MATH 4044 – Statistics for Data Science

Practical Week 7 Solutions

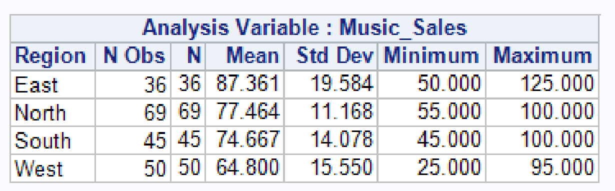
Question 1

The data for this practical is stored in a SAS data file called `store.sas7bdat` located in `mydata` library on the SAS OnDemand server. Variables in that file are as follows:

Variable name	Description
Region	Region of the country (North, East, South, West)
Advertising	Advertising (Yes or No)
Gender	Gender of shopper (M or F)
Book_Sales	Amount spent on books
Music_Sales	Amount spent on music
Electronics_Sales	Amount spent ion electronics
Total_Sales	Total sales

(a) Check the necessary assumptions and perform an ANOVA test to determine whether there is statistically significant difference in average music sales by region. Interpret the results.

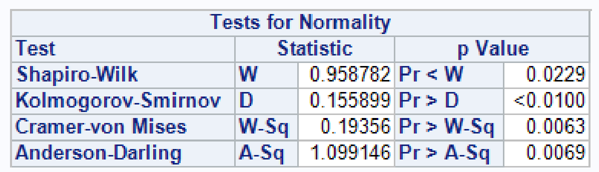
We begin by examining descriptive statistics in Table 1. Sample music sales for the East region are much higher than for the other regions and this difference may prove to be statistically significant. Standard deviations are also quite different, suggesting that the assumption of equal variance may be violated.



Analysis Variable : Music_Sales						
Region	N Obs	N	Mean	Std Dev	Minimum	Maximum
East	36	36	87.361	19.584	50.000	125.000
North	69	69	77.464	11.168	55.000	100.000
South	45	45	74.667	14.078	45.000	100.000
West	50	50	64.800	15.550	25.000	95.000

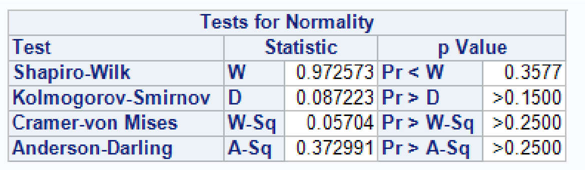
Table 1. Descriptive Statistics for music sales by region.

Boxplots in Figure 3 also suggest that variances may not be equal. There is also some evidence of lack of symmetry in the distribution of music sales for East and North regions. Histograms of music sales by region shown in Figure 2 suggest distributions that are approximately symmetric for North, South and West, but not for the East region. The latter distribution appears to be negatively skewed.



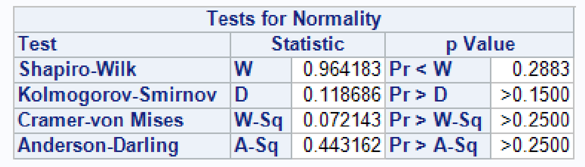
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.958782	Pr < W	0.0229
Kolmogorov-Smirnov	D	0.155899	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.19356	Pr > W-Sq	0.0063
Anderson-Darling	A-Sq	1.099146	Pr > A-Sq	0.0069

Table 2. Normality test results, North region.



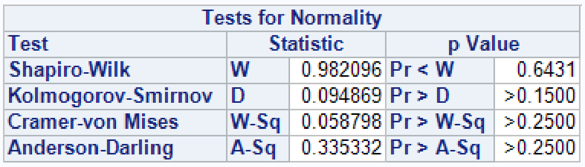
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.972573	Pr < W	0.3577
Kolmogorov-Smirnov	D	0.087223	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.05704	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.372991	Pr > A-Sq	>0.2500

Table 3. Normality test results, South region.



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.964183	Pr < W	0.2883
Kolmogorov-Smirnov	D	0.118686	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.072143	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.443162	Pr > A-Sq	>0.2500

Table 4. Normality test results, East region.



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.982096	Pr < W	0.6431
Kolmogorov-Smirnov	D	0.094869	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.058798	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.335332	Pr > A-Sq	>0.2500

Table 5. Normality test results, West region.

