STA 111: Probability & Statistical Inference

Extra Topics – Designed Experiments and Observational Study; Multicollinearity, Variable Selection and Cross-Validation

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Designed Experiments
- Observational Study
- Weighted Averages
- Nonparametric Regression
- Multicollinearity
- Variable Selection
- Cross-Validation

Introduction to DEs and OSs

- DE - Designed Experiments
- OS - Observational Study
- The usual purpose of both kinds of research is to draw conclusions about causation. For example:
  - Does smoking cause cancer?
  - Does premarital sex cause higher divorce rates?
  - Does college partying cause low grades?
- A double-blind, randomized, controlled experiment gives more accurate conclusions than an observational study.

Double-Blind, Controlled and Randomized

- The gold standard for a statistical study is the double-blind, randomized, controlled experiment.
- A study is double-blind if neither the subjects nor the scientists know who is assigned to which group until after the data are collected. This prevents subjects in different groups from behaving in different ways; prevents scientists from introducing any unconscious bias into the data collection process.
- A study is controlled if one group receives the treatment and another group does not. (In medicine, that group usually gets either a placebo, or standard medical care, or both.)

Double-Blind, Controlled and Randomized (Cont'd)

- A study is randomized if the control group and the treatment group are chosen at random.
- Without randomization, the groups may differ in a systematic way. For example, surgeons used to assign only the healthiest patients to receive an experimental new surgical treatment, since those patients could best withstand the invasive procedure. But the outcomes for those patients are not a reliable forecast for how normal patients would respond.

Double-Blind, Controlled and Randomized (Cont'd)

- Historical controls do not give a randomized experiment, which is one reason their use is problematic. The FDA is very reluctant to approve drugs in which all patients in the trial receive the drug, while the control group are patients who were treated before the drug was invented. One concern is that the standard of basic care constantly improves, so the drug may appear effective when, in fact, the only difference is that current patients get, say, better nursing care.

Double-Blind, Controlled and Randomized (Cont'd)

- Studies of the portacaval shunt, a treatment for cirrhosis of the liver, is telling. Physicians reported 50 experiments on the procedure in the medical literature (most of these experiments were small, involving only about ten or so patients).

	Degree of Enthusiasm		
	High	Moderate	Low
Design			
No Control	24	7	1
Control, Not Randomized	10	3	2
Randomized, Controlled	0	1	3

Confounding Factors

- In an observational study, the researcher does not get to determine who receives the treatment. For example, people who smoke get lung cancer at a higher rate than those who do not smoke. Does smoking cause lung cancer?

  The tobacco lobby used to say no, arguing that: there might be a gene that predisposes people to both enjoy smoking and get cancer; people who like to smoke may tend to follow unhealthy lifestyles (e.g., alcohol use), and that may be the real cause of lung cancer; no randomized, controlled, double-blind experiment (on humans) has shown causation.

Confounding Factors (Cont'd)

- Obviously, it would be ethically problematic to do a randomized controlled experiment (one would have to assign 14 year-olds at random to smoke heavily for the rest of their lives). And it would be hard to make this double-blind—people know if they smoke.

  But animal studies strongly indicate that smoking causes lung cancer in mammals and birds.

- The other two arguments from the tobacco lobby carry more weight. The differences between lung cancer rates in the smokers and non-smokers may be due to smoking, or they may be due to a confounding factor or variable.

  In this case, tobacco lobbies suggested two possible confounding factors: genes and lifestyle.

Confounding Factors (Cont'd)

- A confounding factor is associated with both:
  - outcome
  - group membership

  For example, one might argue that lung cancer is caused by matches, not tobacco.

  Similarly, one might argue that cholesterol does not cause heart disease, but rather is a result of poor circulation or breakdown of heart muscle tissue—so it is associated, but not causal.

Confounding Factors (Cont'd)

- One way to try to handle confounding is to make subgroup comparisons that *control* for possible confounding effects. For example, one could compare the lung cancer rates for smokers who use matches against smokers who use lighters.
- Do seatbelts save lives?

  Seatbelt studies are usually observational (*why*?). One compares the fatality rates in accidents in which seatbelts were worn to the fatality rate in accidents without seatbelts.

  But one has to worry about confounding factors. For example,

  — People who don't wear seatbelts may drive more recklessly.

  — People who don't wear seatbelts may prefer cars that are not designed with safety in mind.

