*Module 4, Week 1, Paper and Pencil Assignment 7*

This assignment focuses on Example 5.3 in your textbook ***A Second Course in Statistics : Regression Analysis***. For your convenience I’ll transcribe the data in Table 5.2 into a .csv file so you can use it in any code you use for analytics. I strongly encourage you to modify the gretl script from the Week 7 gretl assignment. I have written directions to help you do that. Although the textbook includes a lot of analysis/solutions for this example, let’s walk through this to see exactly what is going on. I’ll leave the problem statement as is in the textbook.

Keep in mind that the Week 7 gretl assignment does not cover second-order polynomial models. However, these are quite easy to do in gretl. Another good reason to reuse your Week 7 gretl script is that you might want to incorporate the code included in this P&P assignment to get everything in one place – before the Final Examination!

Also, to test assumptions you need to **change** the modtest –logs to modtest --square to test the model with second-order terms. We will not examine the specific outcome of these tests of the assumptions in this assignment. But, if you look at them the results do not look good…

Example 5.3 (page 288)

First using the data in the example5.3.csv data file, develop an OLS model for this example. The questions that are on Blackboard are preceded by a bolded, blue designation, e.g. **BBQ1**.

Notes and steps to run gretl for the Week 7 P&P Assignment:

1. You will need to go through your gretl model and change the variables from the original “sales,” “price,” and “adv\_cost,” to “y,” “x1,” “x2,” etc.
2. You’ll want to comment out things that you might not need immediately so they do not hamper running the script until you get everything setup as you want it. For example, I commented out the part to compute the sales given price=5.50 and adv\_cost=1200.
3. I noticed that when I was working on the gretl script for Week 7 I must have done some cut/paste and the open statement for the data file is not in a good place. You should move that as soon as possible so you don’t get too many errors.
4. You are given the model is a second-order model. So, you will have to add the required terms to the gretl model. That is,

series x1sq = x1^2

series x2sq = x2^2

series x1x2 = x1\*x2

1. You’ll need to make sure your command to build the OLS model includes all the relevant terms. I haven’t checked to see if the order of the terms makes a difference to gretl or not. So, I’ve just put them in the same order as they appear in Example 5.3 in the textbook.
2. Check your model before proceeding. I ran the script and obtained an second-order OLS model but the values were slightly different than the textbook reports from using SAS. SAS is another common code used to conduct statistical analyses. The URL for SAS is: <https://www.sas.com/en_us/home.html>. As I/we have “discussed” before, it is not unusual to get slight differences between codes due to the way codes build models, rounding, etc. However, if you get a significantly different values than are reported in the textbook something is definitely wrong. If you need help just ask!
3. **BBQ1:** NONE of the variables in the second-order model are statistically significant.
   1. True
   2. False
4. **BBQ2:** The R-squared value for the second-order OLS model is \_\_\_\_\_\_\_\_\_.

Your textbook has a figure illustrating he surface created by this model. I haven’t spent a lot of time on it, but I don’t see a way to produce this graph in gretl. Again, if anyone knows then please let me know! I am sure we can produce it in other codes besides SAS and Minitab, such as R/RStudio and Octave.

Now, let’s play with this model a little from the “trial and error” approach I mentioned before in the gretl assignment, i.e. let’s go through the variables and see how they affect how well the model represents the data. First, let’s go back and just build a multivariable linear model. (Hint: you should immediately know what I mean. If you don’t you should go back and review the differences between simple, multivariable, and multivariate linear models.)

1. **BBQ3:** The R-squared value for the simple linear OLS model is \_\_\_\_\_\_\_\_.
2. **BBQ4**: Other than the intercept, which term or terms is/are statistically significant in the simple linear OLS model?
   1. x1
   2. x2
   3. Neither x1 or x2
   4. Both x1 and x2

Now, let’s add and subtract terms to the model using a brute force method (all possible permutations) to combine terms.

Let’s try adding non-linear or second-order terms for the variables one at a time. First, let’s add the interaction term, i.e. x1\*x2. Then, let’s add an x1-squared term without changing x2. Last, let’s add an x2-squared term without changing x1. That is,

Remember we’ve already built the model with all terms, i.e.

If you setup a separate line to generate all these models and rename them, e.g. ml for the first model with all terms, ml2 for the simple linear model, ml3 for the model with the interaction term and so on. For example,

and so on. Then, you will have all these available throughout your analysis by call on ml, ml2, ml3, or whatever model you want without having to regenerate it.

Examine the models that are output and enter the results including the R-squared values for each model in the following questions.

