Econometrics I, WS 2019/20, 1. Partial, 5.11.2019

Time: 105 min

Name:

Student-ID:

Course number:

1 Case Study 1 – Simple Regression

The following monthly data describes the consumption of ice tea in New Austin from 1951–03–18 until 1953–07–11. Y is the per capita consumption of ice tea (in gallons) and X is the average daily maximum temperature (in degrees Celsius).

1. The simple regression model

$$Y = \beta_0 + \beta_1 X + u, \qquad u|X \sim \text{Normal}\left(0, \sigma^2\right), \tag{1}$$

is fitted with EViews/R. The results can be found in Figure 1 (estimated in R) on page 11 or in Figure 7 (estimated in EViews) on page 15.

(a) What is the value of the OLS estimate of β_0? How can this number be interpreted?

(b) What is the value of the OLS estimate of β_1? How can this number be interpreted?

(c) What is the expected ice tea consumption when the average daily maximum temperature is 20 degrees Celsius?

(d) What is the expected change in ice tea consumption when the average daily maximum temperature falls by 5 degrees celsius?

(e) What is $\sum_{i=1}^{N} y_i - \hat{y}_i$ equal to in this example?

(f) Consider the scatterplot of the data (including the regression line) in Figure 2 (R) or Figure 8 (EViews) and the scatterplot of the residuals in Figure 3 (R) or Figure 9 (EViews). Which of the usual assumptions for the linear regression model seem fulfilled (or not fulfilled)?

2. Now let us examine the linear regression model where the logarithm of Y (but not of X) was taken before estimating it:

$$\log Y = \gamma_0 + \gamma_1 X + \epsilon, \qquad \epsilon | X \sim \text{Normal}\left(0, \sigma_\epsilon^2\right). \tag{2}$$

The results of the OLS estimation can be found in Figure 4 (R) or Figure 10 (EViews).

(a) What *relative* change in ice tea consumption is to be expected if the average daily maximum temperature rises by 1 degree Celsius?

Hint: Determine the predicted value $\widehat{\log}(y)$ *for* $X = x$ *and for* $X = x + 1$ *using that approximately* $\widehat{\log}(y) \approx \log(\hat{y})$*. Additionally, use that approximately* $\log(\hat{y}) \approx \hat{y} - 1$.

(b) Is model (1) or (2) better suited for describing the data? Reason based on the scatter plots in Figures 2 and 3 (R) or Figures 8 und 9 (EViews).

2 Case Study 2 – Multiple Regression

The topic of energy performance of buildings (EPB) is very timely, due to the growing concerns about energy waste and its impact on the environment. Air conditioning accounts for most of the energy use in buildings during the summer. Therefore, it is crucial to design buildings that have good energy conservation properties, such as the cooling load. The cooling load is the rate at which heat energy would need to be removed from a space to maintain the temperature in an acceptable range.

In this case study, we use a multiple linear regression framework to study the effect of the following four input variables on the cooling load (`coolload`) in kW of a building: the relative compactness index (`relcomp`) in %, the wall area (`wallarea`) in m^2, the roof area (`roofarea`) in m^2, and the glazing index (`glazing`) in %. Note that the glazing index (`glazing`) is a percentage between 0 and 100, indicating the percentage of the wall area made of glass. (E.g., a value of 0 % means that there are no windows and a value of 100 % means that all walls are made of glass). The sample size is $N = 768$.

Answer the following questions based on Figure 13 (EViews output) or on Table 1 (R output) .

1. What is the OLS estimate of the regression coefficient associated with the predictor variable roof area? How do you interpret it?

2. What is the OLS estimate of the intercept? What is its usual interpretation and does it make sense in this example? Explain!

3. Determine the expected cooling load for a building with a $220\,m^2$ roof area, that has a relative compactness index of $70\,\%$, a wall area of $310\,m^2$, and a glazing index of $40\,\%$.

