Topic 1 Part II: Multiple Regression

Note Title 11/29/2006

We now focus on the linear regression problem where we have several explanatory (independent) variables.

Example 1 (from Albright & Winston)

A factory manufactures various types of parts for automobils. The factory manager wish to see factors (variables) that affect overhead costs. While there may be many variables affecting these costs, we focus on two of these:.

- Machine hours (MachHrs): number of machine hours used per month.
- Production runs (ProdRuns): number of separate production runs per month.

(The production run involves setting up and reconfiguring machines.)

In this example we work with the following linear model:

$$\underset{\text{Cost}}{\text{Overhead}} = b_0 + b_1 * \text{MachHrs} + b_2 * \text{ProdRuns}$$

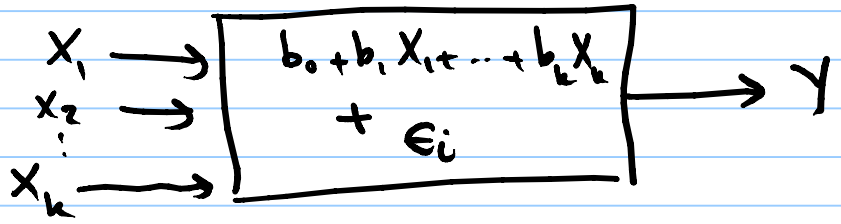
The probabilistic Model: This model is quite similar to the simple regression model:

- We have a number of **independent** variables $X_1 \dots X_k$
  Their values are **determined by experimenter** (they're **fixed**)
- We have one **dependent** variable Y
- We assume that for each combinations of $x_{1i}, x_{2i}, \dots, x_{ki}$ the **observed** value Y_i is given by

$$Y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki} + \epsilon_i$$

Where $b_0, b_1, \dots, b_k$ are **unknown parameters**.
and ϵ_i is a **random error term** following normal distribution $\epsilon_i \sim N(0, \sigma)$
with mean zero and variance σ^2 **independent** of $x_1, \dots, x_k$.

In this model we choose various values for $X_1, X_2, \dots, X_k$ and observe Y



We do not know the values of parameters $b_0, b_1, \dots, b_k$, and we do not know the variance σ^2.

- The distribution of Y_i is also normal:

$$Y_i \sim N(\underbrace{b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}}_{\text{mean of } Y_i}, \underbrace{\sigma}_{\text{stdev of } Y_i})$$

Estimating $b_0, b_1, \dots, b_k$

Suppose $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k$ are our estimates of $b_0, b_1, \dots b_k$

Then for choice $X_{1i}, X_{2i}, \dots, X_{ki}$ our <u>estimate</u> of

$y_i - \hat{y}_i =$
$y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_k x_{1k}$

y_i is $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \cdots + \hat{b}_k x_{ki}$

Also the observed value is y_i.

- Just like simple regression we select $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$ in such a way that sum of squares:

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{b}_0 - \hat{b}_1 x_{1i} + \cdots + \hat{b}_k x_{ki})^2$$

is minimized.

- As in simple regression, taking derivatives with respect to each $\hat{b}_i$ and setting them equal to zero leads to a $k+1$ linear system of $k+1$ unknowns that can be solved.
- The formulas for each $\hat{b}_i$ involves matrix algebra.
- Statistical software routinely calculate the b_i

Estimating σ: Since ϵ_i are normal, we can estimate them by residuals, that is $\hat{\epsilon}_i = y_i - \hat{y}_i$. From $\hat{\epsilon}_i$ we can

estimate σ^2 and σ:

$$\sigma^2 \approx \frac{SSE}{n-k-1} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-k-1} \qquad \sigma \approx S_e = \sqrt{\frac{SSE}{n-k-1}}$$

where n is the number of observations.

- Since we are estimating $k+1$ parameters $(b_0, b_1, \ldots, b_k)$ the degrees of freedom is $n-(k+1) = n-k-1$

Distribution of $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$

Since $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$ depend on observations $Y_1, Y_2, \ldots, Y_n$, which in turn are affected by random errors ϵ_i, the $\hat{b}_i$ are statistics estimating parameters b_i.

It can be shown that each $\hat{b}_i$ is normal

$$\hat{b}_i \sim N(b_i, \sigma(b_i))$$

The $\sigma(b_i)$ depend on σ, the common standard deviation of ϵ_i, which is unknown to us.

. The standard deviations $\sigma(b_i)$ can be estimated $s(b_i)$.

