Linear Regression Analysis

PHOK Ponna

01/01/2025

Case Study 1: Biomass

Data and ideas for this case study come from (Goicoa et al., 2011).

To estimate the amount of carbon dioxide retained in a tree, its biomass needs to be known and multiplied by an expansion factor (there are several alternatives in the literature). To calculate the biomass, specific regression equations by species are frequently used. These regression equations, called allometric equations, estimate the biomass of the tree by means of some known characteristics, typically diameter and/or height of the stem and branches. The BIOMASS file contains data of 42 beeches (Fagus Sylvatica) from a forest of Navarra (Spain) in 2006, where

- **diameter**: diameter of the stem in centimeters
- **height**: height of the tree in meters
- **stemweight**: weight of the stem in kilograms
- **aboveweight**: aboveground weight in kilograms

(a) Create a scatterplot of aboveweight versus diameter. Is the relationship linear? Superimpose a regression line over the plot just created.

(b) Create a scatterplot of log(aboveweight) versus log(diameter). Is the relationship linear? Superimpose a regression line over the plot just created.

(c) Fit the regression model log(aboveweight) $= \beta_0 + \beta_1$log(diameter), and compute R^2, R_a^2, and the variance of the residuals.

(d) Introduce log(height) as an explanatory variable and fit the model log(aboveweight) $=\beta_0 + \beta_1$ log(diameter) + β_2 log(height). What is the effect of introducing log(height)in the model?

(e) Complete the Analysis questions for the model in (d).

Analysis questions:

(1) Estimate the model's parameters and their standard errors. Provide an interpretation for the model's parameters.

(2) Compute the variance-covariance matrix of the $\hat{\beta}_s$.

(3) Provide 95% confidence intervals for β_1 and β_2.

(4) Compute the R^2, R_a^2, and the residual variance.

(5) Construct a graph with the default diagnostics plots of R.

(6) Can homogeneity of variance be assumed?

(7) Do the residuals appear to follow a normal distribution?

(8) Are there any outliers in the data?

(9) Are there any influential observations in the data?

(f) Obtain predictions of the aboveground biomass of trees with diameters diameter = seq(12.5, 42.5, 5) and heights height = seq(10, 40, 5). Note that the weight predictions are obtained from back transforming the logarithm. The bias correction is obtained by means of the lognormal distribution: If $\hat{Y}_{pred}$ is the prediction, the corrected(back-transformed) prediction $\tilde{Y}_{pred}$ is given by

$$\tilde{Y}_{\text{pred}} = \exp(\hat{Y}_{\text{pred}} + \hat{\sigma}^2/2)$$

where $\hat{\sigma}^2$ is the variance of the error term.

Case Study 2: Fruit Trees

Data and ideas for this case study come from Militino et al. (2006).

To estimate the total surface occupied by fruit trees in three small areas (R63, R67, and R68) of Navarra in 2001, a sample of 47 square segments has been taken. The experimental units are square segments or quadrats of 4 hectares, obtained by random sampling after overlaying a square grid on the study domain. The focus of this case study is illustrating two different techniques used to obtain estimates: direct estimation and small area estimation. The direct technique estimates the total surface area by multiplying the mean of the observed surface area in the sampled segments by the total number of segments in every small area. The small area technique consists of creating a regression model where the dependent variable is the observed surface area occupied by fruit trees in every segment and the explanatory variables are the classified cultivars by satellite in the same segment and the small areas to which they belong. The final surface area totals are obtained by multiplying the total classified surface area of every small area by the β's parameter estimates obtained from the regression model (observed surface area $\sim$ classified surface area + small areas). The surface variables in the data frame SATFRUIT are given in m^2:

- **quadrat** is the number of the sampled segment or quadrat
- **smallarea** are the small areas' labels
- **wheat** is the classified surface of wheat in the sampled segment
- **barley** is the classified surface of barley in the sampled segment
- **nonarable** is the classified surface of fallow or non-arable land in the sampled segment
- **corn** is the classified surface of corn in the sampled segment
- **sunflower** is the classified surface of sunflowers in the sampled segment
- **vineyard** is the classified surface of vineyards in the sampled segment
- **grass** is the classified surface of grass in the sampled segment
- **asparagus** is the classified surface of asparagus in the sampled segment
- **alfalfa** is the classified surface of lucerne (type of alfalfa) in the sampled segment
- **rape** is the classified surface of rape Brassica napus in the sampled segment
- **rice** is the classified surface of rice in the sampled segment
- **almonds** is the classified surface of almonds in the sampled segment
- **olives** is the classified surface of olives in the sampled segment
- **fruit** is the classified surface of fruit trees in the sampled segment

- **observed** is the observed surface of fruit trees in the sampled segment

(a) Characterize the shape, center, and spread for the variable **fruit**.