Examining P-values for tests of Normality shown in Table 2 we find that the distribution of music sales in the North region cannot be assumed to be Normal as all four tests of Normality indicate significant departures from Normality (all P-values are < 0.05). For the other three regions, the assumption of Normality cannot be rejected (P-values > 0.05).

However, as sample sizes are reasonably large (greater than 30) and none of the distributions is severely skewed, we can proceed with a one-way ANOVA test.

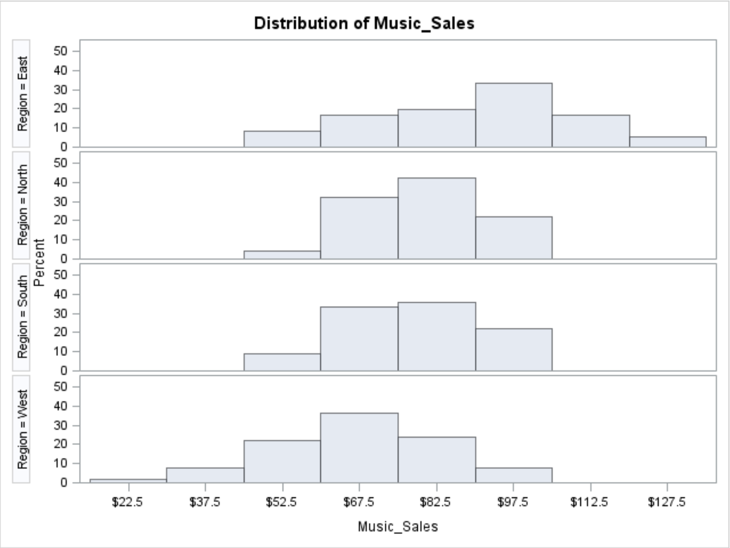


Figure 1. Histograms of music sales by region.

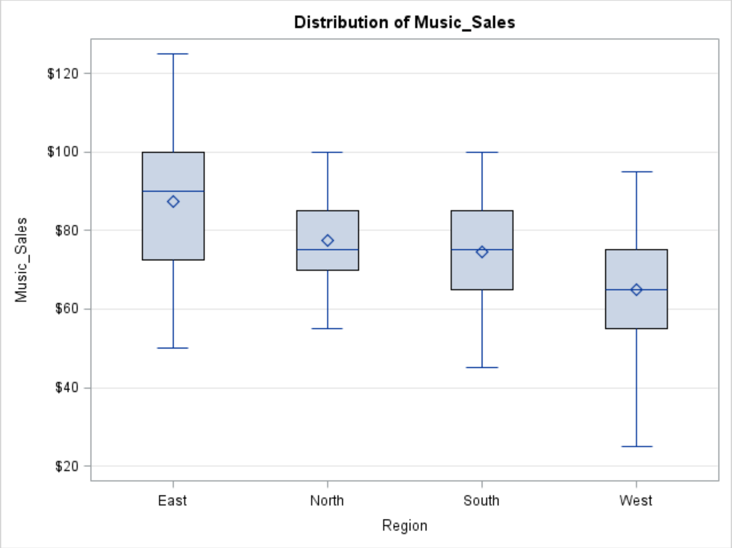


Figure 2. Boxlots of music sales by region.

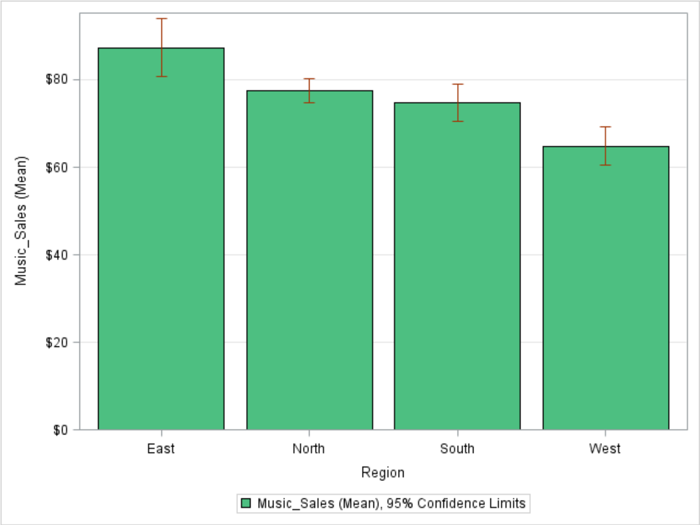
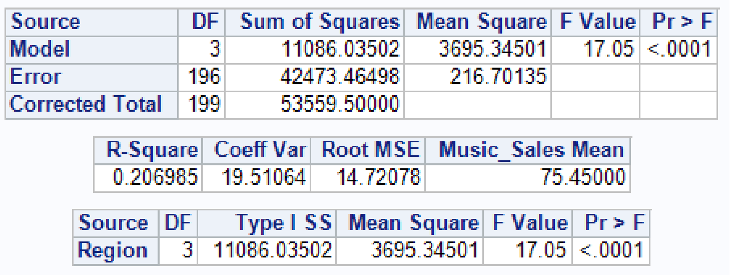


