Log linear model with dummy variables

This example is from Exercise 12.10, Griffith, Hill and The Judge [1993, p. 427-429]. The data set contains weekly sales of a major brand of canned tuna by a supermarket chain in a large Midwest city in the United States. The slope equation of interest is: ln (SALES) = 0 + 1 PRICE1 + 2 PRICE2 + 3 PRICE3 + 4 D1 + 5 D2 + e where D1 and D2 are imaginary variables for two different advertising systems. The child variable is in the log form. What is the impact of phantom variables on weekly sales of canned tuna? Discussion on the interpretation of transactions of phantom variables when the dependent variable is converted to log, in: Halvorsen, R. and Palmquist, P., interpretation of phantom variables in the American Economic Review, Vol. 70, 1980, pp. 474-475. Kennedy, P., an estimate with properly interpreted phantom variables in the semelogretham equations, American Economic Review, Volume 71, 1981, p. 801. The result developed in the above papers is that if b is the estimated coefficient of an imaginary variable, V (b) is the estimated variation of b, the g = 100 (exp (b-V)/1) gives an estimate of the effect of the imaginary variable percentage on the variable being explained. It is also important: how do we interpret price variable coefficients? Price variables in the levels and the child variable in the log form. In this case, 100 (1) gives the rate of change in canned tuna sales to 1 unit change in PRICE1 (holding each other constant). ShazAM commands (file name: tuna. SHA) below estimate slopslope equation coefficients and calculate some test statistics. The impact ratio on each fake ad variable is also calculated on canned tuna sales. Sample 1 52 reading (tuna.txt) sales price1 PRICE2 PRICE2 PRICE3 D1 D2 GENR LSALES = LOG (sales) * OLS LSALES PRICE1 PRICE1 PRICE2 D1 D2 / LOGLIN COEF = BETA STDERR = SE * Test hypothesis Test D1 test = 0 D2 test = 0 END D1 test = D2 * Estimate the effect of the d1 fake variable percentage on sales GEN1 C1 = BETA: 4 GEN1 SE1 = SE: 4 General 1 G1 = 1 *EXP (EXP 1- SE1 * SE1 /2) - 1) * Estimate the impact of the d2 fake variable percentage on sales GEN1 C2 = BETA: 5 GEN1 SE2 = SE: 5 G2 100 * () EXP (C2 - SE2 * SE2/2) - 1) PRINT G2 off the COEF option = BETA on the OLS command saves estimated transactions in the new variable beta and STDERR = SE option saves estimated standard errors from estimated transactions in the new variable SE. These results are later used to calculate the proportion of effects of fake advertising variables on sales. The LOGLIN option is selected on the OLS command. When using this option, flexibility is calculated in sample methods assuming the specifications of a semi-logarithmic model where the child variable is in the form of a record but the explanatory variables are at the levels. Let's say b1 is the coefficient estimated to be Variable price1 and MP1 is average PRICE1. The flexibility assessed on average is: b1 (MP1) the flexibility reported in the last column of the SHAZAM OLS estimate should be interpreted with caution. That is, they may not be suitable for some explanatory variables. For example, the reported flexibility of phantom variables is likely to have no meaningful explanation. ShazAM output can be displayed. The price flexibility assessed in the sample method (close to 2 decimal positions) is: the flexibility price variable1-2.93 PRICE2 0.93 PRICE3 1.02 Positive flexibility for PRICE2 and PRICE3 giveevidence that brand 2 and brand 3 are alternatives to brand 1. The negative flexibility of your PRICE1 price as expected - canned tuna sales will drop 1 in response to any price increase. The results of the estimate show that the estimated coefficients of the d1 and D2 phantom variables are significantly different from 0. A common test of the hypothesis: H0: 4 = 5 = 0 gives an F test statistic of 42.0. 