Statistical Modeling with Linear and Mixed-Effects Models

Fotios Siannis, Stylianos Tzortzakis

Department of Mathematics
National and Kapodistrian University of Athens, Greece

CODEJAM / 2nd Workshop on Computational Biology
31 March 2026, Paris, France

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Presentation Outline

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

General Information

- All information related to this course can be found at **https://eclass.uoa.gr/courses/MATH861/**
- Data used for the examples are part of several R packages, from Kaggle and from our own data simulations
- Introductory notes on R can be found on the above site
- For further questions email at **fsiannis@math.uoa.gr** or contact through the above site

Introduction to Presentation

In this short course we will present statistical methodology for the analysis of data of the following forms

- Cross-sectional data with
  - Continuous response variable (Linear Regression)
  - Binary response variable (Logistic Regression)
- Longitudinal data
  - Data with repeated measurements with a continuous response variable (Random Effects Models)

For these purposes, we will base our presentation in the analysis of three different datasets.

Famous Quotes

George Box

"All models are wrong, but some are useful".

Famous Quotes

George Box

"All models are wrong, but some are useful".

W. Edwards Deming

"In God we trust. All others must bring data".

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Famous Quotes

George Box

"All models are wrong, but some are useful".

W. Edwards Deming

"In God we trust. All others must bring data".

Ronald Coase

"If you torture the data long enough, it will confess to anything".

Famous Quotes

George Box

"All models are wrong, but some are useful".

W. Edwards Deming

"In God we trust. All others must bring data".

Ronald Coase

"If you torture the data long enough, it will confess to anything".

Sherlock Holmes (Arthur Conan Doyle)

"It is a capital mistake to theorize before one has data".

Presentation Outline

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Simple Linear Model

- It is the simplest version of the linear models, with one dependent variable (Y) and only one independent (prognostic) variable (X)
- The model takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where:

- β_0: the *intercept*
- β_1: the *slope*
- $i = 1, 2, \ldots, n$ where n the size of the sample

Main Assumptions

We have pairs of observations (x_i, y_i) for $i = 1, 2, ..., n$, where

- X: fixed variable (not random), called covariate or explanatory variable
- Y: random variable, called response, from population with mean

$$E(Y) = \beta_0 + \beta_1 X,$$

called the response, and

- ϵ: random errors, for which we assume
  - (a) $E(\epsilon_i) = 0$
  - (b) $Var(\epsilon_i) = \sigma^2$, ie. common variance for all i
  - (c) $Cov(\epsilon_i, \epsilon_j) = 0$, ie. independence for i, j
  - (d) Usual assumption:

$$\epsilon_i \sim N(0, \sigma^2)$$

Ordinary Least Squares Estimation (OLS)

- In simple linear model we have to estimate β_0, β_1 (and σ^2)
- We take $\hat{\beta}_0$ and $\hat{\beta}_1$ to be the parameters estimates. Then:

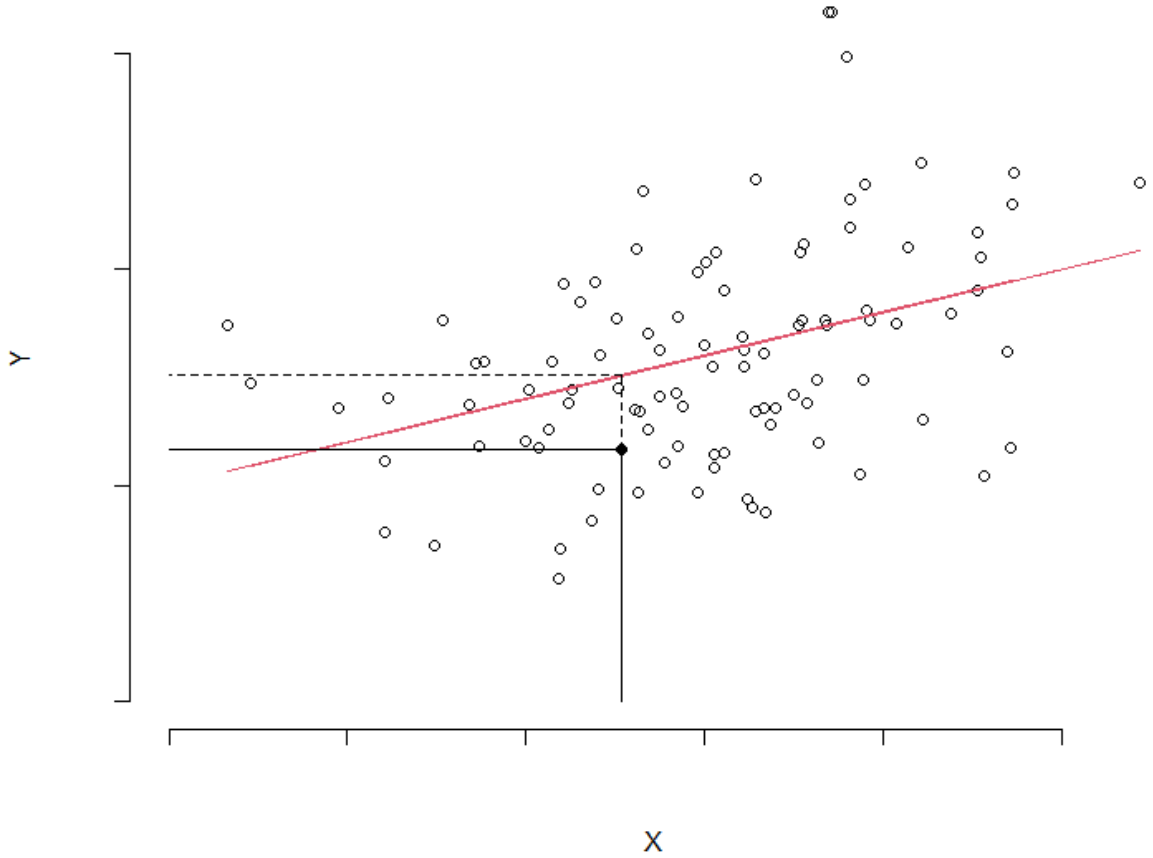
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  is an estimate of the mean for Y for every value of x_i
- The OLS priciple is to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the square distance between y_i and $\hat{y}_i$ to become minimum
- Therefore, we want to minimize the following quantity

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SSE$$

(Residual Sum of Square or Sum of Square Error)

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

The Multiple Linear Regression Model

The General Population Model

The relationship between response (Y) and the k predictors is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ represents the random (biological) error.

- β_0 **(Intercept):** The theoretical response content if all markers (predictors) were zero (often non-physical in biology).
- β_i **(Partial Coefficients):** The change in response for a 1-unit increase in X_i, *holding all other predictors constant*.
- ϵ **(Residuals):** The "unexplained" biological variance (soil micro-differences, measurement noise).

Dataset: Italian Wine Analysis (rattle)

- **Source:** UCI Machine Learning Repository, available via `library(rattle)`.
- **Scope:** Chemical analysis of 178 wine samples from 3 Italian cultivars.
- **Structure:** 13 continuous chemical variables (+ 1 categorical Type).

Target Variable (Y)

Alcohol: The ethanol percentage by volume.

Key Physicochemical Predictors (X)

- **Flavanoids/Phenols:** Antioxidant compounds affecting color and taste.
- **Magnesium:** Mineral content from the soil/vineyard.
- **Proline:** An amino acid used as a marker for wine nitrogen content.
- **Color Intensity:** Spectral analysis of the wine's appearance.

```
> head(wine_reg)
  Alcohol Malic  Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids
1   14.23  1.71 2.43       15.6       127    2.80       3.06          0.28
2   13.20  1.78 2.14       11.2       100    2.65       2.76          0.26
3   13.16  2.36 2.67       18.6       101    2.80       3.24          0.30
4   14.37  1.95 2.50       16.8       113    3.85       3.49          0.24
5   13.24  2.59 2.87       21.0       118    2.80       2.69          0.39
6   14.20  1.76 2.45       15.2       112    3.27       3.39          0.34
  Proanthocyanins Color  Hue Dilution Proline
1            2.29  5.64 1.04     3.92    1065
2            1.28  4.38 1.05     3.40    1050
3            2.81  5.68 1.03     3.17    1185
4            2.18  7.80 0.86     3.45    1480
5            1.82  4.32 1.04     2.93     735
6            1.97  6.75 1.05     2.85    1450
```

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

The Fitted Multiple Linear Regression Model

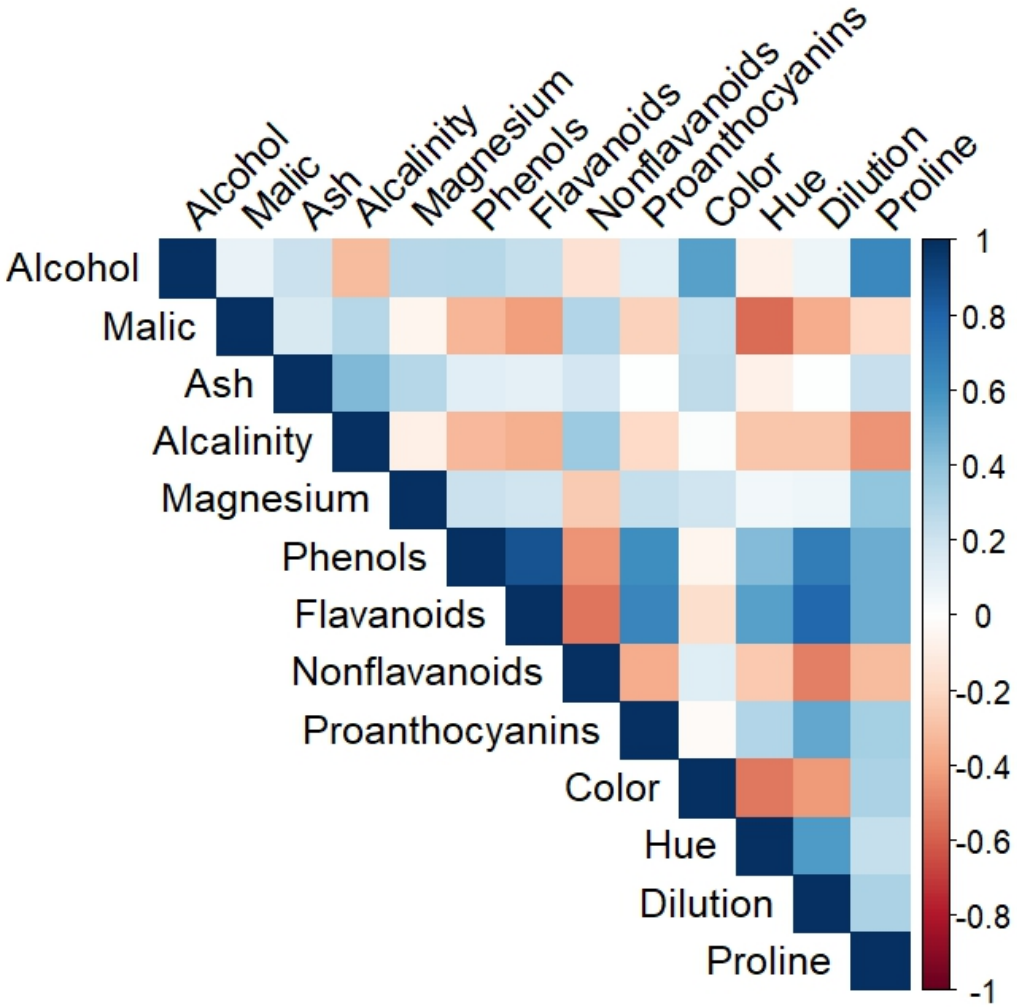
The Estimated Prediction Equation

For the Italian Wine dataset, the alcohol content ($\hat{Y}$) is estimated by:

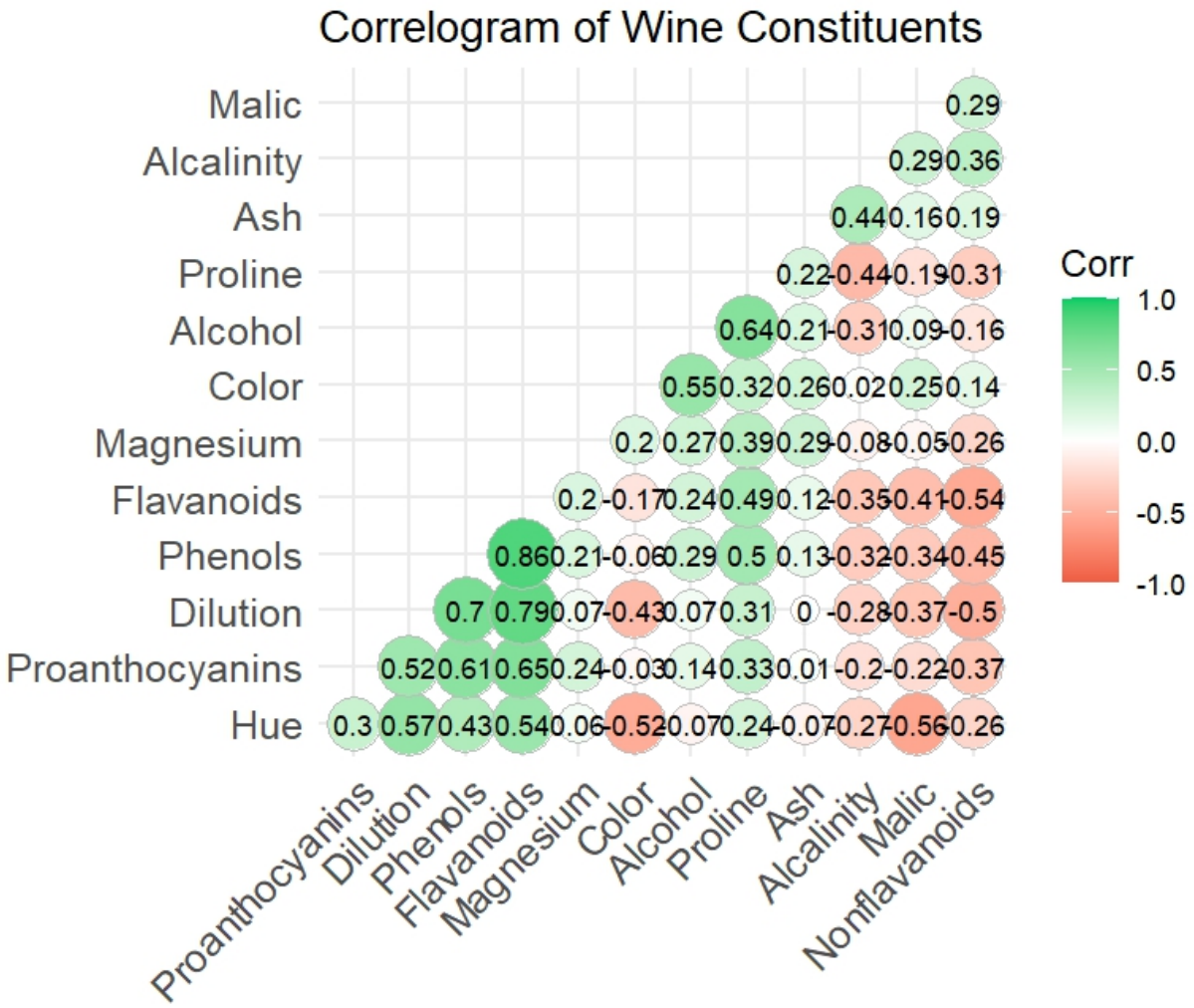
$$\widehat{\text{Alcohol}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Malic} + \hat{\beta}_2 \text{Ash} + \hat{\beta}_3 \text{Alcalinity} + \cdots + \hat{\beta}_{12} \text{Proline}$$

- $\hat{Y}$ (**Alcohol**): The predicted ethanol percentage.
- $\hat{\beta}_0$ (**Intercept**): The value of Alcohol when all predictors are zero.
- $X_{1\ldots12}$ (**Predictors**): The chemical constituents (e.g., Phenols, Magnesium).
- $\hat{\beta}_{1\ldots12}$ (**Slopes**): The change in Alcohol per unit change in the predictor, *adjusting for all other variables* in the model.

```
> summary(full_model)

Call:
lm(formula = Alcohol ~ ., data = wine_reg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45180 -0.30646 -0.02277  0.33195  1.54407

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.107e+01  5.963e-01  18.567  < 2e-16 ***
Malic           1.316e-01  4.528e-02   2.907  0.00415 **
Ash             1.379e-01  2.169e-01   0.636  0.52585
Alcalinity     -3.779e-02  1.781e-02  -2.122  0.03537 *
Magnesium       4.179e-06  3.359e-03   0.001  0.99901
Phenols         5.208e-02  1.340e-01   0.389  0.69796
Flavanoids      9.125e-03  1.069e-01   0.085  0.93211
Nonflavanoids  -2.078e-01  4.336e-01  -0.479  0.63242
Proanthocyanins -1.525e-01  9.823e-02  -1.552  0.12249
Color           1.630e-01  2.744e-02   5.941 1.63e-08 ***
Hue             2.169e-01  2.811e-01   0.772  0.44144
Dilution        1.608e-01  1.097e-01   1.466  0.14462
Proline         1.016e-03  1.999e-04   5.081 1.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5361 on 165 degrees of freedom
Multiple R-squared:  0.5936,    Adjusted R-squared:  0.564
F-statistic: 20.08 on 12 and 165 DF,  p-value: < 2.2e-16
```



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Model Selection

- In multiple regression we have several explanatory variables to consider
- Not all of them are important (explanatory for Y)
- We can adopt model selection strategies so we can come up with the optimal model
  - Forward Selection
  - Backward Elimination
  - Stepwise Selection
- The idea is to get the simplest possible model that describes the data as well as the more complicated models
- Select the criterion of our choice (AIC, BIC,p-value, R^2, R^2_{Adj}, etc)
- If the number of covariates is not big, we can run a grid search (fit all the models) and choose the optimal one (not here)

The step function in R

- R has the built-in function step that uses AIC as the criterion

AIC

The Akaike Information Criterion (AIC) is a metric used to compare different statistical models to see which one best explains your biological data without being unnecessarily complex. It balances "goodness of fit" with a penalty for adding too many parameters, helping you avoid overfitting and choose the most efficient model for your results.

