STAT 3A03 Applied Regression Analysis With SAS

Fall 2017

Assignment 4 Solution Set

Q. 1 **a)** The assumptions for this model are that the error terms ε_i are independently and normally distributed with zero mean and equal variance for each level of the categorical variable X. **[3 marks]**

If this is true then

$$\mathrm{E}[Y \mid X = k] \;=\; \mathrm{E}[Y \mid E_k = 1; E_j = 0 \; j \neq k] \;=\; \beta_k.$$

Hence the β_k is the population mean of the response variable Y when $X = k$. In other words $\beta_1, \ldots, \beta_p$ are the means of the response for the p populations defined by the values of X. **[3 marks]**

b) Let $\boldsymbol{X}$ be the design matrix for this model. Then each row of $\mathbf{X}$ will have exactly one 1 and $p-1$ 0s. If we let n_j be the number of observations with $X = j$ then column j of the matrix will have n_j 1s and $n - n_j$ 0s. The diagonal entries in $\boldsymbol{X}^t\boldsymbol{X}$ are simply the sum of squares of the entries in each column. Since the entries in each column are 0 or 1, this is just the sum of the column or the number of 1s and so the the j^{th} diagonal entry of $\boldsymbol{X}^t\boldsymbol{X}$ is n_j. For the off-diagonal entries, we simply note that since it is impossible for $X = j$ and $X = k$ at the same time, hence at most one of E_j or E_k can be 1 so the product is 0.
Hence we have that

$$\boldsymbol{X}^t\boldsymbol{X} \;=\; \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_p \end{pmatrix}$$

Since this is a diagonal matrix, its inverse is also a diagonal matrix with elements equal to the reciprocals of the diagonal elements in the original matrix. **[5 marks]**
Also we have that

$$\boldsymbol{X}^t\boldsymbol{Y} \;=\; \begin{pmatrix} n_1\overline{y}_1 \\ n_2\overline{y}_2 \\ \vdots \\ n_p\overline{y}_p \end{pmatrix}.$$

where $\overline{y}_j$ is the sample mean of the response variable for all those observations with $X = j$.
Hence the least squares estimates of $\beta_1, \ldots, \beta_p$ are

$$\hat{\boldsymbol{\beta}} \;=\; \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{Y} \;=\; \begin{pmatrix} n_1^{-1} & 0 & \cdots & 0 \\ 0 & n_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_p^{-1} \end{pmatrix} \begin{pmatrix} n_1\overline{y}_1 \\ n_2\overline{y}_2 \\ \vdots \\ n_p\overline{y}_p \end{pmatrix} \;=\; \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_p \end{pmatrix}.$$

Or we can write $\hat{\beta}_j \;=\; \overline{y}_j$ for $j = 1, \ldots, p$. **[5 marks]**

c) From our standard linear model results we know that the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\text{Var}(\hat{\boldsymbol{\beta}}) \;=\; \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\sigma^2 \;=\; \boldsymbol{C}\sigma^2.$$

From the work above we know that $\boldsymbol{C}$ is a diagonal matrix with $C_{jj} \;=\; n_j^{-1}$. Hence we have that

$$\text{Var}(\hat{\beta}_j) \;=\; C_{jj}\sigma^2 \;=\; \frac{\sigma^2}{n_j}$$

Furthermore if $j \neq k$ then

$$\text{Cov}(\hat{\beta}_j, \hat{\beta}_k) \;=\; C_{jk}\sigma^2 \;=\; 0.$$

[4 marks]

d) The usual t-test for the parameter β_j tests $H_0 : \beta_j = 0$. Since we know that $\beta_j \;=\; \text{E}\big(Y \mid X = j\big)$ this is testing whether the response has mean 0 in category j.

The null hypothesis for the F-test is $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$. Again this is testing if the population mean of the response variable in each of the p populations indexed by the categorical variable are all equal to 0. These are typically not very meaningful. It is usually more meaningful to test the less restrictive hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p$ (without the added constraint that the common value be 0). This cannot be tested from directly from this model. Also in the t-test it is usually more useful to test if the mean of Y when $X = j$ is different from the global mean of Y and that global mean need not be 0. This is why we use the somewhat more complicated coding that I described in class for categorical variables which allows us to test these hypotheses directly.

[5 marks]

Q. 2 **a)** This is a simple ANOVA analysis with one categorical variable. If we let μ_i be the mean expression for mice of strain i, $i = 1, 2, 3$ then we wish to test $H_0 : \mu_1 = \mu_2 = \mu_3$ against the alternative that the three population means are not all equal. There are a number of ways to do this in SAS but I will use `PROC GLM` since it generalizes better for the rest of the question.

```
PROC GLM Data=GeneExp;
   Class Strain;
   Model Expr=Strain;
run;
```

The GLM Procedure

Dependent Variable: Expr

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	71.9749290	35.9874645	9.47	0.0022
Error	15	57.0280350	3.8018690		
Corrected Total	17	129.0029640			

R-Square	Coeff Var	Root MSE	Expr Mean
0.557932	26.88937	1.949838	7.251333

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Strain	2	71.97492900	35.98746450	9.47	0.0022

Hence the F-statistic to test this null hypothesis is $F = 9.47$ which has an $F(2, 15)$ distribution if H_0 is true and the p-value for the test is

$$p \;=\; \mathrm{P}(F(2, 15) > 9, 47) \;=\; 0.0022.$$

Since this p-value is less than 0.01, we can reject H_0 at the 1% level and conclude that there is a significant difference in expression of this gene across the three strains. **[4 marks]**

b) We now wish to include gender in the model within each strain. This means including the interaction between gender and strain in the model. To test the hypothesis of a gender effect within strain we would then compare this fitted model (full model) to the reduced model given in part (a).

```
PROC GLM Data=GeneExp;
  CLASS Strain Sex;
  Model Expr=Strain Sex Strain*Sex /E3;
run;
```

The GLM Procedure

Dependent Variable: Expr

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	90.3041227	18.0608245	5.60	0.0068
Error	12	38.6988413	3.2249034		
Corrected Total	17	129.0029640			

R-Square	Coeff Var	Root MSE	Expr Mean
0.700016	24.76512	1.795802	7.251333

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Strain	2	71.97492900	35.98746450	11.16	0.0018
Sex	1	16.61569089	16.61569089	5.15	0.0424
Strain*Sex	2	1.71350278	0.85675139	0.27	0.7711

We now need to calculate the F-statistic to compare the reduced model using strain alone with this full model which includes sex.

