

Lecture 17: Multiple Regression and ANOVA

Review Questions + Possibility of *Combining* the Two Models

Part I: Meddicorp (revised data)

- The Meddicorp Data is discussed in the text in Chapter 4 and again in Chapter 7. [*Chapter 7 was not assigned as required reading.*] We will repeat some of these analyses from the JMP perspective, and continue with a few additional ones.
- **The data we have here is a modified version of the one used in the class. The results of all the analyses may differ from what is in the text.** (The modifications make the output a little more interesting for our discussion.)

- Meddicorp Company sells medical supplies. The company markets in three regions of the United State.
- Meddicorp's management is concerned with the effectiveness of a new bonus program and the effectiveness of the advertisement. The bonus is provided to the sales people based on the performance. The variables considered in this study include:
 - **Sales:** Meddicorp's sales (in thousands of dollars) in each territory for 1999.
 - **Advert:** The amount spent on advertising in each territory (in hundreds) in 1999.
 - **Bonus:** The total amount paid in bonus in each territory (in hundreds) in 1999.
 - **Region:** *South*, *West* and *Midwest*. These regions are divided into territories of approximately equal sales potential, and the data comes from the territories within each region

One-Way ANOVA

1. As a background analysis, let's look at whether **Sales** are related to **Region**. For now, ignore the **Advert** and **Bonus** variables. Here are the standard tables for the analysis of **Sales** on **Region**:

Summary of Fit

Rsquare	0.669
Root Mean Square Error	????
Observations	26

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Region name	??	1045660	522830	????	<.0001
Error	??	517333	22493		
C. Total	??	1562993			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
MIDWEST	11	1498.6	45.2	1405.1	1592.2
SOUTH	7	1037.5	56.7	????	????
WEST	8	1162.5	53.0	1052.8	1272.2

Std Error uses a pooled estimate of error variance

Tables continued on next page.

Here is the first part of some output from the **Comparisons for each pair**:

Comparisons for each pair using Student's t

t	Alpha	t	Alpha
2.069	0.05	1.319	0.2

Here is the first part of some output from the **Comparisons for all pairs**:

Comparisons for all pairs using Tukey-Kramer HSD

q*	Alpha	t	Alpha
2.504	0.05	1.776	0.2

- Fill in the missing entries in the output tables.
- Is there significant evidence of a difference among the regions with respect to Sales?
- If so, which pair(s) of regions are significantly different?
[Including, which procedure should best be used for answering such a question?]
- Suppose that before looking at this data the manager of the **West** made a bet with the manager of the **South** that her region had higher **Sales** than his. Does looking at the data now provide significant evidence that her claim is valid at $\alpha = 0.05$? At $\alpha = 0.2$?

Multiple Regression

2. How are sales affected by Advertising and by Bonus paid? Based on outputs below for the regression of **Sales** vs. **Advert** and **Bonus** answer the following questions. [For now, ignore the variable, “Region”.]

	Correlations		
	Sales	Advert	Bonus
Sales	1.0000	0.9176	0.5876
Advert	0.9176	1.0000	0.4312
Bonus	0.5876	0.4312	1.0000

a) Based on the above table, does Advert or Bonus seem to have more impact on Sales?

Summary of Fit		
RSquare		????
Root Mean Square Error		87.55
Mean of Response		1271.1
Observations		26

b) Without looking further I know that here $R^2 > .841$. Why?

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1386712	693356	90.46
Error	23	????	7664	Prob > F
C. Total	25	????		<.0001

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-682.37	179.4	-3.80	0.0009
Advert	2.695	0.2563	10.51	<.0001
Bonus	2.097	0.6904	3.04	0.0059

Effect Tests				
Source	DF	Sum of Squares	F Ratio	Prob > F
Advert	1	847120	110.53	<.0001
Bonus	1	70701	9.22	????