- It checks all the covariates in the model and from those who induce a drop in AIC, it removes the one the generates the biggest drop
- It repeats the process until no other covariate can be removed
- This is the final (optimal) model based on AIC (with min AIC)

Analysis Starting Point

```
> optimal_model <- step(full_model, direction = "backward", trace = 1)
Start:  AIC=-209.47
Alcohol ~ Malic + Ash + Alcalinity + Magnesium + Phenols + Flavanoids +
    Nonflavanoids + Proanthocyanins + Color + Hue + Dilution +
    Proline + Predicted


Step:  AIC=-209.47
Alcohol ~ Malic + Ash + Alcalinity + Magnesium + Phenols + Flavanoids +
    Nonflavanoids + Proanthocyanins + Color + Hue + Dilution +
    Proline

                Df Sum of Sq    RSS     AIC
- Magnesium      1    0.0000 47.413 -211.47
- Flavanoids     1    0.0021 47.415 -211.47
- Phenols        1    0.0434 47.457 -211.31
- Nonflavanoids  1    0.0660 47.479 -211.23
- Ash            1    0.1161 47.529 -211.04
- Hue            1    0.1711 47.584 -210.83
<none>                      47.413 -209.47
- Dilution       1    0.6173 48.031 -209.17
- Proanthocyanins 1   0.6925 48.106 -208.89
- Alcalinity     1    1.2934 48.707 -206.68
- Malic          1    2.4289 49.842 -202.58
- Proline        1    7.4181 54.831 -185.60
- Color          1   10.1425 57.556 -176.97
```

Step 1

```
Step:  AIC=-211.47
Alcohol ~ Malic + Ash + Alcalinity + Phenols + Flavanoids + Nonflavanoids +
    Proanthocyanins + Color + Hue + Dilution + Proline

                Df Sum of Sq    RSS     AIC
- Flavanoids     1    0.0021 47.415 -213.47
- Phenols        1    0.0435 47.457 -213.31
- Nonflavanoids  1    0.0720 47.485 -213.20
- Ash            1    0.1263 47.540 -213.00
- Hue            1    0.1715 47.585 -212.83
<none>                       47.413 -211.47
- Dilution       1    0.6275 48.041 -211.13
- Proanthocyanins 1   0.7193 48.132 -210.79
- Alcalinity     1    1.2975 48.711 -208.67
- Malic          1    2.4298 49.843 -204.58
- Proline        1    7.6985 55.112 -186.69
- Color          1   10.1431 57.556 -178.97
```

Step 2

```
Step:  AIC=-213.47
Alcohol ~ Malic + Ash + Alcalinity + Phenols + Nonflavanoids +
    Proanthocyanins + Color + Hue + Dilution + Proline

                Df Sum of Sq    RSS     AIC
- Phenols        1    0.0805 47.496 -215.16
- Nonflavanoids  1    0.0818 47.497 -215.16
- Ash            1    0.1406 47.556 -214.94
- Hue            1    0.1826 47.598 -214.78
<none>                      47.415 -213.47
- Dilution       1    0.7333 48.149 -212.73
- Proanthocyanins 1   0.7624 48.178 -212.63
- Alcalinity     1    1.3363 48.752 -210.52
- Malic          1    2.4387 49.854 -206.54
- Proline        1    7.7153 55.131 -188.63
- Color          1   10.1446 57.560 -180.95
```

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Step 3

```
Step:  AIC=-215.16
Alcohol ~ Malic + Ash + Alcalinity + Nonflavanoids + Proanthocyanins +
    Color + Hue + Dilution + Proline

                Df Sum of Sq    RSS     AIC
- Nonflavanoids  1    0.0961 47.592 -216.80
- Ash            1    0.1828 47.679 -216.48
- Hue            1    0.2091 47.705 -216.38
<none>                       47.496 -215.16
- Proanthocyanins 1   0.6820 48.178 -214.63
- Dilution       1    1.2359 48.732 -212.59
- Alcalinity     1    1.4341 48.930 -211.87
- Malic          1    2.3892 49.885 -208.43
- Proline        1    7.9439 55.440 -189.63
- Color          1   10.8310 58.327 -180.60
```

Step 4

```
Step:  AIC=-216.8
Alcohol ~ Malic + Ash + Alcalinity + Proanthocyanins + Color +
    Hue + Dilution + Proline

                Df Sum of Sq    RSS     AIC
- Ash            1    0.1455 47.737 -218.26
- Hue            1    0.1798 47.772 -218.13
<none>                       47.592 -216.80
- Proanthocyanins 1    0.6379 48.230 -216.43
- Alcalinity     1    1.4922 49.084 -213.31
- Dilution       1    1.6322 49.224 -212.80
- Malic          1    2.3122 49.904 -210.36
- Proline        1    8.3602 55.952 -190.00
- Color          1   10.8081 58.400 -182.38
```

Step 5

```
Step:  AIC=-218.26
Alcohol ~ Malic + Alcalinity + Proanthocyanins + Color + Hue +
    Dilution + Proline

                 Df Sum of Sq    RSS     AIC
- Hue             1    0.2135 47.951 -219.47
<none>                       47.737 -218.26
- Proanthocyanins 1    0.7079 48.445 -217.64
- Alcalinity      1    1.5452 49.283 -214.59
- Dilution        1    1.8297 49.567 -213.56
- Malic           1    2.4802 50.218 -211.24
- Proline         1   10.5986 58.336 -184.57
- Color           1   11.7151 59.453 -181.20
```

Step 6

```
Step:  AIC=-219.47
Alcohol ~ Malic + Alcalinity + Proanthocyanins + Color + Dilution +
    Proline

                Df Sum of Sq    RSS     AIC
<none>                       47.951 -219.47
- Proanthocyanins  1    0.6583 48.609 -219.04
- Alcalinity       1    1.5601 49.511 -215.77
- Dilution         1    2.0486 50.000 -214.02
- Malic            1    2.3147 50.266 -213.07
- Proline          1   12.4058 60.357 -180.51
- Color            1   13.1338 61.085 -178.38
```

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Summary: List of Steps Taken (Removed Covariates)

```
> # To see a summary table of the steps taken:
> stepwise_process$anova
              Step Df     Deviance Resid. Df Resid. Dev       AIC
1                  NA           NA       165   47.41318 -209.4732
2     - Magnesium  1 4.447508e-07       166   47.41318 -211.4732
3    - Flavanoids  1 2.112225e-03       167   47.41529 -213.4653
4       - Phenols  1 8.052066e-02       168   47.49581 -215.1633
5 - Nonflavanoids  1 9.607345e-02       169   47.59188 -216.8036
6           - Ash  1 1.454973e-01       170   47.73738 -218.2602
7           - Hue  1 2.134824e-01       171   47.95086 -219.4660
```

Final Model (step function): AIC = -219.47

```
> summary(optimal_model)

Call:
lm(formula = Alcohol ~ Malic + Alcalinity + Proanthocyanins +
    Color + Dilution + Proline, data = wine_reg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50233 -0.34225  0.00116  0.33005  1.69364

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.3332831  0.3943623  28.738  < 2e-16 ***
Malic           0.1143127  0.0397878   2.873  0.00458 **
Alcalinity     -0.0324405  0.0137533  -2.359  0.01947 *
Proanthocyanins -0.1296362  0.0846088  -1.532  0.12732
Color           0.1585201  0.0231627   6.844 1.32e-10 ***
Dilution        0.2254528  0.0834109   2.703  0.00757 **
Proline         0.0011358  0.0001708   6.651 3.76e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5295 on 171 degrees of freedom
Multiple R-squared:  0.5889,    Adjusted R-squared:  0.5745
F-statistic: 40.83 on 6 and 171 DF,  p-value: < 2.2e-16
```

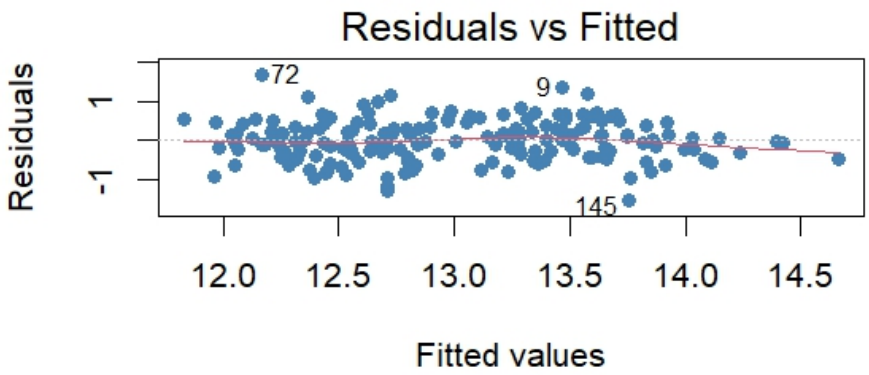
Final Model (Based on *p-value*): AIC = -219.04

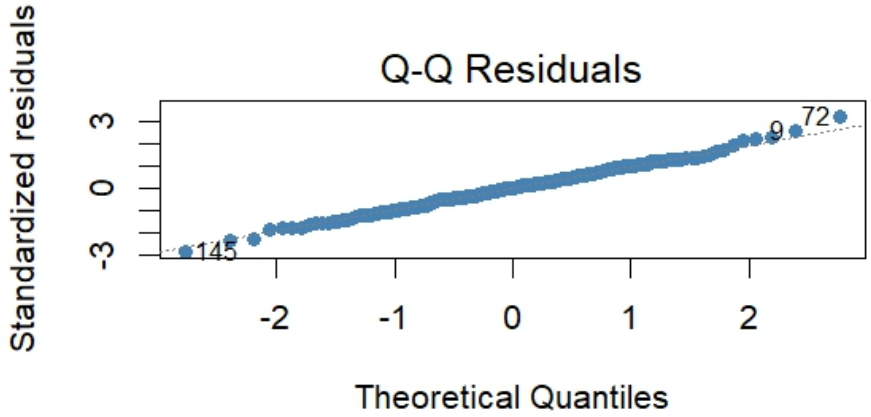
```
> final_model <- lm(Alcohol ~ Malic +       Alcalinity +
ution + Proline, data = wine_reg);   summary(final_model)

Call:
lm(formula = Alcohol ~ Malic + Alcalinity + Color + Dilution +
    Proline, data = wine_reg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46252 -0.36275  0.01755  0.31653  1.67562

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.3229479  0.3958463  28.604  < 2e-16 ***
Malic        0.1175677  0.0398864   2.948  0.00365 **
Alcalinity  -0.0325671  0.0138068  -2.359  0.01946 *
Color        0.1520614  0.0228649   6.650 3.73e-10 ***
Dilution     0.1666219  0.0743373   2.241  0.02628 *
Proline      0.0011161  0.0001709   6.529 7.17e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5316 on 172 degrees of freedom
Multiple R-squared:  0.5833,    Adjusted R-squared:  0.5712
F-statistic: 48.15 on 5 and 172 DF,  p-value: < 2.2e-16
```

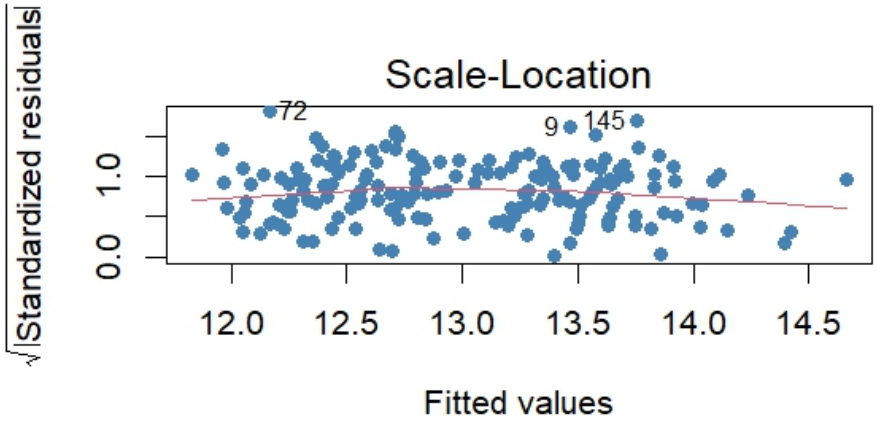
HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

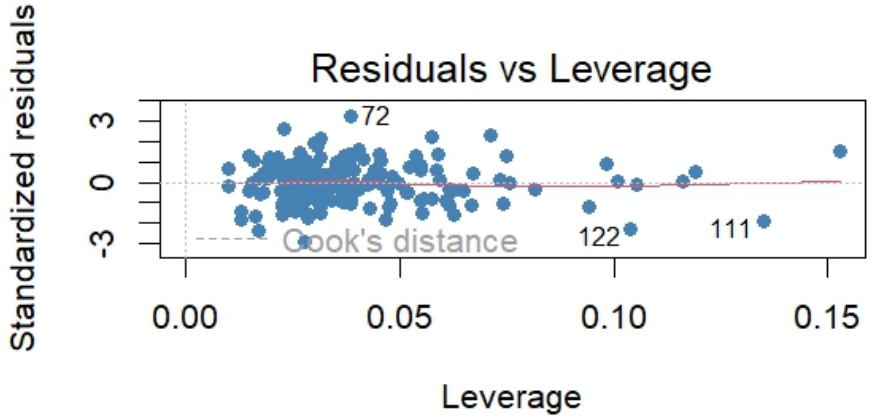
```
> confint(optimal_model)
                              2.5 %        97.5 %
(Intercept)          10.5548379023 12.111728329
Malic                 0.0357741959  0.192851143
Alcalinity           -0.0595885887 -0.005292358
Proanthocyanins      -0.2966483219  0.037375869
Color                 0.1127984476  0.204241654
Dilution              0.0608051437  0.390100536
Proline               0.0007987109  0.001472841
> confint(p_final_model)
                         2.5 %        97.5 %
(Intercept) 10.5416058252 12.104289928
Malic        0.0388378771  0.196297492
Alcalinity  -0.0598197091 -0.005314435
Color        0.1069293781  0.197193352
Dilution     0.0198910799  0.313352715
Proline      0.0007786563  0.001453471
```

Regression Diagnostics (from `step` function)









National and Kapodistrian University of Athens
EST. 1837

Interpreting Regression Diagnostic Plots

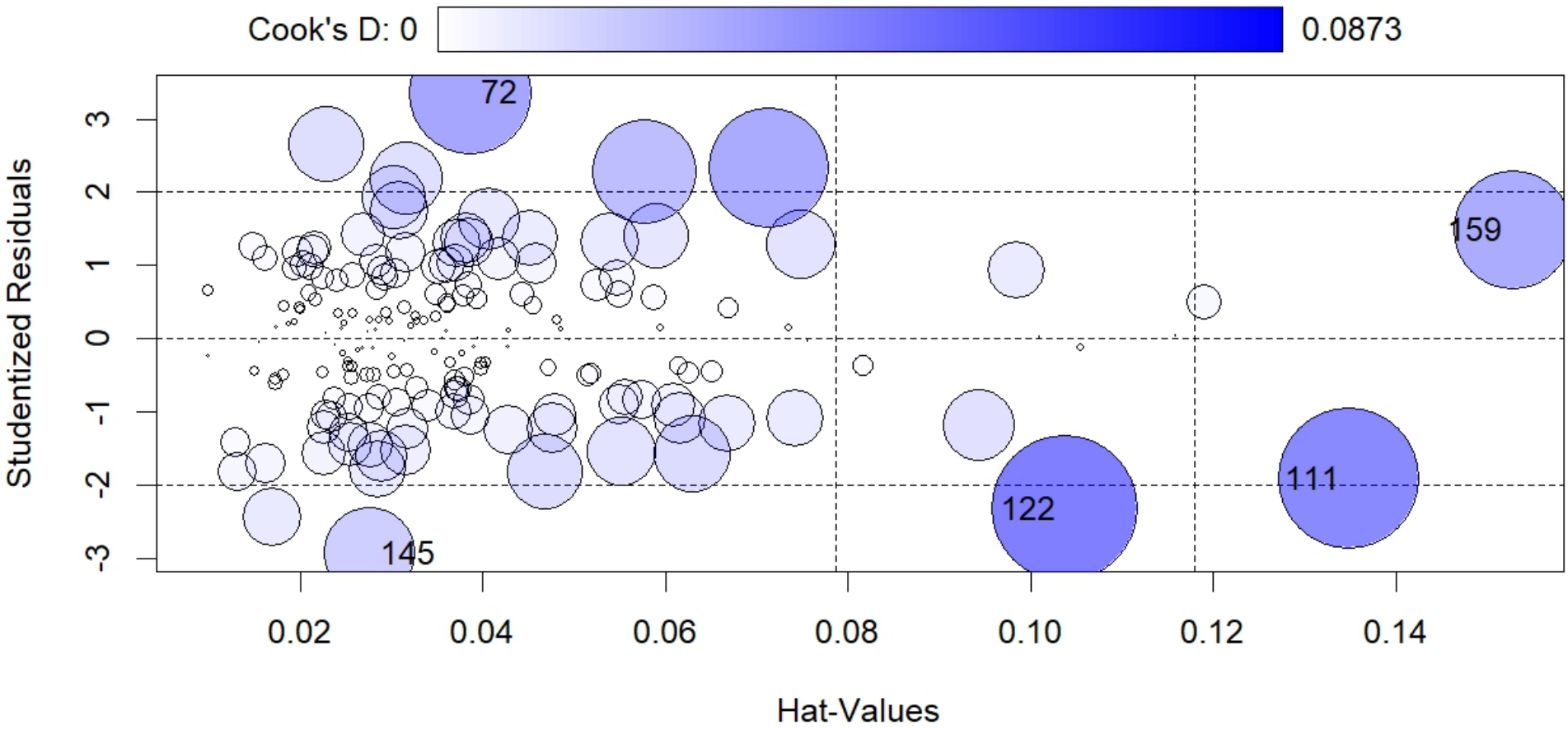
- **Residuals vs Fitted:** Tests *Linearity* and $E(\epsilon) = 0$.
  If points are randomly scattered around the zero line, the relationship is linear. A "U" shape suggests a missing non-linear term.
- **Normal Q-Q:** Tests *Normality of Residuals*.
  Points should follow the dashed diagonal line. Deviations at the ends indicate "heavy tails" or outliers.
- **Scale-Location:** Tests *Homoscedasticity*.
  If the red line is flat and the "spread" of points is constant, the variance of your errors is stable across all alcohol levels.
- **Residuals vs Leverage:** Identifies *Influential Observations*.
  Points outside "Cook's Distance" (red dashed lines) are outliers that disproportionately pull the regression line.