Figure 3. Bar diagram of mean music sales by region, with 95% confidence limits.

The bar diagram in Figure 3 suggests that there are differences in mean sales among some of the regions. In particular, mean sales in the West were much lower than in the East. Results of an ANOVA test for the differences in means are shown in Table 6.



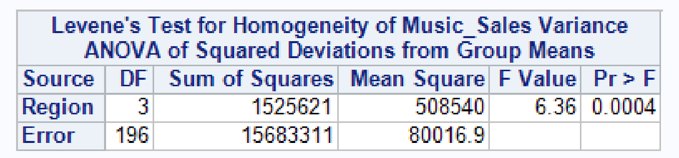
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11086.03502	3695.34501	17.05	<.0001
Error	196	42473.46498	216.70135		
Corrected Total	199	53559.50000			

R-Square	Coeff Var	Root MSE	Music_Sales Mean
0.206985	19.51064	14.72078	75.45000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Region	3	11086.03502	3695.34501	17.05	<.0001

Table 6. Main ANOVA table for music sales by region.

However, before interpreting Table 6, we first examine results of a test of homogeneity of variance in Table 7. There are significant differences in variance across the four regions, F(3,196) = 6.36, P-value = 0.0004 < 0.01.

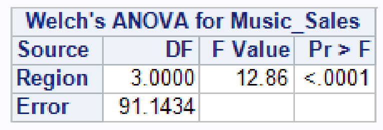


Levene's Test for Homogeneity of Music_Sales Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Region	3	1525621	508540	6.36	0.0004
Error	196	15683311	80016.9		

Table 7. Equality of variance test results for music sales by region.

Since the assumption of homogeneity of variance has been violated, we report Welch's corrected F-ratio shown in Table 8, instead of the one given in the main ANOVA table. Since Welch's F(3,91.14) = 12.86 with P-value < 0.0001, there are significant differences among mean music sales by region.

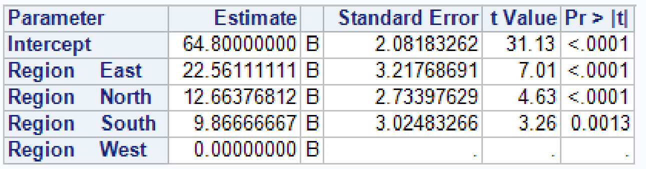


Welch's ANOVA for Music_Sales

Source	DF	F Value	Pr > F
Region	3.0000	12.86	<.0001
Error	91.1434		

Table 8. Welch's F-ratio corrected for departures
from homogeneity of variance.

Table 9 shows parameter estimates (differences in means) relative to the mean for the West region. All parameters are statistically significant (P-values < 0.01). The intercept of 64.80 represents the sample mean for the West region, and the other parameters show the differences between means for other regions and the West. All differences are positive and statistically significant, therefore mean music sales in other regions were significantly higher than in the West.



Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		64.80000000	B	2.08183262	31.13	<.0001
Region	East	22.56111111	B	3.21768691	7.01	<.0001
Region	North	12.66376812	B	2.73397629	4.63	<.0001
Region	South	9.86666667	B	3.02483266	3.26	0.0013
Region	West	0.00000000	B	.	.	.

Table 9. Solution including parameter estimates for the ANOVA test in Table 6.

