STP 531

Analysis of Variance

FINAL PROJECT REPORT

Kushal Kapadia 1213714799

Mitravinda Harish 1210268394

Block	Variety			
	1	2	3	4
1	9.7	11.8	6.3	4.6
2	6.6	9.7	5.3	3.4
3	7.6	10.9	4.7	2.3
4	8.1	11.3	5.5	3.6
5	6.4	10.7	4.5	2.8

Question 1)

As from the table, we can see that there are 5 blocks and 4 different varieties of wheat. A Randomized Complete Block Design (RCBD) may be very appropriate to use here and can be viewed as corresponding to a two-factor study (blocks and treatments are the factors).

PART-A:

Here, as described, the appropriate ANOVA model is Randomized Complete Block Design (RCBD) which can be given as:

$Y_{ij} = \mu_{..} + p_i + \tau_j + \varepsilon_{ij}$

Talking about assumptions, there are two important assumptions that we can check here. They are:

1) The error term is distributed as $N(0, \sigma^2)$ or not. In this, we can check:

- Constancy of variances

- Normality

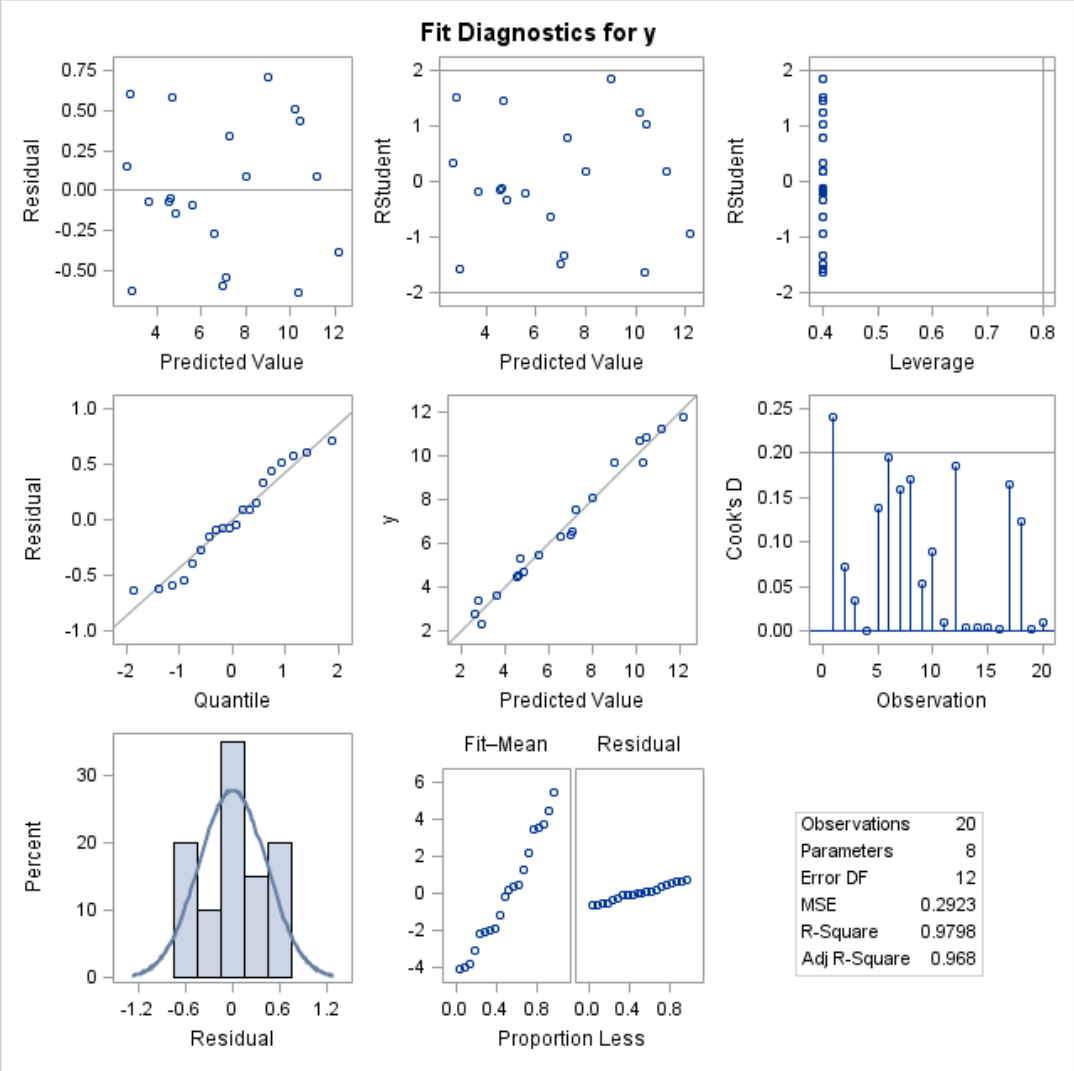
- Independence of error terms

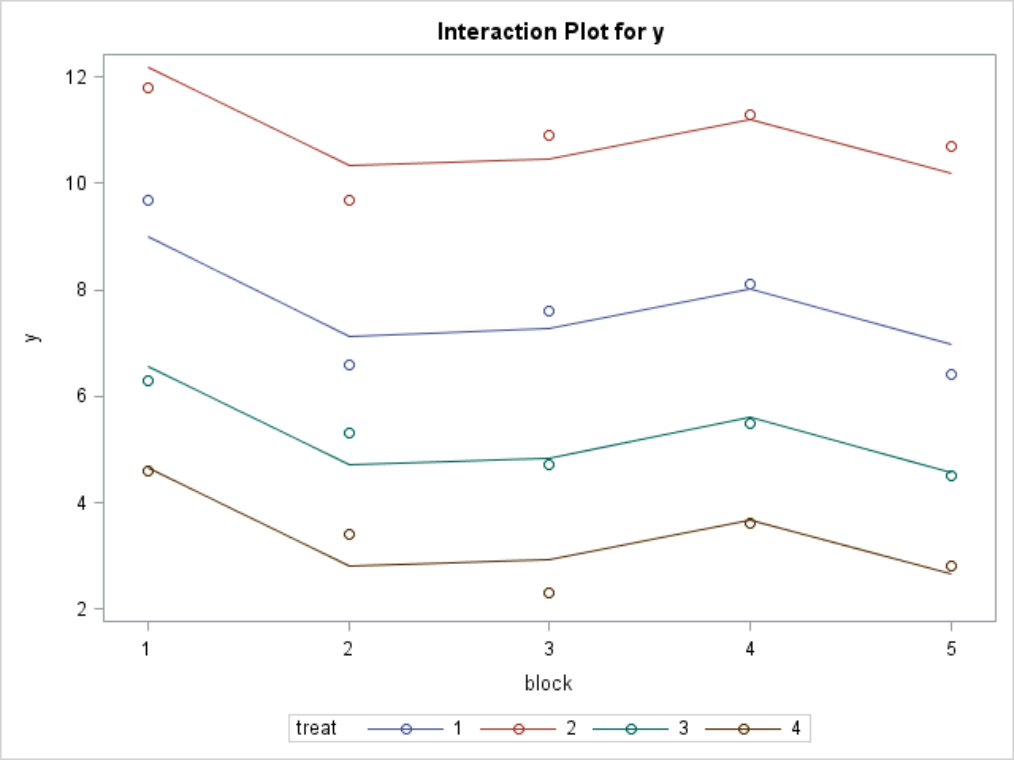
2) No important interactions

Now, for checking the assumptions we will require the SAS code. Here it is:

```
*** F-test, pairwise comparison & diagnostics;
ods graphics on;
proc glm data=wheat plot= DIAGNOSTICS;
 class block treat;
 model y = block treat;
 lsmeans treat/adjust=tukey cl;
 lsmeans treat/adjust=scheffe cl;
 lsmeans treat/adjust=bon cl;
run;
ods graphics off;
```

Figure for checking the assumptions:



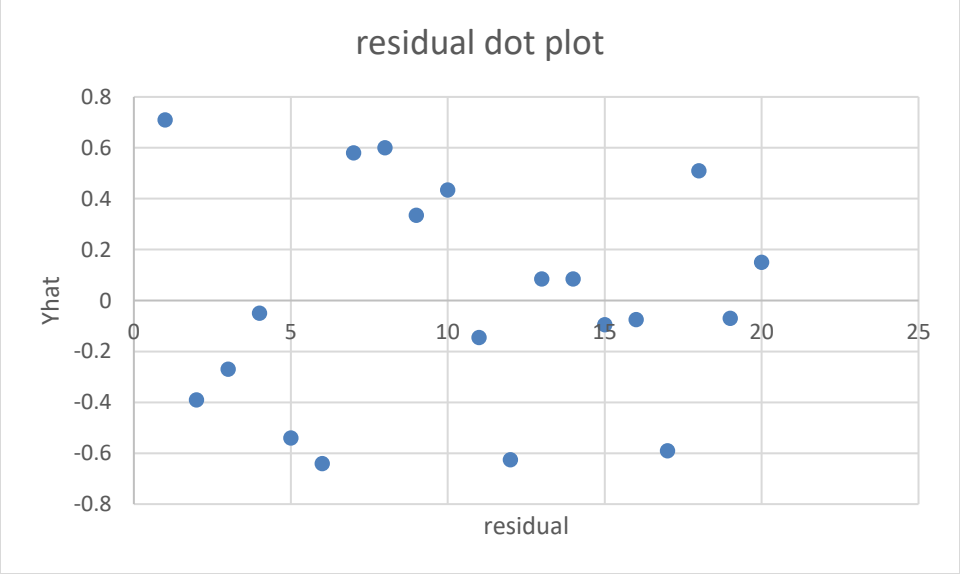


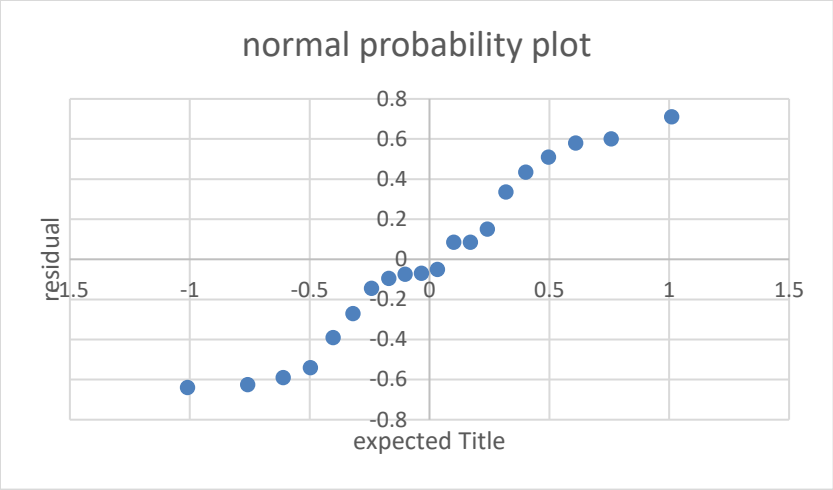
Above are the two relevant figures to check the assumptions:

From the first figure, specifically from the plot of residual vs Predicted Value, we can see the constancy of error variance and from the normal probability plot (Residual vs quantile), we can see that the line is almost straight with slight departures which is alright.

From the second figure, we can see that there is no sign of interaction present between the treatments.

A better way to visualize both the plots:





PART-B:

Here, we need to test whether the effect of variety is present or not. For that, we'll have to use F-test. Also, the ANOVA model from the SAS obtained is:

The SAS System

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	169.8910000	24.2701429	83.05	<.0001
Error	12	3.5070000	0.2922500		
Corrected Total	19	173.3980000			

R-Square	Coeff Var	Root MSE	y Mean
0.979775	7.961731	0.540602	6.790000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	4	11.0730000	2.7682500	9.47	0.0011
treat	3	158.8180000	52.9393333	181.14	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
block	4	11.0730000	2.7682500	9.47	0.0011
treat	3	158.8180000	52.9393333	181.14	<.0001

Alternatives: H_0: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ and H_a: Not all τ_j equal zero (j=1,2,3,4)

Test Statistic F* = MSTR/MSBL.TR = 52.93/0.292 = 181.14 as we can see from the above table as well.

Also, α=0.05 and F(0.95,3,12) = 3.49

Decision rule for controlling error rate at α is:

If $F^* \leq F(0.95,3,12)$, conclude H_0, otherwise conclude H_a. Here, $F^* > F(0.95,3,12)$ and thus we conclude H_a.

P-value here is <0.0001.

PART-C:

To obtain the confidence intervals from all pairwise comparisons between the treatment means, we'll take into consideration 3 procedures here namely Tukey, Scheffe and Bonferroni. The SAS code is already provided during the start of the question and so I'm not copying the same thing here. The output of SAS for each of the procedure is shown below:

First the LSmean confidence intervals of each mean:

treat	y LSMEAN	90% Confidence Limits	
1	7.680000	7.249106	8.110894
2	10.880000	10.449106	11.310894
3	5.260000	4.829106	5.690894
4	3.340000	2.909106	3.770894

Tukey procedure:

Least Squares Means for Effect treat

i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-3.200000	-4.075351	-2.324649
1	3	2.420000	1.544649	3.295351
1	4	4.340000	3.464649	5.215351
2	3	5.620000	4.744649	6.495351

Least Squares Means for Effect treat

i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
2	4	7.540000	6.664649	8.415351
3	4	1.920000	1.044649	2.795351

Scheffe procedure:

Least Squares Means for Effect treat

i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-3.200000	-4.155907	-2.244093
1	3	2.420000	1.464093	3.375907
1	4	4.340000	3.384093	5.295907
2	3	5.620000	4.664093	6.575907
2	4	7.540000	6.584093	8.495907
3	4	1.920000	0.964093	2.875907

Bonferroni procedure:

Least Squares Means for Effect treat

i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-3.200000	-4.150320	-2.249680
1	3	2.420000	1.469680	3.370320
1	4	4.340000	3.389680	5.290320
2	3	5.620000	4.669680	6.570320
2	4	7.540000	6.589680	8.490320
3	4	1.920000	0.969680	2.870320

From all the 3 procedures above, we can see the confidence intervals for Tukey procedure is the narrowest for the pairwise comparisons and so Tukey procedure is the most efficient here comparison procedure with 90% family confidence coefficient.