4. On average, how much higher would the cooling load of the building in Question 3 be if all walls were made of glass?

5. Calculate the (unbiased) estimate of the error variance ($\hat{\sigma}^2$) using the SSR reported.

6. What is the value of the coefficient of determination and how can you interpret it?

7. The boss of a building company wonders why the surface area (`surfacearea`) is not used as an additional predictor. Notice that the surface area can be defined as

   `surfacearea = wallarea + 2*roofarea`.

   Would you follow the boss' suggestion and include the surface area as an additional predictor? Explain your answer!

8. We fit another model, including the additional predictor 'wall height' (`height`) in m. The resulting coefficient of determination is $R^2 = 0.8876$. How would you decide which model to pick?

3 True or false?

True or false? Where is the error in a false statement? The question is considered to be answered correctly if you classify a true statement as true or if you correct a false statement such that it becomes true.

1. Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + u,$$

where $\mathbb{E}(u|X) = 0$. Suppose further that $\beta_0 = 2$, $\beta_1 = -3$ and $\mathbb{E}(X) = 1$.

(a) If X decreases by 2 units, we expect Y to increase by 6 units.

(b) The expected value of Y is 2.

2. Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + u, \qquad \text{where} \quad \mathbb{E}(u|X) = 0, \quad \mathbb{V}(u|X) = \sigma^2 > 0.$$

Suppose you have a random sample (y_i, x_i), $i = 1, \ldots, N$, from this population. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimators of β_0 and β_1.

(a) The smaller the sample variation $s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ of the regressor, the smaller is the variance of the estimator $\hat{\beta}_1$.

(b) Now consider another model

$$Y = \gamma_0 + \gamma_1 Z + u,$$

where $Z = -X$. The OLS estimators $\hat{\gamma}_0$ and $\hat{\gamma}_1$ for γ_0, γ_1 satisfy the following relation

$$\hat{\gamma}_0 = \hat{\beta}_0, \qquad \hat{\gamma}_1 = -\hat{\beta}_1.$$

3. Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u,$$

where $\mathbb{E}(u|X_1, X_2, X_3) = 0$, $\mathbb{V}(u|X_1, X_2, X_3) = \sigma^2 > 0$. Let $(y_i, x_{1,i}, x_{2,i}, x_{3,i})$, $i = 1, \ldots, N$, be a random sample.

Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$. Define the residuals $\hat{u}_i = y_i - \hat{y}_i$, $i = 1, \ldots, N$, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \hat{\beta}_3 x_{3,i}$.

(a) If N is even, the number of positive residuals equals the number of negative residuals.

(b) It holds that $\mathbb{E}(\hat{\beta}_j) = 0$ for all $j = 0, 1, 2, 3$.

(c) Now, suppose Y corresponds to an individual's income, X_1 indicates their mother's income, X_2 their father's income and X_3 is the average of the two parental incomes. Let $(y_i, x_{1,i}, x_{2,i}, x_{3,i})$, $i = 1, \dots, N$, be a random sample. Then the OLS estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ can be computed as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \qquad \text{where} \quad \boldsymbol{y} = (y_1, y_2, \dots, y_N)'.$$

(d) The coefficient of determination R^2 always decreases when omitting a regressor variable.

A R output

A.1 Case Study 1

```
Call:
lm(formula = cons ~ temp_c, data = iceT)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0393 -0.7996 -0.2193  0.2396  3.2941 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.24832    0.33122   0.750  0.45967   
temp_c       0.08779    0.02535   3.463  0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.245 on 28 degrees of freedom
Multiple R-squared:  0.2999,	Adjusted R-squared:  0.2749 
F-statistic: 11.99 on 1 and 28 DF,  p-value: 0.001735
```

