Statistical Analysis of Honey Production in an Environmental Framework

Goar Shaboian
5217162

1. Implications of decline of bee population for global ecosystems

In the current ecological situation, it becomes exceedingly urgent to pay closer attention to various factors influencing the global ecosystems. The role the pollinators play is vast: the great majority of crops and wildflowers depend on bees. Research shows that 35% of global agricultural land is affected by pollinators ("Why bees matter" (2018)). The crops that are dependent on pollinators are five times more valuable for human use. What is more, among 115 crops that have the highest world-wide production rate, 75% of them demonstrate greater yields due to animal, mostly bee, pollination (Klein et al. (2007)). Additionally, many species of wildlife would also be at risk if the pollinators were missing from the ecosystem. Unfortunately, as Figure 1 demonstrates, a negative trend for the population of bees was observed since the 1990s (Zattara and Aizen (2021)). This research aims at attempting to discern how various environmental factors may affect the well-being of the bee population.

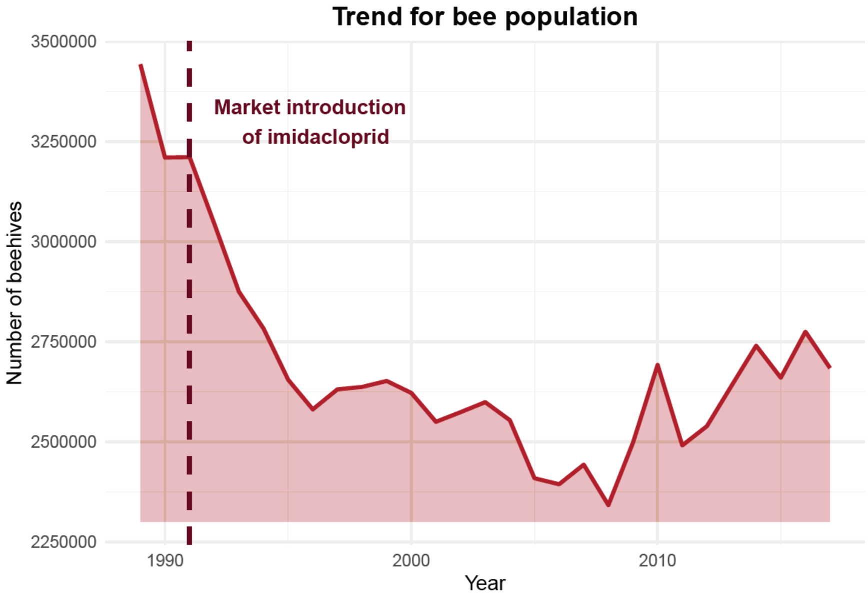

Figure 1: *Number[1] of beehives in the United states, 1989-2021*

1.1. Factors influencing bee population

This analysis focuses on thirty-four continental United States of America states for the year 2016. Total honey production (kilotonnes) is regarded as an indicator for the current state of honeybee population and their ability to pollinate. To conduct meaningful analysis, literature review has been conducted to determine factors that may influence the honeybee population.

[1] Data retrieved from Food and Agriculture Organization of the United Nations.

Table 1: Variable descriptions

Variable	Name	Units	Source
Total honey production	prod	kilotonnes	United States Department of Agriculture
Clothianidin usage	cloth	tonnes	Kaggle
Imidacloprid usage	imid	tonnes	Kaggle
Thiamethoxam usage	thiam	tonnes	Kaggle
Acetamiprid usage	ace	tonnes	Kaggle
Colonies affected by disease	disease	% of colonies	United States Department of Agriculture
PM2.5 annual average concentration	pm	mcg/m3	Centre for Disease Control and Prevention
Energy-related carbon dioxide emissions	co2	MT	Energy Information Administration of the United States of America

The variables included in the analysis are demonstrated in Table 1[2]. From Figure 1, it can be observed that a rapid decline in bee population coincides with the introduction of a neonicotinoid pesticide imidaploprid to the market in the United States (Jeschke and Nauen (2008)). Additionally, studies show that neonicotinoids application has acute lethal toxicity for bees when there is contact exposure, as well as devastating effects on bee reproduction (Blacquiere et al. (2012)). Hence, predictors for usage of four neonicotinoid pesticides were introduced to the model. For each pesticide, the scope of usage varies: acetamiprid is usually used for broadleaf plants and lawns, thiamethoxam for corn, bean, cotton, and turf, imidacloprid is often used for landscapes, pet pests and gardening, and clotianidin has a wide agricultural scope for usage.

Pollination is severely affected by air pollution since it is the reason that bees sometimes fail to recognize and locate the flora that they aim to pollinate (Capitani et al. (2021)). Additionally, deposition of respirable particular matter $PM_{2.5}$ and $PM_{5.0}$ has been found on bees' waxy layers in areas with higher air pollution rates (Thimmegowda et al. (2020)). Thus, two variables representing air pollution rates were added: concentration of particular matter 2.5 particles (micrograms per cubic meter) and Carbon Dioxide emissions values (metric tons).

Variable representing disease affecting the colonies was added to attempt to account for effects not represented by other variables included in the analysis.

1.2. Discovering relations between variables

To obtain a general idea on how the covariates are related with each other, as well as with the response, preliminary analysis of the variables is necessary, namely data visualisation.

[2] Total honey production was rescaled from pounds to kilotonnes; pesticide variables were rescaled from kilogrammes to tonnes.

