STAT 5320 - ASSIGNMENT 2

Pham Thi Thai - T00727094
John Joshua Bardelosa - T00728432

March 23, 2024

1 Electricity consumption

1.1. There are multiple temperature measurements from the weather data. State which one you'll use as your outdoor temperature measurement for this investigation.

Answer

The chosen outdoor temperature measurement for this investigation will be the Mean Temperature (^{0}C).

1.2. Create a simple linear regression model where the response is the consumption and the predictor is the temperature variable. Present the equation of the line from your results and justify whether consumption increases, decreases, or is not significantly related to the outside temperature.

Answer

The equation of the simple linear regression where the response is the Net consumption and the predictor is Mean temperature:

$$y = 15.429 - 0.196x$$

(refer to Figure 1 for more detail.)

The coefficient of slope, -0.196, indicates that for every degree C increase in Mean temperature, the Net consumption decreases by 0.196 KWh. The p-value, which is approximately equal to 0, indicating that Mean temperature is a *significant* predictor of Net consumption.

1.3. Find the point estimate (not intervals, just give me the $\hat{y}$ values for the average electricity consumption) when outside temperatures are -40^{0}C, -20^{0}C, 0^{0}C, 20^{0}C, 40^{0}C.

Answer

Outside Temperature (^{0}C)	Estimated consumption (KWh)
-40	23.269
-20	19.349
0	15.429
20	11.509
40	7.589

1.4. Create a plot of the data and overlay the equation of the line. Just based on the graph, you'll likely feel unsatisfied. In a couple sentences, discuss whether which of the estimated values from the previous questions seem reasonable and which you feel unsatisfied with.

Answer

- Refer to Figure 2. for a plot of the data and overlay the equation of the line.

- Since the scatter plot forms a curved shape, it suggests that the relationship between the variables may not be linear, which could indicate that a linear regression model may not be the best fit for the data. In this case, the estimated values provided by the linear regression equation might not accurately represent the true relationship between the mean temperature and net consumption. Thus, while the point estimates provided by the linear regression model may give a rough approximation, they may not fully capture the complexity of the relationship, leaving room for uncertainty and potentially unsatisfactory predictions, especially at extreme temperature values.

1.5. As in lecture 2.2, create a new regression model with an additional term that does a better (but not necessarily the best) job to addresses your dissatisfaction from the previous model.

a. Identify the equation of your new linear model

Answer

The equation of new linear model is a quadratic function of MeanTemp:

$$y = 13.775 - 0.657x + 0.029x^2$$

(refer to Figure 3 for more detail.)

b. Very carefully use the regression model to answer the naïve question "does the resident uses more or less electricty when the outdoor temperature rises?".

Answer

Taking the derivative of y with respect to x of the above quadratic model, obtain the result

$$\frac{\partial y}{\partial x} = -0.657 + 0.058x$$

Hence, for a specific value of the outside temperature (x), the change in the net consumption associated with an additional 1^0C change in the mean temperature is $-0.657 + 0.058x$. It is not feasible to definitively ascertain whether an increase in outside temperature would result in higher or lower electricity usage by residents.

1.6. Estimate the average electricity use for -40^0C, -20^0C, 0^0C, 20^0C, 40^0C again, but with your new model. Discuss which provide better or worse estimates than the first model.

Answer

Outside Temperature (^{0}C)	Estimated consumption (KWh)
-40	87.110
-20	38.674
0	13.775
20	12.412
40	34.585

Model 2, the quadratic regression, outperforms Model 1, the linear regression. It exhibits a lower residual standard error (3.321 vs. 5.533), indicating better fit. Additionally, Model 2 demonstrates substantially higher R-squared values (0.685 vs. 0.125), implying greater variance explanation. Consequently, the quadratic model is better for predicting Net.consumption based on MeanTemp.

1.7. Create another scatterplot and this time overlay your new linear model. Clearly identify your 5 estimated values and discuss whether the fit is better.

