Multiple Regression

AGEC 317: Economic Analysis for Agribusiness Management

Instructor: Michael Black

TEXAS A&M UNIVERSITY
Agricultural
Economics

Multiple regression

Recall the example where we regressed miles-per-gallon of cars on vehicle weight. The theoretical model was:

$$MPG_i = \beta_0 + \beta_1 weight_i + \varepsilon_i$$

Do you really believe the weight of a car is the only important variable in predicting the MPG of a car?

Multiple regression

Both of these vehicles weigh about 4500lbs:


Do they have the same MPG? Our original model said: YES

Multiple regression

How can we improve our theoretical model to better predict MPG?

Multiple regression

New model:

$$MPG_i = \beta_0 + \beta_1 weight_i + \beta_2 cyl_i + \beta_3 disp_i + \varepsilon_i$$

Multiple regression

Building models with multiple explanatory (independent) variables is often much better than a univariate model.

- Is quantity demand for Pepsi a function of *only* the price?
- Is the quantity supplied of electricity a function of *only* the cost of production?

Multiple regression: Interpretation

Multiple regression is a relatively simple extension of univariate linear regression: we are just adding variables! But what does this *mean*? Does it change our interpretation of the variables?

Yes.

Multiple regression: Interpretation

Recall our discussion from early in the semester about interpreting partial effects. Suppose we have the following model:

$$y_i = \alpha + \beta x_i + \gamma z_i$$

And suppose we move from one point on that function to another point. We will move a distance of Δ:

$$\Delta y_i = \beta \Delta x_i + \gamma \Delta z_i$$

The effect from a movement of x_i on y_i is then:

$$\frac{\Delta y_i}{\Delta x_i} = \beta + \gamma \frac{\Delta z_i}{\Delta x_i}$$

Multiple regression: Interpretation

$$\frac{\Delta y_i}{\Delta x_i} = \beta + \gamma \frac{\Delta z_i}{\Delta x_i}$$

As we let the change (Δ) become very very small, we get:

$$\frac{\partial y_i}{\partial x_i} = \beta + \gamma \frac{\partial z_i}{\partial x_i}$$

So the partial effect of x_i on y_i in a multiple regression framework is β **if** z_i is held constant so there is no change.

Multiple regression: Interpretation

In our MPG model:

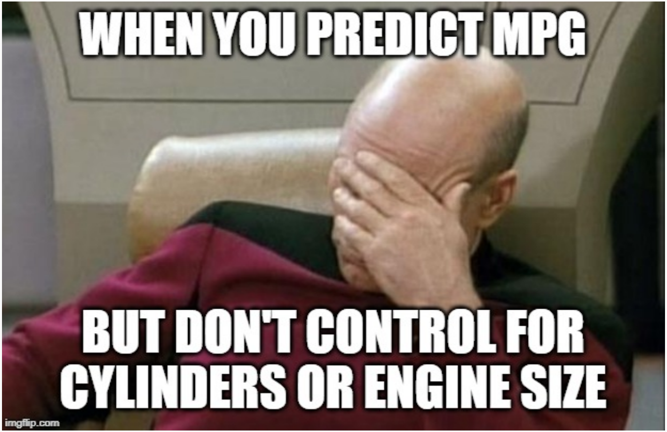
$$MPG_i = \beta_0 + \beta_1 weight_i + \beta_2 cyl_i + \beta_3 disp_i + \varepsilon_i$$

We say that β_1 is the partial effect of vehicle weight on MPG, **holding the number of cylinders and engine size constant**. In our naive model:

$$MPG_i = \beta_0 + \beta_1 weight_i + \varepsilon_i$$

β_1 is the partial effect, but we aren't holding anything constant! That means the estimated effect could be because of weight, or something else we aren't observing.

