Statistics 206

Homework 8 Solution

Due : Nov. 30, 2015, In Class

1. **Regression formulations of one-way ANOVA model**.

   (a) Consider the *cell means formulation* discussed in class:

   $$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i.$$

   Express this model as a linear regression model:

   $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

   Specify $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. What is $\mathbf{X}^T\mathbf{X}$ and what is $\mathbf{X}^T\mathbf{Y}$? What is the LS estimator of $\boldsymbol{\beta}$? Derive the fitted values and residuals. Specify what are $\mathbf{Y}, \boldsymbol{\epsilon}, \mathbf{X}$ and $\boldsymbol{\beta}$.

ANS

$$\underset{n_T \times 1}{\mathbf{Y}} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix}; \quad \underset{n_T \times I}{\mathbf{X}} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix}; \quad \underset{n_T \times 1}{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{I1} \\ \vdots \\ \epsilon_{In_I} \end{pmatrix}$$

where for $\mathbf{X}$, column X_i has n_i 1's and the other entries are 0.

$$\mathbf{X}^T\mathbf{X} = \text{diag}(n_1, n_2, ..., n_I)$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} n_1\bar{Y}_{1.} \\ n_2\bar{Y}_{2.} \\ \vdots \\ n_I\bar{Y}_{I.} \end{pmatrix}$$

The LS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{I.} \end{pmatrix}$$

The fitted values

$$\underset{n_T\times 1}{\hat{\mathbf{Y}}} = \begin{pmatrix} \bar{Y}_{1.} \\ \vdots \\ \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{I.} \\ \vdots \\ \bar{Y}_{I.} \end{pmatrix}$$

where each $\bar{Y}_{i.}$ appears n_i times.
The residuals are

$$\underset{n_T\times 1}{\mathbf{Y}} - \underset{n_T\times 1}{\hat{\mathbf{Y}}} = \begin{pmatrix} Y_{11} - \bar{Y}_{1.} \\ \vdots \\ Y_{1n_1} - \bar{Y}_{1.} \\ Y_{21} - \bar{Y}_{2.} \\ \vdots \\ Y_{2n_2} - \bar{Y}_{2.} \\ \vdots \\ Y_{I1} - \bar{Y}_{I.} \\ \vdots \\ Y_{In_I} - \bar{Y}_{I.} \end{pmatrix}$$

(b) Consider the alternative formulation used by R:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i, \quad \alpha_1 = 0.$$

Express this model as a linear regression model by specifying $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. Compare it with the linear regression model with $I-1$ indicator variables for factor levels (with level 1 as the reference class). What do you find?

ANS

$$\underset{n_T\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix}; \quad \underset{n_T\times I}{\mathbf{X}} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \cdots & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \cdots & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \cdots & \cdots & 0 & 1 \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_I \end{pmatrix}; \quad \underset{n_T\times 1}{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{I1} \\ \vdots \\ \epsilon_{In_I} \end{pmatrix}.$$

It is the same as the linear regression model which sets factor level 1 as baseline, and define $I-1$ dummy variables $X_2, \cdots, X_I$:

$$X_2 = \begin{cases} 1 & \text{if } i = 2 \\ 0 & \text{otherwise} \end{cases}, \cdots, X_I = \begin{cases} 1 & \text{if } i = I \\ 0 & \text{otherwise} \end{cases}$$

2. **Decomposition of total sum of squares in one-way ANOVA**.

(a) Show that $SSTO = SSTR + SSE$, where

- $SSTO := \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2$.
- $SSTR := \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$.
- $SSE := \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$.

ANS

$$\begin{aligned} SSTO &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\Big((Y_{ij} - \overline{Y}_{i\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\Big)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\Big((Y_{ij} - \overline{Y}_{i\cdot})^2 + 2(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2\Big) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\Big((Y_{ij} - \overline{Y}_{i\cdot})^2 + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2\Big) + \sum_{i=1}^{I}\sum_{j=1}^{n_i} 2(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 + \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 + \sum_{i=1}^{I}\left(2(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\right)\left(\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})\right) \\ &= SSE + SSTR + 0. \end{aligned}$$

The last term is 0 since $\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot}) = \sum_{j=1}^{n_i} Y_{ij} - n_i\overline{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij} - \sum_{j=1}^{n_i} Y_{ij} = 0$ for $i = 1, \cdots, I$.