Answer

- For Scatter plot and the linear regression line for new model, refer to Figure 4. Note that 5 estimated values are clearly identified by red color.

- The quadratic model provides a better fit to the data compared to the linear model, as it captures the non-linear relationship between Mean Temperature and Net Consumption more effectively, as observed in both the scatter plot and the model curve.

2 Bushes Data

An analysis of various ornamental bushes is conducted on plants which are left alone anduntrimmed with common light, water, and nutrients after 3 years.

2.1. Ignore the find the equation of the simple linear regression line modelling height of the plant as a function of the width of the plant. Create a plot that again ignores the species and overlay your regression line. Discuss whether the result of the line matches what you see in this plot, how to interpret the slope, whether it's significant.

Answer

The scatter plot (refer to Figure 5) suggests no clear linear relationship between height and width, with data points forming clusters and showing linear patterns within each cluster. However, the regression line indicates a negative slope, implying that as height decreases, width tends to increase. This regression line accurately reflects the observed relationship between height and width on the scatter plot, hence the slope is significant.

2.2. Create a new regression line using two main effects: the width and the species. This time, create the plot distinguishing species. Draw each of the submodels for each species.

Answer

- Following is the equation of new regression line using two main effects, the width and the species:

$$y = 3.028 + 0.961x_1 - 2.062x_2 - 3.020x_3 - 6.363x_4 - 7.242x_5$$

where x_1 is the Width, x_2, x_3, x_4, x_5 are the species B, C, D, E, respectively (species A is the reference level group). (Refer to Figure 6 for more detail.)

- For the plot distinguishing species, refer to Figure 7 and Figure 8.

2.3. Interpret the coeffcient corresponding to the width of the second regression line, whether it's significant, and what makes the answer very very weird compared to your first regression line.

Answer

- The coefficient corresponding to the width in the second regression line is 0.96119. This means that for each unit increase in width, the height is expected to increase by approximately 0.96119 units, holding other variables constant. The coefficient is highly significant, as indicated by the very low p-value ($< 2.2e^{-16}$), suggesting that the relationship between width and height is statistically significant.
- In the first regression line, the coefficient for width is negative, suggesting that as the width increases, the height tends to decrease. However, in the second regression line, which includes species as an additional predictor, the coefficient for width is positive. This indicates that as the width increases, the height tends to increase.

2.4. How would you answer "do wider bushes correspond to taller or shorter bushes?" in these data?

Answer

The relationship between width and height of bushes is not straightforward and depends on the variables considered in the analysis. In the model considering only width, wider bushes tend to be shorter. However, when considering species as well, wider bushes tend to be taller. Therefore, the relationship between width and height varies depending on the context of the analysis.

3 Appendix

```
Call:
lm(formula = Net.consumption ~ MeanTemp, data = edata)

Residuals:
   Min     1Q Median     3Q    Max
-9.721 -3.929 -1.100  2.995 31.968

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.42925    0.22868   67.47   <2e-16 ***
MeanTemp    -0.19592    0.01581  -12.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.533 on 1075 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.125,     Adjusted R-squared:  0.1242
F-statistic: 153.6 on 1 and 1075 DF,  p-value: < 2.2e-16
```