(b) What is the maximum number of m^2 of classified fruits by segment?

(c) How many observations are there by small area?

(d) Use **scatterplotMatrix()** from car or **pairs()** to explore the linear relationships between **observed** and the remainder of the numerical variables. Comment on the results.

(e) Create density plots of the observed fruits' surface area (**observed**) by small areas (**smallarea**).

(f) Use boxplots and barplots with standard errors to compare the observed surface area (**observed**) and the classified surface area (**fruit**) by small areas (**smallarea**).

(g) Compute the correlation between **observed** and all other numerical variables. List the three variables in order along with their correlation coe!cients that have the highestcorrelation with **observed**.

Model (A)

Use backward elimination to develop a model that predicts `observed` using the data frame `SATFRUIT` without considering `smallarea`. Start the backward elimination process by considering all of the numerical variables in `SATFRUIT` as potential predictors. Use a p-value-to-remove of 10%. Store the final model in the object `modelA`.

(i) Compute CV_n, the leave-one-out cross-validation error, for `modelA`. Set the seed to 5 and compute CV_5, the five-fold cross-validation error, for `modelA`. The cross-validation error for a generalized linear model can be computed using the `cv.glm()` function from the `boot` package. Using the function `glm()` without passing a family argument is equivalent to using the function `lm()`. R Code 1 provides a template for how to use the `cv.glm()` function. Note that CV_n is returned with `cv.error$delta[1]`. To compute CV_5, pass the value 5 to the argument K inside the `cv.glm()` function.

R Code 1:

```
> mod.glm <- glm(y ~ x1 + x2, data = DF)
> library(boot)
> cv.error <- cv.glm(data = DF, glmfit = mod.glm)
> cv.error$delta[1]
```

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (A). What is the proportion of total variability explained by Model (A)?

Model (B)

Use the criterion-based procedure AIC, which for linear regression is equivalent to Mallow's C_p, to develop a model that predicts `observed` using all of the numerical variables in `SATFRUIT`. Store the model in the object `modelB`. Verify that the model suggested using BIC is the same model as the one suggested by AIC or Mallow's C_p, which are all the same as Model (A).

Model (C)

Use mean squared prediction error (MSPE) to select a model using all of the numerical variables in `SATFRUIT` as potential predictors for predicting `observed`. Store the model in the object `modelC`. Specifically, select a model using both leave-one-out cross validation (LOOCV) and five-fold cross validation.

1. Compute CV_n for `modelC`. Set the seed to 5 and compute CV_5 for `modelC`.
2. Compute R^2, R_a^2, the AIC, and the BIC for Model (C). What is the proportion of total variability explained by Model (C)?

Model (D)

Use whichever of Model (A) or (C) has the smaller cross-validation error, and introduce `smallarea` into the chosen model. Store the new model that includes `smallarea` in `modelD`.

(i.) Eliminate any variables from `modelD` that are not statistically significant ($\alpha = 0.10$). Store the resulting model in `modelD`.

(ii.) Compute CV_n for `modelD`. Set the seed to 5 and compute CV_5 for `modelD`.

(iii.) Compute R^2, R_a^2, the AIC, and the BIC for Model (D). What is the proportion of total variability explained by Model (D)?

(iv.) Does Model (D) have a smaller cross-validation error than the cross-validation error for either Model (A) or Model (C)?

(v.) Plot the Cook distances, the studentized residuals, the diagonal elements of the hat matrix, the DFFITS, and DFBETAS1 of Model (D) versus the index.

(vi.) Are there any leverage points? Justify the answer given.

(vii.) Are there any outliers? Justify the answer given.

(viii) Check normality and homoscedasticity for Model (D) using graphics and hypothesis tests.

(ix.) Calculate a 95% confidence interval for the fruit coefficient.

(h) How many hectares of observed fruits are expected to be incremented if the classified hectares of fruit trees by the satellite are increased by 10,000 m^2 (1 ha)?