Identifying Influential Wines: The Influence Plot

Purpose: Detects individual samples that disproportionately "pull" the regression line, potentially biasing biological conclusions.

The Three Dimensions of the Plot - What to Look For:

1. **Y-Axis (Studentized Residuals):** Measures *outliers*. High values ($> |2|$) indicate wines where the observed alcohol differs significantly from the model's prediction.
2. **X-Axis (Hat Values):** Measures *leverage*. Points far to the right have unusual chemical profiles (e.g., extreme Proline levels) compared to the rest of the dataset.
3. **Circle Size (Cook's Distance):** Combines residual and leverage. Large circles represent the most "dangerous" points that could change your β coefficients if removed.



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

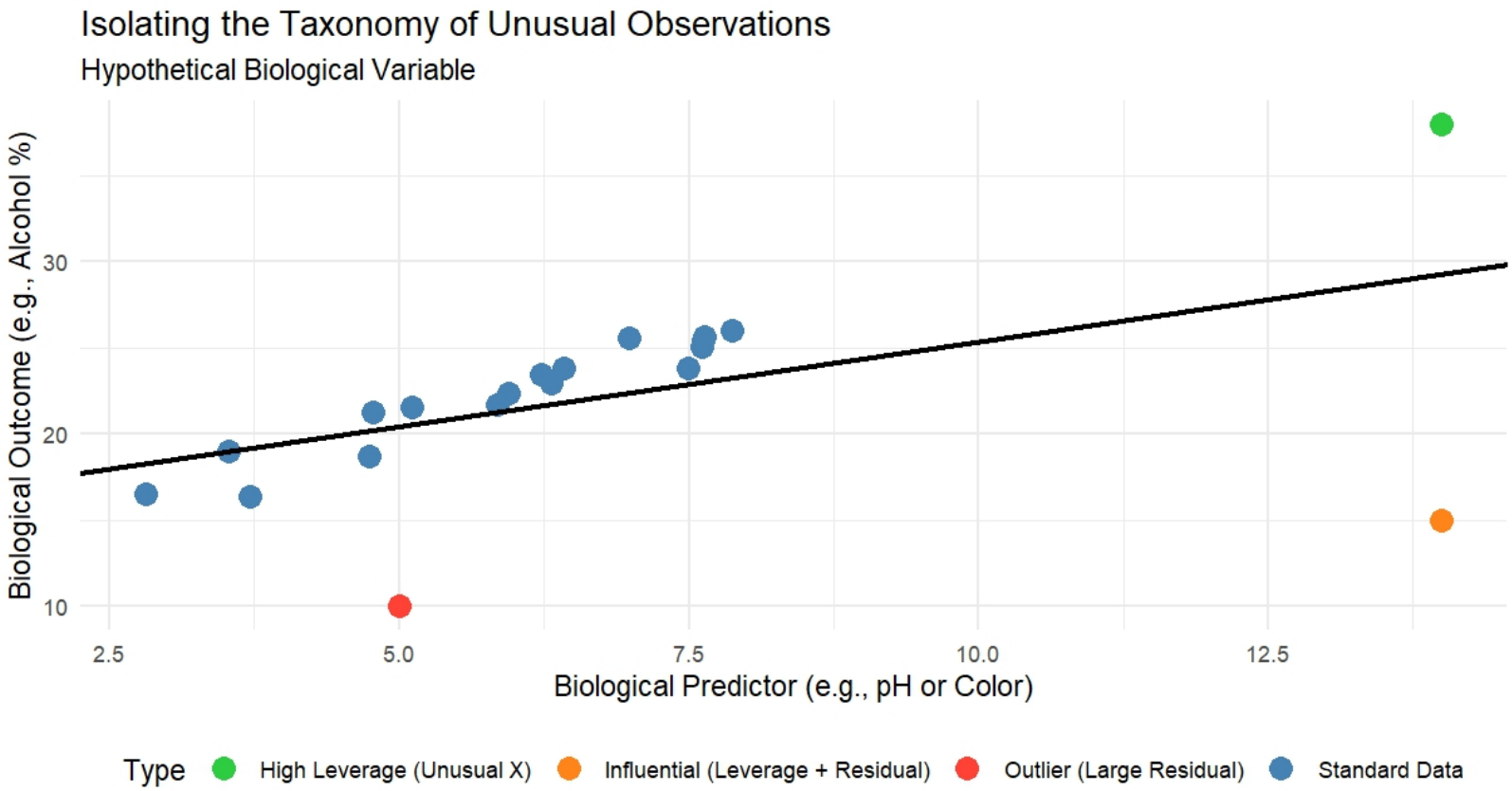
Outliers vs. Influential Points

- **Outlier:** An observation with a large *residual* (the model is wrong).
- **High Leverage:** An observation with extreme *predictor values* (the wine is chemically "weird").
- **Influential Point:** An observation that changes the *slope* of the regression line.

Biological Caution

In biology, we don't always delete these points! A "High Influence" wine might be a new hybrid cultivar or a sample that was fermented under different conditions. It's often the most interesting data point in the set.

Hypothetical Example



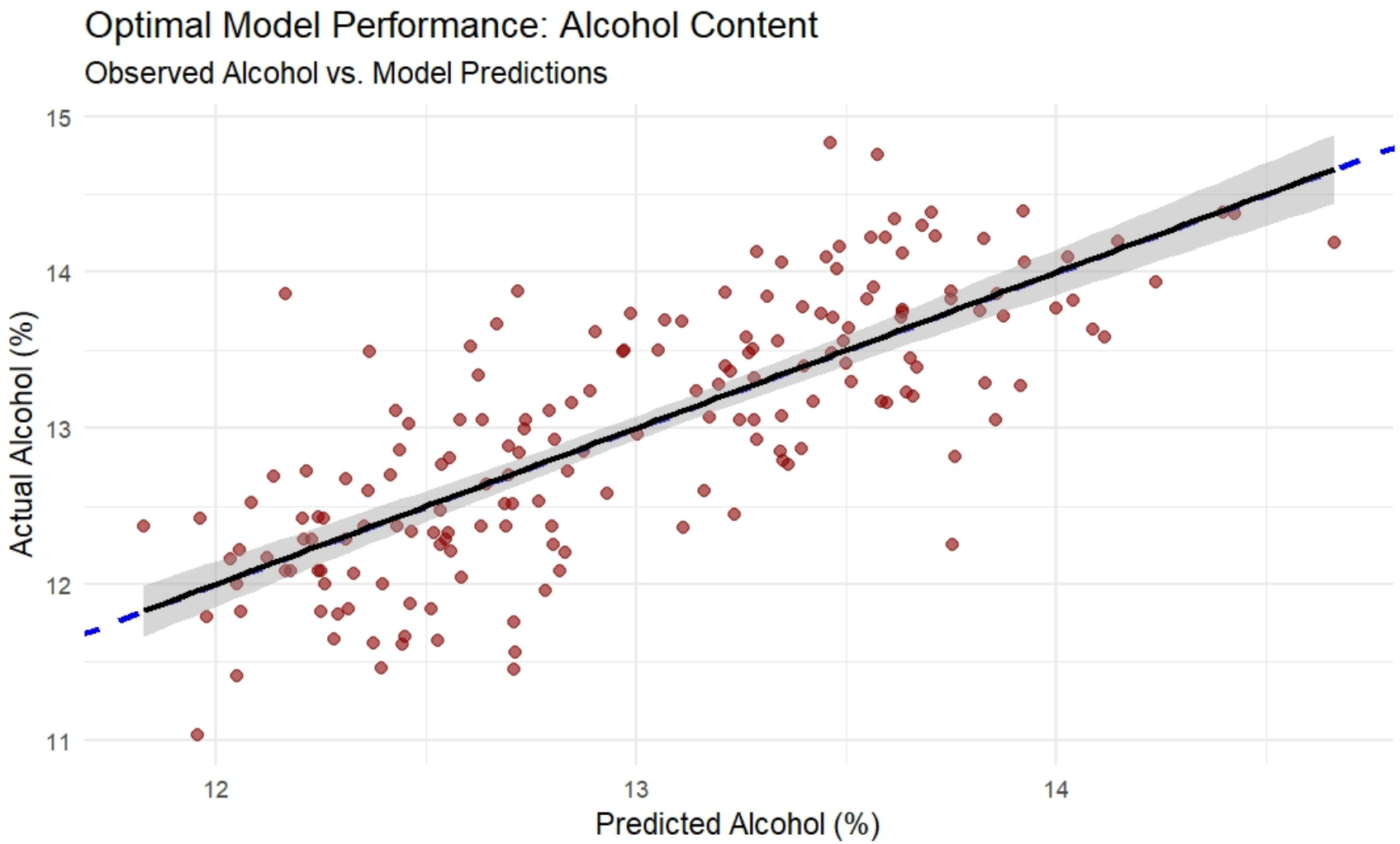
HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Conclusion: Chemical Drivers of Alcohol Content

- **Model Parsimony:** Using AIC-based backward selection, we reduced 12 chemical variables to a core set of robust biological predictors.
- **Key Biological Findings:**
  - **Color Intensity & Proline:** Consistently the strongest positive predictors. In viticulture, high nitrogen (Proline) and pigment synthesis often correlate with higher sugar accumulation/alcohol potential.
  - **Multicollinearity:** Total Phenols and Flavanoids were redundant; the model selection process correctly prioritized the more informative variable.
- **Statistical Validity:** Diagnostic plots confirmed that the assumptions of linearity and normality hold, making our p-values and confidence intervals reliable.

Takeaway for Biologists

Multiple regression allows us to disentangle complex chemical relationships, but we must always check for influential outliers that may represent rare cultivars or lab errors.



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Predicted vs. Observed plot

This is the standard way to visualize how well a statistical model is performing

- **X-axis (Predicted)**: What the mathematical model thought the alcohol content should be
- **Y-axis (Actual)**: What was actually measured in the lab
- **The Black Line & Grey Ribbon**: This is the regression line of the predictions. The grey ribbon represents the 95% CI (the narrower that ribbon, the more certain the model is)

How to Read the Accuracy

- **The Ideal Scenario**: In a perfect model, every red dot would fall exactly on a 45° diagonal line. This would mean *Predicted* = *Actual*
- **Current Performance**: The dots are clustered relatively tightly around the line, indicating a strong positive correlation. However, there is some "scatter" (residuals).
- **Under/Over Estimation**: Dots above the line represent cases where actual alcohol was higher than the model predicted. Dots below the line represent cases where the model overestimated alcohol content.
- **Biological/Practical Interpretation**: Looking at the data, the model seems very reliable between 12% and 14% alcohol.
- **Potential Outliers**: Notice the point in bottom left (Actual 11%, Predicted 12%). This point has high residual and could be an outlier (perhaps a fermentation that stalled or a lab measurement error).
- **Precision**: The spread of dots suggests that while the model captures overall trend well, there is still some unexplained biological/chemical variability (noise) that the model isn't accounting for.

Presentation Outline

Example: Pima Indians Diabetes Dataset

Data: Pima Indians Diabetes Database

Context: This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases in the US. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Source:
- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- R-package: *mlbench*

Objective: Predict whether a woman of Pima Indian heritage has diabetes based on clinical variables.

Response variable: *diabetes* (pos/neg)

Predictive variables:

The Pima Indians Diabetes Dataset

pregnant:	Number of times pregnant
glucose:	Plasma glucose concentration (glucose tolerance test)
pressure:	Diastolic blood pressure (mm Hg)
triceps:	Triceps skin fold thickness (mm)
insulin:	2-Hour serum insulin (mu U/ml)
mass:	Body mass index (BMI)
pedigree:	Diabetes pedigree function
age:	Age (years)

Logistic Model

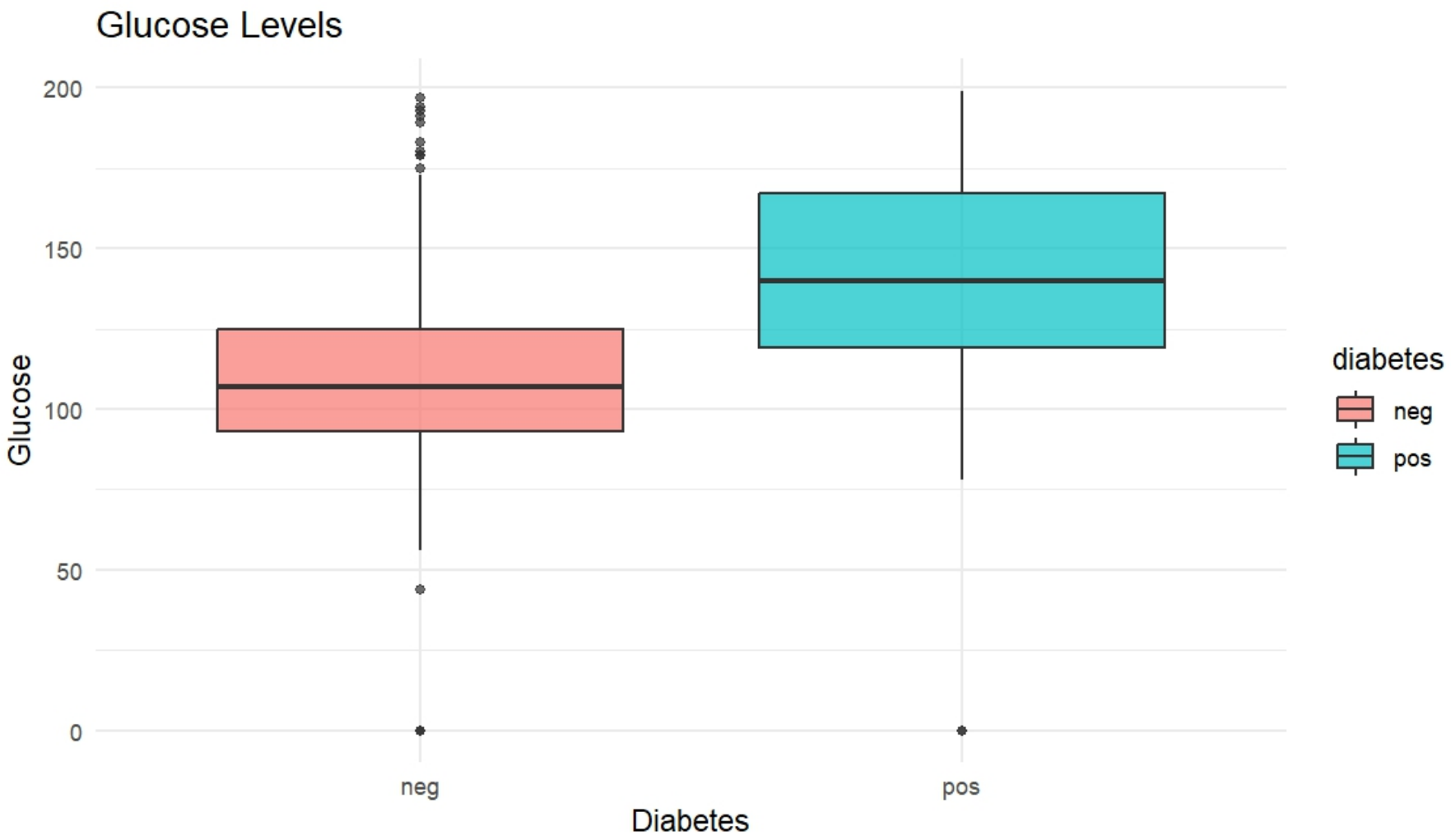
- Many biological and medical outcomes are **binary**: disease / no disease, alive / dead, success / failure.
- Linear regression is not appropriate for binary outcomes because predicted values may fall outside $[0, 1]$.
- **Logistic Regression** models the *probability* of success as a function of continuous and categorical predictors.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

