Q1A

Parameter estimates are given by:
$\alpha_{\text{variety}_A} = 13.44$ with a 95% confidence interval (12.16, 14.73) ;
$\alpha_{\text{variety}_B} = 12.57547$ with a 95% confidence interval (11.29, 13.86);
$\alpha_{\text{variety}_C} = 12.40$ with a 95% confidence interval (11.16, 13.64) ;
$\alpha_{\text{variety}_D} = 12.59$ with a 95% confidence interval (11.34, 13.84);
$\gamma = -1.97$ with a 95% confidence interval (-3.26 ,-0.69);
$\delta = 0.05$ with a 95% confidence interval (-0.03 , 0.14).
The code from R that I used for this is:

```
cs3=read.csv("courseworkdata3.csv")
yi=cs3$yield
va=cs3$variety
so=cs3$soil
ra=cs3$rainfall
clay=1*(so=="clay")
lm1=lm(yi~0+factor(va)+clay+I(ra-mean(ra)))
confint(lm1)["factor(va)A",]
confint(lm1)["factor(va)B",]
confint(lm1)["factor(va)C",]
confint(lm1)["factor(va)D",]
confint(lm1)["clay",]
confint(lm1)["I(ra-mean(ra))",]
```

Q1B

The parameter α_A is the expected yield of wheat in tonnes per hectare for an experimental unit (land) of variety A, with soil type that is not clay, rainfall equal to mean rainfall.

Q1C

We let $\beta_A = \mu = \alpha_A$. We do not really care whether Variety A,B,C or D will be the intercept. The rest of the relationships are:
$\beta_B = \alpha_B - \mu$
$\beta_C = \alpha_C - \mu$
$\beta_D = \alpha_D - \mu$
The code from R that I used for this is:

```
lm2=lm(yi~factor(va)+clay+I(ra-mean(ra)))
```

Q1D

source	d.f	SS	MS	F
regression	5	42.31	8.64	2.52
residual	34	114.10	3.35	
total	39	156.42	19597.1	

The p-value is 0.04 and so we reject H_0 and conclude that the explanatory variables have impact on the outcome.

The code from R that I used for this is:

```
anova(lm)
summary(lm)
```

Q1E

Our hypotheses:

$H_0 : \gamma = 0, H_1 : \gamma \neq 0$

We calculate the p-value:

The corresponding t-value is -3.13, so, we calculate it with the code 2*pt(abs(-3.137),df=34,lower.tail = FALSE) and we get 0.003 which is smaller than 0.025, which is the usual significance level for the t-test. Hence, we reject H_0. We may conclude that the fact whether the experimental unit has clay or not does significantly impact the amount of yield.

Q1F

Our hypotheses:

$H_0 : \beta_A = \beta_B = \beta_C = \beta_D, H_1 : \beta_A \neq \beta_B \neq \beta_C \neq \beta_D$

Using the Analysis of variances test (ANOVA), we see that the p-value is equal to 0.59, so we fail to reject H_0. Alternatively, we calculate the F-value, which is $0.63 < 3.61 = F_{4,34}(0.99)$ so, we fail to reject H_0 and conclude that the varieties of wheat have the same effect on yield.

The code from R that I used for this is:

```
> lm22=lm(yi~clay+I(ra−mean(ra))+factor(va))
> anova(lm22)
Analysis of Variance Table

Response: yi
                   Df  Sum Sq Mean Sq F value   Pr(>F)
clay                1  27.956 27.9560  8.3298 0.00673 **
I(ra − mean(ra))    1   7.937  7.9374  2.3650 0.13334
factor(va)          3   6.421  2.1404  0.6378 0.59591
Residuals          34 114.109  3.3561
—
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1
```

Q1G

After a follow-up test we observe no significant difference in effect on yield between variety A and variety C (difference = 1.04, SE= 0.85), p-value = 1.00.
CODE:

```
install.packages(”emmeans”)
library(”tidyverse”)
library(”emmeans”)
em1 <− emmeans(lm, ”variety”)
em1
em1 %>%
  pairs() %>%
  update(by = NULL) %>%
  summary(adjust = ”holm”)
```