- c) Write the regression equation.
- d) What is the interpretation of R^2 ?
- e) Do the SumSq terms in the Effect table add-up to be the ANOVA SSMModel? {Optional: *When do they do so?*}

f) For an additional territory having Advertising expenditures of 550 and Bonus expenditures of 275, what would you predict to be the Sales?

g) Approximately what would you expect the 95% prediction interval to be for this territory?

Multiple regression and ANOVA Combined

- Controlling on Region may affect the previous conclusions. We fit the data with **Advert**, **Bonus** and **Region**. The relevant output is shown on the next page.

- The model for this analysis is an additive model of the form

$$E(Y) = \beta_0 + \beta_1 x_{\text{Advert}} + \beta_2 x_{\text{Bonus}} + \gamma_{\text{Region}}$$

where x_{Advert} denotes the Advert value corresponding to Y , x_{Bonus} denotes the Bonus value corresponding to Y , and γ_{Region} is a numerical value corresponding to the effect of the Region of Y .

- Other features of this model are that β_1, β_2 are the corresponding slope coefficients, and the γ_{Region} constants have the special property (“side condition”) that $\gamma_1 + \gamma_2 + \gamma_3 = 0$.
- Normality and homoscedasticity of residuals is also assumed (as usual).

Summary of Fit

RSquare	0.939
Root Mean Square Error	67.15
Observations	26

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	1468296	367074	81.40
Error	21	94696	4509	Prob > F
C. Total	25	1562993		<.0001

Effect Tests

Source	DF	Sum of Squares	F Ratio	Prob > F
Advert	1	338385	75.04	<.0001
Bonus	1	35075	7.78	0.0110
Region name	??	81583	9.05	0.0015

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-253.43	177.90	-1.42	0.1690
Advert	2.13	0.25	8.66	<.0001
Bonus	1.53	0.55	2.79	0.0110
Region name[MIDWEST]	101.84	24.61	4.14	0.0005
Region name[SOUTH]	-53.45	24.73	-2.16	0.0424
Region name[WEST]	-48.39	19.31	-2.51	0.0205

a) What null hypothesis and alternative is being tested here by the ANOVA F-Ratio?

b) What is the DF value for **Region** in the Effect Test table?

c) Is **Bonus** significant after controlling for **Advert** and **Region**?
Carry out the hypothesis test at the 0.05 level.

d) Is **Region** significant after controlling for **Advert** and **Bonus**?
Carry out the hypothesis test at the 0.05 level.

