Empirical Research Guide

Noah Lyman
University of North Carolina at Chapel Hill

April 11, 2023

1 Organization

Many ingredients are required for the execution of an empirical project. Without careful organization, it is easy to get lost within the large set of excel files, Stata files, and scripts of code that will inevitably pile up. Before discussing our empirical implementation, we should think about how we want to organize our materials. For a given project, it is good practice to keep a folder on your computer which will house all relevant materials.

1. On your desktop, create a folder to keep all project files in. I will name my folder "Project" and will refer to it throughout the document.

2. Create two subfolders, one called "Data" and one called "Code."

3. Within the "Data" folder, create two additional subfolders. One called "Raw" and another called "Clean."

2 Simulation Exercise

Before working with any actual data, we will simulate some "fake" data which we can use to illustrate how the econometric machinery works. Our application will be modeling the price of chicken. I will conduct the simulation in Stata. Follow along using the file "Simulation.do" for more details. For those who prefer to use Excel, I provide a similar walk through in the file "Simulation.xlsx." For your own empirical projects, I **strongly** recommend not using Excel. Excel can be useful for simple operations, but it quickly becomes inefficient when using more sophisticated empirical techniques. As you may see, implementing the techniques we'll discuss throughout the document can be fairly convoluted in Excel. Stata is built to conduct sophisticated empirical analyses using just a few key strokes.

2.1 A Model of Chicken Prices

Let $sim\_chicken_t$ be the (simulated) price of chicken at time t. Let's first assume that $sim\_chicken_t$ is given by:

$$sim\_chicken_t = \alpha_0 + \alpha_1\, sim\_corn_t + \alpha_2\, sim\_soy_t + \epsilon_t \tag{1}$$

where sim_corn_t and sim_soy_t are the prices of corn and soybeans at time t, respectively.[1] Here I include the $sim\_$ prefixes on all variables to emphasize that these are values which we are simulating. Equation (1) has several components:

1. Parameters: $\alpha_0, \alpha_1, \alpha_2$
2. Independent variables: sim_corn_t, sim_soy_t
3. Error term: ϵ_t

Let's quickly talk through each of these components piece by piece.

Parameters The parameters of Equation (1) describe how the dependent variable, $sim\_chicken_t$ in this case, depends on the independent variables sim_corn_t and sim_soy_t. Almost always, econometric models of this flavor will include an intercept term α_0 which gives the average price of chicken if the sim_corn_t and sim_soy_t were to be equal to zero. Concretely, we have:

$$\mathbb{E}[sim\_chicken_t \,|\, sim\_corn_t = 0, sim\_soy_t = 0\,] = \alpha_0 \tag{3}$$

α_1 is the marginal or *partial* effect of sim_corn_t on the price of chicken. Similarly, α_2 is the partial effect of sim_soy_t on the price of chicken. To understand what this means exactly, first note that given Equation (1):

$$\frac{\partial sim\_chicken_t}{\partial sim\_corn_t} = \alpha_1 \quad \text{and} \quad \frac{\partial sim\_chicken_t}{\partial sim\_soy_t} = \alpha_2 \tag{4}$$

[1] For the purposes of the simulation exercise, I'll make the following distributional assumptions:

$$log(sim\_corn_t) \sim N(0,1) \quad \& \quad log(sim\_soy_t) \sim N(0,1) \tag{2}$$

In words, if the price of corn were to increase by \$1, $sim\_chicken_t$ would increase by $\$\alpha_1$ in response. Similarly, if the price of soybeans were to increase by \$1, $sim\_chicken_t$ would increase by $\$\alpha_2$. Throughout this simulation exercise, let's assume:

$$\alpha_0 = 1$$
$$\alpha_1 = .8$$
$$\alpha_2 = .5$$

Independent Variables Put simply, the set of independent variables included in Equation (1) are the set of things which influence the price of chicken and our available in our data.

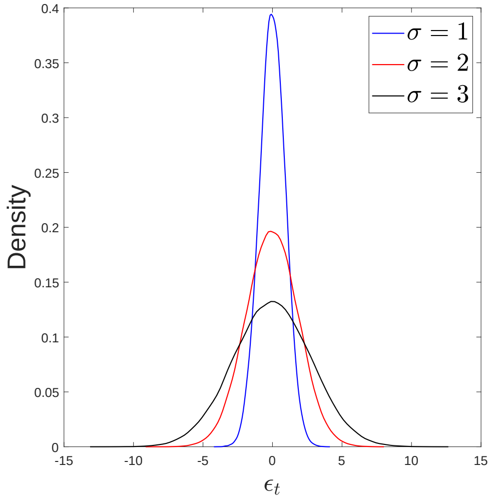
Error Term The price of chicken is unlikely to depend deterministically on corn and soybean prices. There are almost surely some other factors which affects the price of chicken, but we may not observe them in the data. For example: a disease spreads among chickens, decreasing chicken supply and increasing price. Or, a documentary comes out about the environmental impact of factory farming, decreasing demand for chicken and thus its price. We can come up with many examples of things which may impact the price of chicken but are not present in our data. All of these things are captured by this random ϵ_t term. For simplicity, we'll assume:

$$\epsilon_t \sim N(0, \sigma) \tag{5}$$

In English, this means that ϵ_t is normally distributed with mean 0 and standard deviation σ. As long as we include an intercept term, here denoted by α_0, the distribution of ϵ_t will have a mean of zero. This is very convenient, but not something we will discuss in detail here. Drawing many ϵ_t's from this distribution and plotting the resulting probability density will yield the familiar bell curve shape:

Figure 1 shows the distribution of ϵ_t for 3 different choices of the standard deviation σ. Notice that all 3 distributions are centered around 0 (their mean). Varying σ thus has no impact on the mean of σ. What does change is the concentration of the distribution. When σ is very small, most of the realizations of ϵ_t will be fairly close to the mean, hence the high density around 0 for the $\sigma = 1$ case. As σ increases, extreme values of ϵ_t become more likely. As a result, the probability density around the mean is reallocated to values in the tails. The distribution of ϵ_t thus has a larger spread as σ increases. This has important implications for estimation, which we will discuss shortly.

Figure 1: Probability Density Function of ϵ_t



2.2 Simulating the Model

The parametric assumption given by Equation (1) along with the distributional assumption (5) comprise a model for the price of chicken. We can take this assumed model, along with our assumed values of the α's, to simulate some data. I will provide a high-level summary of the simulation procedure in this section of the document. For the initial simulation, I'll set the standard deviation of ϵ_t to 1, so $\sigma = 1$. I'll also denote the size of the sample by T, and start by setting $T = 1000$. The required steps are as follows:

1. Using distributional assumption (2), generate T corn and soybean prices.

2. Using distributional assumption (5) with $\sigma = 1$, draw T ϵ_t shocks.

3. Generate T prices of chicken by simply applying Equation (1) along with the assumed values of the α parameters:

$$chicken_t = 1 + .8 \times corn_t + .5 \times soy_t + \epsilon_t \tag{6}$$

The result will be a simulated data set which looks something like:

The t column lists the time period corresponding to each row. Row 1 corresponds to the first period while row T corresponds to the last period in our sample. We refer to each row as an *observation*. When simulating this sample, I set $T = 1000$, so our data has 1,000 observations. The Chicken, Corn, and Soybeans columns list the realizations of chicken, corn, and soybean prices in each time period. Despite it being "fake," simulated

Table 1: Sample of Simulated Data

t	**Chicken**	**Corn**	**Soybeans**
1	3.13	1.76	2.32
2	.236	.487	.306
3	2.32	.066	.650
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$T-2$	5.72	1.31	4.86
$T-1$	5.45	2.41	5.77
T	3.06	.400	3.85

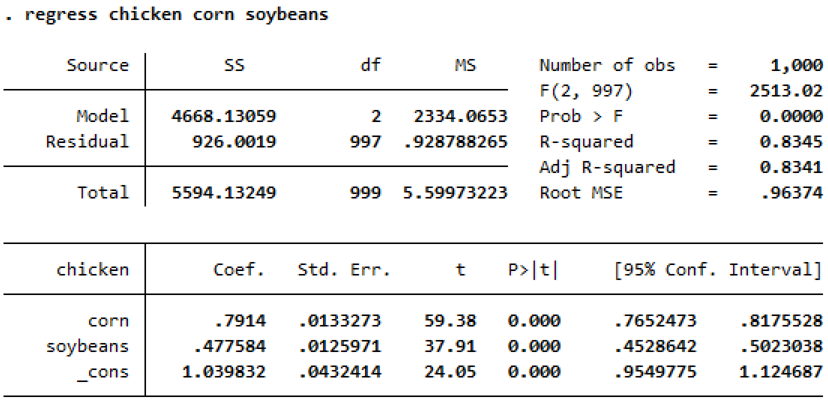
data can serve a multitude of productive purposes. Here, we will use this simulated data to illustrate how to employ a variety of econometric techniques.