Q2A

Suppose the linear model in algebraic form is given by:

$$\text{profit} = \alpha + \beta\text{GRSMIN} + \gamma\text{MIN} + \delta\text{MAX} + \zeta\text{HUM} + \eta\text{RAIN} + \kappa\text{BAR}$$

So the values are:
$\alpha = -116.03$
$\beta = 1.54$
$\gamma = -1.31$
$\delta = 1.11$
$\zeta = 0.18$
$\eta = -1.10$
$\kappa = 0.19$
unbiased estimate for the variance of the errors is given by 378.30
The code from R that I used for this is:

```
cs4=read.csv(”courseworkdata4.csv”)
grsmin=cs4$GRSMIN
min=cs4$MIN
max=cs4$MAX
hum=cs4$HUM
rain=cs4$RAIN
bar=cs4$BAR
profit=cs4$PROFIT
lm2=lm(profit~grsmin+min+max++hum+rain+bar)
sigma(lm2)^2
```

Q2B

Table representing order of statistical significance increasing from top to bottom

variable	$\mathbf{P}(> \|t\|)$	$\mathbf{P}(> F)$
MAX	0.01	0.005
RAIN	0.04	0.009
BAR	0.27	0.27
GRSMIN	0.28	< 0.001
MIN	0.37	0.73
HUM	0.43	0.45

The code from R that I used for this is:

```
lm2=lm(profit~grsmin+min+max++hum+rain+bar)
 anova(lm2)
```

Q2C

variable	$\mathbf{P}(>\lvert t\rvert)$	$\mathbf{P}(>F)$
MAX	1.5e-09	< 0.001
GRSMIN	6.79e-08	< 0.001
MIN	1.77e-07	< 0.001
RAIN	0.004	0.004
HUM	0.301	0.301
BAR	0.394	0.394

The code from R that I used for this is:

```
lmgrsmin=lm(profit~grsmin)
lmmin=lm(profit~min)
lmmax=lm(profit~max)
lmhum=lm(profit~hum)
lmrain=lm(profit~rain)
lmbar=lm(profit~bar)
```

Q2D

The correlation between GRSMIN and MIN is 0.98

The difference in p-values:

In the multivariate model, the p-value is based on the marginal impact of the variable, given all the other variables i.e. probability of the null hypothesis $\beta_{\text{GRSMIN}} = 0$ given that all other parameter values are estimated.

In the univariate model, the p-value is the impact of the variable without considering any other variables.

By the high correlation coefficient, we can see that the variables have high correlation between each other, and in the model in part 2(b) they both impact the outcome significantly. Taking each variable away from each other, as we do in 2(c), we then see that as isolated variables the impact is way less significant.

The code from R that I used for this is:

```
cor.test(grsmin,min,method="pearson")
```

Q2E

Ranking them in decreasing order (1st is best fit)

1)MAX Adjusted R-squared: 0.17 MSE=387.2

2)GRSMIN Adjusted R-squared: 0.14 MSE=403.4

3)MIN Adjusted R-squared: 0.13 MSE=407.6

4)RAIN Adjusted R-squared: 0.039 MSE=452.9

5)HUM Adjusted R-squared: 0.0004 MSE=471.2

6)BAR Adjusted R-squared: -0.001 MSE=472.1

The statistic I use to compare the best fit is the adjusted R squared, which is shown in the lm() output. When the adjusted r squared is greater in value, the goodness of fit is greater. We can also use the residuals sum of square as our statistic.

Q2F

Prediction from univariate model is 78.57 and from best univarite (MAX) model is 116.3. Prediction intervals are $(30.14, 126.99)$ and $(77.37, 155.23)$ respectively.

The code from R that I used for this is:

```
> pred=predict(lm2,newdata=list(grsmin=10.5,min=11,max=16,hum=84,
rain=30.1,bar=987),interval='predict')
> predd=predict(lmmax,newdata=list(grsmin=10.5,min=11,max=16,hum=84,
rain=30.1,bar=987), interval='predict')
```