Assignment-2-CH5440

CH18B020

March 13, 2022

Question-1)

Part a)

We have 4 independent variables: concentrations of CO_2, CH_4, N_2O, O_3. And, we have one independent variable: $T_{avg,\ deviation}$. We can fit a model of the form $X\beta + \beta_0 = y$ where we can estimate the parameters as,

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{\beta}_0 = \bar{y} - \bar{x}^T \hat{\beta}$

The obtained model is given as:

$$T_{deviation} = 11.8 + 0.0607x_1 + 0.00591x_2 - 0.14652x_3 + 0.00804x_4 \tag{1}$$

where,

1. x_1 is concentration of CO_2
2. x_2 is concentration of CH_4
3. x_3 is concentration of N_2O
4. x_4 is concentration of O_3

The temperature deviation is positively correlated with the concentration of all gases other than N_2O, for which it is negatively correlated. This is unexpected because the correlation estimate between that and temperature deviation turns out to be positive (approx. 0.88). The coefficient values were also verified using fitlm() function in MATLAB.

Part b)

Confidence intervals:

Term	Lower Bound	Estimate	Upper Bound
Intercept	-0.2598	11.7998	23.8594
x_1	0.0336	0.0607	0.0878
x_2	0.0039	0.0059	0.0079
x_3	-0.2208	-0.1465	-0.0722
x_4	-0.0025	0.0080	0.0186

If we keep the bound for the residuals as 2, we dont see any residual greater than that value as seen in figure 1. However, the residual corresponding to the 13th data point is somewhat off positioned.

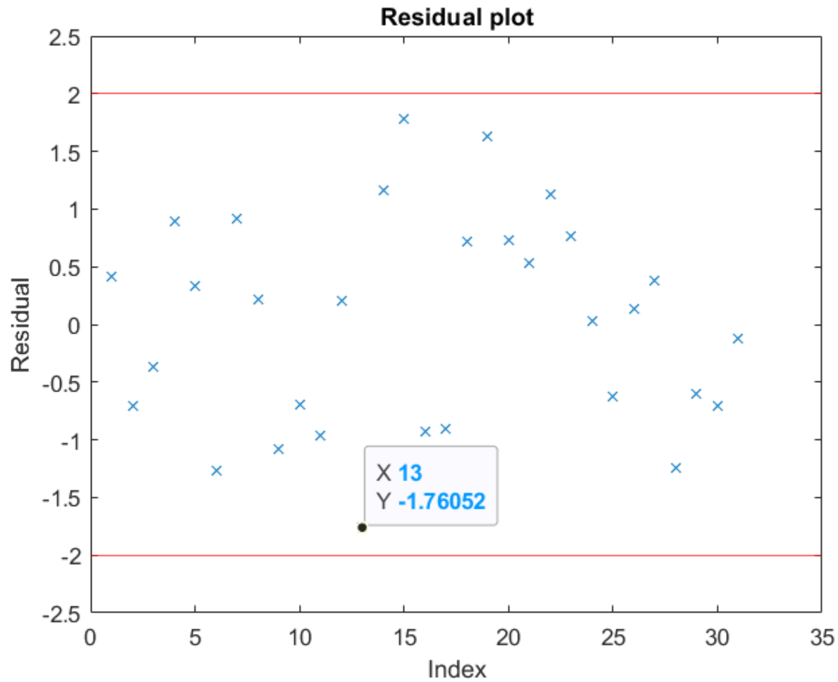

Figure 1: Residual plot for full data

Removing that point and retraining, the residuals are better now.

$$\hat{\beta_{new}} = [0.0631, 0.0066, -0.1573, 0.0086]^T \tag{2}$$

$$\hat{\beta}_{0,new} = 11.8075 \tag{3}$$

No outliers seen in the residuals of the new model.

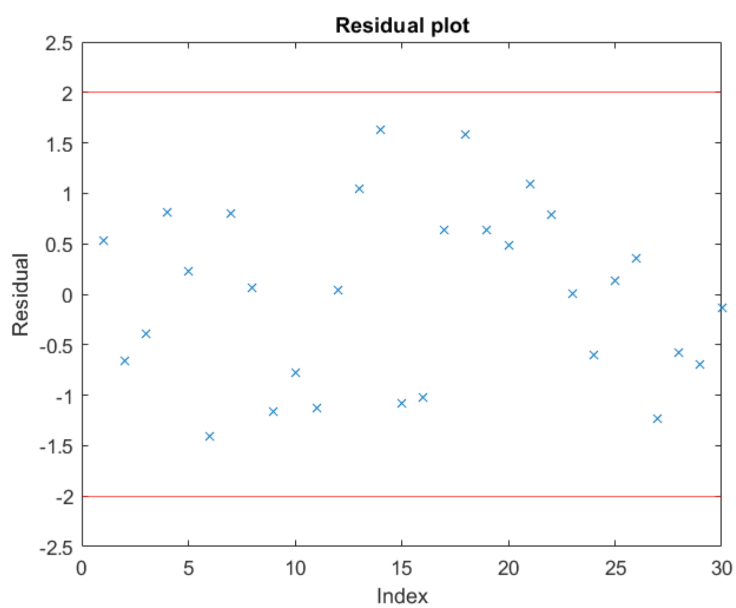

Figure 2: Residual plot after outlier removal

Part c)