2.3 Recovering Model Parameters

At this point, let's completely forget that we assumed a model for the price of chicken and subsequently simulated a fake data set. Instead, let's pretend that we downloaded the data in Table 1 directly from the BLS website. Furthermore, a consulting firm has hired us to use this data to determine the impact of corn and soybean prices on the price of chicken. As we will discuss later, such an exercise can be useful if we want to forecast future prices of chicken.

There is some "true" model of chicken prices which we don't know. What we want to do is use our data to approximate this unknown function as best as we can. There are many ways to do this, but we'll focus on the method of ordinary least squares (OLS for short). I'll give a quick sketch of what this approach entails. First, we recognize that there is some *true* process which generates the price of chicken. This true process is represented by Equation (1) and distributional assumption (5). We are pretending now that we do not know this true model, but wish to approximate it with the data we have at our disposal. While we do not know the true model, we can make a conjecture that chicken prices follow Equation (1). Our conjectured model is often referred to as our *regression* equation. Then using our data, we can try our best to estimate the model parameters α_0, α_1, and α_2. We use $\hat{\alpha}_k$ to denote the estimate of parameter α_k for $k \in \{0,1,2\}$. Is there a best possible estimate of α_k that we can attain? The answer is yes, and conveniently, OLS yields these best possible estimates under a few technical assumptions. The exact formula for the OLS estimator can be quite intimidating. Luckily, we have computers to do the computations for us. I'll focus on Stata as our software of choice. To run a regression in Stata, simply type in the command window:

Figure 2: Regression Output in Stata



```
. regress chicken corn soybeans

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =   2513.02
       Model |  4668.13059         2  2334.0653   Prob > F        =    0.0000
    Residual |    926.0019       997  .928788265   R-squared       =    0.8345
-------------+----------------------------------   Adj R-squared   =    0.8341
       Total |  5594.13249       999  5.59973223   Root MSE        =    .96374

------------------------------------------------------------------------------
     chicken |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        corn |      .7914   .0133273    59.38   0.000     .7652473    .8175528
    soybeans |    .477584   .0125971    37.91   0.000     .4528642    .5023038
       _cons |   1.039832   .0432414    24.05   0.000     .9549775    1.124687
------------------------------------------------------------------------------
```

regress chicken corn soybeans

After you type "regress," Stata will always assume that the first variable you list is the dependent variable, while all subsequent variables are the independent variables. After typing the above statement in your command window, press enter and a table will appear. I provide an image of my version of the table in Figure 2.

Let's highlight a few things. In the bottom half of the table, the dependent variable (chicken) is listed in the top left. The independent variables (corn, soybeans) are listed underneath, along with a term called "_cons". The second column (Coef.) lists the estimated coefficients $\hat{\alpha}_k$ of the model. The coefficient next to corn is our estimate of α_1, while the coefficient next to soybeans is our estimate of α_2. The coefficient next to _cons is our estimate of the intercept parameter α_0. All of these are very close to the values we assumed when we simulated the data, so it seems like this regression has done a good job estimating the parameters of our model. Let's dig a little deeper. The third column (Std. Err.) reports the *standard errors* of each of our estimates, which is a measure of how precise our estimates are. In general, small standard errors suggest precise estimates whereas large standard errors suggest imprecise estimates. The fourth column (t) reports the t-statistic computed as $\frac{\alpha_k}{se_k}$ where se_k is the standard error corresponding to α_k for $k \in \{0, 1, 2\}$. Recall that for each parameter α_k, the null hypothesis is that $\alpha_k = 0$. We can use a standard t-statistic table to determine whether these null hypotheses are rejected or not. An alternative, often more popular way of determining whether the null hypothesis is true or false is to use the fifth column ($P > |t|$). This column reports p-values for each of the coefficients. Stated loosely: the p-value tells us the probability of obtaining the given estimate if the null hypothesis were to be true. Notice that our p-values are extremely small. Taking the corn coefficient as an example, what this says is: if the null hypothesis were true (i.e. $\alpha_1 = 0$), the probability of observing the relationship between chicken and corn present in this data would effectively be zero. So, this strongly suggests that the null hypothesis is not true (i.e. $\alpha_1 \neq 0$) and thus the price of corn has a non-negligible effect on the price of chicken. The last two columns ([95% Conf. Interval]) report the 95% confidence interval centered around our estimates. Notice that the true values of the model (i.e. the α values that we assumed) all lie within the 95% confidence interval of their corresponding estimates. This is further evidence that the linear regression has done a good job recovering the model parameters.

One more component of Figure 2 deserves attention. Look at the top right corner and you'll see a quantity called *R-squared* (or R^2). The R^2 value is the percent of the variation in our dependent variable which can be explained by variation in our independent variables. Figure 2 reports an R^2 of .8345. In words, this means that 83.45% of the variation in simulated chicken prices is attributable to variation in simulated corn and soybean prices. In practice, the R^2 tells us how well our model "fits the data." If our model fits very well, it means that it can do a good job replicating the prices we observe in the data. The higher the R^2, the better our model fits the data. Attaining good fit is important if we want to use this estimated model to make predictions about future simulated chicken prices. How can we make predictions like this? So far, we have estimated (fairly precisely) the impact of corn and soybean prices on the price of chicken. Based upon this estimated relationship, we can compute the predicted value of chicken prices as:

$$sim\_\hat{chicken}_t = \hat{\alpha}_0 + \hat{\alpha}_1\, sim\_corn_t + \hat{\alpha}_2\, sim\_soy_t \tag{7}$$