PART-D:

Now, here we need to determine the difference in mean plant heights for the first two groups of variety ($\mu_{.1}$- $\mu_{.2}$) with 95% confidence interval. As it's a single comparison, we can use t-test here to find the confidence interval as follows:

Here, $\hat{D}$ = 7.68 – 10.88 = -3.2

$s^2\{\hat{D}\} = MSE(1/n_1 + 1/n_2) = 0.292\ (1/5 + 1/5) = 0.292*0.4 = 0.1168$. Thus, $s\{\hat{D}\} = 0.3417$

And t(0.975;12) = 2.179

Thus, the 95% confidence interval is $-3.2-(0.3417)*2.179 \leq \mu_{.1}- \mu_{.2} \leq -3.2+(0.3417)*2.179$

$$-3.944 \leq \mu_{.1}- \mu_{.2} \leq -2.455$$

PART-E:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	169.8910000	24.2701429	83.05	<.0001
Error	12	3.5070000	0.2922500		
Corrected Total	19	173.3980000			

Testing of H_0: $\mu_{.1} = \mu_{.2}$ and H_a: $\mu_{.1} \neq \mu_{.2}$

Here, $\hat{D}$ = 7.68 – 10.88 = -3.2

$s^2\{\hat{D}\} = MSE(1/n_1 + 1/n_2) = 0.292\ (1/5 + 1/5) = 0.292*0.4 = 0.1168$. Thus, $s\{\hat{D}\} = 0.3417$

Also, Test Statistic $t^* = \hat{D}/ s\{\hat{D}\} = -9.364$ and t(0.975;12) = 2.179

Now, if $|t^*| \leq t(0.975; 12)$, conclude H_0, otherwise conclude H_a. Here, $|t^*| > 2.179$ and thus we conclude alternative hypothesis that they are not equal.

P-value is <0.0001. And yes, my conclusion of $\mu_{.1} \neq \mu_{.2}$ matches to that of my conclusion in part (d).

PART-F:

Testing of H_0: $\mu_{.1} \geq \mu_{.2}$ and H_a: $\mu_{.1} < \mu_{.2}$

Here, $\hat{D} = 7.68 - 10.88 = -3.2$

$s^2\{\hat{D}\} = MSE(1/n_1 + 1/n_2) = 0.292\ (1/5 + 1/5) = 0.292*0.4 = 0.1168$. Thus, $s\{\hat{D}\} = 0.3417$

Also, Test Statistic $t^* = \hat{D}/ s\{\hat{D}\} = -9.364$ and $t(0.975;12) = 2.179$

Now, if $|t^*| \leq t(0.975; 12)$, conclude H_0, otherwise conclude H_a. Here, $|t^*| > 2.179$ and thus we conclude alternative hypothesis that $\mu_{.1} < \mu_{.2}$.

PART-G:

The efficiency can be found out as follows:

$\hat{E} = ([(n_b-1)MSBL + n_b(r-1)MSE] / (n_b r - 1)) \ / \ MSE$

$\hat{E} = ([(5-1)*2.768 + 5(4-1)*0.292] / (5*4-1)) \ / \ 0.292$

$\hat{E} = ([4*2.768 + 15*0.292] / 19) / 0.292$

$\hat{E} = ([11.072 + 4.38] / 19) / 0.292$

$\hat{E} = 2.785$

Question 2)

PART-A:

The mixed effects model when the four varieties are randomly selected from a population of of 20 varieties of wheat can be given as follows:

$Y_{ij} = \mu_{..} + p_i + \tau_j + \varepsilon_{ij}$

Assumptions:

- $\mu_{..}$ is a constant

- τ_j are independent $N(0, \sigma_p^2)$

- p_i are constants subject to the restriction $\sum p_i = 0$

- ε_{ij} are independent $N(0, \sigma^2)$ and independent of τ_j

$I=1,\ldots,n_b$ and $j=1,\ldots,r$

PART-B:

The SAS code for the ANOVA mixed effect model is shown below:

```
proc glm data=wheat;
class block treat;
model y = block treat;
random treat;
run;
```

The ANOVA model output from SAS is as follows:

full model: multiple comparisons

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	169.8910000	24.2701429	83.05	<.0001
Error	12	3.5070000	0.2922500		
Corrected Total	19	173.3980000			

