CPH 636 – Spring 2009 – Dr. Charnigo

Lecture 2

Variable types

Beyond continuous and categorical. We often think of variables as either continuous or categorical, yet further distinctions can be made. Fernandez addresses this on page 17, but his definitions are not standard; hence, I use the definitions in Chapter 9 of Rosner. [[The title of this introductory text, used in STA 570 and STA 580, is *Fundamentals of Biostatistics*, Sixth Edition (Thomson, 2005).]]

Nominal variables. We have a nominal variable (or "nominal data") if: the possible values are either categories or artificial numerical designations for categories; and, the categories do not have a natural ordering.

Ordinal variables. We have an ordinal variable (or "ordinal data") if: the possible values are either categories or artificial numerical designations for categories; and, the categories have a natural ordering; but, differences between possible values are not meaningful.

Cardinal variables. We have a cardinal variable (or "cardinal data") if the differences between possible values are meaningful. If the quotients of nonzero possible values are also meaningful, the cardinal variable is called a ratio variable; otherwise it is called an interval variable.

Discussion questions. Classify each of the following as nominal, ordinal, interval, or ratio: number of children in a third-grade class who have a cold; gender; outdoor temperature in degrees Fahrenheit; response to a survey item for which the possible answers are "Strongly Disagree", "Disagree", "Agree", "Strongly Agree". What is the relationship between continuous variables and cardinal variables?

Bias-variance tradeoff

Scenario. Suppose that people in a certain population are classified according to whether they have "Feature A" (e.g., person does not exercise regularly) and whether they have "Feature B" (e.g., person was born in January, March, May, July, September, or November).

Suppose, moreover, that 50% of people in the population have Feature A, 50% have Feature B, and 25% have both Feature A and Feature B. [[Thus, in probabilistic parlance, whether a person has Feature A is "independent" of whether a person has Feature B.]]

A researcher randomly selects 25 people with both Feature A and Feature B, 25 people with Feature A but not Feature B, 25 people with Feature B but not Feature A, and 25 people with neither Feature A nor Feature B. The researcher then obtains a systolic blood pressure measurement for each person in the sample. [[The researcher's approach is referred to as "stratified random sampling", which is different from "simple random sampling" because the researcher did not just randomly select 100 people without regard to Feature A and Feature B. Had the researcher done so, the <u>expected</u> number of people selected within each stratum would have been 25, but the <u>actual</u> number might have been more or less than 25.]]

Unknown to the researcher, systolic blood pressure is characterized by normal distributions whose means and variances are given in Table 1. Note from Table 1 that Feature A, not Feature B, determines the means of the normal distributions.

Table 1:

Subpopulation	Mean	Variance
A, B	135	100
A, not B	135	100
not A, B	125	100
not A, not B	125	100

A prediction problem. Suppose that, after obtaining a systolic blood pressure measurement for each person in the sample, the researcher is asked to make a prediction for a new person not in the sample. This new person is known to have both Feature A and Feature B. Let us denote the new person's systolic blood pressure by SBP_{new}. Consider the following four "prediction rules" that are available to the researcher.

1. Let SBP_{all} denote the mean for all 100 people in the sample. Use SBP_{all} as the predictor of SBP_{new}. Note that using SBP_{all} disregards the information we have about the new person (namely, that this person has Feature A and Feature B).

2. Let SBP_A denote the mean for the 50 people in the sample with Feature A. Use SBP_A as the predictor of SBP_{new}. Note that using SBP_A takes into account some but not all of the information we have about the new person.

3. Let SBP_B denote the mean for the 50 people in the sample with Feature B. Use SBP_B as the predictor of SBP_{new}. Note that using SBP_B takes

into account some but not all of the information we have about the new person.

4. Let SBP_{AB} denote the mean for the 25 people in the sample with both Feature A and Feature B. Use SBP_{AB} as the predictor of SBP_{new}. Note that using SBP_{AB} takes into account all of the information we have about the new person.

Table 2:

Numerical summary	SBP_{all}	SBP_A	SBP_B	SBP_{AB}
1. Expected value of predictor	130	135	130	135
2. Expected value of target	135	135	135	135
3. Bias of predictor [1 - 2]	-5	0	-5	0
4. Squared bias of predictor	25	0	25	0
5. Variance of predictor	1	2	2	4
6. Variance of target	100	100	100	100
7. Mean square error of prediction [4 + 5 + 6]	126	102	127	104

Which predictor is best? Table 2 compares the four prediction rules.

