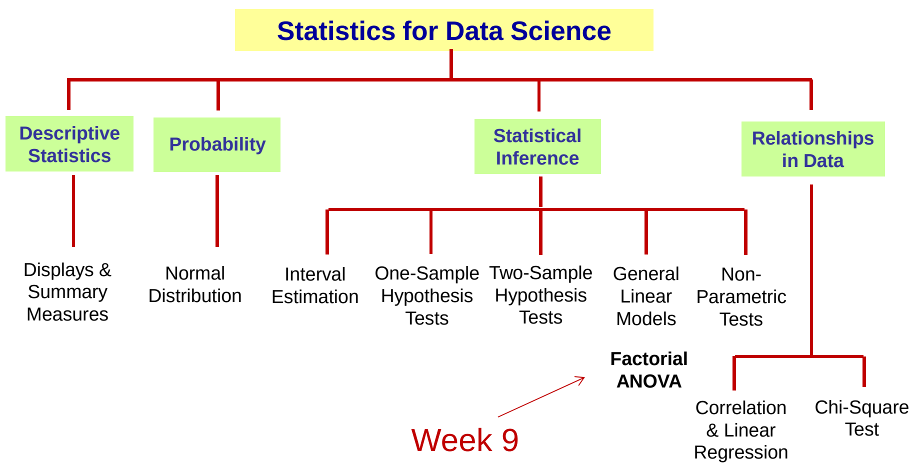
MATH 4044
Statistics for Data Science

Comparing Several Means
Factorial ANOVA

Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 12.

Topics to be covered

- Comparing several means:
  - Factorial ANOVA
  - Main and interaction effects
  - Post-hoc tests
  - Factorial ANOVA as a regression model

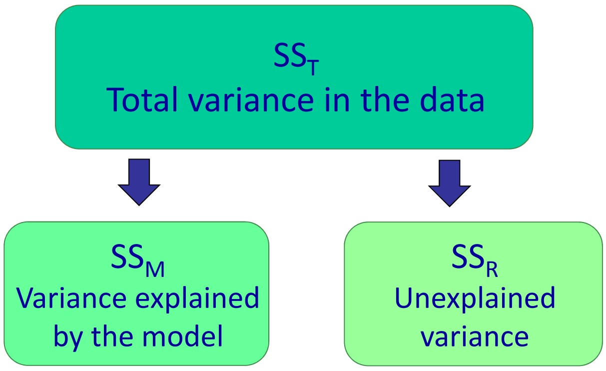
Factorial designs

- In one-way ANOVA, we have considered only the case of investigating whether and how one categorical variable affects a continuous response variable.
  - In many situations, there are at least two categorical variables that could be considered as explanatory variables.
- One of the most important questions is to consider whether and how explanatory variables interact in their effects:
  - Does the effect of one changes as the other changes?
  - It is poor practice just to consider the effects of possible explanatory variables one at a time.

Factorial designs

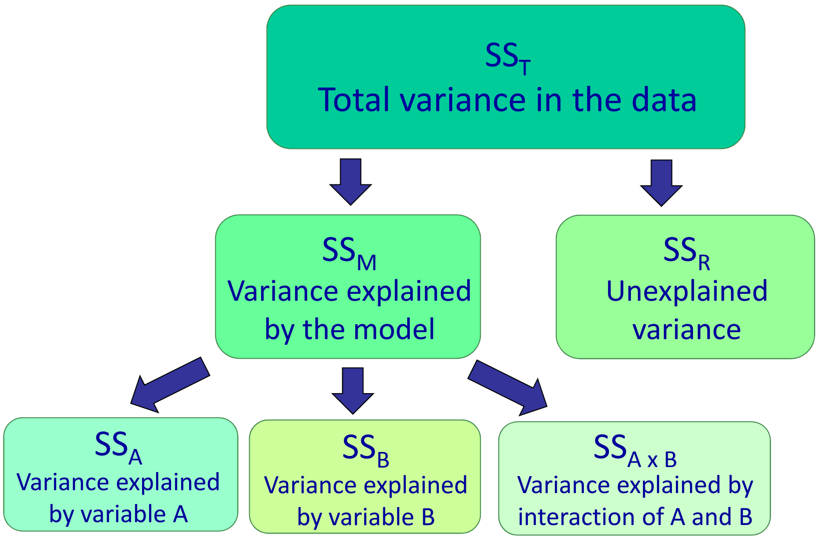
- Independent factorial designs:
  - There are several independent variables or predictors and each has been measured using different subjects.
  - Between groups design.
- Repeated measures (related) factorial design:
  - Several independent variables or factors have been measures, but the same subjects have been used in all conditions.
- Mixed design:
  - Several independent variables have been measured, some for the same subjects and some for different subjects.