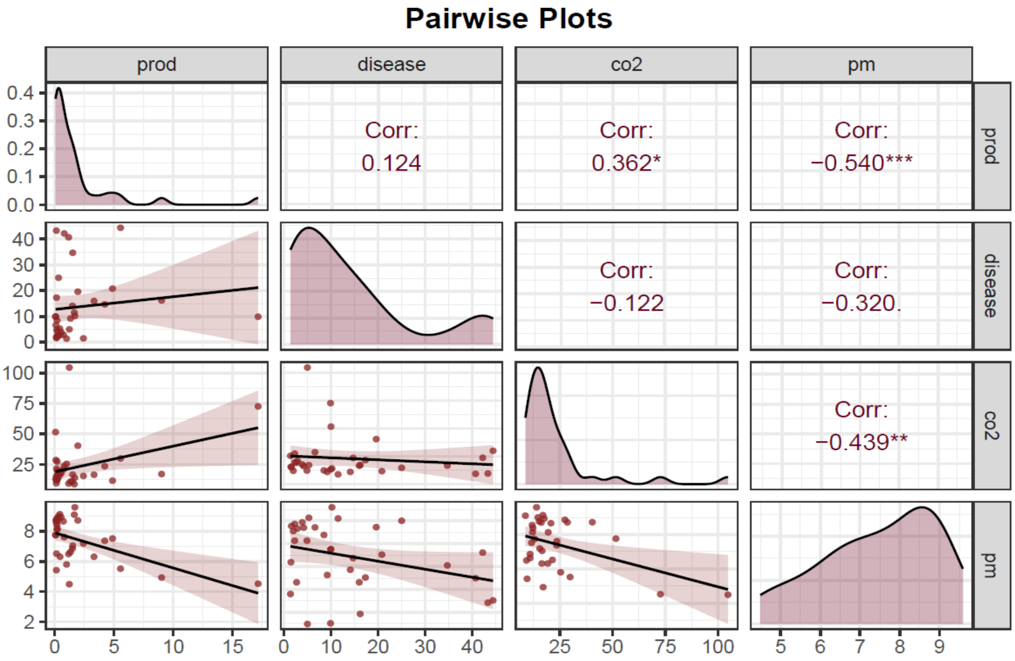

Figure 2: Pair plots for covariates

Figure 2 demonstrates that the values for the response variable are concentrated in the left-hand side of the support, with a few extreme values. Values for PM2.5 concentration are slightly left-skewed, but the degree of skewness relatively low. Additionally, a negative linear relationship is observed between PM2.5 and the response, whereas the relationship between Carbon dioxide emissions and the response appears to be positive.

To investigate the variables for pesticide usage, ridgeline plots were used to study the distribution of the data.

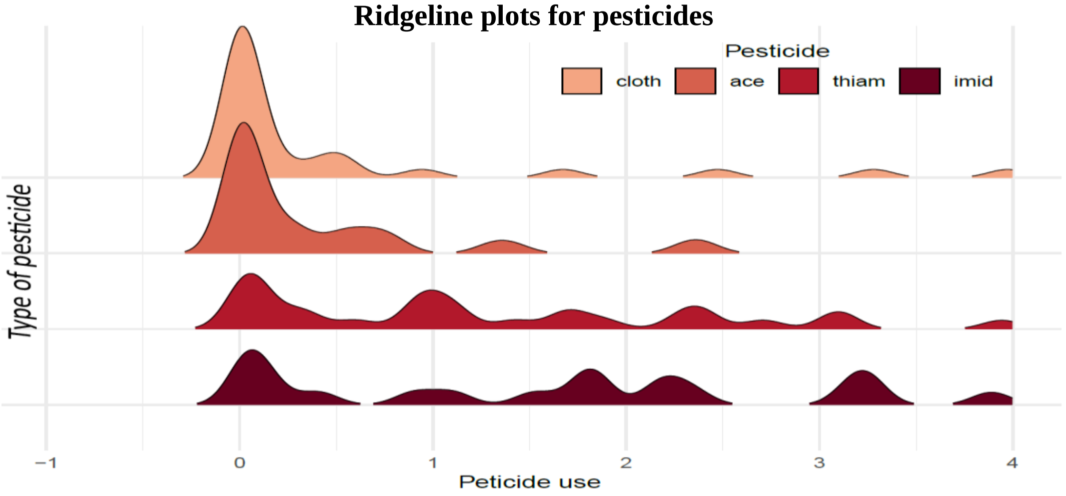

Figure 3: Distribution of pesticide variables

From Figure 3, it is evident that the distributions of clothianidin and acetamiprid are severely right-skewed, with very few observations for high usage. However, for pesticides thiamethoxam and imidacloprid both high and low values of usage are recorded for different states, thus high skewness is not present.

2. Exploring possible models

For conducting the analysis, a linear regression model is assumed, since it provides methodology to study the relationship between the honey production and the environmental variables. Additionally, it allows to make predictions on newly observed values for the covariates. The decision has been made to use the air pollution variables as categorical variables: this will allow to distinguish the effect on the response for states with low and high levels of air pollution indicated by each of them separately. Each variable has been split into two categories, “High” and “Low”, according to the mean of their distributions, with “High” being the reference level for both.

```
data$cat_co2 <- factor (ifelse (data$co2 > mean (data$co2), "High", "Low"))
data$cat_pm <- factor (ifelse (data$pm > mean (data$pm), "High", "Low"))
```

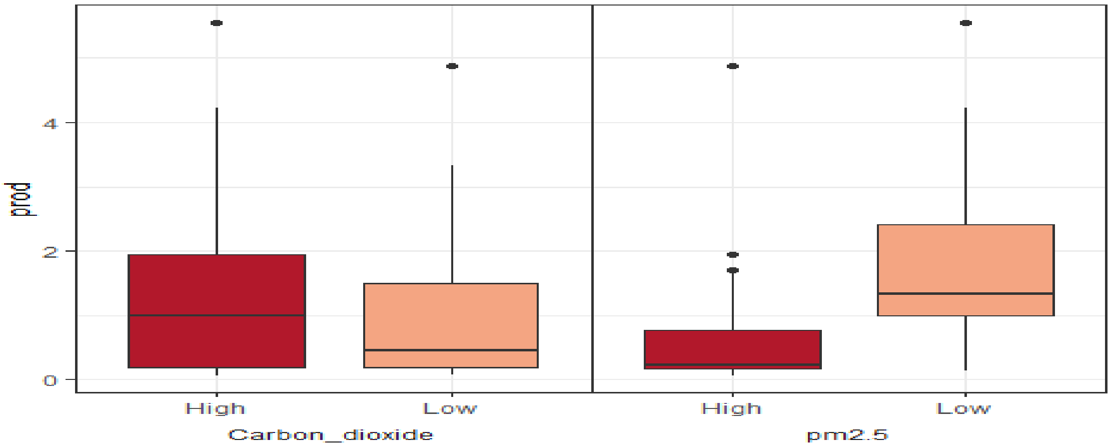

Figure 4: Boxplots for levels of air pollution variables [3]

