STAT0006ICA3_63

Group 63

Student numbers: 24005059, 24001114, 23016889

Task 1: Analysis of the heights of rose plants

Exploratory Data Analysis

Introduction to the data

The roses dataset contains 300 observations with 10 variables. An initial check reveals no missing values across any variables, but one observation contains a negative height value, which represents a clear measurement error. This observation was removed from the analysis, reducing the dataset to 299 valid records.

After removing this observation, the distribution of plant heights is roughly bell-shaped (*Figure 1.a*), with mean 157.7 cm and standard deviation 57.0 cm. *Figure 1.b* indicates minimal outliers and indicates reasonable symmetry, making height an appropriate response variable for regression modelling.

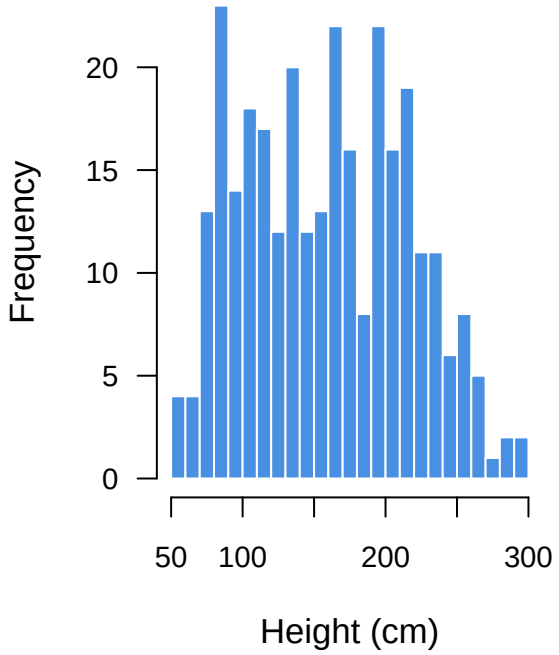

Figure 1.a: Frequency of Plant Heights

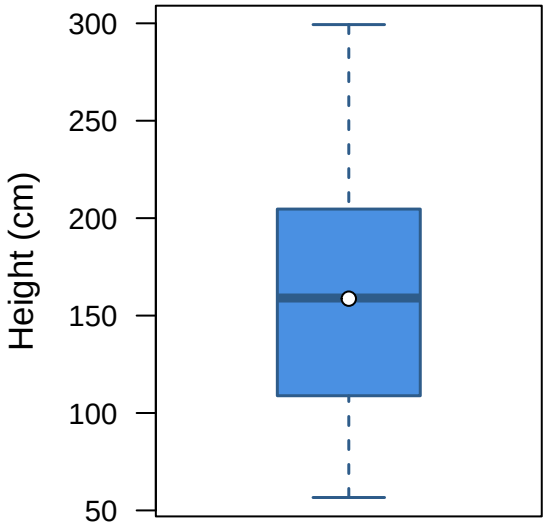

Figure 1.b: Boxplot of Plant Heights

Continuous Predictors: Environmental Factors

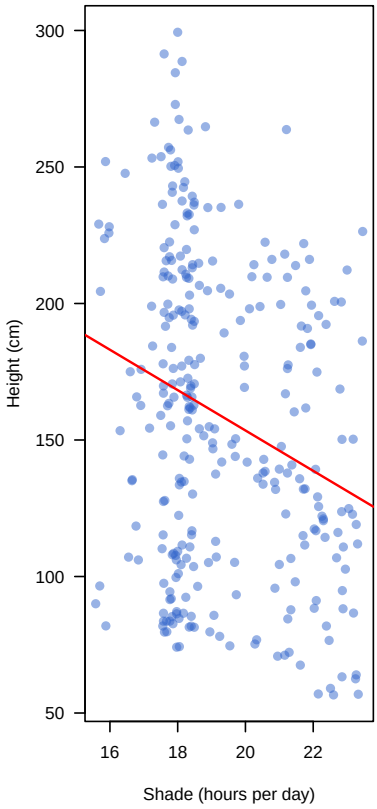

Figure 2.a: Theoritical quantiles vs. sample quantiles

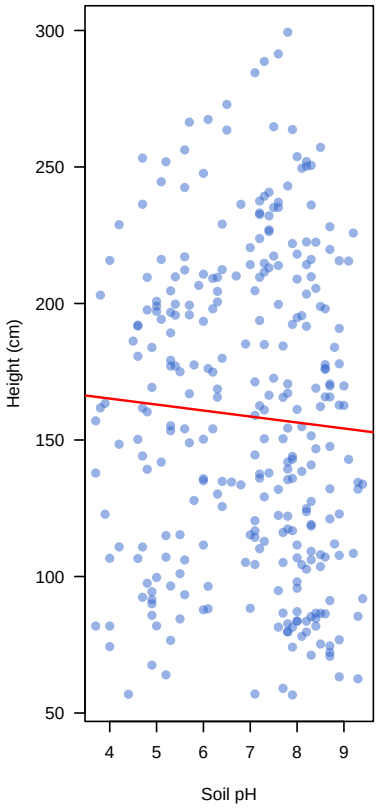

Figure 2.b: Height vs Soil pH

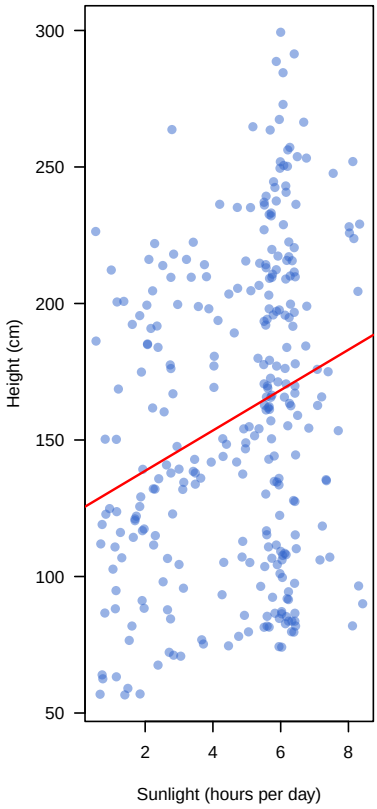

Figure 2.c: Height vs Sunlight Exposure

Shade Exposure: A weak negative association is evident between shade hours and plant height; the fitted regression line slopes downward, suggesting plants that receive more shade tend to be slightly shorter. However, the scatter is considerable, and the relationship is heavily influenced by clustering at 18 hours of shade per day. The uneven distribution of observations across shade levels may need to be looked at when it comes to modelling.

Soil pH: The relationship between soil pH and height is essentially flat, with the regression line showing negligible slope. This scatter plot reveals no meaningful pattern: soil pH may have minimal predictive power for plant height in this dataset.

Sunlight Exposure: The weak positive relationship of sunlight with height mirrors that of shade with height (given sunlight = 24 - shade hours). Since perfect collinearity exists between these two variables, one needs to be excluded from the model.

