STAT-285 Homework 8 Solutions

Section 11.3 Question 30 /11

Study Objective: Determine if there exists an effect of *nitrogen level*, *times of planting*, and *potassium level* on the N content of corn grain.

Formulation: Let

- X_{ijk} denote the (sole) observation attributed with the ith nitrogen level, jth time of planting, and kth potassium level, with $i = 1, 2, 3, 4$, $j = 1, 2$, and $k = 1, 2$ (ie $I = 4$, $J = 2$, and $K = 2$).

We assume that $X_{ijk} \sim N(\mu_{ijk}, \sigma^2)$, where $\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}$, with $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\cdots$

Part A /5

We are to fill out the following ANOVA table:

Source	df	Sum of Squares	Mean Square	F
Nitrogen Level (Factor A)	$I-1$	SSA	MSA	F_A
Planting Time (Factor B)	$J-1$	SSB	MSB	F_B
Potassium Level (Factor C)	$K-1$	SSC	MSC	F_C
AB	$(I-1)(J-1)$	$SSAB$	$MSAB$	F_{AB}
AC	$(I-1)(K-1)$	$SSAC$	$MSAC$	F_{AC}
BC	$(J-1)(K-1)$	$SSBC$	$MSBC$	F_{BC}
Error	$(I-1)(J-1)(K-1)$	SSE	MSE	
Total	$IJK-1$	SST		

- df for nitrogen level is $I - 1 = 3$
- dffor planting time is $J - 1 = 1$
- df for potassium level is $K - 1 = 1$

- df for nitrogen level and planting time interaction is $(I-1)(J-1) = 3$
- df for nitrogen level and potassium level interaction is $(I-1)(K-1) = 3$
- df for planting time and potassium level interaction is $(J-1)(K-1) = 1$
- df for error is $(I-1)(J-1)(K-1) = 3$
- total df is $IJK - 1 = 15$
- $SST = 0.2384$ (given to us)
- $SSA = 0.22625$ (given to us)
- $SSB = 0.000025$ (given to us)
- $SSC = 0.0036$ (given to us)
- $SSAB = 0.004325$ (given to us)
- $SSAC = 0.00065$ (given to us)
- $SSBC = 0.000625$ (given to us)
- $SSE = SST - SSA - SSB - SSC - SSAB - SSAC - SSBC = 0.002925$
- $MSA = SSA/(I-1) = 0.0754$
- $MSB = SSB/(J-1) = 0.00003$
- $MSC = SSC/(J-1) = 0.0036$
- $MSAB = SSAB/((I-1)(J-1)) = 0.0014$
- $MSAC = SSAC/((I-1)(K-1)) = 0.0002$
- $MSBC = SSBC/((J-1)(K-1)) = 0.0006$
- $MSE = SSE/((I-1)(J-1)(K-1)) = 0.0010$
- $F_A = MSA/MSE = 77.35$
- $F_B = MSB/MSE = 0.03$
- $F_C = MSC/MSE = 3.69$
- $F_{AB} = MSAB/MSE = 1.48$
- $F_{AC} = MSAC/MSE = 0.22$
- $F_{BC} = MSBC/MSE = 0.64$

Part B /4

Hypothesis Test: $H_{0A} : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ vs. H_{aA}: At least one $\alpha_i \neq 0$

Test Statistic:

$$F_A = \frac{MSA}{MSE} \sim F(3,3).$$

From **Part A**, we have $F_{A,obs} = 77.35$. Since $P_{H_0}(F > F_{A,obs}) = 0.0024 < 0.05$, we reject H_{0A}.

Hypothesis Test: $H_{0B} : \beta_1 = \beta_2 = 0$ vs. H_{aB}: $\beta_1 \neq 0$ or $\beta_2 \neq 0$

Test Statistic:

$$F_B = \frac{MSB}{MSE} \sim F(1,3).$$

From **Part A**, we have $F_{B,obs} = 0.03$. Since $P_{H_0}(F > F_{B,obs}) = 0.8830 > 0.05$, we fail to reject H_{0B}.

