***Simple Regression—continued***

Regression

* examines the association between variables, and
* builds a “prediction” model

Last time, we looked at the association between Friday advertising expenditures and Saturday store traffic for a chain of yarn shops



Formula for the line (aka regression model:

Traffic = 65.71 people + .35 \* advertising $

* When there is $0 advertising on Friday, 65.71 people on average came into the store on Saturday
* For every $1 increase in advertising, .35 more people come into the store

**Is the regression model valid?**

In this analysis, we saw that in the sample of 20 yarn shops there’s an association between advertising and store traffic

However, just because this association exists in the sample doesn’t mean it also exists in the population of all the yarn shops in the chain

To check on that, we need to run a significance test

In addition, regression has a lot of technical requirements that we’re not going to cover in this class

A regression model is valid only if

• the technical requirements have been met, and

• the model is statistically significant

In this class, we will assume that a statistician has already done all the technical and significance tests and is showing you a regression model that’s accurate and statistically significant

If you want to learn how to run regression analysis yourself, you’ll need advanced training, like that offered in MRKTNG 4900 or STAT 3500

**Another example: soft-serve ice cream**

The dining services manager for a state university wants to know if she can estimate the amount of soft-serve ice cream students will consume in the dining halls, based on the outdoor temperature.

dependent variable = ounces of soft-serve ice cream dispensed per student per day

independent variable = outdoor temperature at 2:00 pm (degrees Fahrenheit)

Data are for 25 randomly sampled days.

Scatterplot of the data



Scatterplot with regression line



Formula for the regression line:

Ounces Dispensed = -1.31 + .12 \* degrees Fahrenheit

According to the formula, how many ounces were dispensed when the outdoor temperature was

• 60 degrees?

• 35 degrees?

• 0 degrees?

Oops!! We can’t dispense a negative amount of ice cream....

**Relevant range of a regression model**

The regression model can only be applied in circumstances that are similar to the data on which it is based

The data for this regression model spans temperatures of 14° to 75°. We should use it to make estimates of soft-serve ounces for temperatures within that range.

The model is ***not valid*** for temperatures below 14° (such as zero!) or above 75°.

**Can we make predictions from a regression model?**

Last time, we learned that regression models describe the past and do not predict the future.

However, in some cases regression models can be useful in making estimates of future behavior. For this reason, regression models are widely used by businesses to make many different kinds of estimates about the future.

These estimates will only be useful if the future is similar to the past. If the environment or circumstances change, the model estimates will be incorrect.

In our example, suppose that the dining service stopped offering cookies as a dessert option after the regression model was developed. This is an important environmental change, because more students will probably choose ice cream as their dessert option or serve themselves larger ice cream portions if cookies are no longer available to them. In this circumstance, the regression model will under-estimate soft-serve consumption.

What are some other examples of environmental changes that might make the regression model inaccurate?

**Model accuracy (model fit)**

Regression models vary widely in how well they fit actual data.

Consider the two scatterplots below:





In the first scatterplot, most of the dots are on or close to the regression line

Remember: the dot represents what actually happened; the regression line is the model’s guess about what happened

The distance between the dot and the line is called “error” (or mis-estimate)

In the top model, error is pretty small, and estimates are pretty close to what actually happened

In the bottom model, there is a lot more distance (on average) between the dots and the line, so there’s a lot more error

An indicator of how much error there is in a regression model is **R2**

When R2 = 1.0 or R2 = -1.0, there is no error in the regression model

The smaller the value of R2, the more error there is

The larger the value of R2, the better the model fits the data

R2 indicates how much variation in the dependent variable is explained by the independent variable

In our soft-serve ice cream model, R2 = .48, which means that 48% of the variation in amount of ice cream dispensed each day is accounted for by variations in outdoor temperature.

The remaining variation in amount of ice cream dispensed is due to other factors not included in the regression model.

Some of these other factors might include:

• day of the week

• whether it’s midterms or final exams week

• whether the ice cream dispenser is malfunctioning or unable to keep up with demand

• and a lot of factors unique to the day, such as whether the football team just lost a big game, the student health center just launched a campaign to encourage students have a more healthy diet, or BTS just released a video showing all the members eating soft-serve ice cream

**Practical application of the regression model:**

In the sample, for each 1 degree increase in temperature, students consumed .12 more ounces of soft-serve

If you’re a dining services manager, you can estimate demand for soft-serve for a given day with this formula:

Total ounces of soft-serve = (-1.31 + .12(temperature)) \* # of dining hall users

For example, if the average number of users per day is 12,534 and the predicted 2:00 pm temperature is 74 degrees, this is how many ounces of soft-serve is likely to be consumed by students:

Total ounces of soft-serve = (-1.31 + .12(74)) \* 12,534

= 94,882 ounces of soft-serve, which is 741 gallons (about 2.8 tons)

This estimate will not be exact because the model contains error. Advanced applications in regression allow the analyst to calculate the maximum amount of likely error in an estimate.

**Remember, regression models cannot predict the future**

But if the environment in the future is similar to the environment when the model was created, the model can be useful in preparing estimates

**Some more regression models to interpret**

**Online retail: The association between return shipping cost and customer returns**

Customer returns are a big headache for online clothing retailers. In many cases, it’s cheaper for the retailer to sell the returned items in bulk at a greatly discounted price to a liquidation service than it is to restock the merchandise to sell at full price to consumers.

One way to reduce returns might be to make it more difficult for customers to return items they’ve ordered, such as making them pay the full cost of return shipping rather than allowing free or low-cost returns.

Research question: Are higher return costs associated with the number of items returned?

Dependent variable: Percentage of purchases by an online retailer that are returned by the customer

Independent variable: Cost to consumer to return an item

The data in this example are hypothetical, but an actual analysis could be done using data from online marketplaces.

Scatterplot with regression line



Regression model

Percentage returned = 8.93 - .52 \* return shipping cost

Applying the regression model:

• For each increase of $1.00 in shipping cost, what is the change in percentage returned?

• Calculate the estimate for percentage returned when the shipping cost is $4.50

Model accuracy (fit):

R2 for this model =.24, meaning that 24% in the variation in returns among online clothing retailers is explained by the retailers’ return shipping costs. The remaining 77% of variation in returns is due to other factors not included in the regression model.

Something to think about:

If lower shipping costs are associated with more returns, why do online retailers offer free returns for their products?

**Another example: Canned cat food sales**

Research question: Are sales of canned cat food in a grocery store associated with the size of the special display for the brand?

dependent variable = sales (# of cans sold)

independent variable = display size (in square feet)

Regression model

Cans sold = 21.69 + 58.98 \* number of square feet

Applying the regression model:

• For increase of 1 square foot in display size, what is the change in number of cans sold?

• Calculate the estimate for number of cans sold when the special display is 2.8 square feet in size

**Last example: Time elapsed between order placement**

A manufacturer of medical supplies for doctors’ offices is trying to model buyer behavior. One question the company has is whether the size of a doctor’s office is associated with how often the office places an order.

dependent variable = time between orders (in days)

independent variable = size of purchaser (measured by number of M.D.s employed)

Regression model

Elapsed time = 20.27 - 0.98 \* number of M.D.s employed

Applying the regression model:

• For increase of 1 M.D. in a medical office, what is the change in number days between orders?

• Calculate the estimate for number of days between orders when a medical office employs 4 M.D.s