R-Square	Coeff Var	Root MSE	y Mean
0.979775	7.961731	0.540602	6.790000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	4	11.0730000	2.7682500	9.47	0.0011
treat	3	158.8180000	52.9393333	181.14	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
block	4	11.0730000	2.7682500	9.47	0.0011
treat	3	158.8180000	52.9393333	181.14	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F

full model: multiple comparisons

The GLM Procedure

Source	Type III Expected Mean Square
block	Var(Error) + Q(block)
treat	Var(Error) + 5 Var(treat)

So, basically here we are trying to make inferences about the difference in plant heights among the 20 varieties of wheat. We are basically trying to find the mean plant height and variation in the mean plan height produced by the 20 different plant heights included as the observations. Now, we'll have to resort to F-test here to find the difference in plant heights across different varieties of wheat. We can get F* directly from the ANOVA table.

PART-C:

Now, we want to check whether or not there is a statiscally significant difference in the plant heights across different varieties of wheat or not. We'll check it as follows:

Alternatives: H_0: $\sigma_u^2 = 0$ and H_a: $\sigma_u^2 > 0$
Test Statistic from the ANOVA table: F* = MSTR/MSBL.TR = 52.93/0.292 = 181.14 and also F(0.95;3;12) = 3.49
Decision rule: If F*≤ F(0.95;3;12), do not reject H_0, otherwise reject H_0.
Conclusion: Here, F*>3.49 and thus we reject H_0 and accept the alternative hypothesis that there is indeed a statistically significant difference in the plant heights across different varieties of wheat.
P-value is <0.0001.

Question 3)

PART-A:

The statistical model for analyzing the data can be given as:

$Y_{ij} = \mu_. + \tau_i + \Upsilon(X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$

Model Assumptions:

- ε_{ij} are independent and have constant variance

- All the treatment regression lines must be identical to each other

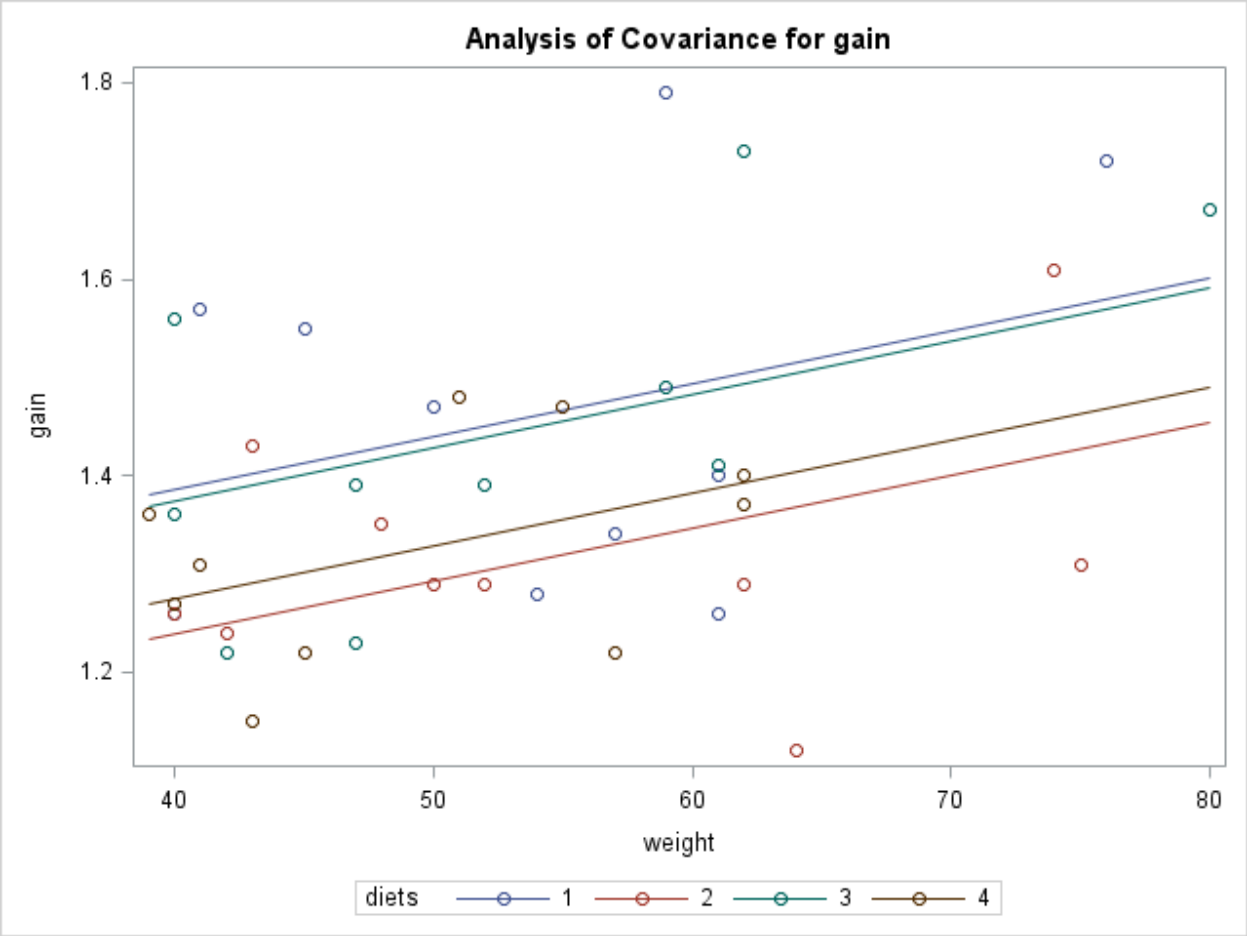
- Linear relation assumption between Y and concomitant variable

