Coursework Submission

University ID: 10667974

November 20, 2022

Report for MATH38141 - Regression Analysis

Question 1.

Cheddar cheese from Australia was analyzed for its chemical composition. The following data is of $n = 30$ observations, each with 4 variables:

- Taste - Subjective test score
- Acetic and H2S - natural log of the concentration of acetic acid and H2S
- Lactic Acid - concentration of lactic acid

a) Scatterplots of Taste Against the Three Explanatory Variables

The first step we'll take in analysing the data, is to plot three scatterplots, one for each explanatory variable. The plots will guide us, when making conclusions about any patterns in the data.

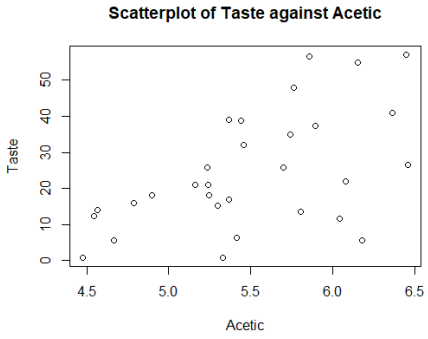

Figure 1: Scatterplot of Taste against Acetic Acid

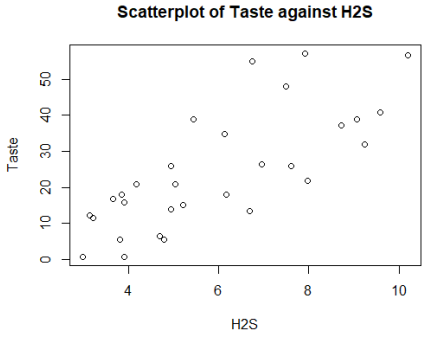

Figure 2: Scatterplot of Taste against H2S

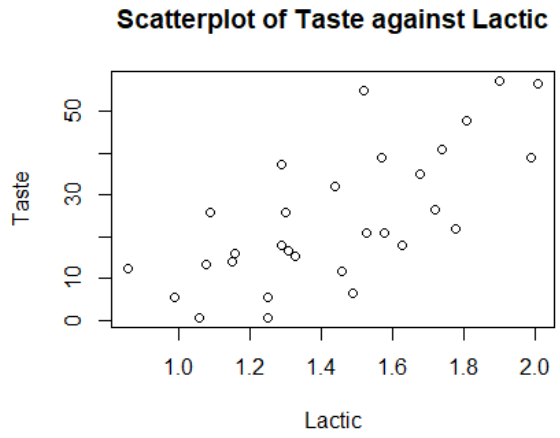

Figure 3: Scatterplot of Taste against Lactic Acid

Now that we've plotted the data, we can start by examining the plots for any signs of linearity. From Figures 2 and 3 we can outright spot a positive linear relationship between **taste** and **H2S**, as well as **taste** and **lactic acid**. When inspecting figure 1, we can note that there is a vague relationship between **taste** and **acidic acid**, but it is not as prevalent as in the other two plots. In Figure 1, most of the data is clustered in between in the range of $[5.2, 6]$, while more than half of the data in Figure 2 is concentrated in the range of $[0, 6]$.

b) Formulatting the Multiple Linear Regression Model

Now that we've seen some positive linear relationships between **taste** and the other three variables, we'll construct a multiple linear regression model, with **taste** being our response and the latter three variables- our regressors:

- Taste - dependent variable
- Acetic Acid, H2S, Lactic Acid - regressors ($\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}$ in vector notation)

Our model will look as such: $Taste = \beta_0 + \beta_1 Acetic + \beta_2 H2S + \beta_3 Lactic + \varepsilon$

Each **taste** observation is a linear combination of the other three variables plus an error term:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

We'll proceed to use the vector notation, hence we'll represent it as: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$.