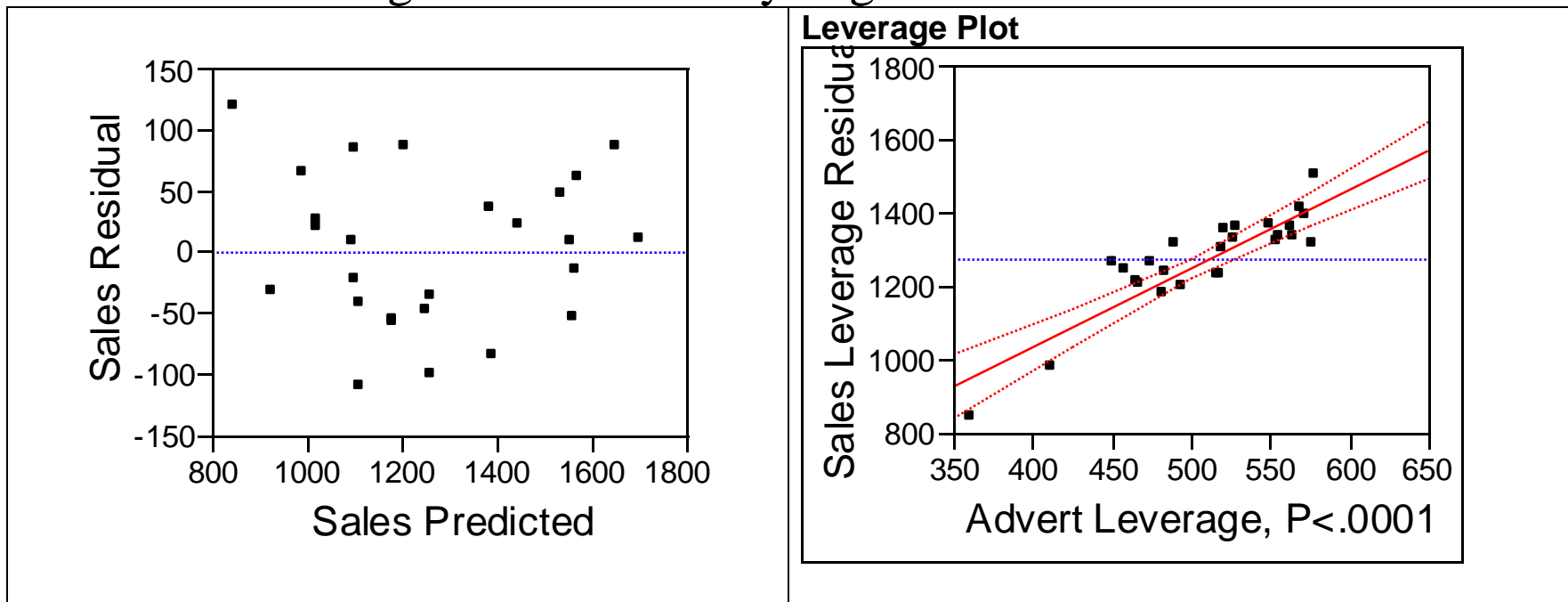
e) Use the model to predict **Sales** for an additional territory in the WEST having **Advertising** expenditures of 550 and **Bonus** expenditures of 275.

f) How would you find a confidence interval for **Sales** for the territory described in part e)? Would this CI be centered on the predicted value in part e)?

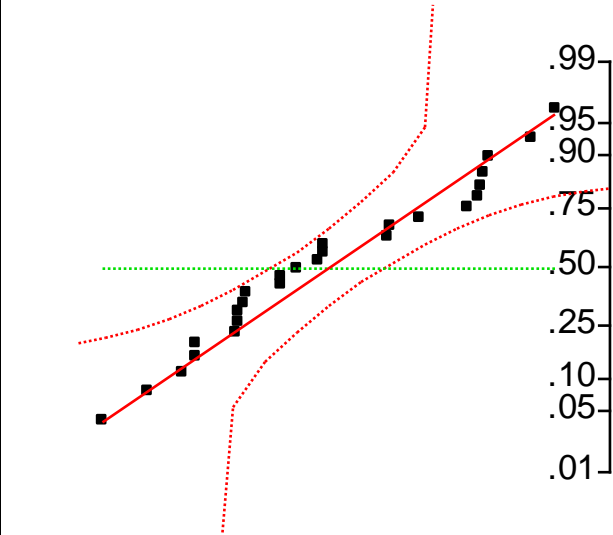
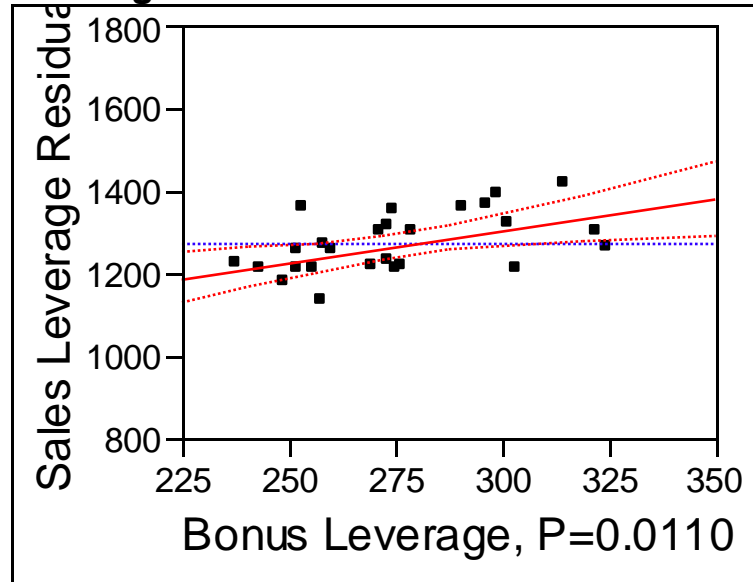
g) Based on the above model what is the estimated difference in the sales for two territories -- one from MIDWEST and one from the WEST, assuming that their **Advert** and **Bonus** are the same?