- All the treatment regression lines have the same slope

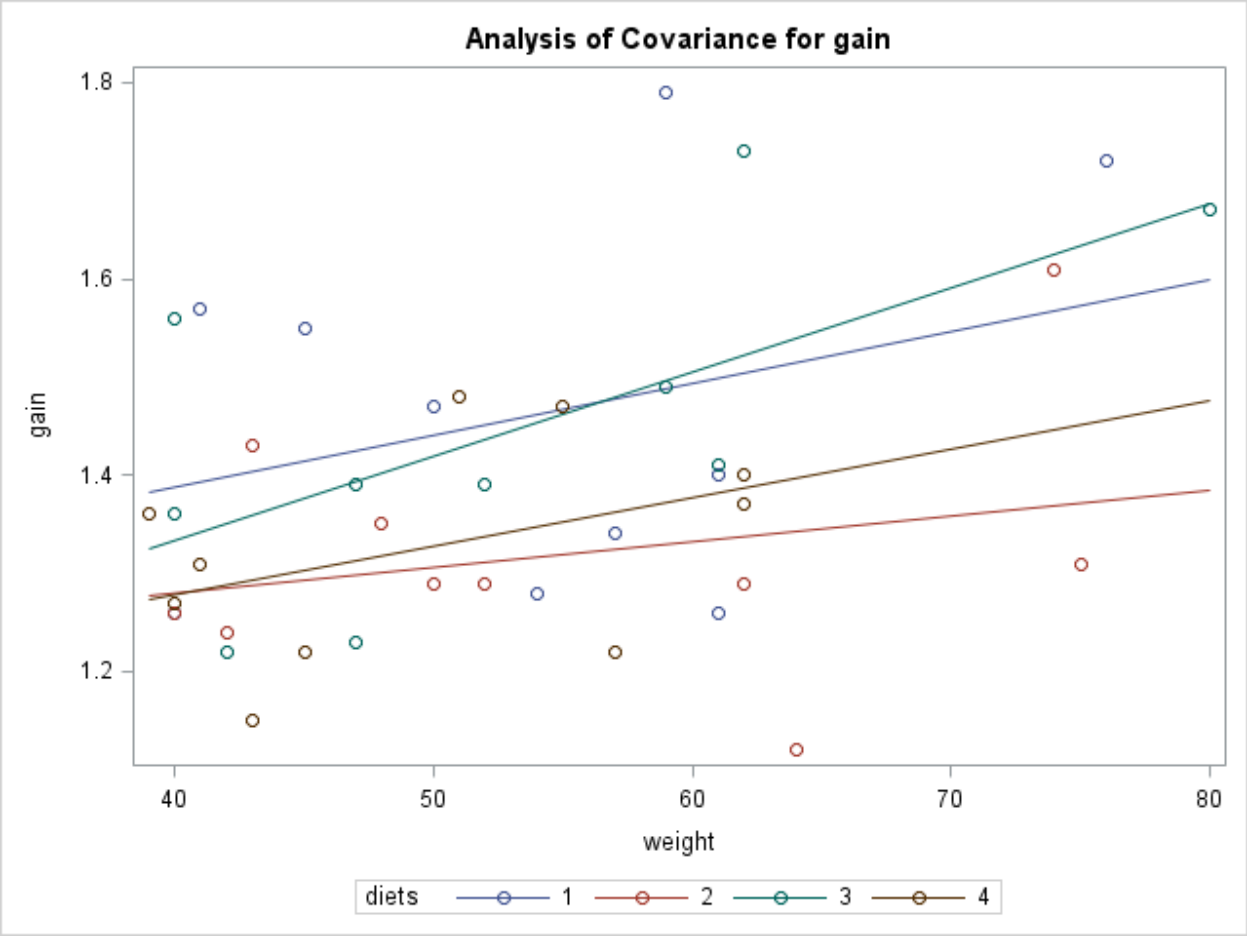
Checking the model assumptions:

The SAS code for ANCOVA model and the output is:

```
ods graphics on;
proc glm data=diet;
class diets;
model gain = weight diets/solution clparm alpha=0.05; *for parameter
estimates;
estimate 'mu_2(xbar)' intercept 1 diets 0 1 0 0 weight 56;
run;
ods graphics off;
```



This plot has been obtained without considering the interaction term and the plot I'm attaching below has been obtained considering the interaction term.