- $\boldsymbol{Y}$ is a vector of responses, with a of length $n \times 1$
- $\boldsymbol{X}$ is a $n \times (p+1)$ design matrix, where p is the number of parameters[1]
- $\boldsymbol{\beta}$ is a vector of coefficients of length $(p+1) \times 1$
- $\boldsymbol{\epsilon}$ is a vector of length $n \times 1$, containing the error terms

[1]In our case, p equals 3. In this report, we will follow the notation in the notes and note $d = p + 1$

The error terms in the model are assumed to be normally distributed with a parameters: $\mu = 0$ and constant variance σ^2. The fitted model does not include error terms and is formulated as follows: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

c) Derivation of LSEs and their respective 95% Confidence Intervals for the Regression Coefficients

To derive the *least squares estimate* of the coefficients vector, we need to compute the following, as per the formula in section 3.2.2 of the written notes: $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$

We start of by defining our design matrix and response vector in R as:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 4.543 & 3.135 & 0.86 \\ 1 & 5.159 & 5.043 & 1.53 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.176 & 4.787 & 1.25 \end{bmatrix} \text{ and our response vector } \boldsymbol{Y} = \begin{bmatrix} 12.3 \\ 20.9 \\ \vdots \\ 5.5 \end{bmatrix}$$

Now by matrix multiplication, we proceed to compute both $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$ and $\boldsymbol{X}^\top \boldsymbol{Y}$, with the matrix $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$ being equal to:

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \begin{bmatrix} 3.79501315 & \vdots & \vdots & \vdots \\ \vdots & 0.19379516 & \vdots & \vdots \\ \vdots & \vdots & 0.01518620 & \vdots \\ \vdots & \vdots & \vdots & 0.72551612 \end{bmatrix}$$

Finally, we compute the vector $\hat{\boldsymbol{\beta}}$ as:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} -28.8767696 \\ 0.3277413 \\ 3.9118411 \\ 19.6705434 \end{bmatrix}$$

From the results, $\hat{\beta}_0$ is the intercept, while $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are the estimated slopes of the each of the explanatory variables. To construct a confidence interval, we'll first note that under the GLM, the vector $\hat{\boldsymbol{\beta}}$ is Normally distributed as $\hat{\boldsymbol{\beta}} \sim N_d\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$. Further, we can pick a vector $\boldsymbol{c} = [c_1, c_2, c_3, c_4]^\top$ with ones and zeroes to single out positions in the coefficients vector:

- Let $\boldsymbol{c_0} = [1, 0, 0, 0]^\top$, $\boldsymbol{c_1} = [0, 1, 0, 0]^\top$, $\boldsymbol{c_2} = [0, 0, 1, 0]^\top$ and $\boldsymbol{c_3} = [0, 0, 0, 1]^\top$

We'll use the following formula from section 3.4.3 of the notes to compute the 95% CI:

$$\hat{\ell} \pm \text{ s.e. } (\hat{\ell} - \ell) t_{n-d}^{(\alpha/2)}. \tag{1}$$

Where $\hat{\ell} = \boldsymbol{c}\hat{\boldsymbol{\beta}}$ and the standard error is defined as:

$$\text{s.e. } (\hat{\ell} - \ell) = s\sqrt{\boldsymbol{c}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{c}}$$

And our critical value, with 26 degrees of freedom at the 5% significance level is:

$$t_{26}^{0.025} = 2.055529 \tag{2}$$

The variance term σ^2 has an unbiased estimator s^2, which is defined as $s^2 = \frac{\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}}}{(n-d)}$. We compute the error as $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. After computation in R, we get that $s^2 = 102.6312$. Hence, the 95% Confidence Interval for $\hat{\beta}_0$ is given by the formula $\boldsymbol{c_0}\hat{\boldsymbol{\beta}} \pm t_{26}^{0.025} s\sqrt{\boldsymbol{c_0}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{c_0}}$ is [-69.44350, 11.68996].

We repeat this process for $i = 1, 2, 3$ $\hat{\beta}_i = \boldsymbol{c_i}\hat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{c_i}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{c_i}}$ to reproduce the following results:

We have the following 95% confidence intervals

- $\hat{\beta}_1 \in [-8.839420, 9.494902]$
- $\hat{\beta}_2 \in [1.345656, 6.478026]$
- $\hat{\beta}_3 \in [1.933267, 37.407820]$

An important observation to make here is that the both intervals for $\hat{\beta}_2$ and $\hat{\beta}_3$ do not include 0, while $\hat{\beta}_0$ and $\hat{\beta}_1$ do include 0, hence we cannot reject the possibility of these two coefficients taking this value.

d) Interpretation of the Estimated Coefficients

As mentioned, the coefficient $\hat{\beta}_0$ represents the estimate of the intercept and the latter three estimates are of the slopes of each explanatory variable. Except the intercept, the latter three all have a positive sign, which correlates with statements made in **a)**, about the positive linear relationship present in the scatterplots. The coefficient $\hat{\beta}_3$ is the most influential of the three, consequently we would expect to see a greater shift in the response, given a single unit of change in a variable in $\boldsymbol{X_3}$ (*lactic acid*), compared to say a single unit of change in a variable from $\boldsymbol{X_1}$.