Hypothesis Test: $H_{0C} : \delta_1 = \delta_2 = 0$ vs. H_{aC}: $\delta_1 \neq 0$ or $\delta_2 \neq 0$

Test Statistic:

$$F_C = \frac{MSC}{MSE} \sim F(1,3).$$

From **Part A**, we have $F_{C,obs} = 3.69$. Since $P_{H_0}(F > F_{C,obs}) = 0.1504 > 0.05$, we fail to reject H_{0C}.

Hypothesis Test: $H_{0AB} : \gamma_{ij}^{AB} = 0$ for all i, j vs. H_{aAB}: At least one $\gamma_{ij}^{AB} \neq 0$

Test Statistic:

$$F_{AB} = \frac{MSAB}{MSE} \sim F(3,3).$$

From **Part A**, we have $F_{AB,obs} = 1.48$. Since $P_{H_0}(F > F_{AB,obs}) = 0.37783 > 0.05$, we fail to reject H_{0AB}.

Hypothesis Test: $H_{0AC} : \gamma_{ik}^{AC} = 0$ for all i, k vs. H_{aAC}: At least one $\gamma_{ik}^{AC} \neq 0$

Test Statistic:

$$F_{AC} = \frac{MSAC}{MSE} \sim F(3, 3).$$

From **Part A**, we have $F_{AC,obs} = 0.22$. Since $P_{H_0}(F > F_{AC,obs}) = 0.8758 > 0.05$, we fail to reject H_{0AC}.

Hypothesis Test: $H_{0BC} : \gamma_{jk}^{BC} = 0$ for all j, k vs. H_{aBC}: At least one $\gamma_{jk}^{BC} \neq 0$

Test Statistic:

$$F_{BC} = \frac{MSBC}{MSE} \sim F(1, 3).$$

From **Part A**, we have $F_{BC,obs} = 0.64$. Since $P_{H_0}(F > F_{BC,obs}) = 0.4819 > 0.05$, we fail to reject H_{0BC}.

Part C /2

Note that we can apply Tukey's method, since we failed to reject H_{0AB} and H_{0AC} in **Part B**, and rejected H_{0A} in **Part B**.

We start by computing

$$W = Q_{0.05,I,(I-1)(J-1)(K-1)}\sqrt{\frac{MSE}{JK}} = 6.82\sqrt{\frac{0.0010}{4}} = 0.1065$$

where JK is the number of observations averaged to obtain each $\bar{X}_{i..}$. We summarize our findings with the following underscoring pattern

i	1	2	3	4
$\bar{X}_{i..}$	1.12	1.3025	1.3875	1.43
		———	———	
			———	———

We can see that the difference between the first level of nitrogen with the others is statistically significant.

Section 12.2 Question 15 /10

Study Objective: Investigate how modulus of elasticity (MOE) is related to flexural strength in concrete beams of a certain type.

Formulation: Let

- Y_i denote the ith measurement of flexural strength, for $i = 1, \cdots, n$, with $n = 27$.
- X_i denote the ith measurement of MOE, for $i = 1, \cdots, n$.

The goal is to establish how Y_i depends on X_i:

$$Y_i = f(X_i) + \varepsilon_i,$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, X_i and ε_i are independent, and $f(X_i) = E(Y_i|X_i)$. We specify $f(\cdot)$ to be a linear function

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

with β_0 and β_1 as unknown parameters.

Part A /2

Table 1 illustrates the stem-and-leaf display. The digit on the left side of the bar "|" is the tens-place digit, and the digit on the right hand side of the bar "|" is the ones-place digit. We can see that the majority of the observations lie in the interval [33, 49], and the distribution of MOE values is skewed to the right.

Table 1: Stem-and-leaf display of the MOE values for Section 12.2 Question 15

2	9
3	335566677889
4	122356689
5	1
6	29
7	9
8	0

Part B /2

If Y was completely determined by X, then the regression model would be

$$Y_i = \beta_0 + \beta_1 X_i.$$

Although we can see a linear relationship between X and Y in Figure 1, the relationship however is not deterministic.