Categorical Predictors: Plant Characteristics and Growing Conditions

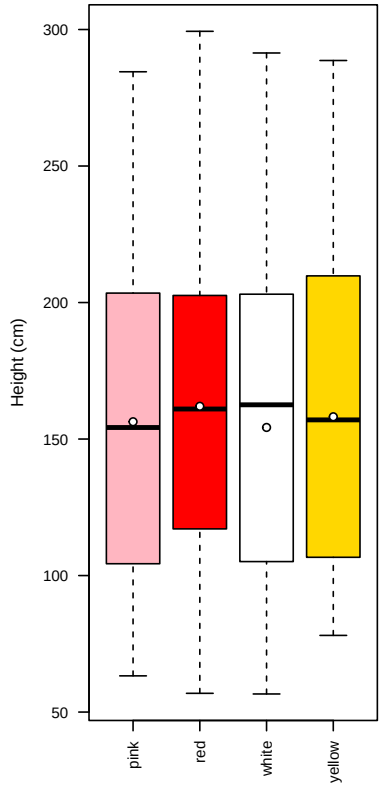

Figure 3.a: Heights by Flower Colour

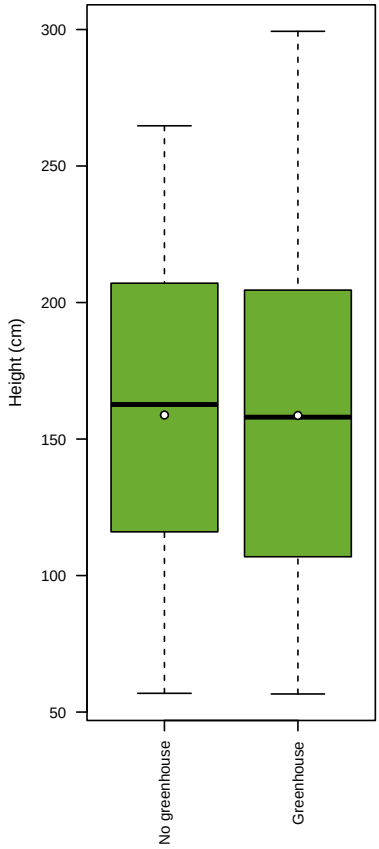

Figure 3.b: Heights by Greenhouse Status

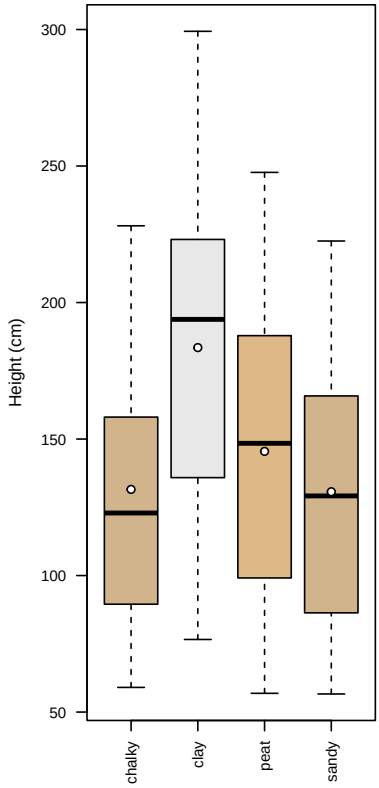

Figure 3.c: Heights by Soil Type

Flower Colour: Height distributions are mostly consistent across all flower colours. Mean and median heights - shown by the white dots and black lines respectively - and interquartile ranges are nearly identical, with no clear differences in variance. This suggests flower colour is not a meaningful predictor of plant height and may be excluded from the model.

Greenhouse Status: Plants grown with and without greenhouse protection show very similar height distributions. Both groups have almost equal medians, quartiles, and spread, indicating that greenhouse status doesn't meaningfully contribute to explaining height variation.

Soil Type: Interestingly, soil type shows clear differences in height distributions. Clay soil is strongly associated with taller plants, with a mean and median height considerably above other groups. Chalky and sandy soils exhibit similar lower height distributions.

Categorical Predictors: Rose Variety, Watering, and Temporal Factors

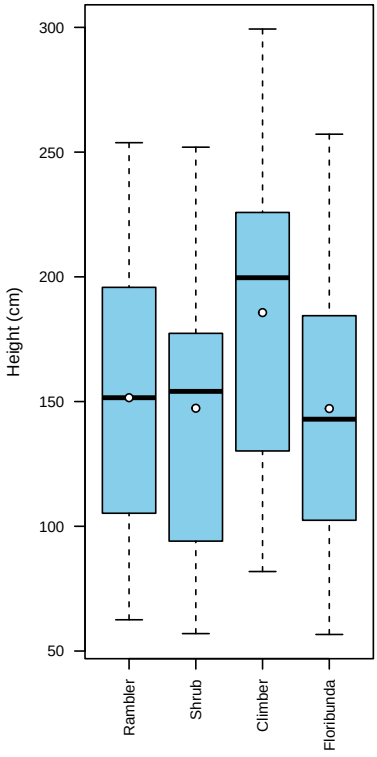

Figure 4.a: Heights by Plant Type

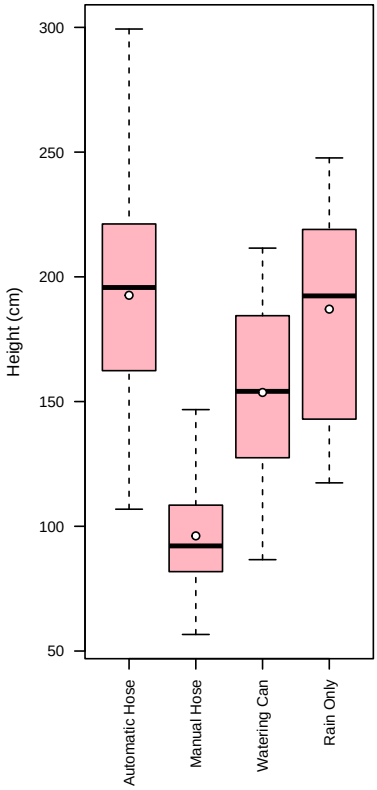

Figure 4.b: Heights by Watering Method

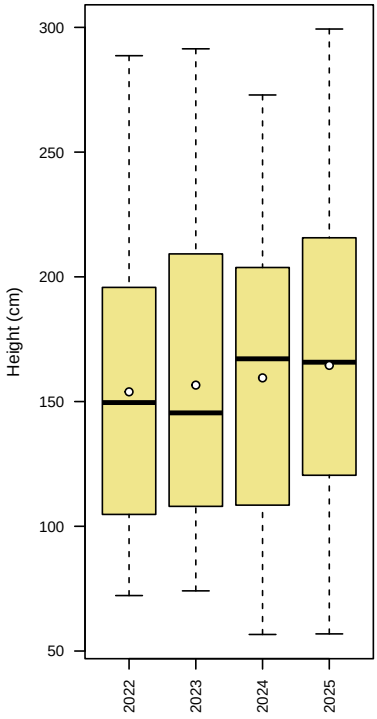

Figure 4.c: Heights by Growing Year

Plant Type (Rose Variety): A few differences emerge across rose varieties: climbers exhibit the tallest expected heights while Ramblers, Shrubs and Floribundas are noticeably shorter and have similar expected heights. The wide interquartile ranges within each category indicate considerable intra-variety variation, but the inter-variety differences are more clear. Plant type seems to be a reasonable predictor of height.

Watering Method: Substantial differences in height distributions arise across watering methods, with manual hosing showing a clearly lower mean and median. Plants receiving automatic watering or rain only tend to have higher expected heights compared to the other methods.

Growing Year: Heights remain relatively stable across the four growing years; all years show similar median heights and comparable spread. No pronounced trend is evident, suggesting that year-to-year environmental variation or systematic changes in growing conditions were minimal.

Model building

Firstly, we start by removing the outlier, since it results from a measurement error. We also change the base category for watering from automatic hose to rain only since nature is the most intuitive basic watering form.

We mentioned that sunlight time = 24 - shade time, such that these two covariates would be linearly dependent. Including both of these in the linear regression would cause perfect collinearity (i.e., the determinant of the observations matrix X will be zero: X will not be invertible, hence $X^T X$ will not be invertible, such that R will not compute the regression coefficients). We thus ignore the shade column in our analysis.

Checking linearity

Next, we check the linearity of our continuous covariates against the response variable, height. Since categorical covariates are linear by construction when plotted against the response variable, we do not need to check these.