Figure 4 demonstrates that for the variable for CO^2 emission the differences in total honey production for different levels of the factor are not significant and are counter-intuitive. However, for $PM_{2.5}$ a significant difference can be observed: values of total honey production are noticeably higher for states where the concentration on particles in the air is lower. In Figure 2, a relationship has been detected between disease in colonies and $PM_{2.5}$ concentrations, thus an interaction between them was included in the model to further investigate the joint effects of the covariates onto the response variable (Weisberg (2005), p. 54). Hence, the interaction will demonstrate if there are any joint effects between the levels of the $PM_{2.5}$ concentration in a certain state and how it can interact with the proportion of the diseased colonies, and influence jointly the response.

Adding redundant regressors to the model may add noise and introduce problem of collinearity, hence it is essential to perform variable selection (Faraway (2016), p. 130). Best subset selection, a common approach, iterates over all possible models that can be fit for each possible number of regressors, and chooses the optimal model for each p according to the lowest RSS value or highest R^2. Afterwards, among them, the optimal model is chosen using information criteria, Mallows’ C_p or R^2_{adj}.

[3] Two of the most extreme values for total honey production were removed for this plot, for better visual representation.

In this framework, the computation burden is not high due to a small number of regressors. Function *regsubsets ()* from *leaps* package was implemented. It should be underlined that the interaction term was first included in the model at step 7, before the variable for disease was added. Hence, the step 7 model was eliminated from the output since it violated the hierarchical principle (James et al. (2013), p. 89). The output for the best subset selection is shown below.

```
bss <- regsubsets (prod ~ . + disease*cat_pm, data = data)
bss_sum <- summary (bss); calls <- as.matrix(bss_sum$which [-7, ])
colnames (calls) [c (7, 8, 9)] <- c ("cat_co2", "cat_pm", "disease * cat_pm")
models <- list (); formulas <- list ()
for (i in 1:7){
  formula <- as.formula (paste ("prod ~ ", paste (names (calls [i,-1]) [calls [i,-1] == TRUE],
collapse = "+")))
  formulas [[i]] <- formula
  models [[i]] <- lm (formula, data = data)}

bic <- numeric (7); aic <- numeric (7); r_adj <- numeric (7); cp_mallows <- numeric (7)
for (i in 1:7){
  model_out <- models [[i]]; sum_out <- summary (model_out)
  aic [i] <- AIC (model_out)
  bic [i] <- BIC (model_out)
  cp_mallows [i] <- ols_mallows_cp (model_out, models [[7]])
  r_adj [i] <- sum_out$adj.r.squared}
bic.null <- BIC (lm (prod ~ 1, data = data)); aic.null <- AIC (lm (prod ~ 1, data = data))
drop_in_BIC <- sapply (bic, function (t) t - bic.null)
drop_in_AIC <- sapply (aic, function (t) t - bic.null)
```

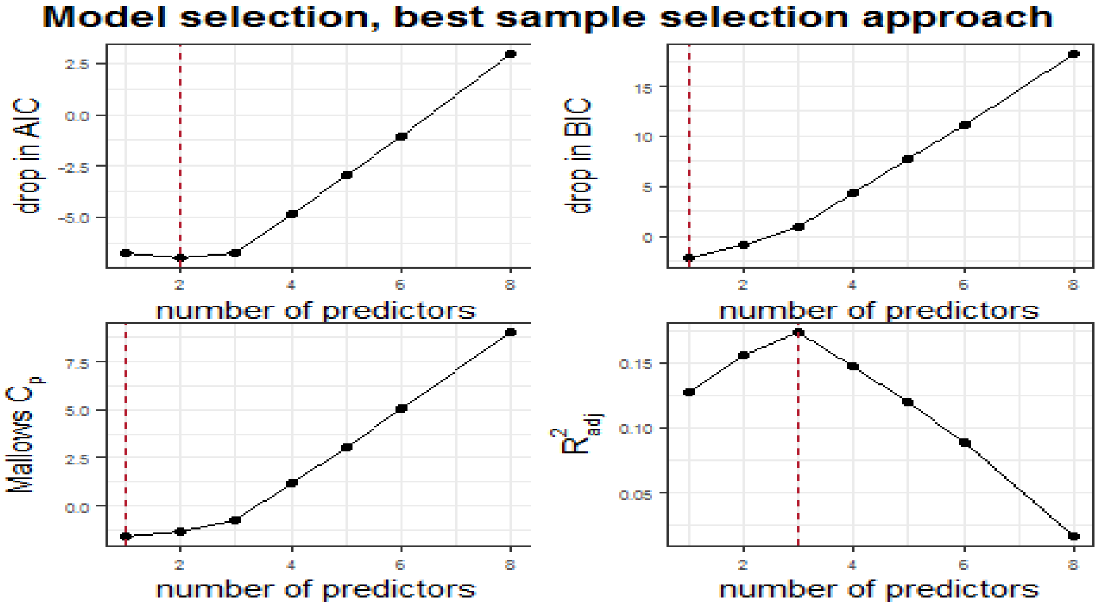

Figure 5: Comparison of models with different number of predictors

Figure 5 shows there is no consensus for the criteria used for selecting the optimal model: according to Bayesian Information criterion and Mallows' C_p, the optimal model should include one parameter, *$PM_{2.5}$ concentrations*. This is because BIC is a very conservative criterion, since it penalizes the model more for higher number of parameters included (Faraway (2016)). Akaike criterion indicates that the most optimal model includes intercept and two regressors: *$PM_{2.5}$ and CO^2 emissions.* The adjusted coefficient of determination, however, indicates that the model with including three covariates is the optimal strategy: *$PM_{2.5}$, CO^2 emissions and imidacloprid*.