Recall – Theory of ANOVA



- If the experiment is successful, then the model will explain more variance than it can't:
  - SS_M will be greater than SS_R.

Theory of two-way ANOVA



Main and interaction effects

- A two-way ANOVA is used to examine how two categorical explanatory variables affect the mean of a continuous variable.
- When there is an interaction between two explanatory variables, the effect on the response variable of one explanatory variable depends on the specific value or level of the other explanatory variable.
- The term main effect describes the mean effect of a single explanatory variable, averaged over other explanatory variables.
- It is usually the interactions between variables that are most interesting in a two-way (or a more general factorial) design.

Example: Electronics sales

- The data set `store` contains the following variables:

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 on electronics**
Total_Sales	Total sales

Example: Electronics sales

- Suppose we want to determine whether the mean of electronics sales varies by region and gender.
- We will check the assumptions and then conduct factorial ANOVA using PROC GLM.

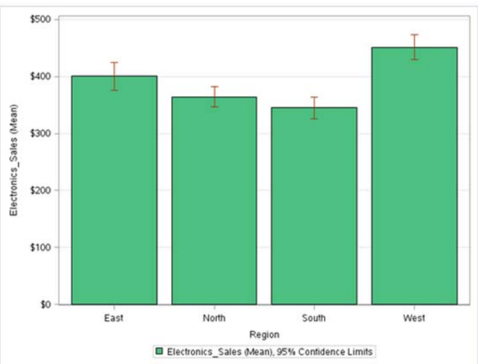
Example: Electronics sales

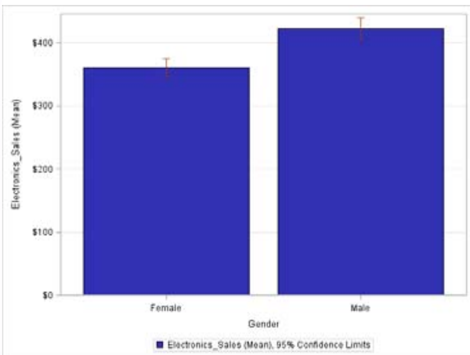
Descriptive Statistics

Analysis Variable : Electronics_Sales							
Region	Gender	N Obs	N	Mean	Std Dev	Minimum	Maximum
East	Female	22	22	364.545	63.526	270.000	480.000
	Male	14	14	457.143	45.814	400.000	570.000
North	Female	39	39	339.231	62.634	220.000	480.000
	Male	30	30	398.000	76.852	250.000	550.000
South	Female	23	23	321.739	53.653	250.000	450.000
	Male	22	22	369.545	66.725	250.000	510.000
West	Female	26	26	422.308	72.350	270.000	550.000
	Male	24	24	483.750	68.513	380.000	610.000

There appear to be some differences by gender across the four regions. Are these differences statistically significant?

Example: Electronics sales





Ignoring gender, sales in the West appear to be higher on average than in any other region.

Ignoring region, males appear to be spending more on electronics on average than females.

Example: Electronics sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	533841.480	76263.069	17.68	<.0001
Error	192	828038.020	4312.698		
Corrected Total	199	1361879.500			

R-Square	Coeff Var	Root MSE	Electronics_Sales Mean
0.391989	16.90159	65.67114	388.5500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Region	3	331827.2810	110609.0937	25.65	<.0001
Gender	1	196917.0078	196917.0078	45.66	<.0001
Region*Gender	3	10422.6951	3474.2317	0.81	0.4922

The overall model is highly significant, F(7,192) = 17.68, P-value < 0.0001. The R-squared is not large at 0.39 so there is considerable variability in electronics sales not accounted for by region and gender.