- Below (on this page and the next one) are the Residual Plot, two leverage plots and a normal quantile plot of residuals.

h) Do the residuals appear acceptably normal and homoscedastic? Are there any outliers or leverage points on these plots? If there are leverage points, do they appear “influential”? And, what would be the effect of excluding them and reanalyzing the data?



Leverage Plot



Normal Quantile Plot of Residuals from Multiple regression [x-axis (omitted) has values of Sales]

Comparison of the ANOVA Effects in a Combined Analysis

- For the one-way ANOVA discussed at the beginning of this lecture an all-pairs Tukey-Kramer test of all pairwise differences is given from JMP as:

Level	Mean
MIDWEST	1498.64
WEST	1162.53
SOUTH	1037.54

Levels not connected by same letter are significantly different

Level	- Level	Difference	Lower CL	Upper CL
MIDWEST	SOUTH	461.10	279.50	642.70
MIDWEST	WEST	336.10	161.58	510.63
WEST	SOUTH	125.00	-69.39	319.38

- A similar table is available in the combined analysis. The interpretation is that it gives all-pairs CI^s (familywise error rate) for the differences in mean Profit of the Regions *after* controlling for **Advert** and **Bonus**. –

- In JMP you can get this table by going to the red triangle at the Leverage Plot for **Region**. Here is the output:

LSMeans Differences Tukey HSD Alpha = 0.050 Q = 2.5206 LSMean[i] By LSMean[j]

Mean[i]-Mean[j] Lower CL Dif Upper CL Dif	MIDWEST	SOUTH	WEST
MIDWEST	0	155.29	150.23
	0	40.85	57.77
	0	269.73	242.70
SOUTH	-155.29	0	-5.06
	-269.73	0	-98.11
	-40.85	0	88.00

Level		Least Sq Mean
MIDWEST	A	1359.11
WEST	B	1208.88
SOUTH	B	1203.82

Levels not connected by same letter are significantly different

- Note that the LSMean (after controlling for **Advert** and **Bonus**) are different from the overall means in the one-way ANOVA analysis. The Hypothesis tests come out the same, but this is somewhat accidental.

Interaction Plots for Slopes in a Combined Analysis

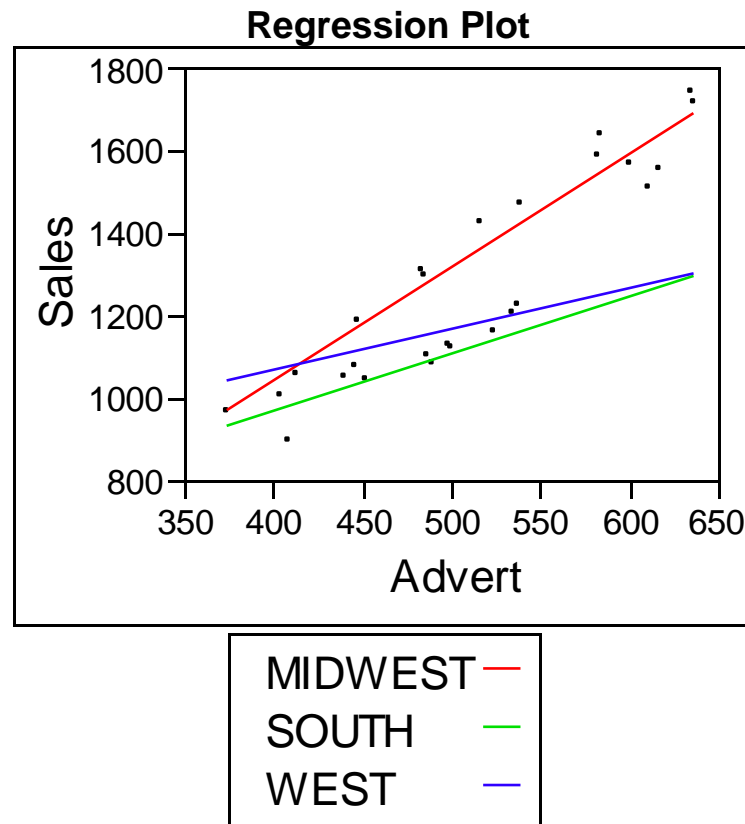
- To simplify the discussion we'll consider a model with only **Advert** and **Region**.