I was somewhat skeptical at this point as to whether we should take interaction term into consideration or not. So, I did the interaction test (not included here. I just used the F-test to conclude that interaction effects are not significant here).

Thus from the figure, linear regression and parallel slopes for the treatment regression lines appear to be reasonable.

PART-B:

The point estimate and a 95% confidence interval for the effect of the initial weight on the weight gain can be given as follows. This code is already included in the SAS code provided above. So, I'm just attaching the output here:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
weight	0.005384841	0.00205854	2.62	0.0130	0.001205780	0.009563903

So, it's very clear from the output that the point estimate for the initial weight on the weight gain is 0.005384 and the 95% confidence interval is [0.00120, 0.00956]. Thus, we can say that the initial weight lies somewhere in the given interval with 95% confidence.

PART-C:

The SAS code and the output to obtain the confidence interval to estimate the mean weight gain for pigs in the second treatment group that have an initial weight of 56 pounds are given below:

```
estimate 'mu_2(xbar)' intercept 1 diets 0 1 0 0 weight 56; (Already included
in the SAS code provided during the start of this question)
```

Output:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
mu_2(xbar)	1.32438484	0.04499173	29.44	<.0001	1.23304676	1.41572292

So, the 95% confidence interval is [1.2330, 1.4157].

PART-D:

Here, in this part, we need to obtain the confidence intervals for all pairwise comparisons between the treatment means. So, we're going to compare to Bonferroni and Scheffe here for multiple pairwise comparisons. We're not going to take into account the Tukey method here because Tukey method does not work so well for covariance analysis.

The SAS code to obtain the 90% confidence intervals of Bonferroni and Tukey methods is given below:

```
proc glm data = diet alpha=0.1;
 class diets;
 model gain = weight diets;
 lsmeans diets / cl adjust=scheffe;
 lsmeans diets / cl adjust=bon;
run;
```

Output for Tukey procedure:

Least Squares Means for Effect diets

i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.148231	-0.016840	0.313302

Least Squares Means for Effect diets

i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	3	0.011461	-0.153748	0.176671
1	4	0.112614	-0.054491	0.279719
2	3	-0.136770	-0.302155	0.028616
2	4	-0.035617	-0.203254	0.132021
3	4	0.101153	-0.064944	0.267250

Output for Bonferroni procedure:

Least Squares Means for Effect diets

i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.148231	-0.011627	0.308089
1	3	0.011461	-0.148531	0.171453
1	4	0.112614	-0.049213	0.274442
2	3	-0.136770	-0.296932	0.023393
2	4	-0.035617	-0.197960	0.126727
3	4	0.101153	-0.059698	0.262004

As very clear from the confidence bounds, Bonferroni is the most efficient here as it has got narrower intervals than Scheffe.

Question 4)

PART-A:

If we consider both the concomitant and blocking variable into the model, then we'll have to resort the ANCOVA model again but including blocking variable this time. Given as:

$Y_{ij} = \mu_. + p_j + \tau_i + \Upsilon(X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$

p_j = blocking variable having j=1,2

τ_i = treatment effects

The SAS code:

Full model:

```
proc glm data=diet;
class age diets;
model gain = weight age diets/solution clparm alpha=0.1; *for parameter
estimates;
estimate 'mu_2(xbar)' intercept 1 diets 0 1 0 0 weight 56;
run;
ods graphics off;
```

Reduced model:

```
ods graphics on;
proc glm data=diet;
class diets;
model gain = weight age/solution clparm alpha=0.1; *for parameter estimates;
estimate 'mu_2(xbar)' intercept 1 diets 0 1 0 0 weight 56;
run;
ods graphics off;
```

The full model and reduced model obtained from SAS are give below:

Full model:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.34765092	0.06953018	3.50	0.0117
Error	34	0.67512658	0.01985666		
Corrected Total	39	1.02277750			

R-Square	Coeff Var	Root MSE	gain Mean
0.339909	10.15045	0.140914	1.388250

Source	DF	Type I SS	Mean Square	F Value	Pr > F
weight	1	0.15345746	0.15345746	7.73	0.0088
age	1	0.01441400	0.01441400	0.73	0.4002
diets	3	0.17977946	0.05992649	3.02	0.0432