Example: Electronics sales

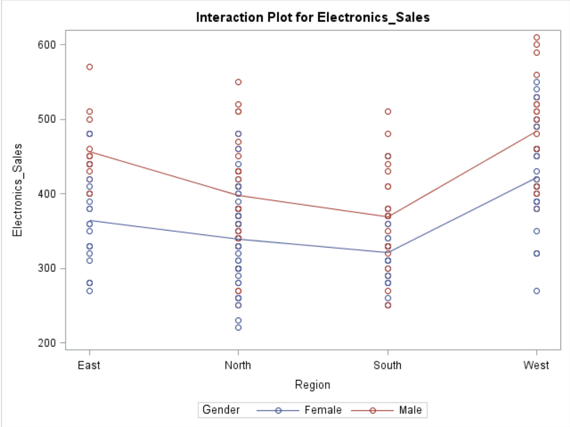
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	533841.480	76263.069	17.68	<.0001
Error	192	828038.020	4312.698		
Corrected Total	199	1361879.500			

R-Square	Coeff Var	Root MSE	Electronics_Sales Mean
0.391989	16.90159	65.67114	388.5500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Region	3	331827.2810	110609.0937	25.65	<.0001
Gender	1	196917.0078	196917.0078	45.66	<.0001
Region*Gender	3	10422.6951	3474.2317	0.81	0.4922

There is a highly significant main effect due to region, F(3,192) = 25.65, P-value < 0.0001, and gender, F(1,192) = 45.66, P-value < 0.0001. There is no evidence of interaction, P(3,192) = 0.81, P-value = 0.4922.

Example: Interaction plot



The interaction plot illustrates interactions between factors.

It plots the different means for each group formed by the combinations of genders and regions.

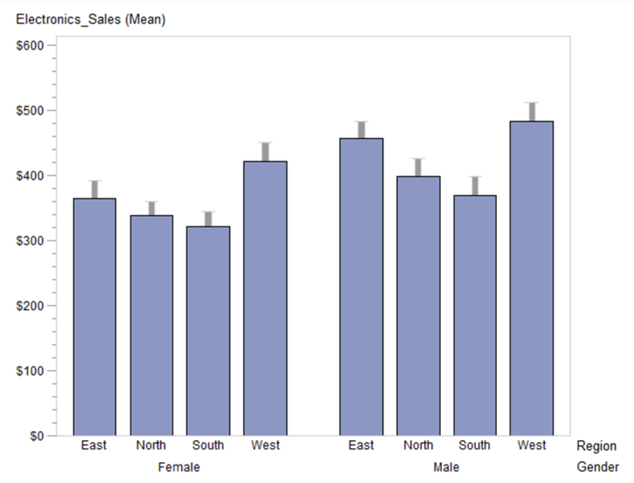
Means for males and for females are connected across regions.

The interaction plot confirms that while there are significant main effects for gender and region, there is no significant interaction. Means for females are lower than for males in all regions, by a similar amount.

Interaction plots – a few observations

- Significant interactions correspond to non-parallel lines on an interactions graph:
  - This does not mean that non-parallel lines automatically mean the interaction is significant.
  - Significance depends on the degree to which the lines are not parallel.
- If the lines on an interaction graph cross, then they are obviously not parallel which means there may be a significant interaction:
  - It is however not always the case that if the lines cross then the interaction is significant.

Example: Interactions and bar charts

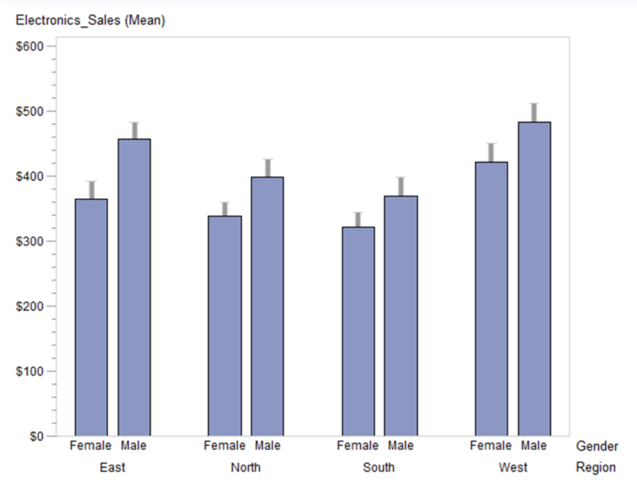


Interactions can also be represented using bar charts.

This one shows gender broken down by region.

There is a similar pattern for both genders.

Example: Interactions and bar charts



This bar chart shows regions broken down by gender.

The difference between genders is quite similar across regions.

