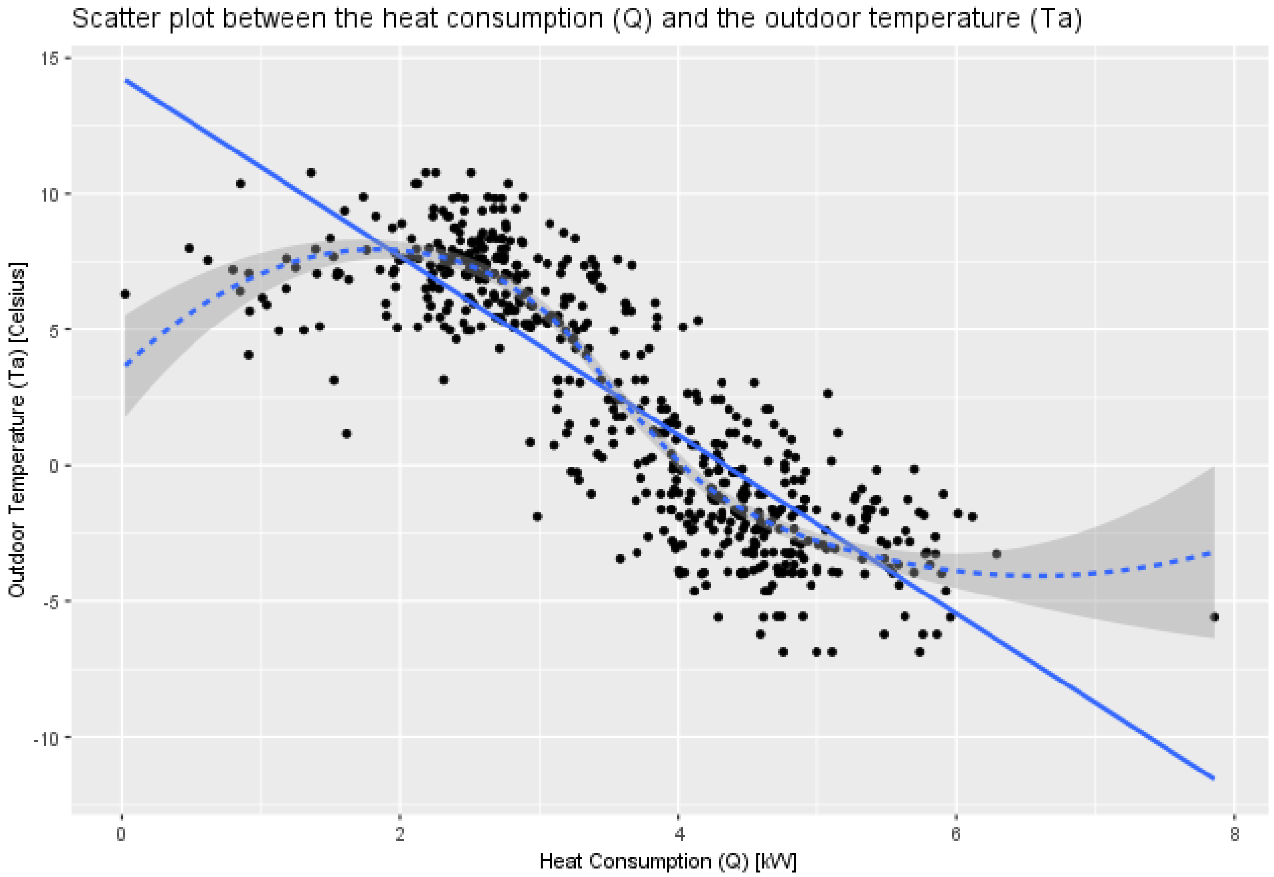


Figure 1: Scatter Plot showing the heat consumption against the outdoor temperature.

As we can see from the scatter plot, the general trend is negative, showing that a higher outdoor temperature generally means a lower heat consumption and that a lower outdoor temperature typically means a higher heat consumption - which would make sense since you would have to spend more energy to heat up and to keep your house warm, when the outdoor temperature is colder. The variables seem to correlate quite well.

Below (Figure 2), you can see a scatter plot between the heat consumption and the solar irradiance.