5% the critical value of F-A distribution with (2,46) degrees of liberty is 3.20. This gives strong evidence to reject an empty hypothesis. That is, advertising of any kind will increase sales of the brand tuna 1 canned. The D2 phantom variable is 1 for both store view and newspaper ad, while the D1 phantom variable is 1 for store view only. Supermarket executives may be interested in knowing whether the newspaper's ad will increase sales more than just a single store offer. The results of the OLS estimate show that the Estimated D2 coefficient is higher than the d1 estimated. So this gives some support to the premise that it is useful to combine a newspaper ad with a store offer. However, to test this we can consider testing the hypothesis: H0: 4 = 5 statistical t-test calculated from the SHAZAM test command is -6.86. SHAZAM reports a value of 0.00000 k. This actually means less than 0.000005 and so the empty premise is rejected at any level of reasonable importance. We conclude that sales increase in the weeks when both forms of advertising are used. We can now ask the question: What is the size of the increase in sales when the store has a store offer and an ad sheet? The calculations show that weekly sales will increase by about 313%. In contrast, when only a store offer is used, weekly sales of brand 1 canned tuna will increase by about 52%. [SHAZAM HOME GUIDE] |_SAMPLE 1 52 |_READ (Tuna .txt) Sale Price1 PRICE2 PRICE3 D2 Unit 88 is now set to: Tuna.txt 6 variables and 52 notes starting from OBS 0 |_GENR Sales =LSALES =LOG =1** Estimate =_OLS PRICE1 PRICE2 PRICE2 D1 D2 / LOGLIN COEF =BETADERR =SE OLS 52 Variable Based Notes = L... Note.. Sample range range to: 1, 52 = .8428 R-SQUARE ADJUSTED = .8257 VARIANCE OF THE ESTIMATE-SIGMA**2 = .11538 STANDARD ERROR OF THE ESTIMATE-SIGMA = .33967 SUM OF SQUARED ERRORS-SSE= 5.3073 MEAN OF DEPENDENT VARIABLE = 8.4372 LOG OF THE LIKELIHOOD FUNCTION(IF DEPVAR LOG) = -453.182 VARIABLE ESTIMATED STANDARD T-RATIO PARTIAL STANDARDIZED ELASTICITY NAME COEFFICIENT ERROR 46 DF P-VALUE CORR. COEFFICIENT AT MEANS PRICE1 -3.7463 .5765 -6.498 .000 -.692 -.4514 -2.9315 PRICE2 1.1495 .4486 2.562 .014 .353 .1584 .9264 PRICE3 1.2880 .6053 2.128 .039 .299 .1268 1.0223 D1 .42374 .1052 4.028 .000 .511 .2612 .1874 D2 1.4313 .1562 9.165 .000 .804 .6720 .2477 CONSTANT 8.9848 .6464 13.90 .000 .899 .0000 8.9848 |_* Hypothesis testing |_TEST |_ TEST D1=0 |_ TEST D2=0 |_END F STATISTIC = 42.015301 WITH 2 AND 46 D.F. P-VALUE= .00000 WALD CHI-SQUARE STATISTIC = 84.030601 WITH 2 D.F. P-VALUE= .00000 UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = .02380 |_TEST D1=D2 TEST VALUE = -1.0075 STD. ERROR OF TEST VALUE .14692 T STATISTIC = -6.8577456 WITH 46 D.F. P-VALUE= .00000 F STATISTIC = 47.028674 WITH 1 AND 46 D.F. P-VALUE= .00000 WALD CHI-SQUARE STATISTIC = 47.028674 WITH 1 D.F. P-VALUE= .00000 UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = .02126 |_* Estimate the percentage effect of dummy variable D1 on SALES |_GEN1 C1=BETA:4 |_GEN1 SE1=SE:4 |_GEN1 G1= 100*(EXP(C1 - SE1*SE1/2) - 1) |_* Estimate the percentage effect of dummy variable D2 on SALES |_GEN1 C2=BETA:5 |_GEN1 SE2=SE:5 |_GEN1 G2=100*(EXP(C2 - SE2*SE2/2) - 1) |_PRINT G1 G2 G1 51.92391 G2 313.3233 |_STOP [SHAZAM Guide home] Posted on Friday, August 17th, 2018 at 8:00 pm. غالباً ما يوصى بتحويلات السجل jcf2d مكتوبة للبيانات المنحرفة، مثل التدابير النقدية أو بعض التدابير البيولوجية والديموغرافية. عادةً ما يكون لبيانات تحويل السجل تأثير نشر كتل البيانات والجمع بين البيانات المنتشرة. على سبيل المثال، يوجد أدناه مدرج بياني لمناطق جميع الولايات الأمريكية الخمسين. ومنحرف إلى اليمين بسبب ألاسكا وكاليفورنيا وتكساس وعدد قليل من الآخرين. بعد تحويل السجل، لاحظ أن الرسم البياني أكثر أو أقل متماثل. لقد نقلنا الولايات الكبيرة أقرب إلى بعضها البعض وتباعدنا خارج الولايات الأصغر. لماذا تفعل هذا؟ أحد الأسباب هو جعل البيانات أكثر طبيعية ، أو متماثل. إذا كنا نقوم بإجراء تحليل إحصائي يفترض الأوضاع الطبيعية، فإن تحويل السجل قد يساعدنا على الوفاء بهذا الافتراض. وهناك سبب آخر للمساعدة في تلبية افتراض التباين الثابت في سياق النمذجة الخطية. آخر هو للمساعدة في جعل علاقة غير خطية أكثر خطية. ولكن في حين أنه من السهل تنفيذ تحويل السجل ، فإنه يمكن أن يعقد التفسير. لنفترض أننا نناسب نموذج خطي مع متغير تابع محول باللوغاريتم. كيف نفسر المعاملات؟ ماذا لو كان لدينا متغيرات مستقلة ومستقلة تعتمد على اللوغاريتم؟ هذا هو موضوع هذه المقالة. أولاً سوف نقدم وصفة ل For those who just want some quick help. Then we'll dig a little deeper into what we say about our model when we record data conversion. Rules for interpretation OK, you run regression/fit linear model and some of your variables are converting the record. The child/response variable is only a log conversion. Base the coefficient, subtract 1 of this number, and multiply by 100 and this gives an increase in percentage (or decrease) in response to each single increase in a separate variable. Example: The plants is 0.198. (EXP (0.198) - 1) * 100 = 21.9. For each single unit an increase in an independent variable, our variable adopts increases by about 22%. Only the independent variable/forecast is to convert the record. Department of factories on 100. This tells us that a 1% increase in an independent variable increases (or reduces) the dependent variable by (coefficient /100) units. Example: The plants is 0.198. 0.198/100 = 0.00198. For each 1% increase in the independent variable, our variable increases by about 0.002. For an x per cent increase, multiply the coefficient in log (1.x). Example: For each 10% increase in the independent variable, the dependent variable increases by about 0.198* log (1.10) = 0.02. Both the child/response variable and the variable/independent/predictive variables are converted by the record. The coefficient is interpreted as an increase in the percentage in the dependent variable for each 1% increase in a separate variable. Example: The plants is 0.198. For each 1% increase in the independent variable, our variable increases by about 0.20%. For an x percent increase, calculate 1.x to the strength of the coefficient, subtract 1, and multiply by 100. Example: For each 20% increase in the independent variable, our variable increases by about (1.20 0.198-1) * 100 = 3.7%. What log conversions really are for your forms it's good to know how to correctly interpret transactions for the data that the record has been converted, but it's important to know exactly what your form means when the data that the record has been converted includes. For a better understanding, let's use R to simulate some data that requires log conversions for correct analysis. We will keep it simple with one independent variable and usually distributed errors. First we will look at a transducer variable from the log. X<- seq (0.1,5,length.out = 100) set.seed (1) e<- rnorm (100, means = 0, sd = 0.2) The first line generates a sequence of 100 values from 0.1 to 5 and assigns it to x. The next line assigns the random number generator to 1. If you do the same, you'll get the same randomly generated data that we got when you run the next line. The rnorm code (100, i.e. 0, SD = 0.2) creates 100 values of normal distribution with an average of 0 and a standard deviation of 0.2. That would be our fault. This is one of the افتراضات من الانحدار الخطي البسيط: يمكن أن تكون على غرار y .ثم نضيف الخطأ العشوائي، ه أخيراً نُسَسّ x، تحويل التابعة. نختار تقاطع (1.2) و ميل (0.2)، الذي نضربه بـ-log الآن نحن على استعداد لإنشاء متغير لدينا e. بياناتنا مع خط مستقيم ولكن سيكون قبالة من قبل بعض المبلغ العشوائي الذي نفترض أنه يأتي من توزيع عادي مع متوسط 0 وبعض الانحراف المعياري. نحن نخصص الخطأ لدينا إلى <- exp(1.2= += 0.2= *= x= += e)= to= see= why= we= exponentiate,= notice= the= following:= $$\text{log}(y)=\beta_0 += \beta_1x$$= $$\text{exp}(\text{log}(y))=\text{exp}(\beta_0 += \beta_1x)$$= $$y=\text{exp}(\beta_0 += \beta_1x)$$= so= a= log-transformed= dependent= variable= implies= our= simple= linear= model= has= been= exponentiated.= recall= from= the= product= rule= of= exponents= that= we= can= re-write= the= last= line= above= as= $$y=\text{exp}(\beta_0) \text{exp}(\beta_1x)$$= this= further= implies= that= our= independent= variable= has= a= multiplicative= relationship= with= our= dependent= variable= instead= of= the= usual= additive= relationship.= hence= the= need= to= express= the= effect= of= a= one-unit= change= in= x= on= y= as= a= percent.= if= we= fit= the= correct= model= to= the= data ,= notice= we= do= a= pretty= good= job= of= recovering= the= true= parameter= values= that= we= used= to= generate= the= data.= lm1=></-> <- lm(log(y)= ~= x)= summary(lm1)= call:= lm(formula=log(y) ~= x)= residuals:= min= 1q= median= 3q= max= -0.4680= -0.1212= 0.0031= 0.1170= 0.4595= coefficients:= estimate= std.