Example: SAS code for compound bar charts

```
proc gchart data=work.store;
    vbar Region / group=Gender type=mean
            sumvar=Electronics_Sales errorbar=top;
    run;

proc gchart data=work.store;
    vbar Gender / group=Region type=mean
            sumvar=Electronics_Sales errorbar=top;
    run;
```


Example: Post-hoc comparisons

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Region	Electronics_Sales LSMEAN	LSMEAN Number
East	410.844156	1
North	368.615385	2
South	345.642292	3
West	453.028846	4

Least Squares Means for effect Region
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Electronics_Sales

i/j	1	2	3	4
1		0.0131	0.0001	0.0219
2	0.0131		0.2675	<.0001
3	0.0001	0.2675		<.0001
4	0.0219	<.0001	<.0001	

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Gender	Electronics_Sales LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
Female	361.955762	<.0001
Male	427.109578	

These post-hoc comparisons are for main effects only; they ignore the interactions between gender and region.

If significant, interaction effects are compared separately.

Example: Post-hoc comparisons

- The Tukey-Kramer post-hoc test reveals a statistically significant difference in means by gender (P-value < 0.0001).
- The only non-significant regional difference is between North and South (P-value = 0.2675).
- All other pairwise comparisons between regions are statistically significant at 5% level.
- As the interaction term was not significant, we disregard the corresponding post-hoc tests results in this scenario.

Example: Simple effects

The GLM Procedure
Least Squares Means

Region*Gender Effect Sliced by Gender for Electronics_Sales					
Gender	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Female	3	151877	50626	11.74	<.0001
Male	3	186505	62168	14.42	<.0001

- This comparison is for the effect of region sliced by gender.
- For both males and females, the effect of region is highly statistically significant, P-value < 0.0001.
  - Differences in mean electronic sales by region are statistically significant for each gender.

Example: Simple effects

The GLM Procedure
Least Squares Means

Region*Gender Effect Sliced by Region for Electronics_Sales					
Region	DF	Sum of Squares	Mean Square	F Value	Pr > F
East	1	73358	73358	17.01	<.0001
North	1	58565	58565	13.58	0.0003
South	1	25699	25699	5.96	0.0156
West	1	47114	47114	10.92	0.0011

- This comparison is for the effect of gender sliced by region.
- For all regions, the effect of gender is statistically significant, all P-values are less than 0.02.
  - Differences in mean electronic sales by gender are statistically significant for each region.

Example: SAS code for factorial ANOVA

This tells SAS to include Region and Gender as explanatory variables with all their interactions

```
ods graphics on;

proc glm data=store;
    class Region Gender;
    model Electronics_Sales=Region | Gender / ss3;
    lsmeans Region | Gender / pdiff adjust=tukey;
    /* Simple interaction effects */
    lsmeans Gender*Region / slice=Gender;
    lsmeans Gender*Region / slice=Region;
    run;
quit;

ods graphics off;
```

Example: Music sales

- The data set `store` contains the following variables:

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 on electronics
Total_Sales	Total sales

Example: Music sales

- Suppose we want to determine whether the mean of music sales varies by region and gender.
- We will check the assumptions and then conduct factorial ANOVA using PROC GLM.

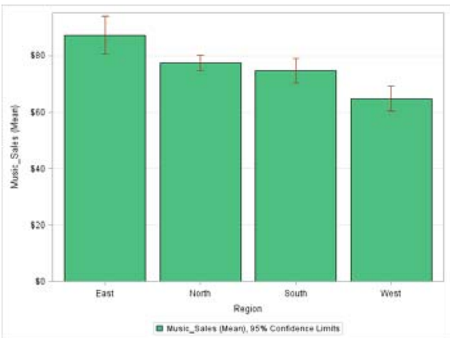
Example: Music sales

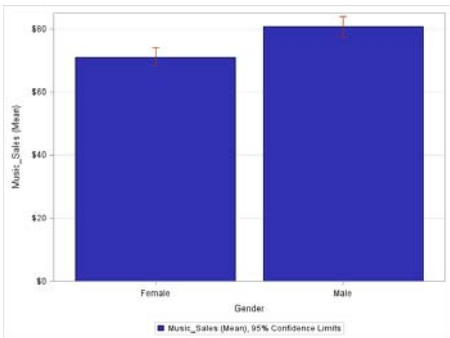
Descriptive Statistics

Analysis Variable : Music_Sales							
Region	Gender	N Obs	N	Mean	Std Dev	Minimum	Maximum
East	Female	22	22	77.045	16.450	50.000	110.000
	Male	14	14	103.571	11.507	85.000	125.000
North	Female	39	39	76.282	11.105	55.000	95.000
	Male	30	30	79.000	11.250	55.000	100.000
South	Female	23	23	73.261	16.488	45.000	100.000
	Male	22	22	76.136	11.226	60.000	100.000
West	Female	26	26	56.346	13.308	25.000	80.000
	Male	24	24	73.958	12.422	55.000	95.000

There appear to be some differences by gender across the four regions. Are these differences statistically significant?