First, we note that the expected values of SBP_{all} and SBP_B differ from the expected value of SBP_{new}. Because of this, we say that SBP_{all} and SBP_B are biased; more specifically, the bias is quantified as the difference between the expected value of the predictor and the expected value of the target. Intuitively, such bias arises because SBP_{all} and SBP_B fail to use the highly relevant information that the new person has Feature A. [[Details: Let $\bar{X}_{AB}, \bar{X}_{AN}, \bar{X}_{NB}, \bar{X}_{NN}$ denote respectively the mean for the 25 people in the sample with Features A and B, the mean for the 25 people in the sample with Feature A only, the mean for the 25 people in the sample with Feature B only, and the mean for the 25 people in the sample with neither Feature A nor Feature B. Then $\bar{X}_{AB} \sim N(135, 4), \bar{X}_{AN} \sim N(135, 4), \bar{X}_{NB} \sim N(125, 4), \bar{X}_{NN} \sim N(125, 4)$ with $\bar{X}_{AB}, \bar{X}_{AN}, \bar{X}_{NB}, \bar{X}_{NN}$

independent. Hence, $SBP_{all} = 0.25\bar{X}_{AB} + 0.25\bar{X}_{AN} + 0.25\bar{X}_{NB} + 0.25\bar{X}_{NN} \sim N(0.25 \times 135 + 0.25 \times 135 + 0.25 \times 125 + 0.25 \times 125, 0.25^2 \times 4 + 0.25^2 \times 4 + 0.25^2 \times 4 + 0.25^2 \times 4) = N(130, 1)$. The other expected values (and variances) can be worked out similarly.]]

Second, we note that SBP_{all} has smaller variance than the other predictors, while SBP_{AB} has the largest variance. Intuitively, this is because SBP_{all} is based on 100 observations, while SBP_{AB} is based on only 25 observations.

Third, the mean square error of prediction for SBP_{all} is defined as $E[(SBP_{all} - SBP_{new})^2]$, the expected value of the squared difference between the predictor and the target. The mean square errors for SBP_A, SBP_B, and SBP_{AB} are defined analogously. Mean square error is an overall measure of predictive accuracy and can be obtained by adding the squared bias of the predictor, the variance of the predictor, and the variance of the target. Clearly, SBP_A is best, though SBP_{AB} is almost as good. On the other hand, SBP_{all} fares quite poorly, and SBP_B is even worse. [[Details: We have $E[(SBP_{all} - SBP_{new})^2] = Var[SBP_{all} - SBP_{new}] + (E[SBP_{all} - SBP_{new}])^2 = Var[SBP_{all}] + Var[SBP_{new}] + (E[SBP_{all}] - E[SBP_{new}])^2$. The first equality rearranges the definition of variance, $Var[X] = E[X^2] - (E[X])^2$. The second equality invokes (additive) linearity of the variance operator for two independent random variables, the fact that $Var[X] = Var[-X]$, and linearity of the expectation operator for two arbitrary random variables.]]

In summary, we pay dearly if we do not use genuinely relevant information (i.e., that the new person has Feature A), but we also incur some cost if we use extraneous information (i.e., that the new person has Feature B).

More on the tradeoff. The example just described was a bit simplistic; this was necessary for the convenient derivation of numerical summaries in Table 2. However, the bias-variance tradeoff is a rather general phenomenon in statistical modeling (Hastie et al, page 37). A more sophisticated example, albeit one for which I cannot readily provide an analogue to Table 2, concerns the selection of explanatory variables for linear regression.

Suppose that I have available $X_1, \ldots, X_{10}$ but that only X_1, X_2, X_3 are truly relevant. If I leave out any of X_1, X_2, X_3, then I will have a problem with bias: my predictions for new people may be systematically understated or overstated. On the other hand, if I include several of $X_4, \ldots, X_{10}$, then I will have a problem with variance: my predictions for new people will not be systematically understated or overstated, but I may be prone to really large errors in either direction. [[This is related to the phenomenon of multicollinearity, which is a sort of extreme case. When the explanatory variables in a linear regression model are highly correlated, so that (arguably) some of them are automatically extraneous, the standard errors of the parameter estimators become very large. Since the predictions are linear combinations of the parameter estimators, via $\widehat{Y} = \widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_j$, the predictions are themselves volatile.]]

Curse of dimensionality

Live demonstration. In anticipation of this lecture, I asked you to complete a personality test that would identify you as introverted ("I") or extroverted ("E"), intuitive ("N") or sensing ("S"), thinking ("T") or feeling ("F"), and perceiving ("P") or judging ("J").

