Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Tutorial sheet 2: Regression – Solution

Prof. Dr. Martin Bichler
Julius Durmann, Markus Ewert, Yutong Chao, Dr. Mete Ahunbay

Exercise T2.1 *Testing Gauss-Markov assumptions*

Please use the provided Jupyter notebook to solve this task.
You might need to install `statsmodels` (see here for the documentation) by running

```
pip install statsmodels
```

(Make sure that your virtual Python environment is active!)

You are given the data set in *gauss-markov.csv*. It contains values for three variables X_1, X_2, X_3 and values for a target variable Y. Our goal is to predict the target variable based on the three input variables.

a) We start by using the simple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Using statsmodels, compute optimal values for the parameters. Let the model predict the values of $\hat{y} \approx y$.

Note: You may want to use `statsmodels.api.add_constant()` to add constant values for the intercept.

b) Interpret the coefficient for $\hat{\beta}_0$ and $\hat{\beta}_1$.

c) Which of the coefficients is statistically significant for a confidence level of $\alpha = 0.05$?

d) Compute the residuals $e = y - \hat{y}$ of the resulting model. Plot the residuals over the input variables x_1 and x_2. What do you observe?

Using a White test, show that we can reject the hypothesis of homoscedastic residuals at an α level of 0.01.

e) Consider the alternative model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2$$

Compute the optimal parameter values. You should observe that the R^2 value improves drastically over the previous model.

Although this model gives a very good fit of the data, there is another problem: Multicollinearity.
Use the Variance inflation factor to check whether the variables are dependent.

f) Consider a third model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2$$

and compute the optimal parameters.

Check if the model has multicollinear input variables using the VIF criterion.

Check if the model satisfies the homoscedasticity assumption using the White test and an α level of 0.01.

Solution

Please also refer to the provided solution notebook.

a) Parameter values:

β_0	31.7239
β_1	-14.0285
β_2	14.3497
β_3	2.5067

$R^2 = 0.663$.

b)
- β_0: If all other variables are equal to zero, i.e. $x_1 = x_2 = x_3 = 0$, then Y is equal to 31.72.
- β_1: If we change x_1 by a single unit and keep all other variables constant (=*ceteris paribus*), then y will *decrease* by 14.03.

c) By *statistical significant*, we usually refer to the t-test for each partial regression coefficient. The hypotheses are given by:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

In our example, only β_1 is statistically significant as the p-value is 0.043, which is smaller than the confidence level $\alpha = 0.05$, and thus, we can reject H_0 in this case.

d) Result of White test ($LM \sim \chi_3^2$): $LM = 73.744$, p-value = $2.8 \times 10^{-12} \ll 0.01$

Since the p-value is very low, we can reject the null hypothesis (homoscedasticity).
$\Longrightarrow$ Heteroscedastic residuals are very likely.

e) Parameter values:

β_0	3.9063
β_1	-3.8780
β_2	-0.4725
β_3	-0.8776
β_4	0.9977

$R^2 \approx 1$.

VIF values:

x_1	355.35
x_2	752.47
x_3	1080.27
x_1^2	1.01

The VIF values are very high ($\gg 10$). Consequently, we may assume that model has multicollinear input variables.

(Bonus) Result of White test ($LM \sim \chi^2_4$): $LM = 16.367$, p-value = $0.23 > 0.01$

f) Parameter values:

β_0	2.3017
β_1	-6.5107
β_2	3.3908
β_3	0.9949

$R^2 = 0.99$.

VIF values:

X_1	1.00
X_2	1.00
X_1^2	1.00

The VIF values are low (< 10). Consequently, the model inputs are likely independent.

Result of White test ($LM \sim \chi^2_3$): $LM = 4.315$, p-value = $0.83 > 0.01$
The p-value for the null hypothesis (homoscedastic residuals) is high. Therefore, the null hypothesis needn't be rejected.