(i) Suppose the total classified fruits by the satellite in area R63 is 97,044.28 m^2, in area R67 is 4,878,603.43 m^2, and in area R68 is 2,883,488.24 m^2. Predict the total area of fruit trees by small areas.

(j) Create a plot of observed versus fruit with the points color coded according to small area. Superimpose the corresponding regression lines for each small area.

(k) Plot the individual predictions for model D versus the observed data. Add a diagonal line to the plot.

(l) Create a bar plot that displays the predicted area occupied by fruit trees based on model D for each small area and the direct estimates of the area occupied by fruit trees by small area knowing that the total number of classified segments in areas R63, R67, and R68 are 119, 703, and 564, respectively.

Case Study 3: Real Estate

Data and ideas for this case study come from (Militino et al., 2004).

The goal of this case study is to walk the user through the creation of a parsimonious multiple linear regression model that can be used to predict the total price (`totalprice`) of apartments by their hedonic (structural) characteristics. The data frame `VIT2005` contains several variables, and further description of the data can be found in the help file.

(a) Characterize the shape, center, and spread of the variable `totalprice`.

(b) Use `scatterplotMatrix()` from the `car` package or `pairs()` to explore the relationships between `totalprice` and the numerical explanatory variables: `area`, `age`, `floor`, `rooms`, `toilets`, `garage`, `elevator`, and `storage`.

(c) Compute the correlation between `totalprice` and all of the other numerical variables. List the three variables in order along with their correlation coefficients that have the highest correlation with `totalprice`.

Model (A)

Use backward elimination to develop a model that predicts `totalprice` using the data frame `VIT2005`. Use a "p-value to remove" of 5%. Store the final model in the object `modelA`.

(i) Compute CV_n, the leave-one-out cross validation error, for `modelA`. Set the seed to 5 and compute CV_5, the five-fold cross validation error, for `modelA`. The cross validation error for a generalized linear model can be computed using the `cv.glm()` function from the `boot` package. Using the function `glm()` without passing a family argument is the same as using the function `lm()`. R Code 2 provides a template of how to use the `cv.glm()` function. Note that CV_n is returned with `cv.error$delta[1]`. To compute CV_5, pass the value 5 to the argument `K` inside the `cv.glm()` function.

– R Code 2

```
> mod.glm <- glm(y ~ x1 + x2, data = DF)
> library(boot)
> cv.error <- cv.glm(data = DF, glmfit = mod.glm)
> cv.error$delta[1]
```

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (A). What is the proportion of total variability explained by Model (A)?

Model (B)

Use the criterion-based procedure AIC, which for linear regression is equivalent to Mallow's C_p, to develop a model that predicts `totalprice` using the variables in `VIT2005`. Store the model in the object `modelB`.

(i) Compute CV_n for `modelB`. Set the seed to 5 and compute CV_5 for `modelB`.

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (B). What is the proportion of total variability explained by Model (B)?

Model (C)

Use the criterion-based procedure BIC to develop a model that predicts `totalprice` using the variables in `VIT2005`. Store the model in the object `modelC`.

(i) Compute CV_n for `modelC`. Set the seed to 5 and compute CV_5 for `modelC`.

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (C). What is the proportion of total variability explained by Model (C)?

Model (D)

Use forward selection to develop a model that predicts `totalprice` using the variables in `VIT2005`. Use a "p-value to add" of 5%. Store the final model in the object `modelD`.

(i) Compute CV_n for `modelD`. Set the seed to 5 and compute CV_5 for `modelD`.

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (D). What is the proportion of total variability explained by Model (D)?

(d) Explore the residuals of the Models (A), (B), (C), and (D) using the function `residualPlot()` or `residualPlots()` from the package `car`. Comment on the results.

(e) Use the function `boxCox()` from `car` to find a suitable transformation for `totalprice`.

Model (E)

Use backward elimination to develop a model that predicts log(`totalprice`) using the data frame `VIT2005`. Use a "p-value to remove" of 5%. Store the final model in the object `modelE`.

(i) Compute CV_n for `modelE`. Set the seed to 5 and compute CV_5 for `modelE`.

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (E). What is the proportion of total variability explained by Model (E)?

Model (F)

Use the criterion-based procedure AIC, which for linear regression is equivalent to Mallow's C_p, to develop a model that predicts log(`totalprice`) using the variables in `VIT2005`. Store the model in the object `modelF`.

