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

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

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

Exercise H2.1 *Retail shop*

The following table displays customer demand for a retail shop.

t	0	1	2	3	4	5	6	7	8
Demand	28.20	37.65	47.28	59.76	73.44	86.19	100.31	112.58	121.63

Note: You can use Python to solve this exercise. Consider using the provided notebook as a template.

a) For the time series above, calculate the forecasted demand value for $t = 10$ using the simple linear regression and the formula below:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t.$$

b) Calculate the RMSE and explain its meaning.

c) For the time series above, calculate the forecasted demand value for $t = 10$, assuming a biannual seasonal component of the following form: Starting from the first period $t = 0$, suppose after every second period a new year begins. Make use of the formula below:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t + \hat{\beta}_2 \cdot Q_1.$$

d) Does the data reflect biannual data?

Solution

a) Forecast:

$$\hat{\beta}_0 = 25.3822, \quad \hat{\beta}_1 = 12.1833$$

and thus

$$\begin{aligned}\hat{y}_{10} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 10 \approx 25.3822 + 12.1833 \cdot 10 \\ &= 147.2152.\end{aligned}$$

b) *Calculation:* We have

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\text{RSS}}{T}}$$

with

$$\text{RSS} = \sum_{t=1}^{T} \hat{e}_t^2, \quad \hat{e}_t = y_t - \hat{y}_t.$$

Now plugging in the values of our table, we get $\hat{e}_0 \approx 28.20 - 25.38222 = 2.81777$, $\hat{e}_1 \approx 37.65 - 37.56556 = 0.08444$, etc. Thus, with RSS ≈ 27.72076, we have MSE ≈ 3.08, and

$$\text{RMSE} \approx \sqrt{3.08} \approx 1.755.$$

Interpretation: The RMSE is the average deviation of the prediction from the actual values of the data. It is useful for comparing different statistical (including machine learning) models for numerical prediction.

c) Forecast with an biannual seasonal component:

$$\hat{\beta}_0 \approx 25.3117, \quad \hat{\beta}_1 \approx 12.1833, \quad \hat{\beta}_2 \approx 0.1270,$$

which allows us the following forecast:

$$\begin{aligned} \hat{y}_{10} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 10 + 1 \cdot \hat{\beta}_2 \\ &\approx 25.3117 + 12.1833 \cdot 10 + 0.1270 \\ &= 147.2717. \end{aligned}$$

Note: Depending on your definition of the seasonal component (exchanged 0s and 1s), you could also end up with the following parameters. They should produce the same forecast for $\hat{y}_{10}$.

$$\hat{\beta}_0 \approx 25.4387, \quad \hat{\beta}_1 \approx 12.1833, \quad \hat{\beta}_2 \approx -0.1270,$$

d) No, because the seasonal variable is not statistically significant.

Exercise H2.2 *OLS implementation*

In this exercise, you will implement your own function for solving OLS regression problems in Python. The function takes the data samples in matrix-form (X, y) as inputs and returns the minimizing solution β as well as the remaining error $\mathcal{L}(\beta)$. You may want to use the provided Notebook as a template.

a) Implement the function. Use the provided template to get started.

b) For our provided toy data set (*ols-implementation-data.csv*), find the optimal regression parameters with the help of your implementation. Don't forget to add a variable for the intercept parameter!

c) Repeat b) with the aid of scikit-learn `LinearRegression` and verify your solution.

d) How much of the total variance can you explain with your model? Compute the R^2 measure. What happens if you forget about the intercept? How does the R^2 measure compare?

e) The computed R^2 value is not very good (even with the intercept). What could be the reason?

Solution

See solution notebook.

e) The model choice could be an inadequate match. Nonlinear transformations of the input variables (i.e. generalized least squares) could provide a better solution. In the solution script, you can find a generalized least-squares model where we added quadratic terms.