To begin with, let's have all the I's stand on one side of the classroom and the E's stand on the other. Next we'll have the I's and E's separate into smaller clusters of IN's, IS's, EN's, and ES's. We'll continue with this process as best we can, given the small room size and the fact that we can only move in two dimensions. What I hope will become clear is that, for many of us, as more characteristics are considered, there are fewer people who are similar to us.

Defining the curse. We have seen that four dichotomous variables separate each one of us from several other people in the classroom. Now imagine what would happen if we considered five variables, or ten, or twenty — especially if they were continuous! Then almost certainly each one of us would be separated from all other people in the classroom.

Further perspective can be acquired by trying to distribute nine points evenly on the interval $[0, 1]$ and on the square $[0, 1] \times [0, 1]$. For the interval $[0, 1]$, we might place the nine points at 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Each point is separated from its nearest neighbor(s) by 0.1 units. For the square $[0, 1] \times [0, 1]$, we might place the nine points at $(0.25, 0.25)$, $(0.25, 0.50)$, $(0.25, 0.75)$, $(0.50, 0.25)$, $(0.50, 0.50)$, $(0.50, 0.75)$, $(0.75, 0.25)$, $(0.75, 0.50)$, and $(0.75, 0.75)$. Now each point is separated from its nearest neighbor(s) by 0.25 units. Moreover, if we wanted to change the 0.25 back to 0.1, we would need to increase the number of points from 9 to $81 = 9^2$.

These considerations motivate us to define the curse of dimensionality (Hastie et al, page 22) as follows: As the number of variables increases, the data points become markedly more separated or spread out unless we dramatically increase the sample size each time we add a variable.

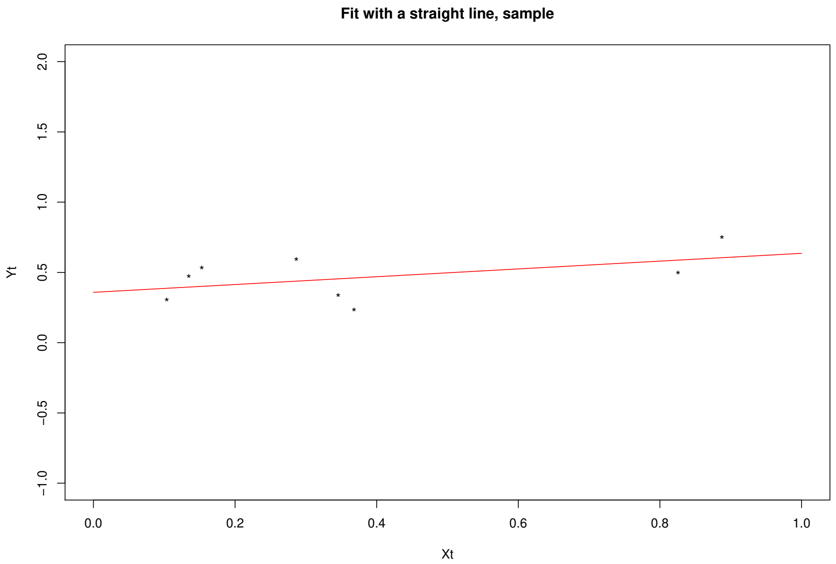
Practical implications of the curse. One practical implication is that, when we have a sample with a large number of explanatory variables and a single continuous response variable, our statistical modeling will require a strong assumption about the mean response. For instance, we may assume that the mean response is linear in the explanatory variables (linear regression) or that it is piecewise constant in the space defined by the explanatory variables (regression tree). Such a strong assumption is necessary because otherwise there is no way to make meaningful predictions for new people; new people are likely to be separated or spread out from the people in the original sample. [[Similar remarks apply when we have a single dichotomous response variable, with the substitutions of "log odds of event", "logistic regression", and "classification tree" for "mean response", "linear regression", and "regression tree".]]

A second practical implication is that, with a large number of explanatory variables, most observations are outliers. [[To illustrate, suppose that $X_1, X_2, \ldots, X_k$ are independent and that each has the uniform distribution on $[0, 1]$. That is, $P(X_1 \leq t) = t$ for $0 \leq t \leq 1$ and similarly for $X_2, \ldots, X_k$. Let us agree that we have an outlier if at least one of $X_1, X_2, \ldots, X_k$ is less than 0.05 or greater than 0.95. At $k = 1$, the probability that we have an outlier is $P(X_1 < 0.05) + P(X_1 > 0.95) = 0.10 = 1 - 0.90^1$. At $k = 5$, the probability increases to $1 - 0.90^5 = 0.410$. At $k = 20$, the probability reaches $1 - 0.90^{20} = 0.878$.]]