Exercise T2.2 *Derivation of closed-form solution*

In this exercise, we will derive the closed-form solution of the regression problem

$$\beta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \tag{1}$$

summarize our input variables $x_i \in \mathbb{R}^d$ and target variables $y_i \in \mathbb{R}$.
The model which we use is

$$\hat{y}_i = \beta^T x_i$$

You may assume that the inputs x_i already contain an entry equal to 1 which allows to include the intercept of our model without further consideration. Also, you may assume that the vectors in $\mathbf{X}$ are not colinear.

a) Formulate the sum of squared errors $e_i = y_i - \hat{y}_i$

$$\mathcal{L} = \sum_{i=1}^{n} e_i^2$$

which we seek to minimize in our analysis:

(i) In terms of the individual elements x_i, y_i

(ii) In terms of the matrix notation $\mathbf{X}, y$

b) Calculate the derivatives (gradients):

(i) $\frac{\partial}{\partial \beta}(\beta^T a) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}(\beta^T a) \\ \vdots \\ \frac{\partial}{\partial \beta_d}(\beta^T a) \end{pmatrix}$ for $\beta, a \in \mathbb{R}^d$

(ii) $\frac{\partial}{\partial \beta}(\beta^T A \beta) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}(\beta^T A\beta) \\ \vdots \\ \frac{\partial}{\partial \beta_d}(\beta^T A\beta) \end{pmatrix}$ for $\beta \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$

Check your results with the matrix cookbook [**MatrixCookbook**], chapter 2.4.

c) Use these derivatives to compute the gradient of the loss: $\frac{\partial}{\partial \beta}\mathcal{L}(\beta)$.

d) Set the derivative to zero (first order condition) to obtain an estimate of β.

e) Why is there no need to check second-order derivatives to prove optimality?

Solution

a)

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - x_i^T \beta)^2 && (i) \\ &= (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) = y^T y - 2\beta^T \mathbf{X}^T y + \beta^T \mathbf{X}^T \mathbf{X}\beta && (ii) \end{aligned}$$

b) (i) Derive the individual components $\frac{\partial \beta^T a}{\partial \beta_i}$ and aggregate to obtain the gradient:

$$\frac{\partial \beta^T a}{\partial \beta_i} = a_i \qquad \Rightarrow \qquad \frac{\partial}{\partial \beta}(\beta^T a) = a$$

(ii) Reformulate the matrix-vector product to a summation of scalars:

$$\beta^T \mathbf{A}\beta = \sum_{i=1}^d \sum_{j=1}^d \beta_i A_{ij} \beta_j$$

Derive the individual components $\frac{\partial \beta^T \mathbf{A}\beta}{\partial \beta_k}$:

$$\begin{aligned} \frac{\partial \beta^T \mathbf{A}\beta}{\partial \beta_k} &= \sum_{i=1}^d \sum_{j=1}^d \frac{\partial(\beta_i A_{ij} \beta_j)}{\partial \beta_k} = \sum_{i=1}^d \sum_{j=1}^d \mathbb{I}_{i=k} A_{ij} \beta_j + \beta_i A_{ij} \mathbb{I}_{j=k} \\ &= \sum_{j=1}^d A_{kj}\beta_j + \sum_{i=1}^d \beta_i A_{ik} = A_{k,:}\beta + \beta^T A_{:,k} = (A_{k,:} + (A_{:,k})^T)\beta \end{aligned}$$

where $A_{k,:}$ denotes the k^{th} row of $\mathbf{A}$ and $A_{:,k}$ denotes the k^{th} column of $\mathbf{A}$.

Aggregate the gradient:

$$\frac{\partial \beta^T \mathbf{A} \beta}{\partial \beta} = (\mathbf{A} + \mathbf{A}^T)\beta$$

c)

$$\begin{aligned}\frac{\partial \mathcal{L}(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta}(\mathbf{X}\beta - y)^T(\mathbf{X}\beta - y) = \frac{\partial}{\partial \beta} y^T y - \frac{\partial}{\partial \beta} 2\beta^T \mathbf{X}^T y + \frac{\partial}{\partial \beta} \beta^T \mathbf{X}^T \mathbf{X} \beta \\ &= -\underbrace{2\mathbf{X}^T y}_{\text{b)(i)}} + \underbrace{(\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T)}_{\text{b)(ii)}}\beta = -2\mathbf{X}^T y + 2\mathbf{X}^T\mathbf{X}\beta\end{aligned}$$

d)

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = 0 \quad \Longleftrightarrow \quad -2\mathbf{X}^T y + 2\mathbf{X}^T\mathbf{X}\beta = 0 \quad \Longleftrightarrow \quad 2\mathbf{X}^T\mathbf{X}\beta = 2\mathbf{X}^T y \quad \Longleftrightarrow \quad \beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

e) The objective is fully convex in β (squared form). The solution from d) therefore is the unique global minimum.