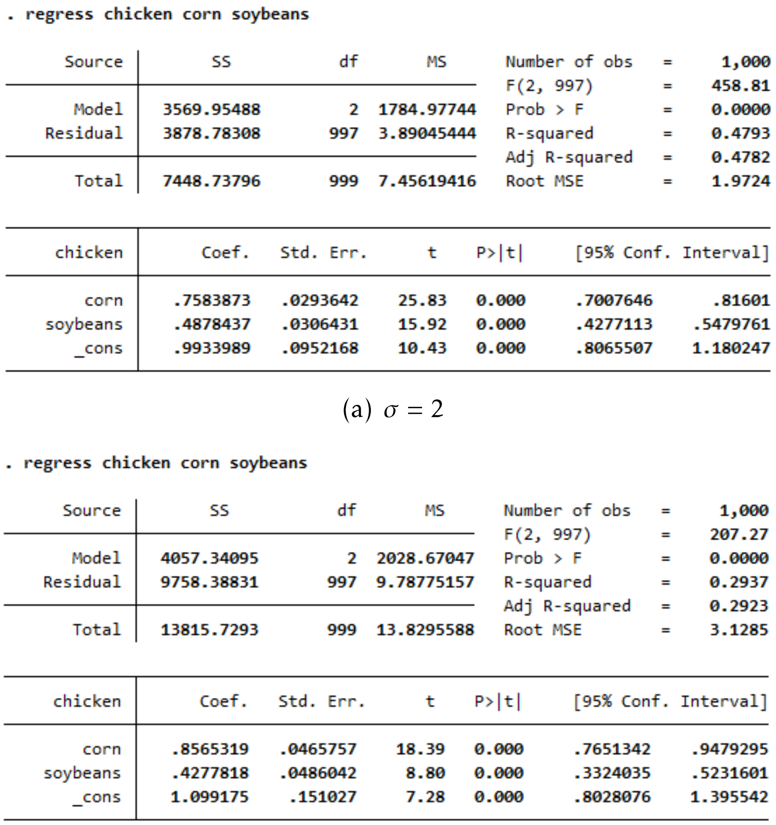
In English, Equation 7 states that given the prices or corn and soybeans, we expect the price of chicken to equal $\hat{\alpha}_0 + \hat{\alpha}_1\, sim\_corn_t + \hat{\alpha}_2\, sim\_soy_t$. In other words, if somebody told you the prices of corn and soybeans, but did not tell you the price of chicken, Equation 7 gives the best prediction of chicken prices based upon the information you were given. How reliable are these predictions? In short, higher R^2 means better predictions. With that being said, it is important to note that a high R^2 is usually not our ultimate goal. In fact, there is often a trade off between obtaining reliable predictions ($sim\_\hat{chicken}_t$ here) and obtaining reliable parameter estimates (the $\hat{\alpha}$'s). In economics, researchers typically care more about obtaining high quality parameter estimates than obtaining a good predictive model. In machine learning, the priority is often reversed. Either way, there are a substantial amount of empirical tools available to researchers interested in advancing either of the two agendas.

In summary, linear regression is a simple yet powerful tool in data science. Many real-world objects have extremely complicated functional relationships with their associated independent variables. In such a case, it is unlikely that we know the exact form of this relationship. Despite not knowing the functional relationship, we can often approximate it very well by collecting data and employing a simple linear regression. The exact specification of our regression can be specially tailored to attack a multitude of problems. We will explore a specific example of this next.

2.4 Revisiting the Effect of σ

As mentioned in Section 2.1, the standard deviation σ of the error term has important implications for our ability to recover model parameters. Recall that to when I simulated the data used to generate Figure 2, I set $\sigma = 1$. Let's see what happens if I conduct the same exact exercise, except I select higher values for σ. Panels (a) and (b) of Figure 3 show the same results when $\sigma = 2$ and $\sigma = 3$, respectively. Let's look closely at how these tables differ from Figure 2, paying particular attention to the standard errors and R^2.

Figure 3: Regression Output and σ



```
. regress chicken corn soybeans

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =    458.81
       Model |  3569.95488         2  1784.97744   Prob > F        =    0.0000
    Residual |  3878.78308       997  3.89045444   R-squared       =    0.4793
-------------+----------------------------------   Adj R-squared   =    0.4782
       Total |  7448.73796       999  7.45619416   Root MSE        =    1.9724

------------------------------------------------------------------------------
     chicken |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        corn |   .7583873   .0293642    25.83   0.000     .7007646      .81601
    soybeans |   .4878437   .0306431    15.92   0.000     .4277113    .5479761
       _cons |   .9933989   .0952168    10.43   0.000     .8065507    1.180247
------------------------------------------------------------------------------
```

(a) $\sigma = 2$

```
. regress chicken corn soybeans

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =    207.27
       Model |  4057.34095         2  2028.67047   Prob > F        =    0.0000
    Residual |  9758.38831       997  9.78775157   R-squared       =    0.2937
-------------+----------------------------------   Adj R-squared   =    0.2923
       Total |  13815.7293       999  13.8295588   Root MSE        =    3.1285

------------------------------------------------------------------------------
     chicken |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        corn |   .8565319   .0465757    18.39   0.000     .7651342    .9479295
    soybeans |   .4277818   .0486042     8.80   0.000     .3324035    .5231601
       _cons |   1.099175    .151027     7.28   0.000     .8028076    1.395542
------------------------------------------------------------------------------
```

(b) $\sigma = 3$

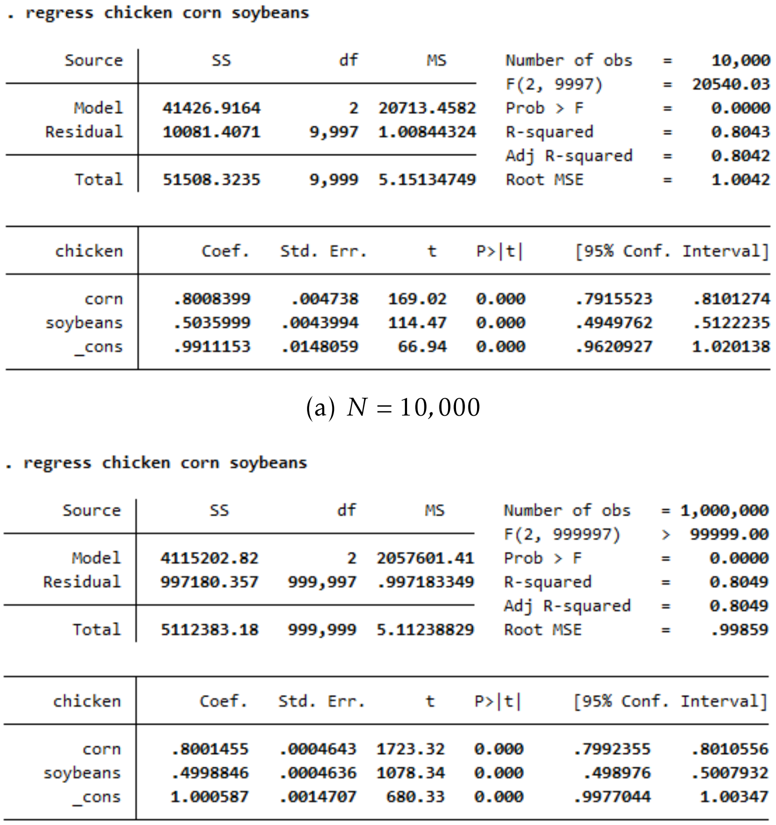
Let's first compare the R^2 values in Figures 2 and 3. Figure 2 has a very large R^2 value of .8345. Panel (a) of Figure 3 has a substantially lower R^2 (.4793). The value in Panel (b) is even lower (.2937). It seems that as σ increases, the resulting R^2 decreases. As the price of chicken becomes “more random,” its realized values may be very far from their predicted values (7). This decreases our model’s ability to fit the data, decreasing the resulting R^2.

Next, lets compare the standard errors reported in Figures 2 and 3. The standard errors in Panel (a) are over twice as large as the standard errors in Figure 2. Furthermore, the standard errors in Panel (b) are larger than those in Panel (a). Recall that the standard errors measure the precision of our parameter estimate, where large standard errors suggest imprecise estimation. It seems that the larger σ is, the larger the standard errors will be, and therefore the less precise our estimates will be. Why does this happen? A high value of σ implies a very noisy relationship between $sim\_chicken_t$ and what we observe in the data (sim_corn_t and sim_soy_t). The more noise present in our data, the more difficult it will be to infer the relationship between $sim\_chicken_t$ and its corresponding independent variables.

One additional point is worth mentioning. We’ve seen that the standard deviation σ or the error term has notable impact on our standard errors (i.e. the precision of our estimates). Another important determinant of the standard errors’ values is the size of our sample. Up until this point, I’ve reported OLS estimates using a simulated data set containing 1,000 observations. Let’s see what happens when if I execute the same exact simulation procedure, but I increase the size of the sample. I will set σ to our original value of 1 for this exercise.