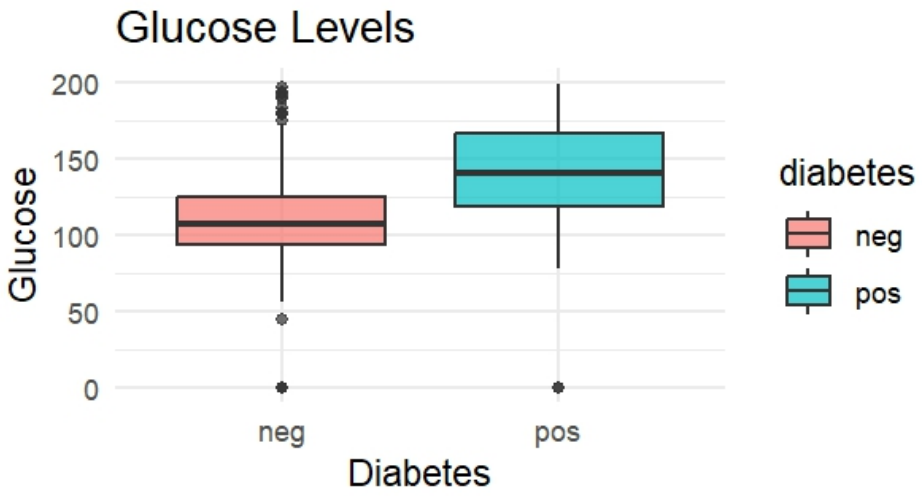
```
> library(mlbench)
> data(PimaIndiansDiabetes)
> head(PimaIndiansDiabetes)
  pregnant glucose pressure triceps insulin mass pedigree age diabetes
1        6     148       72      35       0 33.6    0.627  50      pos
2        1      85       66      29       0 26.6    0.351  31      neg
3        8     183       64       0       0 23.3    0.672  32      pos
4        1      89       66      23      94 28.1    0.167  21      neg
5        0     137       40      35     168 43.1    2.288  33      pos
6        5     116       74       0       0 25.6    0.201  30      neg
```

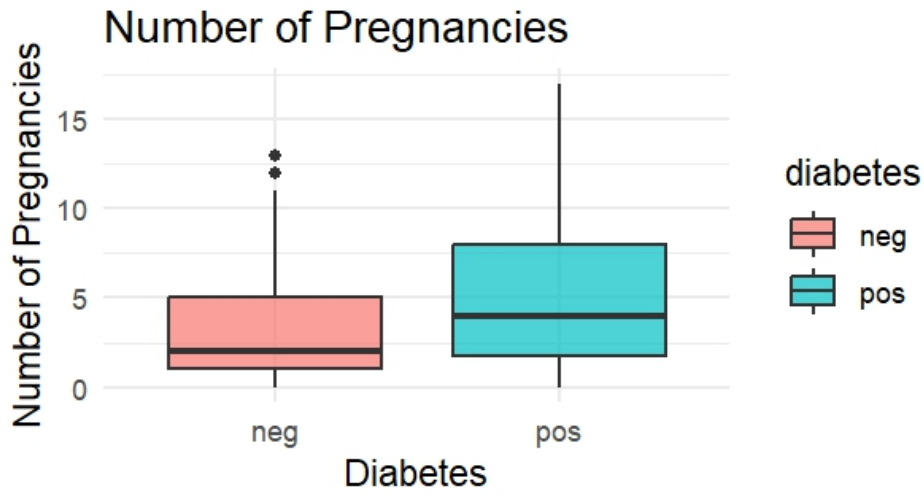
- We will start with some descriptive statistics in order to "take a flavor" of the data.
- This will help us with the modeling part and definitely gives us an idea of what to expect in the analysis that will follow.

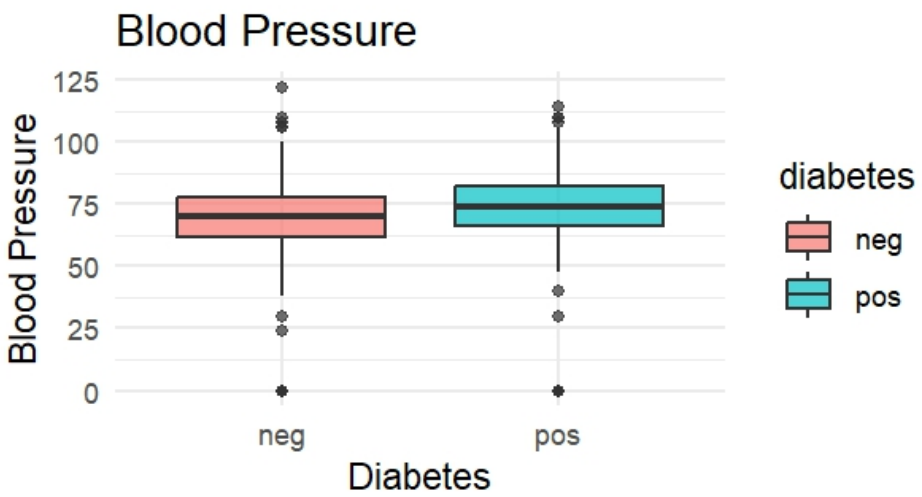
Box Plots

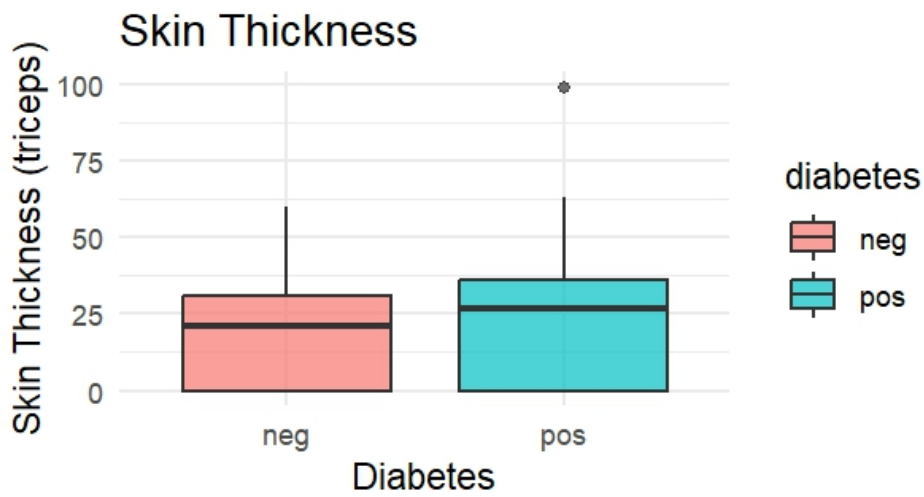


National and Kapodistrian University of Athens
EST. 1837

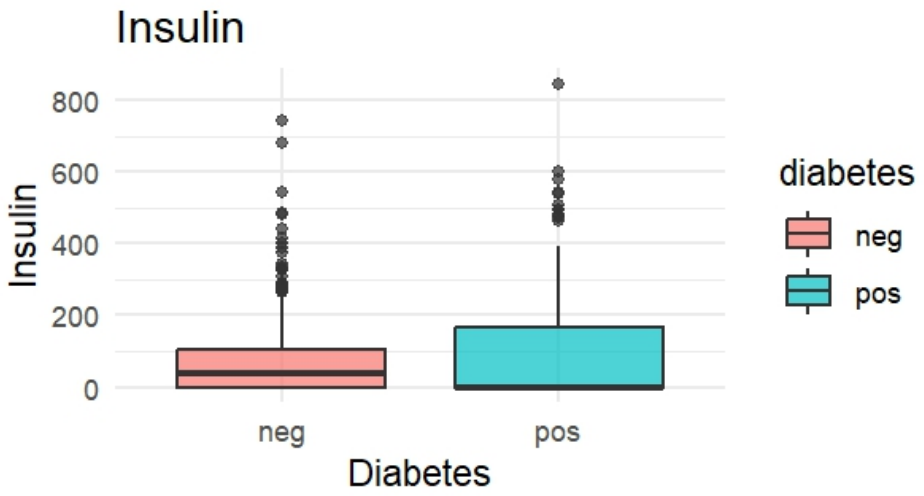


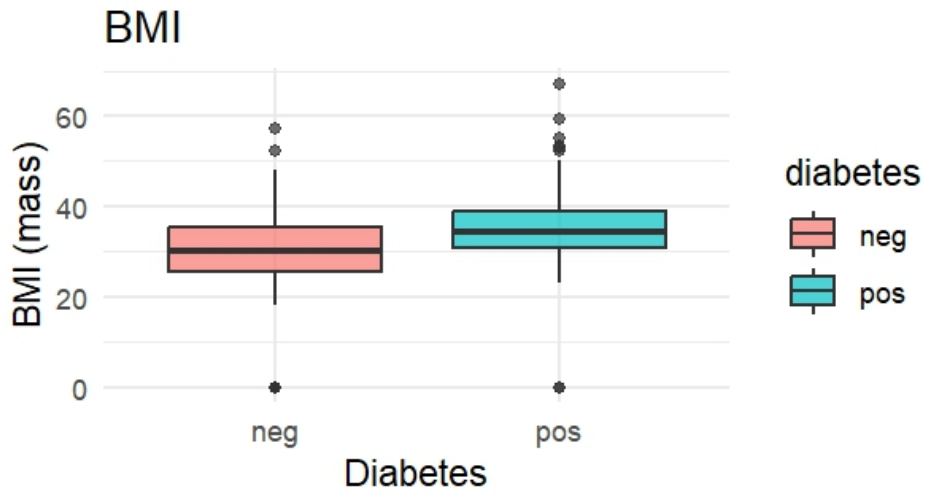


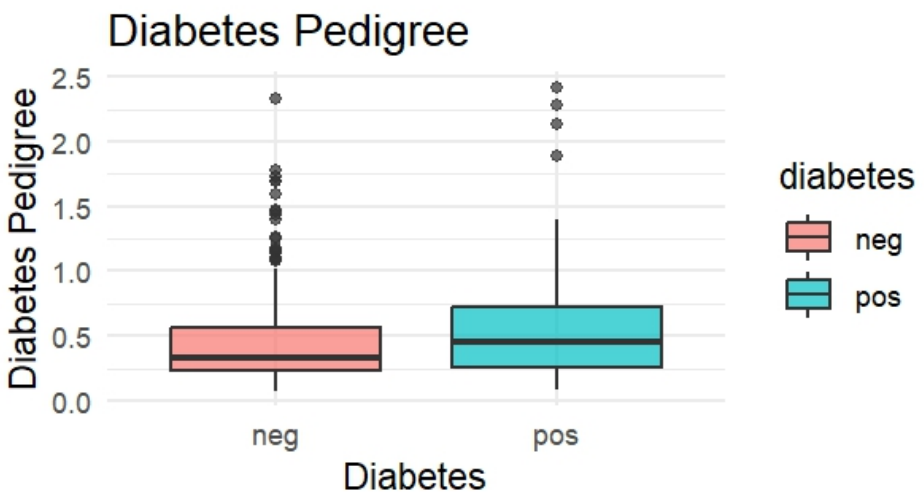


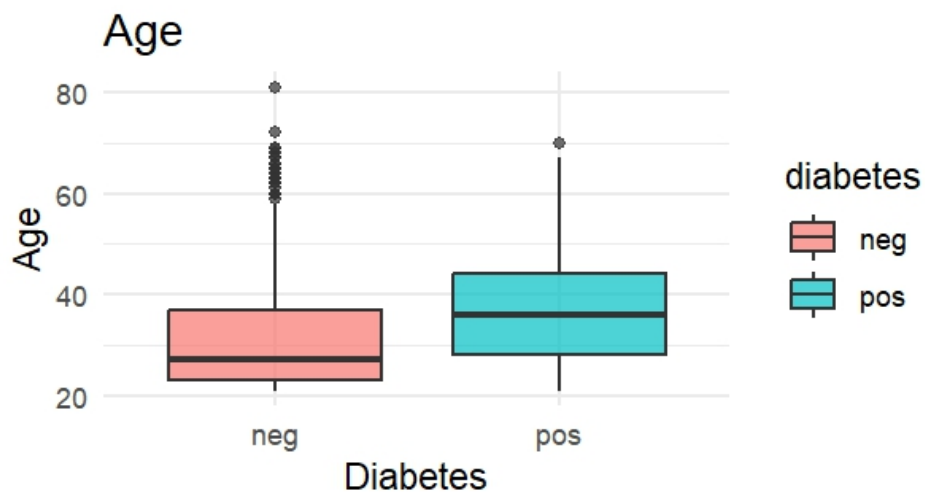


HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837









HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Logistic Regression Model (Pima Indians Data)

Objective: Model the probability that a woman has diabetes ($\texttt{diabetes} = pos$) based on physiological predictors.

Model form:

$$\begin{aligned}\text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\,\texttt{pregnant}_i + \beta_2\,\texttt{glucose}_i \\ &+ \beta_3\,\texttt{pressure}_i + \beta_4\,\texttt{triceps}_i + \beta_5\,\texttt{insulin}_i \\ &+ \beta_6\,\texttt{mass}_i + \beta_7\,\texttt{pedigree}_i + \beta_8\,\texttt{age}_i\end{aligned}$$

$$\eta_i = \beta_0 + \sum_{j=1}^{8}\beta_j X_{ji}, \qquad p_i = P(\texttt{Diabetes}_i = 1) = \frac{1}{1+e^{-\eta_i}}$$

Model in R

```
glm_fit <- glm(diabetes ~ ., data = pid, family = binomial)
summary(glm_fit)
```

```
Call:
glm(formula = diabetes ~ ., family = binomial, data = pid)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.4046964  0.7166359 -11.728  < 2e-16 ***
pregnant     0.1231823  0.0320776   3.840 0.000123 ***
glucose      0.0351637  0.0037087   9.481  < 2e-16 ***
pressure    -0.0132955  0.0052336  -2.540 0.011072 *
triceps      0.0006190  0.0068994   0.090 0.928515
insulin     -0.0011917  0.0009012  -1.322 0.186065
mass         0.0897010  0.0150876   5.945 2.76e-09 ***
pedigree     0.9451797  0.2991475   3.160 0.001580 **
age          0.0148690  0.0093348   1.593 0.111192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

Remove: triceps

```
Call:
glm(formula = diabetes ~ . - triceps, family = binomial, data = pid)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.4051362  0.7167033 -11.727  < 2e-16 ***
pregnant     0.1231724  0.0320688   3.841 0.000123 ***
glucose      0.0351123  0.0036625   9.587  < 2e-16 ***
pressure    -0.0132136  0.0051537  -2.564 0.010350 *
insulin     -0.0011570  0.0008142  -1.421 0.155275
mass         0.0900886  0.0144619   6.229 4.68e-10 ***
pedigree     0.9475954  0.2980063   3.180 0.001474 **
age          0.0147888  0.0092897   1.592 0.111393
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 760  degrees of freedom
AIC: 739.45

Number of Fisher Scoring iterations: 5
```

Remove: insulin

```
Call:
glm(formula = diabetes ~ . - triceps - insulin, family = binomial,
    data = pid)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.239812   0.701970 -11.738  < 2e-16 ***
pregnant     0.124919   0.031972   3.907 9.34e-05 ***
glucose      0.033492   0.003440   9.736  < 2e-16 ***
pressure    -0.013487   0.005114  -2.637  0.00836 **
mass         0.087676   0.014268   6.145 7.99e-10 ***
pedigree     0.896150   0.294862   3.039  0.00237 **
age          0.016325   0.009237   1.767  0.07719 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 725.46  on 761  degrees of freedom
AIC: 739.46