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(57.0280 - 38.6988)/(15 - 12)}{3.2249} \\ &= 1.8945 \end{aligned}$$

The null hypothesis here states that there is no gender effect in any of the strains, that is

$$H_0 : \mu_{1M} = \mu_{1F}, \mu_{2M} = \mu_{2F}, \mu_{3M} = \mu_{3F}$$

against the alternative that at least one of these three equalities does not hold. Under H_0 $F \sim F(3, 12)$. Looking in the table in our book we see that the p-value will be greater than 0.05 since the 5% critical values for the $F(2, 12)$ and $F(4, 12)$ distributions are 3.89 and 3.26 respectively so the critical value for the required distribution will be between these two and certainly greater than the observed F. Hence there is no evidence of a sex effect within strains.

[6 marks]

c) In this question we wish to test if we can model the sex effect using the same effect for each of the three strains. Once again this is a reduced model relative to the full model used in part (b). In this case the null hypothesis is

$$H_0 : \mu_{1M} - \mu_{1F} = \mu_{2M} - \mu_{2F} = \mu_{3M} - \mu_{3F}$$

which corresponds to a model with Strain and Sex but without any interaction term.

```
PROC GLM Data=S3A3.GeneExp plots=none;
   Class Strain Sex;
   Model Expr=Strain Sex /E3;
run;
```

The GLM Procedure

Dependent Variable: Expr

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	88.5906199	29.5302066	10.23	0.0008
Error	14	40.4123441	2.8865960		
Corrected Total	17	129.0029640			

R-Square	Coeff Var	Root MSE	Expr Mean
0.686733	23.43015	1.698999	7.251333

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Strain	2	71.97492900	35.98746450	12.47	0.0008
Sex	1	16.61569089	16.61569089	5.76	0.0309

The F-statistic for comparing this to the full model in part (b) is

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(40.4123 - 38.6988)/(14 - 12)}{3.2249} \\ &= 0.26567 \end{aligned}$$

Without any tables, it is very clear that such an F value will be insignificant and so we can conclude that the sex effect is the same for all three strains. **[6 marks]**

d) The best model here appears to be the model with an additive effect of both Strain and Sex (as fitted in part (c)). The evidence that this model is best is that looking at the individual p-values in that model, we cannot conclude that neither the strain effect adjusted for sex ($F = 12.47$, $df = (2, 14)$, $p = 0.0008$) nor the sex effect adjusted for strain ($F = 5.76$, $df = (1, 14)$, $p = 0.0309$) can be removed from the model at the 5% level of significance. Also, including sex increases the R^2 from 55.8% to 68.7% which is quite a large jump. **[3 marks]**

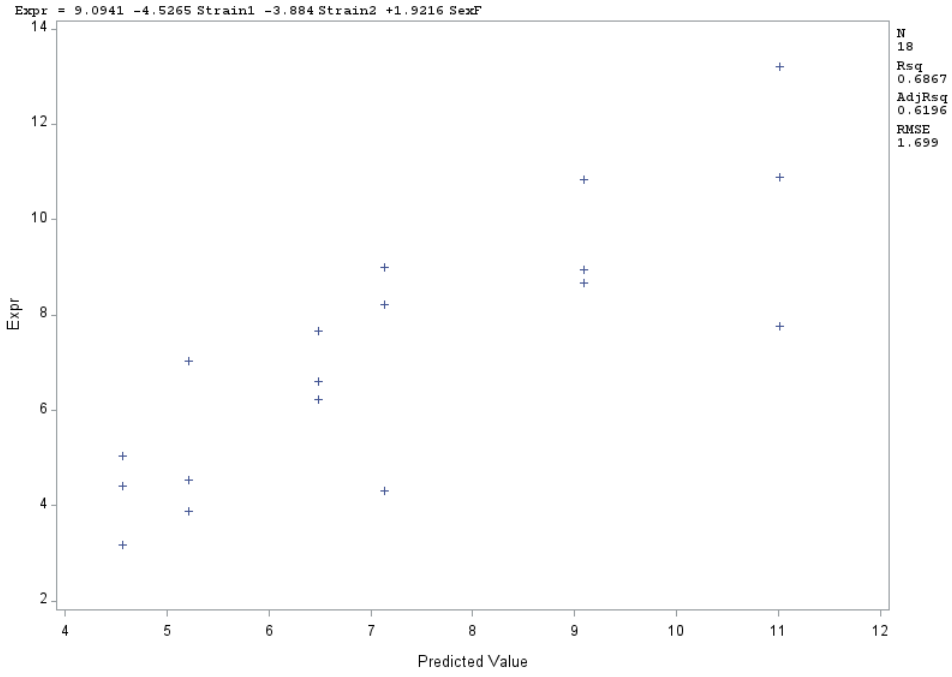
e) Before writing the report, I will refit the chosen model and get the estimates of the effects and the standard diagnostic plots. It is easier to produce the diagnostic plots using `PROC REG` so I will code the required dummy variables and use that method.