. Then the standardized $$\frac{\hat{b}_i - b_i}{s(b_i)} \sim t(n-k-1)$$

that is the b_i have the t- distribution with $df = n-k-1$.

. Statistical software (including Excel's data Analysis → Regression) compute $\hat{b}_i$ and confidence intervals for each

Testing relevance of the variables X_j in determining Y.

In the formula $$Y = b_0 + b_1 X_1 + \cdots + b_k X_k$$

- the jth variable X_j is irrelevant to Y if and only if its coefficient $b_j = 0$
- We can test this easily since we know the distribution

of estimates $\hat{b}_j$.

- For each b_j we have the following hypothesis test:

Null hypothesis H_0: X_j is irrelevant to Y ($b_j = 0$)

Alt. hypothesis H_a: X_j is relevant to Y ($b_j \neq 0$)

Excel and most statistical software produce p-values of the for each b_j.

Note: We have $k+1$ simultaneous tests, and thus $k+1$ p-values, one for each of $b_0, b_1, \ldots, b_k$.

Example 1 Continued In the spreadsheet regressionMultiple.xls we have used tools → data analysis → regression to estimate b_0, b_1 and b_2 in the model

$$\underbrace{\text{overhead}}_{Y} = b_0 + b_1 * \underbrace{\text{MachHrs}}_{X_1} + b_2 * \underbrace{\text{ProdRuns}}_{X_2}$$

Here the estimated $\hat{b}_0 = 3996.68$ $\hat{b}_1 = 43.54$ $\hat{b}_2 = 883.62$

This can be interpreted as follows:

overhead costs on average are 3996.68 plus, for each hour machines are in use, 43.54 is added to overhead, and for each production run 893.62 is added to overhead.

- The p-values of both b_1 and b_2 are very small indicating that b_1 and b_2 have explanatory value to overhead costs.
- The 95% confidence interval for b_1 is $36.23 \leq b_1 \leq 50.84$
- The 95% confidence interval for b_2 is $716.27 \leq b_2 \leq 1050.96$

Sum of squares, R^2 and R for multiple regression

Similar to simple regression we have three important sums of squares:

- Total variation: $SST = \sum_i (Y_i - \bar{Y})^2$
- Variation due to residuals: $SSE = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \hat{b}_0 - \hat{b}_1 X_1 - \cdots - \hat{b}_k X_k)^2$
- Variation due to regression: $SSR = \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (\hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_k X_k - \bar{Y})^2$

Again it can be shown that $SST = SSE + SSR$

- If there is perfect match then $\hat{Y}_i = Y_i$ and in this case $SSE = 0$ and $SST = SSR$
- If None of the $X_1, X_2, \ldots, X_k$ are relevant to Y and if all $\hat{b}_1 = \hat{b}_2 = \cdots = \hat{b}_k = 0$ then $E(\hat{Y}_i) = b_0$. In this case $SSR = 0$ and $SST = SSE$
- The coefficient of determination R^2 is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

This describes the percentage of variation in Y described by the model.

- In general it is possible that some or all X_i are relevant to Y (low p-values for each b_i) but R^2 is small, indicating there are more variables needed to be considered for explaining Y's variation.
- If R^2 is large (close to 1) it indicates the model explains Y closely.

Note: If R^2 is small (close to 0) it only cast doubt on the existance of linear relation between $X_1, \ldots, X_k$ and Y. There may still be strong non linear relation connecting Y to $X_1, \ldots, X_k$.

The relevance of the entire model or subsets of variables:

- R^2 is useful to see if the entire model has sufficient explantory power. But there is a precise test to see if the entire model is relevant or not:

Null hypothesis H_0: $b_1 = b_2 = b_3 = \dots = b_k = 0$ (the variables $x_1, \dots, x_k$ are irrelevant in determining y)

Alternative hypothesis H_a: at least one of $b_1, b_2, \dots, b_k$ is different from zero (The model as a whole has some relevance in determining y.)

Note: b_0 is not involved!

Before stating how we calculate the p-value of this test let's make the following observations:

- It is possible that the p-values of some or most of each individual b_i be large, but the p-value of the above test be small indicating significance of the model as a whole.

