

Lab 11: Multicollinearity

Objectives:

In today's lab we will investigate the existence of a problem in the sample data - **Multicollinearity**. Multicollinearity is the existence of linear association among two or more independent variables. The mere existence of multicollinearity does not mean you have a problem; you only have a problem when the degree of multicollinearity is high. The exercises below illustrate diagnosing the problem, and applying a couple of possible remedies.

Data: The data for today's lab are those that we used last week, the Roses sales data, and the data in the Minitab worksheet *Chickendata.mtw*. You should have a Minitab project file for the Roses data.

◆ **Multicollinearity – diagnosis of the problem in the roses data.**

1. Check for multicollinearity in the model we estimated for the sales of roses. These are time series data collected quarterly. We always suspect times-series data of multicollinearity because common inflationary trends are often part of price and income data. Thus, the variables *prose*, *pcarn* and *dinc* may be highly correlated. In addition, we created the variables *D2* and *ProseD2*. These may also have strong correlation. **Open your Minitab roses project from last week and let's check for multicollinearity problems.**

2. **Estimate the following model:**

$$Sales_t = \beta_0 + \beta_1 Prose + \beta_2 Pearn + \beta_3 Disinc + \delta D2 + \gamma ProseD2 + u;$$

and ask Minitab to calculate the variance inflation factors (VIFs). You'll find a check box for **Variance Inflation Factors** in the **Options** menu of **Regression**.

3. First, review your regression results. Look for the "tell-tale signs" of multicollinearity.
 - What is the R^2 for the model? Is it high suggesting a good model?
 - What is the calculated F-statistic for the model? Is it large suggesting statistical significance?
 - These two measures indicate how well the independent variables explained the dependent variable. A high R^2 and a large F-statistic indicate the model did well explaining the dependent variable.
 - Review the calculated t-statistics for each parameter estimate. How many appear statistically significant? (Recall that we did an F-test in class and found the two variables *D2* and *ProseD2* jointly explained a significant portion of the variation in sales. But individually they are not statistically important. This contradiction suggests these data may have a problem with multicollinearity.)
4. Have Minitab calculate the correlation matrix for the independent variables: *Prose*, *Pearn*, *Disinc*, *D2* and *ProseD2*. (*Correlation* is found under **Stat** and **Basic Statistics**.) If any correlation coefficient is greater than 0.80 in absolute value, we worry about multicollinearity. Are there any "culprits?" Which independent variables are most highly correlated?
5. Next review the VIFs given by Minitab. The rule of thumb is that $VIF > 10$ spells problems. What do you conclude about the independent variables? What do these results suggest for the lack of significance for the dummy variable and interaction term?
6. It is possible that using a natural logarithm transformation will mitigate some of the multicollinearity. Anyway, economists love the log-log model (why?). Estimate the following model and include the VIFs:
$$\ln Sales_t = \beta_0 + \beta_1 \ln Prose + \beta_2 \ln Pearn + \beta_3 \ln Disinc + \delta D2 + \gamma ProseD2 + u$$
7. Were there any improvements?

◆ **Multicollinearity – Chicken Demand data.**

1. Follow the same procedure using the chicken demand data in the file *Chickendata.mtw*. Estimate the following model and again ask Minitab to provide the VIFs:

$$qchik_t = \beta_0 + \beta_1 pchik_t + \beta_2 ppork_t + \beta_3 pbeef_t + \beta_4 dinc_t + u_t$$

2. Review the regression results. Look for the “tell-tale signs” of multicollinearity.
 - What is the R^2 for the model? Is it high?
 - What is the calculated F-statistic for the model? Is it large suggesting statistical significance?
 - Review the calculated t-statistics for each parameter estimate. How many appear statistically significant?
 - What are the VIFs? Are they in the “problem range?”
3. **Calculate correlation coefficients** for the independent variables in model (1): *qchik*, *pchik*, *ppork*, *pbeef* and *dinc*. Is there a potential problem here? How did you decide?
4. **Auxiliary Regressions:** Multicollinearity may be a relationship between two or more independent variables. To check, you can estimate **auxiliary regressions**, regressions of the independent variables on each other. The independent variables are related if **the R^2 is high** for these regressions. Try one. Regress the independent variable *pchik* on *ppork*, *pbeef*, and *dinc*.