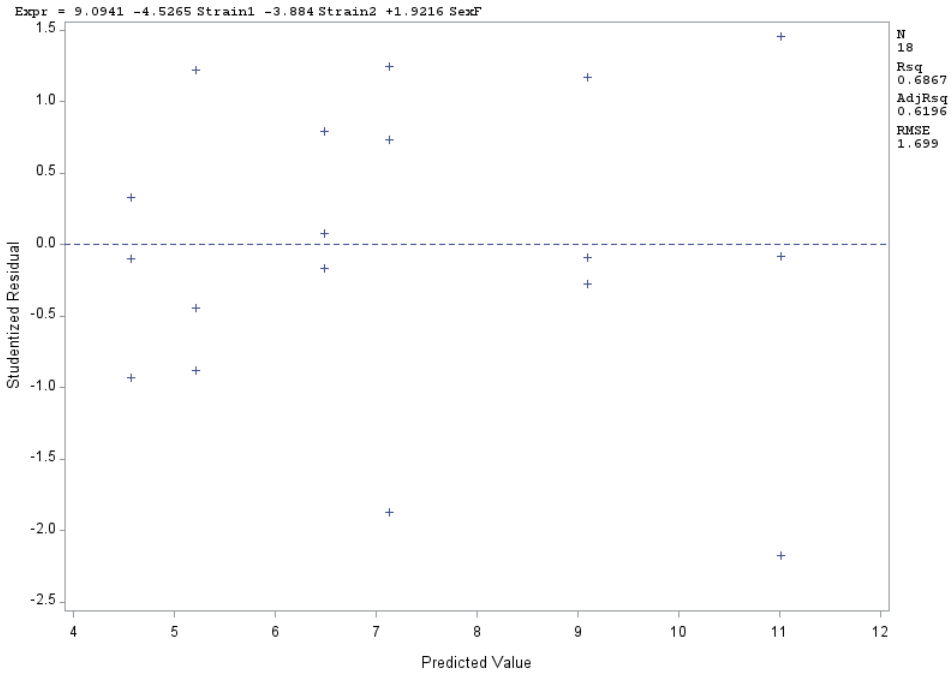
```
Data GeneExp1;
    Set S3A3.GeneExp;
    If Strain=1 then Strain1=1; else Strain1=0;
    If Strain=2 then Strain2=1; else Strain2=0;
    If Sex='F' then SexF=1; else SexF=0;
run;

PROC REG Data=GeneExp1;
    Model Expr=Strain1 Strain2 SexF;
    Plot Expr*Predicted.;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;
```

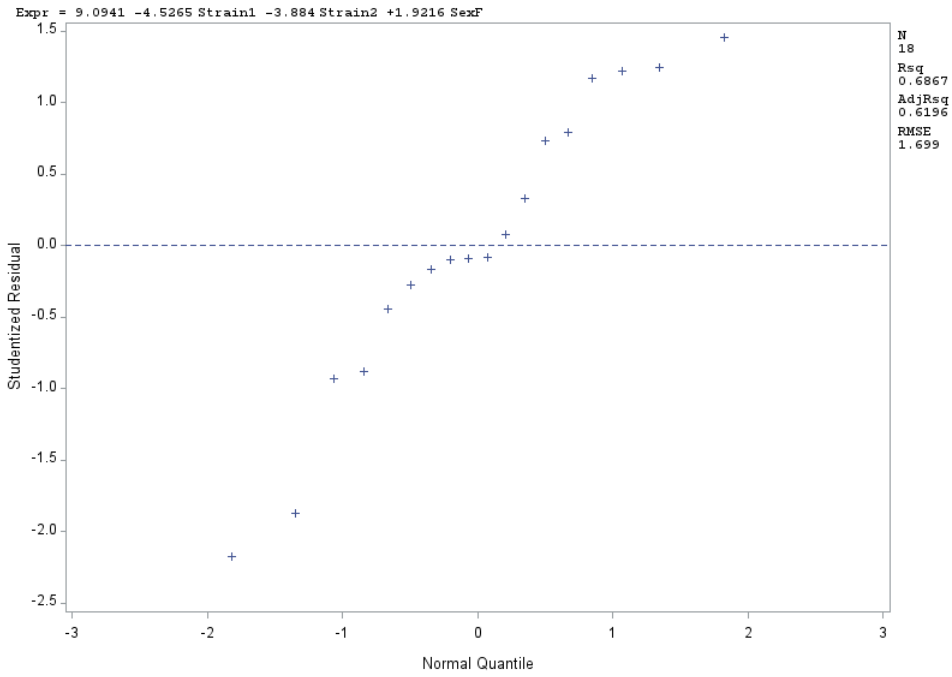
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	88.59062	29.53021	10.23	0.0008
Error	14	40.41234	2.88660		
Corrected Total	17	129.00296			

Root MSE	1.69900	**R-Square**	0.6867
Dependent Mean	7.25133	**Adj R-Sq**	0.6196
Coeff Var	23.43015		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	9.09406	0.80092	11.35	<.0001
Strain1	1	-4.52650	0.98092	-4.61	0.0004
Strain2	1	-3.88400	0.98092	-3.96	0.0014
SexF	1	1.92156	0.80092	2.40	0.0309


Expr = 9.0941 -4.5265 Strain1 -3.884 Strain2 +1.9216 SexF
N
18
Rsq
0.6867
AdjRsq
0.6196
RMSE
1.699
Expr
Predicted Value


Expr = 9.0941 -4.5265 Strain1 -3.884 Strain2 +1.9216 SexF
N
18
Rsq
0.6867
AdjRsq
0.6196
RMSE
1.699
Studentized Residual
Predicted Value



[3 marks]

Report: On examining various models for the expression of this gene we conclude that there is a very strong effect of strain which explains 55.8% of the variability observed in the gene expressions. However, even when adjusting for this difference between strains, there is still an effect due to gender which is statistically significant at the 5% level. The effect of gender is the same for all three strains. Including a gender effect increases the proportion of variance in gene expression explained by the model to 68.7%.