(b) Show that

- $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{\cdot\cdot})=0$. (So $d.f.(SSTO)=n_T-1$)
- $\sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})=0$. (So $d.f.(SSTR)=I-1$)
- For each $i=1,\cdots,I$, $\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\cdot})=0$. (So $d.f.(SSE)=n_T-I$)

ANS

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{\cdot\cdot})=\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij}-\sum_{i=1}^{I}n_i\overline{Y}_{\cdot\cdot}=\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij}-\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij}}{\sum_{i=1}^{I}n_i}\sum_{i=1}^{I}n_i=0.$$

$$\sum_{i=1}^{I}n_i(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})=\sum_{i=1}^{I}n_i\left(\frac{\sum_{j=1}^{n_i}Y_{ij}}{n_i}\right)-\left(\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij}}{\sum_{i=1}^{I}n_i}\right)\sum_{i=1}^{I}n_i=\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij}-\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij}=0.$$

$$\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\cdot})=\sum_{j=1}^{n_i}Y_{ij}-\sum_{j=1}^{n_i}\frac{\sum_{j=1}^{n_i}Y_{ij}}{n_i}=\sum_{j=1}^{n_i}Y_{ij}-\frac{\sum_{j=1}^{n_i}Y_{ij}}{n_i}\sum_{j=1}^{n_i}1=\sum_{j=1}^{n_i}Y_{ij}-\frac{\sum_{j=1}^{n_i}Y_{ij}}{n_i}n_i=0.$$

3. **Expectation and Sampling distribution of SS in one-way ANOVA**.

(a) Show that $E[SSTR]=(I-1)\sigma^2+\sum_{i=1}^{I}n_i(\mu_i-\mu_\cdot)^2$, where

$$\mu_\cdot=\frac{1}{n_T}\sum_{i=1}^{I}n_i\mu_i,\quad n_T=\sum_{i=1}^{I}n_i.$$

(Hints: (i) $\overline{Y}_{i\cdot}\sim N(\mu_i,\sigma^2/n_i)$ *and are independent with each other; (ii)* $\bar{Y}_{\cdot\cdot}=\frac{1}{n_T}\sum_{i=1}^{I}n_i\bar{Y}_{i\cdot}\sim N(\mu_\cdot,\frac{\sigma^2}{n_T})$*; (iii) For a random variable* Z*:* $E(Z^2)=Var(Z)+(E(Z))^2$.

ANS

$$\begin{aligned}
SSTR&=\sum_{i=1}^{I}n_i\left(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}\right)^2=\sum_{i=1}^{I}n_i\left[(\overline{Y}_{i\cdot}-\mu_\cdot)-(\overline{Y}_{\cdot\cdot}-\mu_\cdot)\right]^2\\
&=\sum_{i=1}^{I}n_i\left[\left(\overline{Y}_{i\cdot}-\mu_\cdot\right)^2+\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)^2-2\left(\overline{Y}_{i\cdot}-\mu_\cdot\right)\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)\right]\\
&=\sum_{i=1}^{I}n_i\left(\overline{Y}_{i\cdot}-\mu_\cdot\right)^2+\sum_{i=1}^{I}n_i\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)^2-2\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)\sum_{i=1}^{I}n_i\left(\overline{Y}_{i\cdot}-\mu_\cdot\right)\\
&=\sum_{i=1}^{I}n_i\left(\overline{Y}_{i\cdot}-\mu_\cdot\right)^2+n_T\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)^2-2\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)n_T\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)\\
&=\sum_{i=1}^{I}n_i\left(\overline{Y}_{i\cdot}-\mu_\cdot\right)^2-n_T\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)^2\ \text{ since }\ \sum_{i=1}^{I}n_i\left(\overline{Y}_{i\cdot}-\mu_\cdot\right)=n_T\left(\overline{Y}_{\cdot\cdot}-\mu_\cdot\right)
\end{aligned}$$

$$E(SSTR) = \sum_{i=1}^{I} n_i E\left[\left(\overline{Y}_{i\cdot} - \mu_\cdot\right)^2\right] - n_T E\left[\left(\overline{Y}_{\cdot\cdot} - \mu_\cdot\right)^2\right]$$