Running the command **summary()** on a linear model in R will output a table of data, with one of the columns being the associated p-value of a given variable. The p-value of H2S is 0.00425, which indicates it is highly significant, in contrast to the p-value of Acetic, that is 0.94198, which is above the threshold- signifying it doesn't contribute much to the model.

e) Determining the Fit of the Linear Model

The R2 statistic is the *coefficient of determination* and the formula for it is the explained variance over the total variance. The statistic gives an idea of how good a fit is our proposed multiple linear model is to the data. To calculate R2, we'll first output the SSE and then proceed to output S_{yy}.

$$SSE = \hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}} = 2668.411 \tag{3}$$

$$S_{yy} = \boldsymbol{Y}^\top \boldsymbol{Y} - n\bar{y} = 7662.887 \tag{4}$$

We know that $SSR = S_{yy} - SSE$, hence:

$$SSR = S_{yy} - SSE = 4994.476 \tag{5}$$

From equations (4) and (5):

$$R^2 = \frac{SSR}{S_{yy}} = 0.6517747 \tag{6}$$

The R2 statistic suggests that the multiple linear model fits the data well, as it does a moderate job, of explaining 65.17%, or about ⅔-rds, of the total variance.

f) Should the Intercept be set to zero?

By setting the intercept ($\hat{\beta}_0$) to 0, we state at when the explanatory variables are 0, then *Taste* should also be zero. Intuitively, this make sense, as when we strip a chemical off of its components, then it ought to have no taste. To support this argument, we can refer back to our findings in **c)**, where we found the 95% confidence interval of $\hat{\beta}_0$, [-69.44350, 11.68996], to include zero in the interval. Therefore, it is reasonable to have the intercept set to 0.

g) Reduced model and ANOVA table

Model	Explanatory Variables
Full Model Ω	Acetic, H2S, Lactic
Reduced Model ω	H2S, Lactic

Table 1: Full model Ω and Reduced Model ω with their variables

Let our reduced model be[1]: $Taste = \beta_0 + \beta_1 H2S + \beta_2 Lactic + \varepsilon$

With each error term again being normally distributed with $N \sim \left(0, \sigma^2\right)$

Now that we've stated the reduced model, we can continue filling the table with our values. Because our response and full model are in both cases the same, we can reuse our values for $S_{yy} = 7662.887$ and $SSE_\Omega = 2668.411$. To calculate the $SSEXT$, we make use of the following formula: $SSEXT = SSE_\omega - SEE_\Omega$. SSE_ω is found by the same derivation process as in **b)**, by computing $\hat{\boldsymbol{y}}$ and subsequently finding the error ($\hat{\boldsymbol{\varepsilon}}$) of the reduced model.

$$SSEXT = SSE_\omega - SEE_\Omega = 2668.965 - 2668.411 = 0.5542675 \tag{7}$$

The degrees of freedom for the Full model Ω is 26, calculated as:
d.f Residual Fitting Ω = Number of Observations - (Number of Total Parameters + 1)

In our reduced model, we have kept 2 parameters, hence the degrees of freedom, corresponding to the Regression fitting ω is 2, while the degrees of freedom for Extra is the number of "dropped" parameters:

d.f Extra = Number of Total Parameters - Number of Kept Parameters $= 3 - 2 = 1$

[1] In vector notation the model is: $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X_2} + \beta_2 \boldsymbol{X_3} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N\left(0, \sigma^2 I_n\right)$