Figure 1: SLR for Electricity consumption

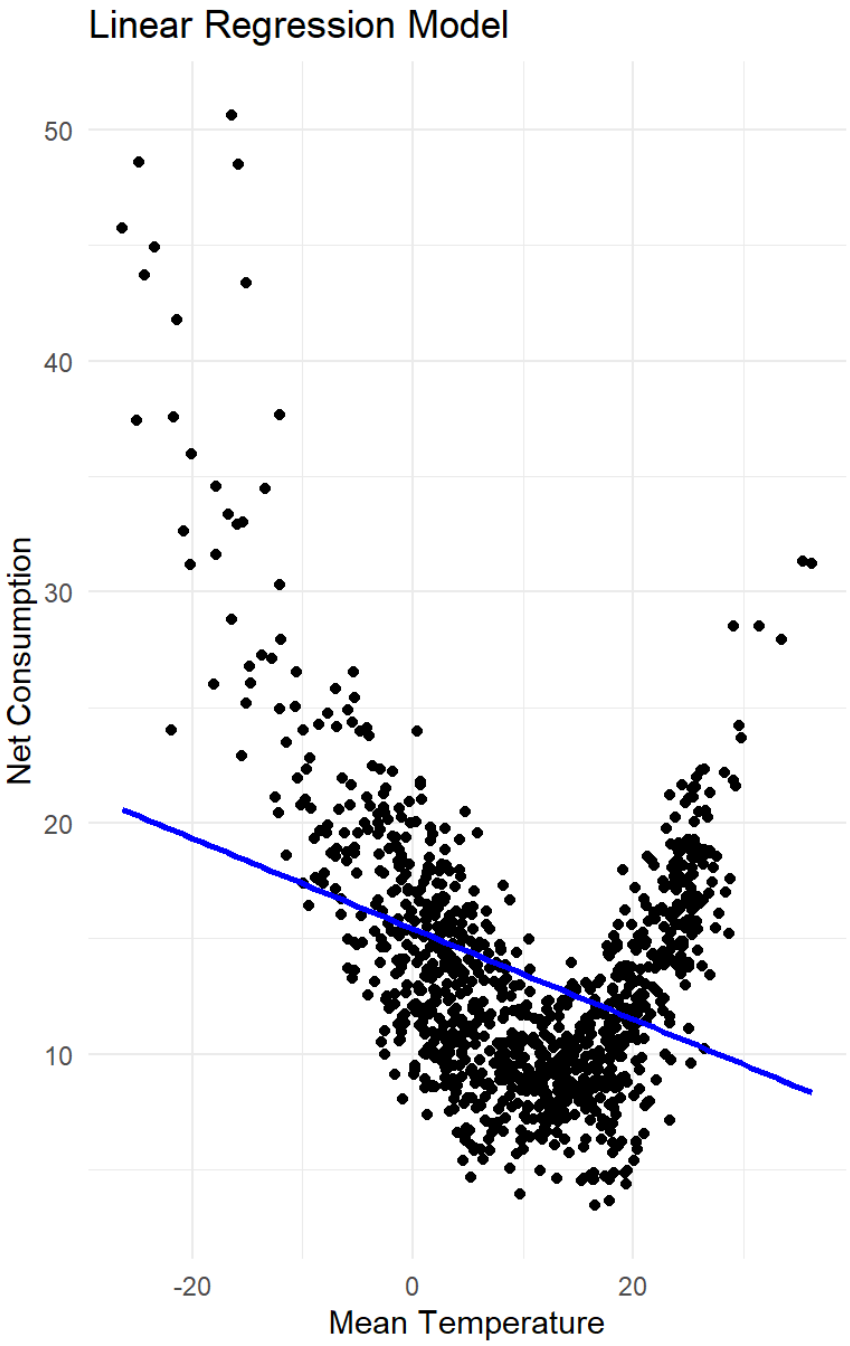

Figure 2: Scatter plot and LR line for electricity consumption

```
Call:
lm(formula = Net.consumption ~ MeanTemp + I(MeanTemp^2), data = edata)

Residuals:
     Min       1Q   Median       3Q      Max
-18.2637  -2.0953  -0.2644   2.1329  18.1548

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.7747085  0.1423973   96.73   <2e-16 ***
MeanTemp      -0.6565577  0.0141839  -46.29   <2e-16 ***
I(MeanTemp^2)  0.0294206  0.0006733   43.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.321 on 1074 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.685,     Adjusted R-squared:  0.6844
F-statistic:  1168 on 2 and 1074 DF,  p-value: < 2.2e-16
```