Figure 1: Case Study 1

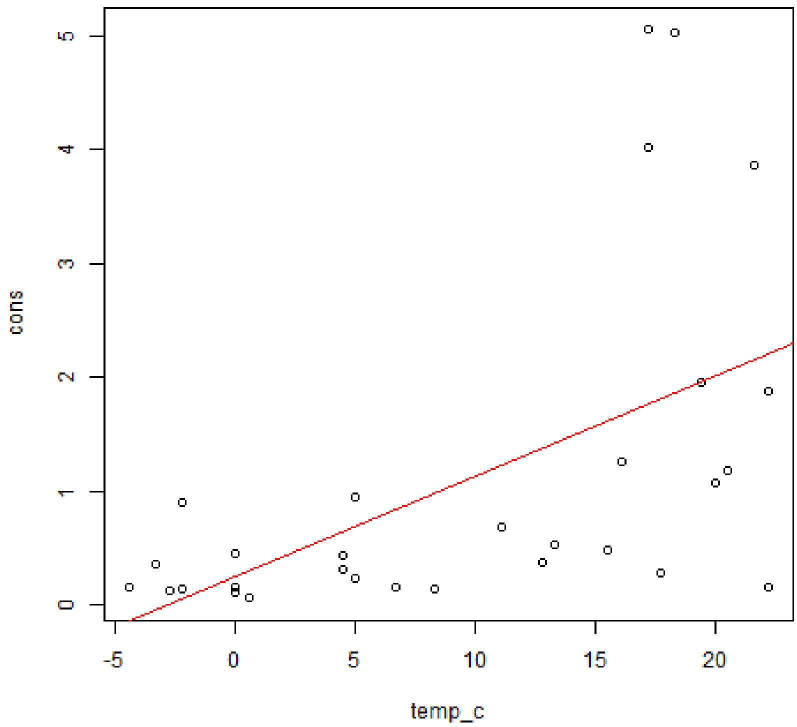

Figure 2: Case Study 1

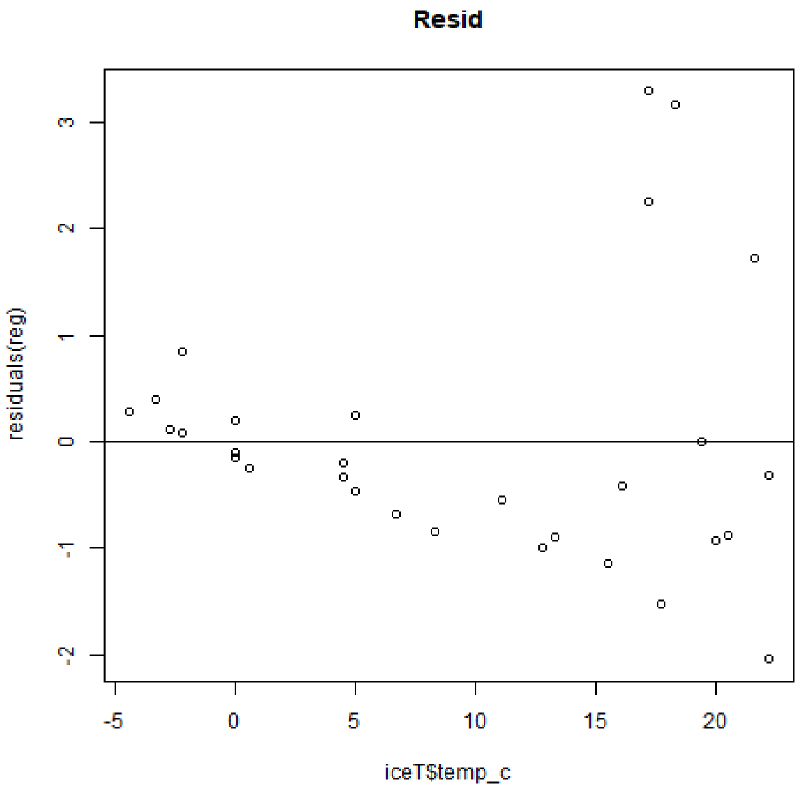

Figure 3: Case Study 1

```
Call:
lm(formula = log(cons) ~ temp_c, data = iceT)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.28780 -0.54652 -0.08408  0.65731  1.61052 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.52211    0.25571  -5.953 2.08e-06 ***
temp_c       0.08895    0.01957   4.545 9.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9615 on 28 degrees of freedom
Multiple R-squared:  0.4245,	Adjusted R-squared:  0.404 
F-statistic: 20.66 on 1 and 28 DF,  p-value: 9.615e-05
```