Since $E\left(\overline{Y}_{\cdot\cdot}\right) = \mu_\cdot$,

$$E\left[\left(\overline{Y}_{\cdot\cdot} - \mu_\cdot\right)^2\right] = \text{Var}\left(\overline{Y}_{\cdot\cdot}\right) = \frac{\sigma^2}{n_T}$$

$$\begin{aligned} E\left[\left(\overline{Y}_{i\cdot} - \mu_\cdot\right)^2\right] &= E\left[\left((\overline{Y}_{i\cdot} - \mu_i) + (\mu_i - \mu_\cdot)\right)^2\right] \\ &= E\left[(\overline{Y}_{i\cdot} - \mu_i)^2\right] + E\left[(\mu_i - \mu_\cdot)^2\right] + 2(\mu_i - \mu_\cdot)E\left[\overline{Y}_{i\cdot} - \mu_i\right] \\ &= Var(\overline{Y}_{i\cdot}) + (\mu_i - \mu_\cdot)^2 + 0 \quad \text{since } E(\overline{Y}_{i\cdot}) = \mu_i \\ &= \frac{\sigma^2}{n_i} + (\mu_i - \mu_\cdot)^2. \end{aligned}$$

Substituting in to the expression for $E(SSTR)$ yields:

$$E(SSTR) = \sum_{i=1}^{I} n_i \left[\frac{\sigma^2}{n_i} + (\mu_i - \mu_\cdot)^2\right] - n_T \times \frac{\sigma^2}{n_T} = (I-1)\sigma^2 + \sum_{i=1}^{I} n_i(\mu_i - \mu_\cdot)^2.$$

(b) Show that $E[SSTO] = (n_T - 1)\sigma^2 + \sum_{i=1}^{I} n_i(\mu_i - \mu_\cdot)^2$. (Hint: Recall $SSE \sim \sigma^2\chi^2_{n_T - I}$).

$$SSTO = SSE + SSR$$

$$\therefore E(SSTO) = E(SSE) + E(SSR) = (n_T - I)\sigma^2 + (I-1)\sigma^2 + \sum_{i=1}^{I} n_i(\mu_i - \mu_\cdot)^2$$

$$= (n_T - 1)\sigma^2 + \sum_{i=1}^{I} n_i(\mu_i - \mu_\cdot)^2$$

(c) **(Optional Problem).** Under $H_0 : \mu_1 = \cdots = \mu_I$, show that SSTO $\sim \sigma^2\chi^2_{(n_T-1)}$. Due to independence of SSE and $SSTR$ (see Problem 4), therefore under H_0, SSTR $\sim \sigma^2\chi^2_{(I-1)}$.

(Hints: (i) Under H_0, $\mu_i \equiv \mu_\cdot$ for $i = 1, \cdots, I$; (ii) So $Y_{ij} - \overline{Y}_{\cdot\cdot} = (Y_{ij} - \mu_i) - (\overline{Y}_{\cdot\cdot} - \mu_\cdot)$. Therefore, without loss of generality, we can assume $\mu_i \equiv 0$ for $i = 1, \cdots, I$, i.e., $Y_{ij} \sim_{i.i.d.} N(0, \sigma^2)$; (iii) Note $\overline{Y}_{\cdot\cdot} = \frac{1}{n_T}\sum_{ij} Y_{ij}$.)

ANS Under H_0, $Y_{ij} \overset{iid}{\sim} N(\mu, \sigma^2)$ and $\bar{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}}{n_T}$. Then based on the known facts:

$$\frac{\sum(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{(n-1)} \text{ Under assumption that } X_i \overset{iid}{\sim} N(\mu, \sigma^2)$$

we have,

$$\frac{SSTO}{\sigma^2} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2}{\sigma^2} \sim \chi^2_{(n_T-1)}$$

4. **Independence of SSE and SSTR in one-way ANOVA.** Show the following:

   (a) For each $i = 1, \cdots, I$, $Y_{ij} - \overline{Y}_{i\cdot}$ $(j = 1, \cdots, n_i)$ and $\overline{Y}_{i\cdot}$ are independent. *(Hint: Calculate their covariance.)*