Figure 3: New LR model for Electricity consumption

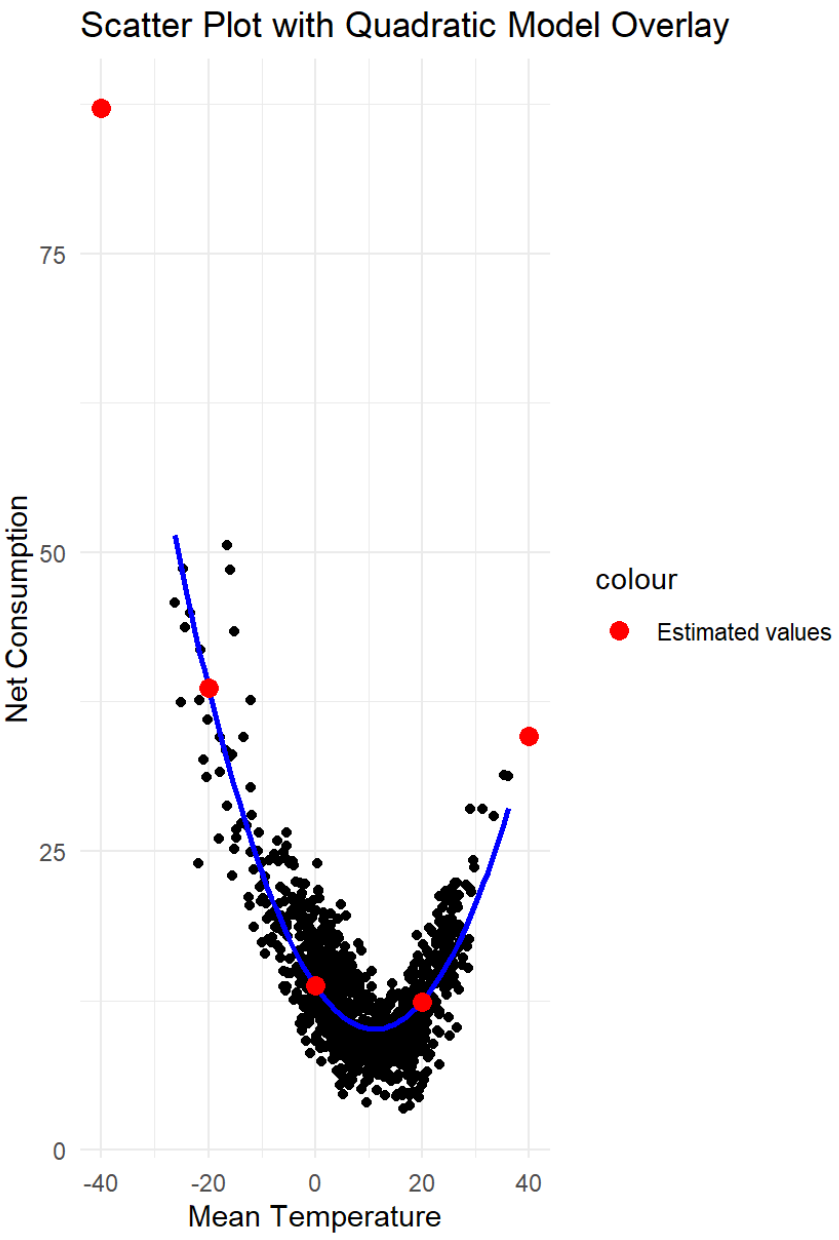

Figure 4: Scatter plot and LR curve with 5 estimated points

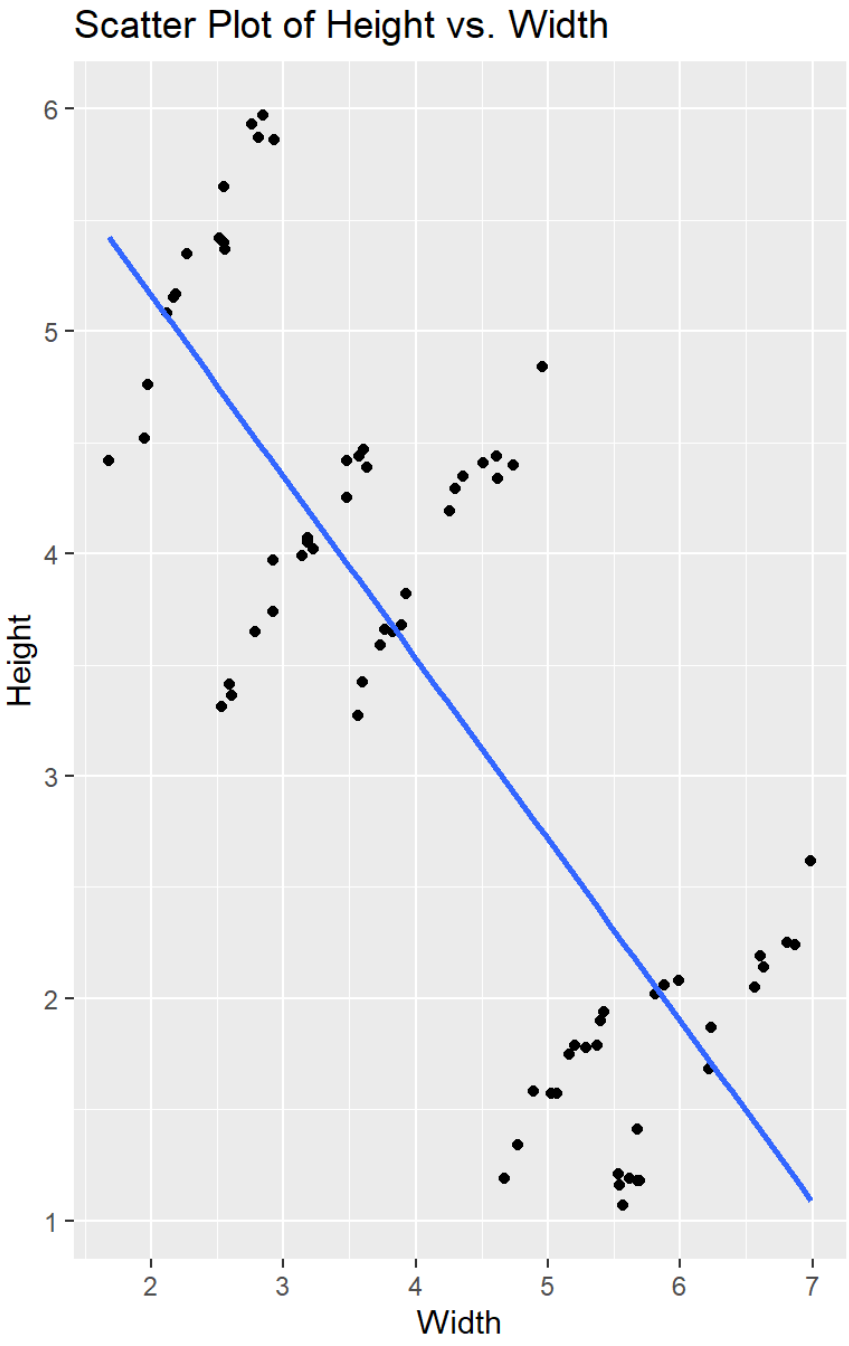

Figure 5: Scatter plot of Height vs. Width

```
Call:
lm(formula = height ~ width + species, data = plants)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38251 -0.06610  0.01552  0.06882  0.24892

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.02819    0.08933   33.90   <2e-16 ***
width        0.96119    0.03474   27.66   <2e-16 ***
speciesB    -2.06162    0.05277  -39.07   <2e-16 ***
speciesC    -3.02072    0.07737  -39.04   <2e-16 ***
speciesD    -6.36308    0.11009  -57.80   <2e-16 ***
speciesE    -7.24224    0.13845  -52.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1266 on 69 degrees of freedom
Multiple R-squared:  0.9933,    Adjusted R-squared:  0.9928
F-statistic:  2045 on 5 and 69 DF,  p-value: < 2.2e-16
```