- It is possible that the p-value of the test above be small indicating high relevance of the entire model, but that R^2 be also small indicating that still there are more variables that could be added to the model to improve it.
- The similarity to ANOVA tests should be obvious: we're trying to see if several parameters $(b_1, b_2 \ldots, b_k)$ are equal to zero (and each other). In fact Analysis of Variance and the F- distribution play a central role:

The F- test For equality $b_1 = b_2 = \cdots = b_k = 0$;

- Under the null hypothesis that $b_1 = b_2 = \cdots = b_k$, and remembering that Variance of Y_i is σ^2 regardless of X_i's, the quantity

$$\frac{SSR}{\sigma^2} = \frac{1}{\sigma^2}\sum (\hat{Y}_i - \bar{Y})^2 \quad \text{follows} \quad \chi^2(k) \text{ distribution.} \quad (k = (k+1) - 1)$$

- Under all circumstances, i.e. whether H_0 is true or false, the quantity $\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}\sum (y_i - \hat{y}_i)^2$ follows $\chi^2(n-k-1)$ distribution.

- Therefore if the null hypothesis is true, then

$$F\text{-Statistic} = \frac{\frac{SSR/\sigma^2}{k}}{\frac{SSE/\sigma^2}{n-k-1}} = \frac{MSR}{MSE} \sim F(k, n-k-1).$$

Therefore, the p-value of the F-stat (also called F-ratio or F-value) can be calculated by

$$p\text{-value} = FDIST(F\text{-value}, k, n-k-1)$$

- Most software that carry out regression analysis, also provide ANOVA tables that look something like this:

ANOVA table

	df (degrees of freedom)	SS (sum of sq)	MS (Mean Sq)	F-value	p-value
SSR (regression)	k	SSR	$MSR = \frac{SSR}{k}$	MSR/MSE	
SSE (residual)	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$		
SST (Total)	$n-1$	SST			

$F_{dist}(F_{value}, k, n-k-1)$

Example 1 continued In regressionMultiple.xls file, "manufacturing overhead", the model was

$$overhead = b_0 + b_1 MachineHrs + b_2 NumRuns$$

- The fact that F-value = 107 and p value = 3.75×10^{-15} shows that the model as a whole has explanatory value
- The fact that $R^2 = 87\%$ shows that the model is a very good fit to the data.

Testing relerance of some of the variables: partial F-test.

- Thus far we have seen how to evaluate the relevance or contribution of individual variables, and the relevance of the entire set of variables.
- Now we will study the effect of a subset of variables in the model.

Example Suppose the manager of a commuter rail transportation is interested in determining the effects of various variables on the number of riders per week. An obvious variable is the price per ride. Using simple regression (see regressionMultiple.xls file) we see that price per ride is quite relevant (p-value = 2.7×10^{-10}) and $R^2 = 80\%$ shows that 80% of variation on rides is explained by the price per ride. However, the manager suspects that other

factors are important. She believes that population, average income and parking rates in the region also affect the ridership. She adds these variables to the model and runs the regression analysis. The results show that all new variables are also quite relevant and the $R^2 = 95\%$. The question is how do we measure collectively the additional relevance

Reduced and complete models:

Consider model 1:

$$Y = b_0 + b_1 X_1 + \cdots + b_k X_k$$

and model 2:

$$Y = b_0 + b_1 X_1 + \cdots + b_k X_k + b_{k+1} X_{k+1} + \cdots + b_{k+m} X_{k+m}$$

- Our goal is to see whether the additional m variables $X_{k+1}, \ldots, X_{k+m}$ collectively have explanatory power or not.

- This question is interesting because by themselves these m variables may be significant. But in the presence of other variables (i.e $x_1, \ldots, x_k$) their significance may be diminished.
- Model 1 is called the **reduced model**, while model 2 is called the **complete model**.
- Formally we wish to do the following test:
  - Null hypothesis H_0: The new variables $x_{k+1}, x_{k+2}, \ldots, x_{k+m}$ are irrelevant
    $(b_{k+1} = b_{k+2} = \cdots = b_{k+m} = 0)$
  - Alt hypothesis H_a: At least one of variables $x_{k+1}, \ldots, x_{k+m}$ is relevant
    (at least one of $b_{k+1}, b_{k+2}, \ldots, b_{k+m}$ is not zero)
- Let SSE_R be the SSE of the reduced model, that is the SSE of the regression $Y = b_0 + b_1 x_1 + \cdots + b_k x_k$

- Let SSE_C and MSE_C be the SSE and MSE the complete model, that is the SSE and MSE of the regression $Y = b_0 + b_1 X_1 + \dots + b_k X_k + b_{k+1} X_{k+1} + \dots + b_{k+m} X_{k+m}$
- If the null hypothesis, $b_{k+1} = \dots = b_{k+m}$, is true then it can be shown that

$$\text{partial F-ratio} = \frac{(SSE_R - SSE_C)/m}{MSE_C}$$

follows the F- distribution with $df_1 = m$ and $df_2 = n - (k+m+1)$.