To provide another algorithm for model comparison, cross-validation was implemented. Since the number of observations is low *(n = 34)*, it is possible to implement the most exhaustive algorithm: Leave-One-Out Cross-Validation. This implies training each model 34 times by omitting one observation each time, then predicting the value of that particular observation using the trained model, and calculating the Mean Square Error at each iteration, the average of which provides the estimate for cross-validation error. This approach does not incorporate randomness in the procedure and guarantees that there is no bias: it does not overestimate the test error (James et al. (2013)).

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i\,, \qquad where\ MSE_i = (y_i - \hat{y}_i)^2$$

```
CV <- NULL
for (i in 1:7){
  glm.fit <- glm (formulas [[i]], data = data)
  CV[i] <- cv.glm (data, glm.fit)$delta [1]}
```

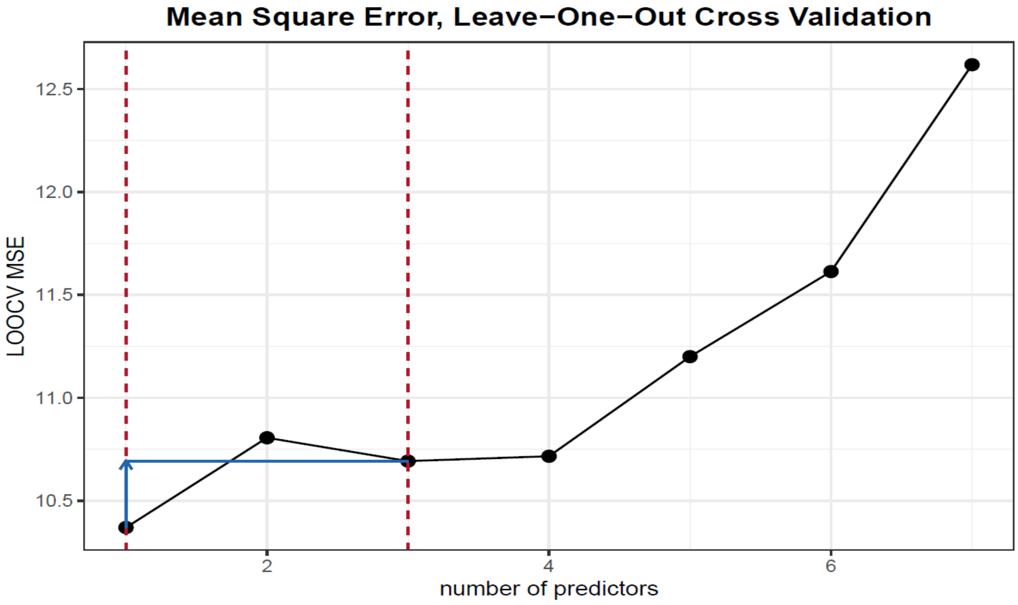

Figure 6: Leave-One-Out Cross-Validation results

LOOCV results demonstrate that model with 1 regressor has the lowest MSE value. However, it can be observed that the difference with the error for the model with 3 variables is relatively small. For the purpose of maximizing the proportion of explained variance, and for obtaining predictions based on as much information as possible with statistical justification, the model suggested by R^2_{adj} with 3 regressors *($PM_{2.5}$ concentrations, CO^2 emissions and imidacloprid pesticide)* was chosen. It should be underlined that imidacloprid was the pesticide for which high values of usage were recorded for more states, relative to the other neonicotinoid pesticides. It is also the most commonly used one (Jeschke and Nauen (2008)).

```
fit <- models [[3]]
```

3. Collinearity issues

Collinearity is a problem that arises when the predictors have high levels of correlation between each other: in such case, the predictors explain similar share of the variance of the response, inflating the variance of the coefficients, thus lowering the precision of the regression coefficients. To make preliminary conclusions on the presence of collinearity, the table of Pearson correlations is analysed.

$$\rho = \frac{cov\ (X, Y)}{sd\ (X)\ sd\ (Y)}$$

Although two of the variables chosen for the model are categorical, it is possible, for thoroughness of the analysis and to make preliminary conclusions, to investigate the Pearson's correlation values *(function cor ())* between the original variables to determine if there are any linear relationship between them.

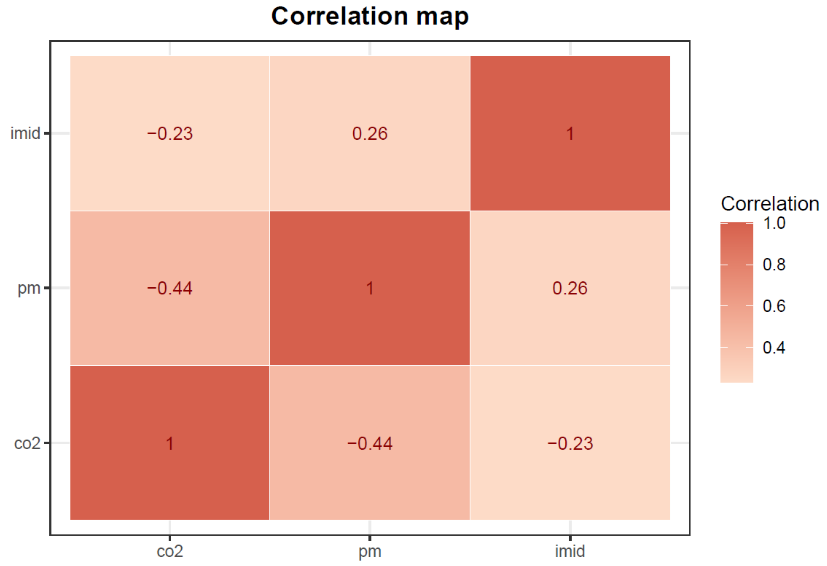

Figure 7: Heatmap for correlation matrix

Form Figure 7, it can be observed that the highest value for correlation between covariates is −0.44, indicating moderate negative correlation, which generally does not cause serious collinearity problems (Dormann et al. (2013)). Additional and more precise measures of correlation are Variance Inflation Factors: they take into consideration not only the relationship between two variables, but also the relationships between all covariates. The function *vif ()* from the *car* package provides an algorithm that is effective in calculating the variance inflation factors for models with categorical variables and is not sensitive to coding of the categorical variables (Fox and Weisberg (2019)).

