1 growth

We fit a second-order model using the following equation, check the summary.

```
> h<-lm(Yield~x1+x2+x3+I(x1^2)+I(x2^2)+I(x3^2)+x1*x2+x2*x3+x1*x3,data = growth)
> summary(h)
> pure.error.anova(h)
```

The summay and anova table are shown as shown in figure **??** and figure **??**

```
> summary(h)

Call:
lm(formula = Yield ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^
 +
    x1 * x2 + x2 * x3 + x1 * x3, data = growth)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6661  -9.1577  -0.6661   9.1718  17.3339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  100.666      5.564  18.093  5.7e-09 ***
x1             1.271      3.691   0.344  0.73765
x2             1.361      3.691   0.369  0.71998
x3            -1.494      3.691  -0.405  0.69411
I(x1^2)       -3.767      3.593  -1.048  0.31912
I(x2^2)      -12.430      3.593  -3.459  0.00613 **
I(x3^2)       -9.601      3.593  -2.672  0.02342 *
x1:x2          2.875      4.823   0.596  0.56436
x2:x3         -4.625      4.823  -0.959  0.36020
x1:x3         -2.625      4.823  -0.544  0.59819
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.64 on 10 degrees of freedom
Multiple R-squared:  0.6631,    Adjusted R-squared:  0.3598
F-statistic: 2.186 on 9 and 10 DF,  p-value: 0.1194
```

Figure 1: summary of second order model

Based on the p-values in figure 1, we could tell that x1 is significant non-important and we can simply remove x1 from our model. We refit the model using the following code, check the summary and analysis of variance.

```
>h<-lm(Yield~x2+x3+I(x2^2)+I(x3^2)+x2*x3,data = growth)
>summary(h)
>pure.error.anova(h)
```

The p-value of lack of fit is 0.373181, which is greater than 0.05. Thus, we could conclude that the second model is adequate to represent the data. Fitted model: $\hat{y} = 97.583 + 1.361\hat{x_2} - 1.494\hat{x_3} - 12.055\hat{x_2}^2 - 9.227\hat{x_3}^2 - 4.625\hat{x_2}\hat{x_3}$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   97.583      4.349  22.440 2.25e-12 ***
x2             1.361      3.399   0.401  0.69483    
x3            -1.494      3.399  -0.440  0.66688    
I(x2^2)      -12.055      3.292  -3.662  0.00257 ** 
I(x3^2)       -9.227      3.292  -2.803  0.01411 *  
x2:x3         -4.625      4.441  -1.041  0.31532    
---
```

Figure 2: summary of second order model