- Thus if $\text{p-value} = \text{FDIST}(\text{F-ratio}, m, n-(k+m+1))$ is smaller than α we reject H_0 and declare, with $1-\alpha$ confidence that at least some of the new variables $X_{k+1}, \dots, X_{k+m}$ are relevant.

Example: Ridership in commuter rail system (see regressionMultiple.xls)
We wish to study which factors affect the volume of ridership in a metropolitan commuter system. First, we consider

only the average price per ride is considered. Thus in this example the reduced model has the estimated equation:

Reduced Model: $\text{Weekly\_riders} = 1167 - 266 \times \text{Price\_Per\_Ride}$

This means that on average 1167 riders per week, and for each \$1 increase in price 266 riders are lost per week. Here $SSE_R = 45325.45$ and $R^2 = 42.6\%$

$$p\text{-value}(\text{price\_Per\_Ride}) = 2.7 \times 10^{-10}$$

The small p-value shows that price per ride is quite relevant but R^2 of 42.6% shows that probably other variables also affect ridership.

Now we add population, income, and parking_rate as possible explanatory variables. The complete model has the following estimated equation:

Complete Model:

weekly_rides = 124.4 − 167 × price_per_ride + 0.62 × population − 0.05 × income + 194 × parking rate

Here again all new variables, individually, have significance because each have small p-values. We can also calculate the p-value of the new variables:

$SSE_C = 1015693$ $\quad df1 = m$ = number of new variables = 3

$SSE_R = 45325.45$ $\quad df2 = n-(m+k+1) = 22$

$MSE_C = 461.67$

$$\text{Partial } F = \frac{(SSE_R - SSE_C)/m}{MSE_C} = \frac{(45325.25 - 1015693)/3}{461.67}$$

$$= 25.39$$

and $FDIST(25.39, 3, 22) = 2.5 \times 10^{-7}$

indicating the new variables are significant.

More general Modeling using Multiple Regression

The multiple regression framework is extremely powerful and encompasses several types of models. Here are some examples:

Polynomial models:

Suppose we beleive that we have one explanatory variable X and one response variable Y. But instead of a linear relationship we believe a quadratic or higher order polynomial relation holds:

$$Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3 + \cdots$$

Our data is given by

Y	X	X^2	X^3	$\cdots$
Y_1	X_1	X_1^2	X_1^3	$\cdots$
Y_2	X_2	X_2^2	X_2^3	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$
Y_n	X_n	X_n^2	X_n^3	$\cdots$

$$y = b_0 + b_1 x + b_2 x^2$$

$$y = b_0 + b_2 x^2$$

This can be very easily modeled by multiple regression: Treat

X^2 as a new variable U, X^3 as a new variable V, and so on.

- The significance of each additional power of X can be measured by the p-value of its coefficient.
- If we wish to see the polynomial model results in significant improvement over the linear model, we can simply run the partial F-test with the reduced model the linear model: $Y = b_0 + b_1 X$

and the complete model the polynomial model: $Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3 + \ldots$

Example (see regressionMultiple.xls file, the powa worksheet.)

An electric power company wish to relate cost of producing electricity as a function of amount (units) of electricity produced. We consider two models:

Model 1: $Cost = b_0 + b_1 Units$

Model 2: $Cost = b_0 + b_1 Units + b_2 Units^2$

Compare the two models in the regressionMultiple.xls file. The partial F-test results in a p-value = .000356, Thus the $Units^2$ term adds significantly to the model.

observation: Note that the p-value of the F-test is the same as the p-value of the coefficient of $Units^2$. This is the case when the complete model has only one more variable than the reduced model.

Qualitative (Categorical) variables:

In regression models one or more variables may be qualitative.

Example: We wish to model the sales volume of cereal boxes to their price (a quantitative variable) and the color of box (a qualitative variable) so we wish to find the function

$$sales = f(price, color)$$

Suppose the cereal boxes come in three colors red, blue and green.

These three colors are called the **levels** of color variable. Also qualitative variables are called **factors** in regression analysis. Thus, we say: the factor color has three levels.