Figure 4: Case Study 1

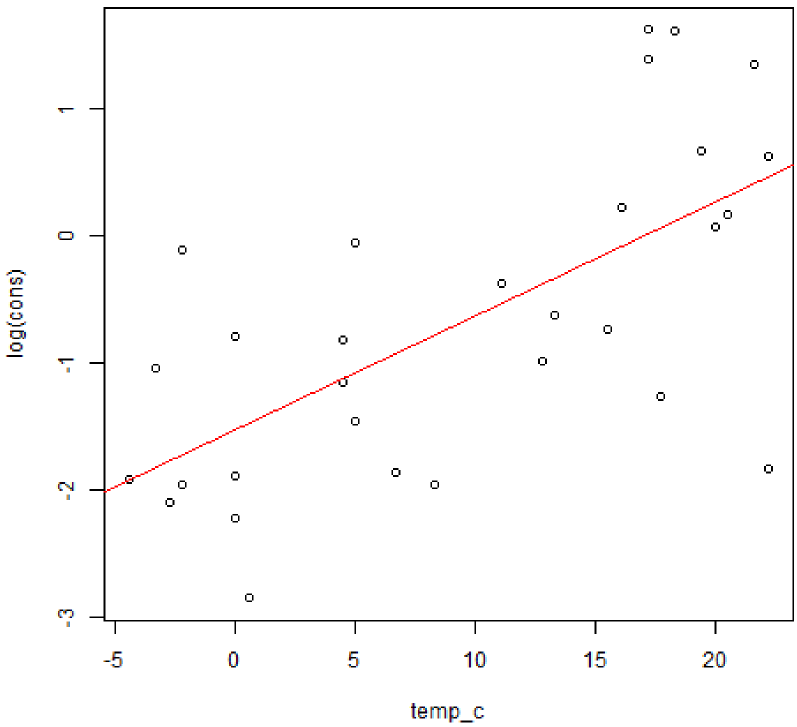

Figure 5: Case Study 1

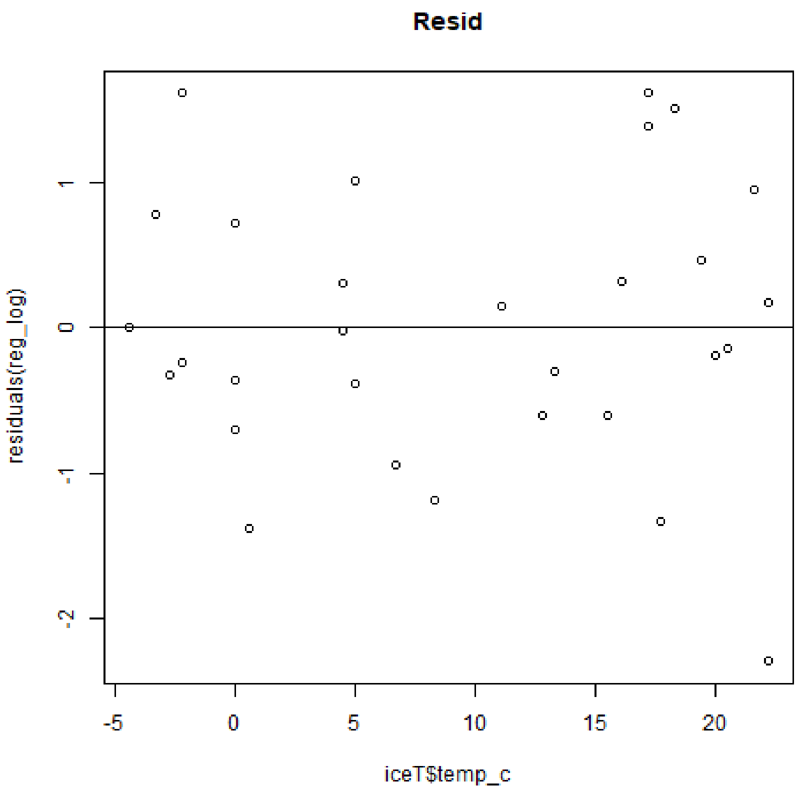

Figure 6: Case Study 1

A.2 Case Study 2

```
> fit <- lm(coolload ~ relcomp + wallarea + roofarea +  glazing, data = energy)
> summary(fit)

Call:
lm(formula = coolload ~ relcomp + wallarea + roofarea + glazing,
data = energy)

Residuals:
Min      1Q  Median      3Q     Max
-7.2274 -1.7729 -0.8312  1.0400 12.8147

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 279.970327  14.746220   18.99   <2e-16 ***
relcomp      -1.541746   0.093561  -16.48   <2e-16 ***
wallarea     -0.145381   0.011745  -12.38   <2e-16 ***
roofarea     -0.536435   0.022432  -23.91   <2e-16 ***
glazing       0.148180   0.009406   15.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.47 on xxx degrees of freedom
Multiple R-squared:  0.8676,Adjusted R-squared:  0.8669
F-statistic:  1250 on 4 and 763 DF,  p-value: < 2.2e-16

> SSR <- sum(fit$residuals^2)
> SSR
[1] 9188.46
```

Table 1: Case Study 2 Output

B EViews output

B.1 Case Study 1

Dependent Variable: CONS
Method: Least Squares
Date: 10/29/19 Time: 16:26
Sample: 1 30
Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.248324	0.331215	0.749736	0.4597
TEMP_C	0.087789	0.025350	3.463053	0.0017

R-squared	0.299873	Mean dependent var	1.082320
Adjusted R-squared	0.274868	S.D. dependent var	1.462591
S.E. of regression	1.245465	Akaike info criterion	3.341235
Sum squared resid	43.43310	Schwarz criterion	3.434648
Log likelihood	-48.11852	Hannan-Quinn criter.	3.371119
F-statistic	11.99274	Durbin-Watson stat	2.406628
Prob(F-statistic)	0.001735		