Strain 3 has the highest level of expression ($\hat{\mu}_{M3} = 9.09$, $\hat{\mu}_{F3} = 11.02$). The effect of Strain 2 is to lower these average expressions by 3.88 units and the effect of Strain 1 is to lower them by 4.53 units. Females have higher expression of this gene by 1.92 units for each of the strains.

Diagnostic plots suggest that there might be some violation of the assumption of homoscedasticity since the residual variability seems to increase somewhat for larger fitted values. It is possible that a transformation of the expression could correct this. There does not seem to be any violation of the assumption of linearity based on a plot of the observed against fitted values. A normal quantile-quantile plot of the studentized residuals does not show any major violations from normality. **[3 marks]**

Q. 3 Textbook 5.7

a) The following code will create the required dummy variables in SAS

```
Data Fertilizer1;
  Set S3A3.Fertilizer;
  If Fertilizer=1 then F1=1; else F1=0;
  If Fertilizer=2 then F2=1; else F2=0;
  If Fertilizer=3 then F3=1; else F3=0;
run;
```

[3 marks]

b) Here is the fitted model using `PROC REG`.

```
PROC REG Data=Fertilizer1 plots=none;
    Model Yield=F1 F2 F3;
run;
```

[5 marks]

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	362.60000	120.86667	5.14	0.0046
Error	36	845.80000	23.49444		
Corrected Total	39	1208.40000			

Root MSE	4.84711	**R-Square**	0.3001
Dependent Mean	32.80000	**Adj R-Sq**	0.2417
Coeff Var	14.77776		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	29.80000	1.53279	19.44	<.0001
F1	1	6.80000	2.16769	3.14	0.0034
F2	1	0.10000	2.16769	0.05	0.9635
F3	1	5.10000	2.16769	2.35	0.0242

c) The null hypothesis to be tested here is $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$.
We can test this using the F test from the Analysis of Variance Table. The value of the test statistic is $F = 5.14$ which has an $F(3, 36)$ distribution if H_0 is true. From the ANOVA table we see that the p-value is

$$p \;=\; \mathrm{P}(F(3,36) > 5.14) \;=\; 0.0046.$$

Hence at the 5% level of significance we would reject the null hypothesis and conclude that at least one of the three fertilizers has an effect on corn yield compared to the control group not given any fertilizer. **[3 marks]**

d) In this case we want to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$. Under H_0 the true model would be

$$y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_1 F_{i2} + \mu_1 F_{i3} + \varepsilon_{ij} = \mu_0 + \mu_1(F_{i1} + F_{i2} + F_{i3}) + \varepsilon_{ij}$$

