

# Week 11 Live Session

*w203 Instructional Team*

*July 25, 2016*

---

## Multiple linear regression (MLR)

A simple linear regression uses a single predictor variable (X) to model the response variable (Y). In the real world, relationships are often more complex and more than one factor can influence the outcome of an event (i.e. the response variable). A multiple linear regression describes how a single response variable Y depends linearly on a number of predictor variables.

For example, a single linear regression can be used to show the impact of location on housing prices. A multiple linear regression can also factor in other factors that may have a stronger impact on the price, such as the number of bathrooms, number bedrooms, the number of bathrooms, the year the house was built, the square footage, etc.

We will use the term multiple regression for models with two or more predictors and one response (There are also regression models with two or more response variables, which are called multivariate regression models).

## Interpreting coefficients in multiple regression correctly

Consider the following estimated regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

where  $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2, \hat{\beta}_2 = 3$

Q1.1 Correctly interpret the coefficient associated with  $x_1, \hat{\beta}_1$ . Bonus: use the words “ceteris paribus” in your answer.

---

## OLS Estimators

Remember that a regression coefficient based on sample data is an estimate of a true regression parameter for the population the sample is drawn from. There are several desirable properties that we might want our coefficients to have. These properties require different sets of assumptions to hold.

Q2.1: What does it mean for an estimator to be unbiased?

Q2.2: What does it mean for an estimator to be consistent?

Q2.3: What does it mean for an estimator to be efficient?

Q2.4: There are still more properties we might want our estimator to have. For example, why might we want to know what the shape of the estimator’s sampling distribution is?

---

To show that the coefficients in a multiple regression are unbiased, we need four assumptions, which are extension of the assumptions we learned for simple regression.

Q3.1: What are these four assumptions?

Q3.2: What are the implications of these four assumptions, both conceptually and mathematically?

## Associative versus Causal models

Q4.1 What is the difference between the following two conditions?

1.  $cov(x, \hat{u}) = 0$
2.  $cov(x, u) = 0$

### Associative models

There are (rare) times when we don't care about causality. Instead, we just want to know the line of best fit. That is, we want  $\beta_0$  and  $\beta_1$  that minimize,

$$E\left((Y - \beta_0 - \beta_1 X)^2\right)$$

Notice that we're talking about the population, not the sample here. That means that we can never find this line exactly.

In fact, if this line of best fit is all you can about, you could just *define* your errors,  $v$ , as the distance from this line. You can then prove that  $cov(x, v) = 0$ , the exogeneity assumption.

In Figure 1, we have a heatmap showing the joint distribution of umbrella sales and rainfall. This is a toy example that's problematic in many ways (do we have measurements across time? across locations?) but it helps to clarify what's happening.

Q4.2 How would you draw the line of best fit in the population on the Umbrella heatmap?

Q4.3 Suppose you collect a set of observations from this population model and fit an ols regression line to it. How does this line compare to the line above? What happens as  $n$  grows large?

Q4.4 Does the regression line represent the causal effect of umbrella sales on rain?

### Causal models

Two things to remember from the above discussion:

1. OLS always consistently finds the line of best fit.
2. But usually, the line of best fit is not actually the line we're interested in.

Causal modeling is a huge topic and we will explain just a little about how we think about it intuitively. Note that the Wooldridge text generally takes a causal approach to modeling.

In a causal/structural approach, we can't just define the errors to be whatever we want. The errors represent real-world factors that affect the outcome.

We believe that if we could just measure all the factors that are out there and put them into a regression equation (in the right way), our parameters will have a causal interpretation. In the umbrella example, the

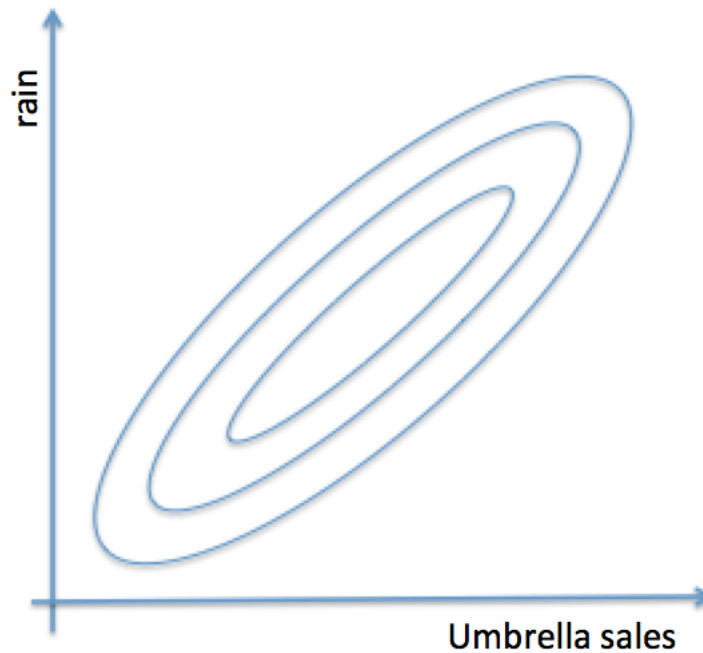


Figure 1: Predicting Rainfall from Umbrella Sales

problem would be that there are atmospheric factors that seem to be missing from the model. Suppose the real population model is the following,

$$rainfall = 0 \cdot umbrellas + 20 \cdot humidity - 10 \cdot airpressure + u$$

where  $u$  is uncorrelated with all other predictor variables. This model tells us that buying an extra umbrella has zero effect on rainfall.