ANS

$$\begin{aligned}\operatorname{cov}\left(Y_{ij} - \overline{Y}_{i\cdot}, \overline{Y}_{i\cdot}\right) &= \operatorname{cov}(Y_{ij}, \overline{Y}_{i\cdot}) - \operatorname{cov}(\overline{Y}_{i\cdot}, \overline{Y}_{i\cdot}) = \frac{1}{n_i}\operatorname{cov}(Y_{ij}, \sum_{j=1}^{n_i} Y_{ij}) - \frac{\sigma^2}{n_i} \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i}\operatorname{cov}(Y_{ij}, Y_{ij'}) - \frac{\sigma^2}{n_i} \\ &= \frac{1}{n_i}\operatorname{Var}(Y_{ij}) - \frac{\sigma^2}{n_i} \qquad (\text{ Since } Y'_{ij}s \text{ are independent }, \operatorname{cov}(Y_{ij}, Y_{ij'}) = 0 \text{ when } j \neq j') \\ &= \frac{\sigma^2}{n_i} - \frac{\sigma^2}{n_i} = 0\end{aligned}$$

   Under the normal assumption, $Y_{ij} - \overline{Y}_{i\cdot}$ and $\overline{Y}_{i\cdot}$ are independent for each $i = 1, \cdots, I$.

   (b) For $i \neq i'$, $Y_{i'j} - \overline{Y}_{i'\cdot}$ $(j = 1, \cdots, n_{i'})$ and $\overline{Y}_{i\cdot}$ are independent .

ANS $Y_{ij} - \overline{Y}_{i\cdot}$ is a function of the observations in the i-th group, while $\overline{Y}_{i'\cdot}$ is a function of the observations in the i'-th group. The observations from two different groups are independent, thus $Y_{ij} - \overline{Y}_{i\cdot}$ and $\overline{Y}_{i'\cdot}$ are independent.

   (c) SSE and $\overline{Y}_{i\cdot}$ $(i = 1, \cdots, I)$ are independent. *(Hint: Use the fact that if two sets of random variables are independent, then any function of the first set is independent with any function of the second set.)*

ANS We have already proved in part (a) and (b) that $Y_{ij} - \overline{Y}_{i\cdot}$ and $\overline{Y}_{i\cdot}$ are independent and $Y_{i'j} - \overline{Y}_{i'\cdot}$ and $\overline{Y}_{i\cdot}$ are independent for each $i' \neq i$. Hence SSE is independent with $\overline{Y}_{i\cdot}$ and this is true for $i = 1, \cdots, I$.

   (d) SSE and $SSTR$ are independent.

ANS $SSTR = \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$ is a function of $\overline{Y}_{i\cdot}$'s. Therefore by part (c) we know that SSE and $SSTR$ are independent.

Notes: Indeed we have already had the results in problems 2, 3, 4 from linear regression theories (Are you able to recognize them?). Here, we show that they can be derived without using matrix algebra. This is due to the special structure of the ANOVA model: Recall that the design matrices in problem 1 are of very special forms.

5. **One-way ANOVA case study in R.**

   Notes: You do not need to hand in this problem on Monday.

   A company uses six filling machines of the same make and model to place detergent into cartons that show a label weight of 32 ounces. The production manger has complained

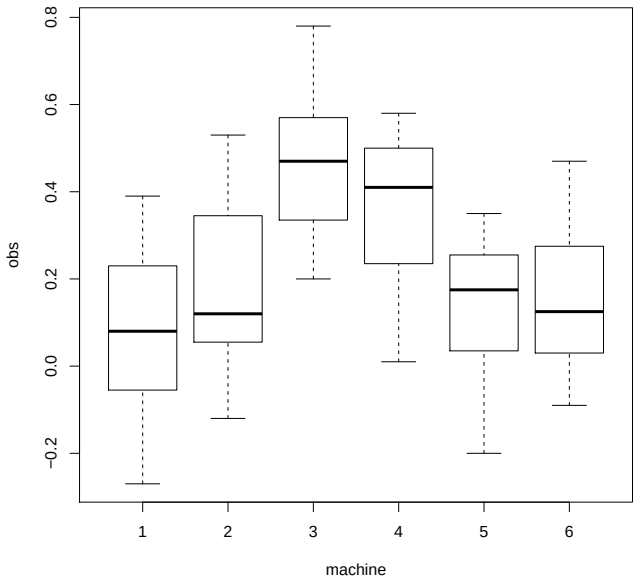
that the six machines do not place the same amount of fill into the cartons. A consultant requested that 20 filled cartons be selected randomly from each of the six machines and the content of each carton weighted. The observations were recorded in terms of the deviations of weights from 32 ounces. The data is under Resources/Homework/filling.txt: The first column is the observation, the second column is the index for the filling machine and the third column is the index for the carton. Consider fitting the one-way ANOVA model to this data.