Exercise H2.3 *Determinants of Wages Data*

This exercise performs regression on the CPS1988 data set [**AER**].
Note: Use Python and statsmodels to solve this exercise. Have a look at the provided template notebook.

a) Load the data set from the provided file (*CPS1988.csv*). Briefly describe the data set:
   i) Name the dependent variable and the independent variables.
   ii) Which scales of measurement do the variables belong to (e.g., nominal, ordinal, interval or ratio)?
   iii) Does the data set consist of cross-sectional, time-series or panel data?

b) Plot the dependent variable against each independent variable and transform the variables if necessary.
   i) Which transformations would you carry out and why?
   ii) Estimate the following model:

$$\begin{aligned}\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{experience}_i \\ + \hat{\beta}_4 \cdot \text{experience}_i^2.\end{aligned} \tag{MR1}$$

c) Interpret the model from above (Equation MR1):
   i) Which variables are statistically significant?
   ii) Is the entire model statistically significant?
   iii) What is the explanatory power of the model and why?
   iv) Interpret each regression coefficient.

d) Now consider the following alternative model:

$$\begin{aligned}\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{education}_i \cdot \text{ethnicity}_i \\ + \hat{\beta}_4 \cdot \text{experience}_i + \hat{\beta}_5 \cdot \text{experience}_i^2.\end{aligned} \tag{MR2}$$

   What is the difference between both models from above (Equation MR1 and Equation MR2)?

e) Repeat c) with the model from Equation MR2.

Solution

a) i) Dependent variable: wage. Independent variables: education, experience, ethnicity, smsa, region, parttime.
   ii) Ratio: wage, education, experience. Nominal: ethnicity, smsa, region, parttime.
   iii) Cross-sectional: 28,155 different men in 1988.

b) See solution notebook.

c) Interpret the first model (Equation MR1):
   i) All variables, including the intercept, are statistically significant at level $\alpha = 0.01$ (look at Pr$[>|t|]$).
   ii) The entire model is statistically significant (F-statistic) at level $\alpha = 0.001$.

iii) Adjusted R-squared: 0.3346 (rather low explanatory power). Reason: too many important variables missing (e.g., ability).

iv) $\hat{\beta}_0 \approx 4.321$ and $\ln(\widehat{\text{wage}}_i) = 4.321$, $\widehat{\text{wage}}_i = e^{4.321} = 75.26$. The wage per week for Caucasian-American worker is \$75.26 with no education and no experience.

We have $\hat{\beta}_1 \approx 0.08567$. With education at x:

$$\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{experience}_i + \hat{\beta}_4 \cdot \text{experience}_i^2 \tag{1}$$

and education at $x + 1$:

$$\ln(\widehat{\text{wage}}_i') = \hat{\beta}_0 + \hat{\beta}_1 \cdot (x+1) + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{experience}_i + \hat{\beta}_4 \cdot \text{experience}_i^2. \tag{2}$$

we can combine them (Equation 2 $-$ Equation 1) as:

$$\ln(\widehat{\text{wage}}_i'/\widehat{\text{wage}}_i) = \hat{\beta}_1,$$

where

$$\ln(\widehat{\text{wage}}_i'/\widehat{\text{wage}}_i) = \ln\left(1 + \frac{\widehat{\text{wage}}_i' - \widehat{\text{wage}}_i}{\widehat{\text{wage}}_i}\right) \approx \frac{\widehat{\text{wage}}_i' - \widehat{\text{wage}}_i}{\widehat{\text{wage}}_i}.$$

In our case, we get

$$\hat{\beta}_1 \approx \frac{\widehat{\text{wage}}_i' - \widehat{\text{wage}}_i}{\widehat{\text{wage}}_i} = 0.08567.$$

Wage in dollars increases by 8.6 % for each additional year of education, keeping ethnicity and experience constant.

$\hat{\beta}_2 \approx -0.2434$. Wage is 24.34 % lesser for African-America worker ($\text{ethnicity}_i = 1$) as compared to Caucasian-American ($\text{ethnicity}_i = 0$), keeping education and experience constant.

For $\hat{\beta}_3 \approx 0.07747$ and $\hat{\beta}_4 \approx -0.001316$, we have

$$\ln(\widehat{\text{wage}}_i) = \mu + \hat{\beta}_3 \cdot \text{experience}_i + \hat{\beta}_4 \cdot \text{experience}_i^2 \tag{3}$$

and

$$\ln(\widehat{\text{wage}}_i') = \mu + \hat{\beta}_3 \cdot (\text{experience}_i + 1) + \hat{\beta}_4 \cdot (\text{experience}_i + 1)^2. \tag{4}$$

Here μ contains all other variables and their coefficients (cancels out in next step). Now we consider Equation 4 $-$ Equation 3:

$$\ln(\widehat{\text{wage}}_i'/\widehat{\text{wage}}_i) \approx 0.07747 - 2 \cdot 0.001316 \cdot \text{experience}_i - 0.001316. \tag{5}$$

Note that because experience enters the linear regression as a linear and a quadratic term, the effect of an increase in experience on wage depends on the level of experience.