```
Analysis of Variance Table

Response: Yield
             Df  Sum Sq Mean Sq F value   Pr(>F)   
x2            1   25.31   25.31  0.1655 0.691981   
x3            1   30.50   30.50  0.1994 0.663859   
I(x2^2)       1 1848.02 1848.02 12.0821 0.005185 **
I(x3^2)       1 1239.17 1239.17  8.1015 0.015901 * 
x2:x3         1  171.13  171.13  1.1188 0.312853   
Residuals    14 2208.83  157.77                    
 Lack of fit  3  526.33  175.44  1.1470 0.373181   
 Pure Error  11 1682.50  152.95                    
```

Figure 3: anova of second order model

$B_{2,2} = \begin{pmatrix} -12.055 & -2.3125 \\ -2.3125 & -12.055 \end{pmatrix}$,$b = \begin{pmatrix} 1.361 \\ -1.494 \end{pmatrix}$,$x_s = -\frac{1}{2}B^{-1}b = \begin{pmatrix} 0.07561505 \\ -0.09990894 \end{pmatrix}$, and the eigenvalues of matrix B are $\begin{pmatrix} -7.930455 \\ -13.351545 \end{pmatrix}$. All eigenvalues are negative, which makes sure that X_s is the maximum point.

Based on the result, we can couclude that the optimal setting is x2 = 0.076, x3 = -0.1, while x1 can be any value since it is not important.

2 average age

a) The sample design is simple random sampling without replacement. Under SRSWOR, the sample mean $\bar{y}$ is an unbiased estimator of $\bar{Y}$, thus the estimator of mean age for children is $\bar{y} = \frac{9*13+10*35+11*44+12*69+13*36+14*24+15*7+16*3+17*2+18*5}{240} = 12.08$ The $v(\bar{y})$ is an unbiased estimator of $V(\bar{y})$, and $v(\bar{y}) = \frac{s^2}{n} = \frac{3.705}{240} = 0.015$, thus, the standard error $se(\bar{y}) = \sqrt{v(\bar{y})} = 0.124$. And the 95% confidence interval for the average age is $\bar{y} \pm Z_{\alpha/2} s \sqrt{\frac{1}{n}} = 12.08 \pm 0.243$ **b)** We determine the sample size based on this formula: $n = \frac{{Z_{\alpha/2}}^2 S^2}{e^2} = \frac{1.96^2 * 3.705}{0.5^2} = 56.93$, hence, the minimun sample size is 57.

3 clams

First, we calculate $N_h h = 1, , , 4$, $N_1 = 222.81 * 25.6 = 5704, N_2 = 49.61 * 25.6 = 1270, N_3 = 50.25 * 25.6 = 1287, N_4 = 197.81 * 25.6 = 5064, N = N_1 + N_2 + N_3 + N_4 = 13325$. Then we obtain the $\bar{y_{st}} = \sum_{h=1}^{H} W_h \bar{y_h} = 1.36$. After that, we can have the estimator of the total number of bushels $\hat{t_{st}} = N\bar{y_{st}} = 13325 * 1.36 = 18122$.

The variance of $\hat{y_{st}}$: $v(\hat{y_{st}}) = \sum_{h=1}^{H} {W_h}^2(1 - n_h/N_h){s_h}^2/n_h = 0.0327$, thus, the variance of $\hat{t_{st}}$: $v(\hat{t_{st}}) = N^2 v(y_{st}) = 13325^2 * 0.0327 = 5806069$ the standard error is $se(\hat{t_{st}}) = \sqrt{v(\hat{t_{st}})} = 2410$

4 totoal number of acres

a) Use ratio estimation to estimate the total number of acres:

$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{mean(acres92)}{mean(farms87)} = 459.8975$

$\hat{t_{yr}} = \hat{R}t_x = 459.8975 * 2087759 = 960,155,061$

b) Use the regression estimation:

$\hat{\beta_0} = 267029.81, \hat{\beta_1} = 47.65$

$\hat{\bar{y}}_{req} = \hat{\beta_0} + \hat{\beta_1}\bar{x} = 267029.81 + 47.65 * 2087759/3078 = 299350.1$

$\hat{t}_{yreq} = N\hat{\bar{y}}_{req} = 3078 * 299350.1 = 921,399,608$

c) In order to find the method with most precision, we calculate the standard variance of $\hat{t}_y$.

ratio estimation with auxiliary variable acres87, $se(\hat{t}_{yra87}) = \sqrt{var(\hat{t}_y)} = \sqrt{N^2(1 - \frac{n}{N})\frac{1}{n}\frac{1}{n-1}\sum_{i \in s}(y_i - \hat{R}x_i)^2} = 5,344,567$

ratio estimation with auxiliary variable farms87,

$se(\hat{t}_{yrf87}) = \sqrt{var(\hat{t}_y)} = \sqrt{N^2(1 - \frac{n}{N})\frac{1}{n}\frac{1}{n-1}\sum_{i \in s}(y_i - \hat{R}x_i)^2} = 65,364,822$

regression estimation with auxiliary variable farms87,

$se(\hat{t}_{yregf87}) = \sqrt{var(\hat{t}_y)} = \sqrt{N^2(1 - \frac{n}{N})\frac{1}{n}\frac{1}{n-1}\sum_{i \in s}(y_i - \beta_0 - \beta_1 * x_i)^2} = 58,065,813$

Based on the variances, we can tell that ratio estimation has the most precision since its variance is minimum among these three methods.

5 Neyman allocation

a)

$$\begin{aligned}
V_{Neyman}(\hat{t}_{str}) &= N^2 V(\bar{y}_{st}) = N^2 \sum_{h=1}^{H} {W_h}^2 (1 - \frac{n_h}{N_h}) {S_h}^2 / n_h \\
&= \sum_{h=1}^{H} {N_h}^2 (1 - \frac{n_h}{N_h}) {S_h}^2 / n_h \\
&= \sum_{h=1}^{H} {N_h}^2 (1 - \frac{\frac{N_h S_h n}{\sum_{h=1}^{H} N_l S_l}}{N_h}) {S_h}^2 \frac{\sum_{h=1}^{H} N_l S_l}{N_h S_h n} \\
&= \sum_{h=1}^{H} N_h S_h (1 - \frac{S_h n}{\sum_{h=1}^{H} N_l S_l}) \frac{\sum_{h=1}^{H} N_l S_l}{n} \\
&= \sum_{h=1}^{H} N_h S_h (\frac{\sum_{h=1}^{H} N_l S_l}{n} - S_h) \\
&= \frac{1}{n} \sum_{h=1}^{H} N_l S_l \sum_{h=1}^{H} N_h S_h - \sum_{h=1}^{H} N_h {S_h}^2 \\
&= \frac{1}{n} (\sum_{h=1}^{H} N_h S_h)^2 - \sum_{h=1}^{H} N_h {S_h}^2
\end{aligned}$$

b)

$$\begin{aligned}
V_{prop}(\hat{t}_{str}) - V_{Neyman}(\hat{t}_{str}) &= \frac{N}{n} \sum_{h=1}^{H} N_h {S_h}^2 - \sum_{h=1}^{H} N_h {S_h}^2 - \frac{1}{n} (\sum_{h=1}^{H} N_h S_h)^2 + \sum_{h=1}^{H} N_h {S_h}^2 \\
&= \frac{N}{n} \sum_{h=1}^{H} N_h {S_h}^2 - \frac{1}{n} (\sum_{h=1}^{H} N_h S_h)^2 \\
&= \frac{N^2}{n} \sum_{h=1}^{H} \frac{N_h}{N} {S_h}^2 - \frac{N^2}{n} (\sum_{h=1}^{H} \frac{N_h}{N} S_h)^2 \\
&= \frac{N^2}{n} [\sum_{h=1}^{H} \frac{N_h}{N} {S_h}^2 - (\sum_{h=1}^{H} \frac{N_h}{N} S_h)^2] \\
&= \frac{N^2}{n} \sum_{h=1}^{H} \frac{N_h}{N} (S_h - \sum_{l=1}^{H} \frac{N_l}{N} S_l)^2
\end{aligned}$$