(a) What is the response variable? What is the factor? How many levels are there for this factor? Name the design of this study.

ANS The response variable is the the deviation of weight from 32 ounces.
The factor is the filling machine. It has 6 levels.
This is a balanced complete randomized design.

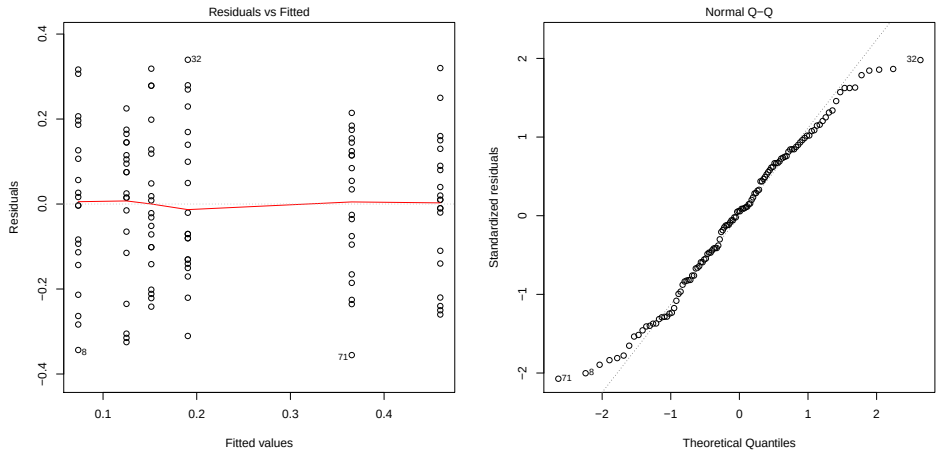
(b) Draw side-by-side box plots of the response for the factor levels. Do the factor level means appear to differ? Does the variability of the observations within each factor level appear to be approximately the same for all factor levels?

ANS The box plots of the response for the factor levels are shown below. The factor level means appear to differ. The sizes of the boxes are roughly equal so there is no obvious sign of unequal variance



(c) Fit the one-way ANOVA model. Draw residual versus fitted value plot and residual Q-Q plot. Comment on model assumptions. Are remedial measures needed? *(Hint: When using the* `lm` *function, remember to declare the factor as* `factor`*.)*

ANS The normal QQ plot and residuals plots are shown below. The normal QQ plot indicates the slightly light tailed of the residuals' distribution; while the residuals versus fitted values plots shows no sign for unequal variance. Hence we could conclude that the model assumptions hold.



(d) Obtain the estimated factor levels means and obtain the ANOVA table.

ANS We fit the model under the restriction $\alpha_1 = 0$, i.e. we set level 1 as our baseline

```
> data=read.table("filling.txt")
> obs = data[,1]
> machine = as.factor(data[,2])
> n = nrow(data)

> fit=lm(obs~machine)
> summary(fit)

Call:
lm(formula = obs ~ machine)

Residuals:
Min      1Q  Median      3Q     Max
-0.3555 -0.1305  0.0100  0.1289  0.3395

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07350    0.03935   1.868   0.0644 .
machine2     0.11700    0.05565   2.102   0.0377 *
machine3     0.38650    0.05565   6.945 2.47e-10 ***
machine4     0.29200    0.05565   5.247 7.22e-07 ***
```

```
machine5     0.05150    0.05565   0.925   0.3567
machine6     0.07800    0.05565   1.402   0.1638
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.176 on 114 degrees of freedom
Multiple R-squared:  0.3934,    Adjusted R-squared:  0.3668
F-statistic: 14.78 on 5 and 114 DF,  p-value: 3.636e-11
```