Example: Music sales





Ignoring gender, sales in the East appear to be higher on average than in any other region, while sales in the West appear to be the lowest.

Ignoring region, males appear to be spending more on music on average than females.

Example: Music sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	21195.35081	3027.90726	17.96	<.0001
Error	192	32364.14919	168.56328		
Corrected Total	199	53559.50000			

R-Square	Coeff Var	Root MSE	Music_Sales Mean
0.395735	17.20768	12.98319	75.45000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Region	3	13139.46726	4379.82242	25.98	<.0001
Gender	1	7170.48161	7170.48161	42.54	<.0001
Region*Gender	3	4507.82347	1502.60782	8.91	<.0001

The overall model is highly significant, F(7,192) = 17.96, P-value < 0.0001. The R-squared is not large at 0.40 so there is considerable variability in music sales not accounted for by region and gender.

Example: Music sales

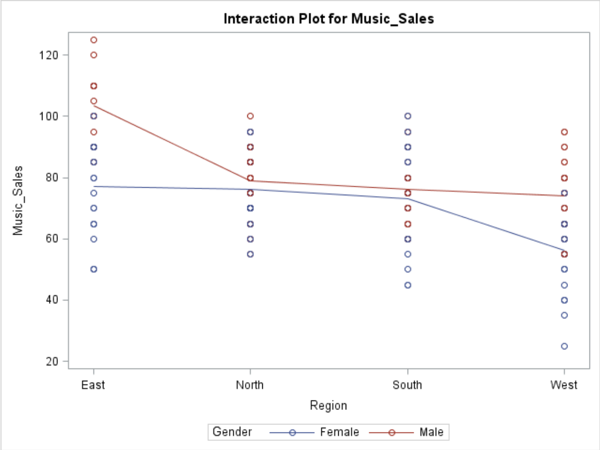
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	21195.35081	3027.90726	17.96	<.0001
Error	192	32364.14919	168.56328		
Corrected Total	199	53559.50000			

R-Square	Coeff Var	Root MSE	Music_Sales Mean
0.395735	17.20768	12.98319	75.45000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Region	3	13139.46726	4379.82242	25.98	<.0001
Gender	1	7170.48161	7170.48161	42.54	<.0001
Region*Gender	3	4507.82347	1502.60782	8.91	<.0001

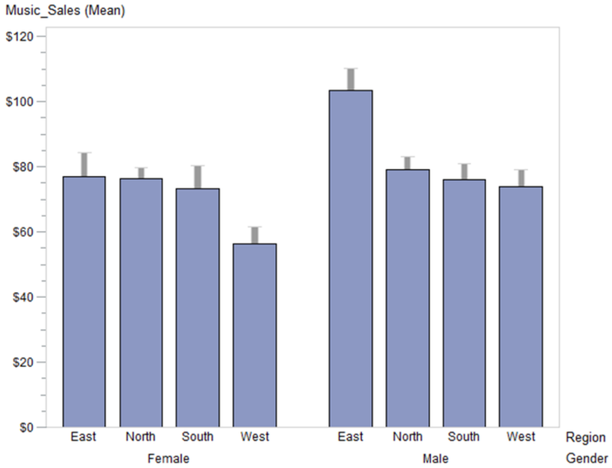
There is a highly significant main effect due to region, F(3,192) = 25.98, P-value < 0.0001, and gender, F(1,192) = 42.54, P-value < 0.0001. There is also a significant interaction, P(3,192) = 8.91, P-value < 0.0001.

Example: Interaction plot



The interaction plot confirms that in addition to significant main effects for gender and region, there is interaction. Difference in means for males and females are much greater for East and West.

