

Read M&M Chapters 2 and 11 again. Section leaders will decide how much of Chapters 12 and 13 to cover formally; they will assign the readings.
Variations on regression: explanatory categorical variables.

Today I will concentrate on the way categorical variables are handled by least squares. It is traditional to think of problems where the predictor variables are all categorical—that is, where each predictor merely indicates the presence or absence of some effect—as examples of *analysis of variance*, as described in Chapters 12 and 13 of M&M.

You will see that analysis of variance is just a special form of regression. There are a few small technical points to note, and a few slight differences in interpretation, but the main ideas are the same. To make the point that categorical variables can be handled in much the same way as other predictors, I will present an example where a categorical predictor is added to improve (slightly) an ordinary regression fit. The data, which I believe had their origin in a 1979 article in *Consumer Reports*, come packaged with the Splus statistical program.

The full data set contains a large amount of information about different models of automobile: length, wheel base, seating room, gear ratios, mileage, . . . , price, horsepower, and so on.

Somewhat arbitrarily, I chose to look at the relationship between price and horsepower (variables Price and HP). The categorical variable *Country* indicates the country where the car was manufactured. I extracted from the full data set the subset corresponding to only four countries. For reasons that you will soon understand, I added columns *logPrice* (the logarithm to base 10 of the price) and *logHP* (the logarithm to base 10 of the horsepower), and four columns that indicate the country in a slightly more verbose way than the Country column:

```
MTB > let c26 = ('Country' = "Germany")
MTB > let c27 = ('Country' = "Japan")
MTB > let c28 = ('Country' = "Japan/USA")
MTB > let c29 = ('Country' = "/USA")
MTB > names c26 'Germany' c27 'Japan' c28 'Japan/USA' c29 'USA'
```

Row	name	HP	Price	Country	logPrice	logHP	Germany	Japan	Japan/USA	USA
1	Acura Integra	130	11950	Japan	4.07737	2.11394	0	1	0	0
2	Acura Legend	160	24760	Japan	4.39375	2.20412	0	1	0	0
3	Audi 100	130	26900	Germany	4.42975	2.11394	1	0	0	0
4	Audi 80	108	18900	Germany	4.27646	2.03342	1	0	0	0
5	BMW 325i	168	24650	Germany	4.39182	2.22531	1	0	0	0
6	BMW 535i	208	33200	Germany	4.52114	2.31806	1	0	0	0
7	Buick Century	110	13150	USA	4.11893	2.04139	0	0	0	1
8	Buick Electra	165	20225	USA	4.30589	2.21748	0	0	0	1
...
42	GEO Metro	55	6695	Japan	3.82575	1.74036	0	1	0	0
...
83	Toyota Corolla	102	8748	Japan/USA	3.94191	2.00860	0	0	1	0
84	Toyota Cressida	190	21498	Japan	4.33240	2.27875	0	1	0	0
85	Toyota Supra	200	22860	Japan	4.35908	2.30103	0	1	0	0
86	Toyota Tercel	78	6488	Japan	3.81211	1.89209	0	1	0	0
87	Volkswagen Corrado	158	17900	Germany	4.25285	2.19866	1	0	0	0
88	Volkswagen Jetta	100	9995	Germany	3.99978	2.00000	1	0	0	0
89	Volkswagen Vanagon	90	14080	Germany	4.14860	1.95424	1	0	0	0

1. Transformations

How well is the price predicted by the horsepower?