The estimated group means can be obtained from the summary output:

$$\hat{\mu}_1 = 0.0735 \qquad \hat{\mu}_2 = \hat{\mu}_1 + 0.1170 = 0.1905 \qquad \hat{\mu}_3 = \hat{\mu}_1 + 0.38650 = 0.4600$$

$$\hat{\mu}_4 = \hat{\mu}_1 + 0.29200 = 0.3655 \qquad \hat{\mu}_5 = \hat{\mu}_1 + 0.05150 = 0.1250 \qquad \hat{\mu}_6 = \hat{\mu}_1 + 0.07800 = 0.1515$$

ANS ANOVA output from R is shown as bellow:

```
> anova(fit)
Analysis of Variance Table

Response: obs
Df Sum Sq Mean Sq F value    Pr(>F)
machine     5 2.2893 0.45787  14.784 3.636e-11 ***
Residuals 114 3.5306 0.03097
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The SSR is 2.2893, the SSE is 3.5306 from the output. The degrees of freedom for SSR is $I-1=5$, the degrees of freedom for SSE is $n-I=114$. Hence the ANOVA table is:

Source of variation	Sum of Squares (SS)	Degree of Freedom (df)	MS
Between treatments	$SSTR = 2.2893$	$I - 1 = 5$	$MSTR = 0.45787$
Within treatments	$SSE = 3.5306$	$n_T - I = 114$	$MSE = 0.03097$
Total	$SSTO = 5.8200$	$n_T - 1 = 119$	

(e) Test whether or not the mean fill differs among the six machines at level 0.05. State the null and alternative hypotheses, the decision rule and the conclusion.

ANS

$$H_0 : \mu_1 = \cdots = \mu_6 \qquad \text{versus} \qquad H_a : \text{not all factor level means are the same}$$

$$F^* = \frac{MSTR}{MSE} = 14.784$$

Under the null hypothesis:$F^* \overset{iid}{\sim} F_{(5,114)}$
We can reject the null hypothesis if $F^* > F(1-\alpha, I-1, n_T-1)$ or if $p = P(F_{(5,114)} > 14.784) < \alpha$

$$p = P(F_{(5,114)} > 14.784) = 3.636 \times 10^{-11} < \alpha$$

Therefore, we can reject the null hypothesis at 5% level.

(f) Construct a 99% confidence interval for μ_2. If we are interested in all factor level means, what multiple comparison procedure shall we use? What would be the corresponding 99% confidence interval for μ_2?

ANS The $(1-\alpha)$-confidnce interval of μ_i is:

$$\widehat{\mu}_2 \pm s\left(\widehat{\mu}_2\right) t\left(1-\frac{\alpha}{2}; n_T - I\right)$$

$$\widehat{\mu}_2 = 0.1905 \qquad s\left(\widehat{\mu}_2\right) = \sqrt{\frac{MSE}{n_i}} = \frac{0.176}{\sqrt{20}} = 0.0393511 \qquad t\left(1-\frac{\alpha}{2}; n_T - I\right) = 2.6189$$

Therefore, the 99% confidence interval for μ_2 is:

$$[0.0874, 0.2936]$$

If we are interested in all factor level means, we should use Bonferroni procedure (here $g = 6$). The corresponding Bonferroni's multiplier is:

$$B = t\left(1-\frac{\alpha}{2g}; n_T - I\right) = t(1-\frac{0.01}{2\times 6}; 114) = 3.2207$$

The corresponding confidence interval for μ_2 is:

$$\widehat{\mu}_2 \pm s\left(\widehat{\mu}_2\right) \times B = [0.0638, 0.3172]$$

(g) How many pairwise comparisons of factor levels means are there? If we are interested in all these pairwise comparisons, what multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? Construct the corresponding 95% confidence intervals for all pairwise comparisons. At familywise significance level 0.05, which pairs of factor level means should be declared as being different?

ANS There are 15 pairwise comparisons of factor levels means. If we are interested in all these pairwise comparisons, we should use Tukey's procedure. The corresponding Tukey's multiplier is:

$$T = \frac{1}{\sqrt{2}} q\left(1-\alpha; I, n_T - I\right) = \frac{1}{\sqrt{2}} q\left(0.95; 6, 114\right) = 2.8988$$

The Tukey's 95% confidence intervals for all pairwise comparisons were obtained by using the function `TukeyHSD`, and the output from R is shown as below:

```
> TukeyHSD(aov(fit))
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = fit)