Figure 4 shows fit diagnostics for the model, which are the same as those produced for regression. The residual versus predicted value plot shows unequal vertical differences, which is consistent with our earlier conclusion that the assumption of equal variances has been violated in this case. Studentised residuals plot shows a few points outside the -2 and 2 bounds, but most are close to the bounds and the number is not too large (less than 5%). There are no points of high leverage, but a number of observations have Cook's D values above the 4/n cut-off, however they are not too extreme. From the histogram and a Q-Q plot, residuals are reasonably close to Normal. Apart from unequal variances, there are no serious assumption violations.

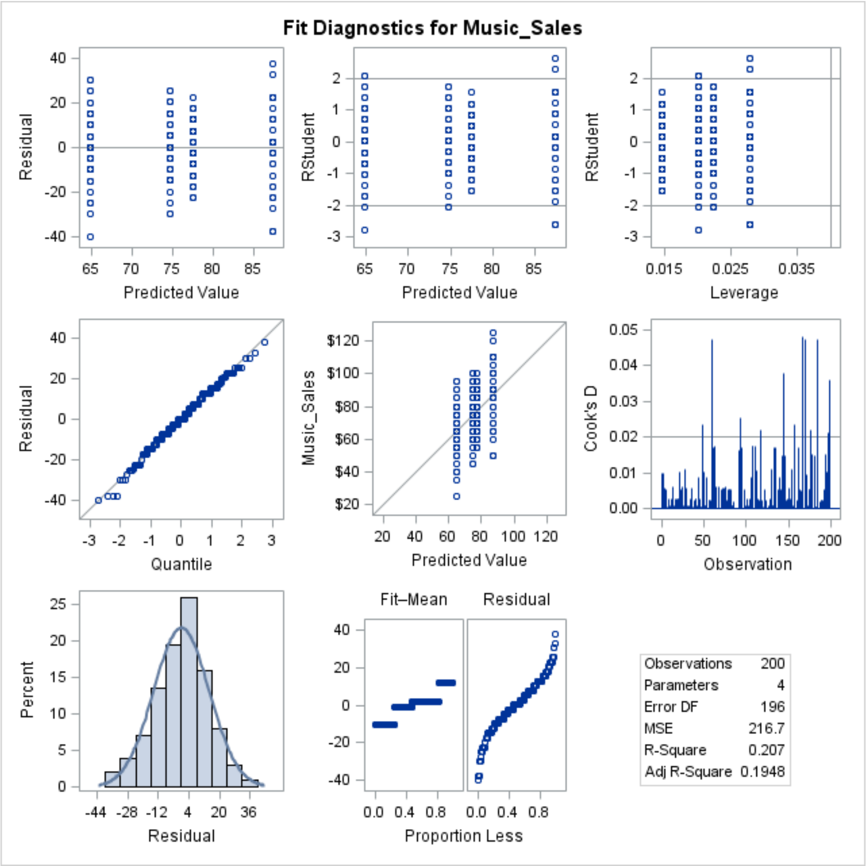
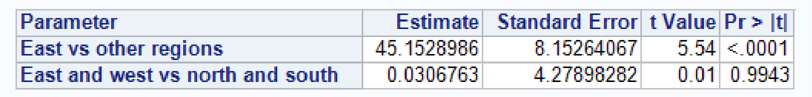


Figure 4. Diagnostic plots for one-way ANOVA test in Table 6.

(b) Suppose we want to test the hypothesis that music sales in the East region are different than in the rest of the country. Obtain relevant SAS output and interpret your results.



Parameter	Estimate	Standard Error	t Value	Pr > |t|
East vs other regions	45.1528986	8.15264067	5.54	<.0001
East and west vs north and south	0.0306763	4.27898282	0.01	0.9943

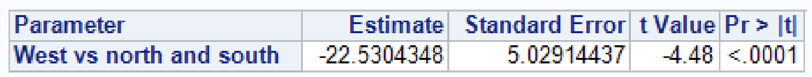
Table 10. Planned contrast estimates for the ANOVA test in Table 6.

Weights for requested comparisons were as follows:

- For East versus other regions, 3, -1, -1, -1. This contrast compares the mean for East to the average of the means for the other regions.
- For East and West versus North and South, 1, -1, -1, 1. This contrast compares the average of the means for East and West to the average of the means for North and South.