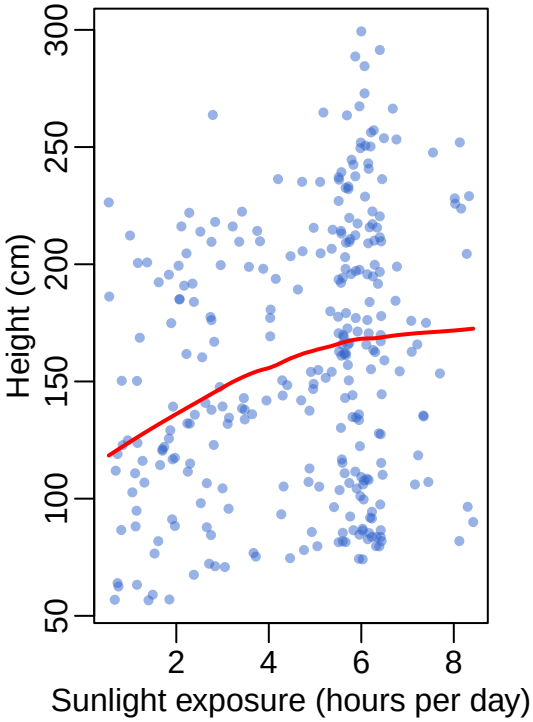

Figure 5.a: Height vs. Sunlight

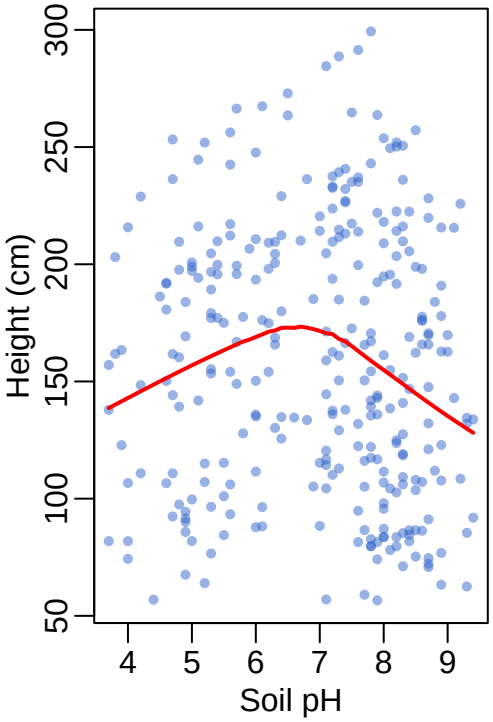

Figure 5.b: Height vs. Soil pH

The trend line on *Figure 5.a* shows a mild upward trend, with a flattening at higher values from around 5 hours of sunlight/day. However, this flattening may be misleading since there are fewer observations at higher values of sunlight exposure. Though the sunlight-height relationship seems reasonably acceptable as linear, we could consider a quadratic term.

Meanwhile, the trend line on *Figure 5.b* shows clear curvature, with the maximum turning point at around pH = 7. To avoid violations of linearity, we shall attempt to include soil pH in the linear regression quadratically.

Fitting the first model

Now, we fit a full 'naive' linear model that includes all our covariates in the dataset except shade, and with the quadratic terms on soil pH and sunlight:

```
##
## Call:
## lm(formula = height ~ sunlight + I(sunlight^2) + soil_pH + I(soil_pH^2) +
##     soil_type + watering + plant_type + greenhouse + colour +
##     year, data = roses1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.892  -7.631  -0.350   7.298  35.370
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -913.0606  1257.0178  -0.726  0.46822
## sunlight                      11.8597     1.5717   7.546 6.34e-13 ***
## I(sunlight^2)                 -0.5133     0.1845  -2.782  0.00577 **
## soil_pH                       20.0788     5.0598   3.968 9.20e-05 ***
## I(soil_pH^2)                  -0.6058     0.3831  -1.581  0.11492
## soil_typeclay                 67.1890     2.6902  24.975  < 2e-16 ***
## soil_typepeat                 52.9545     3.6607  14.466  < 2e-16 ***
## soil_typesandy                 2.5735     2.3245   1.107  0.26920
## wateringautomatic_hose        -1.9520     3.6439  -0.536  0.59260
## wateringmanual_hose          -99.1178     3.7533 -26.408  < 2e-16 ***
## wateringmanual_watering_can  -51.9812     4.1924 -12.399  < 2e-16 ***
## plant_typeshrubs              -0.1713     1.9565  -0.088  0.93028
## plant_typeclimbers            38.7910     1.8465  21.008  < 2e-16 ***
## plant_typefloribunda          -0.6238     1.8364  -0.340  0.73434
## greenhouseY                   12.4761     1.6430   7.593 4.67e-13 ***
## colourred                     -1.6680     1.6310  -1.023  0.30734
## colourwhite                   -0.7653     2.0784  -0.368  0.71298
## colouryellow                  -0.2070     2.1178  -0.098  0.92221
## year                           0.4438     0.6202   0.716  0.47485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.17 on 280 degrees of freedom
## Multiple R-squared:  0.964,  Adjusted R-squared:  0.9616
## F-statistic:   416 on 18 and 280 DF,  p-value: < 2.2e-16
```

We immediately notice that colour and years have enormous p-values - we would fail to reject, at any common significance level, that their true regression coefficient is zero. This is unsurprising given the earlier EDA revealed no relationship between these covariates and height.

Removing these covariates yields the following model:

```
##
## Call:
## lm(formula = height ~ sunlight + I(sunlight^2) + soil_pH + I(soil_pH^2) +
##     soil_type + watering + plant_type + greenhouse, data = roses1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.946  -7.671   0.039   7.554  35.476
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -15.6577    16.6367  -0.941  0.34743
## sunlight                      11.9389     1.5596   7.655 3.04e-13 ***
## I(sunlight^2)                 -0.5224     0.1828  -2.857  0.00459 **
## soil_pH                       19.9871     4.9564   4.033 7.09e-05 ***
## I(soil_pH^2)                  -0.5945     0.3748  -1.586  0.11386
## soil_typeclay                 67.1846     2.6640  25.219  < 2e-16 ***
## soil_typepeat                 53.1892     3.6370  14.624  < 2e-16 ***
## soil_typesandy                 2.5638     2.3127   1.109  0.26855
## wateringautomatic_hose        -2.4955     3.5970  -0.694  0.48838
## wateringmanual_hose          -99.6562     3.6981 -26.948  < 2e-16 ***
## wateringmanual_watering_can -52.3410     4.1575 -12.589  < 2e-16 ***
## plant_typeshrubs              -0.1511     1.9488  -0.078  0.93826
## plant_typeclimbers            38.6927     1.8313  21.129  < 2e-16 ***
## plant_typefloribunda          -0.6799     1.8269  -0.372  0.71004
## greenhouseY                   12.7554     1.6223   7.863 7.88e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.13 on 284 degrees of freedom
## Multiple R-squared:  0.9637, Adjusted R-squared:  0.9619
## F-statistic: 538.7 on 14 and 284 DF,  p-value: < 2.2e-16
```

Notably, the p-value of sunlight squared is significant (though we would fail to reject its coefficient is 0 at the 0.1% significance level), but the p-value of the quadratic term on soil pH is still quite large (0.11386: we would fail to reject the true coefficient is 0 at any standard significance level). Meanwhile, the p-values of soil pH and sunlight are significant at any common significance level.

Given this t-test result, we consider omitting the quadratic term on soil pH, but first check this will not violate the linearity assumption.

To do this, we plot the standardised residuals against soil pH for two models: the above model that includes soil_pH^2, and another (otherwise identical) model without soil_pH^2.

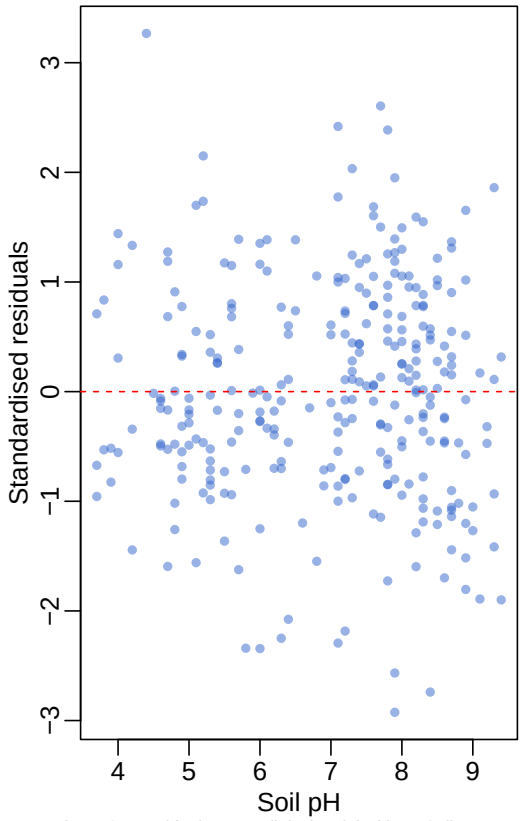

Figure 6.a: residuals vs. sunlight (model without Soil pH square)