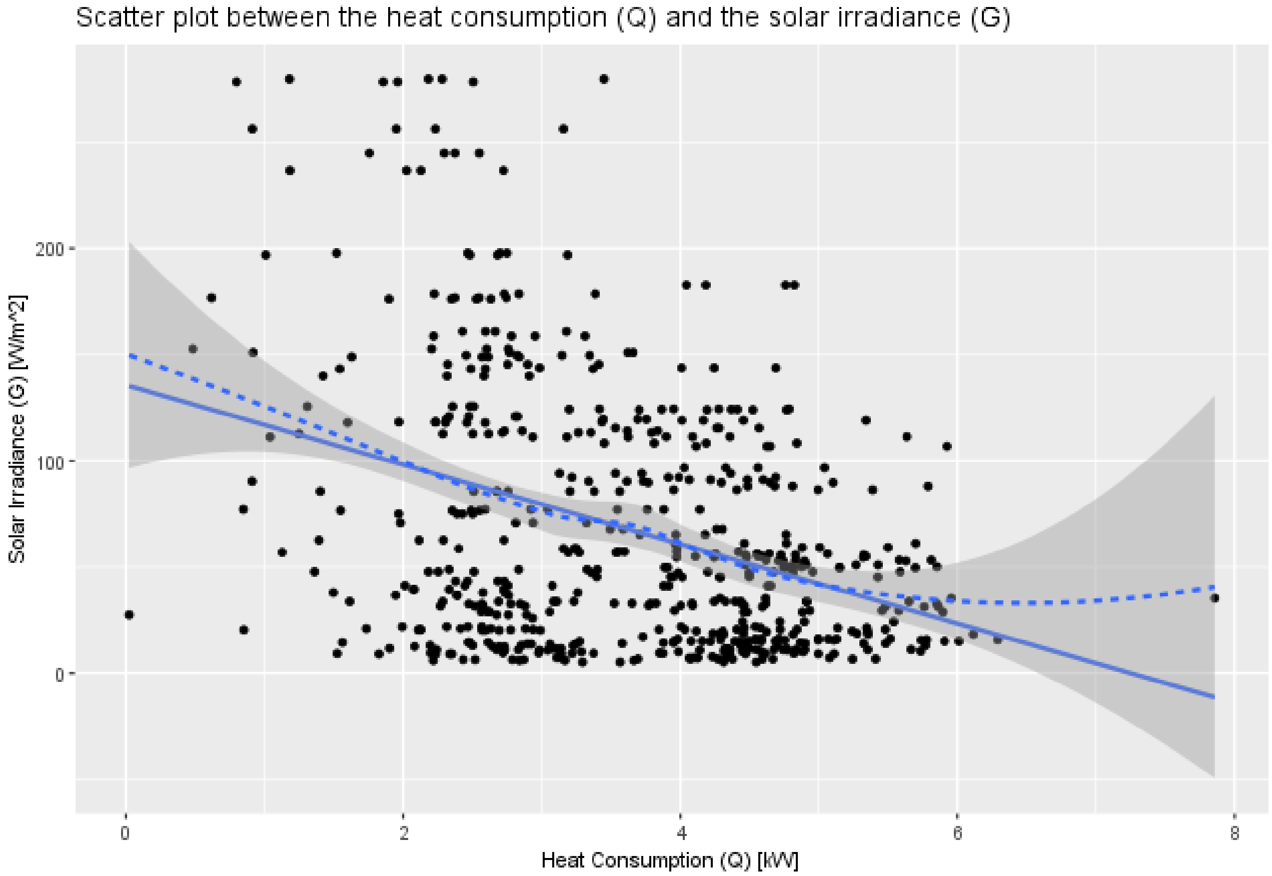


Figure 2: Scatter Plot showing the heat consumption against the solar irradiance.

As we can see on the scatter plot, the general trend is negative, showing that more solar irradiance generally leads to less heat consumption, and less solar irradiance generally leads to more heat consumption. The solar irradiance seems to have quite a big spread - but this is likely due to how fluctuating solar irradiance is. We can still see from the scatter plot that if we take the average of all the values, we tend to stay somewhat close to the regression line (though we will return to that later when we perform the actual linear regression model). We will be able to be a bit more certain about this after looking at the other figures.

2.1.3 Histograms

Now we will look at histograms for all the variables to learn more about the distribution of their empirical density. We start by looking at the empirical density of the heat consumption in the figure below (Figure 3).

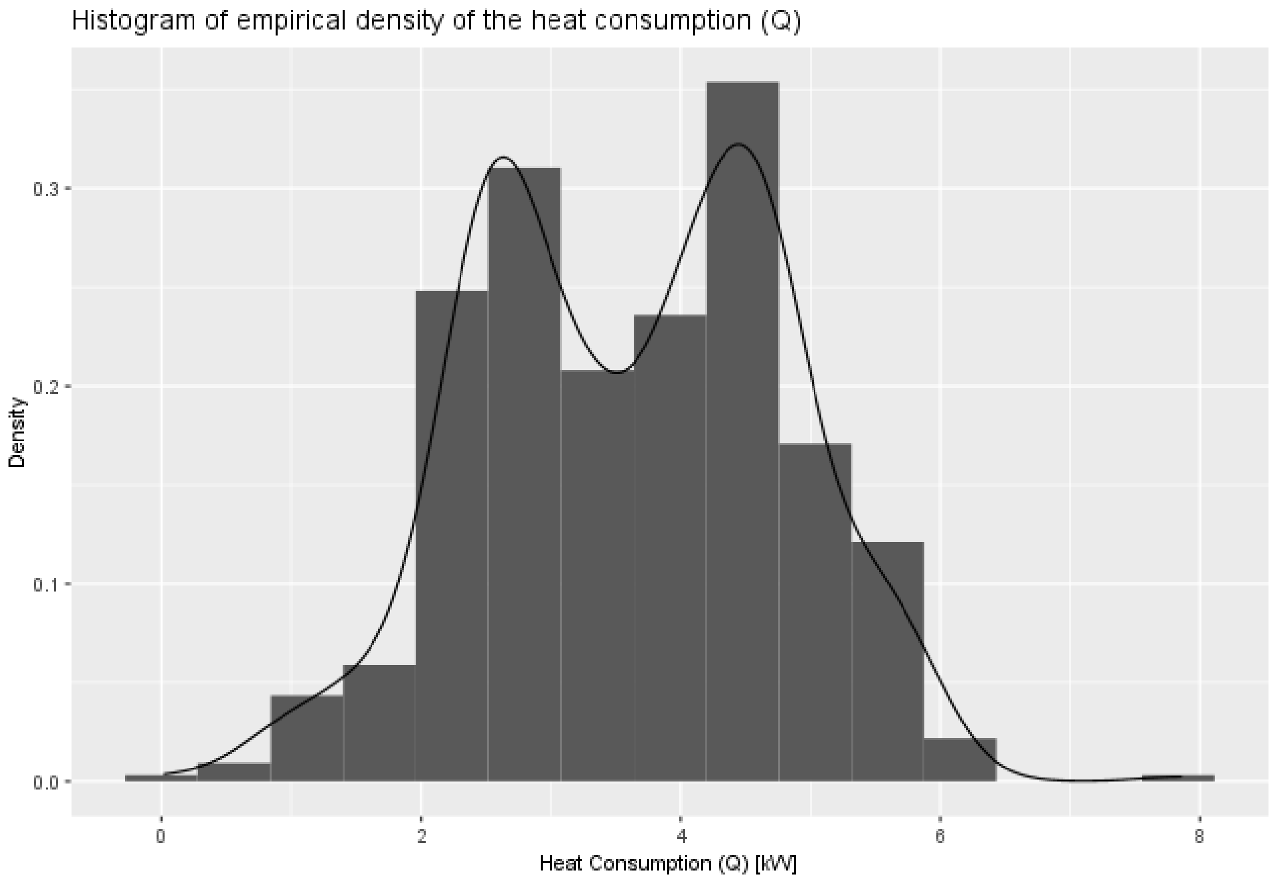


Figure 3: Histogram of the empirical density of the heat consumption.

As we can see on the histogram, the empirical distribution of the heat consumption is bimodal, with the two distinct "peaks" either being between $2kW$ and $3kW$ or between $4kW$ and $5kW$. The values range from around $0kW$ (having used no heat) to around $7.9kW$ (that small outlier at the very end of the histogram).

Next, we will look at the outdoor temperature's empirical density, which can be seen in the figure below (Figure 4).