Figure 7: Case Study 1

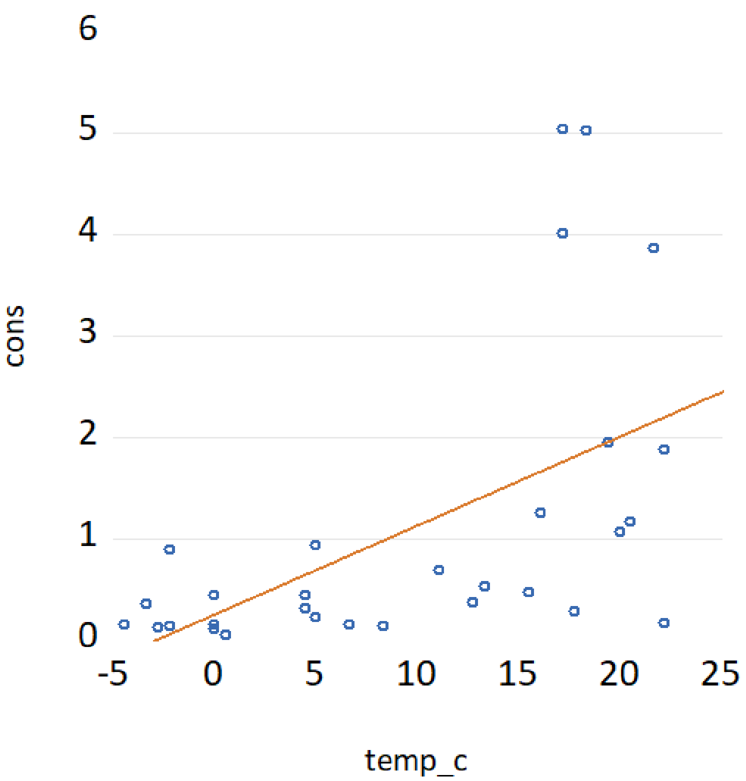

Figure 8: Case Study 1

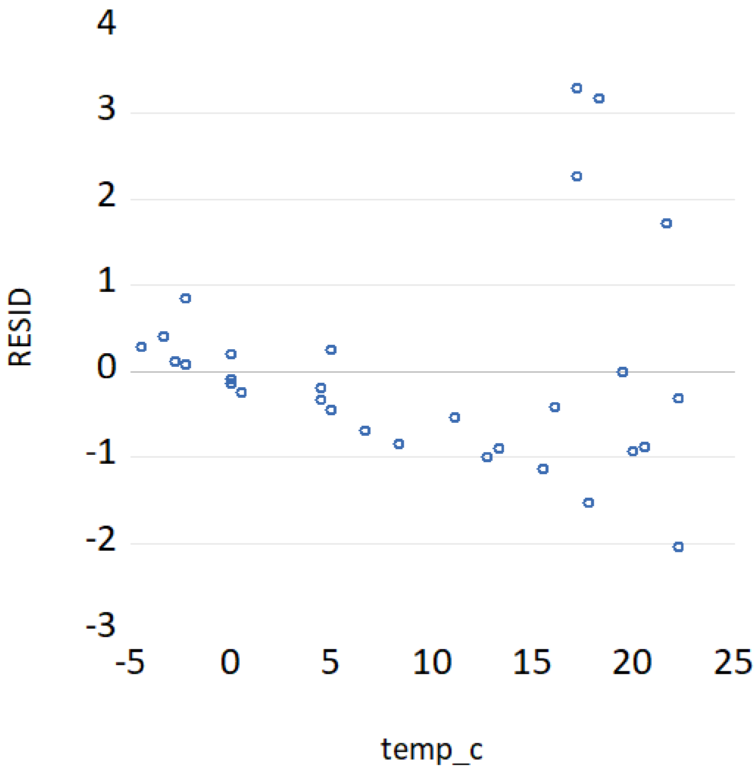

Figure 9: Case Study 1

Dependent Variable: LOG(CONS)
Method: Least Squares
Date: 10/29/19 Time: 15:52
Sample: 1 30
Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.522113	0.255708	-5.952540	0.0000
TEMP_C	0.088948	0.019571	4.544882	0.0001

R-squared	0.424531	Mean dependent var	-0.677104
Adjusted R-squared	0.403978	S.D. dependent var	1.245474
S.E. of regression	0.961536	Akaike info criterion	2.823771
Sum squared resid	25.88744	Schwarz criterion	2.917184
Log likelihood	-40.35656	Hannan-Quinn criter.	2.853655
F-statistic	20.65595	Durbin-Watson stat	2.132731
Prob(F-statistic)	0.000096		