This is a reduced model relative to the full model we fitted in part (b) so we need to fit this reduced model and then compare the error sums of squares for the two models. To fit this model we will construct the new indicator variable $F = F_1 + F_2 + F3$ and use that in the regression model.

```
Data Fertilizer1;
    Set Fertilizer1;
    F=F1+F2+F3;
run;

PROC REG Data=Fertilizer1 plots=none;
    Model Yield=F;
run;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	120.00000	120.00000	4.19	0.0476
Error	38	1088.40000	28.64211		
Corrected Total	39	1208.40000			

Root MSE	5.35183	**R-Square**	0.0993
Dependent Mean	32.80000	**Adj R-Sq**	0.0756
Coeff Var	16.31656		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	29.80000	1.69240	17.61	<.0001
F	1	4.00000	1.95421	2.05	0.0476

[6 marks]

We use the F test comparing the reduced model to the full model to test H_0.

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(1088.4 - 845.8)/(38 - 36)}{23.4944} \\ &= 5.16 \end{aligned}$$

If H_0 is true then $F \sim F(2, 36)$. From Table A.4 we see that the 5% critical value for the $F(2, 36)$ distribution will be between $F(0.05; 2, 30) = 3.32$ and $F(0.05; 2, 40) = 3.23$. Since our observed F statistic is greater than either of these we can conclude that $p = \text{P}(F(2, 36) > 5.16) < 0.05$ and so, at the 5% level of significance we can reject H_0 and conclude that the three fertilizers do not have the same effect. **[4 marks]**

e) Looking at the parameter estimates from the model fitted in part (b) we see that the $\hat{\mu}_1 - \hat{\mu}_{\text{control}} = 6.8$ which is the largest of the three coefficients of the F_j. Hence we would conclude that Fertilizer 1 has the largest effect and that the effect is to increase the corn yield by 6.8 units on average. **[4 marks]**

Q. 4 a)

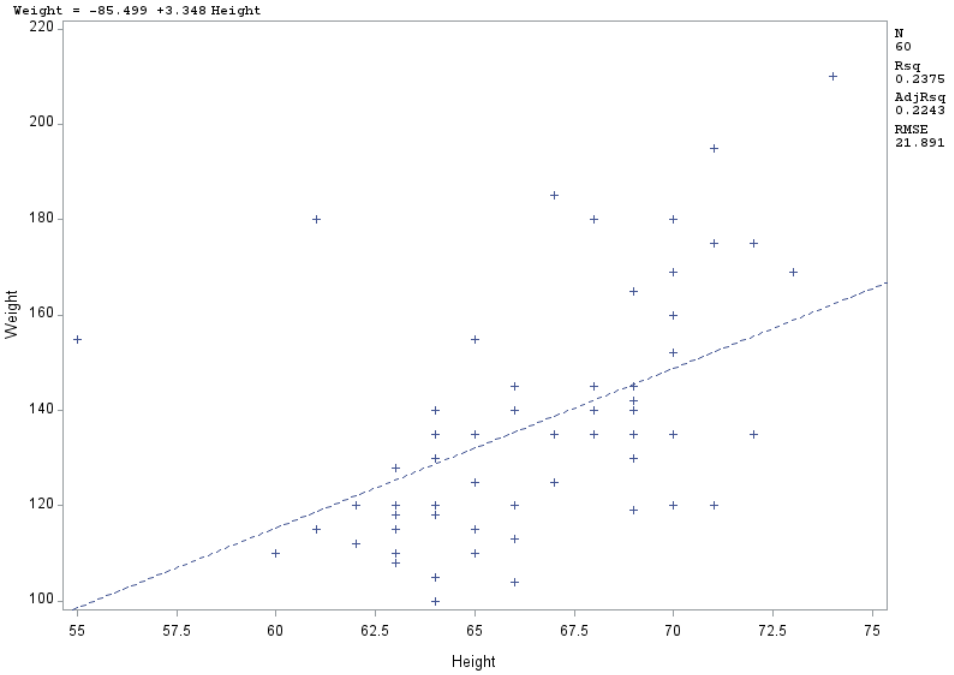
```
PROC REG Data=Students;
    Model Weight=Height;
    Output out=Studentout
        Predicted=Fitted