Number of Fisher Scoring iterations: 5
```

Interpretation

- The model estimates the **log-odds** of diabetes.
- A positive coefficient increases the probability of diabetes.
- Each β_j represents the change in log-odds per one-unit increase in the corresponding variable, holding others constant.
- $\exp(\beta_j)$ gives the **odds ratio**.

- The estimated coefficient for `glucose` from the final model is $\hat{\beta}_{glu} = 0.033$.
- Then for each additional unit of `glucose`:

$$\text{Odds Ratio} = \exp(0.033) \approx 1.034$$

Each unit increase in Glucose Level raises the odds of having diabetes by about 3.6%.

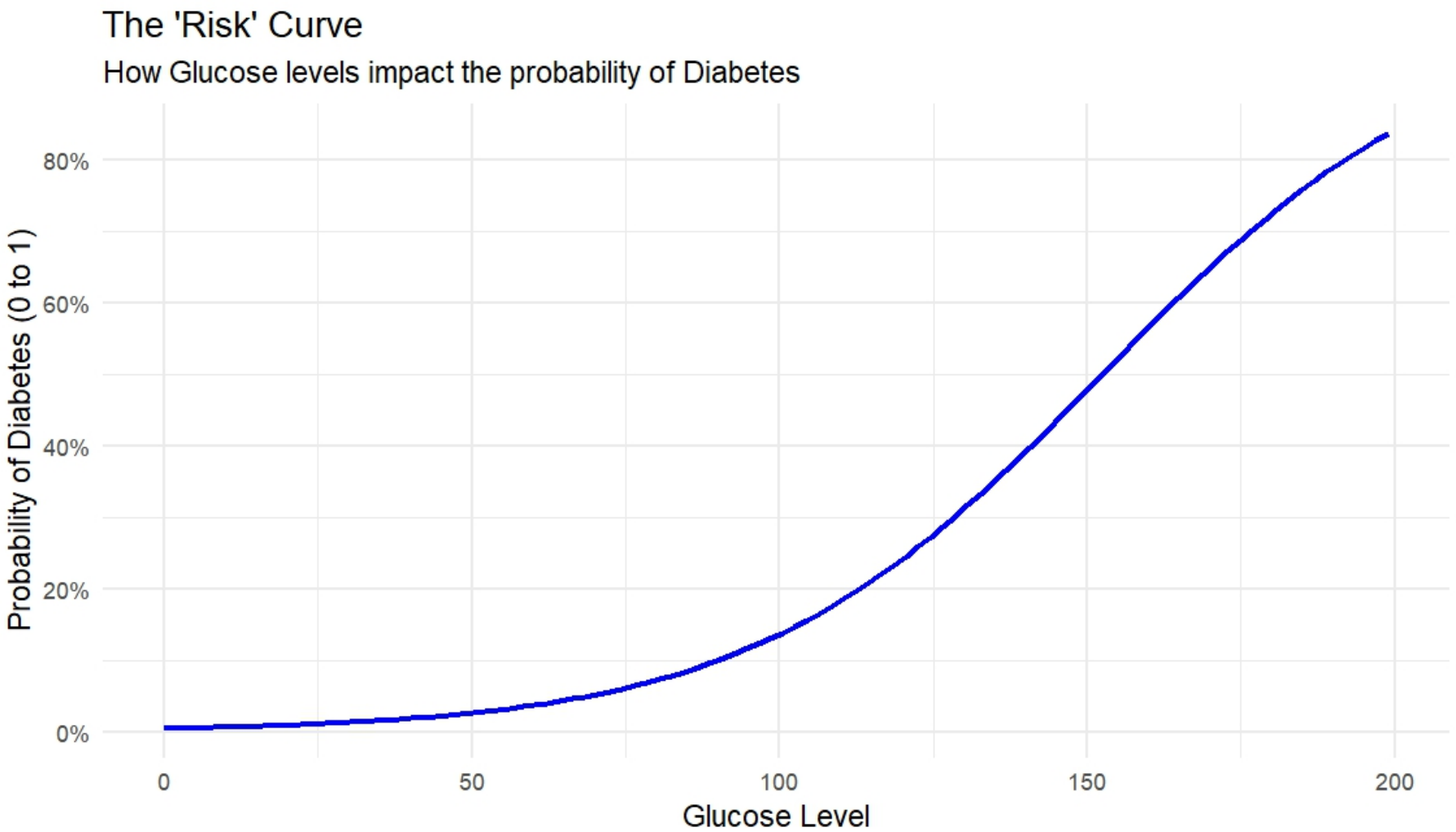
- Similarly, negative coefficients correspond to protective effects. For `pressure`:

$$\text{Odds Ratio} = \exp(-0.013) \approx 0.987$$

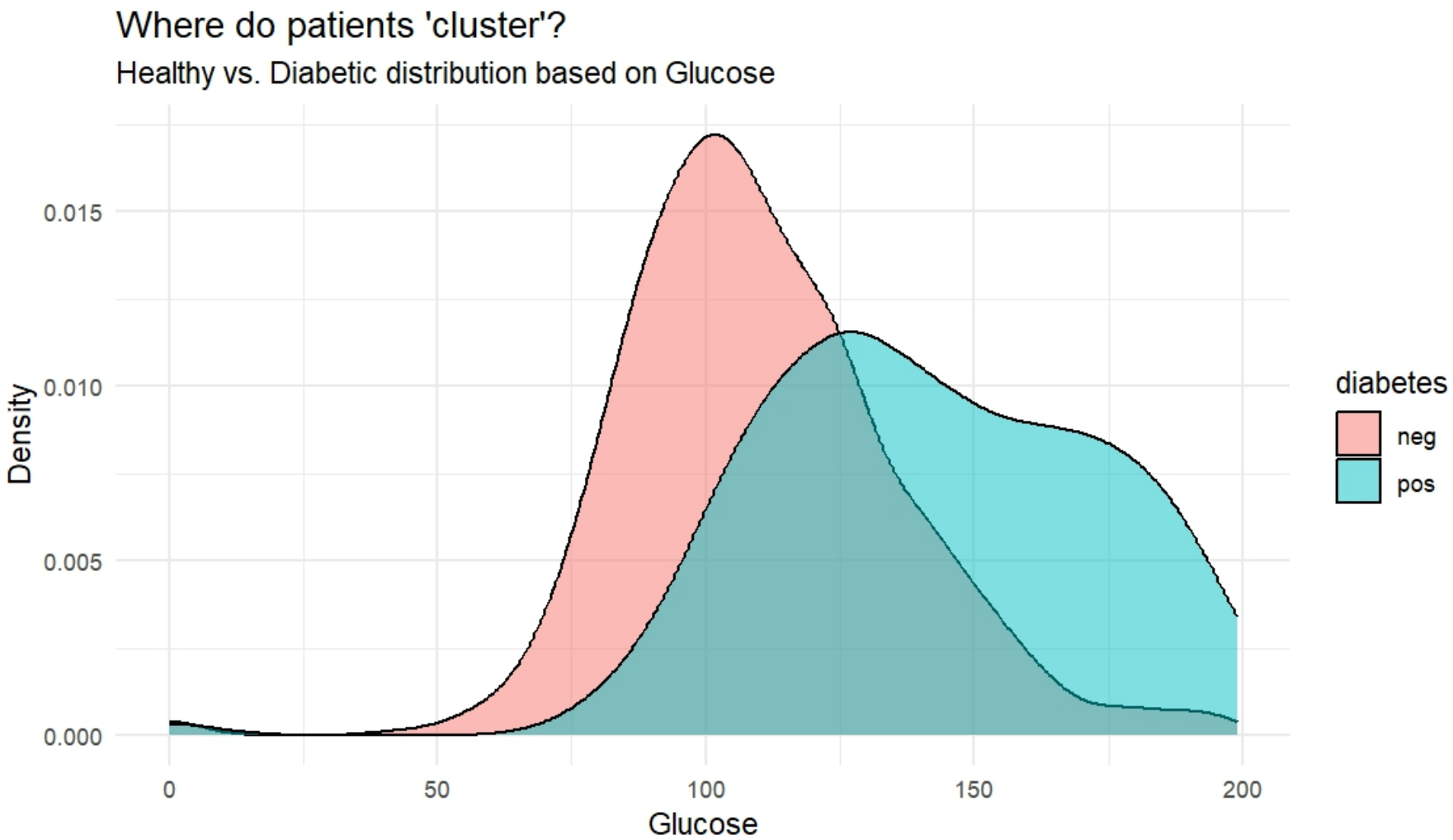
Each unit increase in Blood Pressure reduces the odds of having diabetes by about 1.3%.

Predicted Probabilities and Visualization