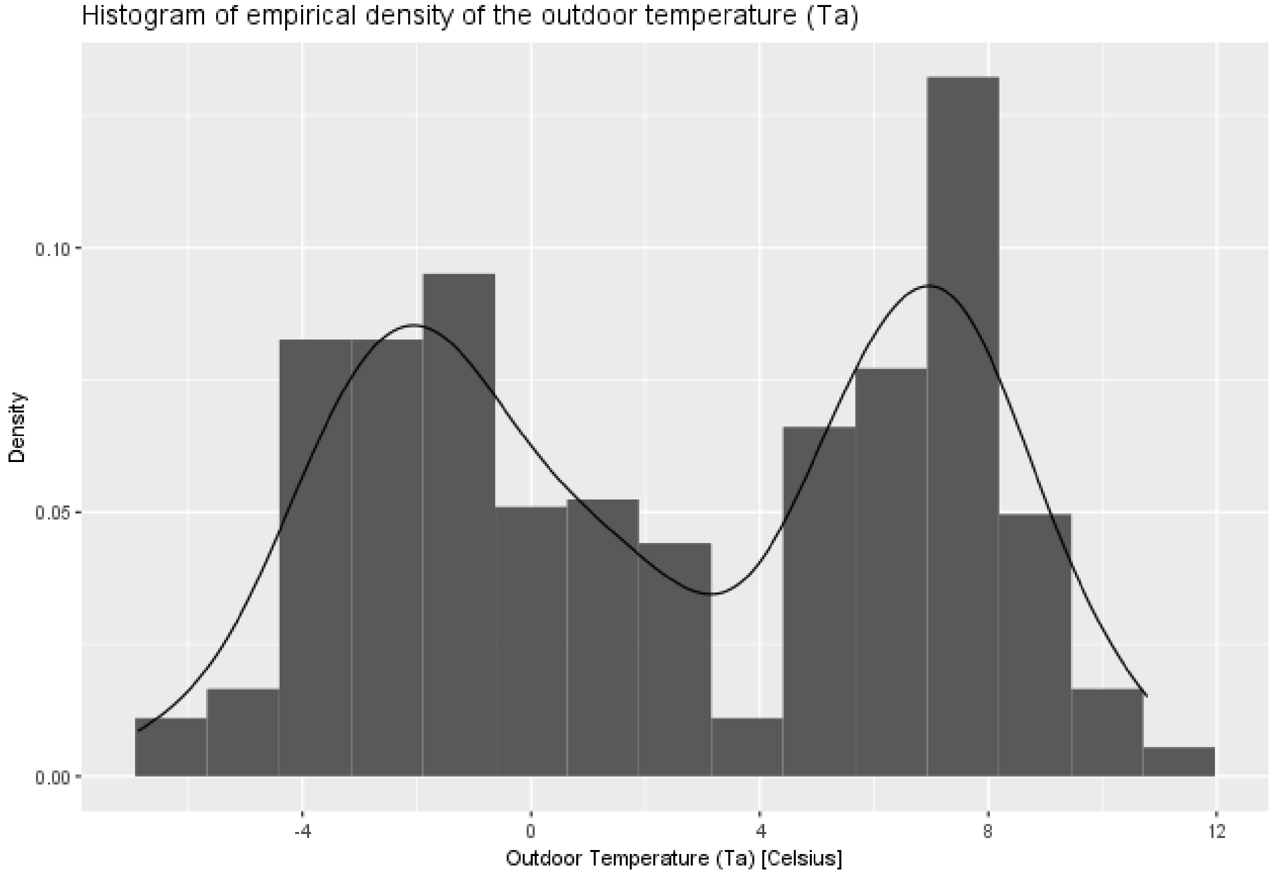


Figure 4: Histogram of the empirical density of the outdoor temperature.

As we can see on the histogram, this empirical distribution is also bimodal, with the two "peaks" in the density of the outdoor temperature being around $-2°C$ and $7°C$. This makes sense since it's relatively cold during half of the year, and the other half of the year, it's pretty warmer. The values range from around $-6.9°C$ to around $10.8°C$ (which is not that hot compared to our summers now, but that's probably due to climate change).

Finally, we will be looking at the empirical density of the solar irradiance, which can be seen in the figure below (Figure 5).

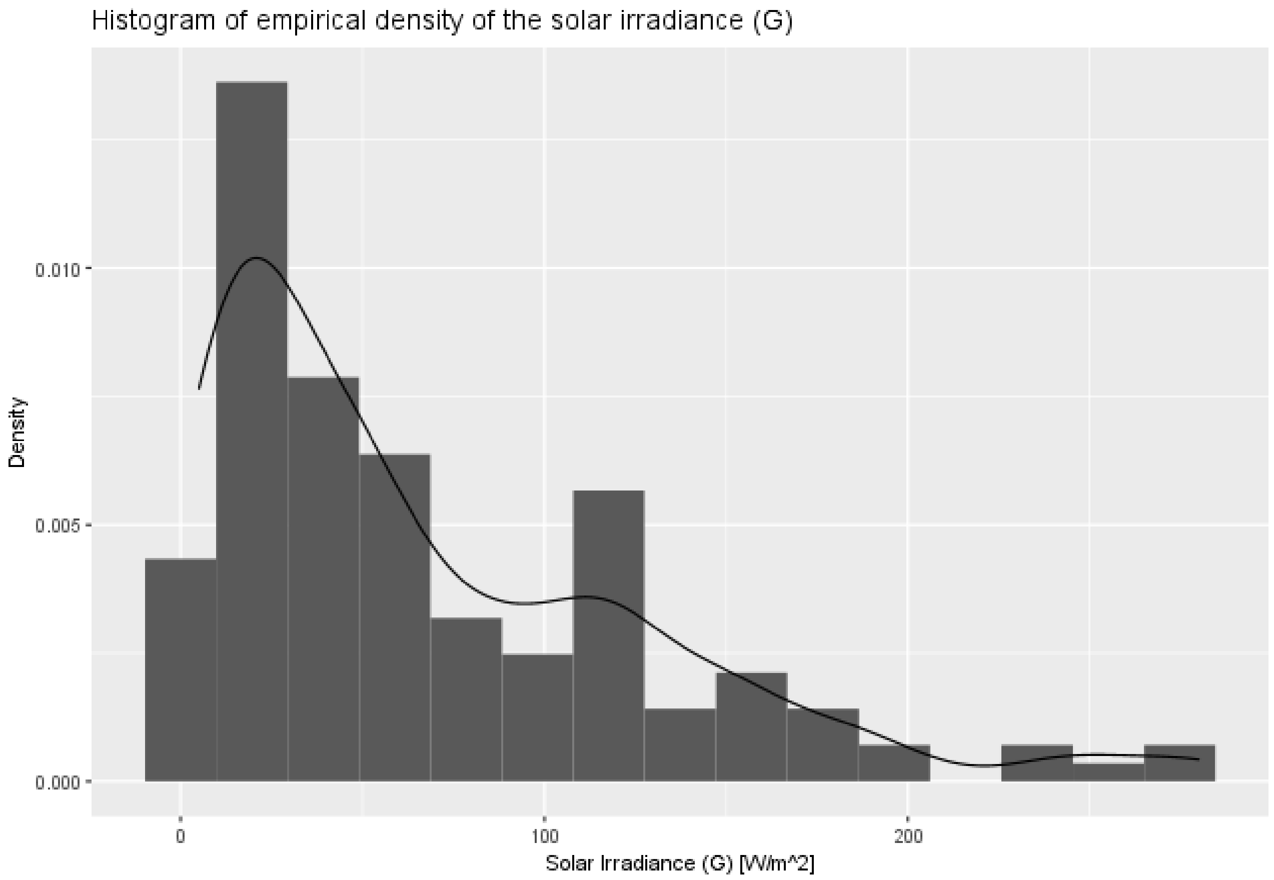
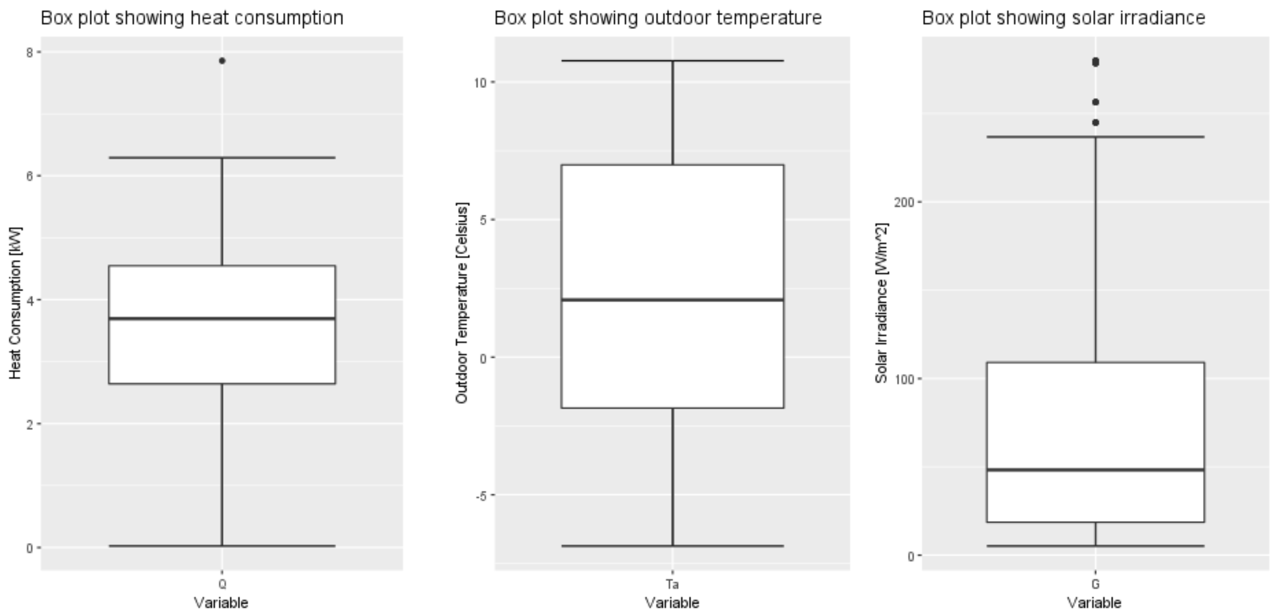