fitlm() function is used to find out the pValues and standard errors.

```
mdl_c =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate        SE         tStat       pValue  
                   _________    _________    _______    _________

    (Intercept)        12.95        11.46       1.13      0.26919
    x1              0.063059     0.025766     2.4473      0.02176
    x2             0.0066088    0.0019334     3.4183    0.0021657
    x3              -0.15733      0.07074    -2.2241     0.035407
    x4             0.0086162     0.010032    0.85883      0.39859


Number of observations: 30, Error degrees of freedom: 25
Root Mean Squared Error: 0.122
R-squared: 0.858,  Adjusted R-Squared: 0.835
F-statistic vs. constant model: 37.7, p-value = 3.02e-10
```

We can see that Pvalue for concentration of Ozone coefficient is high (> 0.05) meaning it is insignificant. So we drop that first and rebuild a model.

```
mdl_c1 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate        SE         tStat       pValue  
                   _________    _________    _______    _________

    (Intercept)       17.266       10.247     1.6849      0.10397
    x1              0.067786     0.025044     2.7067     0.011845
    x2             0.0064572    0.0019155     3.3709    0.0023508
    x3              -0.16813     0.069261    -2.4274     0.022435


Number of observations: 30, Error degrees of freedom: 26
Root Mean Squared Error: 0.122
R-squared: 0.854,  Adjusted R-Squared: 0.837
F-statistic vs. constant model: 50.5, p-value = 5.51e-11
```

We can see that Pvalue for intercept is high (> 0.05) meaning it is insignificant. So we drop that and rebuild the model.

```
mdl_c2 =


Linear regression model:
    y ~ x1 + x2 + x3
```

```
Estimated Coefficients:
          Estimate        SE          tStat       pValue
          _________    _________    _______    __________

    x1     0.025869    0.0029784     8.6857    2.6637e-09
    x2    0.0043522    0.0015007     2.9001     0.0073299
    x3    -0.052118    0.0077743    -6.7039    3.3882e-07


Number of observations: 30, Error degrees of freedom: 27
Root Mean Squared Error: 0.126
```

We can see that Pvalue for all coefficients is very low ($<< 0.05$) which means all coefficients are highly significant.

Part d)

GWP of the gases is simply ratio of its regression coefficient (adjusted for units) with the regression coefficient of CO_2.

1. GWP of $CO_2 = 1$ (by definition)
2. GWP of $CH_4 = \frac{0.0059}{0.0607} * 10^3 = 97.405$
3. GWP of $N_2O = \frac{-0.1465}{0.0607} * 10^3 = -2413.3$

We observe that CH_4 GWP is close to the values observed over a 20 year horizon, but N_2O is not.

Question-2)

Part a)

Considering,

$$y = ln(P^{sat}) \tag{4}$$

$$x = \frac{1}{T} \tag{5}$$

we perform OLS and obtain,

$$A' = 4.7607 \tag{6}$$

$$B' = -37.896 \tag{7}$$

Part b)

The optimization problem is set similar to OLS since measurements of y is noise-free, but of course, it is nonlinear in this case.

$$\min_{\hat{P}_1^{sat},..,\hat{P}_{100}^{sat},A,B,C} \quad \sum_{i=1}^{100} (P_i^{sat} - \hat{P}_i^{sat})^2 \tag{8}$$

$$\text{s.t.} \quad ln(\hat{P}_i^{sat}) = A - \frac{B}{T_i + C} \tag{9}$$

$$\tag{10}$$

We can eliminate the equality constraint by substituting P_i^{sat} back in the objective. After doing so, using lsqnonlin() to solve the problem, one obtains,

$$A = 14.1018 \tag{11}$$

$$B = 2821.4489 \tag{12}$$

$$C = 228.7554 \tag{13}$$

Part c)

We are given that,

$$\sigma_{\epsilon_x} = 0.18 \tag{14}$$
$$\sigma_{\epsilon_y} = 2 \tag{15}$$

So we can use these values to set up a WTLS style optimization problem as given below.

$$\min_{\hat{P}_1^{sat},..,\hat{P}_{100}^{sat},\hat{T}_1,...,\hat{T_{100}},A,B,C} \quad \sum_{i=1}^{100} \frac{(P_i^{sat} - \hat{P}_i^{sat})^2}{\sigma_{\epsilon_y}^2} + \frac{(T_i - \hat{T}_i)^2}{\sigma_{\epsilon_x}^2} \tag{16}$$
$$\text{s.t.} \quad ln(P_i^{sat}) = A - \frac{B}{T_i + C} \tag{17}$$
$$\tag{18}$$

To improve convergence, we have the initial guess for A,B,C as the solution of part b) (OLS problem), and the initial guess for temperature as the temperature measurement. We obtain the following estimates,

$$A = 14.1217 \tag{19}$$
$$B = 2835.2165 \tag{20}$$
$$C = 229.4130 \tag{21}$$

Part d)

Listed below are the maximum absolute error in each case.

1. Part a): 39.8234
2. Part b): 4.4844
3. Part c): 4.2112