If you examine the standard errors in Figure 4, you’ll see that they are much smaller than what we obtained in Figure 2. Furthermore, the parameter estimates are almost exactly equal to their true values in both cases. It seems that the precision of our estimates is phenomenal. I won’t show it in detail here, but the standard errors of our estimates will always strictly decrease as sample size increases. The practical implication for researchers is that larger samples permit more precise parameter estimates. This makes sense intuitively. Each observation represents a piece of information we have about the relationship between chicken, corn, and soy. The more information we have (i.e. the more observations are in our sample), the easier it should be for us to infer the relationship between chicken prices and their determinants. A large sample size is by no means a solution to any estimation issue which may arise, but there is no doubt that it helps us with precision.

Figure 4: Regression Output and Sample Size



```
. regress chicken corn soybeans

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =  20540.03
       Model |  41426.9164         2  20713.4582   Prob > F        =    0.0000
    Residual |  10081.4071     9,997  1.00844324   R-squared       =    0.8043
-------------+----------------------------------   Adj R-squared   =    0.8042
       Total |  51508.3235     9,999  5.15134749   Root MSE        =    1.0042

------------------------------------------------------------------------------
     chicken |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        corn |   .8008399    .004738   169.02   0.000     .7915523    .8101274
    soybeans |   .5035999   .0043994   114.47   0.000     .4949762    .5122235
       _cons |   .9911153   .0148059    66.94   0.000     .9620927    1.020138
------------------------------------------------------------------------------
```

(a) $N = 10,000$

```
. regress chicken corn soybeans

      Source |       SS           df       MS      Number of obs   = 1,000,000
-------------+----------------------------------   F(2, 999997)    >  99999.00
       Model |  4115202.82         2  2057601.41   Prob > F        =    0.0000
    Residual |  997180.357   999,997  .997183349   R-squared       =    0.8049
-------------+----------------------------------   Adj R-squared   =    0.8049
       Total |  5112383.18   999,999  5.11238829   Root MSE        =    .99859

------------------------------------------------------------------------------
     chicken |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        corn |   .8001455   .0004643  1723.32   0.000     .7992355    .8010556
    soybeans |   .4998846   .0004636  1078.34   0.000      .498976    .5007932
       _cons |   1.000587   .0014707   680.33   0.000     .9977044     1.00347
------------------------------------------------------------------------------
```

(b) $N = 1,000,000$

2.5 Application: Collusion in the Market for Chicken

Very often, economists (and doctors, data scientists, psychologists, etc.) are interested in estimating *treatment effects*. Put simply, a treatment effect is the effect of some "treatment" on some outcome of interest. Examples include:

- Effect of a medication on blood pressure
- Effect of a minimum wage on employment
- Effect of social media usage on mental health

In industrial organization, economists are often interested in estimating the effect of collusion on market prices. In this section, we will again use simulated data to illustrate how such an analysis may be conducted. Let's begin by setting the stage. Suppose that at some known point in time, call it t^*, chicken producers decide to collude and fix prices above their perfectly-competitive level. The price of chicken can then be expressed as:

$$sim\_chicken_t = \begin{cases} \alpha_0 + \alpha_1\, sim\_corn_t + \alpha_2\, sim\_soy_t + \epsilon_t & \text{if} \quad t < t^* \\ \alpha_0 + \alpha_1\, sim\_corn_t + \alpha_2\, sim\_soy_t + \gamma + \epsilon_t & \text{if} \quad t \geq t^* \end{cases} \tag{8}$$

Here γ is the effect of collusion on the price of chicken. To simplify notation, we can express (8) equivalently as:

$$sim\_chicken_t = \alpha_0 + \alpha_1\, sim\_corn_t + \alpha_2\, sim\_soy_t + \gamma \times \mathbf{1}[t \geq t^*] + \epsilon_t \tag{9}$$

$\mathbf{1}[t \geq t^*]$ is an indicator = 1 if $t \geq t^*$ (so there is collusion) and = 0 if $t < t^*$ (there is no collusion). Our ultimate goal will be to determine the effect of collusion on chicken prices. This amounts to estimating the parameter γ. To do this, we'll first need a control group. For this, we can use the price of turkey. Assume that the turkey market is perfectly competitive (i.e. no collusion), and turkey prices are given by:

$$sim\_turkey_t = \lambda_0 + \lambda_1\, sim\_corn_t + \lambda_2\, sim\_soy_t + \delta_t \tag{10}$$

δ_t is an error term just like the ϵ_t term in our chicken equation. We'll make the exact same distributional assumption for δ_t as we did for ϵ_t:

$$\delta_t \sim N(0, \nu) \tag{11}$$

For simplicity we will assume $\nu = \sigma = 1$ for this exercise. We'll assume the following values for the λ parameters:

$$\lambda_0 = 1$$
$$\lambda_1 = .3$$
$$\lambda_2 = .9$$

Let's simulate a new set of data which contains both chicken and turkey prices. We will assume the effect of collusion $\gamma = 1.5$ for the purposes of simulation. The process is effectively the same as before:

1. Using distributional assumptions X and Y, generate corn and soybean price data.

2. Using distributional assumption (5), draw T ϵ_t shocks.

3. Using distributional assumption (11), draw T δ_t shocks.

4. Generate T prices of chicken following Equation (9) along with the assumed values of the α parameters:

$$sim\_chicken_t = 1 + .8 \times sim\_corn_t + .5 \times sim\_soy_t + 1.5 \times \mathbf{1}[t \geq t^*] + \epsilon_t \tag{12}$$

5. Generate T prices of turkey following Equation (10) along with the assumed values of the λ parameters:

$$sim\_turkey_t = 1 + .3 \times sim\_corn_t + .9 \times sim\_soy_t + \delta_t \tag{13}$$

The result will be a simulated panel data set which looks something like: As before, the t column lists the time period corresponding to each row. Notice the additional i column. The value of i tells us whether the row is listing the price of chicken or turkey. Here, $i = 1$ corresponds to chicken and $i = 2$ corresponds to turkey.

Table 2: Sample of Simulated Data

t	i	**Price - Chicken/Turkey**	**Price - Corn**	**Price - Soybeans**
1	1	.288	.356	.354
2	1	.785	.313	.264
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$T-1$	1	5.63	2.64	1.18
T	1	.769	1.39	1.07
1	2	.799	.356	.354
2	2	1.01	.313	.264
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$T-1$	2	3.18	2.64	1.18
T	2	2.56	1.39	1.07

We need to define two new variables before proceeding. First, let $Post_t$ be an indicator = 1 if we are in the treatment period and = 0 otherwise. Concretely:

$$Post_t = \begin{cases} 0 & \text{if} \quad t < t^* \\ 1 & \text{if} \quad t \geq t^* \end{cases} \tag{14}$$

Next, let $Treat_i$ be an indicator = 1 if observation i is in the treatment group (i.e. chicken) and = 0 otherwise:

$$Treat_i = \begin{cases} 0 & \text{if} \quad i = 2 \\ 1 & \text{if} \quad i = 1 \end{cases} \tag{15}$$

$Treat_i$ simply tells us whether a given observation belongs to the chicken group or not, while $Post_t$ tells us whether collusion has begun or not. To recover the impact of collusion (i.e. estimate the γ parameter), we can run the following regression:

$$p_{it} = \beta_0 + \beta_1 \, Post_t + \beta_2 \, Treat_i + \beta_3 \, Post_t \times Treat_i + \beta_4 \, corn_t + \beta_5 \, soy_t + \epsilon_{it} \tag{16}$$

There are several ways to run this regression in Stata. First, you can manually create the interaction term by creating a new variable defined as $interact_{it} = Post_t \times Treat_i$. Then, type the following code in your command line:

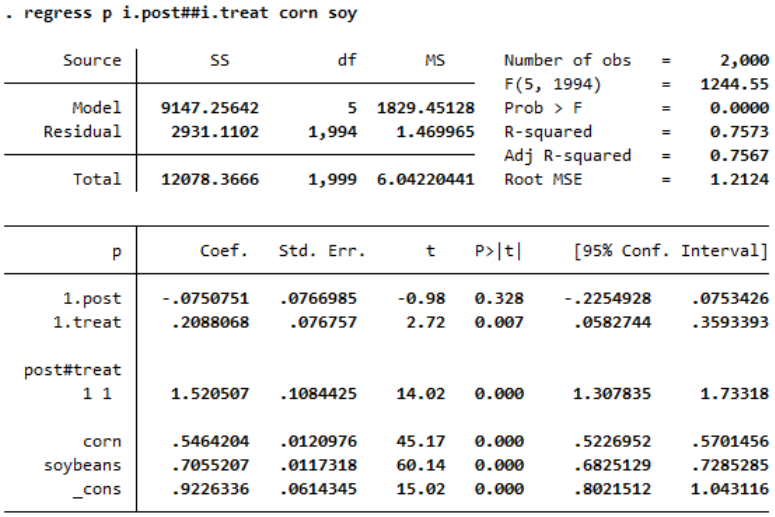
regress p post treat interact corn soy

We can avoid manually creating the interaction term by typing the following:

regress p i.post##i.treat corn soy

In this application, it will not matter which of the two approaches you take. In more sophisticated settings, the second approach is generally preferable. Typing either of the two lines of code in your command window will generate a table containing results.