Table 2: Variance Inflation Factors

Variable	imid	co2	pm
VIF	1.095	1.05	1.05

Table 2 confirms there are no multicollinearity issues to tackle in this framework. However, if it would have been detected, the following remedies may have been applied:

- Removing one of the variables that inflates the variance from the regression analysis. Much information would not have been lost since the variable correlated with it explains approximately the same amount of the response variability.

- Combining the variables, accounting for differences of measurement scales, and including the new variable into the model specification. This method is preferable since it ensures that no information is lost.

4. Model diagnostics

Making inference for the multiple linear regression model obtained using the ordinary least squares method relies on complying with several important assumptions, which are investigated using model diagnostics. Residual-based diagnostics allow to make conclusions on the errors, due to the assumption that the behavior of residuals and errors is somewhat similar, and they are related through the hat matrix (Weisberg (2005), p. 206). Table 3 provides description of the diagnostics performed.

Table 3: Diagnostics for linear regression

	Assumption	Detection	Consequences	Remedies
errors	Constant variance	Fitted-residual plot $E[\underline{e}\vert\underline{U}]$	➢ Not reliable inference ➢ Inefficient estimation	➢ Variance-stabilising transformation ➢ Weighted Least Squares
	No correlation	Durbin-Watson test $cov\left[\epsilon_i;\epsilon_j\right] = 0$	➢ Confidence intervals narrower; p-values lower	➢ Generalised Least Squares
	Normality	QQ-plot Shapiro-Wilk test $\epsilon \sim N(0, \sigma^2)$	➢ OLS estimates unbiased, but not optimal ➢ Not reliable inference	➢ For long-tailed distributions, use resampling methods
model	Linearity	Fitted-residual plot $E[\underline{Y}\vert\underline{X}] = \underline{X}\,\underline{\beta}$	➢ Suspicious conclusions ➢ Reduced predictive accuracy	➢ Apply covariate transformation (for example, power transformations)
observations	Outliers	Studentised residuals $r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}$	➢ May affect the residual standard error	➢ Careful treatment for each separate case
	High-leverage points	Leverage $h_{ij} = \underline{x}_i\left(\underline{X}^T\underline{X}\right)^{-1}\underline{x_j}$	➢ May affect the fit	
	Influential points	Cook's distance $D_j = \frac{(\underline{\hat{y}}_{-j} - \underline{\hat{y}})^T(\underline{\hat{y}}_{-j} - \underline{\hat{y}})}{(p+1)\hat{\sigma}^2}$	➢ Affect the fit of the model	

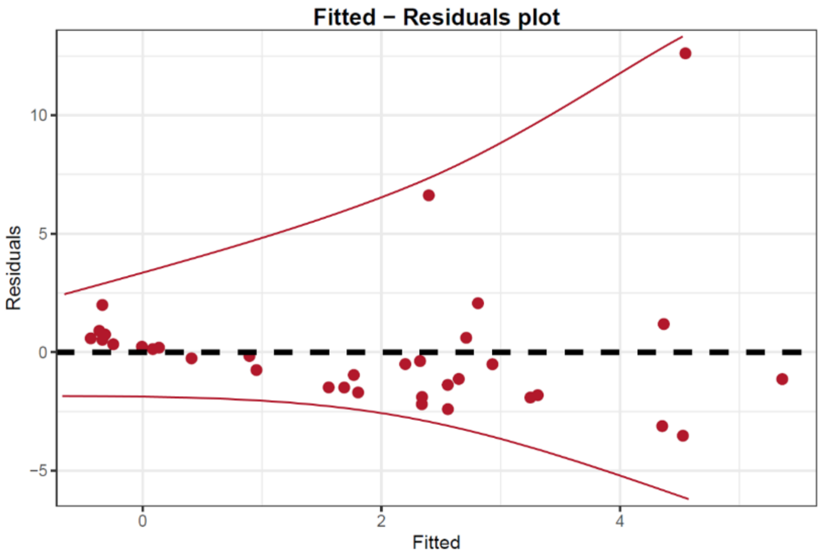

Figure 8: Fitted and residual values for the optimal model

First step implemented was investigating the fitted-residual plot, for it allows to make conclusions simultaneously on two assumptions: constant variance of errors and linearity of the model. From Figure 8, a funnel shape can be observed in the scatterplot, which implies some degree of non-constant variance in the errors. To remedy this problem, a variance-stabilising transformation *(log transformation)* was implemented. Further diagnostics tests were performed on the model with the transformed response.

```r
data_vst <- data.frame (cbind (prod = log (data_model$prod), data_model [, 2:4]))
fit <- lm (prod ~ ., data = data_vst)
sw <- shapiro.test (residuals (fit)) # Shapiro-Wilk test
dw <- durbinWatsonTest (fit) # Durbin-Watson test
# HLP:
hat <- influence (fit)$hat; thr <- 2*4/n # 0.24 threshold
hn_quantiles <- sapply (seq(1/(n+1), n/(n+1), length.out = n), function (t) qhalfnorm(t, theta=
sqrt(pi/2), lower.tail = TRUE, log.p = FALSE)) # theoretical quantiles
# Outliers:
rstad <- rstandard (fit) # studentised resids
alpha <- 0.05; b_threshold <- qt (1-alpha/(2*n), df = fit$df.residual) # Bonferroni correction
cook <- cooks.distance (fit) # Cook's distance
```