Training, validation, and test data

A motivating illustration. Consider Figure 1, which shows the result of fitting a simple model to a set of eight data points. The simple model reflects a belief that the vertical coordinate of each data point (viewed as the response variable) should be a linear function of the horizontal coordinate (viewed as the explanatory variable) plus some random error.

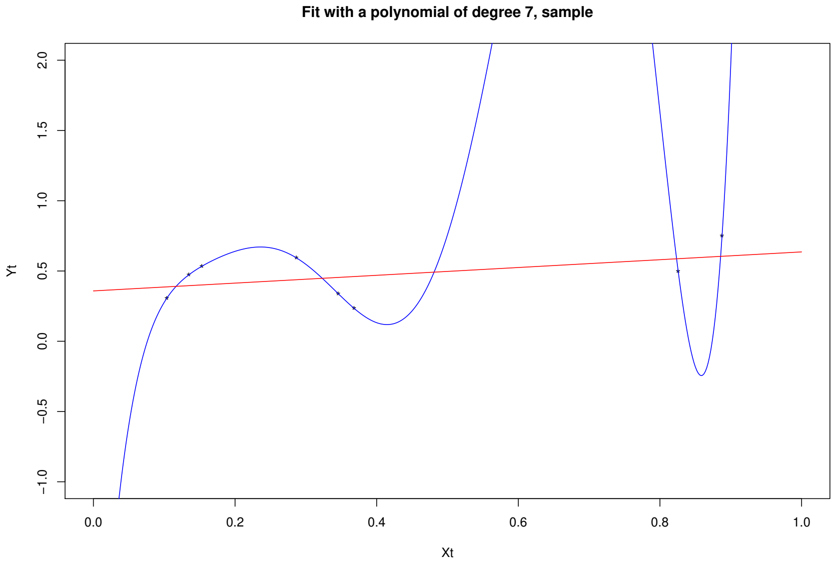
Figure 1:



Next consider Figure 2, which shows the result of fitting a complicated model to the same eight data points. Note that the complicated model fits the eight data points perfectly.

Now suppose that the eight data points are just a sample from a larger population. Figure 3 shows that the simple model provides a decent description of the relationship between the response variable and the explanatory variable in the population. Figure 4 shows that the complex model is a complete failure when we generalize from sample to population.

Figure 2:



The problem: double use of sample data. Two key points are as follows.

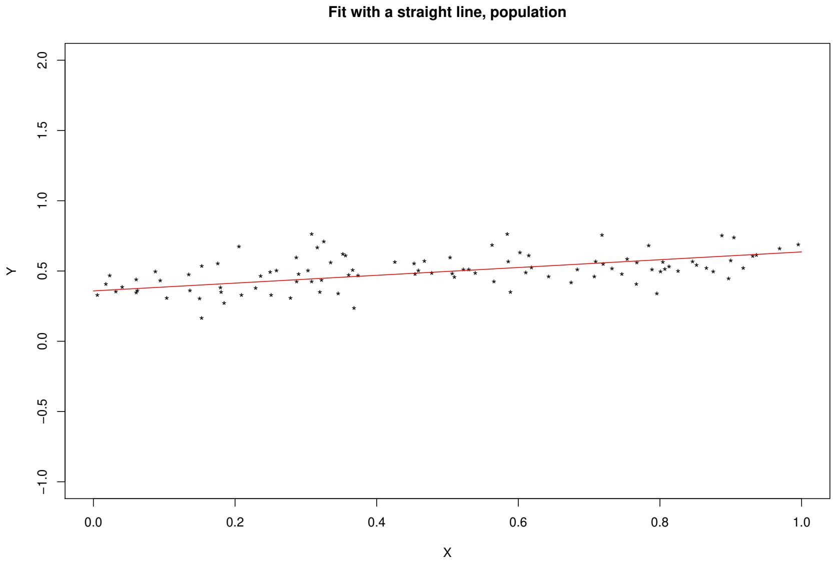
1. If you use the same data set twice, both to fit a model and then to evaluate it, your evaluation will be optimistic: the model will tend to appear better than it really is.

2. The optimism increases as the model becomes more complex.

For a simple model, the optimism is mild. The quality of the fit in Figure 1 is not visibly different from the quality of the fit in Figure 3. For a complex model, the optimism is considerable. The quality of the fit in Figure 2 <u>is</u> visibly different from the quality of the fit in Figure 4. Since the quality of the fit in Figure 4 is what we really care about, we feel badly misled by the considerable optimism in Figure 2.