◆ **Multicollinearity – Correcting the Problem**

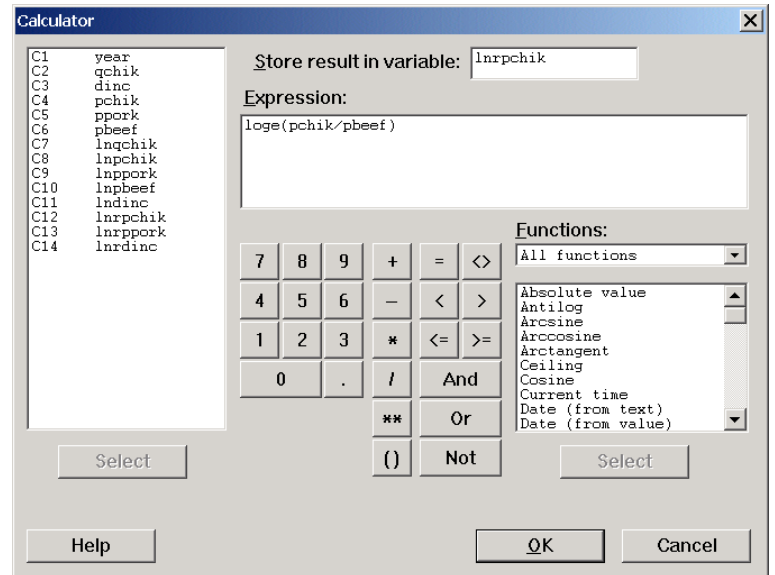
1. One possible solution to multicollinearity is to estimate a “non-linear” model. Our log-log model, model fits in this category:

$$\ln qchik_t = \beta_0 + \beta_1 \ln pchik_t + \beta_2 \ln ppork_t + \beta_3 \ln pbeef_t + \beta_4 \ln dinc_t + u_t$$

2. Quickly check to see if transforming the independent variables into natural logarithms help cure the problem. If you haven’t already done so, include the **Variance Inflation Factors** in your regression results for this model. How are they? Compare them to the linear demand function presented in class. Any improvements?
3. Check the correlation coefficients for the variables: *lnpchik*, *lnppork*, *lnpbeef* and *lndinc*. What do you conclude; is there “an offensive variable?”
4. With price and income data there is often a lot of multicollinearity as you can see here. One way to break up multicollinearity is to create “relative price and income variables.” Compute the following set of relative prices and income:

$$rpchik = \frac{pchik}{pbeef}; \quad rppork = \frac{ppork}{pbeef}; \quad \text{and} \quad rdinc = \frac{dinc}{pbeef}.$$

Create these new variables, in **log form**, using **Calc** and **Calculator**. I've included a screen capture at the right to illustrate the creation of the natural log of the relative price of chicken. Use the natural log transformation in Minitab. (Note: it doesn't make sense to calculate the "relative price of beef – all the values will be "1."



5. Are these new variables still plagued by multicollinearity? Check the correlation coefficients for *lnrpchik*, *lnrppork*, and *lnrdinc*.

6. Estimate the following regression model (include the VIFs to further convince yourself that relative prices are good):

$$\ln qchik_t = \beta_0 + \beta_1 \ln rpchik_t + \beta_2 \ln rppork_t + \beta_4 \ln rdinc_t + u_t$$

The parameter estimates are still interpreted as elasticities. For example, $\hat{\beta}_1$ is the elasticity of demand for chicken with respect to the price of chicken.

7. Sadly, it appears that we lost one of our price elasticities. Not really. There is microeconomic theory result that says the sum of all the price and income demand elasticities must be zero:

$$\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 = 0; \text{ or } \hat{\beta}_3 = -\hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_4.$$

Using this equation, solve for the missing estimate, the elasticity of chicken demand with respect to the price of beef.