Having enstablished our degrees of freedom, we can compute $MSEXT$ and MSE_Ω, with the corresponding formulae:

$$MSEXT = \frac{SSEXT}{\text{d.f Extra}} = \frac{0.5542675}{1} = 0.5542675 \tag{8}$$

$$MSE_\Omega = \frac{SSE_\Omega}{\text{d.f Residual Fitting } \Omega} = \frac{2668.411}{26} = 102.6312 \tag{9}$$

Following the results from equations (8) and (9), we can now compute the F-ratio as:

$$F = \frac{MSEXT}{MSE_\Omega} = 0.005400575 \tag{10}$$

Combining all results from the above paragraph, we can now fill out the summarised ANOVA table:

Source	s.s	d.f	m.s	F-ratio
Regression fitting reduced model	4993.921	2	-	-
Extra	0.5542675	1	0.5542675	0.005400575
Residual fitting full model	2668.411	26	102.6312	-
Total	7662.887	29	-	-

Table 2: Summarised ANOVA table

h) Does Acetic Acid warrant inclusion?

To answer, we can do a formal test. With the F-ratio and degrees of freedom from **g)**, we can compute a p-value from an F-Test for the coefficient β_1 using the following formula:

$$p\text{-value} = \mathrm{P}\left(F_{\text{p-k, n-p-1}} > F\right) \tag{11}$$

Let:

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

Where we reject H_0 if and only if the p-value of equation (11) is smaller than significance level.

The degrees of freedom for F are taken as 1 and 26, because $p = 3$ and $k = 2$, while $n = 30$. Directly substituting into equation (11):

$$p\text{-value} = \mathrm{P}\left(F_{1,26} > 0.005400575\right) = 1 - \mathrm{P}\left(F_{1,26} \leq 0.005400575\right) = 0.9419798$$

As the p-value here is greater than 0.05, we cannot deny the option of $\beta_1 = 0$. Moreover, the coefficient is even greater than 0.9, which means the regressor is not significant enough, hence we can remove it from our model.

i) Regressing Taste on Acetic alone

Our simple linear model is: $Taste = \beta_0 + \beta_1 Acetic + \varepsilon$. We can formulate a similar test, as we did in h):

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

Rejecting H_0 if and only if the p-value of the F-test $\mathrm{P}\left(F_{1,\text{ n-2}} > F\right) < \alpha$, where α is significance level.

Source	s.s	d.f	m.s	F-ratio
Regression	2314.142	1	2314.142	12.11424
Residual	5348.745	28	191.0266	
Total	7662.887	29		

Table 3: Summarised ANOVA table for the Simple Linear Regression Model

Using the table, we compute:

- p-value $= \mathrm{P}\left(F_{1,28} > 12.11424\right) = 1 - \mathrm{P}\left(F_{1,28} \leq 12.11424\right) = 0.001658192$

At the 5% significance level, we reject H_0, because our p-value is less than 0.05.

Observe, here we reject the possibility of the coefficient β_1 being 0. Hence in the SLR model, β_1 is significant in explaining the variance. Although it contradicts our findings in part **h)**, where we found β_1 to be insignificant, hence it can take the value of 0, it must be mentioned that β_1 is not informative in the case, where the other two explanatory variables, *H2S* and *Lactic Acid*, are already present in the model.

To support the above calculation, the formulae used:

$$S_{xx} = \sum_{i=1}^{i=30} x_i^2 - n\bar{x}^2 = 9.451161 \tag{12}$$

$$S_{xy} = \sum_{i=1}^{i=30} (x_i - \bar{x})(y_i - \bar{y}) = 147.8896 \tag{13}$$

$$SSR = \frac{S_{xy}^2}{S_{xx}} = 2314.142 = MSR = \frac{SSR}{1} \tag{14}$$

$$MSE = \frac{SSE}{n-2} = \frac{5348.745}{30-2} = 191.0266 \tag{15}$$

$$F = \frac{MSR}{MSE} = 12.11424 \tag{16}$$

Question 2.

We're given a dataset from years 1925 till 1941[2], which includes 7 variables. From the captured data, we're to make a *MLR* model Ω, regressing the following six variables on PBE:

- YEAR - Year of observation
- PFO, DINC, CFO, RDINC, RFP - Indices of retail food price, disposable income per capita, food consumption per capita, real consumption per capita and retail food price, adjusted for CPI
- PBE - Price of beef (in cents per lb).