Figure 5: Histogram of the empirical density of the solar irradiance.

This empirical distribution of the solar irradiance is right-skewed (the mean and median are not centered and the tail is on the right side), and it ranges from around $5.1W/m^2$ to approximately $279.9W/m^2$, which is a seemingly big difference in values. But, as we can see on the histogram, most of the values seem to be around $50W/m^2$, just like we saw on the scatter plot.

2.1.4 Box plots

Now we will be taking a look at box plots for all three variables. The box plots can be found in the figure below (Figure 6).



(a) Box plot for the heat consumption. **(b)** Box plot for outdoor temperature. **(c)** Box plot for the solar irradiance.

Figure 6: Box plots for all variables.

Through the box plots, we can read the minimum value (within a distance of $1.5 \times IQR$ from Q_1), lower quartile (Q_1), median (Q_2), upper quartile (Q_3) and the maximum value (within a distance of $1.5 \times IQR$ from Q_3) for every variable.

The first box plot (Figure 6a) shows the distribution of the values for heat consumption. The box plot seems almost symmetrical, and from the plot, we can also see the extreme value that we also noticed from the histogram.

The second box plot (Figure 6b) is showing the distribution of values for the outdoor temperature. This plot also seems to be symmetrical.

The third box plot (Figure 6c) shows the distribution of the values for the solar irradiance. This plot seems asymmetrical, and we can also see the extreme values/outliers we noticed from the scatter plot and the histogram.

2.1.5 Table with summary statistics

Below you can find a table with the relevant summary statistics for the values covered in this question (Table 2).

	Q	Ta	G
Number of obs.	576	576	576
Sample mean	3.61	2.39	68.00
Sample variance	1.44	21.60	3815.78
Sample Std. dev.	1.20	4.65	61.77
Lower quartile	2.64	-1.85	18.59
Median	3.69	2.08	48.23
Upper quartile	4.55	6.99	109.05

Table 2: Relevant summary statistics for the variables Q, Ta, and G.

The table provides us with more "precise" information compared to the figures covered in the earlier parts of this question, but it is hard to visualize how the variables are distributed without proper visual figures. Here we can also see that the sample means and medians of the heat consumption and outdoor temperature are almost equal, which supports our earlier deduction that they are symmetrical - and likewise, the solar irradiance is not very symmetrical.

2.2 Question b

Formulate a multiple linear regression model with the heat consumption as the dependent/outcome variable (Y_i), and outdoor temperature and solar irradiance as the independent/explanatory variables ($x_{1,i}$ and $x_{2,i}$, respectively).

We start by formulating the multiple linear regression model, where Y_i is the heat consumption, $x_{1,i}$ is the outdoor temperature, and $x_{2,i}$ is the solar irradiance. We assume that the residuals (ε_i) are independent and identically distributed (i.i.d.) normal random variables with zero mean and some unknown constant variance (σ), and that $\varepsilon_i \sim N(0, \sigma^2)$ [1]:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

2.3 Question c

Estimate the parameters of the model, give an interpretation of the estimates, present the estimated standard deviations of the estimates, the degrees of freedom used for the estimated residual variance $\hat{\sigma}^2$, and the explained variation, R^2.

We will be estimating the parameters of the model by using the *lm(formula, data, ...)* function in R, which will perform the linear least squares regression for us, and then we will use the *summary()* function in R to view the parameters. In the tables below (Table 3 and Table 4) you can see an overview of the relevant values for this question which have been extracted from the mentioned functions:

Coefficient	Estimated effect	$\hat{\sigma}$	$\hat{\sigma}^2$
β_0	4.2788	0.0389	$1.5092 \cdot 10^{-3}$
β_1	-0.2090	0.0058	$3.3942 \cdot 10^{-5}$
β_2	-0.0025	0.0004	$1.9219 \cdot 10^{-7}$

Table 3: Model coefficients and their estimated regression coefficients, estimated standard deviations/errors, $\hat{\sigma}$, and estimated standard variations, $\hat{\sigma}^2$.