- Suppose $\text{experience}_i = 40$: Then,

$$\ln(\widehat{\text{wage}}_i'/\widehat{\text{wage}}_i) = -0.029.$$

Wage decreases by 2.9 % when workers with at least 40 years of experience accumulate one additional year of experience, keeping other independent variables constant.

- Suppose $\text{experience}_i = 10$: Then,

$$\ln(\widehat{\text{wage}}_i'/\widehat{\text{wage}}_i) = 0.05.$$

Wage increases by 5 % when workers with at least 10 years of experience accumulate one additional year of experience, keeping other independent variables constant.

To find the number of years of experience at which further experience decreases the wage, we have to set the term of Equation 5 to zero:

$$\ln(\widehat{\text{wage}}_i'/\widehat{\text{wage}}_i) \approx 0.07747 - 2 \cdot 0.001316 \cdot \text{experience}_i - 0.001316 \stackrel{!}{=} 0.$$

Solving this equation gives $\text{experience}_i = 28.934$ (years).

d) The second model from Equation MR2 contains an interaction term in addition to the first model. The interaction term between education and ethnicity allows us to distinguish between the marginal effect of education on the wage of an African-American worker and on the wage of a Caucasian-American worker.

Model 1:

$$\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{experience}_i + \hat{\beta}_4 \cdot \text{experience}_i^2.$$

Model 2:

$$\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{education}_i \cdot \text{ethnicity}_i + \hat{\beta}_4 \cdot \text{experience}_i + \hat{\beta}_5 \cdot \text{experience}_i^2.$$

The interaction term in Model 2 is captured in $\text{education}_i \cdot \text{ethnicity}_i$ with coefficient $\hat{\beta}_3$.

e) Interpret the second model from Equation MR2:

i) All variables, including the intercept, are statistically significant at a level $\alpha = 0.05$. The effect of being African-American on wage now splits up between the dummy and the interaction effect and therefore is weaker for each variable.

ii) The entire model is statistically significant (F-statistic) at level $\alpha = 0.001$.

iii) Adjusted R-squared: 0.3347 has increased slightly. Reason: still far too many important variables missing (e.g., ability).

iv) To interpret the coefficients of the second model, let us define the following simplified version:

$$\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{education}_i \cdot \text{ethnicity}_i + \mu.$$

Again, μ contains all other variables and their coefficients. Their interpretation does not differ from c). We treat both ethnicities separately:

- For $\text{ethnicity}_i = 0$ (Caucasian-American worker), we have $\ln(\widehat{wage}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \mu$.

$\hat{\beta}_0 \approx 4.313$, $\widehat{\text{wage}}_i = e^{4.313} = 74.66$. Wage per week for Caucasian-American worker with no education and no experience is \$74.66.

$\hat{\beta}_1 \approx 0.08631$. Wage increases by 8.6 % for each additional year of education for Caucasian-American worker, keeping experience constant.

- For $\text{ethnicity}_i = 1$ (African-American worker), it follows $\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 + \hat{\beta}_3 \cdot \text{education}_i + \mu$ which is equivalent to $\ln(\widehat{\text{wage}}_i) = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) \cdot \text{education}_i + \mu$.

  $(\hat{\beta}_0 + \hat{\beta}_2) = 4.313 - 0.124 = 4.189$, $\widehat{\text{wage}}_i = e^{4.189} = 65.96$. Wage per week for African-American worker with no education and no experience is \$65.96.

  $\hat{\beta}_2 = -0.124$. Wage is 12.4 % lesser if worker with no education and no experience is African-American instead of Caucasian-American: $74.66 \cdot (1 - 0.124) = 65.4 (\approx 65.96)$.

  $\hat{\beta}_1 + \hat{\beta}_3 = 0.08631 - 0.00965 = 0.07666$. Wage increases by 7.6 % for each additional year of education for African-American worker, keeping experience constant.

$\hat{\beta}_3 = -0.00965$. Wage is 0.965 % lesser for African-American worker than for Caucasian-American worker for each additional year of education, keeping experience constant.