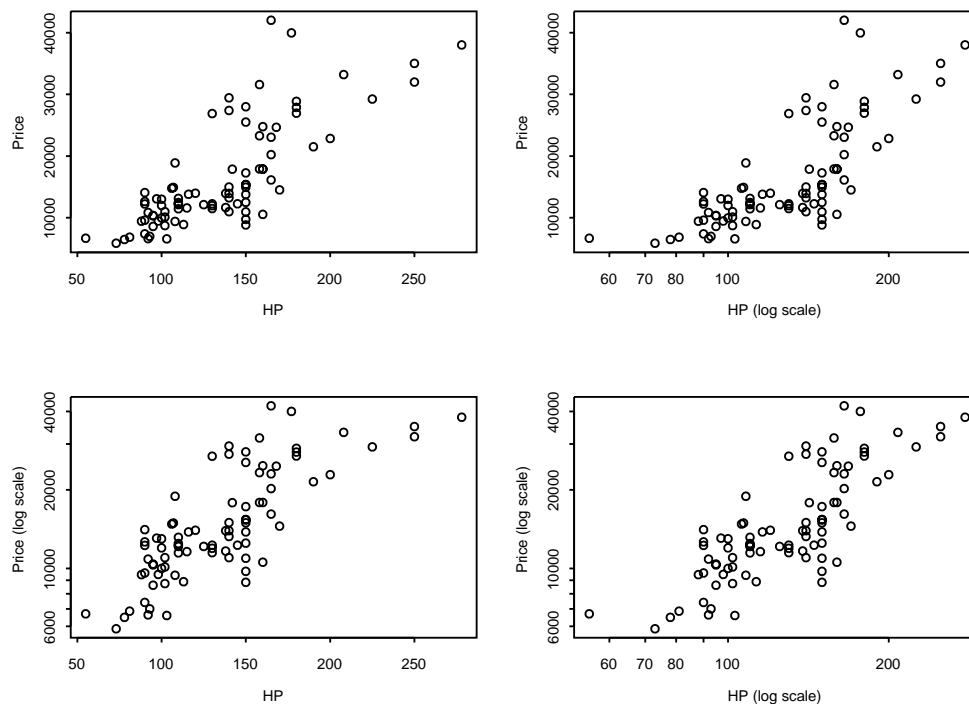
There is quite a large variation in price, as you can see from the plot of Price versus HP (the picture in the north-west corner of the next display). Notice the wedge shape of the region where most of the points fall. The variability in price seems to increase with price. (Does that phenomenon make sense? Would a customer be more sensitive to a five hundred dollar price difference in a \$7000 car or in a \$40,000 car?)

If we merely regressed Price on HP the higher-priced cars would have a disproportionate control over the fit. Also, a model assuming independent $N(0, \sigma)$ errors, with the same σ for each car, would clearly be suspect.

The variability becomes more consistent with a fixed- σ model if we plot Price on a log scale, as in the lower two pictures. The vertical axes in those two pictures are still labelled with prices, but the vertical position for each point is determined by the logarithm (to base 10) of the price. Notice that the vertical distance between the \$10,000 and \$20,000 prices is now the same as the vertical distance between the \$20,000 and \$40,000 prices.

Equivalently, I could have used *logPrice* on the vertical axis, but then I would have had some fiddling to get the tick marks corresponding to easily comprehended prices. Which would you find easier to interpret: a price of \$20,000, or a logPrice of 4.3?

It is often a good idea to work with logarithms of variables whose distributions spread over a large range (that is, ratio of largest to smallest is much greater than 1).



The two pictures on the right-hand side show the effect of working with horsepower on a log scale.

It seemed to me that I would have most luck with a linear fit to the points shown in the south-east corner. Thus I decided to work with the logarithms (to base 10) of both Price and HP.

```
MTB > let c24 = logten('Price')
MTB > let c25 = logten('HP')
MTB > names c24 'logPrice' c25 'logHP'
```

2. Country effects

You should be able to interpret the printed output from a regression of logPrice on logHP:

```
MTB > Regress 'logPrice' 1 'logHP';
SUBC> Constant;
SUBC> Brief 2.
```

Regression Analysis

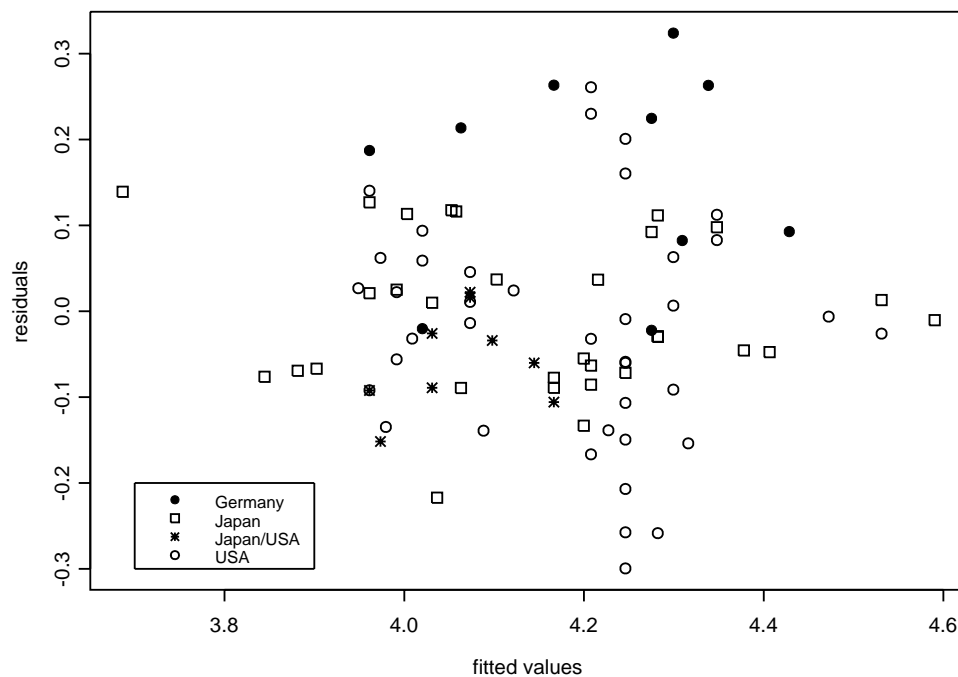
The regression equation is $\text{logPrice} = 1.45 + 1.28 \text{ logHP}$