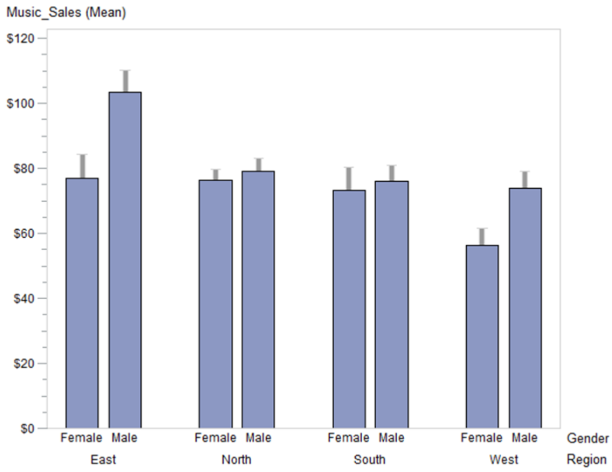
Example: Interactions and bar charts



This bar chart shows gender broken down by region.

There are different patterns for males and females.

Example: Interactions and bar charts



This bar chart shows regions broken down by gender.

North and South show similar means for males and females.

East and West have quite different patterns.

Example: Parameter estimates

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	73.95833333	B	2.65018299	27.91	<.0001
Region East	29.61309524	B	4.36620017	6.78	<.0001
Region North	5.04166667	B	3.55559359	1.42	0.1578
Region South	2.17803030	B	3.83215827	0.57	0.5705
Region West	0.00000000	B	.	.	.
Gender Female	-17.61217949	B	3.67514256	-4.79	<.0001
Gender Male	0.00000000	B	.	.	.
Region*Gender East Female	-8.91379454	B	5.76271412	-1.55	0.1236
Region*Gender East Male	0.00000000	B	.	.	.
Region*Gender North Female	14.89423077	B	4.84227055	3.08	0.0024
Region*Gender North Male	0.00000000	B	.	.	.
Region*Gender South Female	14.73668542	B	5.33830292	2.76	0.0063
Region*Gender South Male	0.00000000	B	.	.	.
Region*Gender West Female	0.00000000	B	.	.	.
Region*Gender West Male	0.00000000	B	.	.	.

These parameter estimates are not statistically significant.

Example: Parameter estimates

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	73.95833333	B	2.65018299	27.91	<.0001
Region East	29.61309524	B	4.36620017	6.78	<.0001
Region North	5.04166667	B	3.55559359	1.42	0.1578
Region South	2.17803030	B	3.83215827	0.57	0.5705
Region West	0.00000000	B	.	.	.
Gender Female	-17.61217949	B	3.67514256	-4.79	<.0001
Gender Male	0.00000000	B	.	.	.
Region*Gender East Female	-8.91379454	B	5.76271412	-1.55	0.1236
Region*Gender East Male	0.00000000	B	.	.	.
Region*Gender North Female	14.89423077	B	4.84227855	3.08	0.0024
Region*Gender North Male	0.00000000	B	.	.	.
Region*Gender South Female	14.73668542	B	5.33830292	2.76	0.0063
Region*Gender South Male	0.00000000	B	.	.	.
Region*Gender West Female	0.00000000	B	.	.	.
Region*Gender West Male	0.00000000	B	.	.	.

Mean for sales to males in the West

Differences in mean sales to males in other regions compared to the West

Difference between mean sales to females and males in the West

Differences between mean sales to females and males in a given region relative to the difference between them in the West.

Example: Post-hoc comparisons

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Region	Music_Sales LSMEAN	LSMEAN Number
East	90.3084416	1
North	77.6410256	2
South	74.6986166	3
West	65.1522436	4

Least Squares Means for effect Region
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Music_Sales

i/j	1	2	3	4
1		<.0001	<.0001	<.0001
2	<.0001		0.6411	<.0001
3	<.0001	0.6411		0.0025
4	<.0001	<.0001	0.0025	

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Gender	Music_Sales LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
Female	70.7336323	<.0001
Male	83.1665314	

There is a statistically significant difference in means by gender (P-value < 0.0001).

The only non-significant regional difference is between North and South (P-value = 0.6411).