Multiple regression: Interpretation



Derivation encore

Last lecture, we used calculus to show that a model of the form:

$$y_i = \beta_0 + \beta_1 x_i$$

is best estimated as the minimization of the sum of squared residuals, resulting in the following parameter estimates:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

Derivation encore

In this lecture, we want to perform linear regression on a more complicated model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}$$

...and we want to be explicit about the assumptions we make when we run regressions. We also want to apply multiple to regression in a meaningful way.

Derivation of OLS: Multiple Regression

Suppose we start with a general multivariate regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}$$

The **residual** associated with an estimation of the above true model is:

$$\begin{aligned} y_i - \hat{y}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}) \\ &= \hat{u}_i \end{aligned}$$

Derivation of OLS: Multiple Regression

Our optimization problem is then:

$$\min_{\hat{\beta}_0,\cdots,\hat{\beta}_n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni})]$$

If we have a model with 10 variables, that means we will have 11 FOCs...

Solving a system of 11 equations with 11 unknowns. Does that sound fun?

Derivation of OLS: Multiple Regression

Lucky for us, we can use **linear algebra** to solve for the 11 coefficients in the model.

Derivation of OLS: Multiple Regression

Remember this?

	A	B	C
1	ID	AGEC 317 Grade	Starting Salary
2	1	91	89
3	2	79	24
4	3	99	100

$$\downarrow$$

$$\begin{bmatrix} 1 & 91 & 89 \\ 2 & 79 & 24 \\ 3 & 99 & 100 \end{bmatrix}$$

Derivation of OLS: Multiple Regression

Let's derive the optimal estimates for $\hat{\beta}_0, \cdots, \hat{\beta}_k$. Some assumptions before we start:

- The true model is: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$
- The model has $k + 1$ unknowns: all slope coefficients plus the intercept. Let $k + 1 = j$.
- There are n observations

Derivation of OLS: Multiple Regression

y	x_1	x_2	$\cdots$	x_k
45	300	0	$\cdots$	10
100	300	15	$\cdots$	5
80	280	10	$\cdots$	15

$$\downarrow$$

$$\begin{aligned} 45 &= 300x_{1i} + 0x_{2i} + \cdots + 10x_{ki} + u_i \\ 100 &= 300x_{1i} + 15x_{2i} + \cdots + 5x_{ki} + u_i \\ 80 &= 280x_{1i} + 10x_{2i} + \cdots + 15x_{ki} + u_i \end{aligned}$$

$$\downarrow$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

Derivation of OLS: Multiple Regression

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \ U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix}$$

Derivation of OLS: Multiple Regression

In matrix form:

$$
\begin{aligned}
Y &= X\beta + U \\
(n \times 1) &= (n \times j)(j \times 1) + (n \times 1)
\end{aligned}
$$

Derivation of OLS: Multiple Regression

Example:

$$Y = \begin{bmatrix} 45 \\ 100 \\ 80 \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \ U = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & 300 & 0 & 10 \\ 1 & 300 & 15 & 5 \\ 1 & 280 & 10 & 15 \end{bmatrix}$$

Derivation of OLS: Multiple Regression

Example:

$$Y = X\beta + U$$

$$\begin{bmatrix} 45 \\ 100 \\ 80 \end{bmatrix} = \begin{bmatrix} 1 & 300 & 0 & 10 \\ 1 & 300 & 15 & 5 \\ 1 & 280 & 10 & 15 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

$$Y = X\hat{\beta} + U$$

$$\begin{bmatrix} 45 \\ 100 \\ 80 \end{bmatrix} = \begin{bmatrix} 1 & 300 & 0 & 10 \\ 1 & 300 & 15 & 5 \\ 1 & 280 & 10 & 15 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

Derivation of OLS: Multiple Regression

In matrix form, note that:

$$Y = X\beta + U$$

$$\hat{Y} = X\hat{\beta}$$

$$\hat{U} = Y - \hat{Y} = Y - X\hat{\beta}$$

Derivation of OLS: Multiple Regression

Using linear algebra this time, let's take another look at our minimization problem:

$$\min_{\hat{\beta}} \hat{U}'\hat{U} = \min_{\hat{\beta}} (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