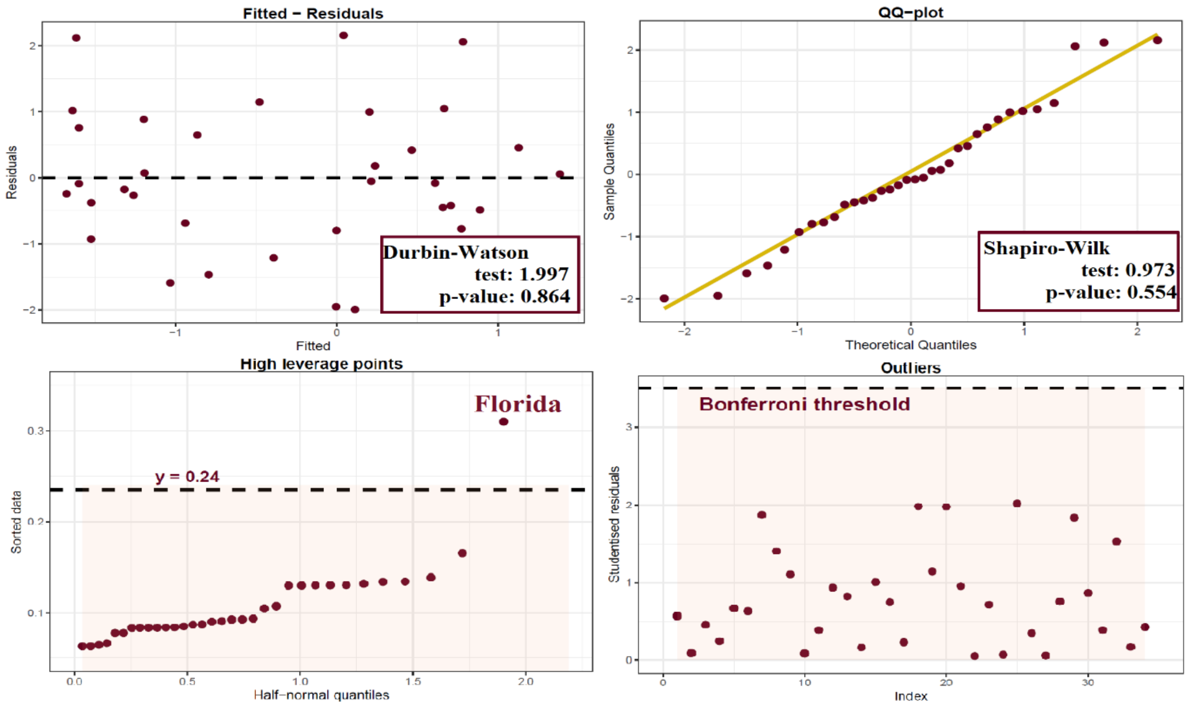

Figure 9: Plots for regression diagnostics

From Figure 9, several conclusions can be drawn. Firstly, the trend observed in the residuals is eliminated, indicating constant variance in errors. There is no violation of the linearity assumption. Normality assumption is confirmed both by visual analysis and by the test statistic. Durbin_Watson[4] outcome suggests that there is no statistical evidence against the null hypothesis of the absence of correlation. No outliers are detected; however, the state of Florida appears to be a high leverage observation. To test whether it affects the fit in a significant way, the Cook's distance was computed.

[4] For testing for error correlation, methodology suggested by Fox (2015), namely using the Durbin-Watson test, was implemented.

Table 4: Cook's distance: descriptive statistics

Min	1 Quant	Median	Mean	3 Quant	Max
0	0.002	0.015	0.031	0.045	0.146

From Table 4, it can be observed that for none of the observations, including Florida, Cook's distance exceeds the rule-of-thumb value of one. Additionally, the maximum value does not appear extreme with respect to the rest of the distribution, hence it can be concluded that the high-leverage point observed above does not influence the fit significantly and thus, to avoid losing information, its exclusion from the model may be avoided.

5. Model interpretation and inference

The obtained linear regression model contains the optimal number of variables and it adheres to the assumptions of the linear regression model. Hence, the estimates it provides are reliable, and inference can be made. The aims for the analysis are discovering the nature of the relationship between covariates and the response, explaining the variability of the response and making predictions, with less focus placed on quantifying the effects of regressors.

5.1. Regression estimates

The results of the *lm ()* fit are presented below.

Table 5: Summary of linear regression fit

Variable	Estimate	se	t-stat
intercept	-1.035	0.417	-2.482 *(0.019)*
imid	0.054	0.017	3.071 *(0.004)*
cat_co2	-0.663	0.431	-1.539 *(0.134)*
cat_pm	1.691	0.395	4.277 *(0.000)*

RSS	R2	F-statistic
1.117	0.453	8.27 *(0.0003)*

Table 5 provides coefficient estimates, their uncertainty and interpretation, measures of goodness of fit for the model, as well as test statistic values that will be used for inference further in the analysis. It should be stressed that the estimate for imidacloprid does not correspond to the initial hypothesis: the estimation suggests that increasing pesticide usage increases honey production. For CO^2 emissions, the result violates the initial hypothesis as well. To illustrate the uncertainty associated with the estimators, effects plots are used. They allow to observe the estimate of the effect of the covariate on the response holding other variables constant, together with the uncertainty associated with the estimates. For categorical variables, they demonstrate how the levels for them influence the response, with other variables held constant, also with the degree of uncertainty.

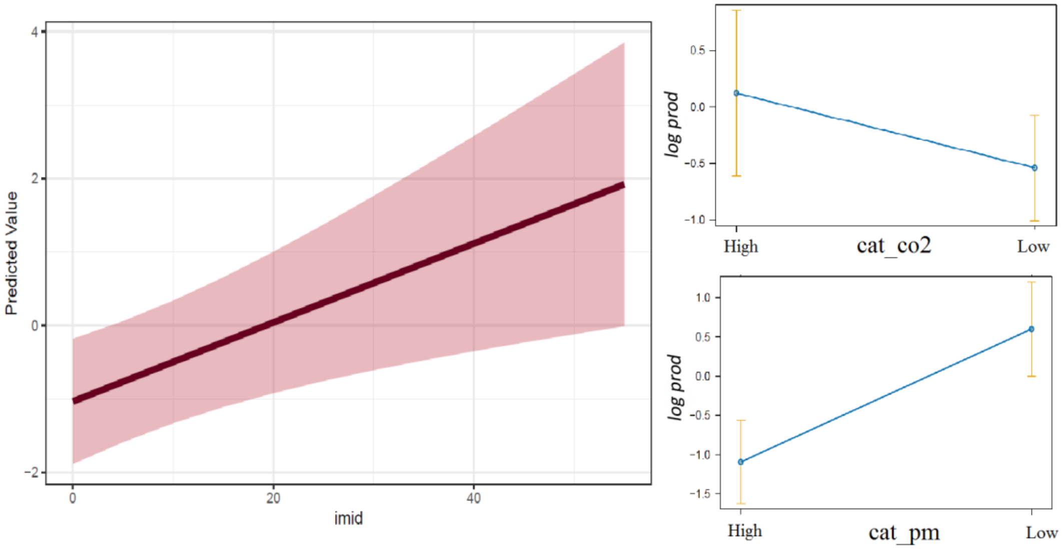

Figure 10: Effective plots

From the effective plots, it can be observed that, while the estimate for $PM_{2.5}$ demonstrates negative relationship between air pollution and honey production, confirming the initial hypothesis, it also has smaller uncertainty associated with it.