Example: Post-hoc comparisons for interactions

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Region	Gender	Music_Sales LSMEAN	LSMEAN Number
East	Female	77.045455	1
East	Male	103.571429	2
North	Female	76.282051	3
North	Male	79.000000	4
South	Female	73.260870	5
South	Male	76.136364	6
West	Female	56.346154	7
West	Male	73.958333	8

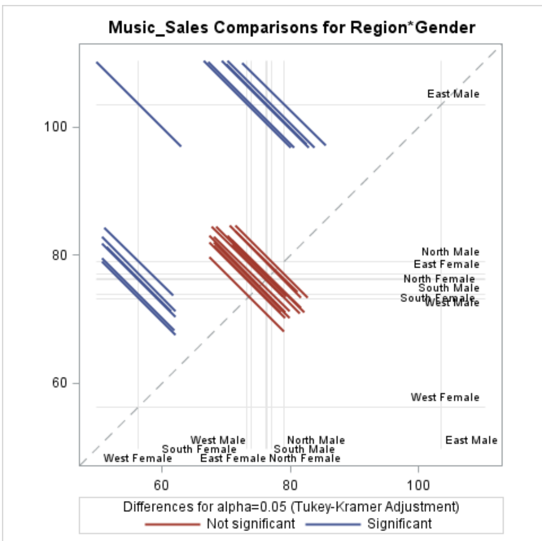
Least Squares Means for effect Region*Gender
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Music_Sales

i/j	1	2	3	4	5	6	7	8
1		<.0001	1.0000	0.9994	0.9771	1.0000	<.0001	0.9927
2	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
3	1.0000	<.0001		0.9890	0.9871	1.0000	<.0001	0.9972
4	0.9994	<.0001	0.9890		0.7529	0.9937	<.0001	0.8480
5	0.9771	<.0001	0.9871	0.7529		0.9955	0.0003	1.0000
6	1.0000	<.0001	1.0000	0.9937	0.9955		<.0001	0.9992
7	<.0001	<.0001	<.0001	<.0001	0.0003	<.0001		<.0001
8	0.9927	<.0001	0.9972	0.8480	1.0000	0.9992	<.0001	

E.g. 1: Mean sales for females in the North are significantly different only from mean sales for males in the East and for females in the West.

E.g. 2: Means sales for females in the West are highly statistically different from mean sales for all other groups.

Example: Post-hoc comparisons for interactions



Same information in a diffogram

Easier to understand?

Example: Simple effects

The GLM Procedure
Least Squares Means

Region*Gender Effect Sliced by Gender for Music_Sales					
Gender	DF	Sum of Squares	Mean Square	F Value	Pr > F
Female	3	7591.919530	2530.639843	15.01	<.0001
Male	3	8958.577742	2986.192581	17.72	<.0001

- For both males and females, differences in mean music sales by region are highly statistically significant, P-value < 0.0001.

Example: Simple effects

The GLM Procedure
Least Squares Means

Region*Gender Effect Sliced by Region for Music_Sales					
Region	DF	Sum of Squares	Mean Square	F Value	Pr > F
East	1	6019.922439	6019.922439	35.71	<.0001
North	1	125.261984	125.261984	0.74	0.3897
South	1	92.974308	92.974308	0.55	0.4586
West	1	3871.157051	3871.157051	22.97	<.0001

- The effect of gender is statistically significant for East and West only, P-values < 0.0001.
- Differences in mean music sales by gender for North (P-value = 0.3897) and South (P-value = 0.4586) are not statistically significant.

Example: Contrasts

- In factorial ANOVA, we can estimate differences of interest using contrasts, according to the same rules as for one-way ANOVA:
  - It does however get quite complicated with more than one factor! Other approaches may be a better way to go.
- Suppose we wish to compare music sales in the East to other regions:
  - Weights for this comparison are 3 -1 -1 -1.
- Suppose also that we are interested in the difference in mean music sales between males and females in the East:
  - Weights for this comparison are -1 1 on Gender, and -1 1 0 0 0 0 0 0 on interactions.

Example: Contrasts

Parameter	Estimate	Standard Error	t Value	Pr > |t|
East vs other regions	53.4334388	7.34435498	7.28	<.0001
Gender difference in the East	26.5259740	4.43871617	5.98	<.0001

- There is a highly statistically significant difference between mean music sales between East and the other regions (P-value < 0.0001).
- There is also a highly statistically significant difference between mean music sales for males and females in the East (P-value < 0.0001).
  - Mean music sales for males are significantly higher than for females in that region.