Figure 5: Difference in Difference Results



```
. regress p i.post##i.treat corn soy

      Source |       SS           df       MS      Number of obs   =     2,000
-------------+----------------------------------   F(5, 1994)      =   1244.55
       Model |  9147.25642         5  1829.45128   Prob > F        =    0.0000
    Residual |  2931.1102      1,994    1.469965   R-squared       =    0.7573
-------------+----------------------------------   Adj R-squared   =    0.7567
       Total |  12078.3666     1,999  6.04220441   Root MSE        =    1.2124

------------------------------------------------------------------------------
           p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.post |  -.0750751   .0766985    -0.98   0.328    -.2254928    .0753426
     1.treat |   .2088068    .076757     2.72   0.007     .0582744    .3593393
             |
  post#treat |
        1 1  |   1.520507   .1084425    14.02   0.000     1.307835     1.73318
             |
        corn |   .5464204   .0120976    45.17   0.000     .5226952    .5701456
    soybeans |   .7055207   .0117318    60.14   0.000     .6825129    .7285285
       _cons |   .9226336   .0614345    15.02   0.000     .8021512    1.043116
------------------------------------------------------------------------------
```

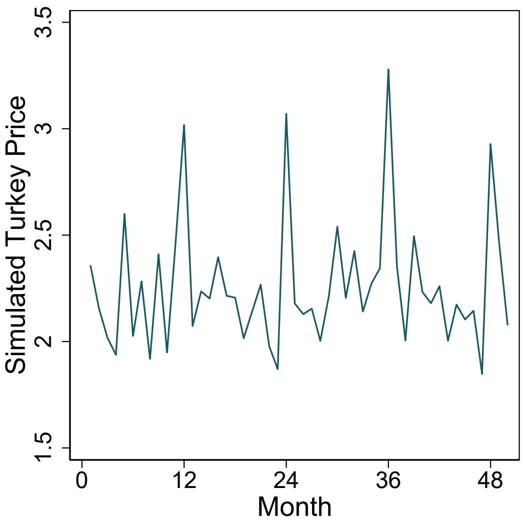
The *post#treat* row contains our estimate of the treatment effect γ. Note that it is very close to our assumed value of 1.5. Furthermore, the standard error is small relative to the estimate, t-stat is high, p-value is low, and the true value lies within the 95% confidence interval of our estimate. We can conclude that the difference in difference implementation was a success.

2.6 Practical Concerns

2.6.1 Seasonality

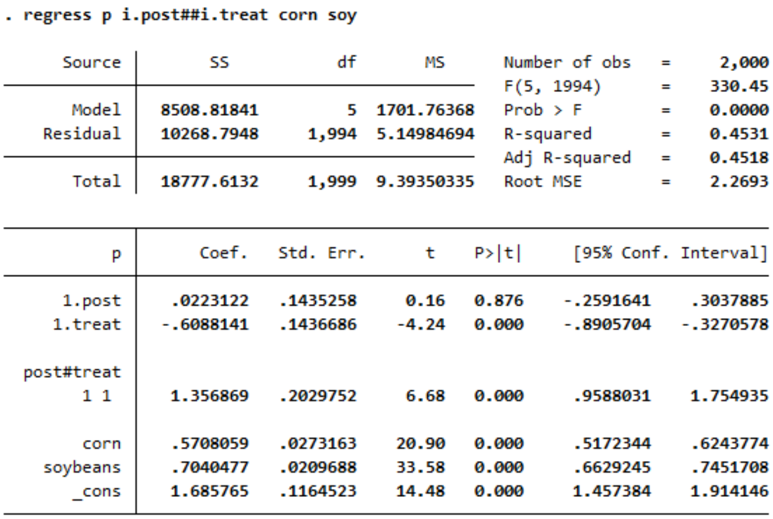
Though the previous implementation went smoothly, many problems can arise which can cause researchers headaches. Let's consider a specific example of one. Let's first recognize that once per year (Thanksgiving), demand for turkeys surges. Turkey prices increase in respond to this increased demand. Such a phenomenon will lead to some seasonal effects contaminating our data. Let's re-simulate our turkey data, this time artificially increasing the price once every 12 months.

Figure 6: Simulated Seasonality of Turkey



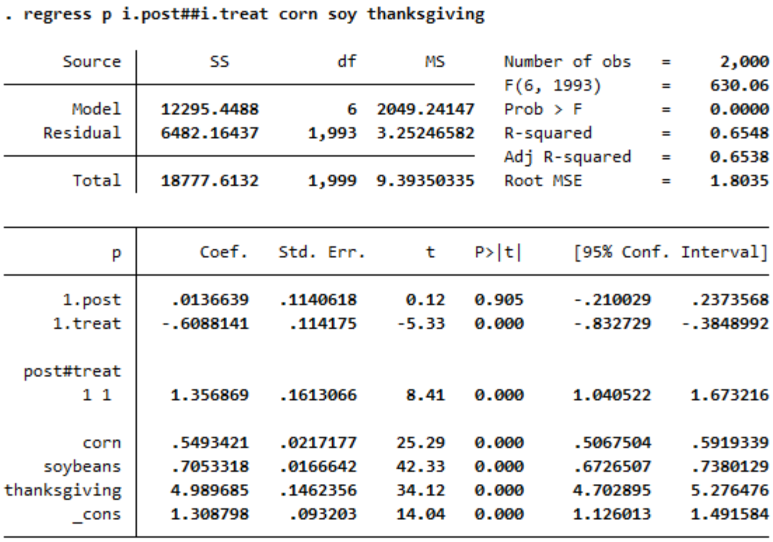
Seasonality of this flavor can easily be seen graphically. Figure 6 plots the first 50 observations of the new simulated turkey data, where we see a large spike once every 12 months. Severe seasonality can pose problems for our ability to recover the effect of collusion γ. To illustrate this, I'll report the estimates of Equation 16 using the new data containing a seasonal trend: Let's compare these results with our original results in Figure 5. Luckily, we were still able to detect a significant impact of collusion on prices. But, the precision of our estimate of γ has decreased substantially. The new standard error of our estimate $\hat{\gamma}$ is nearly twice what we obtained in Figure 5. Additionally, our R^2 (i.e. model fit) has significantly decreased. Luckily, there is a very easy solution. We can "control" for the seasonality of turkey by defining an indicator variable $thanksgiving_t$ which takes a value of 1 in the month of November, and 0 otherwise. Next, we can include this new variable as an independent variable in our regression equation. Estimating the resulting equation yields:

Figure 7: Difference in Difference Results



```
. regress p i.post##i.treat corn soy

      Source |       SS           df       MS      Number of obs   =     2,000
-------------+----------------------------------   F(5, 1994)      =    330.45
       Model |  8508.81841         5  1701.76368   Prob > F        =    0.0000
    Residual |  10268.7948     1,994  5.14984694   R-squared       =    0.4531
-------------+----------------------------------   Adj R-squared   =    0.4518
       Total |  18777.6132     1,999  9.39350335   Root MSE        =    2.2693

------------------------------------------------------------------------------
           p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.post |   .0223122   .1435258     0.16   0.876    -.2591641    .3037885
     1.treat |  -.6088141   .1436686    -4.24   0.000    -.8905704   -.3270578
             |
  post#treat |
        1 1  |   1.356869   .2029752     6.68   0.000     .9588031    1.754935
             |
        corn |   .5708059   .0273163    20.90   0.000     .5172344    .6243774
    soybeans |   .7040477   .0209688    33.58   0.000     .6629245    .7451708
       _cons |   1.685765   .1164523    14.48   0.000     1.457384    1.914146
------------------------------------------------------------------------------
```

Figure 8: Results Controlling for Seasonality



```
. regress p i.post##i.treat corn soy thanksgiving

      Source |       SS           df       MS      Number of obs   =     2,000
-------------+----------------------------------   F(6, 1993)      =    630.06
       Model |  12295.4488         6  2049.24147   Prob > F        =    0.0000
    Residual |  6482.16437     1,993  3.25246582   R-squared       =    0.6548
-------------+----------------------------------   Adj R-squared   =    0.6538
       Total |  18777.6132     1,999  9.39350335   Root MSE        =    1.8035

------------------------------------------------------------------------------
           p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.post |   .0136639   .1140618     0.12   0.905     -.210029    .2373568
     1.treat |  -.6088141    .114175    -5.33   0.000     -.832729   -.3848992
             |
  post#treat |
        1 1  |   1.356869   .1613066     8.41   0.000     1.040522    1.673216
             |
        corn |   .5493421   .0217177    25.29   0.000     .5067504    .5919339
    soybeans |   .7053318   .0166642    42.33   0.000     .6726507    .7380129
thanksgiving |   4.989685   .1462356    34.12   0.000     4.702895    5.276476
       _cons |   1.308798    .093203    14.04   0.000     1.126013    1.491584
------------------------------------------------------------------------------
```

Controlling for the month of thanksgiving has improved both our model fit (increased R^2) and the precision of our estimate $\hat{\gamma}$ (decreased its standard error). In practice, it is typically good to control for as much as possible given what is contained in your data. Omitting relevant variables from your regression equation can decrease precision, model fit, and lead us to make erroneous conclusions about the relationship between our dependent and independent variables.

2.6.2 Trends & Spurious Regression

I mentioned earlier that a high R^2 is not our ultimate goal. In fact, there are many instances where a particular model has a high R^2, but yields results which are utterly meaningless. For example, suppose that we regressed US GDP on cumulative rainfall in the US. Both variables have a steep upward time trend, so it is likely that Stata would report a high R^2 if we were to run this regression. As a result, one may think that cumulative rainfall is a good predictor of GDP. Obviously this is nonsense. When thinking about time series, as we are when we think about prices over time, it is important to recognize that two series moving together does not imply any causal relation between the two. When two series move together but are not causally related, we say there is a *spurious* relationship between the two. Such relationships appear often, and can easily contaminate results.

Let's think about our previous example. In the real world, it is likely to be the case that both chicken and turkey trend upwards over time. Let's see how such a trend may impact our difference in difference results. Suppose that there is some time trend τt which impacts turkey and chicken prices according to:

$$sim\_chicken_t = \alpha_0 + \alpha_1 sim\_corn_t + \alpha_2 sim\_soy_t + \gamma \times \mathbf{1}[t \geq t^*] + \tau t + \epsilon_t \tag{17}$$

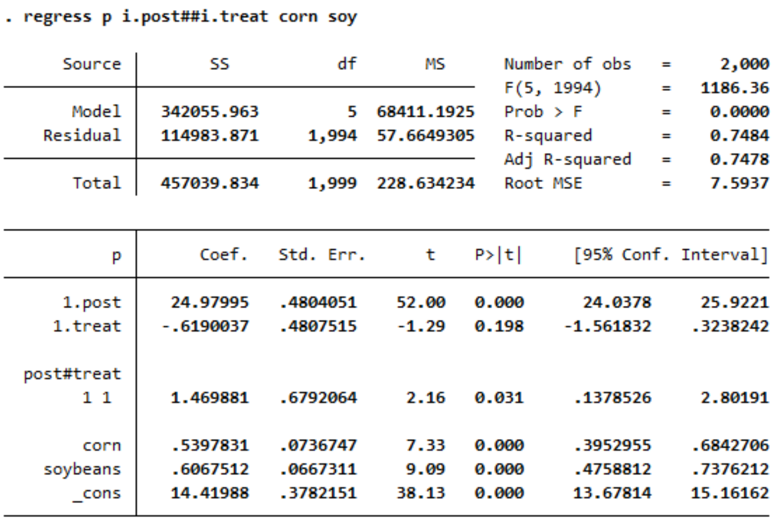
$$sim\_turkey_t = \lambda_0 + \lambda_1 sim\_corn_t + \lambda_2 sim\_soy_t + \tau t + \delta_t \tag{18}$$

Notice that the above equations are identical to what we had previously (Equations (12) and (13)), with the addition on the time trend τt. The parameter τ governs the slope of simulated chicken and turkey with respect to time. If $\tau = .05$ for example, this means that the prices increase by \$0.05 per month (all else equal). I'll adjust our simulated chicken and turkey variables to account for this time trend (with $\tau = .05$) and re-estimate Equation (16). I report results in Figure 9.

If we ignore this time trend, we lose a significant amount of statistical power. What I mean by this is that our ability to detect the treatment effect γ has fallen substantially. This is reflected, for example, by noticing the very large standard error and p-value for the interaction term relative to what we obtained before we added the time trend. The p-value here is .031, so luckily the time trend was mild enough so that we retain some amount of statistical significance. Even still, more precise estimates of γ are preferred to less precise estimates.

One approach for handling a time trend like this is to take the difference between simulated turkey and chicken prices. What do I mean by this? If you look at (17) and (18), you may notice that subtracting (18) from (17) would completely eliminate the time

Figure 9: Diff-in-Diff (Ignoring Time Trend)



```
. regress p i.post##i.treat corn soy

      Source |       SS           df       MS      Number of obs   =     2,000
-------------+----------------------------------   F(5, 1994)      =   1186.36
       Model |  342055.963         5  68411.1925   Prob > F        =    0.0000
    Residual |  114983.871     1,994  57.6649305   R-squared       =    0.7484
-------------+----------------------------------   Adj R-squared   =    0.7478
       Total |  457039.834     1,999  228.634234   Root MSE        =    7.5937

------------------------------------------------------------------------------
           p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.post |   24.97995   .4804051    52.00   0.000      24.0378     25.9221
     1.treat |  -.6190037   .4807515    -1.29   0.198    -1.561832    .3238242
             |
   post#treat |
        1 1  |   1.469881   .6792064     2.16   0.031     .1378526     2.80191
             |
        corn |   .5397831   .0736747     7.33   0.000     .3952955    .6842706
    soybeans |   .6067512   .0667311     9.09   0.000     .4758812    .7376212
       _cons |   14.41988   .3782151    38.13   0.000     13.67814    15.16162
------------------------------------------------------------------------------
```