(i) Compute CV_n for `modelF`. Set the seed to 5 and compute CV_5 for `modelF`.

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (F). What is the proportion of total variability explained by Model (F)?

Model (G)

Use the criterion-based procedure BIC to develop a model that predicts log(`totalprice`) using the variables in `VIT2005`. Store the model in the object `modelG`.

(i) Compute CV_n for `modelG`. Set the seed to 5 and compute CV_5 for `modelG`.

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (G). What is the proportion of total variability explained by Model (G)?

Model (H)

Use forward selection to develop a model that predicts log(`totalprice`) using the variables in `VIT2005`. Use a "p-value to add" of 5%. Store the final model in the object `modelH`.

(i) Compute CV_n for `modelH`. Set the seed to 5 and compute CV_5 for `modelH`.

(ii) Compute R^2, R_a^2, the AIC, and the BIC for Model (H). What is the proportion of total variability explained by Model (H)?

(f) Which model has the smallest CV_5 as well as the smallest CV_n error among Models (E), (F), (G), and (H)?

(g) Use the model selected from part (f) and explore its residuals using the function `residualPlots()` from `car`. Comment on the results.

Model (I)

Refer to the model selected in part (e) as `modelI`.

(i) Plot the Cook distances, the studentized residuals, and the diagonal elements of the hat matrix of Model (I) versus the index. Based on the graphs, are there any outliers?

(ii) Create a bubble-plot of the studentized residuals versus the hat values with the function `influencePlot()`. Are any of the points influential?

(iii) The original researchers evaluated the apartments in rows 3 and 93 and decided they were not representative and decided to remove them from the study. Remove observations 3 and 93 from consideration in `modelI`.

(iv) Check normality and homoscedasticity for `modelI` using graphs and hypothesis tests.

(v) Find the variance inflation factors for Model (I). Is multicollinearity a problem?

(vi) Find the parameter estimates, and compute 95% confidence intervals for the parameters of Model (I).

(vii) Find the relative contribution of the explanatory variables to explaining the variability of the prices in Model (I).

(viii) What is the variable that explains the most variability in Model (I)?

(ix) What variables jointly explain 80% of the total variability of log(`totalprice`)?

(x) Find the predictions of Model (I) with bias correction and without bias correction. The bias correction is obtained by means of the lognormal distribution: If $\hat{Y}_{\text{pred}}$ is the prediction of Model (I), the corrected (backtransformed) prediction $\tilde{Y}_{\text{pred}}$ of Model (I) is given by

$$\tilde{Y}_{\text{pred}} = \exp\left(\hat{Y}_{\text{pred}} + \hat{\sigma}^2/2\right)$$

where $\hat{\sigma}^2$ is the variance of the error term, and the confidence interval is given by

$$l_{\text{inf}} = \exp\left(\widehat{Y}_{\text{pred}} + \hat{\sigma}^2/2 - z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left(\widehat{Y}_{\text{pred}}\right) + \widehat{\text{Var}}\left(\hat{\sigma}^2\right)/4}\right)$$

$$l_{\text{sup}} = \exp\left(\widehat{Y}_{\text{pred}} + \hat{\sigma}^2/2 + z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left(\widehat{Y}_{\text{pred}}\right) + \widehat{\text{Var}}\left(\hat{\sigma}^2\right)/4}\right)$$

and

$$\widehat{\text{Var}}\left(\hat{\sigma}^2\right) = \frac{2\hat{\sigma}^4}{df_{\text{residual}}}.$$

(xi) For Model (I), plot the predicted values (with and without bias correction) versus observed values. Comment on the results.

(xii) Show that in Model (I) an increment of 10 m^2 in the area of a flat implies an increment of roughly 4% in the predicted total price. To verify this, find the predicted price of three apartments with areas of 80, 90, and 100 m^2, respectively, and keep the rest of the explanatory variables fixed. For example, assign the following values to the explanatory variables: `zone = Z32`, `elevator = 1`, `toilets = 1`, `garage = 1`, `category = 3B`, `out = E50`, `storage = 1`, `heating = 3A`, and `streetcategory = S3`. Compute the corresponding 90% prediction intervals.

(xiii) What is the percentage change in the total price of an apartment when the number of garages changes from one to two?

(xiv) What is the percentage change in the total price of an apartment when the heating type changes from "1A" to "3B"?