Predictor	Coef	StDev	T	P
Constant	1.4520	0.2197	6.61	0.000
logHP	1.2840	0.1040	12.35	0.000

S = 0.1256 R-Sq = 63.7% R-Sq(adj) = 63.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2.4060	2.4060	152.44	0.000
Residual Error	87	1.3731	0.0158		
Total	88	3.7791			



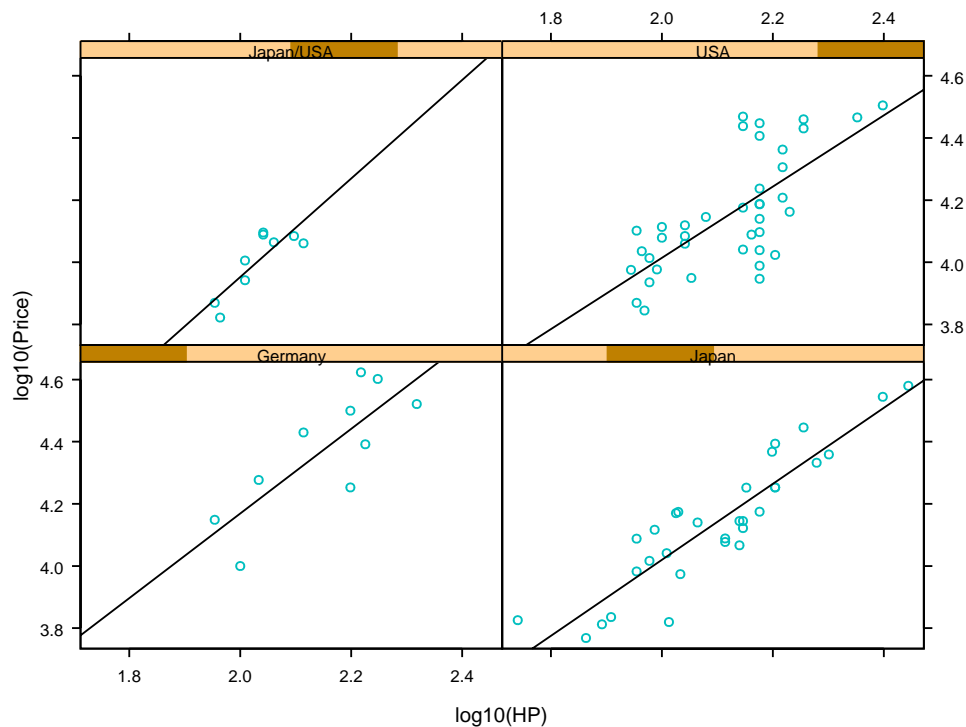
Does the point near the left edge correspond to a car or to an expensive lawn mower?

Obs	logHP	logPrice	Fit	StDev Fit	Residual	St Resid
42	1.74	3.8258	3.6867	0.0406	0.1391	1.17 X

I have used different plotting symbols to draw your attention to the effects of Country on the residuals. Doesn't it seem that the residuals for Germany are consistently higher? Do German cars cost more than cars of similar horsepower from other countries? The other variables in the data set might shed some light on the pricing, but I won't investigate them in this lecture.

3. Separate regressions

If we make separate least squares fits for the cars from each country, the German effect becomes even more apparent:



Notice that the slopes of the least squares lines are similar for each country, but the intercepts are somewhat different.