- To incorporate factors into a regression model we create one **dummy variable** for each level. In this example, we create one for red, one for blue and one for green.
- These new variable are either 0 or 1 depending on whether the given cereal box is that color or not. For example, suppose our data is as follows:

box	color	price	sales(100)/week
1	red	2.95	21
2	green	3.95	15
3	blue	3.15	29
4	red	2.99	31
⋮	⋮	⋮	⋮

See MultipleRegression.xls file "categorical factors" sheet

We create three new dummy variables so our data is augmented

box	color	price	sales (100)/week	red	blue	green
1	red	2.95	21	1	0	0
2	green	3.95	15	0	0	1
3	blue	3.15	29	0	1	0
4	red	2.99	31	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Observations:

1) There is exactly one 1 in each row.
2) If we know the zero-one values of two of the three dummy variables, we can determine the third one. Thus, one of the dummy variables is redundant and must be removed.
3) The removed dummy variable corresponds to a level which is called the base level.

With these considerations, let us choose red as our base level and thus remove it. Then our data is as follows:

blue	green	price	sales (100)/week
0	0	2.95	21
0	1	3.95	15
1	0	3.15	29
0	0	2.99	31
⋮	⋮	⋮	⋮

independent variables

response variable

The linear regression model is then:

$$\text{Sales} = \underset{1000}{b_0} + \underset{-20}{b_1 * \text{price}} + \underset{-100}{b_2 * \text{blue}} + \underset{+300}{b_3 * \text{green}}$$

from here we run the multiple regression analysis as usual.

- Suppose after conducting our analysis we obtain the following estimates:

	estimate	p-value
$\hat{b}_0$	1000	—
$\hat{b}_1$	-20	.0003
$\hat{b}_2$	-100	.001
$\hat{b}_3$	300	.09

Thus the estimated equation is:

$$\text{sales} = 1000 - 20 * \text{price} - 100 * \text{blue} + 300 * \text{green}$$

- The interpretation is as follows:
  - the base level is red:
  - the coefficient -3 of blue says:

    On average blue boxes sell -3 hundred boxes fewer than red boxes, with the same price. The p-value of .001 says that this is significant
  - The coefficient 1 of gree says:

On average green boxes sell 1 hundred boxes more than red boxes, with the same price. The p-value of 0.09 says that this is insignificant. (so we cannot be 95% sure that red and green boxes sales are significantly different.

Example: vacation expenditure

The worksheet "vacation expenditure" in regressionMultiple.xls file on another example with two categorical factors.

Here we wish to relate the response variable "amount spent in a vacation" to three explanatory variables: Income, career, and education, the last two are categorical.

we look at several models:

Model 1: Consider only income

Model 2: Income and career

model 3: Income and education

model 4: Income and career and education

model 5: Consider only education

For categorical factors career and education, we create dummy variables.

- Comparing model 1 and model 2: To see whether the addition of career to income was significant or not we must run a partial F-test for the relevance of all dummy variables coming from career
- Comparing model 1 and model 3: To see if adding education has a significant effect we run a partial F

test for relevance of dummy variables comming from education factor

- Comparing model 3 and model 5: Note that when we look at education alone (model 5) there is significant difference among various levels of education. We have the following estimated equation for model 5:

travelExpense = 1412. + 664 Educ2 + 3461 Educ3 + 6769 Educ4

p-values .59 .008 1.3×10^{-5}

Here is the interpretation:

- the base level (Educ1) spend $1412 on travel expense.
- people with education level 2 spend $664 more than educ1. But because of high p-value this is not significant.
- people with educ level 3 spend $3461 more than educ1.

This is significant because of small pvalue (.003)

- Finally people with education level 4 spend $6769 more than educ 1. This is also significant because of tiny p-value

Now let us consider model 3, where income is added to educ

The estimated equation is:

travelExpense = -2356 + 299 Educ2 -309 Educ3 -4946 Educ4 + 0.12 income

p-values .9 6.8 .698 .0004

This model says that the average travel expense for education level 1 is given by travel expense = -2356 + 0.12 x income.

At education level 2 travel expense = (-2356 + 299) + 0.12 income

At education level 3 travel expense = (-2356 - 309) + 0.12 income

At education level 4 travel expense = (-2356 - 4946) + 0.12 income

Note: When we add income to the model, the significance

of all education level evaporates (all p-values of education levels are large.) In fact the partial F-test with reduced model income only and with complete model income and education yields a p-value of 0.14, suggesting that education levels, collectively, are not significant when added to income.

- Interpretation The reason that in model 5, where we had only education as a factor, higher educated people had a higher travel expenditure was probably because they tended to have higher income. Once income was added to the model education's significance was already incorporated in the income and no more additional information seems to be added by the education factor.