        Student=Res_stud;
    Plot Weight*Height;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;
```

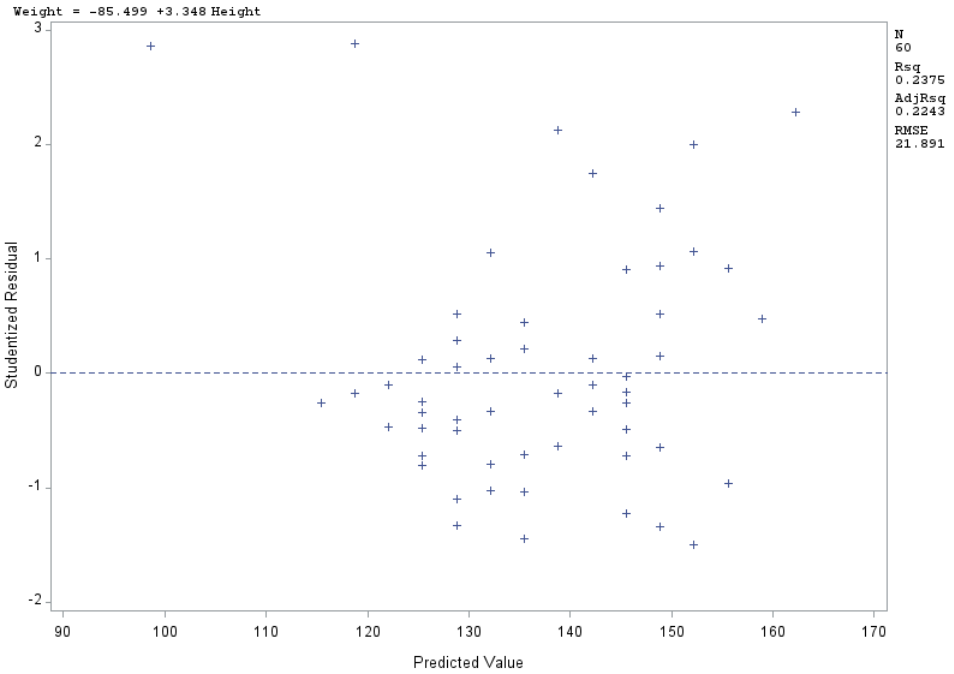
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	8655.46231	8655.46231	18.06	<.0001
Error	58	27795	479.23226		
Corrected Total	59	36451			

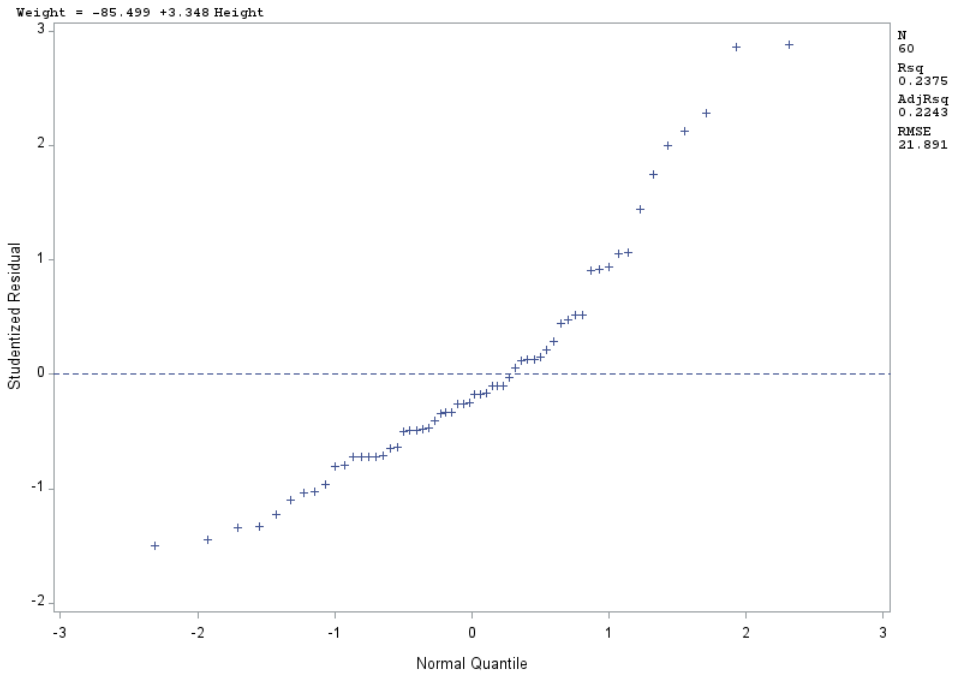
Root MSE	21.89137	**R-Square**	0.2375
Dependent Mean	137.53333	**Adj R-Sq**	0.2243
Coeff Var	15.91714		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-85.49900	52.55627	-1.63	0.1092
Height	1	3.34800	0.78779	4.25	<.0001

[2 marks]


Weight = -85.499 +3.348 Height
N
60
Rsq
0.2375
AdjRsq
0.2243
RMSE
21.891
Weight
Height
220
200
180
160
140
120
100
55
57.5
60
62.5
65
67.5
70
72.5
75


Weight = -85.499 +3.348 Height
N
60
Rsq
0.2375
AdjRsq
0.2243
RMSE
21.891
Studentized Residual
Predicted Value
3
2
1
0
-1
-2
90
100
110
120
130
140
150
160
170

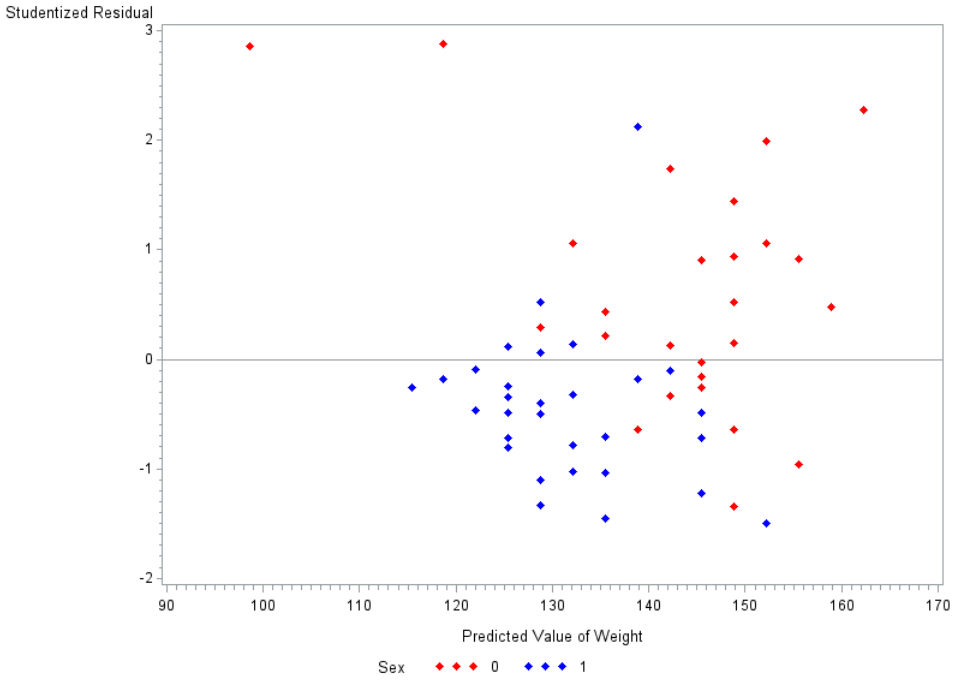


The three diagnostic plots show that there may be some issues with the assumptions of the linear model. There appear to be a couple of quite large outliers with small predicted values and also there is some evidence of heteroscedasticity. The normality assumptions does not seem to be badly violated from the normal quantile-quantile plot. **[3 marks]**

The fitted model says that weight increases by about 3.35 pounds for every increase of 1 inch in height. The effect of height is significant at a level of significance of 0.01%. The value of R^2 tells us that height explains about 23.75% of the variability in weight. **[2 marks]**

b) The following code will reproduce the plot of studentized residuals against the fitted values but with different colours for the two sexes (denoted by adding `=Sex` after the usual specification of the plot). For convenience of interpretation I have also added a reference line at $y = 0$. This is accomplished using the option `VREF=0` after the forward slash in the `PLOT` statement.