Figure 6: Summary for new regression model

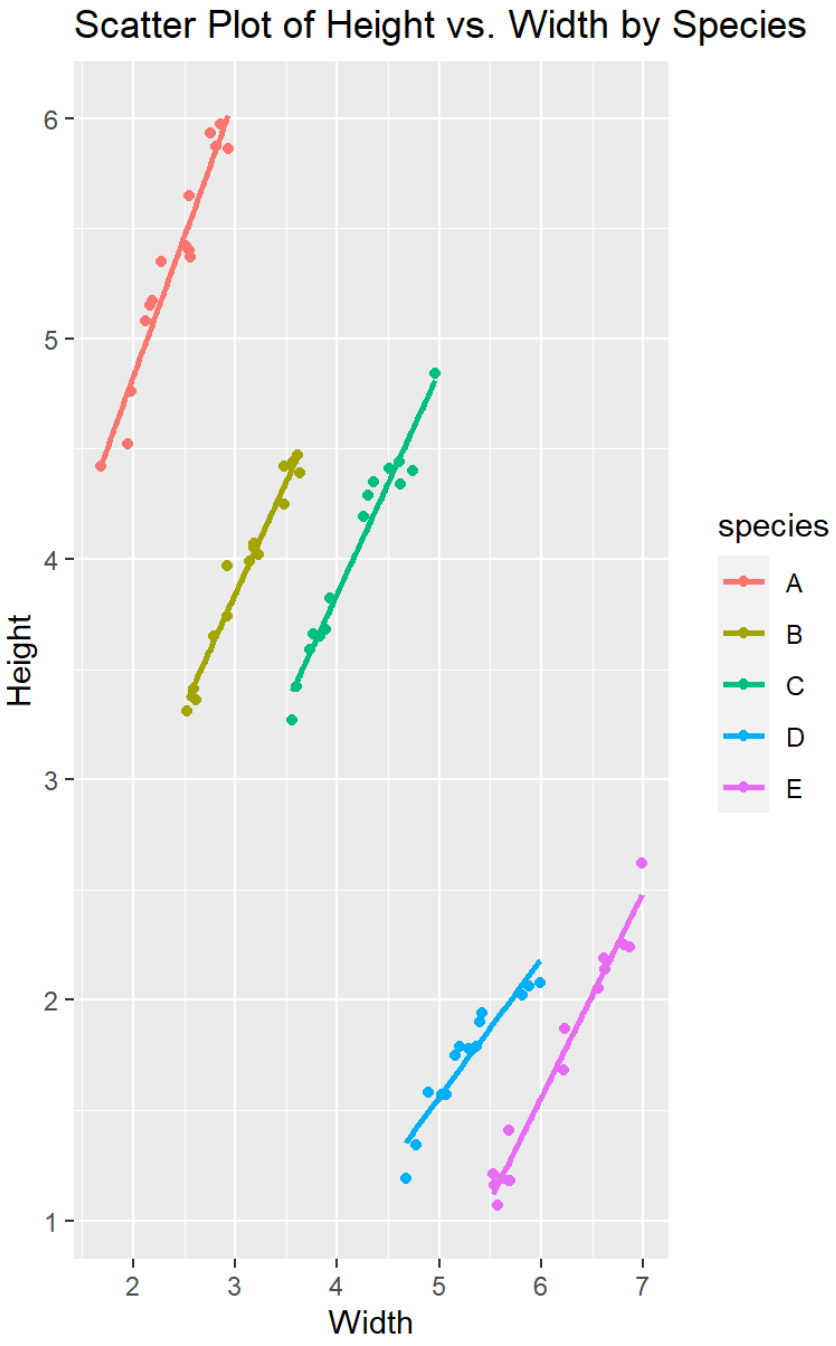

Figure 7: Scatter plot of height against width, distinguishing species