These planned comparisons reveal that music sales in the East region differ significantly from the sales in the rest of the country, t(196) = 5.54, P-value < 0.0001. However, the difference between mean sales in East and West relative to mean sales in North and South is not significantly different from zero, t(196) = 4.28, P-value = 0.9943 > 0.05.

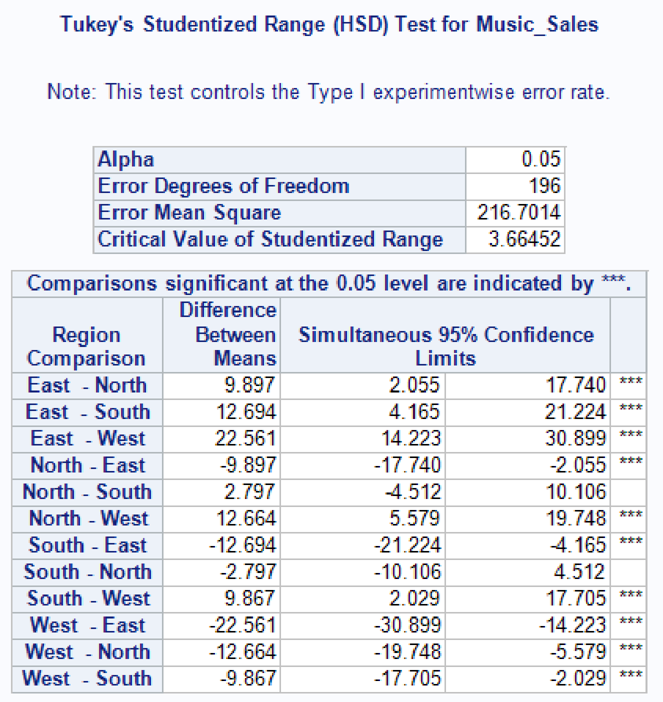
Note: Following the rules for defining contrasts, we should have actually excluded East from the second comparison as it was singled out in the previous one. A better follow-up comparison would have been West versus North and South. The weights for this comparison would be 0, -1, -1, 2 and the result would be as follows:



Parameter	Estimate	Standard Error	t Value	Pr > |t|
West vs north and south	-22.5304348	5.02914437	-4.48	<.0001

Table 10a. Planned contrast estimate for West versus North and South

(c) Perform appropriate post-hoc tests and interpret the results.



Tukey's Studentized Range (HSD) Test for Music_Sales

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	196
Error Mean Square	216.7014
Critical Value of Studentized Range	3.66452

Comparisons significant at the 0.05 level are indicated by ***.

Region Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
East - North	9.897	2.055	17.740	***
East - South	12.694	4.165	21.224	***
East - West	22.561	14.223	30.899	***
North - East	-9.897	-17.740	-2.055	***
North - South	2.797	-4.512	10.106	
North - West	12.664	5.579	19.748	***
South - East	-12.694	-21.224	-4.165	***
South - North	-2.797	-10.106	4.512	
South - West	9.867	2.029	17.705	***
West - East	-22.561	-30.899	-14.223	***
West - North	-12.664	-19.748	-5.579	***
West - South	-9.867	-17.705	-2.029	***

(d)

(e) **Table 11.** Tukey's post-hoc comparison results for the ANOVA test in Table 6

From the results of Tukey's post-hoc procedure shown in Table 11, the only means that are not significantly different are for North and South. This is confirmed by the diffogram shown in Figure 5. Therefore, we have established reasons behind the significant result of the 'omnibus' ANOVA test from part (a).