Along with an SLR model, regressing PBE on CFO alone.

a) Stating the Full Model Ω and Reduced Model ω, with Relevant Assumptions

Model	Explanatory Variables
Full Model Ω	YEAR, PFO, DINC, CFO, RDINC, RFP
Reduced Model ω	CFO

Table 4: Proposed MLR Full model Ω and SLR ω with their variables

Our full model is:

$$PBE = \beta_0 + \beta_1 YEAR + \beta_2 PFO + \beta_3 DINC + \beta_4 CFO + \beta_5 RDINC + \beta_6 RFP + \varepsilon \quad (17)$$

And our reduced model is:

$$PBE = \beta_0 + \beta_1 CFO + \varepsilon \quad (18)$$

Our assumptions for both models are:

- There exists a linear relationship between PBE and our corresponding regressors.
- Errors are normally distributed, with zero mean, and constant σ^2 variance and are independent of each other. Further, that the error is independent of $\hat{y}$.
- Observations are independent from one another. Additionally, the regressors are independent of each other.

  For our MLR model, we also assume that there is no multicollinearity.

For our calculations, we'll again use the vector approach. For our full model Ω, we formulate the design matrix $\boldsymbol{X}$, with the first column being ones and each of the other 6 columns- the 6 explanatory variables. $\boldsymbol{Y}$ is our response vector, the variable PBE. Hence, $\boldsymbol{X}$ is n by d matrix. $\boldsymbol{Y}$ is n by 1 vector. $\boldsymbol{\beta}$ is of length d by 1 and $\boldsymbol{\varepsilon}$ is a vector of length n by 1.

[2]This implies the number of observations is $n = 17$

b) Calculating the Unexplained Variation of the Two Models

Using the same vector approach as in Question 01, we formulate the design matrix $\boldsymbol{X}$ as mentioned in **a)**, along with a vector $\boldsymbol{Y}$, with the response entries. Hence, we compute $\hat{\boldsymbol{\beta}}$ and then the error vector to find SSE_Ω:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\Omega}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \begin{bmatrix} -3.621417e+03 \\ 2.120140e+00 \\ \vdots \\ -3.562661e-02 \end{bmatrix} \text{ and } \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\Omega}} = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}}_{\boldsymbol{\Omega}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} \tag{19}$$

$$SSE_\Omega = \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\Omega}}^\top \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\Omega}} = 132.8499 \tag{20}$$

Let $\boldsymbol{W}$ be a 2 by n matrix, with the first column ones and the latter one the variable CFO:

$$\boldsymbol{W} = \begin{bmatrix} 1 & 90.9 \\ 1 & 92.1 \\ \vdots & \vdots \\ 1 & 97.5 \end{bmatrix} \tag{21}$$

Hence, by following the same calculation as in equation 19,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\omega}} = \left(\boldsymbol{W}^\top \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{Y} = \begin{bmatrix} 254.750111 \\ -2.040432 \end{bmatrix} \tag{22}$$

$$SSE_\omega = \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\omega}}^\top \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\omega}} = 307.7355 \tag{23}$$

c) Why is SSE_ω larger than SSE_Ω?

With the inclusion of more parameters, we can explain more and more of the variance. It's a trade-off. When adding the parameters, we "fine-tune" our model to the current data, in turn explaining the variance in it, but we run the risk of overfitting it[3]. The formula is $S_{yy} = SSE + SSR$ and S_{yy} is constant in both models. Hence, by adding more parameters to "explain" the data- we in turn increase our SSR, which decreases our SSE. Therefore, a model with less parameters will have a greater SSE.

d) Can the coefficient of DINC be equal to 2?

We can do a t test and calculate the test statistic T and check its value against the critical value. Let:

$$H_0 : \beta_4 = 2 \text{ vs } H_A : \beta_4 \neq 2$$

Further, define $\boldsymbol{c_4} = [0, 0, 0, 1, 0, 0, 0]^\top$ and $\hat{\ell_4} = \boldsymbol{c_4}\hat{\boldsymbol{\beta}}$. We'll reject the null hypothesis, if:

$$T = \left| \frac{\hat{\ell_4} - 2}{s.e(\hat{\ell} - \ell)} \right| > t_{n-d}^{(\alpha/2)} \tag{24}$$

Standard error is defined as:

$$s.e(\hat{\ell_4} - \ell_4) = \sqrt{s^2 \boldsymbol{c_4}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{c_4}}$$

and our given $\alpha = 0.10$.