- Recall that the basic model here has

$$E(Y) = \beta_0 + \beta_1 x_{\text{Advert}} + \gamma_{\text{Region}}.$$

- This model has one slope coefficient for Advert. **It is the same slope coefficient for every region.**
- It is possible to have a different slope coefficient for each region.
- In JMP you get this by entering Advert and Region in the Fit Model Model Effects and then also adding the “Cross” of those two.
- The most eye-catching part of this output is the Regression plot that shows the different lines for each region. Here is that plot:

Output from Model having **Advert**, **Region** and **Region*Advert**



- This shows that the regression slope in the Midwest is larger than in the South or West, etc.

- You can get the usual types of tables for this analysis.
- The ANOVA table tests the overall hypothesis that none of the parameters matter:

Summary of Fit

RSquare	0.941
Root Mean Square Error	67.999
Observations	26

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1470514	294103	63.60
Error	20	92478	4624	Prob > F
C. Total	25	1562993		<.0001

- Of course we reject the null hypothesis that none of the parameters matter. (P-value << .0001).
- We can also get a Parameter Estimates table. (Use the Expanded Estimates option):

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	355.00	155.53	2.28	0.0335
Advert	1.69	0.32	5.30	<.0001
Region name[MIDWEST]	129.47	24.62	5.26	<.0001
Region name[SOUTH]	-91.17	31.98	-2.85	0.0099
Region name[WEST]	-38.30	25.08	-1.53	0.1425
Region name[MIDWEST]*(Advert-510.322)	1.03	0.36	2.83	0.0104
Region name[SOUTH]*(Advert-510.322)	-0.32	0.45	-0.72	0.4811
Region name[WEST]*(Advert-510.322)	-0.70	0.53	-1.33	0.1989

- It's easiest to understand these estimates by looking at an example. So, consider a territory in the **Region = WEST** with **Advert = 550**.

- The estimated **Profit** is

$$\widehat{\text{Profit}} = 355 + 1.69 \times 550 - 0.70 \times (550 - 510.322) + (-38.30)$$

$$= 1456$$

- The analysis also has **Effect Tests**. Here is the table and its interpretation:

Interpretation of Effect Test Table

Source	Effect Tests			
	DF	Sum of Squares	F Ratio	Prob > F
Advert	1	129871.66	28.09	<.0001
Region name	2	129674.23	14.02	0.0002
Region name*Advert	2	37293.82	4.03	0.0338

- The first line tests the null hypothesis that the **average** slope coefficient for **Advert** is non-zero (after controlling for Region).
- We conclude there is significant evidence that it is not **0**.
- The next line tests the null hypothesis that there are no differences among the region effects (after controlling for all the slope coefficients on Advert).
- We conclude there is significant evidence that the regions have different effect on **Profits** (after controlling for **Advert** in this model that allows different slope coefficients for advert).
- The last line tests that the slope coefficients for **Advert** differ in the different regions.
- We conclude at $\alpha = 0.05$ that there are (but not at $\alpha = 0.01$).