1. **BBQ5**: The value of R-squared for the OLS model containing an interaction (x1\*x2) term is \_\_\_\_\_\_\_\_\_.
2. **BBQ6**: Which term or terms is/are statistically significant in this model? Check all that apply.
   1. The intercept
   2. x1
   3. x2
   4. x1x2
   5. None of the terms are statistically significant
3. **BBQ7**: The value of R-squared for the OLS model containing a x1-squared term is \_\_\_\_\_\_\_\_\_.
4. **BBQ8**: Which term or terms is/are statistically significant in the simple linear OLS model? Check all that apply.
   1. The intercept
   2. x1
   3. x2
   4. x1sq
   5. None of the terms are statistically significant
5. **BBQ9**: The value of R-squared for the OLS model containing a x2-squared term is \_\_\_\_\_\_\_\_\_.
6. **BBQ10**: Which term or terms is/are statistically significant in ths model? Check all that apply.
   1. The intercept
   2. x1
   3. x2
   4. x2sq
   5. None of the terms are statistically significant
7. **BBQ11**: The value of R-squared for the OLS model containing interaction as well as the x1-squared and x2-squared terms is \_\_\_\_\_\_\_\_\_.
8. **BBQ12**: Which term or terms is/are statistically significant in this model? Check all that apply.
   1. The intercept
   2. x1
   3. x2
   4. x1x2
   5. x1sq
   6. None of the terms are statistically significant
9. **BBQ13**: The value of R-squared for the OLS model containing interaction and x2-squared terms is \_\_\_\_\_\_\_\_\_.
10. **BBQ14**: Which term or terms is/are statistically significant in the simple linear OLS model? Check all that apply.
    1. The intercept
    2. x1
    3. x2
    4. x1x2
    5. x2sq
    6. None of the terms are statistically significant

If all variables appear to be statistically significant, you can see how tedious this would be if this were the only way to determine which should really be kept in a model. Based on these results my “guess” would be that the x2 pressure term is more important than the x1 temperature term. However, without additional knowledge about what is being produced as well as the chemicals involved it is really impossible to say more.

This is why it is important in these situations to get subject matter experts (SME’s) involved as early as possible, and to use more sophisticated analysis techniques such as the design of experiments to conduct analyses.

However, this has given you some experience in building higher-order, non-linear models.

1. **BBQ15**: Returning to the initial model including all terms (interaction and second-order terms), what is the expected quality of the product if the temperature is 85 degrees F and the pressure is 57 psi? (Hint: what do you do with the interaction and second-order terms now? Since x1=85 and x2=57, then x1x2=4845. Similarly, x1sq=7225. I’ll leave it to you to calculate x2sq. Once you have these values you predict the point estimate the same way we did in the gretl exercise when we predicted the sales for price=$5.50 and adv\_cost=$1200, without the conversion for units to $1,000 USD.)
2. **BBQ16**: Based on the data available and model developed this is a reasonable value for product quality.
   1. True
   2. False
3. **BBQ17**: Now compute the expected value when the pressure is increased to 60 psi. The computed product quality is \_\_\_\_\_\_\_\_\_.
4. **BBQ18**: How does the product quality change if the pressure is increased to 60 psi?
   1. The quality increases.
   2. The quality remains the same.
   3. The quality decreases.
   4. There is no change in the quality.
5. **BBQ19**: Based on the data available and model developed this is a reasonable outcome.
   1. True
   2. False
6. **BBQ20**: Consider the F-value and related P-value for testing the entire model. Does the enter model with interaction and second-order terms appear to be a good model for this data?
   1. Yes
   2. No
7. **BBQ21**: One way to determine whether or not all the terms (including the interaction and second-order terms) should be retained in the model is to compare the “complete” model with a reduced model. The easiest way to complete this comparison is to open the model window for Model ml and record F(5,21) and its associated P-value. Then open the model window for the Model ml2, the multivariable OLS model, and record F(2,24) and its associated P-value. **Based on your comparison of the complete and reduced models, should all terms in the complete model be retained?**  (Hint: see section 4.13 and Example 4.11 in your textbook.)
   1. Yes
   2. No

Just to be thorough I’m going to paste the comparable model computed using R/RStudio below. We’ll continue the rest of Chapter 5 in your textbook, A Second Course in Statistics: Regression Analysis, next week in preparation for the Final Examination. If you have any questions – please ask now to give me time to get back to you!

>

> x1x2 <- x1\*x2

> x1sq <- x1^2;

> x2sq <- x2^2;

> ml <- lm(y~x1+x2+x1x2+x1sq+x2sq, data = example5.3)

> summary(ml)

Call:

lm(formula = y ~ x1 + x2 + x1x2 + x1sq + x2sq, data = example5.3)

Residuals:

Min 1Q Median 3Q Max

-2.8370 -0.9648 0.0963 1.0074 2.9407

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.133e+03 1.098e+02 -46.73 < 2e-16 \*\*\*

x1 3.115e+01 1.339e+00 23.27 < 2e-16 \*\*\*

x2 1.399e+02 3.127e+00 44.72 < 2e-16 \*\*\*

x1x2 -1.460e-01 9.652e-03 -15.13 9.16e-13 \*\*\*

x1sq -1.336e-01 6.825e-03 -19.57 5.80e-15 \*\*\*

x2sq -1.145e+00 2.730e-02 -41.94 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.672 on 21 degrees of freedom

Multiple R-squared: 0.9931, Adjusted R-squared: 0.9914

F-statistic: 602.4 on 5 and 21 DF, p-value: < 2.2e-16