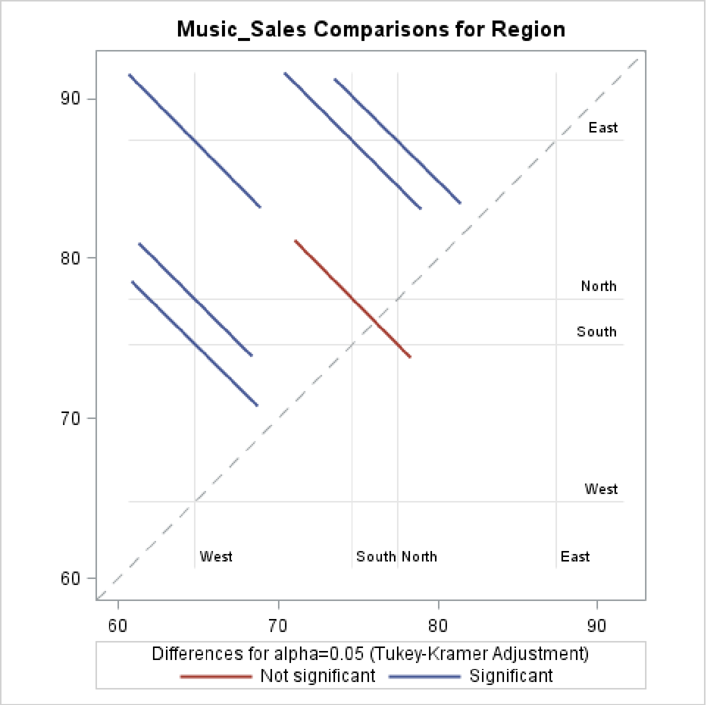
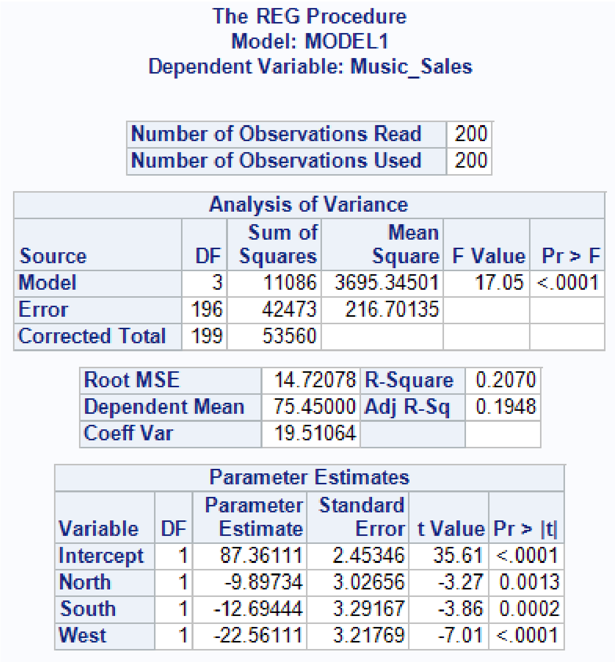


Figure 5. Diffogram for Tukey's post-hoc comparisons in Table 11.

(f) Define and estimate a multiple regression model for music sales with Region as the explanatory variable. Define appropriate dummy variables using *Region* = *East* as the baseline. Interpret the results.



The REG Procedure
Model: MODEL1
Dependent Variable: Music_Sales

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11086	3695.34501	17.05	<.0001
Error	196	42473	216.70135		
Corrected Total	199	53560			

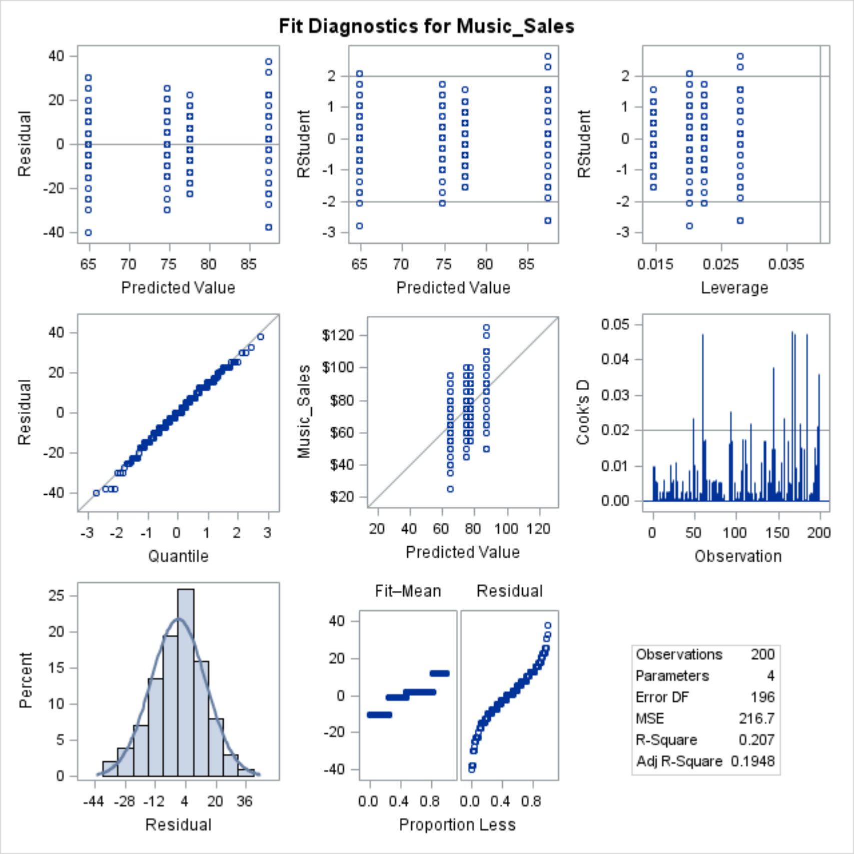
Root MSE	14.72078	R-Square	0.2070
Dependent Mean	75.45000	Adj R-Sq	0.1948
Coeff Var	19.51064		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	87.36111	2.45346	35.61	<.0001
North	1	-9.89734	3.02656	-3.27	0.0013
South	1	-12.69444	3.29167	-3.86	0.0002
West	1	-22.56111	3.21769	-7.01	<.0001