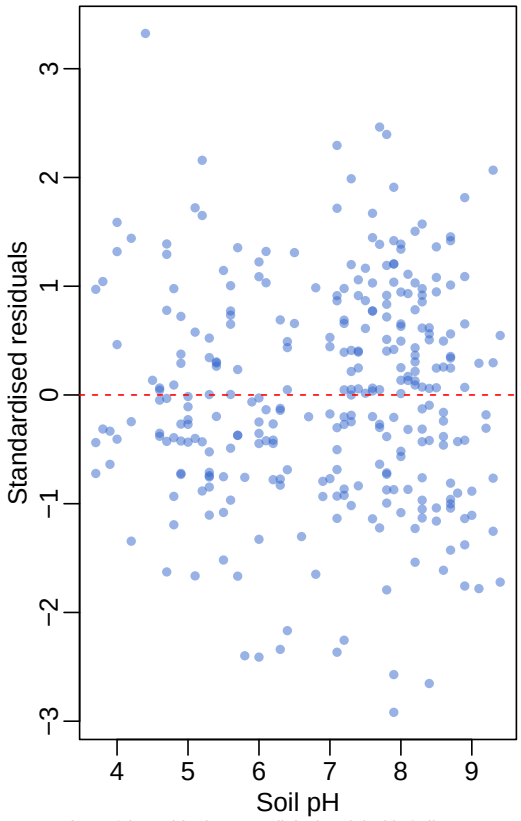

Figure 6.b: residuals vs. sunlight (model with Soil pH squared)

In both figures, residuals scatter close to randomly around 0. Since there is no obvious structure or curvature, the linearity assumption seems to be satisfied even in the model without the quadratic term, and we thus omit soil_pH^2 from now on. It may seem surprising that both figures are virtually identical, despite the clear curvature identified in *Figure 5.b*. This is likely explained by the short range of values for pH, which attenuates the impact of the quadratic term on the fit.

First check of model assumptions

We may check for possible violations of linearity, homoscedasticity and normality in our model that contains sunlight squared:

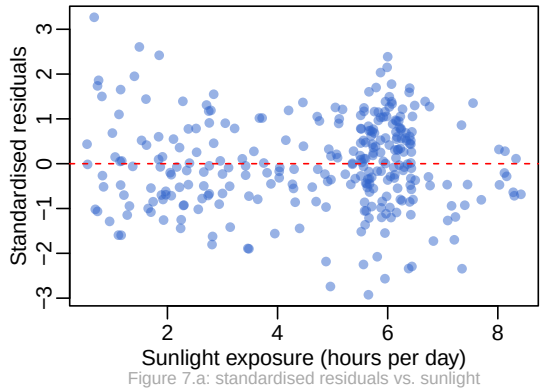

Figure 7.a: standardised residuals vs. sunlight

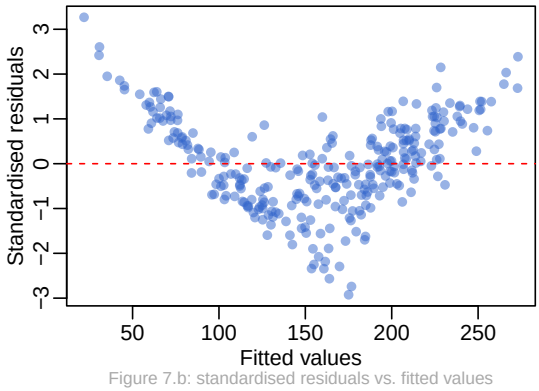

Figure 7.b: standardised residuals vs. fitted values

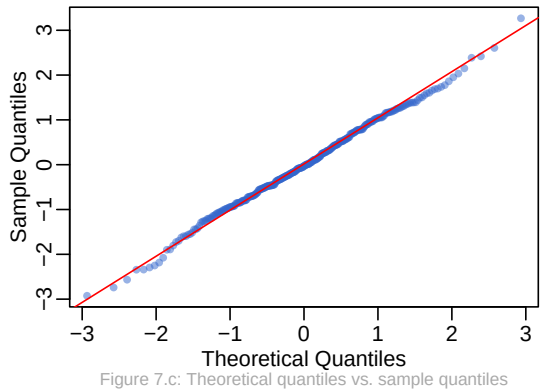

Figure 7.c: Theoretical quantiles vs. sample quantiles

In *Figure 7.a* standardised residuals scatter almost randomly around 0, with no clear structure. The scatter looks slightly less random without the quadratic term on sunlight (graph not shown), such that we choose to retain it for now.

The V-shape shape in *Figure 7.b* indicates failure of the linear specification and of the zero conditional mean assumption. The plot also provides evidence against homoscedasticity, as the variance of residuals is not constant across fitted values; we will try to transform the response to alleviate this.

Normality is approximately satisfied in *Figure 7.c* since the points lie close to the line, indicating the standardised residuals follow the theoretical normal distribution. Thus, there is here no significant evidence against the normality assumption.

We do not conduct a check for independence, as we were told we could safely assume our sample was drawn at random. Given that our data is not a time series, not spatially indexed (no knowledge of clusters) and has generally no known dependence structure, we assume independence and will not check for it further.

Choosing a transformation

To choose the 'optimal' power to satisfy the assumptions, we use a Box-Cox plot:

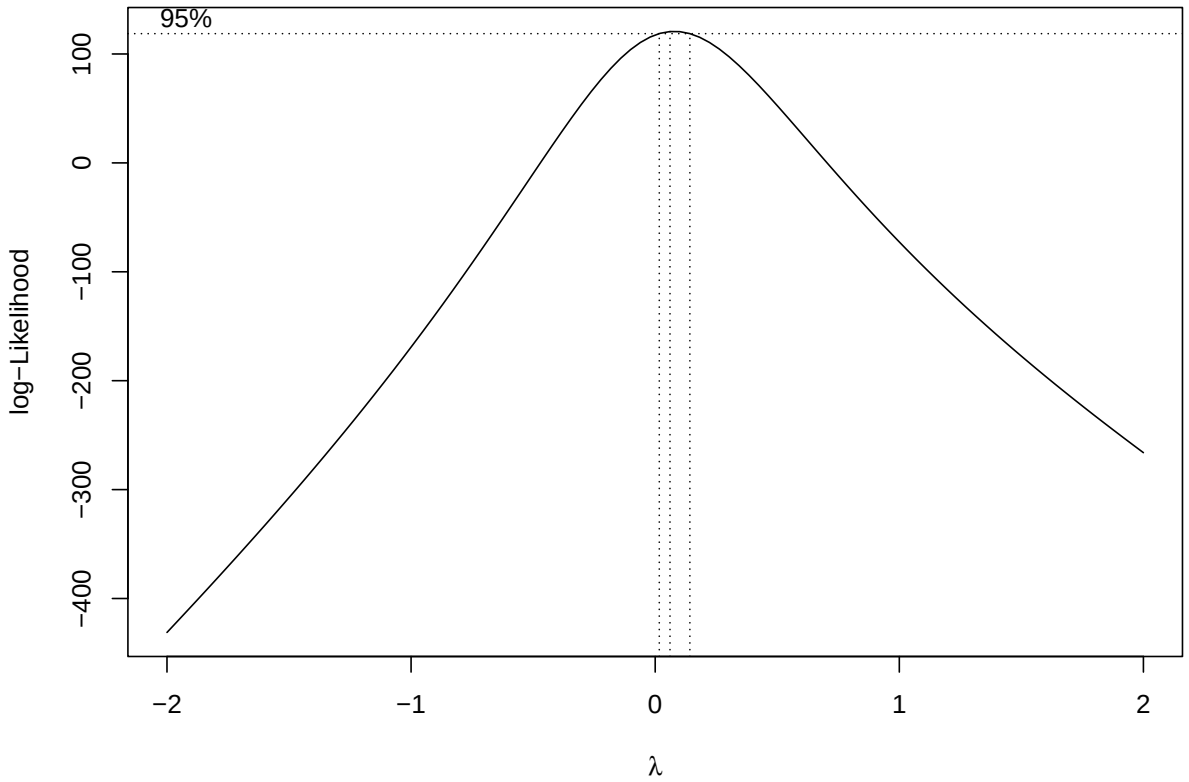

Figure 8: Log−likelihood as a function of lambda (height's exponent)

The 95% confidence interval (CI) stretches approximately between 0 and 0.20. We choose $\lambda = 0$ (the log transformation) for interpretability. A transformation seems recommended since the CI does not span $\lambda = 1$ (i.e., no transformation).