5.2. Significance of regression coefficients

To determine whether the effect of a covariate on the response is significant, a null hypothesis of the lack of effect is proposed:

$$H_0\colon \beta_j = 0;\ \ H_1\colon \beta_j \neq 0;\ \ \beta_{!j}\ \ arbitrary$$

To test this hypothesis, a standard t-test is run, the results for which with corresponding p-values are recorded in Table 5. For imidacloprid and $PM_{2.5}$, it can be concluded that the effects of those covariates on the response are statistically significant at 1% significance level. For the CO^2 variable, however, the null of the absence of the effect is not rejected.

5.3. Statistical testing for several regressors

From previous inference, it can be concluded the effect of $PM_{2.5}$ variable on the response is significant and has smaller uncertainty associated with it. Hence, it is important to test a hypothesis if the contribution of the remaining two covariates improves the amount of the explained variability of the response enough to provide a significant result of an ANOVA testing: this implies testing a pair of covariates to assess if their contribution to the explanation of variance is discernable enough with respect to the amount of variance explained by the full model. The following hypothesis is tested:

$$H_0\colon \beta_1 = \beta_2 = 0;\ \ \beta_3\ \ arbitrary$$

```
fit_nested <- lm (prod ~ imid + cat_co2, data = data_vst)
anova_res <- round (anova(fit_nested, fit), digits = 2)
```

Table 6: Testing multiple regression coefficients

Res.Df	RSS	Df	Sum of Sq	F	P-value
31	60.30				
30	37.45	1	22.84	18.3	< 2.2e-16

Outcome of the ANOVA analysis suggests that CO^2 and imidacloprid have statistically significant effect at 1% level in the model when considered together, hence should not be excluded from the model. However, this effect can likely be attributed to the variable for imidacloprid, for it is statistically significant according to the results of the t-testing above, whereas CO^2 variable was deemed not significant in the model.

Furthermore, the global F-test can be analysed, which takes into account all the predictors, and tests the specification of the mean functions as such (Weisberg (2005), p. 134):

$$H_0: E(Y\,|X) = \beta_0;\; H_1: E(Y\,|\,X) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

The result shown in Table 5 suggests that there is statistical evidence against the null at 1% level, which implies that the predictors significantly improve the fit of the model, compared to the null model.

5.4. Goodness of fit

Common measures of goodness of fit for linear regression model are the coefficient of determination and the residual standard error. As evidenced by Table 5, R^2 amounts to approximately 0.45, which implies that model explains around 45% of the variability of the response variable. Residual standard error is the measure of lack of fit of the model (James et al. (2013), p. 69), its value suggesting that total honey production in each states deviates from the true regression line by approximately 1.12 units on the log scale, which, considering the regression estimation results and the summary statistics of the data, appears to be a relatively large value.

6. Prediction for an omitted state

At the step of data collection, some states were omitted from the analysis because the data on pesticide usage was not collected for them for that year. Using the obtained regression fit, it is possible to estimate the value for total honey production for a state previously omitted from the analysis. The state of California was chosen: the data for pesticide imidacloprid was used from 2015, whereas for the air pollution variables actual data for 2016 was input.

```r
new_data <- data.frame ("imid" = (150569.3 / 1000), "cat_co2" = ifelse (9 > mean (data_full$co2
), "High", "Low"), "cat_pm" = ifelse (36.7 > mean (data_full$pm), "High", "Low"))
cali_pred <- predict (fit, newdata = new_data, interval = "prediction", level= .95, se.fit = T)
```

Table 7: Prediction for California

True value	Prediction	Standard error	Confidence interval
1.622	6.395	2.455	(0.885,11.904)

The true value for total honey production in California is available in the original dataset, hence it is possible to meaningfully compare obtained prediction with the actual value. Table 7 demonstrates that the model has poor predictive ability: the predicted value does not correspond to the true value; however, the true value is included in the confidence interval. It must be noted that the confidence interval is quite wide, relative to the values estimated. This is stressed by the fact that the predictive confidence interval incorporates two sources of uncertainty: uncertainty associated with the $\hat{\beta}$ estimates, and uncertainty associated with observing a new value.

7. Simulating new values

Using the model specification, it is possible to simulate new observation for the response variable using the predictor variables. The predictors are assumed non-stochastic and measured without error. The estimates of the linear regression (the $\hat{\beta}$ coefficients and the residual standard error, which is the estimate for the square root of the variance of the errors) are assumed to be the true population parameters. Hence, the values obtained by matrix multiplication of the model matrix to the vector of $\hat{\beta}$ coefficients would provide the mean response for each state; by adding to it a stochastic component, an error term, distributed $N(0, \hat{\sigma}^2)$, it would be possible to simulate new values for the log of honey production for the population.

```
sims <- function (coefficients, seed = Sys.time ()) {
  set.seed (seed)
  sigma <- sum_reg$sigma
  preds <- model.matrix (fit)
  xb <- preds %*% coefficients
  epsilons <- rnorm (n, mean = 0, sd = sigma)
  y_sim <- xb + epsilons
  return (y_sim)}
coefs <- sum_coefs [, 1]; simulations <- sims (coefs, 17)
```