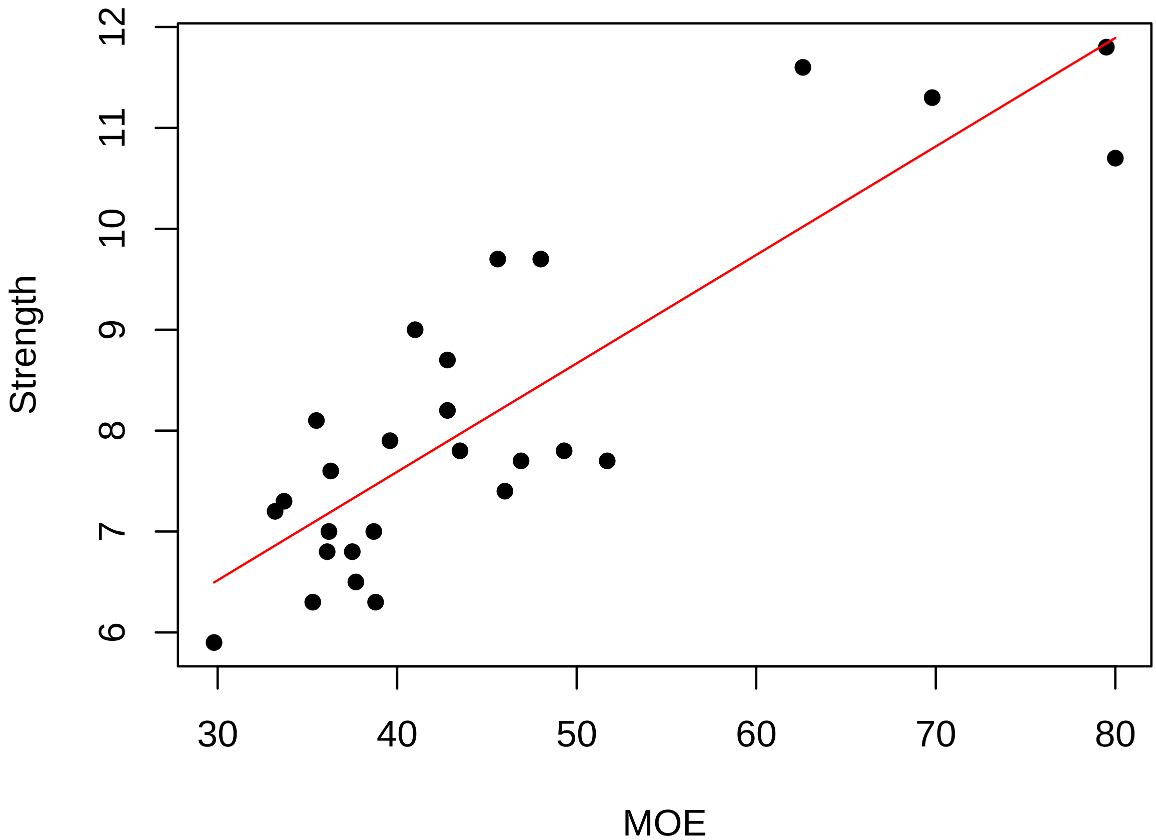

Figure 1: Scatter plot of Y vs X for data from Section 12.2 Question 15, and the estimated least squares line.

Part C /4

We can see that $\hat{\beta}_0 = 3.2925$ and $\hat{\beta}_1 = 0.10748$, so the estimated regression line is

$$\hat{Y}_i = 3.2925 + 0.10748X_i.$$

When $X = 40$, the predicted value of Y is

$$\hat{Y} = \hat{E}(Y|X = 40) = 3.2925 + 0.10748(40) = 7.5917$$

We **should not** use the regression line to predict Y when $X = 100$, since we do not have any observations near $X = 100$ in our sample. In other words, we cannot ensure that the linear relationship between X and Y holds for values of $X > 80$.

Part D /2

We can see that $SSE = 18.736$, $SST = 71.605$, and $R^2 = 1 - (SSE/SST) = 0.738$
Since the value of R^2 is quite large, we can conclude that the simple linear regression effectively describes the relationship between X and Y.