Log transformation on height

We now plot the same model with a log transformation on height, and again check the assumptions.

```
##
## Call:
## lm(formula = log(height) ~ sunlight + I(sunlight^2) + soil_pH +
##     soil_type + watering + plant_type + greenhouse, data = roses1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.106436 -0.023993  0.001069  0.024160  0.113977
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.0745214  0.0269057 151.437   <2e-16 ***
## sunlight                      0.0998537  0.0056009  17.828   <2e-16 ***
## I(sunlight^2)                -0.0057964  0.0006564  -8.831   <2e-16 ***
## soil_pH                       0.0709299  0.0024290  29.201   <2e-16 ***
## soil_typeclay                 0.4258658  0.0088423  48.162   <2e-16 ***
## soil_typepeat                 0.3147718  0.0130458  24.128   <2e-16 ***
## soil_typesandy                0.0162080  0.0080847   2.005   0.0459 *
## wateringautomatic_hose       -0.0203721  0.0128365  -1.587   0.1136
## wateringmanual_hose          -0.7184761  0.0131996 -54.432   <2e-16 ***
## wateringmanual_watering_can -0.3135961  0.0149153 -21.025   <2e-16 ***
## plant_typeshrubs             -0.0060064  0.0070011  -0.858   0.3917
## plant_typeclimbers            0.2529225  0.0065782  38.448   <2e-16 ***
## plant_typefloribunda         -0.0052344  0.0065605  -0.798   0.4256
## greenhouseY                   0.0734381  0.0058173  12.624   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03999 on 285 degrees of freedom
## Multiple R-squared:  0.9899, Adjusted R-squared:  0.9895
## F-statistic:  2159 on 13 and 285 DF,  p-value: < 2.2e-16
```

Note that all our regression coefficients are highly significant, except for two categories of the watering and soil type covariates.

Linearity is equally satisfied for soil pH (graph not shown) and better satisfied for sunlight.

We take this opportunity to fit the same model without the quadratic term on sunlight, to see what impact this has on the linearity assumption given the log transformation.

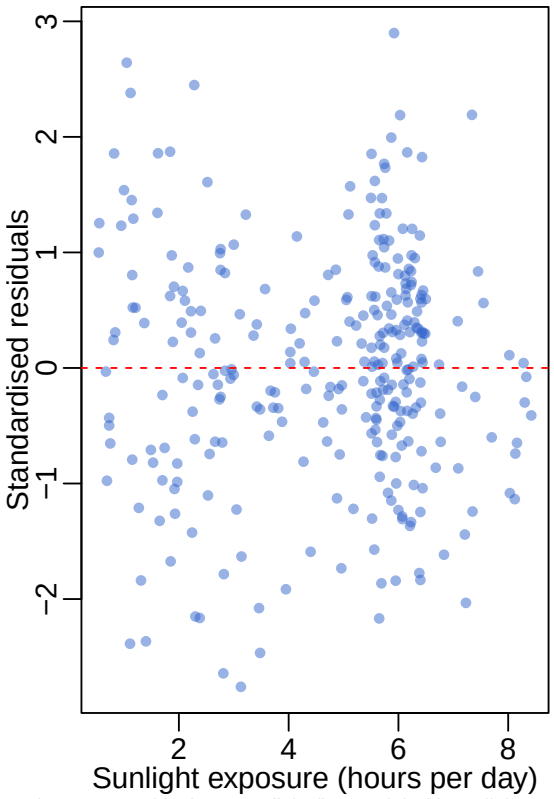

Figure 9.a: residuals vs. sunlight (log model with sunlight squared)

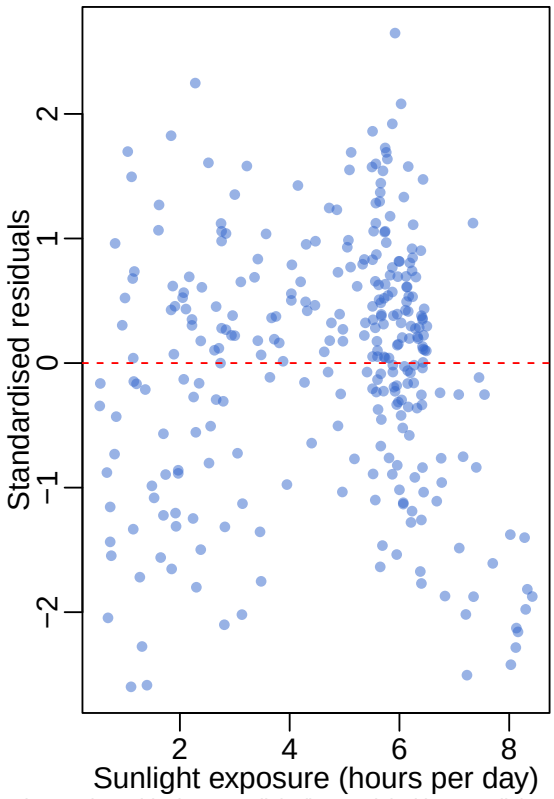

Figure 9.b: residuals vs. sunlight (log model without sunlight squared)

The scatter on *Figure 9.b* (log model without the quadratic term on sunlight) shows an inverted U shape that would provide evidence against the linearity assumption. On the other hand, the scatter on *Figure 9.a* (same log model, with the quadratic term on sunlight) is significantly more random-looking. It seems the log transformation on height has made the quadratic term on sunlight more crucial to the linearity assumption (since the log transformation compresses larger height values, it can reveal previously masked curvature in the sunlight-height relationship).

We notice that homoscedasticity has significantly improved and that normality is equally well satisfied (analysed further below).

Trying interactions

We now try various interaction terms to see if those can improve model fit. We try watering x soil_type because both their main effects are strong and it makes scientific sense (clay retains water better than sandy...)

```
##
## Call:
## lm(formula = log(height) ~ sunlight + I(sunlight^2) + soil_pH +
##     soil_type + soil_type * watering + watering + plant_type +
##     greenhouse, data = roses1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09356 -0.02433  0.00000  0.02425  0.10976
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                               4.0794488  0.0321828 126.759
## sunlight                                  0.0987726  0.0055913  17.666
## I(sunlight^2)                            -0.0056629  0.0006567  -8.624
## soil_pH                                   0.0706156  0.0024222  29.154
## soil_typeclay                             0.4248598  0.0275771  15.406
## soil_typepeat                             0.3328744  0.0466271   7.139
## soil_typesandy                           -0.0073694  0.0455433  -0.162
## wateringautomatic_hose                   -0.0405116  0.0249559  -1.623
## wateringmanual_hose                      -0.6926184  0.0256450 -27.008
## wateringmanual_watering_can              -0.2961647  0.0364249  -8.131
## plant_typeshrubs                         -0.0066839  0.0069803  -0.958
## plant_typeclimbers                        0.2531486  0.0065351  38.737
## plant_typefloribunda                     -0.0051410  0.0065509  -0.785
## greenhouseY                               0.0711563  0.0057458  12.384
## soil_typeclay:wateringautomatic_hose      0.0238155  0.0294536   0.809
## soil_typepeat:wateringautomatic_hose      0.0016966  0.0481877   0.035
## soil_typesandy:wateringautomatic_hose     0.0501169  0.0468340   1.070
## soil_typeclay:wateringmanual_hose        -0.0297126  0.0300784  -0.988
## soil_typepeat:wateringmanual_hose        -0.0520020  0.0488674  -1.064
## soil_typesandy:wateringmanual_hose       -0.0033446  0.0473633  -0.071
## soil_typeclay:wateringmanual_watering_can -0.0141605  0.0401049  -0.353
## soil_typepeat:wateringmanual_watering_can  0.0366585  0.0672313   0.545
## soil_typesandy:wateringmanual_watering_can -0.0187781  0.0561499  -0.334
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## sunlight                                   < 2e-16 ***
## I(sunlight^2)                             5.16e-16 ***
## soil_pH                                    < 2e-16 ***
## soil_typeclay                              < 2e-16 ***
## soil_typepeat                             8.33e-12 ***
## soil_typesandy                               0.872
## wateringautomatic_hose                       0.106
## wateringmanual_hose                        < 2e-16 ***
## wateringmanual_watering_can               1.45e-14 ***
## plant_typeshrubs                             0.339
## plant_typeclimbers                         < 2e-16 ***
```

```
## plant_typefloribunda                            0.433
## greenhouseY                                   < 2e-16 ***
## soil_typeclay:wateringautomatic_hose            0.419
## soil_typepeat:wateringautomatic_hose            0.972
## soil_typesandy:wateringautomatic_hose           0.286
## soil_typeclay:wateringmanual_hose               0.324
## soil_typepeat:wateringmanual_hose               0.288
## soil_typesandy:wateringmanual_hose              0.944
## soil_typeclay:wateringmanual_watering_can       0.724
## soil_typepeat:wateringmanual_watering_can       0.586
## soil_typesandy:wateringmanual_watering_can      0.738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03922 on 276 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9899
## F-statistic:  1327 on 22 and 276 DF,  p-value: < 2.2e-16
```