Example: SAS code for PROC GLM

```
ods graphics on;

proc glm data=store;
    class Region Gender;
      /* model Music_Sales=Region | Gender / ss3; */
    model Music_Sales=Region | Gender / ss3 solution;
    estimate 'East vs other regions' Region 3 -1 -1 -1;
    estimate 'Gender difference in the East' Gender -1 1
                     Region*Gender -1 1 0 0 0 0 0 0;
      lsmeans Region / adjust=tukey;
    lsmeans Region | Gender / pdiff adjust=tukey;
    lsmeans Gender*Region / slice=Region;
    lsmeans Gender*Region / slice=Gender;
    run;
quit;

ods graphics off;
```


Factorial ANOVA as a GLM

- Is there a relationship between a numerical variable and categorical variables of interest?
- Recall from linear regression:

  outcome = (model) + error

$$\hat{y} = b_0 + \underbrace{b_1 x_1 + \ldots + b_p x_p}_{\text{model}} + error$$

Multiple linear regression

- In the music sales example:
  - The response variable is music sales.
  - Predictors are dummy variables representing gender, region and interactions between groups formed by regions and gender.

Example: SAS code for dummy variables

```
data work.store_dummies;
    set work.store;

    if Gender='Female' then Female=1; else Female=0;

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

    if Region='East' and Gender='Female' then East_Female = 1;
    else East_Female = 0;
    if Region='North' and Gender='Female' then North_Female = 1;
    else North_Female = 0;
    if Region='South' and Gender='Female' then South_Female = 1;
    else South_Female = 0;
run;
```


Example: SAS code for regression

```
ods graphics on;

proc reg data=work.store_dummies plots=diagnostics;
    model Music_Sales=Female East North South
          East_Female North_Female South_Female;
    run;
quit;

ods graphics off;
```

Example: Music sales

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	7	21195	3027.90726	17.96	<.0001
Error	192	32364	168.56328		
Corrected Total	199	53560			

Root MSE	12.98319	R-Square	0.3957
Dependent Mean	75.45000	Adj R-Sq	0.3737
Coeff Var	17.20768		

The model is statistically significant at 1% level.

Regions, gender and interaction terms explain 37% of variability in music sales.

Example: Music sales

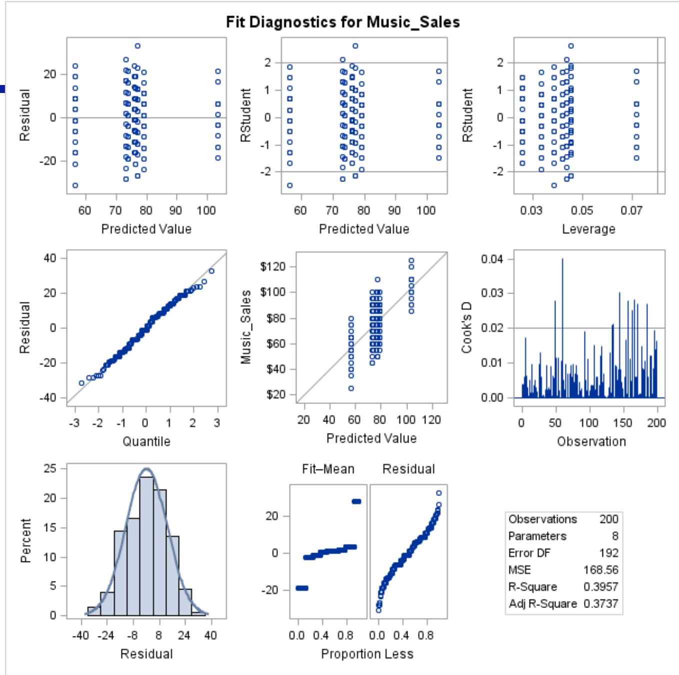
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	73.95833	2.65018	27.91	<.0001
Female	1	-17.61218	3.67514	-4.79	<.0001
East	1	29.61310	4.36620	6.78	<.0001
North	1	5.04167	3.55559	1.42	0.1578
South	1	2.17803	3.83216	0.57	0.5705
East_Female	1	-8.91379	5.76271	-1.55	0.1236
North_Female	1	14.89423	4.84227	3.08	0.0024
South_Female	1	14.73669	5.33830	2.76	0.0063

Base category are males in the West, represented by the intercept.

Slopes represent differences from the base category.

Note that parameter estimates and P-values presented here are the same as in the parameter table from PROC GLM.



How do we interpret these diagnostic plots?

49