Table 12. Regression results for music sales vs region.

Overall, the model is statistically significant; F(3,196) = 17.05, P-value < 0.0001. All coefficients, including the intercept, are statistically significant (all P-values < 0.01). The intercept represents the mean music sales level in the East region; the slopes are the differences between the mean sales in the East and the other regions. Since all slopes are negative, the mean sales in the East were significantly higher than in the other regions.

The adjusted R-squared is quite low (19.48%) which is to be expected since there are likely to be other factors, in addition to region, that affect music sales. Fit diagnostics plots in Figure 12 are identical to those from Figure 4.



Observations	200
Parameters	4
Error DF	196
MSE	216.7
R-Square	0.207
Adj R-Square	0.1948

Figure 6. Diagnostics plots for the regression model in Table 12.

Appendix – SAS code

```
ods graphics on;

proc means data=work.store maxdec=3;
var Music_Sales;
class Region;
run;

proc univariate data=work.store normal;
    var Music_Sales;
    class Region;
    histogram / nrows=4;
run;

%let stat=Mean;

/*--Get variable names or labels--*/
data _null_;
    array x(1) Music_Sales;
    set work.store;
    call symputx ("Label", vlabel(x(1)));
run;

/*--Put variabel name/label or custom label into macro variable--*/
data _null_;
    call symputx ("respLabel", "&Label");
run;

/*--Combine label and stat into statRespLabel--*/
 %let statRespLabel=&respLabel (&stat);

proc sgplot data=work.store noautolegend;
    /*--TITLE and FOOTNOTE--*/
    /*--Bar chart settings--*/
    vbar Region / response=Music_Sales fillattrs=(color=big)
limits=Both
        limitstat=CLM numstd=1 transparency=0.00 stat=Mean
dataskin=None
        name='Bar';

    /*--Category Axis--*/
    xaxis;

    /*--Response Axis--*/
    yaxis grid label="&statRespLabel";

    /*--Legend Settings--*/
    keylegend 'Bar' / location=Outside;
run;

/* Perform one-way ANOVA using proc GLM */

proc glm data=store plots=diagnostics;
    class Region;
    model Music_Sales=Region/ solution;
    estimate 'East vs other regions' Region 3 -1 -1 -1;
    estimate 'East and west vs north and south' Region 1 -1 -1 1;
    means Region / hovtest Welch Tukey;
    lsmeans Region / pdiff adjust=Tukey;
    run;
quit;

/* Create dummy variable for level of region */

data work.store_dummies;
    set work.store;

    if Gender='Male' then
        Gender_dummy=1;
    else
        Gender_dummy=0;

    if Region='North' then
        North=1;
    else
        North=0;

    if Region='South' then
        South=1;
    else
        South=0;

    if Region='West' then
        West=1;
    else
        West=0;
run;

/* Fit a linear regresion model */

proc reg data=work.store_dummies plots=diagnostics;
    model Music_Sales=North South West;
    run;
quit;

ods graphics off;
```