Virtually none of the interaction terms are significant, with all p-values but one being greater than 0.1.

Before removing the interaction, we test whether our interaction terms are collectively useful to the model fit. We do not solely rely on t tests as we would risk missing how the interaction terms collectively explain variation in the outcome and instead use a partial F test where the nested model has no interaction.

```
## Analysis of Variance Table
##
## Model 1: log(height) ~ sunlight + I(sunlight^2) + soil_pH + soil_type +
##     watering + plant_type + greenhouse
## Model 2: log(height) ~ sunlight + I(sunlight^2) + soil_pH + soil_type +
##     soil_type * watering + watering + plant_type + greenhouse
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1    285 0.45569
## 2    276 0.42448  9  0.031207 2.2545 0.01902 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Given the p-value 0.5202, we would favour the model without an interaction. We thus do not retain the watering x soil_type interaction.

We come to a similar conclusion after trying the greenhouse x sunlight interaction.

We conclude that the best model we found uses log height as the response and sunlight, sunlight squared, soil_pH, soil_type, watering, plant_type and greenhouse as covariates.

Model checking for the final chosen model

Determination

The final model fits the data well. The R-squared of 0.9899 shows 99% (to 2 significant figures) of the variability in the height is explained by the model, and indicates that the full model is considerably better in explaining the outcome than the intercept-only model - which is also shown by the minuscule p-value of 2.2e-16 on the F statistic. However, the high R-squared does not necessarily indicate a strong linear model, so we shouldn't purely base our model assessment off this metric. As a note of warning to any prospective garden centres, the model only measures correlation between variables and does not imply any causation so although some variables occur with taller plants, we do not show that these variable cause differences in height.

Linearity in (continuous) Covariates

We find no evidence against the linearity assumption by plotting standardised residuals against each of the continuous covariates. This was already referenced when discussing *Figures 9.b* (regarding sunlight), and we come to the same conclusion for soil_pH.

Homoscedasticity and Normality

Furthermore, since residuals are not independent of the observed responses, we plot the standardised residuals against the fitted values instead of observed values.

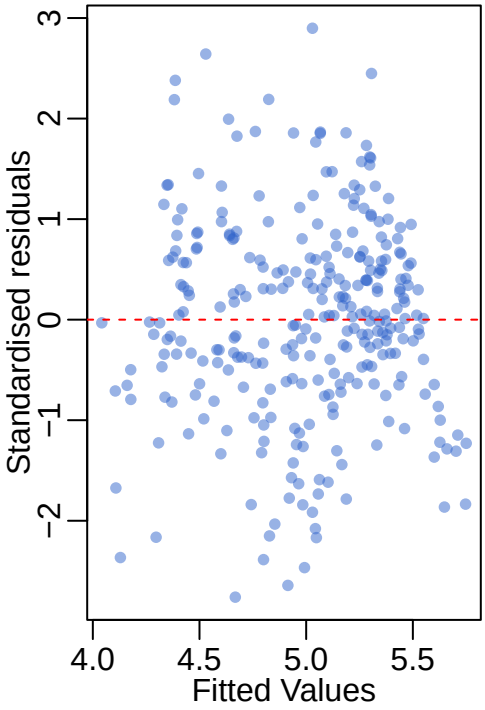

Figure 10.a: standardised residuals vs. fitted values

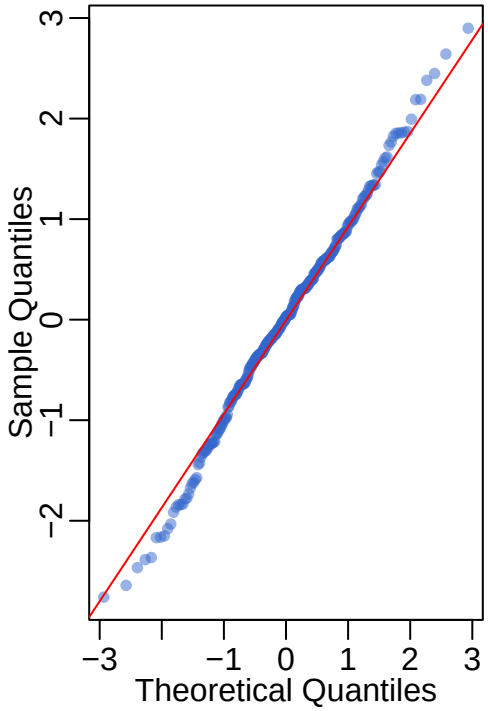

Figure 10.b: theoritical quantiles vs. sample quantiles

Looking at *Figure 10.a*, there is generally constant variance of the standardised residuals, but with a noticeable compression at higher fitted values. To mitigate the potential bias in the variance estimator, heteroscedasticity-robust standard errors can be computed and used. The plot further suggests no significant evidence against the linearity assumption or support for any visual autocorrelation. On *Figure 10.b*, the vast majority if not all of the points lie on or very close to the normal line (the sample quantiles closely track the theoretical quantiles), providing no evidence against the normality assumption. Hence, further inference based upon confidence intervals and hypothesis tests seems valid only when using heteroscedasticity-robust standard errors.

Conclusion

To conclude, a rambler that was grown outside of a greenhouse, has soil pH of zero, is grown in chalk, received zero hours of average direct sunlight per day, and was watered with rain only has an expected log height of approximately 4.07, or an expected height of approximately 55.6 centimetres.

Soil type and watering method correlate with plant heights the most. Most notably, using clay or peat over chalk increases the expected height of a plant by approximately $(e^{0.426} - 1) \times 100 \approx 53.1\%$ and 37.0 % respectively. Using a manual hose or a manual watering can instead of only rain decreases the expected height of a plant by approximately 105.0 % and 36.9 % respectively. This seems counterintuitive, since we would expect manual watering (that is, giving additional water intentionally to each plant as opposed to a general scattering of rain) to be more effective at providing sufficient water for a plant and hence maximise its height. In addition, another notable effect is plant type: the expected height for climbers is approximately 28.8 % taller than that of ramblers.

Discussion of limitations

The main issue with the data is mild heteroscedasticity, making the conventional OLS standard error estimator inconsistent, hence leading to invalid inference. We could instead use heteroscedasticity-robust standard errors. These are valid asymptotically given the consistency of the variance estimator is supported by the Law of Large Numbers and inference is justified by the asymptotic normality of the OLS estimator under the Central Limit Theorem, which seems reasonable given our 299 valid observations.

Another issue with the model is that it does contain quite a few covariates, increasing the risk of overfitting. To mitigate this, we have adopted a parsimonious specification that only includes the covariates with the highest significance levels.

Finally, the log transformation of the response makes the interpretation of the regression coefficients a little complicated: a one-unit increase in a covariate is associated with a proportional change in plant height of $(e^{\hat{\beta}} - 1) \times 100\%$, while it may have been more intuitive to measure effects in absolute centimetres.

Task 2: Analysis of the number of flowers produced

Model Selection Rationale

The dependent variable (number of flowers produced) is a count variable with values ranging from 0 to 11, with a mean of 4.25 and a median of 4.00. Two models were developed and compared to identify the best approach.