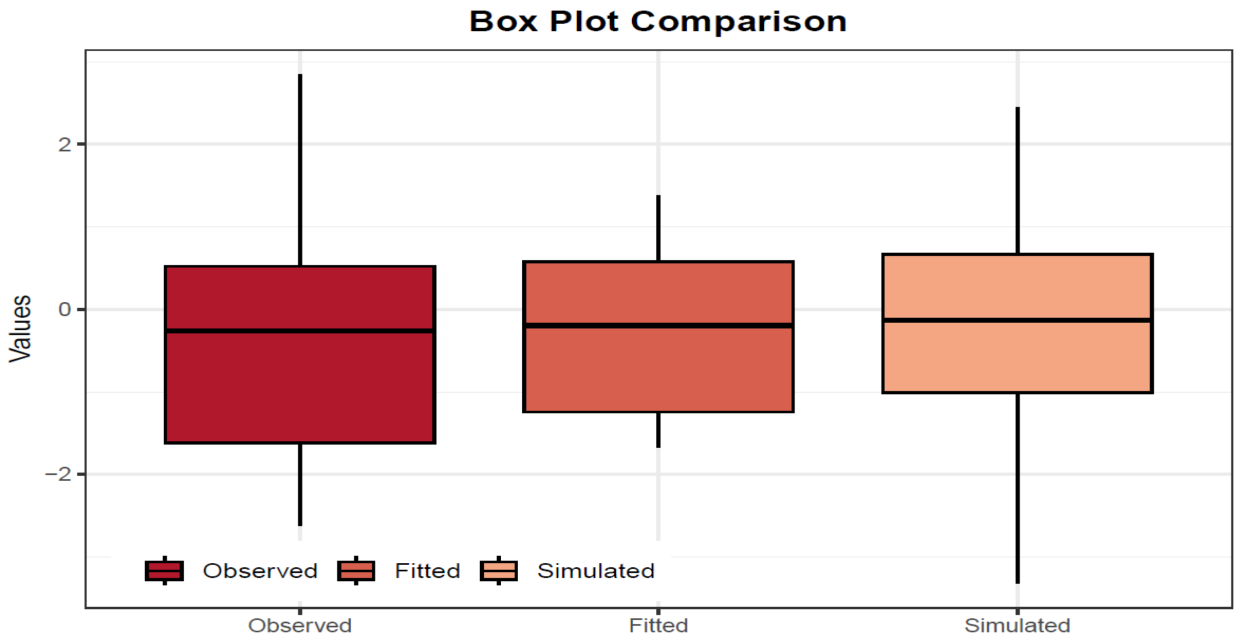

Figure 11: Comparison of distributions of observed, fitted and simulated values for log of total honey production

From Figure 11 it can be observed that the distribution of simulated and fitted values is consistent with the distribution of observed values, with a slight difference in the tails.

8. Conclusions

The estimation of the linear regression model allowed to obtain some insights into the nature of the relationship between honey production and certain environmental factors. As such, a significant negative linear relationship was found between $PM_{2.5}$ and the response, which suggests that, for states with higher concentrations of particular matter in the air, lower values for honey production are observed. However, the relationship found between the usage of a neonicotinoid pesticide and imidacloprid and honey production was determined to be positive. It should be underlined that both results are associated with some uncertainty, and the model itself demonstrated a quite high residual standard error, despite explaining almost half of the overall variability of the response.

Nevertheless, there is a wide range of possibilities for improving the analysis in the future. As such, inclusion of additional predictors, such as traffic pollution, the presence of human-made structures and weather conditions should be also accounted for in the analysis. What is more, it would be beneficial to expand the research to include a longer time frame to be able to observe some trends that may occur over time. Additionally, other regressors, such as interaction terms between covariates for air pollution, may be included in the model to study their joint effects on the response. Further ways to improve the analysis may include investigating the relationships within observed units and attempting to account for any patterns discovered.

References

1. Blacquiere, Tjeerd, Guy Smagghe, Cornelis AM Van Gestel, and Veerle Mommaerts. 2012. “Neonicotinoids in Bees: A Review on Concentrations, Side-Effects and Risk Assessment.” *Ecotoxicology* 21: 973–92.

2. Capitani, Giancarlo, Giulia Papa, Marco Pellecchia, and Ilaria Negri. 2021. “Disentangling Multiple PM Emission Sources in the Po Valley (Italy) Using Honey Bees.” *Heliyon* 7 (2): e06194.

3. Dormann, Carsten F, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R, García Marquéz, et al. 2013. “Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance.” Ecography 36 (1): 27–46.

4. Faraway, Julian J. 2016. *Extending the Linear Model with r: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC press.

5. Fox, John, and Sanford Weisberg. 2019. An R Companion to Applied Regression. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

6. Fox, John. 2015. *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications.

7. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer Science & Business Media.

8. Jeschke, Peter, and Ralf Nauen. 2008. “Neonicotinoids—from Zero to Hero in Insecticide Chemistry.” *Pest Management Science: Formerly Pesticide Science* 64 (11): 1084–98.

9. Klein, Alexandra-Maria, Bernard E Vaissiere, James H Cane, Ingolf Steffan-Dewenter, Saul A Cunningham, Claire Kremen, and Teja Tscharntke. 2007. “Importance of Pollinators in Changing Landscapes for World Crops.” *Proceedings of the Royal Society B: Biological Sciences* 274 (1608): 303–13.

10. Thimmegowda, Geetha G, Susan Mullen, Katie Sottilare, Ankit Sharma, Saptashi Soham Mohanta, Axel Brockmann, Perundurai S Dhandapany, and Shannon B Olsson. 2020. “A Field-Based Quantitative Analysis of Sublethal Effects of Air Pollution on Pollinators.” *Proceedings of the National Academy of Sciences* 117 (34): 20653–61.

11. Weisberg, Sanford. 2005. *Applied Linear Regression*. Vol. 528. John Wiley & Sons.

12. “Why bees matter.” 2018. https://www.fao.org/3/I9527EN/i9527en.PDF.

13. Zattara, Eduardo E, and Marcelo A Aizen. 2021. “Worldwide Occurrence Records Suggest a Global Decline in Bee Species Richness.” *One Earth* 4 (1): 114–23.