Confounding Factors (Cont'd)

- Some researchers try to control for this by comparing the fatality rates among seatbelt wearers and non-wearers in similar cars, or cars that are thought to have been traveling at the same speed. But this is awkward to do and invites criticism.
- In order to control for a confounding factor, one has to guess what it is. But that can be hard and you are never sure that you have thought of everything.

Confounding Factors (Cont'd)

- In contrast, with a randomized design, the random assignment of people to the treatment and control groups ensures that there is almost no chance of a systematic difference between the groups. You are unlikely to get most of the safe drivers in one group and the reckless in the other, or most of the people with good genes for lung cancer in one group and all those with bad genes in the other.
- Health experts say that exercise increases one's lifespan. What kinds of data might they have, and what would be the statistical issues regarding the validity of their claim?

Weighted Averages

- Subgroup analysis is one way to control for a potential confounding factor. Here one studies each group defined by the confounder separately. Another way to control for a confounder is to use a weighted average.
- In the 1960s, the University of California at Berkeley was embarrassed. It was rejecting a larger proportion of women than men, and applicants claimed there was gender bias. But when the Dean asked each department to report their admission rates separately, it turned out that each department accepted a larger proportion of women than men. (The Dean was doing a subgroup analysis without realizing it.)

Weighted Averages (Cont'd)

This apparent reversal of a pattern is sometimes called Simpson's Paradox. It happens when there is a third confounding variable (major) which affects the other two (admission and gender).

The Dean asked Professor Betty Scott to study the problem. She showed that women tended to apply to the majors that were most selective, whereas the men applied to majors that were less selective. So overall, the women had higher rejection rates.

Weighted Averages (Cont'd)

To put such comparisons on a fair footing, she calculated the weighted average admission rates for women and men, where the weights are determined by the proportion of people applying to each of the different majors. This controls for the confounding variable.

To see how the weighted average works, we focus on just two majors. Assume major A accepts 80% of all applicants, but Major B accepts just 10%. Suppose 100 men and 200 women apply. Consider two scenarios:

Weighted Averages (Cont'd)

Scenario 1: Half the men and half the women apply to A, the rest apply to B.

Scenario 2: 90 men apply to A, 10 to B; but 180 women apply to B, 20 to A.

In the first case, major is not a confounding variable. Men and women show the same major preferences. (Note: They do not have to apply in 50-50 ratios—it would still not be a confounder if both genders applied in 25-75 ratios, for example.)

In the second case, major is a confounder. Men prefer A, but women prefer B.

Weighted Averages (Cont'd)

In Scenario 1, the raw number of men who are accepted is

$$.8 * 50 + .1 * 50 = 45$$

and for women the percentage is the same: (80+10)/200 is 45%.

In Scenario 2, the raw number of men who are accepted is

$$.8 * 90 + .1 * 10 = 73$$

or 73%. And the raw number of women accepted is

$$.1 * 180 + .8 * 20 = 34$$

so their acceptance rate is 34/200 or 17%. This looks like gender bias, but actually it is not—the admission policy is completely gender blind.

Weighted Averages (Cont'd)

To make a fair comparison, weight the acceptance rates for men in each major by the fraction of people applying to that major:

$$\frac{90+20}{300} * \frac{72}{90} + \frac{10+180}{300} * \frac{1}{10} = .357$$

and the weighted average proportion of women accepted is

$$\frac{90+20}{300} * \frac{16}{20} + \frac{10+180}{300} * \frac{18}{180} = .357$$

The weighted average shows that the acceptance rates for men and women, controlling for major, are equal.

Weighted Averages (Cont'd)

The general formula for finding the weighted average correction for the acceptance rate of men is:

$$\text{wtd avg} = \sum_i \quad (\text{prop. of people applying to major } i) * \\ (\text{acceptance rate for men at major } i)$$

Recap

Recall that the multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{V}[\epsilon_i] = \sigma^2$, and the ϵ_i's are independent.

The model is useful because:

- it is interpretable—the effect of each explanatory variable is captured by a single coefficient
- theory supports inference and prediction is easy
- simple interactions and transformations are easy (how?)
- dummy variables allow use of categorical information
- computation is fast.