$\hat{\sigma}$	$\hat{\sigma}^2$	R^2
0.62	0.39	$0.73 = 73.25\%$

Table 4: Estimated residual standard error/deviation, $\hat{\sigma}$, estimated residual variance, $\hat{\sigma}^2$, with 573 degrees of freedom, and the explained variation, R^2.

We insert the estimated parameters of the model into the model:

$$\begin{array}{rcllllllllll} \hat{y}_i & = & \hat{\beta}_0 & + & \hat{\beta}_1 x_{1,i} & + & \hat{\beta}_2 x_{2,i} & + & \varepsilon_i & , & \varepsilon_i \sim N(0,\sigma^2) & \text{and i.i.d.} \\ & = & 4.279 & - & 0.209 x_{1,i} & - & 0.002 x_{2,i} & + & \varepsilon_i & , & \varepsilon_i \sim N(0,\sigma^2) & \text{and i.i.d.} \end{array}$$

We interpret the estimate $\hat{\beta}_0$ as the interception with the y-axis. We intercept $\hat{\beta}_1$ and $\hat{\beta}_2$ as the effect of the variables given the other variables, i.e., for $\hat{\beta}_1$, the effect of $x_{1,i}$ (the outdoor temperature) accounting for the effect of $x_{2,i}$ (the solar irradiance).

$\hat{\beta}_1$ tells us that, in this model, the outdoor temperature has a more significant influence on heat consumption than solar irradiance. Due to both values being negative, we can tell that when the explanatory variables ($\hat{\beta}_1$ & $\hat{\beta}_2$) increases, our response variable ($\hat{y}_i$) decreases. According to our model, for every one percent increase in outdoor temperature, there should be a -0.2% decrease in heat consumption.

Now we will take a look at the estimated standard deviations/errors of $\hat{\beta}_0$ ($\hat{\sigma}_{\beta_0}$), $\hat{\beta}_1$ ($\hat{\sigma}_{\beta_1}$) and $\hat{\beta}_2$ ($\hat{\sigma}_{\beta_2}$), the degrees of freedom used for the estimated residual variance, $\hat{\sigma}^2$, and the explained variation, R^2.

The standard deviations tell us how much variation there is around our coefficient's estimated effects.

The degrees of freedom used for the estimated residual variance, $\hat{\sigma}^2$, is calculated using Theorem 6.2 from the book [1], which means that our degrees of freedom become $n-(p+1)$, where p is the number of explanatory variables. So, with 576 observations and 2 explanatory variables, we get 573 degrees of freedom.

The value of the explained variation, R^2, explains "how well" the model is able to place a regression line that is able to estimate our response variable. 73.25% is an acceptable amount of variation and shows that there is quite a bit of correlation between our explanatory variables and our response variables.

2.4 Question d

Perform model validation with the purpose of assessing whether the model assumptions hold.

We first validate our normal assumption with a Q-Q plot (Figure 7):

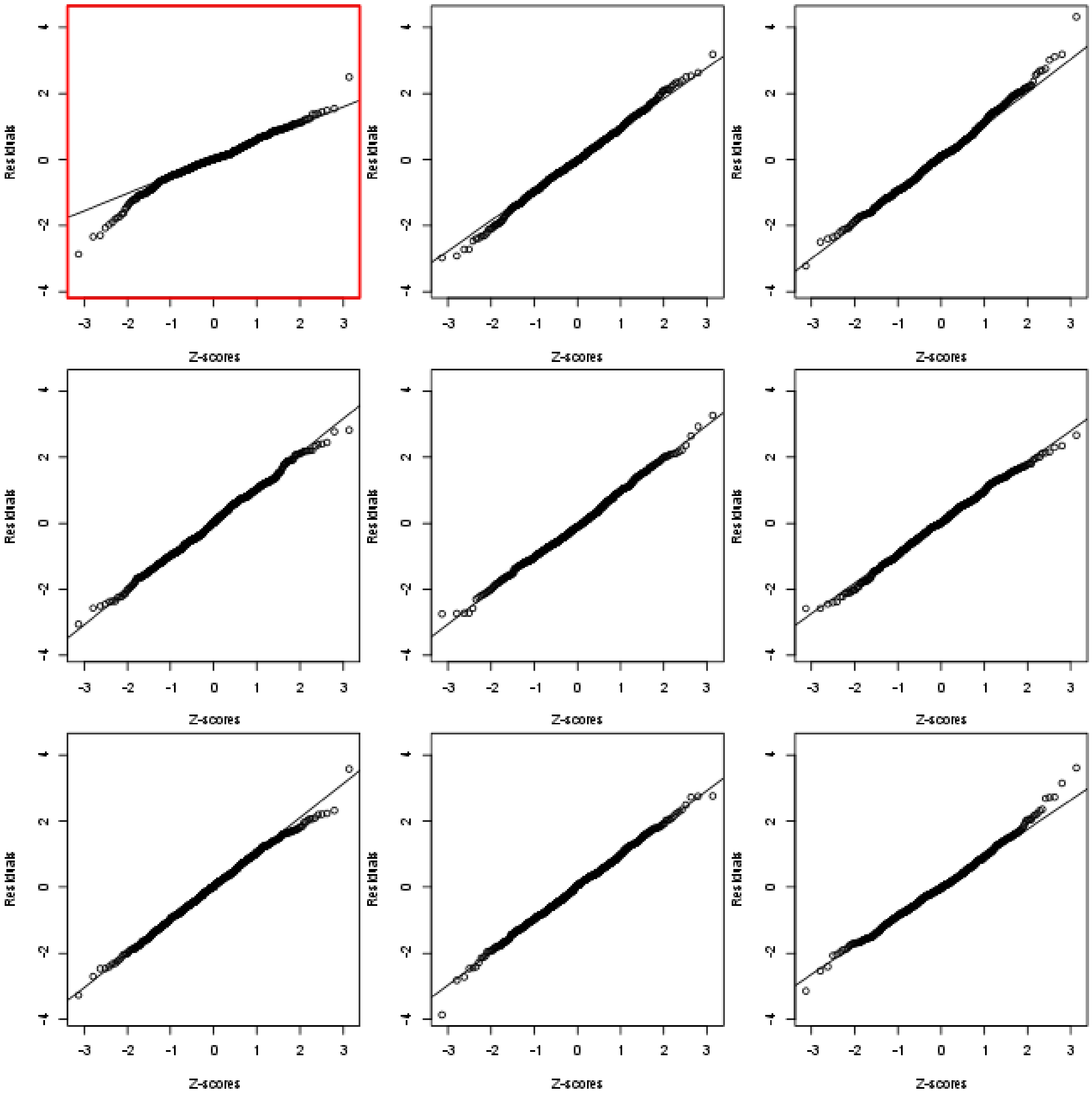


Figure 7: Multiple (simulated) Q-Q Plots for the residuals.