Section 12.2 Question 16 /14

Study Objective: Investigate how rainfall volume (m^3) is related to runoff volume (m^3) in a particular location.

Formulation: Let

- Y_i denote the ith measurement of runoff volume, for $i = 1, \cdots, n$, with $n = 15$.
- X_i denote the ith measurement of rainfall volume, for $i = 1, \cdots, n$.

The goal is to establish how Y_i depends on X_i:

$$Y_i = f(X_i) + \varepsilon_i,$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, X_i and ε_i are independent, and $f(X_i) = E(Y_i|X_i)$. We specify $f(\cdot)$ to be a linear function

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

with β_0 and β_1 as unknown parameters.

Part A /2

We can see in Figure 2 that there is a strong linear relationship between X and Y, which supports the use of the simple linear regression model.

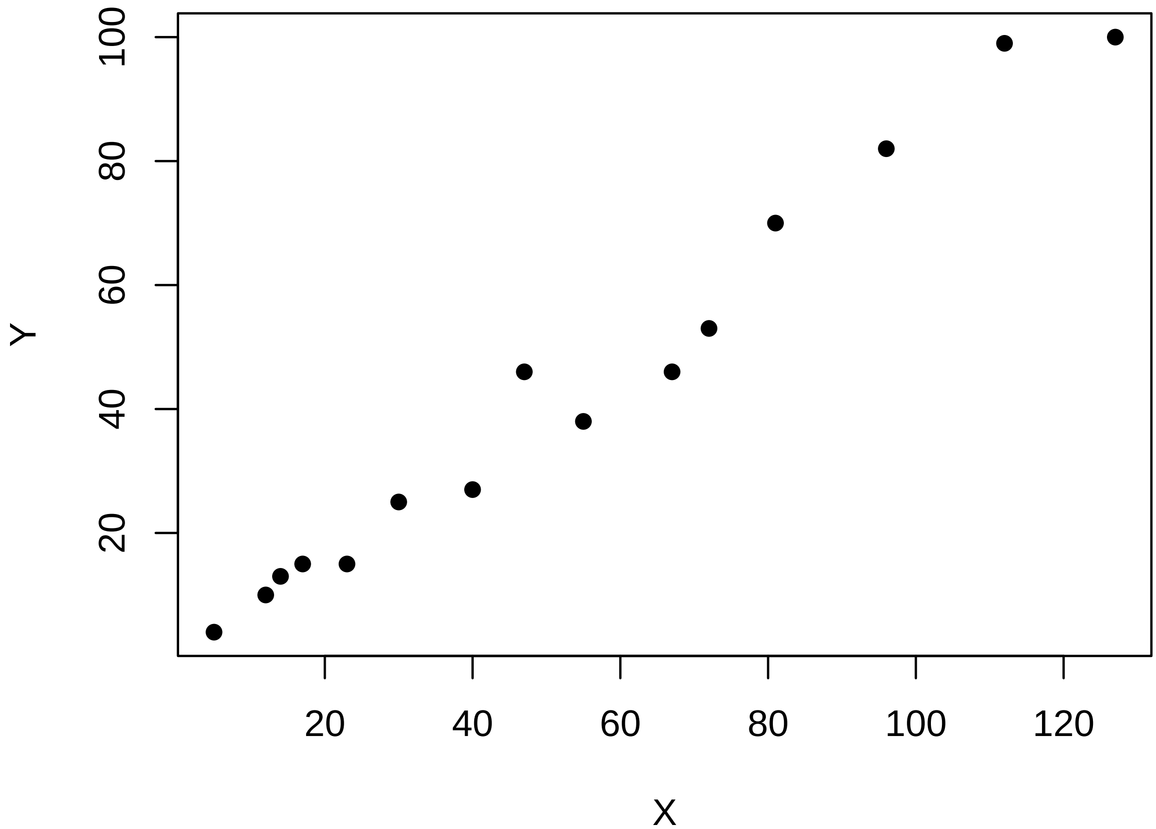

Figure 2: Scatter plot of Y vs X for data from Section 12.2 Question 16.