Derivation of OLS: Multiple Regression

$$
\begin{aligned}
\min \hat{U}'\hat{U} &= \min(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
&= \min(Y' - \hat{\beta}'X')(Y - X\hat{\beta}) \\
&= \min Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta} \\
&= \min Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}
\end{aligned}
$$

Derivation of OLS: Multiple Regression

Now we need to take the FOC wrt the $\hat{\beta}$ vector:

$$
\begin{aligned}
\frac{\partial Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}}{\partial\hat{\beta}} &= 0 \\
0 - 2X'Y + 2X'X\hat{\beta} &= 0 \\
2X'X\hat{\beta} &= 2X'Y \\
X'X\hat{\beta} &= X'Y \\
(X'X)^{-1}X'X\hat{\beta} &= (X'X)^{-1}X'Y \\
\hat{\beta} &= (X'X)^{-1}X'Y
\end{aligned}
$$

Derivation of OLS: Multiple Regression

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Derivation of OLS: Multiple Regression

See the "matrixOLS.xlsx" for an example of how we can use our new equation for $\hat{\beta}$.

Regression: Technical Model Fit

When we perform regression, we minimize the sum of squared residuals. That results in the best fit, right? Well, it results in the best fit *given the model specification*. If we add or subtract variables, the SSR could go up or down, and if it goes down (decreases) the model fit improves.

But what about just looking at a single regression. Can we assess the goodness of fit without comparing to another model?

Regression: Technical Model Fit

Recall the following:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Recall that: $SST = SSE + SSR$.

Regression: Technical Model Fit

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

"R-squared" is the fraction of the sample variation of the dependent variable explained by the independent variable(s). R^2 is bounded by 0 and 1, and as $R^2 \to 1$, the model does a better job at "explaining" the dependent variable.

Regression: Technical Model Fit

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Recall that $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. If we add additional explanatory variables, SSR will fall (as long as there is some error in the model), which suggests that R^2 improves with more variables. Indeed it does. We can even drive R^2 to 1 if we include as many variables as observations!

Regression: Technical Model Fit

We can also look at **adjusted-**R^2:

$$\bar{R}^2 = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}}$$

where n is the number of observations, and k is the number of independent (explanatory) variables.

Regression: Technical Model Fit

How you should use R^2 and $\bar{R}^2$:

1. Is the model a multivariate regression?
   - Yes: Use $\bar{R}^2$, move to (2)
   - No: Use R^2, move to (2)
2. Do you have another model you are comparing to?
   - Yes: the model with the higher R^2 or $\bar{R}^2$ is better.
   - No: check to make sure the R^2 or $\bar{R}^2$ is not low*.

* "Low" is subjective; if the R^2 or $\bar{R}^2$ is 0.01, that's not great.

Regression: Assumptions

When we run a regression, we are making implicit assumptions. These are:

- **Linear in parameters**: the true (population) model *can* be written as:
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$
- **Random sampling**: the sample we use for regression is randomly selected from the population of interest.

Regression: Assumptions

Continued:

- **No perfect co-linearity**: there can be no exact relationship between independent (explanatory) variables.
- **Zero conditional mean**: the expected value (average) of the error term is zero, given the values of the independent variables.
- **Homoskedasticity**: the variance of the error term is constant, and *not* a function of anything.

Under the first four assumptions, $E[\hat{\beta}_j] = \beta_j$, meaning the OLS estimate is **unbiased**.

Under all five assumptions, for any other type of estimate that is linear and unbiased, $\tilde{\beta}_j$, $var(\hat{\beta}_j) \leq var(\tilde{\beta}_j)$, meaning any other type of estimator is not as **efficient** (has higher variance) than our $\hat{\beta}_j$ from OLS.

The first theorem (and thus four assumptions) tells us our OLS estimate $\hat{\beta}_j$ is linear and unbiased, and the second theorem tells us we have the *best* linear and unbiased estimator. That is, our estimate is BLUE. This second theorem is better known as the **Gauss-Markov Theorem**.