```
Symbol1 color=red Value=Squarefilled;
Symbol2 color=blue Value=Diamondfilled;
PROC GPLOT Data=Studentout;
    PLOT Res_Stud*Fitted=Sex / VREF=0;
run;
```



From this plot, we do see that there is an issue related to sex. Most of the residuals for females are negative whereas most of those for males are positive. There also seems to be a difference in variability of the residuals for the two genders. **[3 marks]**

c) To fit an interaction using `PROC REG` we need to compute the product of the indicator variable for `Sex` with `Height` which we do in a `Data` step as follows.

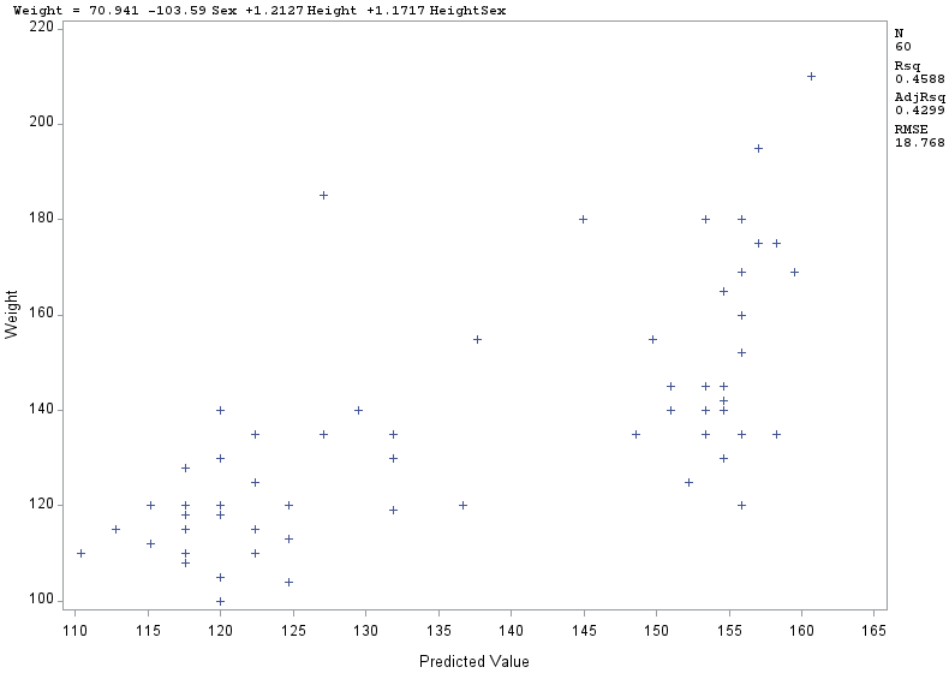
```
Data Students1;
    Set S3A3.Students;
    HeightSex=Height*Sex;
run;