We can see that the distribution is relatively symmetric, though there is a slightly heavy tail. As can be observed from the figures, the non-simulated plot (marked with a red border) doesn't stick out from the crowd of simulated Q-Q plots by a lot. We will therefore assume that the normality assumption is OK.

We can then check if our observed values look like our fitted values - this can be checked in the figure below (Figure 8):

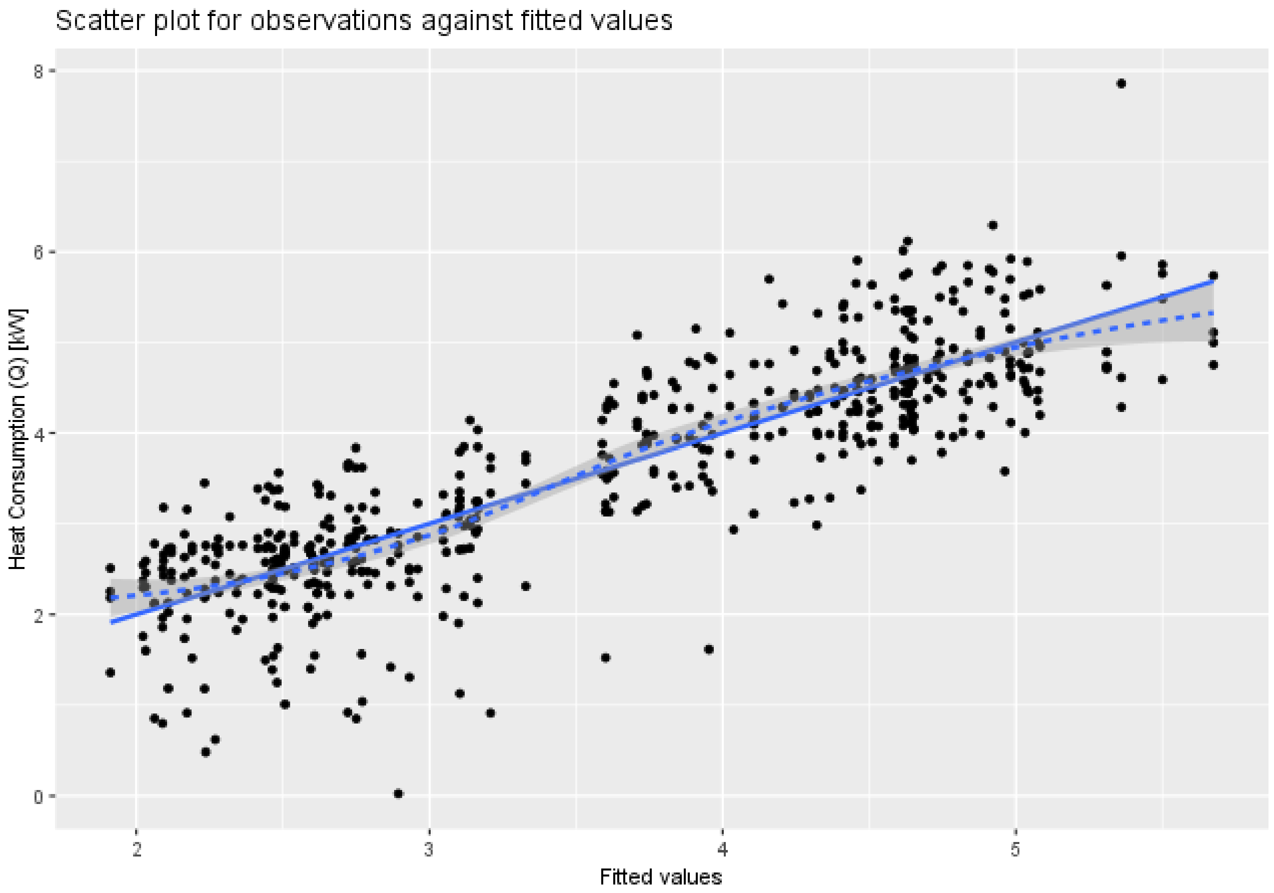
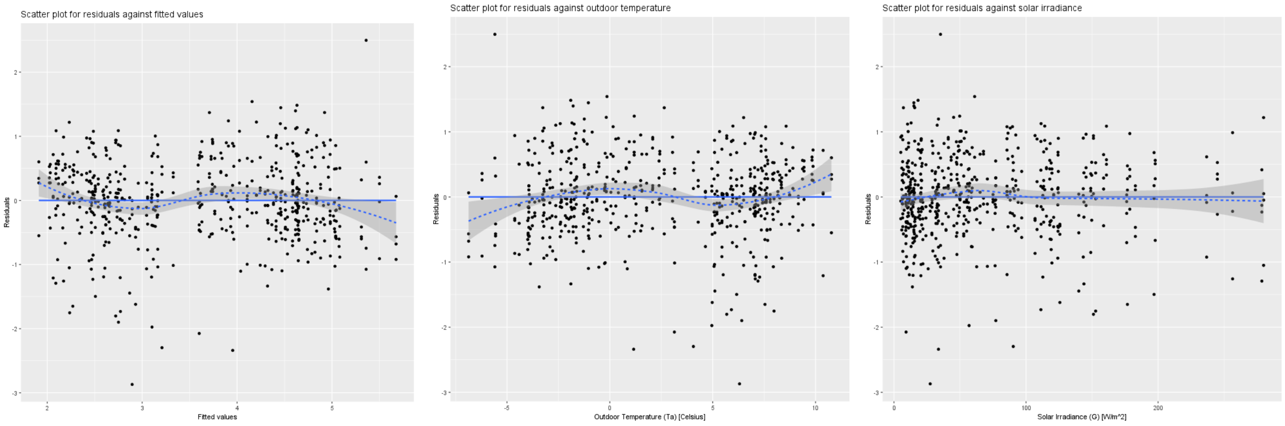


Figure 8: Scatter plot for observations against the fitted values.

From the plot, it looks like the fitted values are very close to our observed values, but there is still some prediction error, as you can see slight deviations between the fitted values and the observed values.

We will now be checking the systematic behavior by plotting the residuals, ε_i as a function of the fitted valued, $\hat{y}_i$, and the explanatory variables, *Ta* & G, to check for a constant range and mean since we assumed earlier that ε_i is i.i.d. and that $\varepsilon_i \sim N(0, \sigma^2)$. The plots can be found below (Figure 9):



(a) Residuals vs. fitted values.

(b) Residuals vs. outdoor temperature.

(c) Residuals vs. solar irradiance.

Figure 9: Scatter plots for checking systematic behavior.

We can see that there is no systematic dependence in the residuals, as the range and mean values are constant. We will therefore assume that the model is valid.

2.5 Question e

State the formula for a 95% confidence interval for the outdoor temperature coefficient, here denoted by β_1. Insert numbers into the formula, and compute the confidence interval.