Source	DF	Type III SS	Mean Square	F Value	Pr > F
weight	1	0.15561489	0.15561489	7.84	0.0084
age	1	0.03187993	0.03187993	1.61	0.2137
diets	3	0.17977946	0.05992649	3.02	0.0432

Reduced model:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.16787147	0.08393573	3.63	0.0363
Error	37	0.85490603	0.02310557		
Corrected Total	39	1.02277750			

R-Square	Coeff Var	Root MSE	gain Mean
0.164133	10.94941	0.152005	1.388250

Source	DF	Type I SS	Mean Square	F Value	Pr > F
weight	1	0.15345746	0.15345746	6.64	0.0141
age	1	0.01441400	0.01441400	0.62	0.4347

Source	DF	Type III SS	Mean Square	F Value	Pr > F
weight	1	0.13492755	0.13492755	5.84	0.0207
age	1	0.01441400	0.01441400	0.62	0.4347

Now, we have to test whether the main effects of diet are present by assuming that all level combinations of diet and initial age are equally important or not. We'll use F-test to compare that as follows:

Alternatives: H_0: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ and H_a: Not all τ_i are zero where i=1,2,3,4

Here, Test Statistic F* from the ANCOVA model is 3.02 and F(0.90;3,34) = 2.20

If $F^* \leq 2.20$, conclude H_0. Otherwise conclude H_a. Here, $F^* > 2.20$ and thus we conclude alternative hypothesis that the not all treatment effects are zero i.e. they are present in the model.

PART-B:

Now, here we need to divide up our data into 75% of 1's in the age (blocking variable) and 25% of 2's in the same variable. So, the SAS code for this is:

```
proc glm data=diet order=data alpha=0.1;
class age diets;
model gain = weight age diets age|diets/ss3 clparm; *for parameter estimates;
estimate 'L1' intercept 1 diets 1 0 0 0 age 0.75 0.25 age*diets 0.75 0 0 0
0.25 0 0 0 ;
estimate 'L2' intercept 1 diets 0 1 0 0 age 0.75 0.25 age*diets 0 0.75 0 0 0
0.25 0 0 ;
estimate 'L3' intercept 1 diets 0 0 1 0 age 0.75 0.25 age*diets 0 0 0.75 0 0
0 0.25 0 ;
```

```
estimate 'L4' intercept 1 diets 0 0 0 1 age 0.75 0.25 age*diets 0 0 0 0.75 0
0 0 0.25 ;
run;
```

The output:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
L1	1.07036466	0.14939913	7.16	<.0001	0.81705564	1.32367369
L2	0.90869534	0.15424780	5.89	<.0001	0.64716530	1.17022539
L3	1.01650967	0.14965446	6.79	<.0001	0.76276773	1.27025161
L4	0.92465711	0.14769794	6.26	<.0001	0.67423248	1.17508175

Alternatives: H_0: $\Gamma_1 = \Gamma_2 = \Gamma_3 = \Gamma_4 = 0$ and H_a: Not all Γ_i are zero where i=1,2,3,4

The test Statistic t* of each treatments are already shown in the output under the column "t Value" i.e. t* = L/s{L} and also $t(1-\alpha/2; n_T - ab) = 1.697$

Decision rule: If $t^* \leq t(1-\alpha/2; n_T - ab)$, conclude H_0, otherwise conclude H_a.

Here, t*>1.697 as evident from the output of SAS code. So, we conclude the alternative hypothesis that the main effects of treatments under the given population are present.

PART-C:

So, part-c wants us to find the 95% confidence interval of the second treatment group with the value of initial weight as 56 pounds. The SAS code is given below:

```
estimate 'mu_2(xbar)' intercept 1 diets 0 1 0 0 age 0.75 0.25 age*diets 0
0.75 0 0 0 0.25 0 0  weight 56;
```

The output:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
mu_2(xbar)	1.43701463	0.09071923	15.84	<.0001	1.25199154	1.62203773

So, we are 95% confident that the value of the mean weight gain in the second treatment group is the range [1.251,1.622].

PART-D:

Now, we have to test whether the mean weight gain in the second obtained in part-c is positive or not. Putting the words into action, it looks something like this:

$H_0: \mu_{.2} \leq 0$ and $H_a: \mu_{.2} > 0$

The test statistic t* from the SAS code already provided in the part-c is $t^* = \mu_{.2}/s\{\mu_{.2}\} = 5.89$ and also $t(1-\alpha/2; n_T - ab) = 1.697$

Decision rule: If $t^* \leq t(1-\alpha/2; n_T - ab)$, conclude H_0, otherwise conclude H_a.

Here, $t^* > 1.697$. So, we conclude the alternative hypothesis that the mean weight gain of the second treatment group is positive.

P-value is <0.0001.