4. Indicators for countries

We can force a least squares fit with the same slopes but (possibly) different intercepts by modelling the $\log\text{Price}_{c,j}$ for the j th car in the c th country as

$$\log\text{Price}_{c,j} = \alpha_i + \beta \log\text{HP}_{c,j} + \epsilon_{c,j}$$

The (c, j) subscripts let us indicate the appropriate country for a particular car.

Alternatively, we can indicate the country by means of *indicator variables* to denote whether a car comes from a specific country, or not. That is, the indicator *German* contains a 1 in the i th row if the car for that row comes from Germany, a zero otherwise; and so on. There is some redundancy between the four country indicator variables, because, for each car, exactly one of the indicators takes the value 1. Maybe the names for the indicators are a bit klunky, but at least I can remember what they indicate.

The model with indicators takes the form

$$\begin{aligned} \log\text{Price}_i = & \alpha + \beta \log\text{HP}_i \\ & + \gamma_G \text{Germany}_i + \gamma_J \text{Japan}_i + \gamma_{JU} \text{Japan/USA}_i + \gamma_U \text{USA}_i + \epsilon_i, \end{aligned}$$

where the ϵ_i 's are independent $N(0, \sigma)$ variables. I have also written the theoretical coefficients in a notation that helps me keep track of what each value represents. You could call the five predictor variables x_1, x_2, \dots, x_5 , and write β_1, \dots, β_5 for the theoretical coefficients if you wanted the equations to look more like the ones in M&M. Here is what we get when we add the four indicators in as predictors:

```
MTB > Regress 'logPrice' 5 'logHP' 'Germany' 'Japan' 'Japan/USA' 'USA';
SUBC> Constant;
SUBC> Brief 3.
```

Regression Analysis

- * USA is highly correlated with other X variables
- * USA has been removed from the equation

The regression equation is

$$\text{logPrice} = 1.58 + 1.21 \text{ logHP} + 0.185 \text{ Germany} + 0.0145 \text{ Japan} - 0.0415 \text{ Japan/USA}$$

Predictor	Coef	StDev	T	P
Constant	1.5770	0.2038	7.74	0.000
logHP	1.21451	0.09558	12.71	0.000
Germany	0.18538	0.03983	4.65	0.000
Japan	0.01452	0.02709	0.54	0.594
Japan/US	-0.04150	0.04239	-0.98	0.330

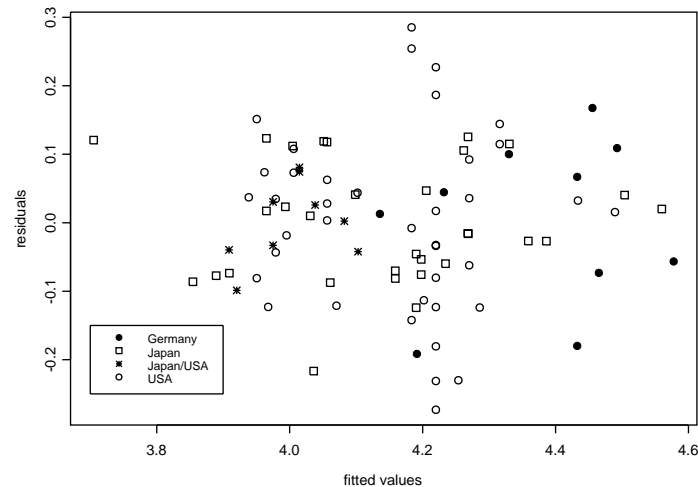
S = 0.1121 R-Sq = 72.0

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	2.72274	0.68069	54.13	0.000
Residual Error	84	1.05635	0.01258		
Total	88	3.77909			

Source	DF	Seq SS
logHP	1	2.40599
Germany	1	0.29522
Japan	1	0.00949
Japan/US	1	0.01205

What's going on? Why doesn't Minitab like my USA indicator? Why can the last predictor be discarded with impunity? The answer is that we have given Minitab more parameters to play with than it needed to get a least squares fit. If we added any constant to the intercept coefficient α , and subtracted the same constant from each γ , we would get exactly the same set of theoretical means. In particular, we could choose γ_U as the subtracted constant, thereby ensuring that the coefficient of the USA predictor is zero. Whatever Minitab wants to do with α , γ_G , γ_J , γ_{JU} , γ_U it can also do with α , γ_G , γ_J , γ_{JU} , taking γ_U as zero.