[3]Overfitting means we can explain our current data, but we'll be not be accurate when predicitng future data.

Under our null hypothesis, $\frac{\hat{\ell_4}-2}{\sqrt{s^2 \boldsymbol{c_4}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c_4}}} \sim t_{n-d}$, which is why we can compare the t-score to the critical value.[4]

Using equation (24), we get $T = |0.06195476|$, which is not greater than our critical value of $t_{10}^{0.05} = 1.812461$. Therefore, we cannot reject H_0, i.e we cannot reject $\beta_4 = 2$.

e) Calculating the prediction interval of the change of PBE in 2015 to 2018

To calculate the prediction interval, we'll make use of the following formula:

$$(\boldsymbol{x_1^\star} - \boldsymbol{x_2^\star})^\top \hat{\boldsymbol{\beta}} \pm t_{n-d}^{(\alpha/2)} s \sqrt{(\boldsymbol{x_1^\star} - \boldsymbol{x_2^\star})^\top (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{x_1^\star} - \boldsymbol{x_2^\star})} \tag{25}$$

Let $\boldsymbol{x_1^\star} = (1, 2015, 200, 200, 200, 220, 2000)$ and $\boldsymbol{x_2^\star} = (1, 2018, 190, 210, 210, 210, 2100)$. Hence, using year 2015 as our basis year, the CI's are as follows:

- At the 10% significance level, the prediction interval is: $[-252.2624, 247.4954]$
- At the 5% significance level, the prediction interval is: $[-309.5707, 304.8037]$
- At the 1% significance level, the prediction interval is: $[-439.3221, 434.5551]$

  Note: $t_{10}^{0.05} = 1.812461$, $t_{10}^{0.025} = 2.228139$, $t_{10}^{0.005} = 3.169273$

f) Regressing PBE on YEAR and CFO

(1) Model Formulation:

We can extend the model in **a)**, by including the variable YEAR as a regressor, along with CFO and a third term- the combination of both Year and CFO. Hence, we'll create a MLR model with PBE as the response and YEAR, CFO and YEAR*CFO as the explanatory variables. Hence, satisfying PBE and CFO being dependant on YEAR. Further, satisfying the condition of nested models, by having CFO appear in both models.

Model	Explanatory Variables
Model Ω_1	YEAR, CFO, YEAR*CFO
Reduced Model ω	CFO

Table 5: Proposed MLR Model Ω_1 and SLR ω with their regressors

Model Ω_1:

$$PBE = \beta_0 + \beta_1 YEAR + \beta_2 CFO + \beta_{12} YEAR * CFO + \varepsilon$$

(2) Comparing the two models: To compare the two models, we can construct an ANOVA table and use an f-test for the hypothesis.

[4] In the notation, we again denote d as the number of parameters + 1, $d = p + 1$

Source	s.s	d.f	m.s	F-ratio
Regression fitting reduced model ω	457.92	1	-	-
Extra	114.45	2	57.225	3.849
Residual fitting full model Ω_1	193.29	13	14.868	-
Total	765.66	16	-	-

Table 6: Summarised ANOVA table for the models Ω_1 and ω

We test:

$$H_0 : \beta_1, \beta_{12} = 0 \text{ vs } H_A : \beta_1, \beta_{12} \neq 0$$

$$p\text{-value} = \mathrm{P}\left(F_{2,13} > 3.849\right) = 1 - \mathrm{P}\left(F_{2,13} \leq 3.849\right) = 0.04866$$

The observed p-value is less than 0.05, therefore we reject the null hypothesis. Including the variables YEAR and YEAR*CFO is a good addition to our original SLR model.