```
##     flowers         colour               time
##  Min.   : 0.00   Length:100         Min.   : 0.00
##  1st Qu.: 2.00   Class :character   1st Qu.: 5.75
##  Median : 4.00   Mode  :character   Median :12.25
##  Mean   : 4.25                      Mean   :11.56
##  3rd Qu.: 6.00                      3rd Qu.:16.38
##  Max.   :11.00                      Max.   :23.75
```

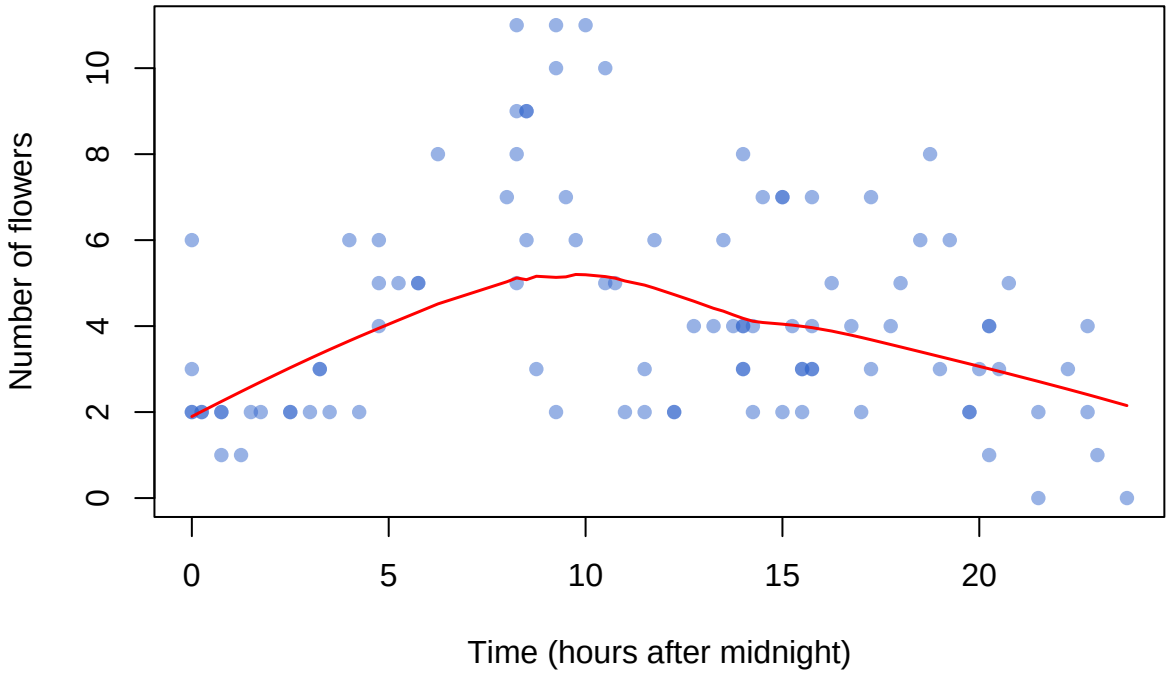

Figure 11: Flowers vs. watering time

Initial Model: Genralised Linear Model (GLM)

A GLM with Poisson as the family and a log link was initially fitted using the formula: flowers ~ time + colour. This represents the standard baseline for (positive) count data, assuming that the mean and variance of flower counts are equal, with a log-linear relationship between predictors and the expected count.

```
##
## Call:
## glm(formula = flowers ~ time + colour, family = poisson(link = "log"),
##     data = watering)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.587544   0.120404  13.185   <2e-16 ***
## time         -0.004622   0.007366  -0.628    0.530
## colourred    -0.092485   0.113487  -0.815    0.415
## colourwhite  -0.236387   0.181065  -1.306    0.192
## colouryellow -0.142909   0.156379  -0.914    0.361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 150.19  on 99  degrees of freedom
## Residual deviance: 147.84  on 95  degrees of freedom
## AIC: 470.74
##
## Number of Fisher Scoring iterations: 5
```

The Poisson model produced an AIC of 470.74, with residual deviance of 147.84 on 95 degrees of freedom. All predictor coefficients were non-significant: time ($p = 0.530$), red ($p = 0.415$), white ($p = 0.192$), and yellow ($p = 0.361$). The null deviance was 150.19, indicating only minimal deviance reduction (147.84 vs. 150.19), suggesting poor model fit.

Refined Model: Generalised Additive Model (GAM)

As seen in (*Figure 11*), there is a non-linear relationship between watering time and flower production. A GAM was therefore fitted using: flowers ~ s(time) + colour with Poisson as the family. The smooth term s(time) with 5.863 effective degrees of freedom allows the time effect to vary flexibly rather than assuming linearity.

```
## Loading required package: nlme
```

```
## This is mgcv 1.9-3. For overview type 'help("mgcv-package")'.
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## flowers ~ s(time) + colour
##
```