Figure 10: Case Study 1

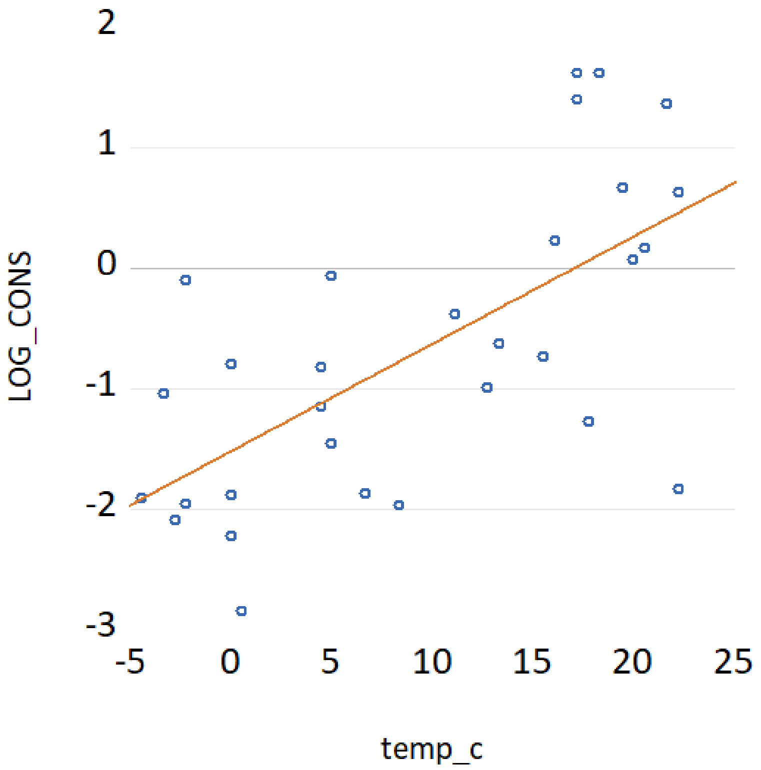

Figure 11: Case Study 1

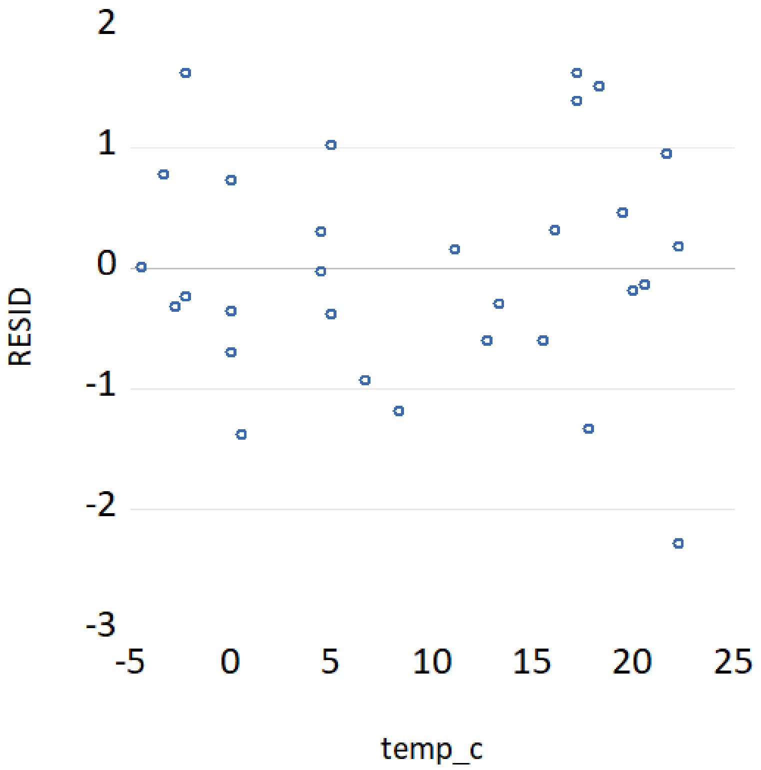

Figure 12: Case Study 1

B.2 Case Study 2

Dependent Variable: COOLLOAD
Method: Least Squares
Date: 10/31/19 Time: 10:26
Sample: 1 768
Included observations: 768

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	279.9703	14.74622	18.98590	0.0000
RELCOMP	-1.541746	0.093561	-16.47844	0.0000
WALLAREA	-0.145381	0.011745	-12.37822	0.0000
ROOFAREA	-0.536435	0.022432	-23.91402	0.0000
GLAZING	0.148180	0.009406	15.75428	0.0000

R-squared	0.867632	Mean dependent var	24.58776
Adjusted R-squared	0.866938	S.D. dependent var	9.513306
S.E. of regression	3.470237	Akaike info criterion	5.332812
Sum squared resid	9188.460	Schwarz criterion	5.363045
Log likelihood	-2042.800	Hannan-Quinn criter.	5.344448
F-statistic	1250.303	Durbin-Watson stat	1.072638
Prob(F-statistic)	0.000000		

Figure 13: Case Study 2 Output