Q4.5 How would you draw a line representing the causal effect of umbrellas on rainfall on the heatmap below?

Q4.6 Assuming that we haven't measured humidity or airpressure, can OLS regression estimate this line?

Since we didn't measure humidity or airpressure, they become part of the error in our simple regression.

$$rainfall = 0 \cdot umbrellas + w, \text{ where } w = 20 \cdot humidity - 10 \cdot airpressure + u$$

Q4.7 If you define error  $w$  as above, will  $cov(x, w)$  be 0, above 0, or below 0?

**Remember:** The exogeneity assumption  $cov(x, u) = 0$  is really about whether you have omitted variables in your regression.

## Model Diagnostics

Model diagnostic procedures, both graphical methods and formal statistical tests, are important for checking our model. These procedures allow us to explore whether the assumptions of the regression model are valid and decide whether we can trust subsequent inference results.

What is the value in examining a scatter plot for a regression analysis?

## Residuals Plots

A residuals plot can be used to assess the assumption that the variables have a linear relationship. The plot is formed by graphing the standardized residuals on the y-axis and the standardized predicted values on the x-axis. An optional horizontal line can be added to aid in interpreting the output.

## Influential Observations

Q4.1 What does it mean for a data point to have high leverage?

Q4.2 What does it mean for a data point to have high influence?

Q4.3 How is it possible for a data point to have low leverage but high influence?

Q4.4 How is it possible for an outlier to have low influence?

## R Exercise

The file `crime1.RData` contains individual-level data on criminal activity and punishment. It was used in the paper, J. Grogger (1991), "Certainty vs. Severity of Punishment," *Economic Inquiry* 29, 297-309.

You are interested in predicting the number of times an individual was arrested in 1986 (`narr86`). Your dependent variables will be the proportion of prior arrests that led to conviction (`pcnv`) and the average previous sentence length in months (`avgsen`).

```
load("crime1.RData")
desc

##      variable                                label
## 1   narr86                # times arrested, 1986
## 2   nfarr86               # felony arrests, 1986
## 3   nparr86              # property crime arr., 1986
## 4   pcnv      proportion of prior convictions
## 5   avgsen      avg sentence length, mos.
## 6   tottime  time in prison since 18 (mos.)
## 7   ptime86      mos. in prison during 1986
## 8   qemp86       # quarters employed, 1986
## 9   inc86       legal income, 1986, $100s
## 10  durat      recent unemp duration
## 11  black              =1 if black
## 12  hispan          =1 if Hispanic
## 13  born60          =1 if born in 1960
## 14  pcnvsq              pcnv^2
## 15  pt86sq             ptime86^2
## 16  inc86sq            inc86^2
```

1. First, you are interested in the simple regression of `narr86` on `pcnv`. Examine these variables and run a linear regression.
2. Generate a residuals vs. fitted values plot. Assess the assumptions for unbiasedness of OLS.
3. Generate a residuals vs. leverage plot. Assess whether any observations are exerting unusual influence on the regression coefficient.
4. Interpret your regression coefficients.

5. Next, you are interested in adding variable `avgsen` to your model. However, instead of fitting a linear model directly in R, use the regression anatomy formula to predict what the slope coefficient for `avgsen` would be in the multiple regression.

Recall that the coefficient on an independent variable  $x_i$  is equal to,

$$\beta_i = cov(y, r_i) / var(r_i)$$

where  $r_i$  is the residual from regressing  $x_i$  on all other independent variables.

- First, regress `avgsen` on `pcnv` and extract the residuals from this regression.
  - Next, regress `narr86` on the residuals from your first stage. What slope coefficient do you get?
6. Now, confirm your previous result by directly fitting the multiple regression model. That is, run the regression of `narr86` on both `pcnv` and `avgsen`. What is the coefficient on `avgsen`?
  7. Comment on the practical significance of your results.
  8. How do you explain the sign of the coefficient on `avgsen`? Explain whether this variable is exogenous. Can your coefficient have a causal interpretation?
  9. Use R to predict what number of arrests an individual would have, if half of their prior arrests led to convictions and their average sentence was 3 months.
  10. Which of your regression models fits the data better? Provide statistics to back up your response.
  11. Place the results from both models in a regression table.

### Cheat Sheet for checking bias assumptions

Linear population model: Nothing to check, because we haven't constrained the errors yet.

Random Sampling: Not interpretable from graphs. You need information about where the data comes from.

No perfect multicollinearity: This is a rare case that will trigger an error in R.

Zero-conditional mean: Check the Residual vs Fitted plot for a systematic departure from the horizontal line.