Note that we face a challenge not just in <u>model evaluation</u> but also in <u>model selection</u>. In particular, we must ask ourselves: how much better does a complex model have to fit the sample data for us to choose it over

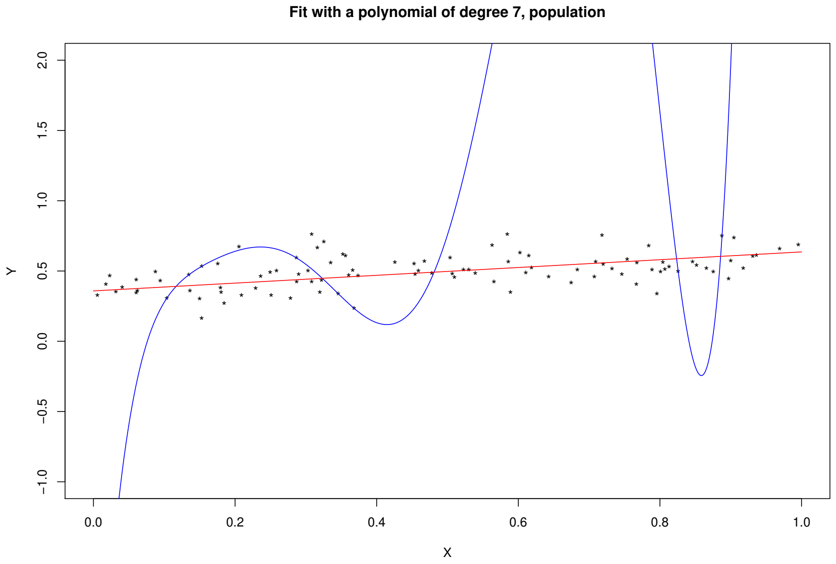
Figure 3:



a simple model? [[The R^2 inflation phenomenon in linear regression illustrates this idea. Recall that R^2 (almost always) goes up when you add more explanatory variables to a linear regression model. Yet, because R^2 is calculated from the same data that have been used for model fitting, R^2 will be optimistic. Since the optimism will increase as more explanatory variables are added, R^2 becomes especially untrustworthy when there are a lot of explanatory variables. In particular, R^2 should never be used to judge between competing models with different numbers of explanatory variables.]]

One solution: training, validation, and test data. Suppose that we want both to select a statistical model and to evaluate it. If we have a large data set, one solution to the problem described above is to divide the data set into three parts: a training subset, a validation subset, and a test subset. Fernandez mentions this on page 17, although he is less than clear about what the three subsets are used for. He is more explicit in the summary on

Figure 4:



page 39, but what he says there does not accord with what leading statisticians advocate. What follows is based on pages 195 and 196 of Hastie et al.

The training subset is used to fit various competing models. The validation subset is used to judge between and ultimately choose from among the competing models, as the training subset would tend to exhibit inappropriate favoritism to the more complex models. The test subset is then used to evaluate the chosen model, as the training subset would tend to evaluate the chosen model too favorably. In summary, the first key point on page 10 is what renders the training subset unsuitable for model evaluation, while the second key point is what renders the training subset unsuitable for model selection.

Dividing the data set into three parts thus permits model fitting, model selection, and model evaluation while circumventing the problem described above.

Hastie et al offer $50\%, 25\%, 25\%$ as a guideline for the creation of training, validation, and test subsets. However, they caution that this may not work in all situations, especially if we have a data set of only moderate size.

Incidentally, you may wonder why a validation subset cannot be used for both model selection and model evaluation. The reason is that an evaluation based on the validation subset may be a bit too optimistic. Why this should be so is not immediately obvious; I will try to provide some intuition through a metaphor. Suppose that I vote to elect politician "Dollar" Bill but that you vote for his opponent. If "Dollar" Bill is elected, which one of us do you think will more favorably evaluate his performance while he is in office? Most likely I will. Similarly, a validation data set that has encouraged us to select a particular statistical model will tend to evaluate that model too favorably.

A second solution: training and test data. Suppose that we have a data set of only moderate size but that we still want both to select a statistical model and to evaluate it. There exist model selection criteria that, when applied to the training subset, mimic the validation step. This obviates creation of a validation subset. [[From your previous coursework you may recognize the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).]] However, we still create a test subset to evaluate the chosen model.

The RANSPLIT macro. This macro can be used to divide a SAS data set into training, validation, and test subsets. You have the option to create only training and test subsets; however, the training and test subsets will be named "TRAIN" and "VALID" instead of "TRAIN" and "TEST". See {SASMacroInstr.txt} for guidance in using the macro.