(d) Choice between full model (a) and Reduced model:

In Full model, adjusted $R^2 = 0.9519$ (model in 4(a))

In Reduced model, adjusted $R^2 = 0.9553$ (model in 4c)

Also, we perform Anova Analysis (in R code) the reduced model have the slightly better adjusted R^2 than the full model.

we use adjusted R^2 as it slightly give better accuracy in terms of model selection.

Also, The Residual Standard error for full model = 0.7035, and residual standard error for reduced model = 0.6783.

The reduce model have the smaller standard error than the full model.

Analyzing All this evidence, I choose reduce model as a better choice

(e) Testing

$$H_0 : \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0 \qquad H_1 : \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \neq 0$$

Here the null hypothesis describe that the coefficient $\beta_1, \beta_2, \beta_3$ will be 0 at a time.

we will test this for two model

Case- (i) Restricted model: ($\beta_1, \beta_2, \beta_3$ will be 0)

is,

$$y_i = \beta_0 + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Case- (ii) Full model (4(a))

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Case -(i) Restricted model:

$$y_i = \beta_0 + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Using R we perform regression analysis, of restricted model.

Case (ii) - Full model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

we also fit this using regression model.

Anova table for this two model

Model 1 : $y \sim x1 + x2 + x3 + x4 + x5$

Model 2 : $y \sim x4 + x5$

Res.DF	RSS	DF	Sum of sq	F	Pr(>F)
16	7.9175				
19	8.9793	-3	-1.0617	0.7152	0.5572

(ii) For second model:

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

From Anova table we see that, P value is 0.5572 which is greater than 0.05 (if we consider 95% significance level). Also, F statistics is 0.7152.

Hence, we don't have enough evidence to reject the null hypothesis.

Overall conclusion: In problem 4, we develop multiple regression model with full model in (a). In 4(a) the residual plot is not fully perfect as we see some high positive and negative y axis (residual) value. In (b) we show the 95% CI & PI. and as usual PI tells where we can see the next data point sampled. In 4(c) a new model is develop with x_2, x_4, x_5. And we choose new reduced model as better choice as it have smaller residual standard error and higher adjusted R^2. Finally we test the hypothesis using Anova approach and we don't able to reject the null hypothesis.

problem 5:

Solution:

(a) Simple linear regression:

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i$$

where y = Analytical Results.

x = quantities of Calcium in carefully prepared solution.

Where $\hat{\beta}_0 = -0.228090$

$\hat{\beta}_1 = 0.994757$ [Details are in R code]

Assumption of linear regression:

(1) Regressor Variable $x_1, x_2, \cdots x_n$ are not random Variables.

(2) $\epsilon_1, \epsilon_2, \cdots \epsilon_n$ are random Variable

and $E(\epsilon_i) = 0$, $i = 1, 2, \cdots n$

(3) $V(\epsilon_i) = \sigma^2$ is constant for all $i = 1, 2, \cdots n$

This means the variance $V(y_i) = \sigma^2$ are all the same and all observation have the same precision.

(4) ε_i and ε_j different errors, hence response y_i & y_j are independent.

This means $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

Also, response Variable (y_i) drawn from probability distribution with means $\mu_i = E(y_i) = \beta_0 + \beta_1 x_i$ for constant variance σ^2

Besides, two observations y_i and y_j $(i \neq j)$ are independent.

(b) <u>95% CI for the intercept</u>:

95% confidence for the intercept is: $(-0.5459503, 0.08977054)$

[R code is given]

This means 95% confident interval is a range of value that we are 95% confident it contains the true value of the intercept.

5 (c) <u>95% Confidence Interval of the Slope</u> :

95% CI for slope = (0.9827204, 1.006793)

This also mean, 95% CI for slope gives us range of value that we are 95% confident it contains the true value of the slope.

<u>5(d)</u>: (i) when $x=0$, then $y=0$ if there is no calcium present, our technique should not find any :

we are looking for whether the intercept is zero or not.

From (a) we find,

$$\hat{\beta}_0 = -0.228090$$

with t value $= -1.655$

p value $= 0.137 > 0.05$

Also, From this we can say that we have not enough evidence to reject the null hypothesis ($\hat{\beta}_0 = 0$).

Also, From 5(b) we see that confidence interval of the intercept is $(-0.5459503, 0.08977054)$

which means that the CI contains the value 0.

So, we can conclude that if $x=0$ then $y=0$ as

the intercept close to 0.

(ii) Here its say that if the empirical technique

is any good then, the slope of regression $=1$.

Let us formulate a hypothesis.

$[\beta_1$ is slope$]$

$$H_0 : \beta_1 = 1$$

$$H_A : \beta_1 \neq 1$$

From (a) we know, $\hat{\beta}_1 = 0.994757$

So, Test statistics $= \dfrac{0.994757 - 1}{s.e.(\hat{\beta}_1)}$

$= \dfrac{0.994757 - 1}{0.005219}$ $[se(\hat{\beta}_1) = 0.005219]$ from R

$= -1.00459$

and the p value is < 0.05.

So, we can not able to reject the null hypothesis.

Also, from (a) we see that $\hat{\beta}_1 = 0.994757$ which is close to 1.

Also the CI of slope : $(0.9827204, 1.0067921)$ contains the value 1.

So, we show the evidence for both d(i) & d(ii).

5(e): if we accept (i) of 5(d) then the model become:

$$y = \beta_1 x + \varepsilon$$

So for this model, $\hat{\beta}_1 = 0.987153$ [R code is given at the end)

Redoing part (c)

95% CI is $= (0.9810362, 0.9932693)$

Reexamine property (ii)

$H_0 : \beta_1 = 1$ (previous)

$H_A : \beta_1 \neq 1$

we see that the CI $= (0.9810362, 0.9932693)$ which doesn't contain the value 1. So, we reject the null hypothesis.

So, the property (ii) does't hold in 5(e).

and t value = 365.1

and p value $= 2 \times 10^{-16} <<< 0.05$

(f) Reason why the result for (d) & (e) are different.

The main reason is that in problem 5(d) we don't have enough evidence to reject the null hypothesis. Also we see that the CI of slope also contains the value 1 in 5(d).

But in 5(e) the CI is (0.9810362, 0.9932693) which doesn't contain the value 1.

As in the problem 5(e) we exclude the intercept which leads to poor fit of the model, which ultimately make the two result different.