= error= t= value= pr(=>|t|) (33.20 0.03693 1.22643 (اعتراض <2e-16 ***= x= 0.19818= 0.01264= 15.68=></2e-16> <2e-16 ***= ---= signif.= codes:= 0= '***'= 0.001= '**'= 0.01= '*'= 0.05= '.' = 0.1= '= ' = 1= residual= standard= error:= 0.1805= on= 98= degrees= of= freedom= multiple= r-squared:= 0.7151,= adjusted= r-squared:= 0.7122= f-statistic:= 246= on= 1= and= 98= df,= p-value:=></2e-16> < 2.2e-16= the= estimated= intercept= of= 1.226= is= close= to= the= true= value= of= 1.2.= the= estimated= slope= of= 0.198= is= very= close= to= the= true= value= of= 0.2.= finally= the= estimated= residual= standard= error= of= 0.1805= is= not= too= far= from= the= true= value= of= 0.2.= recall= that= to= interpret= the= slope= value= we= need= to= exponentiate= it.= exp(coef(lm1)[x])= x= 1.219179= this= says= every= one-unit= increase= in= x= is= multiplied= by= about= 1.22.= or= in= other= words,= for= every= one-unit= increase= in= x,= y= increases= by= about= 22%.= to= get= 22% ,= subtract= 1= and= multiply= by= 100.= (exp(coef(lm1)[x])= -= 1)= *= 100= x= 21.91786= what= if= we= fit= just= y= instead= of= log(y)?= how= might= we= figure= out= that= we= should= consider= a= log= transformation?= just= looking= at= the= coefficients= isn't= going= to= tell= you= much.= lm2=> <- lm(y= ~= x)= summary(lm2)= call:= lm(formula=y ~= x)= min= 1q= median= 3q= max= -2.3868= -0.6886= -0.1060= 0.5298= 3.3383= coefficients:= estimate= std.= error= t= value= pr(=>|t|) ((اعتراض 12.73 0.23643 3.00947 <2e-16 ***= x= 1.16277= 0.08089= 14.38=></2e-16> <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard ***= ---= signif.= codes:= 0= '***'= 0.001= '**'= 0.01= '*'= 0.05= '.' = 0.1= '= ' = 1= residual= standard=></2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard > </-> </-> </-> 1.156 on 98 degrees of multiple freedom R-square: 0.6783, modified R-square: 0.675 F-Statistical: 206.6 on 1 and 98 DF, p-value: < 2.2e-16 surely, since we generate data, we can see transactions are far from the remaining error very high. But in real life you won't know this! That's why we diagnose regression. A key assumption of verification is the constant variation of errors. We can do this with the site scale plot. Here's a plot for the model we just ran without converting the record y. plot (lm2, which = 3) #3 = scale site plot notice unified tailings trending upwards. This is a sign that the fixed contrast assumption has been violated. Compare this plot to the same plot for the correct model. The trend line is up and the residue is uniformly scattered. Does this mean that you should always record your child variable conversion if you suspect the fixed variance assumption has been violated? The contrast may not necessarily be static due to other errors in your form. Also think about what the approved variable modeling means in the record. She says she has a double relationship with the donors. Does that sound right? Use your judgment and experience. Now let's consider the data with an independently anticipated variable converted record. This is easier to generate. We simply record the conversion x. - 1.2 + 0.2 * log (x) + e again we first fit the correct model and note that it does a great job of recovering the real values we used to generate data: lm3 <- lm (y ~record (x)) summary (lm) 3) call: lm (formula = y ~ log () remaining x)) Min 1Q Average 3Q Max -0.46492 -0.12063 0.00112 0.11661 0.45864 Transactions: Std. Estimate. |) (Intercept) 1.22192 0.02308 52.938 < 2e-16*** Record (x) 0.19979 0.02119 9.427 2.12e-15*** signature ---if. Codes: 0'***' 0.001 '**' 0.01 '*' 0.05'. 