Nonparametric Regression

We extended the multiple linear regression model to nonlinear regression, in which we fit a model of the form:

$$y_i = \beta_0 + \beta_1 g_1(x_{i2}) + \ldots + \beta_p g_p(x_{ip}) + \epsilon_i$$

where the g_j's are known transformations of the data, such as the log or $1/x$, and, as before, $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{V}[\epsilon_i] = \sigma^2$, and the ϵ_i's are independent.

This model can be further extended to nonparametric regression, in which case one does not now the functions $g_1, \ldots, g_p$ but instead must estimate these by smoothing the data.

In the real applications, the linear regression model is usually only a locally correct approximation. And it is rare that one has a strong theoretical model that prescribes specific nonlinear transformations. Thus nonparametric regression is a practical tool in many cases.

Curse of Dimensionality

Regression becomes much harder as the number of explanatory variables increases. This is called the Curse of Dimensionality (COD). The term was coined by Richard Bellman in the context of approximation theory.

The COD applies to all multivariate regressions that do not to impose strong modeling assumptions—especially the nonparametric regressions, but also those in which one tests whether a specific variable or transformed variable should be included in the model.

In terms of the sample size n and dimension p, the COD has three nearly equivalent descriptions:

- For fixed n, as p increases, the data become sparse.
- As p increases, the number of possible models explodes.
- For large p, most datasets are multicollinear.

Curse of Dimensionality (Cont'd)

To explain the model explosion aspect, suppose we restrict attention to just linear models of degree 2 or fewer. For $p = 1$ these are:

$$\begin{array}{lll} \mathbb{E}[Y] = \beta_0 & \mathbb{E}[Y] = \beta_1 x_1 & \mathbb{E}[Y] = \beta_2 x_1^2 \\ \mathbb{E}[Y] = \beta_0 + \beta_1 x_1 & \mathbb{E}[Y] = \beta_0 + \beta_2 x_1^2 & \mathbb{E}[Y] = \beta_1 x_1 + \beta_2 x_1^2 \\ \mathbb{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 & & \end{array}$$

For $p = 2$ this set is extended to include expressions with the terms $\alpha_1 x_2$, $\alpha_2 x_2^2$, and $\gamma_{12} x_1 x_2$. For general p, combinatorics shows that the number of possible models is

$$2^{1+2p+\left(\begin{array}{c} p \\ 2 \end{array}\right)} - 1.$$

This increases superexponentially in p, and there is not enough sample to enable the data to discriminate among these models.

Curse of Dimensionality (Cont'd)

For the multicollinearity issue, we note that **multicollinearity** occurs when two or more of the explanatory values are highly correlated. This implies that the predictive value of the fitted model breaks down quickly as one moves away from the subspace in which the data concentrate.

We shall agree that multicollinearity occurs whenever the absolute value of the correlation between two of the explanatory variables exceeds 0.9. But this is a judgment call, and one can have multicollinearity that arises in more complex ways.

For large p with finite n, it is almost certain that two explanatory variables will have high correlation, just by chance.

Variable Selection

One wants to select a multiple regression model that only includes useful variables. Some methods are:

- Forward Selection. One starts with no variables in the model, and sequentially adds the one that best explains the current residuals (or the raw data, at the initial step). One stops when none of the remaining variables provide significant explanation.
- Backwards Elimination. Start with all the variables in the model, and sequentially removes the variable that explains the least, until a t-test shows that no further variables should be removed.
- Stepwise Regression. Alternate use of forward selection and backwards elimination.

None of these is bulletproof.

Cross-Validation

To assess model fit in complex, computer-intensive situations, the ideal strategy is to hold out a random portion of the data, fit a model to the rest, then use the fitted model to predict the response values from the values of the explanatory variables in the hold-out sample.

This allows a straightforward estimate of the error in prediction using regression. But we usually need to compare fits among *many* models. If the same hold-out sample is re-used, then the comparisons are not independent and (worse) the model selection process will tend to choose a model the overfits the hold-out sample, causing spurious optimism.

Cross-Validation (Cont'd)

Cross-validation is a procedure that balances the need to use data to select a model and the need to use data to assess prediction.

Specifically, v-fold cross-validation is as follows:

- randomly divide the sample into v portions;
- for $i = 1, \ldots, v$, hold out portion i and fit the model from the rest of the data;
- for $i = 1, \ldots, v$, use the fitted model to predict the hold-out sample;
- average the predictive mean squared error (PMSE) over the v different fits.

One repeats these steps (including the random division of the sample!) each time a new model is assessed.

The choice of v requires judgment. Often $v = 10$.