```
# Predicted probabilities
predicted_pid <- data.frame(glucose = seq(min(pid$glucose), max(pid$glucose), len=1
# Keep other variables at their average/median
predicted_pid$pregnant <- median(pid$pregnant)
predicted_pid$pressure <- median(pid$pressure)
predicted_pid$triceps  <- median(pid$triceps)
predicted_pid$insulin  <- median(pid$insulin)
predicted_pid$mass     <- median(pid$mass)
predicted_pid$pedigree <- median(pid$pedigree)
predicted_pid$age      <- median(pid$age)
#
predicted_pid$prob <- predict(model, newdata = predicted_pid, type = "response")
#
ggplot(predicted_pid, aes(x = glucose, y = prob)) +
geom_line(color = "blue", size = 1) +
labs(title = "The 'Risk' Curve",
subtitle = "How Glucose levels impact the probability of Diabetes",
x = "Glucose Level", y = "Probability of Diabetes (0 to 1)") +
scale_y_continuous(labels = scales::percent) +
theme_minimal()
```



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Key Takeaways

- Logistic regression models probabilities via the logit link function.
- Coefficients describe how predictors modify the odds of an event.
- Model fit and diagnostics can be assessed using deviance, ROC curves, and classification accuracy.
- The Pima Indians dataset is a standard benchmark for illustrating binary outcomes in medical data.

Model Evaluation: Confusion Matrix

Goal: Assess how well the logistic regression model classifies individuals as diabetic or non-diabetic (probability threshold $= 0.5$).

Output:

	Predicted: Neg	**Predicted: Pos**
Observed: Neg	445	55
Observed: Pos	112	156

Interpretation:

- Accuracy $= (445 + 156)/(445 + 55 + 112 + 156) = 0.78$
- Sensitivity (True Positive Rate) $= 156/(156 + 112) = 0.58$
- Specificity (True Negative Rate) $= 445/(445 + 55) = 0.89$

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

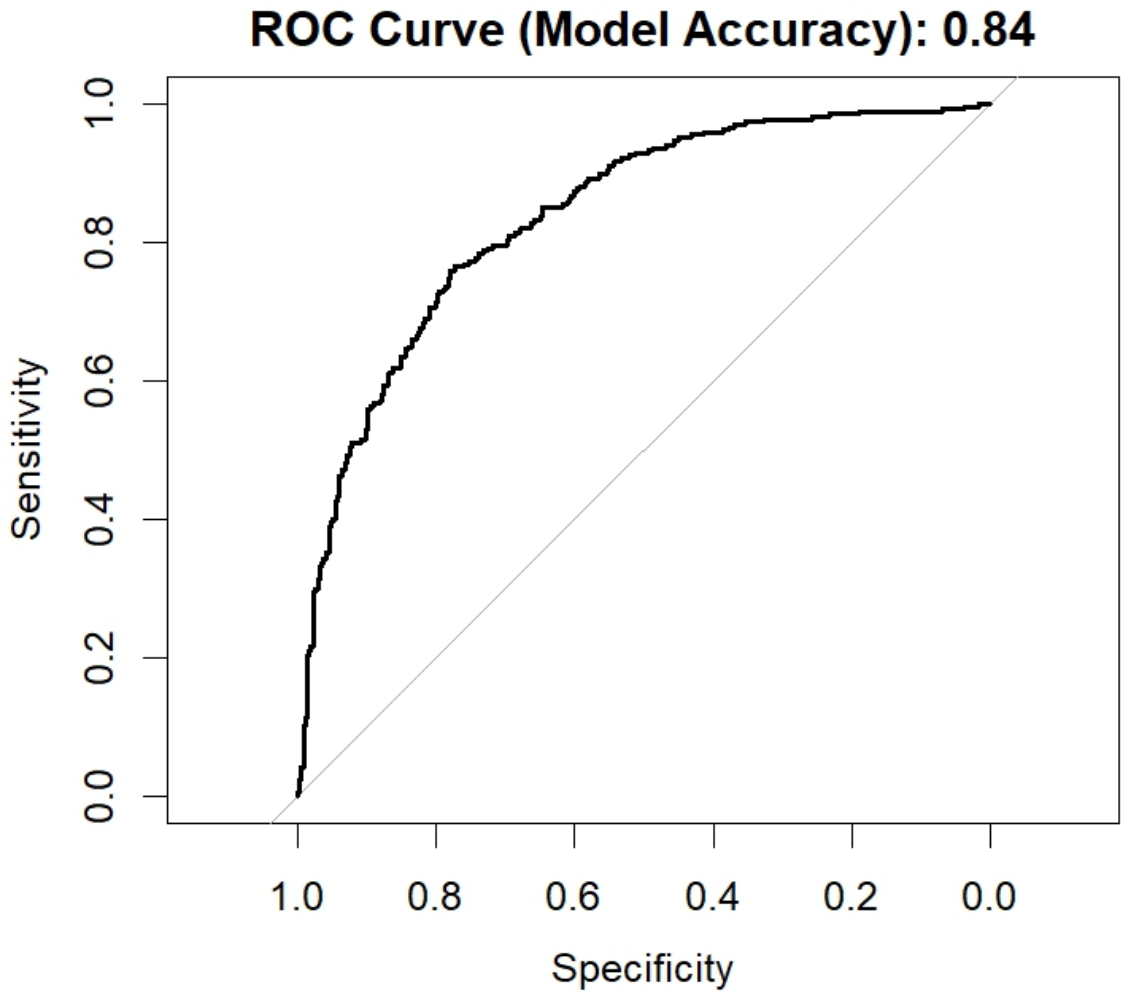
Youden's J Statistic: It can be shown that the optimal probability threshold is not 0.5 but 0.3537. So, if we use this one, we get:

Output:

	Predicted: Neg	**Predicted: Pos**
Observed: Neg	392	108
Observed: Pos	70	198

Interpretation:

- Accuracy $= (392 + 198)/768 = 0.77$
- Sensitivity (True Positive Rate) $= 198/(198 + 70) = 0.74$
- Specificity (True Negative Rate) $= 392/(392 + 108) = 0.78$



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Summary

Risk Curve: Use this to say: "Look how the risk jumps from 25% to 75% once Glucose crosses 150."

Density Plot: Use this to say: "See how the two groups are separated? This is why our model works."

ROC Curve: Use this to say: "Check how well the logistic regression model discriminates between individuals with diabetes (pos) and those without diabetes (neg)".
Note: It is independent of any single classification threshold.

Presentation Outline

Random Effects Models

- We need to understand (at least qualitatively) what are the likely sources of random variation
- One possible source is `Random Effects`, when units are sampled at random from a population and various aspects of their behavior may show stochastic variation between units
- We introduce Linear Random Effects model where
  - the response is assumed to be a linear function of exploratory variables with regression coefficients that vary from one individual to the next
  - variability reflects natural/biological heterogeneity due to unmeasured factors

Linear Mixed Model

- As presented before, the Simple Linear Model is of the form:

$$y_{ij} = \beta_0 + \beta_1 time_{ij} + \epsilon_i,$$

  where:

  - β_0: the *intercept*
  - β_1: the *slope*
  - $i = 1, 2, \ldots, n$ where n the size of the sample
  - $j = 1, 2, \ldots$ is the time instance (#1, #2, etc observation)

- `Time` is a key covariate in longitudinal studies

Linear Mixed Model

- As presented before, the Simple Linear Model is of the form:

$$y_{ij} = \beta_0 + \beta_1 time_{ij} + \epsilon_i,$$

where:

  - β_0: the *intercept*
  - β_1: the *slope*
  - $i = 1, 2, \ldots, n$ where n the size of the sample
  - $j = 1, 2, \ldots$ is the time instance (#1, #2, etc observation)

- `Time` is a key covariate in longitudinal studies

In linear mixed models, random effects are added in parts of the model that are expected to show unmeasured or unpredictable variability, allowing the model to better capture real differences among experimental units

National and Kapodistrian University of Athens
EST. 1837

The plants data

Dataset Description: The dataset comes from a controlled biological experiment designed to study **plant growth over time** under two treatments:

- Control
- Fertilizer

The goals of the study are to understand:

(1) how plant height changes over time

(2) whether treatment affects growth, and

(3) how much variability exists among individual plants

This structure makes the dataset well suited for both linear and mixed-effects modelling.