```
## Parametric coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.41130    0.09235  15.282   <2e-16 ***
## colourred    -0.02566    0.11496  -0.223    0.823
## colourwhite  -0.22065    0.18399  -1.199    0.230
## colouryellow -0.02282    0.16012  -0.143    0.887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(time) 5.863  6.999  59.21  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =    0.4   Deviance explained = 45.8%
## UBRE = 0.010625  Scale est. = 1         n = 100
```

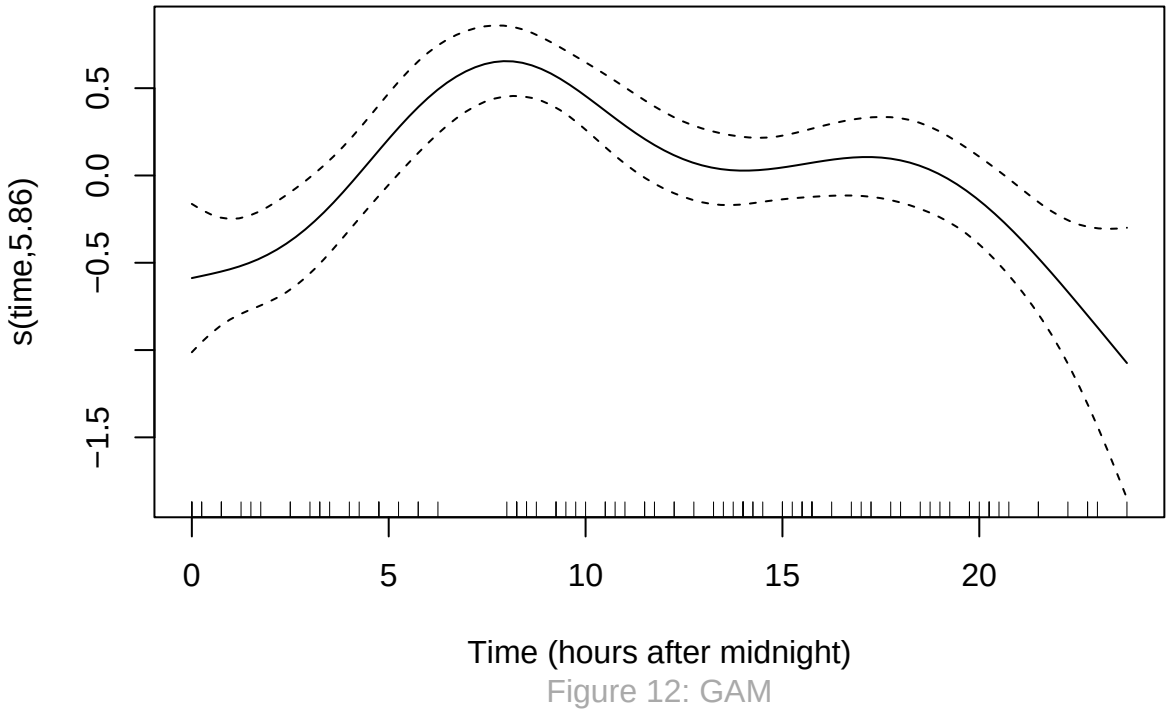

Figure 12: GAM

The GAM substantially improved model performance: AIC decreased to 413.97, and residual deviance dropped to 81.33, compared to 150.19 null deviance, now explaining 45.8% of deviance. The smooth term for time was also highly significant ($\chi^2 = 59.21$, $p < 0.001$).

Figure 13 also shows that the GAM provides a visually superior fit compared to the GLM, whose curve aligns more closely with the observed data.

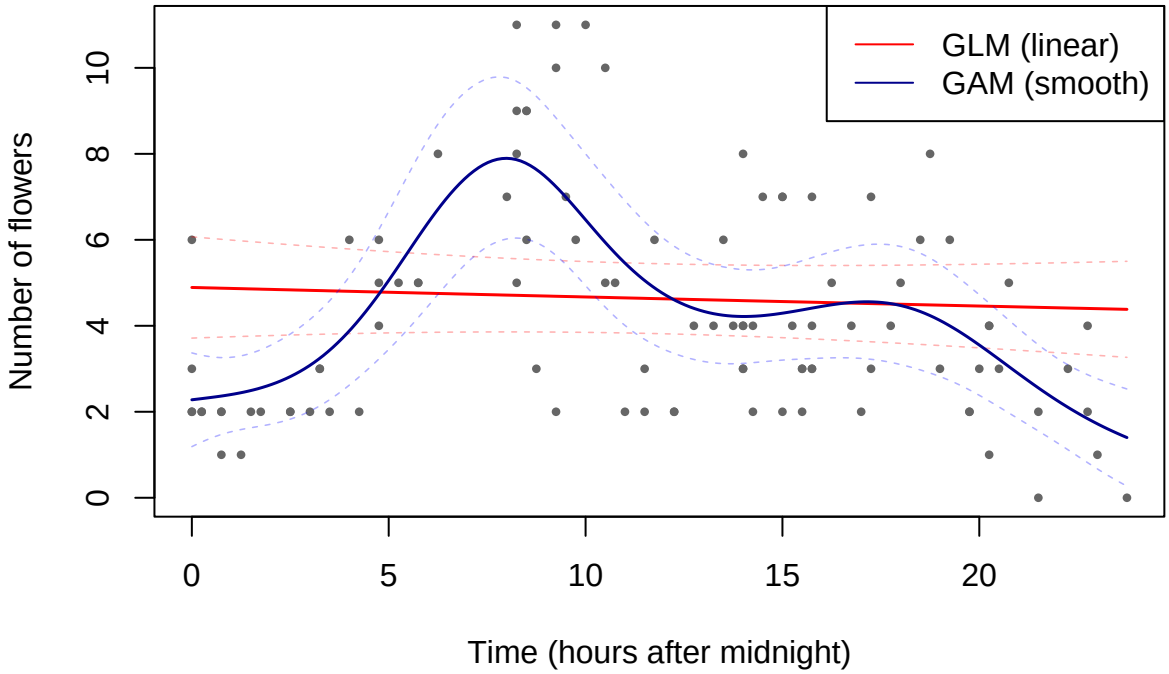

Figure 13: GLM vs. GAM model comparison: flowers vs. time

Model Selection and Assumption Assessment

The GAM was selected as the final model based on three key criteria. Firstly, it demonstrates substantially lower AIC, indicating superior balance between model fit and complexity. Secondly, it achieves a significant reduction in residual deviance, explaining 45.8% of the deviance compared to approximately 1.9% for the GLM. Thirdly, the smooth term for time is highly significant, providing statistical confirmation of the non-linear relationship between watering time and flower production.

Poisson models assume equidispersion (variance = mean). The dispersion parameter for the GAM was estimated as 1.0, consistent with Poisson assumptions. However, *Figure 14.b* reveals some potential over-dispersion at higher fitted values, though not dramatic.

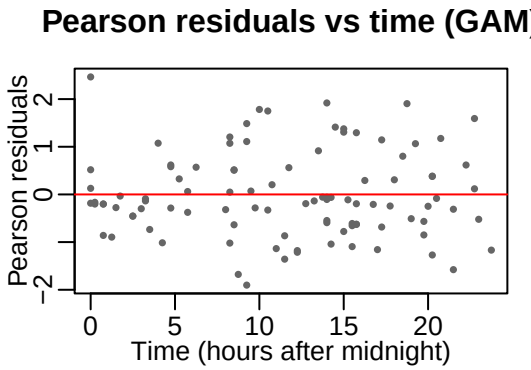

Figure 14.a: standardised residuals vs. fitted values

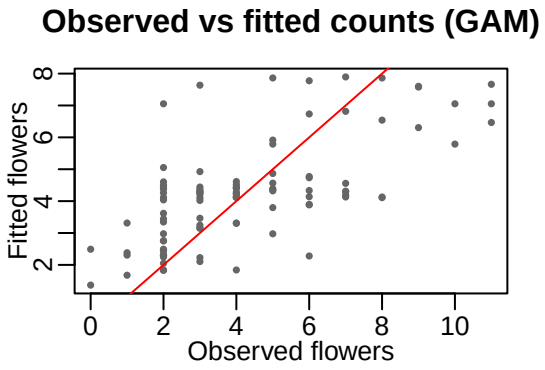

Figure 14.b: standardised residuals vs. fitted values

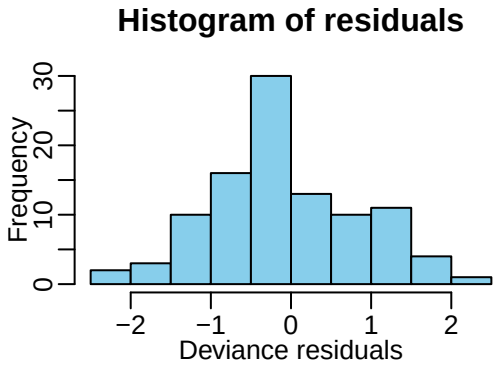

Figure 14.c: standardised residuals vs. fitted values

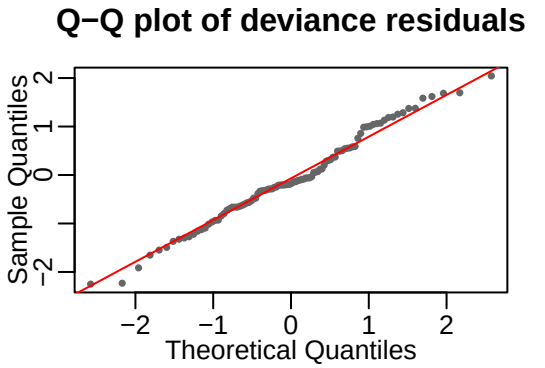

Figure 14.d: standardised residuals vs. fitted values

Count regression models rest on three main assumptions which have been addressed below.

The independence assumption is supported by *Figure 14.a* which displays no systematic patterns or clusters across time values; residuals fluctuate randomly around zero across the watering time range, with no autocorrelation evident.

The linearity assumption was critically violated by the initial GLM. *Figure 13* shows a clear non-linear pattern; the GAM's smooth function captures this pattern well. The histogram of residuals shows an approximately symmetric, bell-shaped distribution, and the Q-Q plot demonstrates reasonably close adherence to the theoretical normal line, with only minor deviations at the extremes consistent with count data.

The GAM's 45.8% deviance explained represents fair performance for horticultural count data, where substantial unexplained variation is typical due to a range of unmeasured factors.

Comments and Recommendations for Garden Centre Staff

Peak flower production (about 7–8 flowers) occurs when plants are watered between 8 AM and 2 PM, yielding roughly 1.5–2 extra flowers per plant compared with dawn or evening watering. However, the model explains only 45.8% of the variation in flower counts, and factors such as genetics, microclimate, soil, pests, and interactions with temperature still cause about ±2 flowers of unexplained variability, and predictions are most reliable for counts between 3 and 7. To strengthen the guidance, multi-year trials and recording temperature and humidity during watering are recommended, and flower colour can be ignored in scheduling since its estimated effects were small and not statistically significant (all $p > 0.2$).

Total word count: 2984.

Statement about use of generative AI tools

No AI tools were used to complete this coursework.