(3) Graphs of the relation between PBE and CFO, in the years 1925, 1930, 1935 and 1940:

The vector containing the coefficients for the model Ω_1 is $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Omega_1}} = \begin{bmatrix} 7316.16363 \\ -3.62330 \\ -90.94968 \\ 0.04567 \end{bmatrix}$

We're asked to plot four fitted regression lines for specific years. When we constrain the variable YEAR in Model Ω_1, it becomes an SLR model, with an intercept and a variable. Hence, we can easily plot it in R:

$$\hat{PBE}_{\Omega_1} = \underbrace{7316.16363 - 3.62330 * YEAR}_{intercept} \overbrace{-90.94968 * CFO + 0.04567 * (YEAR \times CFO)}^{slope} \tag{26}$$

Using equation (26), we can substitute in the the YEAR to be 1925, 1930, 1935 and 1940. Following that, we can produce the following graph:

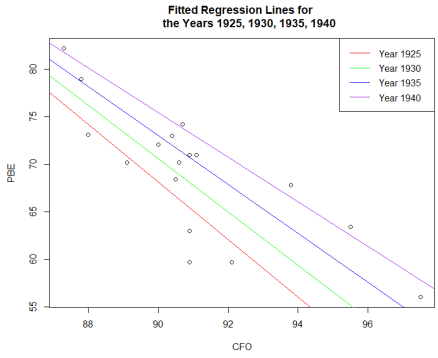

Figure 4: Fitted Regression Lines for the Years 1925, 1930, 1935, 1940

From the plot, in the peaceful years before the Second World War, we can observe a very persistent pattern of increases in prices of beef throughout. If we take note of a point on the plot, say $x = 91$, and draw lines:

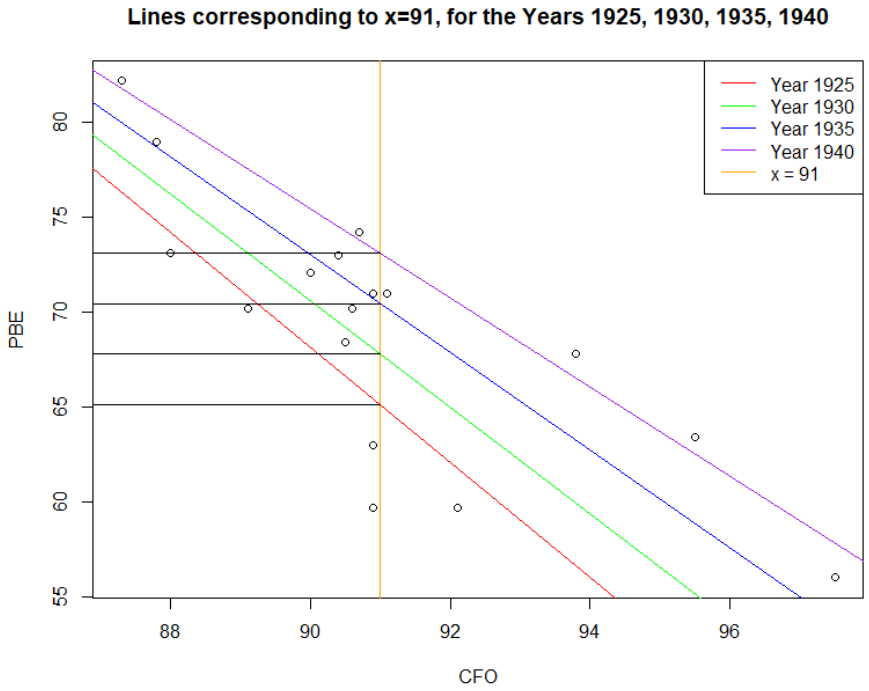

Figure 5: Lines corresponding to x=91, for the Years 1925, 1930, 1935, 1940

We can conclude that for the same level of consumption, one must pay more money for the same amount of beef, compared to 5 years prior. As an example, for around 65 cents, we could have bought a pound of beef in the year of 1925. Yet, an equal amount of beef in the year of 1935, would have costed us around 70 cents.

Formulae used to draw lines for respective years in Figure 4, when substituting into equation (26):

$$\hat{y}_{1925} = 341.3111 - 3.03493 * CFO$$
$$\hat{y}_{1930} = 323.1946 - 2.80658 * CFO$$
$$\hat{y}_{1935} = 305.0781 - 2.57823 * CFO$$
$$\hat{y}_{1940} = 286.9616 - 2.34988 * CFO$$