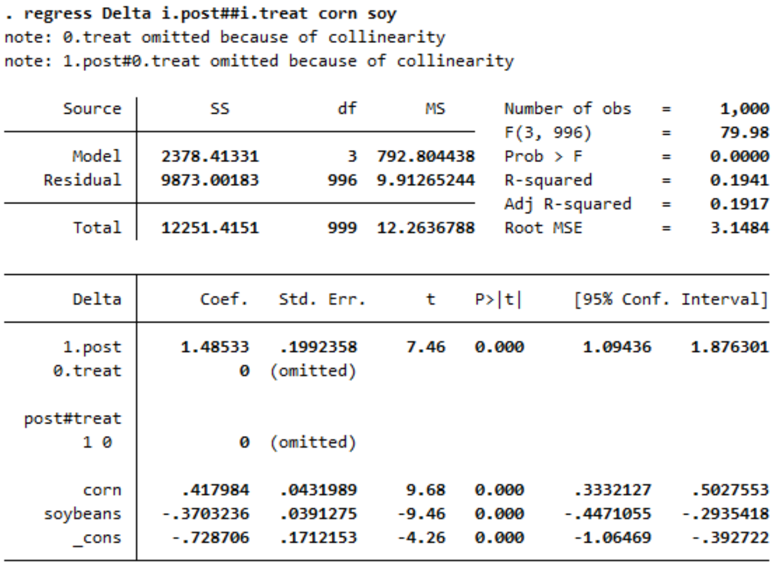
trend τt:

$$\Delta_t = sim\_chicken_t - sim\_turkey_t =$$
$$(\alpha_0 - \lambda_0) + (\alpha_1 - \lambda_1) sim\_corn_t + (\alpha_2 - \lambda_2) sim\_soy_t + \gamma \times \mathbf{1}[t \geq t^*] + \epsilon_t - \delta_t \qquad (19)$$

I've defined the new variable Δ_t as the difference between simulated chicken and turkey prices in period t. The virtue of the variable Δ_t is that it contains no time trend τ_t. Let's try estimating Equation (16), but using Δ_t as the left-hand-side variable. I show results in Figure 10. A few things may look odd at first. Notice that the $Treat_i$ and $Post_t \times Treat_i$ variables have been omitted. When taking the contemporaneous difference between prices, the $Treat_i$ variable no longer has much meaning. $Treat_i$ was meant to classify observations as either belonging to the chicken or turkey group. Which of the two groups does Δ_t belong to? This isn't really a sensible question. Because of this, any term involving the $Treat_i$ variable drops out of the regression. The mathematical reason for this is referred to as *collinearity*. In short, we cannot include independent variables which are perfectly predicted by other variables in the equation.

Given that the $Treat_i$ variable is now meaningless. We must adjust our interpretation of the results. The coefficient representing the treatment effect has been completely absorbed by the $Post_t$ variable. Manipulating Equation (16) may reveal how exactly this happens. This is our new estimate of the treatment effect. Compare this estimate with the estimate of the interaction term coefficient in Figure 9. Using the contemporaneous difference Δ_t as the left-hand-side variable has allowed us to obtain a much more precise estimate of γ. This is reflected, for example, by the much smaller standard errors and

Figure 10: Diff-in-Diff (Using Contemporaneous Difference Δ_t)



```
. regress Delta i.post##i.treat corn soy
note: 0.treat omitted because of collinearity
note: 1.post#0.treat omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(3, 996)       =     79.98
       Model |  2378.41331         3  792.804438   Prob > F        =    0.0000
    Residual |  9873.00183       996  9.91265244   R-squared       =    0.1941
-------------+----------------------------------   Adj R-squared   =    0.1917
       Total |  12251.4151       999  12.2636788   Root MSE        =    3.1484

------------------------------------------------------------------------------
       Delta |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.post |    1.48533   .1992358     7.46   0.000      1.09436    1.876301
     0.treat |          0  (omitted)
             |
  post#treat |
        1 0  |          0  (omitted)
             |
        corn |    .417984   .0431989     9.68   0.000     .3332127    .5027553
    soybeans |  -.3703236   .0391275    -9.46   0.000    -.4471055   -.2935418
       _cons |   -.728706   .1712153    -4.26   0.000     -1.06469    -.392722
------------------------------------------------------------------------------
```

p-value. Our goal was to obtain a more precise estimate of γ than what we obtained in Figure 9, and we were able to achieve this by using the contemporaneous difference Δ_t.

3 Empirical Exercise

In the previous section, we used simulated data to illustrate the implementation of some simple econometric techniques. Now, I will walk through how to create the real sample using commodity data from the Bureau of Labor Statistics (BLS). First, I'll describe how to download and clean the data. Lastly, I'll show how to use Stata to create some simple descriptive statistics. This part of the document is intended for Stata users. To echo my opinion from before, you will save yourself from a lot of headaches by using Stata over Excel. For the cleaning portion of this section, look through the "clean.do" file to see the exact sequence of steps.

3.1 Constructing our Sample

1. Download time series from this link

    - Scroll down to where is says "Commodity Data including "headline" FD-ID indexes." Click the green column labeled "One Screen."
    - A new window will appear with two boxes. In the left box, click "01 Farm products."

- There are three sets of price data we want: corn, soybeans, and slaughter chickens. Above the right box, search “corn” and select “012202 Corn.” Make sure “Not Seasonally Adjusted” is selected (and “Seasonally Adjusted” in not selected), then click “Add to selection.” Repeat for soybeans and slaughter chickens.
- When all three items are in the “Your selection” box, click “Get Data.”
- Three tables will appear, one for each of our price series. Above each table is a download link. Click the link and open the three excel sheets.
- Save the three excel sheets in Folder/Data/Raw/. Name the excel sheets for corn, soybeans, and chicken “corn.xlsx,” “soybeans.xlsx,” and “chicken.xlsx” respectively.

2. Convert excel spreadsheets to Stata files.

    - Open Stata. On the toolbar in the top left corner, there is a tab called “New Do-File Editor.” Click it and a blank file will open. Save this as “clean.do” in your code subfolder. This is where we will type our cleaning code.
    - Set a file directory so Stata knows where to look for the excel files.
    - Next, we’ll need to reshape each of the three file. If you look at the excel files, you’ll see that each columns corresponds to a month while each row corresponds to a year. Essentially, what we want to do is take each row, transpose it (basically rotate it 45 degrees clockwise), and stack them on top of each other. This can be a bit involved at times, but see the “Clean.do” file for step-by-step instructions.

3. Merge the chicken, corn, and soybean data to create the final sample.

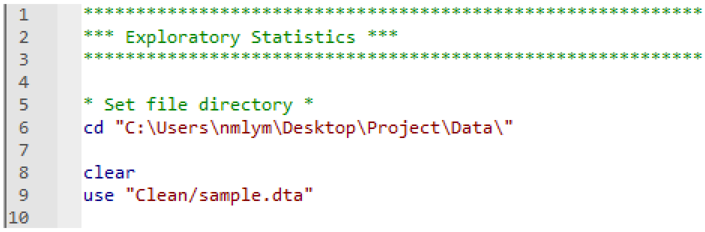
    - Now, we can merge each of the three files to create our final sample. Start by loading one of the three files. I’ll start with the chicken file. Type the following two lines in your Do file:
      clear
      use “Raw/chicken.dta” Next, merge with the corn file by typing:
      merge 1:1 year month using “Raw/corn.dta”
      Finally, merge with the soybeans file by typing:
      merge 1:1 year month using “Raw/soybeans.dta”

- With all three files merged, our construction of the sample is complete. The last thing we need to do is save our final sample. I'll call this final sample "sample.dta" and save it in the "Cleaned" subfolder within my Data folder: save"Clean/sample.dta", replace

3.2 Exploratory Analysis

At this point, we have our final sample "sample.dta" saved in Project/Data/Clean. Now we can begin exploring the data. In this section, I'll quickly give some tips and useful commands which should help you get started. First, start a new do file for the exploratory code. I will call mine "exploratory.do." As before, the first lines of your Do file should: (1) set a file directory, (2) clear, (3) import the sample. Here is what mine looks like:

Figure 11: Beginning of Do File



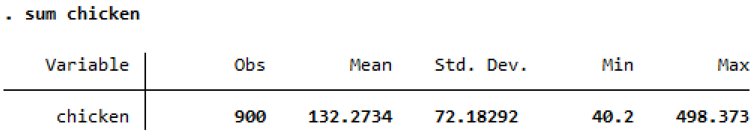
```
***********************************************************
*** Exploratory Statistics ***
***********************************************************

* Set file directory *
cd "C:\Users\nmlym\Desktop\Project\Data\"

clear
use "Clean/sample.dta"

```

Once the sample is imported into Stata. We can begin with the exploratory analysis. A first step we may take is generating summary statistics of our key variable. Taking chicken as an example, type:

sum chicken

This will create a small table with summary statistics:

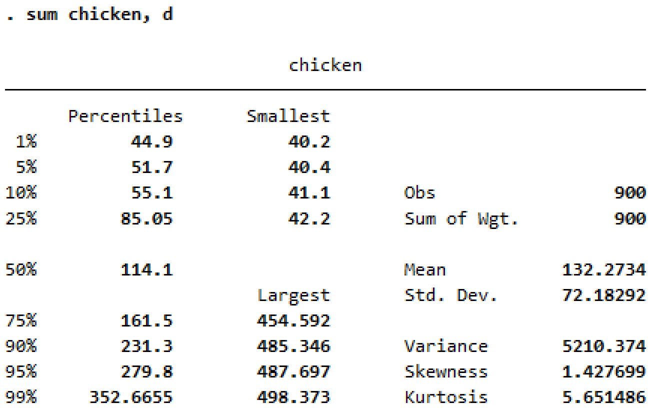
Figure 12: Summary Statistics



```
. sum chicken

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     chicken |        900    132.2734    72.18292       40.2    498.373
```

In order, what Stata gives us is the number of chicken observations, the mean of chicken prices, the standard deviation, and the smallest and largest values in the sample. These are nice, but we may want a bit more detail. For example, what is the median chicken price? To get this, we can generate more detailed summary statistics by adding a ", d" to the previous command:

Figure 13: More Detailed Summary Statistics



```
. sum chicken, d

                           chicken
-------------------------------------------------------------
      Percentiles      Smallest
 1%         44.9           40.2
 5%         51.7           40.4
10%         55.1           41.1       Obs                 900
25%        85.05           42.2       Sum of Wgt.         900

50%        114.1                      Mean           132.2734
                        Largest       Std. Dev.      72.18292
75%        161.5        454.592
90%        231.3        485.346       Variance       5210.374
95%        279.8        487.697       Skewness       1.427699
99%     352.6655        498.373       Kurtosis       5.651486
```

sum chicken, d

Now we have all the same summary statistics as before, plus many more. The leftmost column shows various percentiles of the chicken prices distribution, where the 50th percentile corresponds to the median. The column second to the left shows the 5 smallest and largest values in the sample. Such information can be useful for determining whether or not it is a good idea to drop extreme values. If the maximum value were to be extremely far away from the 99th percentile, it may be a good idea to toss some of these outliers out. Whether or not this is a good course of actions depends heavily on the particular application. We again see the number of observations, mean, and standard deviation. Additionally, we see the variance (std. dev. squared), skewness, and kurtosis (basically a measure of how high the peaks of the distribution are). You can use the exact same code for the corn and soybeans prices by simply swapping the name “chicken” with the variable of interest.

In addition to generating statistics which summarize the distribution of chicken prices (or any other variable), we can look at the distribution itself. To plot (an estimate) of the probability density function for chicken prices, type:

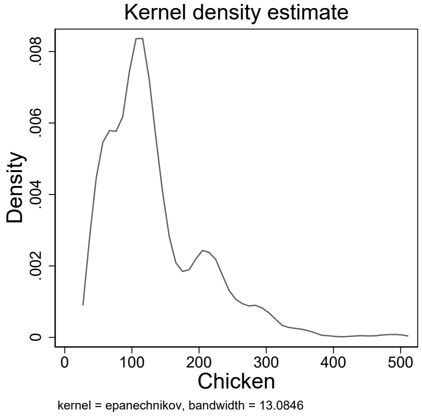
kdensity chicken

This will plot the distribution of chicken prices. As always, the same code can be applied to any of our other variables.

Importantly, prices may change with time. Computing averages across the entire sample will hide this. If we want to look at the evolution of chicken prices over time, any easy way to do this is by simply plotting the price of chicken over time. To do this, simply type:

line chicken t

Figure 14: Chicken Price Distribution



The “line” command will create a line plot where the first variable corresponds to what is on the vertical axis, while the second corresponds to the horizontal axis. We can do the same thing for corn or soybeans. Even better, we can overlay the three graphs using the following line of code:

twoway (line chicken t, color(blue)) (line corn t, color(red)) (line soy t, color(green)), legend(label(1 ”Chicken”) label(2 ”Corn”) label(3 ”Soybeans”)) ytitle(”Price”) xtitle(”Year”)

The color options set the “color” of each line, “legend” defines the labels corresponding to each line, and “ytitle” and “xtitle” create the axis titles. The result is shown in Figure 15. Figures like these are nice because they are easy for an audience to read and despite their simplicity, can highlight important features of the data (ex: the upward trend in prices over time).

Figure 15: Prices over Time

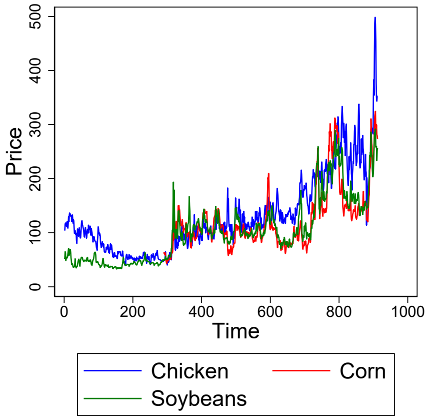
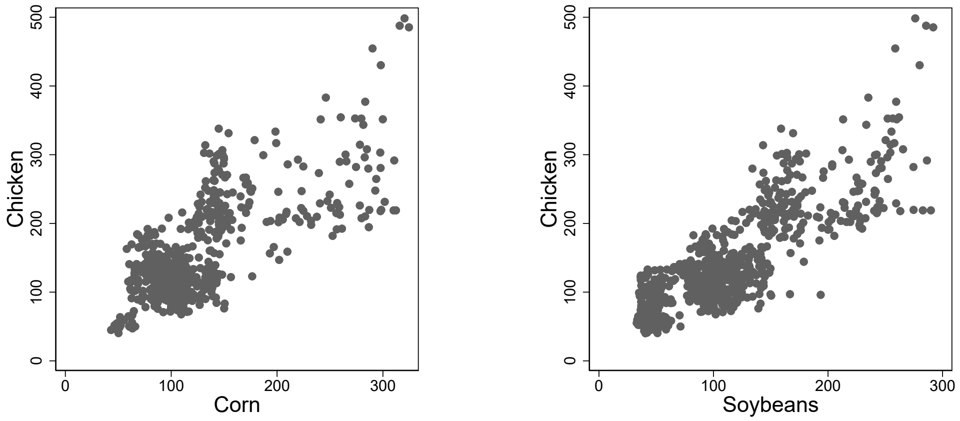


Figure 16: Scatter Plots



(a) Chicken & Corn

(b) Chicken & Soy

Next, we may be interested in the extent to which our variables of interest are correlated. A natural step to take along this path is to create a scatter plot. To scatter chicken prices against corn prices, type:

scatter chicken corn

Just like when using “line,” when using “scatter,” the first variable listed will appear one the vertical axis while the second will appear on the horizontal axis. We can use similar code to create scatter plots using any variables we like. Separately scattering chicken prices against corn and soybean prices will yield Figure 16.

Clearly, chicken prices appear to be positively correlated with both corn and soybean prices. This should come at no surprise, as corn and soybeans are important inputs in the chicken production process. What is the nature of the underlying relationship between these variables? As you may realize, this is not something we know with perfect precision. But, we can apply the methods discussed throughout the simulation exercise to begin estimating the parameters which govern this relationship.