Part B /6

We have

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 42.86667$$
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 53.2$$
$$S_{xy} = \sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y} = 17024.4$$
$$S_{xx} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = 20586.4$$

Then point estimates for β_0 and β_1 are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \approx 0.8270$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \approx -1.1283$$

Part C /2

A point estimate of the $E(Y|X = 50)$ is

$$\hat{E}(Y|X = 50) = -1.1283 + 0.8270(50) = 40.2204$$

Part D /2

With

$$SSE = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 = 357.0117,$$

a point estimate for σ is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{SSE}{n-2}} \approx 5.2405$$

Part E /2

With

$$SST = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = 14435.73,$$

the proportion of variation in runoff volume explained by rainfall volume in the simple linear regression model is

$$R^2 = 1 - \frac{SSE}{SST} = 0.9753$$

Section 12.3 Question 32 /6

Note: This question is a continuation of Section 12.2 Question 16

Hypothesis Test: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

Test Statistic:

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t(n-2)$$

under H_0. Here, note that

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{XX}} Y_i,$$

so that

$$\begin{aligned} Var(\hat{\beta}_1) &= \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_{XX}} \right)^2 Var(Y_i) \\ &= \frac{\sigma^2}{S_{XX}}. \end{aligned}$$

From the model output, we are given that

$$\begin{aligned} \hat{\beta}_1 &= 0.82697, \\ \widehat{SE}(\hat{\beta}_1) &= \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 0.03652, \\ T_{obs} &= \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = 22.64 \end{aligned}$$

Method 1 - p-value: $P_{H_0}(|T| > |T_{obs}|) = P_{H_0}(|T| > 22.64) \approx 7.8961 \times 10^{-12}$, where $T \sim t(n-2)$.
Specifying $\alpha = 0.05$, since $0.05 > 7.8961 \times 10^{-12}$, we reject H_0.

Method 2 - Rejection Region:

$$\mathcal{R}_{0.05} = \{t : |t| > t_{\alpha/2}(n-2)\}$$

$$
\begin{aligned}
&= \{t : |t| > t_{0.025}(13)\} \\
&= \{t : |t| > 2.16\}.
\end{aligned}
$$

Since $T_{obs} \in \mathcal{R}$, we reject H_0.

A 95% confidence interval for β_1 is

$$
\begin{aligned}
\hat{\beta}_1 \pm t_{0.025}(13) \times \widehat{SE}(\hat{\beta}_1) &= 0.82697 \pm 2.16 \times 0.03652 \\
&\approx [0.7481, 0.9059]
\end{aligned}
$$

Section 12.4 Question 44 /9

Note: This question is a continuation of Section 12.2 Question 15.

For this question, let $\mu_{Y|X=x}$ denote $E(Y|X = x) = \beta_0 + \beta_1 x$ (the expected value of Y given $X = x$).

Part A /2

From the stem-and-leaf display from Section 12.2 Question 15, the majority of observations are around $X = 40$, whereas we have few observations around $X = 60$. The increased variability with large values of X is a reflection of possessing comparatively limited information.

Part B /3

A 95% confidence interval for $\mu_{Y|X=40}$ is

$$
\begin{aligned}
\hat{\mu}_{Y|X=40} \pm t_{0.025}(25) \times \widehat{SE}(\hat{\mu}_{Y|X=40}) &= 7.592 \pm 2.06 \times 0.179 \\
&\approx [7.2233, 7.9607]
\end{aligned}
$$

Part C /3

Note that $S^2 = 0.8657$ is reported in Section 12.2 Question 15.

A 95% prediction interval for $Y|X = 40$ is

$$
\begin{aligned}
\hat{\mu}_{Y|X=40} \pm t_{0.025}(25) \times \sqrt{S^2 + \widehat{SE}(\hat{\mu}_{Y|X=40})^2} &= 7.592 \pm 2.06 \times \sqrt{0.8657^2 + 0.179^2} \\
&\approx [5.7713, 9.4127]
\end{aligned}
$$

Part D /1

Using the Bonferroni technique, the simultaneous confidence level for the two intervals is $100(1 - 2 \times 0.05)\% = 90\%$.