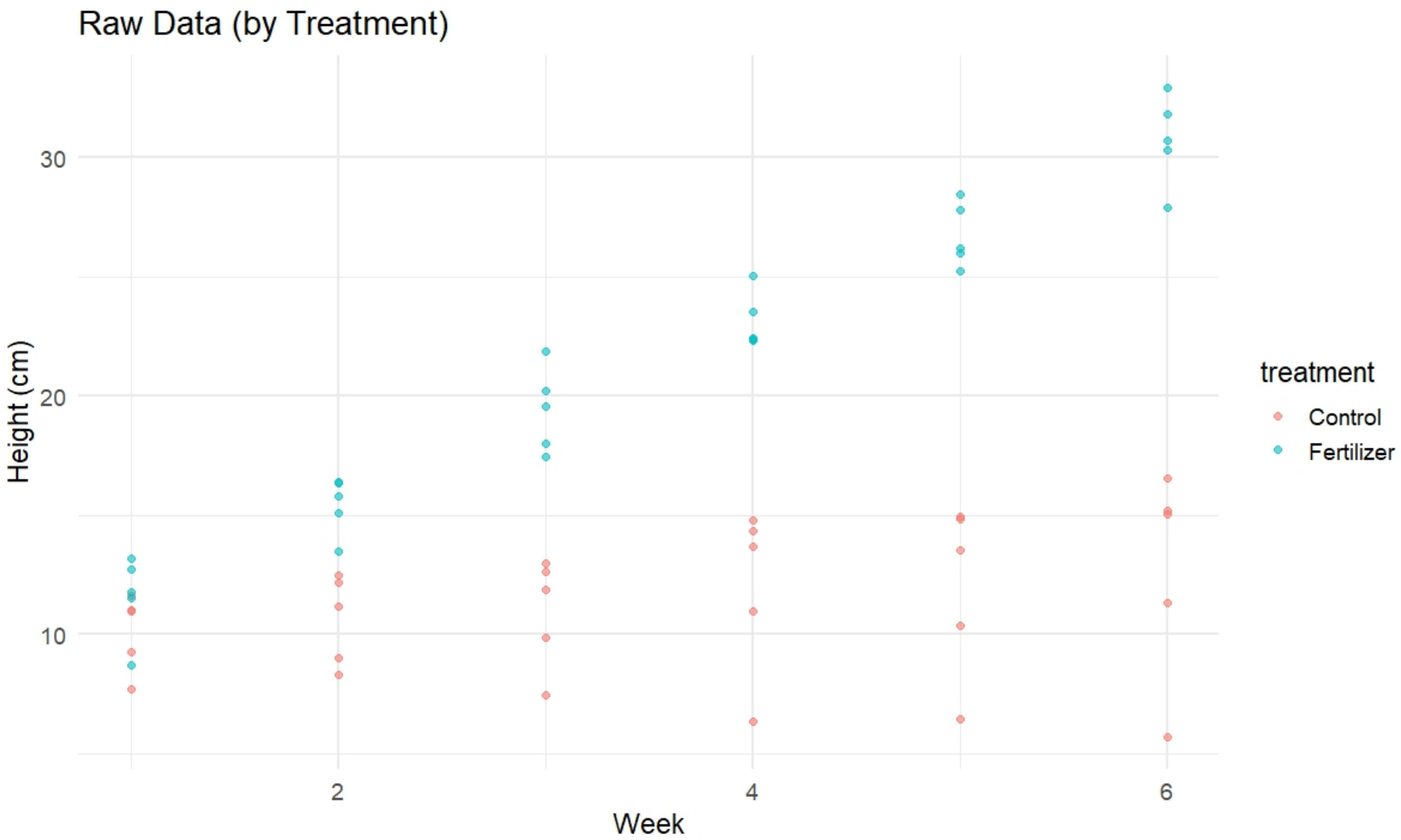
Subjects: Ten individual plants (P1–P10) were included in the study. Plants P1–P5 belong to the Control group, while plants P6–P10 belong to the Fertilizer group.

Measurements: Plant height was measured once per week for six consecutive weeks. Each plant therefore has six repeated measurements, giving a total of sixty observations. This is a balanced dataset.

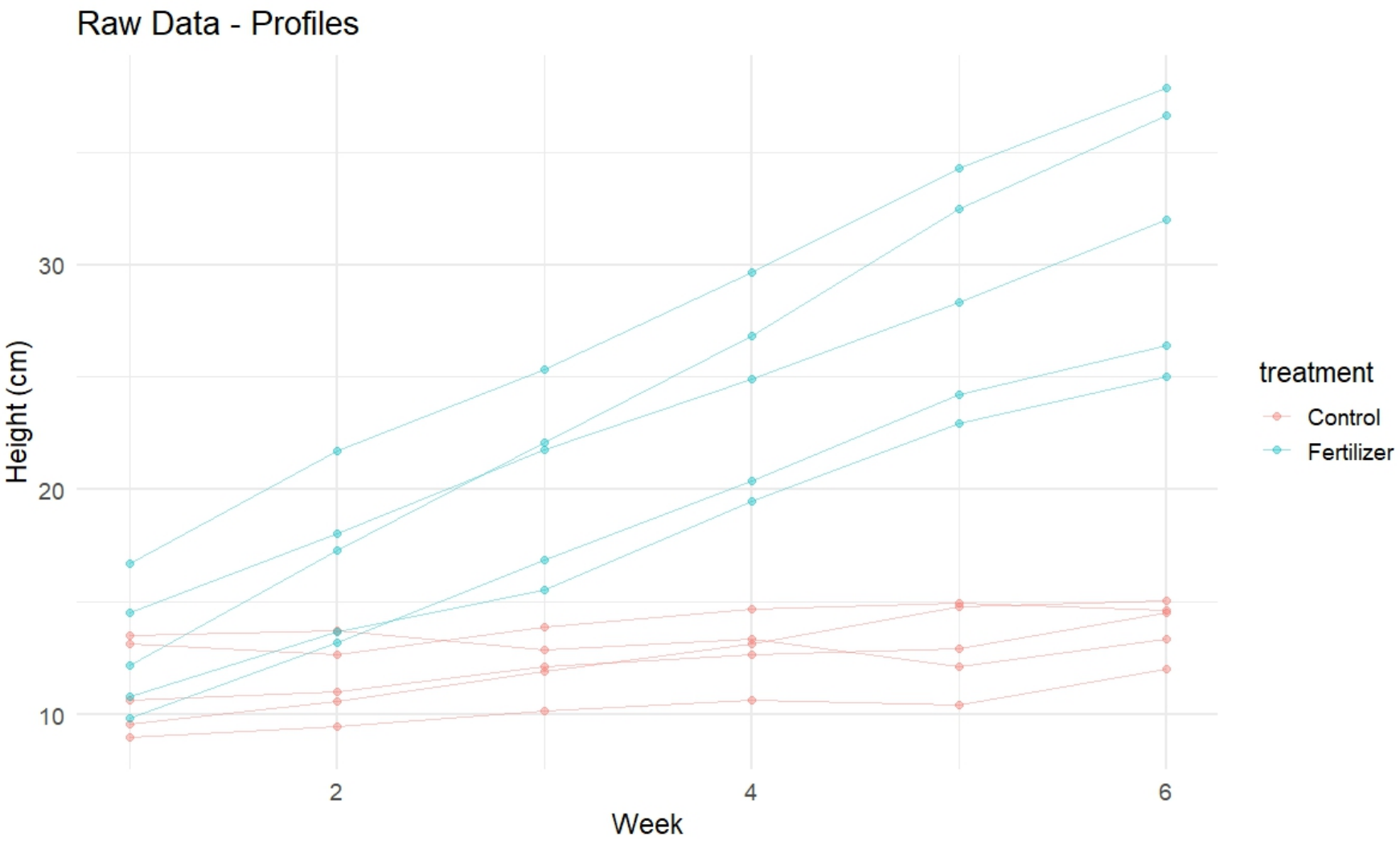
Biological Interpretation

Treatment Effects: Plants in the Fertilizer treatment are expected to grow faster and achieve greater heights than those in the Control group.

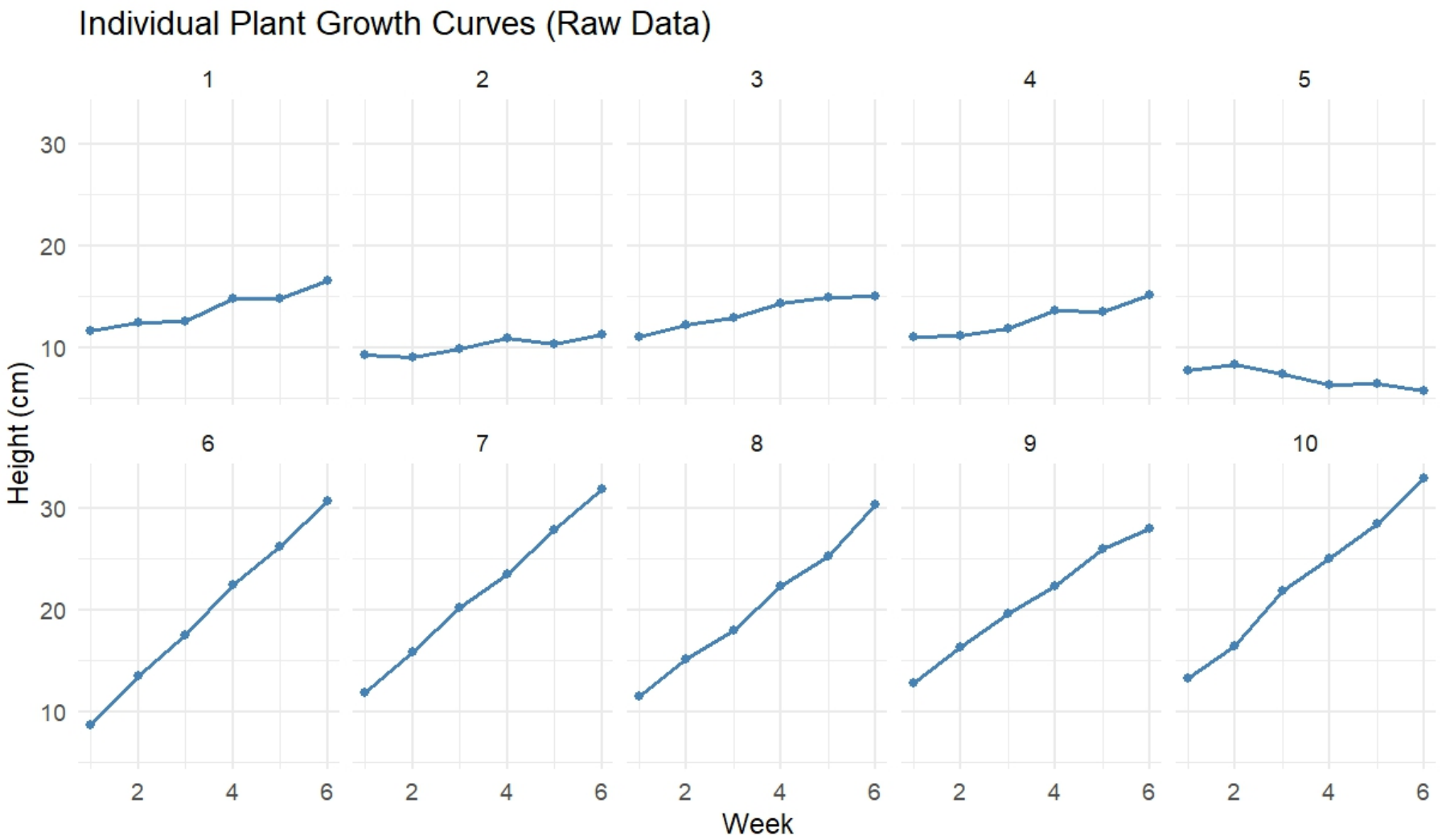
Individual Variability: Even within the same treatment, plants differ in their baseline height and in their growth rates. This variability represents genuine biological differences.



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Random Intercept Model

Consider the model

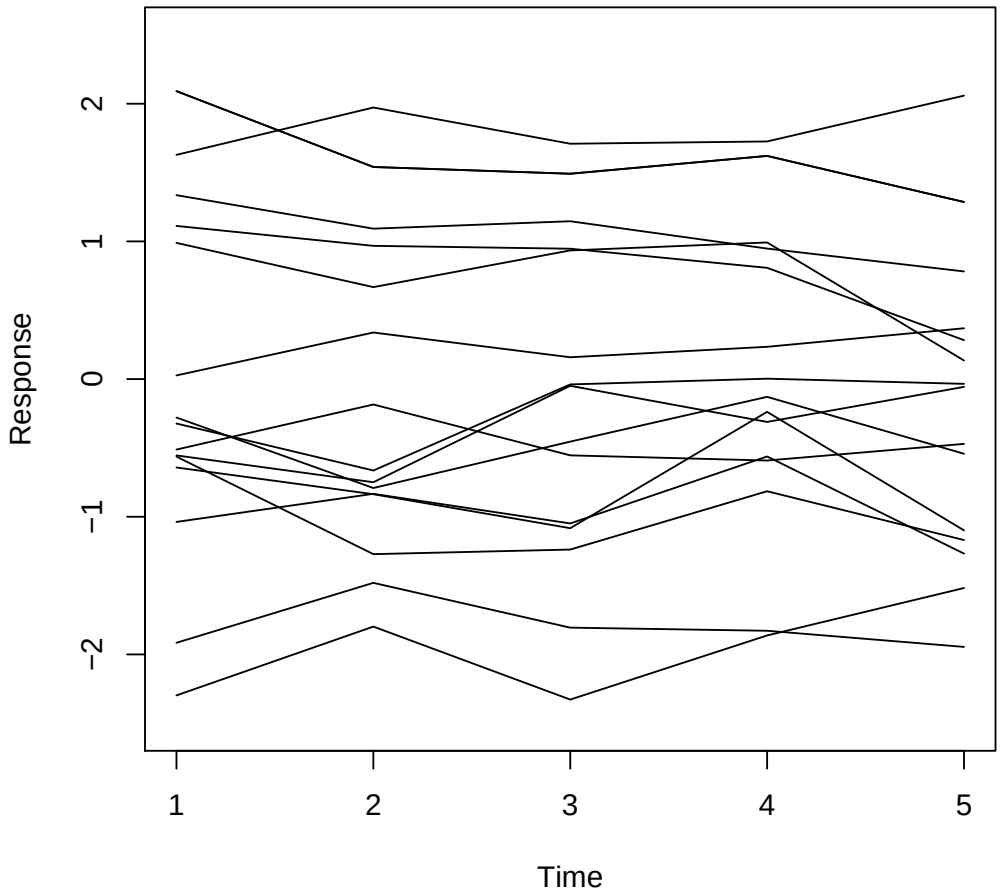
$$Y_{ij} = (\beta_0 + b_i) + \beta_1 treatment_i + \beta_2 time_{ij} + \epsilon_{ij}$$

- Each subject's profile appears flat or parallel (over time)
- Observations Y_{ij} vary around a different value for each subject
- These values are the intercepts of the lines for each subject's responses vary around
- b_i represents the deviations of subject's i intercept from the population one (β_0)
- The set of intercepts are sample from the population of intercepts
- This implies that there is *between-subject variability*

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Example

Consider observations measured in subjects over time.
Such profiles support the assumption of a model with random intercept.



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

- Typically the random term follows a normal distribution

$$b_i \sim N(0, \sigma_b)$$

- As a result, the variance of Y_{ij} takes the form

$$Var(Y_{ij}) = Var(b_i) + Var(e_{ij}) = \sigma_b^2 + \sigma^2$$

- The covariance between any pair of observations of the same subject

$$Cov(Y_{ij}, Y_{ik}) = Cov(b_i, b_i) = \sigma_b^2.$$

- The presence of random effect induce correlation among repeated measurements

$$\rho = Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}.$$

Marginal vs Conditional Model

The random intercept model yields two complementary perspectives.

- The marginal model is of the form

$$E(Y_{ij}) = \beta_0 + \beta_1 treatment_i + \beta_2 time_{ij},$$

and describes the average relationship between predictors and the response across the entire population. It averages the random effects out, so the resulting regression captures the population-level mean response.The coefficients represent population-average effects.

- The conditional model is of the form

$$E(Y_{ij}|b_i) = (\beta_0 + b_i) + \beta_1 treatment_i + \beta_2 time_{ij}$$

and describes the relationship between predictors and the response within a specific individual or experimental unit, given its random effect.

Association and Variability in Longitudinal Data

- The repeated observations from the same subject (plant) are **not independent** — measurements are usually correlated
- The **association** between two measurements tends to be **stronger when the visits are close in time** and weaker as time gap increases

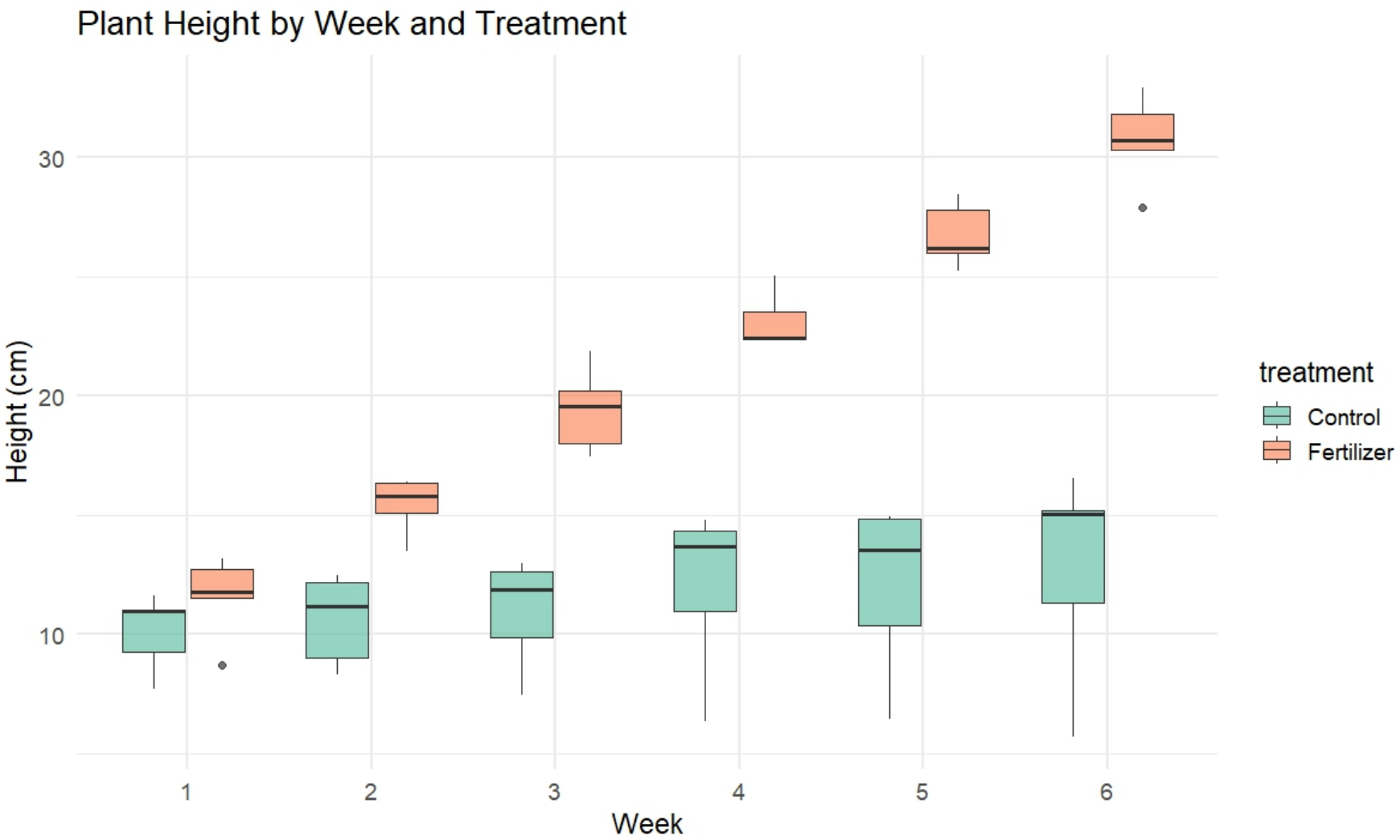
Note:

The random intercept model induces a constant association $\left(\rho = \frac{\sigma_b^2}{\sigma_b^2+\sigma^2}\right)$ between the observations of the same subject at (any) different time points

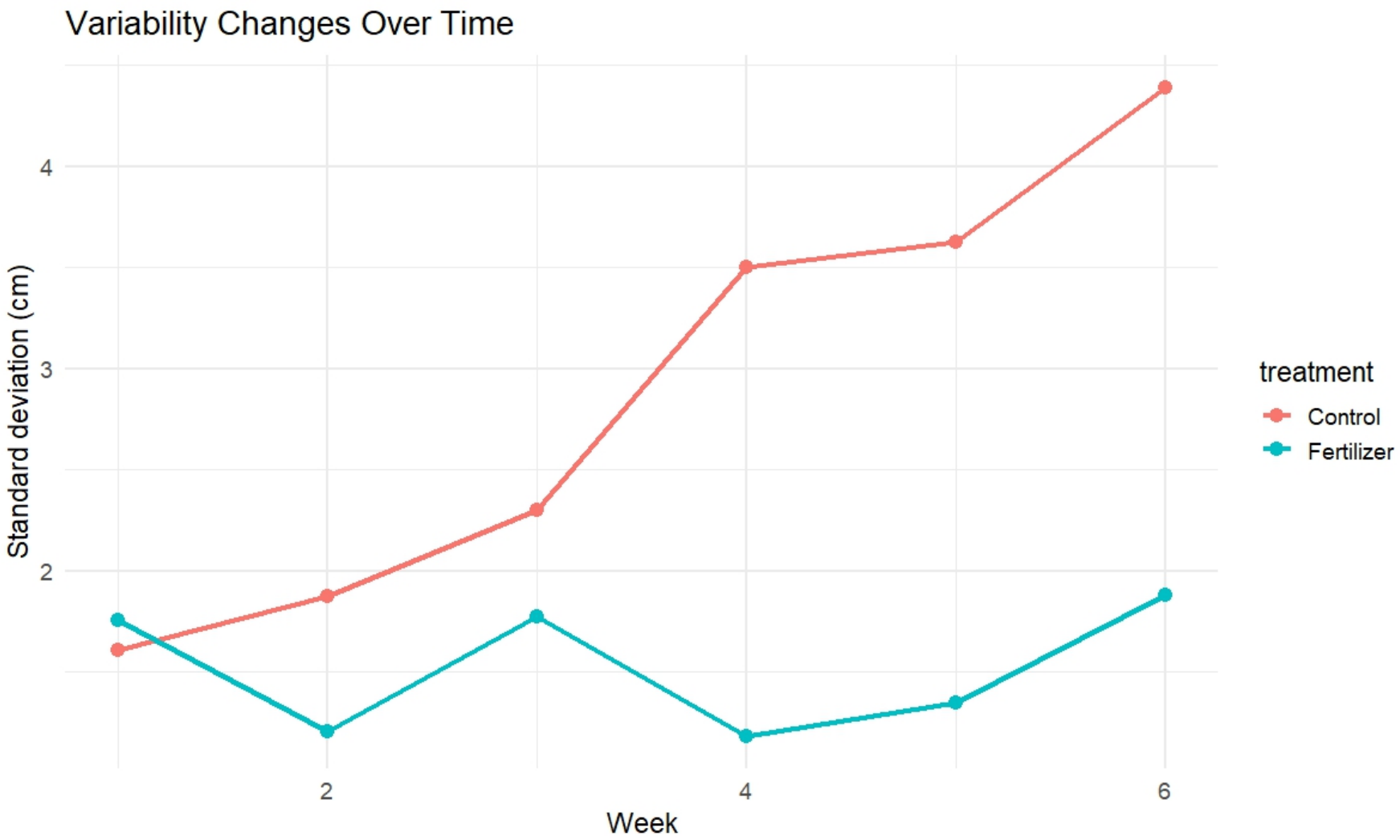
- Over time, **within-subject variability often increases** since individuals tend to diverge as biological processes progress, environmental influences accumulate, or measurement error grows.

Note:

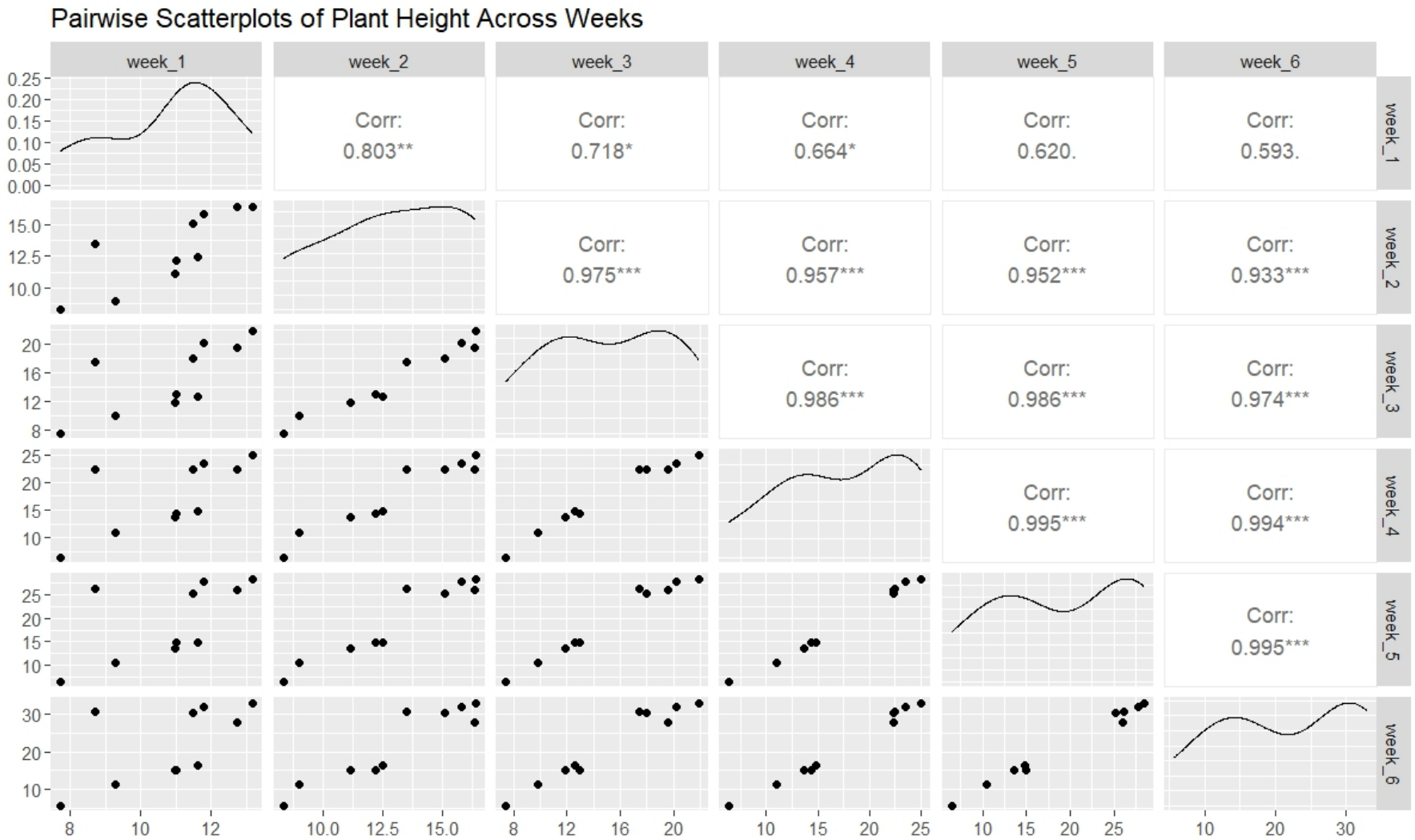
Random intercept model assumes constant variance $\left(\sigma_b^2 + \sigma^2\right)$ over time

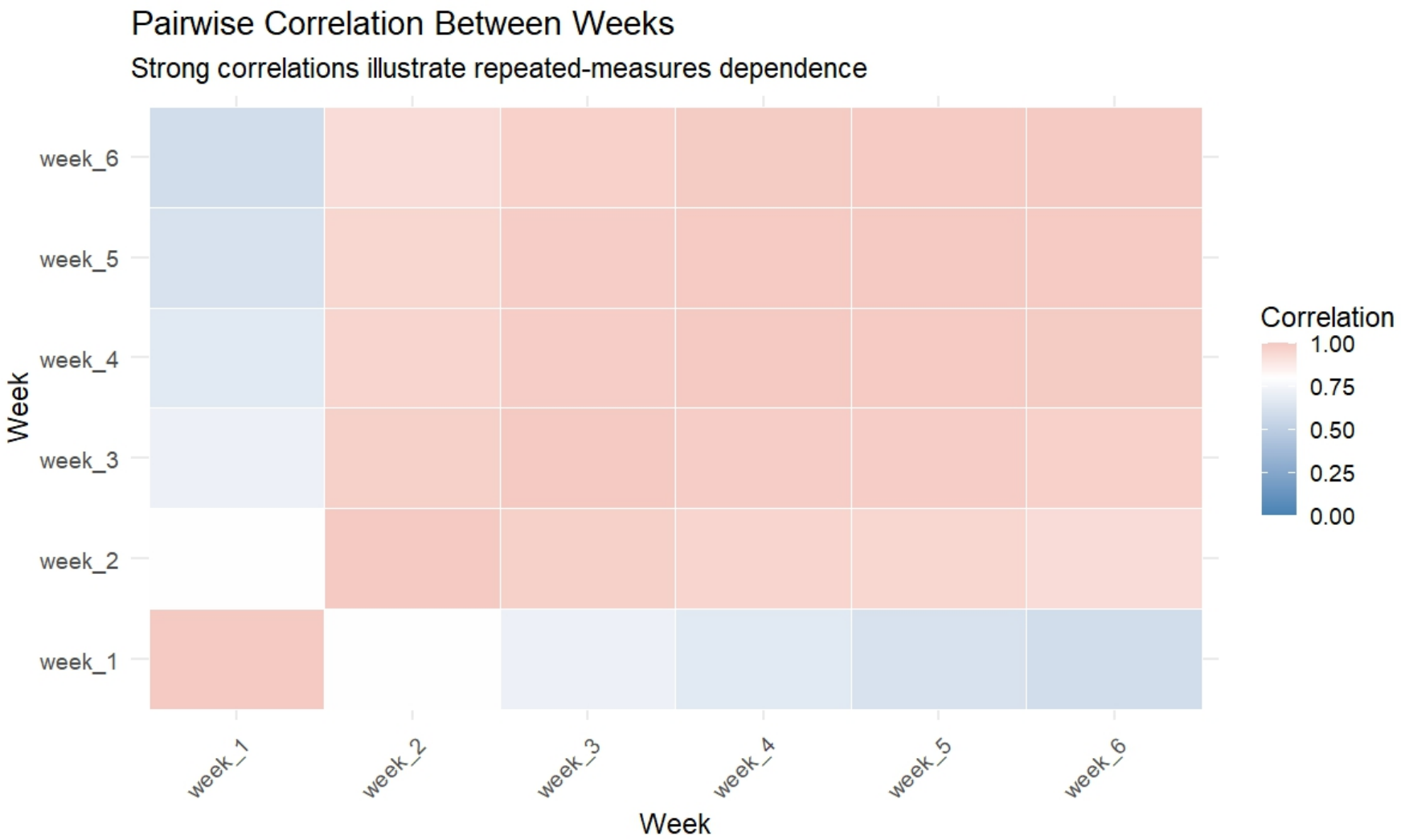


HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837





HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

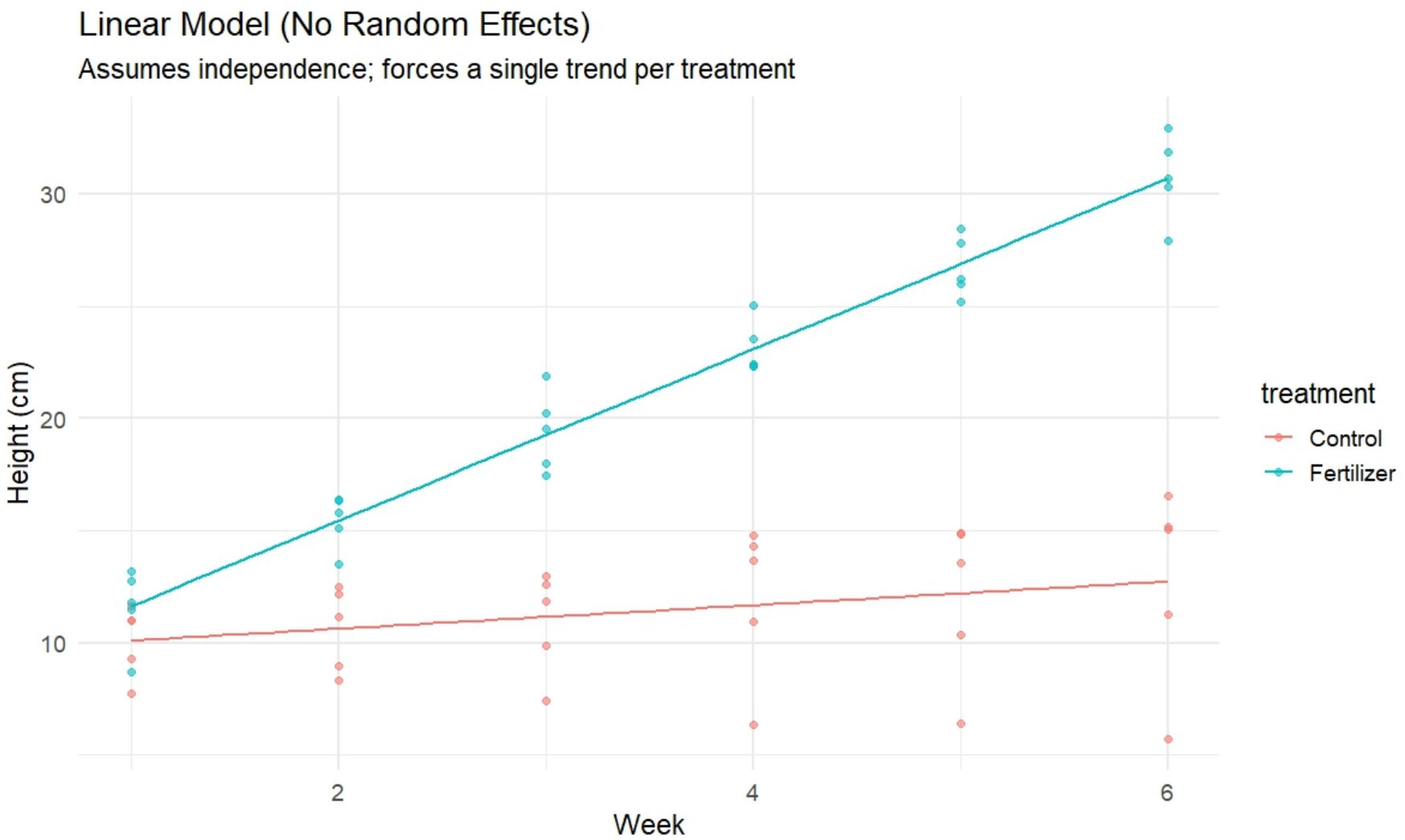
Observed Patterns

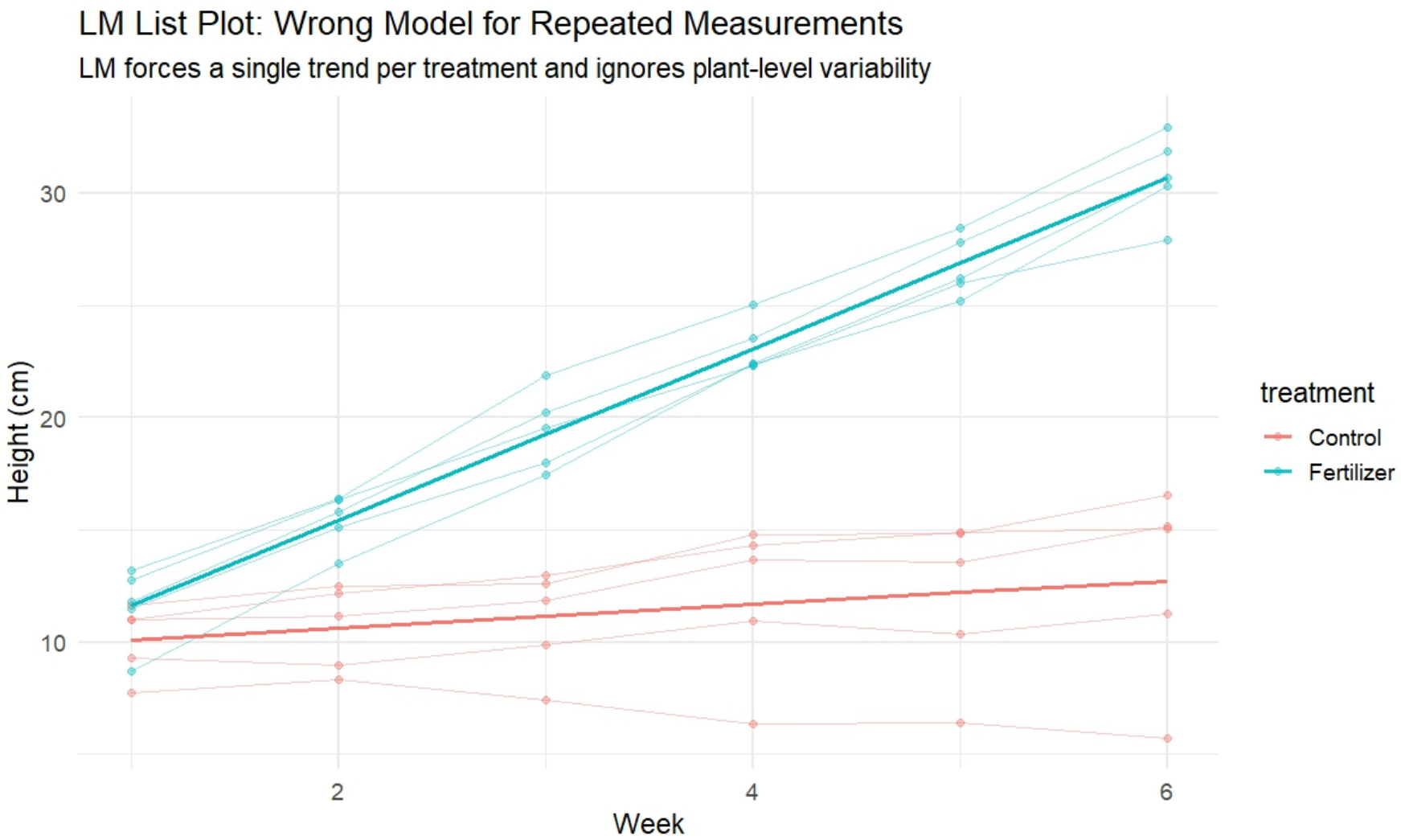
Exploratory analyses show that:

- All plants grow over time.
- Fertilizer plants grow more rapidly.
- Variability among plants increases with time.
- Measurements from adjacent weeks are strongly correlated.
- Correlations decrease as the temporal distance increases.

These characteristics violate the independence assumptions of simple linear models, highlighting the need for mixed-effects approaches.

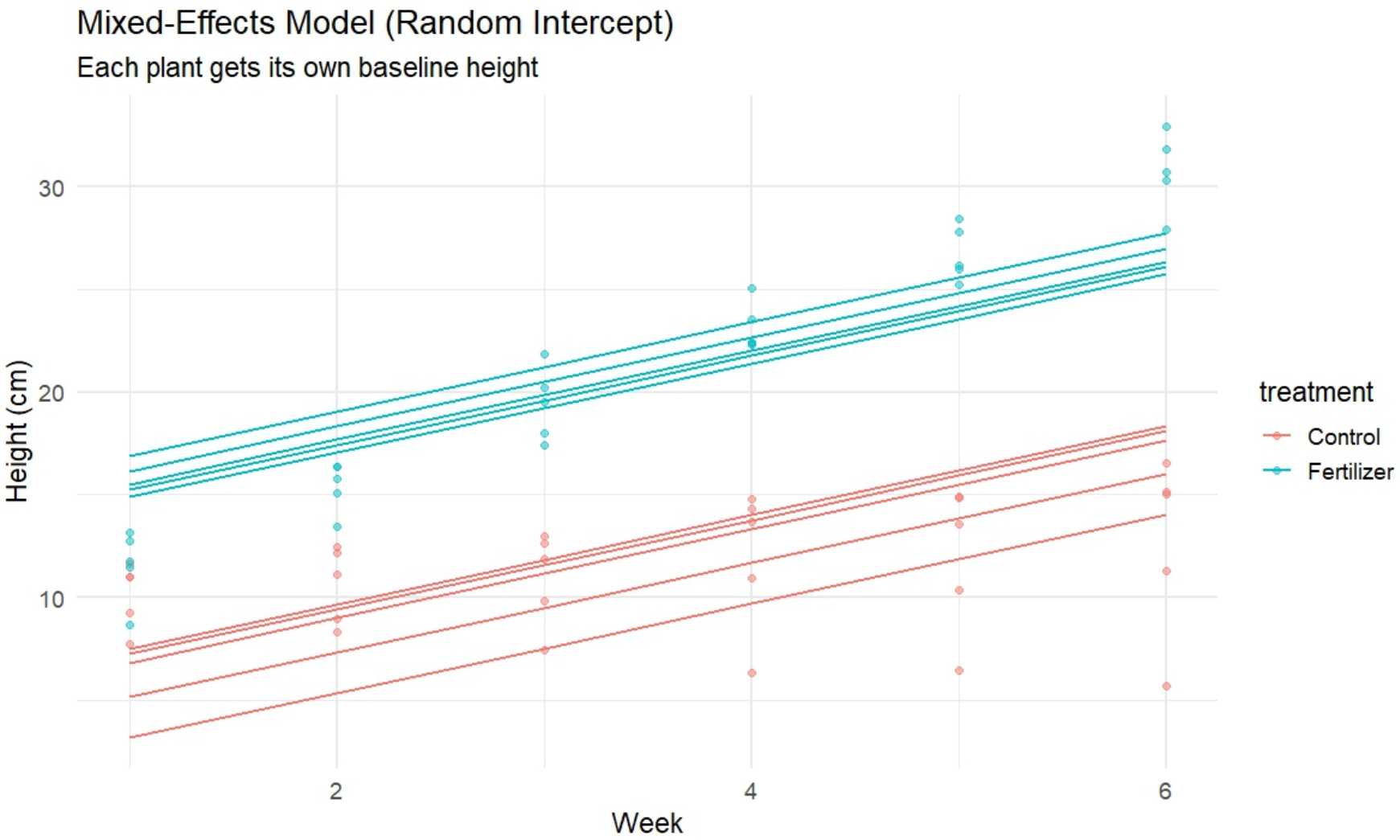
Fit Models


Linear Model (No Random Effects)
Assumes independence; forces a single trend per treatment
Height (cm)
30
20
10
2
4
6
Week
treatment
Control
Fertilizer



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