Did the inclusion of the indicators for countries improve the fit significantly?

5. Cautions regarding interpretation

In general, whenever Minitab finds that the contribution to the fit from a predictor variable can be closely approximated by contributions from other predictor variables, which appear earlier in the list of predictors, it discards the nearly-redundant predictor.

The redundancy makes it a little harder to interpret the individual coefficients. It is better to treat the four country indicators as a single contribution to the fit—a contribution with 3 degrees of freedom—beyond what is already given by the intercept and the logHP terms, without trying to break out the contributions due to each separate country:

Terms added sequentially (first to last)					
	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
log10(HP)	1	2.405987	2.405987	191.3221	0.000000000000
Country	3	0.316754	0.105585	8.3960	0.00006050361
Residuals	84	1.056349	0.012576		

Notice that

$$0.316754 \approx 0.29522 + 0.00949 + 0.01205$$

The Country Sum of Squares is just the sum of the three Sequential Sums of Squares from the last table in the previous Minitab display.

The subtleties in the interpretation of the regression output, when there are redundant predictors, are illustrated by the following two regressions. In each case, Minitab calculates the same fitted values, and the same estimate for σ , but the estimated coefficients are different.

Change order of predictors

Enter the predictor variables in a different order, putting *German* last. Then Minitab regards the indicator for German cars as redundant:

```
MTB > Regress 'logPrice' 5 'logHP' 'Japan' 'Japan/USA' 'USA' 'Germany';
SUBC> Constant;
SUBC> Brief 2.
```

Regression Analysis

```
* Germany is highly correlated with other X variables
* Germany has been removed from the equation
```

The regression equation is

$$\text{logPrice} = 1.76 + 1.21 \text{ logHP} - 0.171 \text{ Japan} - 0.227 \text{ Japan/USA} - 0.185 \text{ USA}$$

Predictor	Coef	StDev	T	P
Constant	1.7623	0.2086	8.45	0.000
logHP	1.21451	0.09558	12.71	0.000
Japan	-0.17086	0.04108	-4.16	0.000
Japan/US	-0.22687	0.05276	-4.30	0.000
USA	-0.18538	0.03983	-4.65	0.000

S = 0.1121 R-Sq = 72.0

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	2.72274	0.68069	54.13	0.000
Residual Error	84	1.05635	0.01258		
Total	88	3.77909			

Source	DF	Seq SS
logHP	1	2.40599

Japan	1	0.00200
Japan/US	1	0.04236
USA	1	0.27240

Omit intercept

If we explicitly force Minitab to omit the intercept term, then all four country indicators are retained:

```
MTB > Regress 'logPrice' 5 'logHP' 'Germany' 'Japan' 'Japan/USA' 'USA';
SUBC> NoConstant;
SUBC> Brief 3.
```

Regression Analysis

The regression equation is

$\text{logPrice} = 1.21 \text{ logHP} + 1.76 \text{ Germany} + 1.59 \text{ Japan} + 1.54 \text{ Japan/USA} + 1.58 \text{ USA}$

Predictor	Coef	StDev	T	P
Noconstant				
logHP	1.21451	0.09558	12.71	0.000
Germany	1.7623	0.2086	8.45	0.000
Japan	1.5915	0.2016	7.89	0.000
Japan/US	1.5355	0.1978	7.76	0.000
USA	1.5770	0.2038	7.74	0.000

S = 0.1121

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	1543.03	308.61	24540.05	0.000
Residual Error	84	1.06	0.01		
Total	89	1544.08			

Source	DF	Seq SS
logHP	1	1542.02
Germany	1	0.21
Japan	1	0.03
Japan/US	1	0.01
USA	1	0.75

You might try to figure out the relationship between the various tables and estimates for the three ways of expressing the model. In particular, try to explain why the fitted regression lines are the same in each case, even though the coefficients differ:

1.58	$+1.21 \times \text{logHP}$	$+0.185 \times \text{Germany}$	$+0.0145 \times \text{Japan}$	$-0.0415 \times \text{Japan/USA}$	
1.76	$+1.21 \times \text{logHP}$		$-0.171 \times \text{Japan}$	$-0.227 \times \text{Japan/USA}$	$-0.185 \times \text{USA}$
	$1.21 \times \text{logHP}$	$+1.76 \times \text{Germany}$	$+1.59 \times \text{Japan}$	$+1.54 \times \text{Japan/USA}$	$+1.58 \times \text{USA}$