$mach
diff        lwr        upr     p adj
2-1  0.1170 -0.0443194  0.2783194 0.2934937
3-1  0.3865  0.2251806  0.5478194 0.0000000*
4-1  0.2920  0.1306806  0.4533194 0.0000106*
5-1  0.0515 -0.1098194  0.2128194 0.9392011
6-1  0.0780 -0.0833194  0.2393194 0.7260015
3-2  0.2695  0.1081806  0.4308194 0.0000588*
4-2  0.1750  0.0136806  0.3363194 0.0252432*
5-2 -0.0655 -0.2268194  0.0958194 0.8469184
6-2 -0.0390 -0.2003194  0.1223194 0.9815028
4-3 -0.0945 -0.2558194  0.0668194 0.5359056
5-3 -0.3350 -0.4963194 -0.1736806 0.0000003*
6-3 -0.3085 -0.4698194 -0.1471806 0.0000029*
5-4 -0.2405 -0.4018194 -0.0791806 0.0004684*
6-4 -0.2140 -0.3753194 -0.0526806 0.0026737*
6-5  0.0265 -0.1348194  0.1878194 0.9968910
```

At familywise significance level 0.05, there are eight pairs should be declared as being different, which are:
1 versus 3, 1 versus 4;
2 versus 3, 2 versus 4;
3 versus 5, 3 versus 6;
4 versus 5, 4 versus 6

(h) What if we are only interested in 6 **pre-specified** pairwise comparisons, which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? What if we are only interested in the 6 pairwise comparisons that show the most differences in the data (i.e corresponding to the six largest $|\widehat{D}|$), which procedure shall we use and what are the corresponding C.Is?

ANS If we are only interested in 6 pre-specified pairwise comparison, Bonferroni's procedure should be applied and the corresponding multiplier at $\alpha = 0.05$ is (note Bonferroni's multiplier in this case is smaller than Tukey's multiplier):

$$B = t\left(1 - \frac{\alpha}{2g}; n_T - I\right) = t(1 - \frac{0.05}{2 \times 6}; 114) = 2.6851$$

If we are only interested in the 6 pairwise comparisons that show the most differences in the data,we should use Tukey's multiplier as Bonferron's procedure is not

applicable anymore: It is only applicable to pre-specified comparisons.

(i) If we are interested in all possible contrasts, which multiple comparison procedure shall we use? Construct the corresponding 95% confidence interval for the contrast:

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}.$$

Test whether or not $L = 0$ with family-wise-error-rate controlled at 0.05. What if we are interested in 20 pre-specified contrasts (including L), which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? Construct the corresponding 95% confidence interval for L (assuming L is in this prespecified set of contrasts). What do you find?

ANS If we are interested in all possible contrasts, we should use Scheffe's procedure. Consider the constrast:

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$
$$c_1 = c_2 = 0.5, c_3 = c_4 = -0.5$$

An unbiased estimator of L:

$$\hat{L} = \frac{\overline{Y}_{1.} + \overline{Y}_{2.}}{2} - \frac{\overline{Y}_{3.} + \overline{Y}_{4.}}{2} = \frac{0.0735 + 0.1905}{2} - \frac{0.4600 + 0.3655}{2} = -0.28075$$

$$s(\hat{L}) = \sqrt{MSE \times \sum_{i=1}^{4} \frac{c_i^2}{n_i}} = 0.03935$$

Since we are interested in all possible constrasts, we choose Scheffe's multipler at $\alpha = 0.05$:

$$S = \sqrt{(I-1)F(1-\alpha; I-1, n_T - I)} = \sqrt{5 \times F(0.95; 5, 114)} = 3.3867$$

So the 95% confidence interval for L is:

$$\hat{L} \pm s(\hat{L}) \times S = [-0.4140, -0.1475]$$

Since 0 is not contained in the above interval, we can reject the null hypothesis that $L = 0$.

If we are interested in 20 pre-specified contrasts, we should use Bonferroni's multiplier. Since att $\alpha = 0.05$, the corresponding Bonferroni's multiplier is:

$$B = t\left(1 - \frac{\alpha}{2g}; n_T - I\right) = t(1 - \frac{0.05}{2 \times 20}; 114) = 3.0919 < S = 3.39.$$

The corresponding 95% confidence interval for L is: $[-0.4024, -0.1591]$. The resulting confidence interval is narrower compared with the one obtain through Scheffe's procedure.