```
> summary(lmer_int)
Linear mixed model fit by REML ['lmerMod']
Formula: height ~ treatment + week + (1 | plant_id)
   Data: plants

REML criterion at convergence: 108.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2681 -0.5995  0.2535  0.7258  1.6058

Random effects:
 Groups   Name        Variance Std.Dev.
 plant_id (Intercept) 0.0000   0.0000
 Residual             0.3197   0.5654
Number of obs: 60, groups:  plant_id, 10

Fixed effects:
                   Estimate Std. Error t value
(Intercept)         10.77267    0.18175   59.27
treatmentFertilizer  1.90000    0.14598   13.02
week                 0.80971    0.04274   18.95

Correlation of Fixed Effects:
            (Intr) trtmnF
trtmntFrtlz -0.402
week        -0.823  0.000
```



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Random Intercept and Slope Model

Consider the more complicated model

$$Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 + b_{2i}) week_{ij} + e_{ij}.$$

- Each subject varies with respect
  (i) baseline level when $week_{i1} = 1$ and
  (ii) rate of change of response over time
- In this case we have the same fixed and random terms
- The variance is a function of time
  $Var(Y_{ij}) = Var(b_{1i}) + 2t_{ij} Cov(b_{1i}, b_{2i}) + t_{ij}^2 Var(b_{2i}) + Var(e_{ij})$ and the covariance too
  $Cov(Y_{ij}, Y_{ik}) = Var(b_{1i}) + (t_{ij} + t_{ik}) Cov(b_{1i}, b_{2i}) + t_{ij} t_{ik} Var(b_{2i})$

Some Characteristics

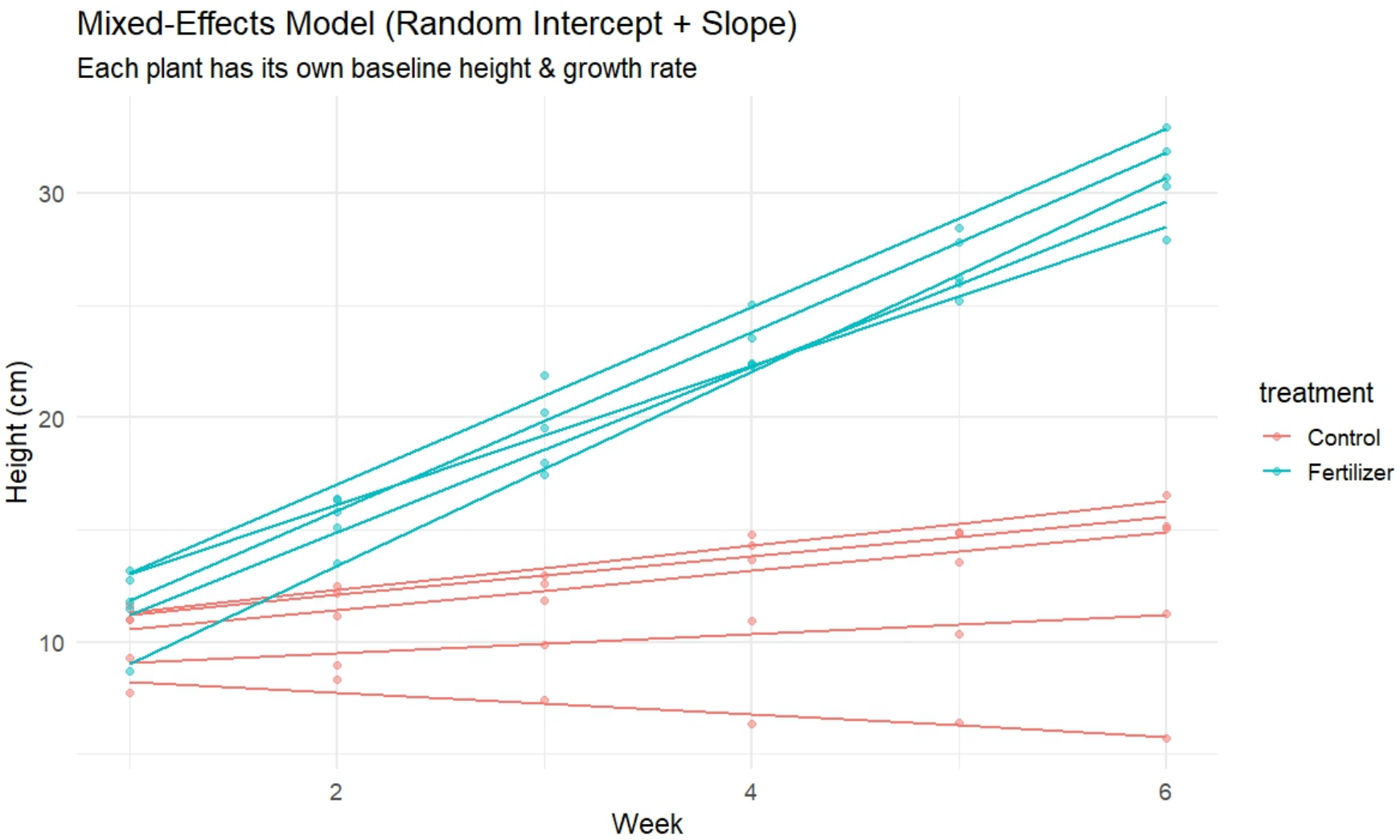
- There is no need of balanced data.
- The covariances are functions of time. As a result, each subject (plant) can have its own sequence of measurement times. This property makes these models suitable for the analysis of *real life* longitudinal data.
- The number of covariance parameters that need to be estimated remains unchanged regardless of the number of measurements.
- The random effects covariance structure allows the variances and covariances to change (increase or decrease) as a function of measurement times, without introducing restrictive structures

Statistical Structure

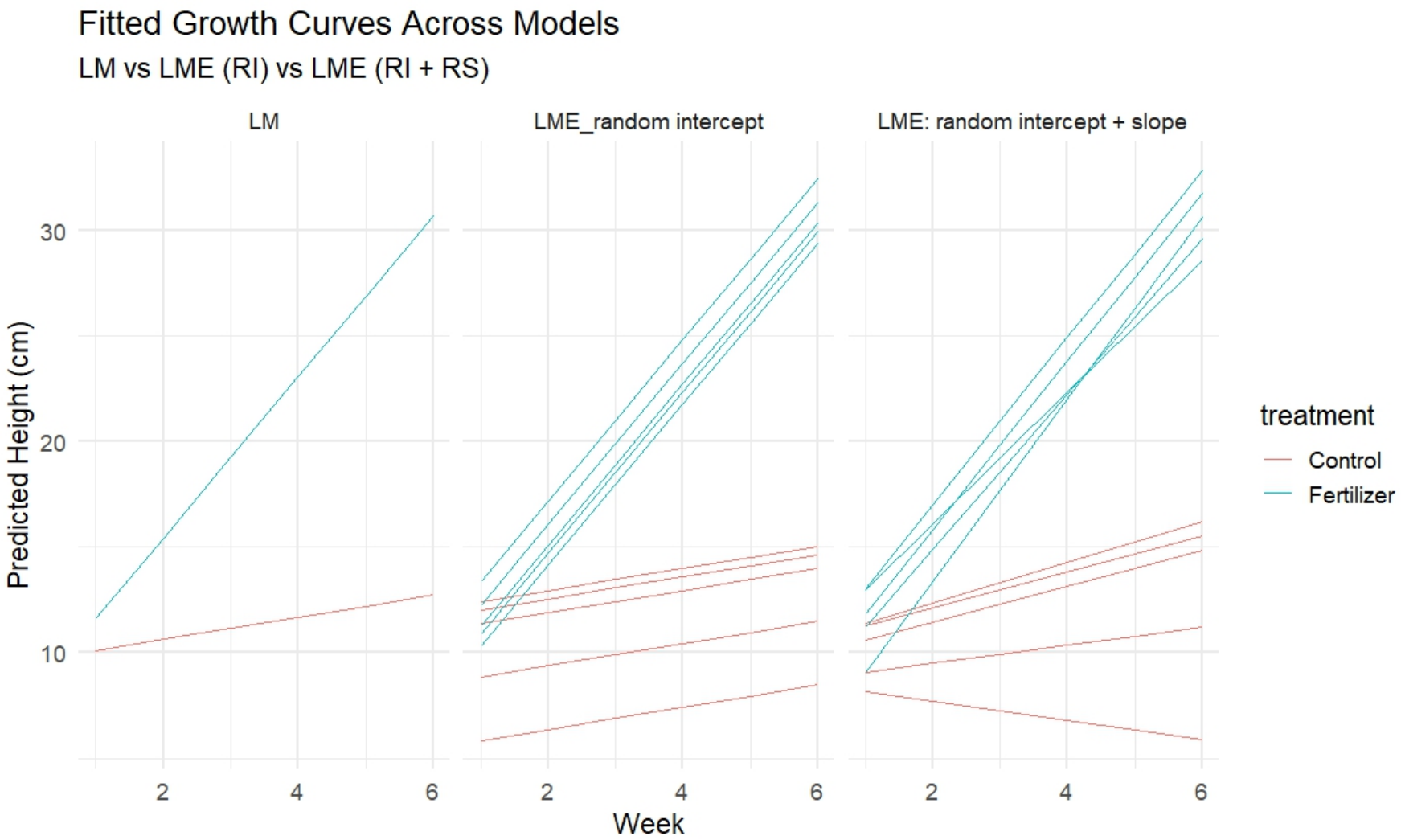
The dataset is ideally suited for linear mixed-effects models because:

- Each plant has **repeated measurements** across time.
- Observations within the same plant are **correlated**.
- Plants differ both in baseline height and in growth rate.
- The structure is hierarchical:
  - repeated measurements (level 1)
  - nested within plant (level 2)

```
> summary(lmer_slope)
Linear mixed model fit by REML ['lmerMod']
Formula: height ~ treatment + week + (week | plant_id)
   Data: plants

REML criterion at convergence: 62.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7807 -0.6116  0.1273  0.5849  1.7126

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 plant_id (Intercept) 0.89154  0.9442
          week        0.07827  0.2798   -0.98
 Residual             0.07434  0.2727
Number of obs: 60, groups:  plant_id, 10

Fixed effects:
                   Estimate Std. Error t value
(Intercept)         10.82173    0.31668  34.173
treatmentFertilizer  1.80187    0.13691  13.161
week                 0.80971    0.09084   8.914

Correlation of Fixed Effects:
            (Intr) trtmnF
trtmntFrtlz -0.216
week        -0.952  0.000
```



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

```
> anova(lmer_int,lmer_slope)
refitting model(s) with ML (instead of REML)
Data: plants
Models:
lmer_int: height ~ treatment * week + (1 | plant_id)
lmer_slope: height ~ treatment * week + (week | plant_id)
           npar    AIC    BIC   logLik -2*log(L)  Chisq Df Pr(>Chisq)
lmer_int      6 213.78 226.35 -100.891    201.78
lmer_slope    8 164.70 181.45  -74.349    148.70 53.083  2  2.972e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Summary

This dataset represents a longitudinal biological experiment in which plant height was measured weekly for six weeks under two treatments. Because each plant has repeated measurements and exhibits unique growth characteristics, the dataset provides an excellent demonstration of both the limitations of simple linear modelling and the advantages of linear mixed-effects models.