PROC REG Data=Students1;
    Model Weight=Sex Height HeightSex;
    Output out=Studentout1
        Predicted=Fitted
  Student=Res_stud;
    Plot Weight*Predicted.;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;

PROC GPLOT Data=Studentout1;
    PLOT Res_Stud*Fitted=Sex / VREF=0;
run;
```

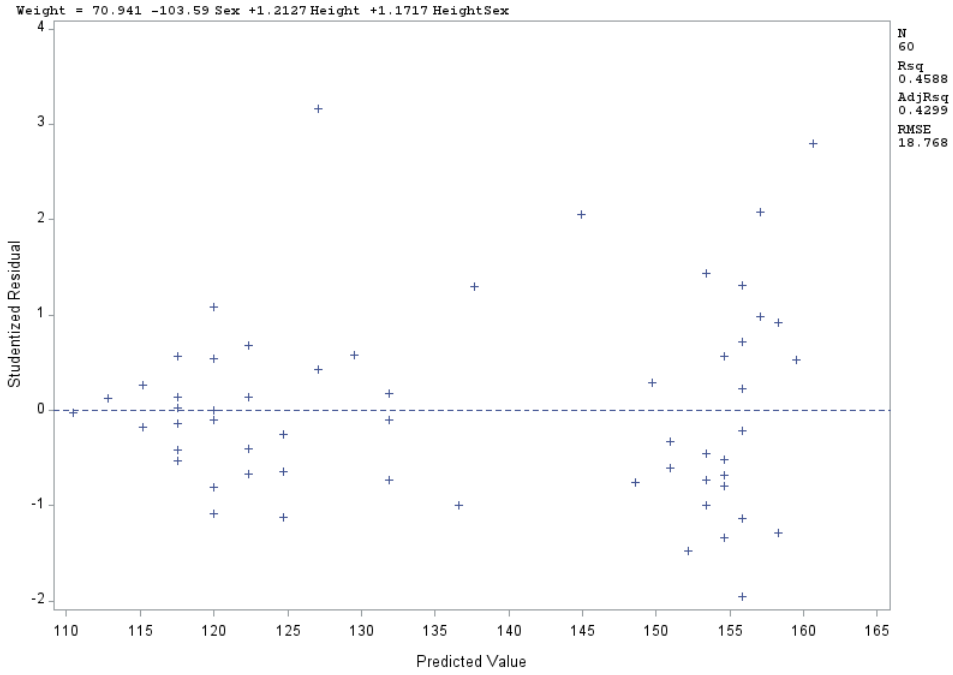
[3 marks]

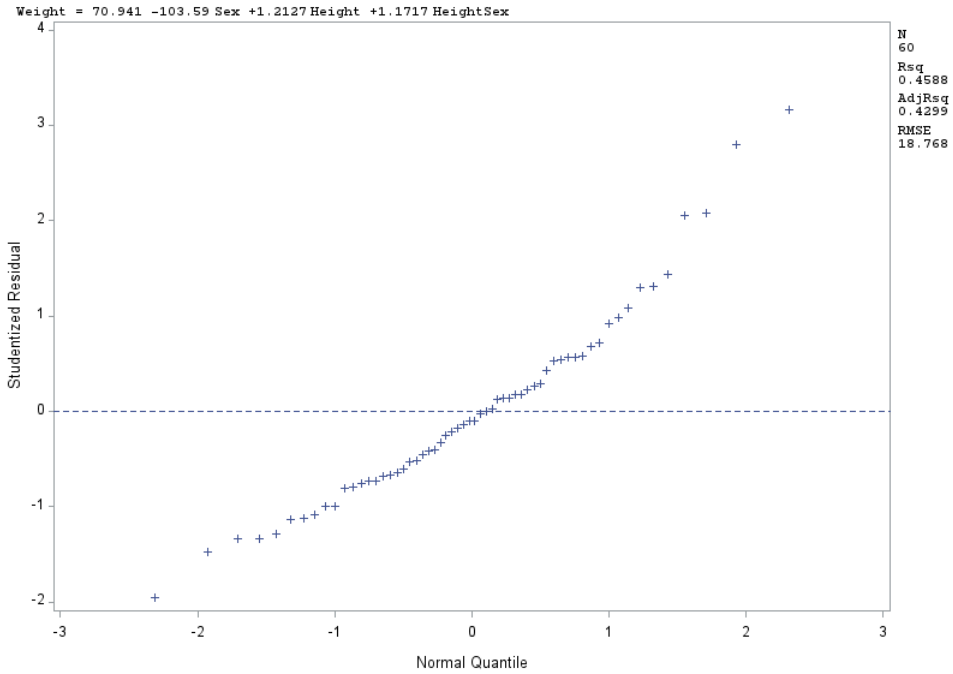
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	16725	5575.12570	15.83	<.0001
Error	56	19726	352.24208		
Corrected Total	59	36451			

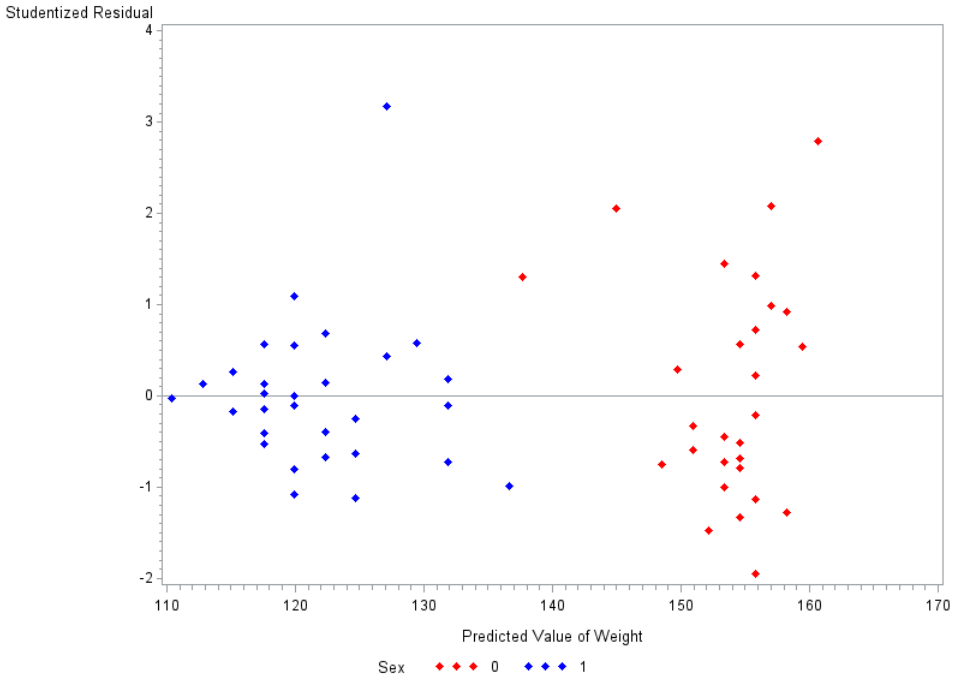
Root MSE	18.76811	**R-Square**	0.4588
Dependent Mean	137.53333	**Adj R-Sq**	0.4299
Coeff Var	13.64623		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	70.94094	65.17613	1.09	0.2811
Sex	1	-103.59052	106.64899	-0.97	0.3356
Height	1	1.21266	0.95179	1.27	0.2079
HeightSex	1	1.17171	1.60983	0.73	0.4697




Weight = 70.941 -103.59 Sex +1.2127 Height +1.1717 HeightSex
N
60
Rsq
0.4588
AdjRsq
0.4299
RMSE
18.768
Studentized Residual
Predicted Value
4
3
2
1
0
-1
-2
110
115
120
125
130
135
140
145
150
155
160
165


Weight = 70.941 -103.59 Sex +1.2127 Height +1.1717 HeightSex
N
60
Rsq
0.4588
AdjRsq
0.4299
RMSE
18.768
Studentized Residual
Normal Quantile
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3



There appear to be fewer issues with the usual assumptions in this model although there is still some evidence of heteroscedasticity. The linearity assumption seems justified except for one outlier. The final plot shows that the issues of different signs of residuals for the two sexes has now been corrected by the full model although there does still seem evidence of differences in variability between the genders.

[4 marks]

This model fits the following regression lines

$$\text{Weight} = 70.941 + 1.213\text{Height} \quad \text{for Males.}$$

$$\text{Weight} = -32.650 + 2.384\text{Height} \quad \text{for Females}$$

We note that both of the estimates for the effect of Height on Weight are smaller than for the pooled analysis. The analysis by gender now explains 45.88% of the variability in Weight which is much higher than in the pooled model. **[2 marks]**

d) To compare the two models fitted in parts (a) and (c) we need to use the F test with the model in (c) as the full model and the pooled model in (a) as a reduced model.

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(27795 - 19726)/(58 - 56)}{352.242} \\ &= 11.454 \end{aligned}$$

The null hypothesis that we are testing here is $H_0 : \beta_{0M} = \beta_{0F}, \beta_{1M} = \beta_{1F}$ where β_{0M} and β_{0F} are the intercepts for Males and Females respectively and β_{1M} and β_{1F} are the two slopes. The alternative hypothesis is that different intercepts and/or different slopes are require. If

the null hypothesis is true then $F \sim F(2, 56)$. From Table A.5 we see that $F(0.01; 2, 40) = 5.18$ Since our observed F statistic is much higher than this we can conclude that the p-value is less than 0.01 and so we would reject the null hypothesis. In conclusion there is strong evidence that the regression lines for predicting weight from height are different for males and females. **[3 marks]**

e) We can test the hypothesis of parallel lines by examining the output for the model in part (c). Let β_S, β_H and β_{HS} denote the coefficients of the Sex dummy variable, the height variable and the interaction. The two lines are

$$\begin{array}{lcll} \text{Weight} & = & \beta_0 + \beta_H \times \text{Height} & \text{for males} \\ \text{Weight} & = & \beta_0 + \beta_S + (\beta_H + \beta_{HS}) \times \text{Height} & \text{for females} \end{array}$$

These two lines will be parallel if, and only if, $\beta_{HS} = 0$ so a test of parallel lines is a test of

$$H_0 : \beta_{HS} = 0 \quad \text{V} \quad H_1 : \beta_{HS} \neq 0$$

We can conduct this test using the regular t test. The test statistic in this case is $t = 0.73$. Under the null hypothesis this is an observation from a t_{56} distribution and the corresponding p-value is $p = 0.4697$ from which we can conclude that there is no significant difference between the slopes of the lines for males and females so we can use parallel lines. **[3 marks]**