Sources of bias

What if we include irrelevant variables or forget to include important ones? Will our estimates be affected?

Including irrelevant variables

Including irrelevant variables does not affect the regression and is not a source of bias. Suppose:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

where $\beta_2 = 0$ in the population model.
Including z in the regression does not affect the identification of β_1.

Excluding relevant variables

Excluding relevant variables **does** affect the regression and is a source of bias. Suppose:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

And suppose $\beta_2 \neq 0$ and $corr(x, z) \neq 0$

Excluding relevant variables

Now suppose instead of:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

We estimate:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

And $z = \delta_0 + \delta_1 x + u$

Excluding relevant variables

$$\begin{aligned} \Rightarrow y &= \beta_0 + \beta_1 x + \beta_2(\delta_0 + \delta_1 x + u) + \varepsilon \\ y &= (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x + (\beta_2 u + \varepsilon) \end{aligned}$$

$\Rightarrow bias = \hat{\beta}_1 - \beta_1 = \beta_1 + \beta_2\delta_1 - \beta_1 = \beta_2\delta_1$

This type of bias is known as **omitted variable bias**.

Direction of OVB

if	$corr(x,z) > 0$ or $\delta_1 > 0$	$corr(x,z)$ or $\delta_1 < 0$
$\beta_2 > 0$	$\beta_2\delta_1 > 0$: Positive bias	$\beta_2\delta_1 < 0$: Negative bias
$\beta_2 < 0$	$\beta_2\delta_1 < 0$: Negative bias	$\beta_2\delta_1 > 0$: Positive bias

OVB

Remember:

- **Including irrelevant variables is fine.**
- **Excluding relevant variables is not fine.**

Regression F-test

In a multivariate regression, we can test the *joint* significance of the model:

$$y = \beta_0 + \beta_1 x + \beta_2 z$$

Are *any* of the variables significant? We can use an F-test for this.

Regression F-test

$$y = \beta_0 + \beta_1 x + \beta_2 z$$

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : any \quad \beta_j \neq 0$$

Regression F-test

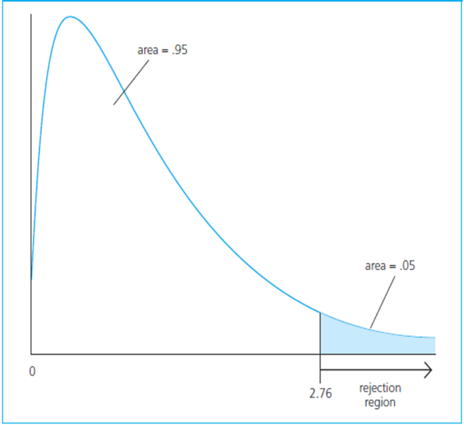
The F-statistic is calculated from the regression results as:

$$F_{stat} = \frac{SSE/k}{SSR/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Regression F-test

The F_{stat} follows a (k, n-k-1) distribution:

FIGURE 4.7 The 5% critical value and rejection region in an $F_{3,60}$ distribution.



Regression F-test

The hypothesis decision follows a similar process to the t-test:

1. Form your null hypothesis
2. Form the alternative hypothesis
3. Compute F-stat
4. Use critical value to make decision to reject or fail-to-reject null.

(1) is always "everything equal to zero", and (2) is always "something is not zero", at least in this class

Excel

Now let's apply multiple regression to a few models...

Key Skills

In this lecture, we discussed how to estimate and interpret a multivariate regression, and some major assumptions and sources of bias. At this point, you should be able to:

- Perform multivariate regression in Excel without using the Data Analysis ToolPak
- Perform multivariate regression in Excel using the Data Analysis ToolPak
- Interpret the results from a multivariate regression
- Confirm that a multivariate regression is BLUE
- Defend a multivariate regression from potential sources of bias
- Test for the joint significance of a multivariate regression