0.1' 1 Remaining standard error: 0.1806 on 98 degrees of multiple freedom R-square: 0.4756, R-square adjusted: 0.4702 F-statistical: 88.87 on 1 and 98 cannons, p-value: 2.121e-15 to interpret slope coefficient divided by 100. Coef (lm3) [log (x)/100 log (x) 0.001997892 This tells us that a 1% increase in x increases the dependent variable by about 0.002. Why does he tell us that? Let's do some of the calculations below we calculate the change in y when you change x from 1 to 1.01 (i.e., an increase of 1%). $beta_0 =$1 beta_1 beta_1.beta_1\text{log}1.01 beta_1 beta_1 beta_0) beta_1 {1} = \beta_1\text{log}1.01$ result multiplied the slope coefficient in a record (1.01), which is roughly equal to 0.01, or \frac{1}{100}\ Hence the explanation is that a 1% increase in x increases the x variable by a factor of 100. Again let's fit the error By failing to determine the conversion of a record for x in the syntax model. A summary of the model will reveal that transaction estimates are far from real values. But in practice we never know the true values. Again, the diagnoses are in order to assess the adequacy of the model. A useful diagnosis in this case is a partial remaining plot that can reveal a departure from sin. Mention that linear models assume that the probies are an addition and have a linear relationship with the response variable. The car package provides the function of crPlot to create partially remaining plots of land quickly. Just give it a model object and determine the variable that you want to create a partial remaining plot for. Library (car) crPlot (lm4, variable = x) the straight line represents the specific relationship between x and y. The curved line is a smooth trend line that summarizes the observed relationship between x and y. We can say that the observed relationship is not linear. Compare this plot to the remaining partial piece of the correct model. crPlot (lm3, variable = log (x)) smooth lines and their installation are right on top of each other reveal ing no serious exit from linear. This does not mean that if you see a departure from sin you should immediately assume the registry conversion is one and only repair! A non-linear relationship can be complex and cannot be easily interpreted with a simple shift. But converting a record may be appropriate in such cases and certainly something to consider. Finally let's look at the data where both the child and independent variables are the converted record. R< - exp (1.2 + 0.2*log (x) + e) look closely at the code above. The relationship between x and y is now both multiplied and nonlinear! As usual we can fit the correct model and note that it does a great job of recovering the real values that we used to generate data: lm5 <- lm (log)~log (x)) summary (lm5) call: lm (formula = log (r) ~ log () Remaining: Min 1Q Average 3Q Max -0.46492 -0.12063 0.00112 0.1166 0.45864 Transactions: Std Estimate. Error value t Pr (>t|) (intercept) 1.22192 0.02308 52.938 < 2e-16*** log (x) 0.19979 0.02119 9.427 2.12e-15*** --- Signif. Codes: 0'***' 0.001 '**' 0.01 '*' 0.05'. 0.1' 1 Remaining standard error: 0.1806 on 98 degrees of multiple Freedom R-squared: 0.4756, R-squared rate: 0.4702 F-Statistical: 88.87 on 1 and 98 guns, p-value: 2.121e-15 X coefficient interpretation as an increase in y percentage per 1% increase in x. In this case, this is a 0.2% increase in y per 1% increase in x. The installation of the wrong model again produces a coefficient and remnants of standard error estimates that are violently out of target. LM6 < - lm (r~o) summary (lm6) in a site -partial site and pieces of remaining plots provide evidence that something is wrong with our model. View site size chart The direction line and the remaining partial plot shows the linear and smooth lines that fail in the game. crPlot (lm6, variable = x) How do we know in real life that the correct model requires independent variables that have been converted from the record and reliability? We won't do that we may have a hunch based on diagnostic plots and modeling experience. Or we may have some experience in the subject on the modeling process we are and have good reason to believe that the relationship is multiplied and non-linear. Hopefully you now have the best handling of not only how to interpret the variables that have turned the record into a linear model but also what the variables that the record has been converted to your model. For questions or clarifications about this article, contact UVA StatLab: statlab@virginia.edu view the full range of UVA StatLab library articles. Clay Ford Statistical Research Consultant University of Virginia Library August 17, 2018 2018