We use Method 6.5. of the book to calculate the confidence interval [1]. Because we want the 95% confidence interval, we pick an a-value of 0.05, which means that we have to find $1 - a/2$, then we get 0.975. This means we need to calculate $t_{0.975}$ to calculate our confidence interval. One can use R to find the value or look the value up in a t-value table since to calculate the t-distribution, there is the need for an unobservable value, μ. The t-distribution function is from Theorem 3.5 [1]. We got the p-value from the *lm(formula, data, ...)* function in R.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n - (p+1))$$

$$t_{0.975} = qt(0.975, n - (p+1)) = qt(0.975, 576 - ((2+1)) \approx 1.96$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_1} \implies \hat{\beta}_1 \pm t_{0.975} \cdot \hat{\sigma}_{\beta_1} \implies -0.21 \pm 1.96 \cdot 0.006 = [-0.22, -0.20]$$

Now we can run *confint(fit, level* = 0.95*)* from R to check our result and to determine confidence intervals for the two other regression coefficients. The confidence intervals can be found in the table below (Table 6).

	Lower bound of CI	**Upper bound of CI**
(Intercept)	4.202	4.355
Ta	-0.220	-0.198
G	-0.003	-0.002

Table 6: 95% confidence interval (CI) for the regression coefficients.

It looks like our calculation was correct.

2.6 Question f

It is of interest whether β_1 might be -0.25. Formulate the corresponding hypothesis. Use the significance level $\alpha = 0.05$. State the formula for the relevant test statistic, insert numbers and compute the test statistic. State the distribution of the test statistic (including the degrees of freedom), compute the p-value, and write a conclusion.

This can be done by testing the following hypothesis:

$$H_{0,1} : \beta_1 = -0.25,$$

$$H_{1,1} : \beta_1 \neq -0.25.$$

We will use $\alpha = 0.05$ (5%) as the significance level. This is our "risk level".

Below is the formula for calculating the value of the test statistic (Method 6.4 [1]):

$$t_{\text{obs},\beta_i} = \frac{\beta_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$$

Now we will be calculating the value of the test statistic, with 573 $(n-(p+1))$ degrees of freedom:

$$\begin{aligned} t_{\text{obs},\beta_1} &= \frac{\beta_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}} \\ &= \frac{-0.21-(-0.25)}{0.01} \\ &= \frac{-0.21+0.25}{0.01} \\ &= \frac{0.04}{0.01} \\ &= 7.05 \end{aligned}$$

Below you can see the distribution of the test statistic, with 573 $(n-(p+1))$ degrees of freedom (Figure 10).

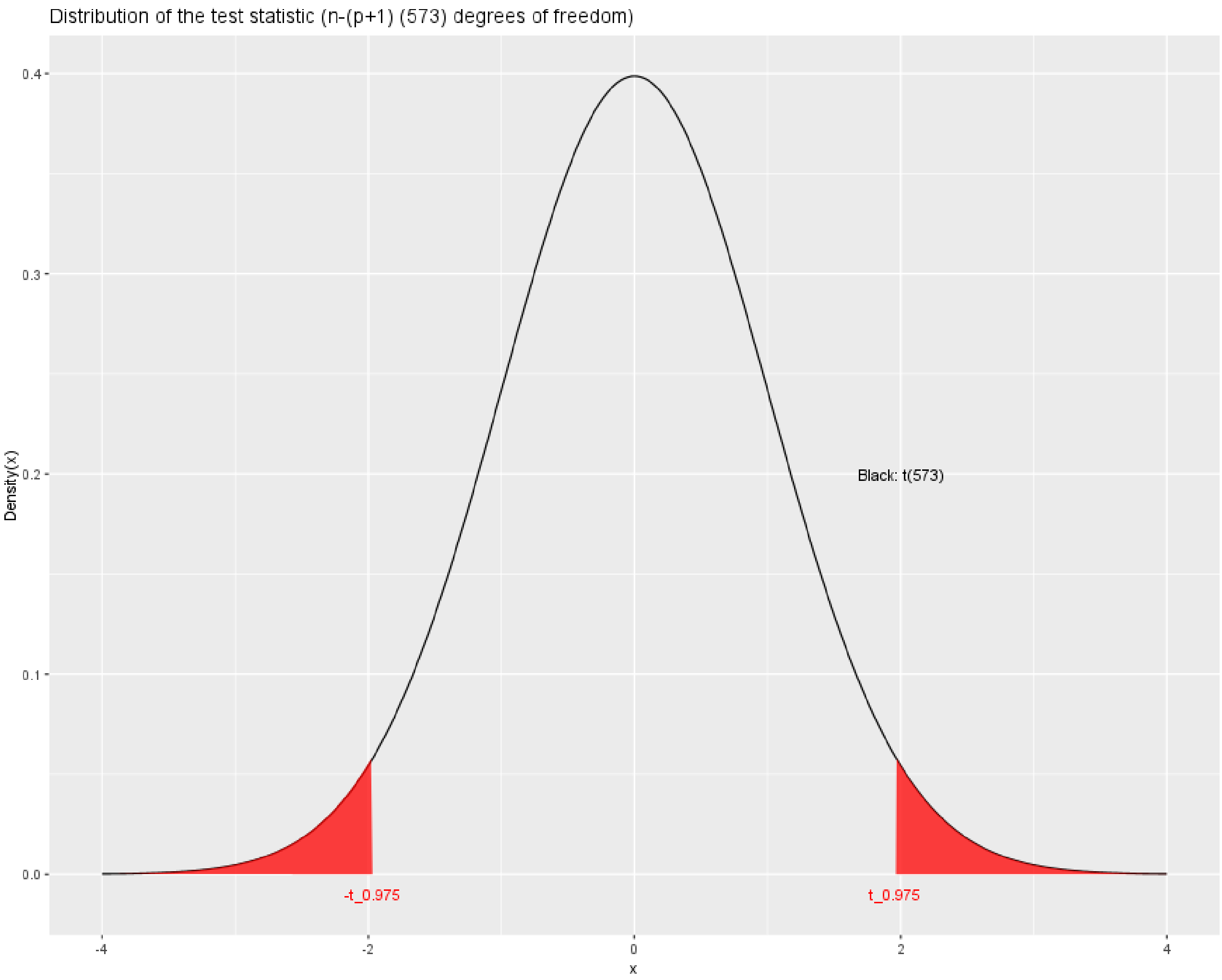


Figure 10: Distribution of the test statistic, with 576 observations, making the degrees of freedom for the distribution 573 $(n-(p+1))$.