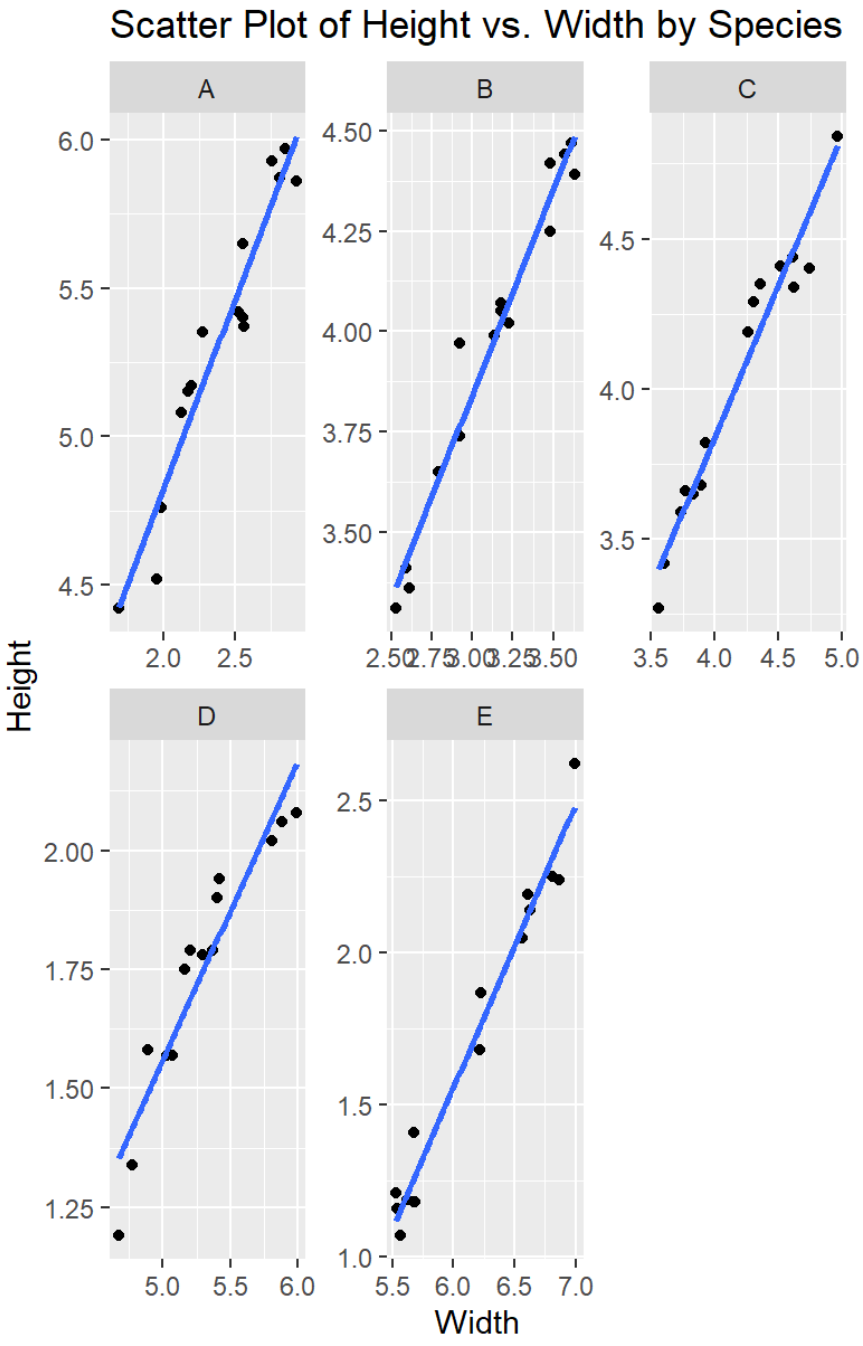

Figure 8: Scatter plot of height against width, distinguishing species and faceting by species