ANOVA Modeling One-way (Completely randomized) as regression

Here we have one qualitative variable X, with say k levels and one response variable Y. The qualitative response variable with k levels results in $k-1$ dummy $(0-1)$ variables $X_1, X_2, \dots, X_{k-1}$, with level k as the reference level. Thus the regression model here is

$$Y = b_0 + b_1 X_1 + \dots + b_{k-1} X_{k-1}$$

Example: The car promotion problem again

Consider the example of promotions and car dealers again we have three types of promotions (prom1, prom2, prom3) and we wish to see if there is significant difference in car sales among the three types. We choose three groups of car dealers randomly and assign each to a different promotion. The car sales during a week are recorded:

Prom1	Prom2	Prom3
4	4	6
2	3	2
7	4	
3	3	
	3	
	1	

We have seen how to model this problem using sum of squares in between and within variables.

The Null hypothesis was that all promotions have equal affect on sales and the alternative was that at least one of the promotions is significantly different, but whether it leads to higher or lower sale we don't, and won't know.

- We now model this problem using regression, relating sales as a function of the categorical variable "promotion" at three levels "prom1", "prom2" and "prom3". Let us choose "prom1"

as the base level and create two dummy variables prom 2 and prom3. Our data can now be written as:

response: sale; explanatory (predictor): prom 2, prom 3

sale	prom 2	prom 3
4	0	0
2	0	0
7	0	0
3	0	0
4	1	0
3	1	0
4	1	0
3	1	0
3	1	0
1	1	0
6	0	1
2	0	1

We run regression on this model and obtain:

$sale = 4.9 - 2.45 \times prom2 - 0.16 \times Prom3$

P-values: 0.005 0.86

(See regressionMultiple.xls file, one way ANOVA regression sheet.)

Interpretation: On average under promotion 1 (the base level) 4.9 ($=\hat{b}_0$) cars are sold. This is the mean of prom1 sales. promotion 2 on average results in 2.45 fewer sales than promotion 1. This is significant because p-value is small. In Prom3 on average sales are 0.15 lower promotion 1, but this is not significant because p-value of prom3 is large.

Note: The coefficient of prom2 and prom3 indicate by how much they are lower or higher than the base level. Thus the null hypothesis that coefficient of prom2 equals zero, means that there is no difference between prom2 and prom1. (In this example this null is rejected.) Similarly the null hypothesis that coefficient of prom3 equals zero means that there is no difference between

sales of base level (prom1) and prom3 (In this example this null is <u>not</u> rejected.) Finally the null hypothesis that

coefficient of prom2 = coefficient of prom3 = 0

means that all promotions have equal sales. (In this case the p-value of the F-ratio is 0.0088, and so it is rejected.) Note that this p-value is exactly the same as the p-value given by traditional ANOVA. Regression analysis, however, gives a lot more information.

Modeling two-way (randomized block) design with regression

Now recall that to control variation caused due to differences among car dealers skills, we add car dealer as the blocking factor. Thus we choose eight dealers and ask each to push one of the

promotions in each of three weeks. The "dealer" at eight levels is a new factor. We take the "dealer 1" as our base level and create seven dummy variables dealer2, ..., dealer8 for dealer factor. The regression model is:

$$Sales = b_0 + b_1 \times prom2 + b_2 * prom3 + b_3 * dealer2 + \cdots + b_9 * dealer8$$

Here b_0 is the average sales under promotion 1 (base promotion) and dealer 1 (base dealer). b_1 and b_2 are the amounts by which promotions 2 and 3 differ from promotion 1 in sales. And b_3 through b_9 are the amounts by which dealers are different in sales. To see if the promotions as a whole are significant we must calculate the partial F test with reduced model: $Sales = b_0 + b_1 * dealer2 + \cdots + b_7 dealer8$ and complete model:

$$\text{sales} = b_0 + b_1 \times \text{promo2} + b_2 \times \text{promo3} + b_3 \times \text{dealer2} + \cdots + b_9 \times \text{dealer8}$$

(See regressionMultiple.xls, two-way ANOVA regression sheet.)

Case Study: Bank gender discrimination

Female employees of a bank brought a lawsuit against the bank alleging gender discrimination. Data for a large group of male and female employees including their salary, gender, the year they were hired, level of education and their job grade. Based on this data we wish to examine statistically the claim of gender discrimination. The complete analysis is in the file bank.xls. Please study this carefully.