And now the p-value:

$$\begin{aligned}
p\text{-value} &= 2 \cdot P(T > |t_{obs}|) \\
&= 2 \cdot (1 - pt(abs(t_{obs}), df = n - (p+1))) \\
&= 2 \cdot (1 - pt(abs(7.05), df = 576 - (2+1))) \\
&= 2 \cdot (1 - pt(abs(7.05), df = 573)) \\
&= 2 \cdot (1 - 1) \\
&= 2 \cdot 0 \\
&= 0
\end{aligned}$$

We don't even have to consult the p-value table (Table 3.1 [1]) to know that this is very strong evidence against H_0. It is way below our significance level, $\alpha = 0.05$ (5%) - and as such, we discard H_0 and go with the alternative hypothesis, that $\beta_1 \neq -0.25$.

This outcome was expected, as -0.25 also was well below the bounds of the outdoor temperature's confidence interval.

2.7 Question g

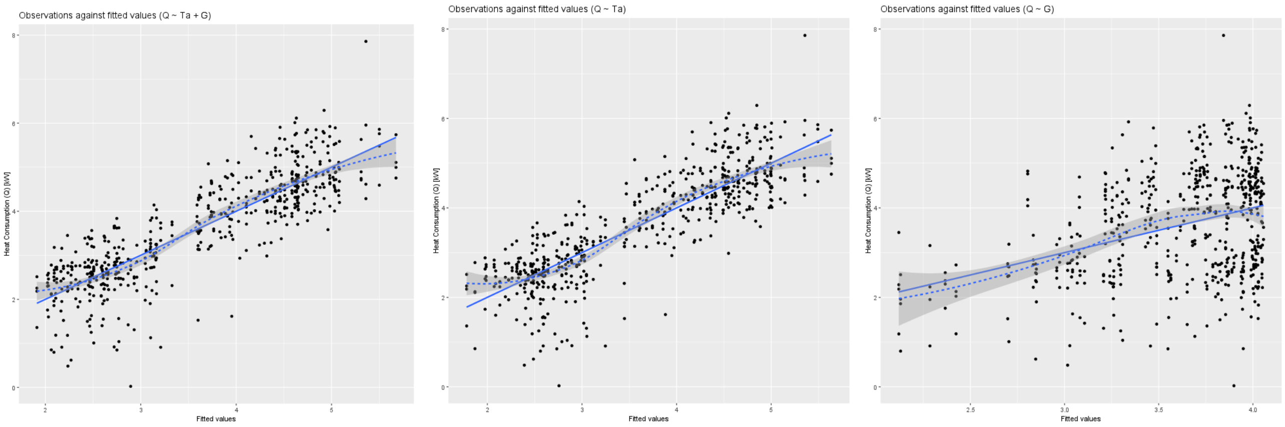
Use backward selection to investigate whether the model can be reduced.

We will start out by looking at the p-values for all variables in the model. The variable with the biggest p-value is the Solar Irradiance, G, so we will be trying to perform backward selection on G (even though all of the p-values are well below the value required to be significant). Keep in mind that the value of R^2 for the current model is 73.25%.

After removing the Solar Irradiance, G, as a variable in our model, the p-values for the other variables remain the same. Our value of R^2, however, drops down to 71.78%.

Just for the sake of it, let's try to perform backward selection on the Outdoor Temperature, Ta, as well. After completing the fit with Q and G, the p-value for G drops down to the same value that Ta had before. However, our value of R^2 has dropped significantly, making the model's value of R^2 become 13.19%.

In the figure below (Figure 11), you can compare the observations and the fitted values of all three models.



(a) Observations vs. fitted values, original model ($Q \sim Ta + G$).

(b) Observations vs. fitted values, second model ($Q \sim Ta$).

(c) Observations vs. fitted values, third model ($Q \sim G$).

Figure 11: Scatter plots for comparing models after performing the backward selections.

From the figures, it is also clear that the first two models have a better fit.

Though it might be out-of-scope for the course, it is also possible to compare the models using AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), where a lower value is better. The AIC and BIC "punish" you for overfitting your model, i.e., they try to make you use a model that is not complex just for the sake of being complex but using variables that positively influence the model. The AIC for a model can be found using the *AIC(<model>)* function in R. Backward selection/elimination can also be performed using the *step(<model>, direction = "backward")* function in R. In the listing below (Listing 1), you can see the values for both AIC and BIC for all three models.

```
> # The first/original model
> fit <- lm(Q ~ Ta + G, data = D_model)
> c(AIC(fit), BIC(fit))
[1] 1090.816 1108.240

> # Backward selection, removing the solar irradiance, G
> fit2 <- lm(Q ~ Ta, data = D_model)
> c(AIC(fit2), BIC(fit2))
[1] 1119.552 1132.620

> # Backward selection, removing the outdoor temperature, Ta
> fit3 <- lm(Q ~ G, data = D_model)
> c(AIC(fit3), BIC(fit3))
[1] 1766.816 1779.885
```

Listing 1: Lexer from hardware.g4

Hence, it would seem that the "best model" (out of these three possible models) would be our original model:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.} \\ &= 4.279 - 0.209 x_{1,i} - 0.002 x_{2,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.} \end{aligned}$$

2.8 Question h

Determine predictions and 95% prediction intervals for the heat consumption for each of the four observations in the validation set (D_{test}). Compare the predictions to the observed heat consumptions in the validation set and assess the prediction capabilities of the final model.

The validation set (D_{test}) is a subset of the original data, which contains observations from the winter of 2008/2009. We'll use this subset to evaluate the prediction capabilities of the final model [2].

By using the *predict(<model>, newdata = <validation set>, interval = "prediction", level = 0.95)* function in R, we can calculate predictions and prediction intervals for the heat consumption from the explanatory variables in the validation set, and then compare them to the actual heat consumption of the validation set. The output of the said function can be found in the table below (Table 10).

Index	*houseId*	Actual value	Predicted value	Difference	Lower bound of PI	Upper bound of PI
66	3	3.53	3.41	−0.12	2.19	4.63
1117	5	4.09	3.85	−0.24	2.63	5.07
5027	10	3.13	3.07	−0.06	1.84	4.29
11858	17	2.60	2.76	+0.16	1.54	3.98

Table 10: Actual values from the validation set, predictions, the difference between actual and predicted values, and the 95% prediction interval (PI) for the heat consumption for each of the four observations in the validation set.

We can now compare the predictions to the observed heat consumption from the validation set - and on average, the difference between the predicted and actual values is "only" 0.14 kW. We can also see that all the actual values are within our prediction interval. So, our model's ability to predict values generally seems satisfactory.

A Appendix

A.1 R Code

Please refer to the R-code, which was handed in with this submission.

B Bibliography

[1] Per B. Brockhoff, Jan K. Møller, Elisabeth W. Andersen, Peder Bacher, and Lasse E. Christiansen. *Introduction to Statistics at DTU*. Polyteknisk Kompendie, May 2022.

[2] DTU. Project 2: Heating in Sønderborg II. https://www2.compute.dtu.dk/courses/introstat/projects/soenderborg2.zip, 2022.