# CPH 636 – Spring 2009 – Dr. Charnigo

## Lecture 4

### Regression and classification

*Back to continuous and categorical.* Recall that a supervised learning problem is one in which we have a well-defined response variable. We can further describe a supervised learning problem as either a "regression problem" or a "classification problem". Most people use "regression problem" to refer to any situation in which the response variable is continuous (or so finely discretized that it may be treated as continuous) and "classification problem" to refer to any situation in which the response variable is categorical.

Fernandez departs from this convention (page 151) by also calling a "regression problem" any situation in which the response variable is dichotomous (or "binary"). Presumably he does this because one of the most common statistical methods for handling a dichotomous response variable is called "logistic regression". In any case, the nature of your response variable determines what data mining methods you can employ.

### Simple linear regression

*Basic idea.* You have a continuous response variable $Y$ and a continuous explanatory variable $X$. Imagine plotting your data points so that the vertical position of each point is a value of $Y$ and the horizontal position of each point is a value of $X$. Now imagine drawing a line through the plotted data points such that most of the data points are close to the line. This is the basic idea behind simple linear regression.

*Statistical formulation.* Let $E(Y|x)$ denote the "conditional expectation" of $Y$ given that $X = x$. Here we adhere to a common notational convention in statistics: a capital letter such as $X$ represents a random variable, while a lower case letter such as $x$ represents a generic numerical value assumed by a random variable. This conditional expectation is the average value for $Y$ when $X = x$. For example, suppose that

$$E(Y|x) = 40 + 5x.$$

Based on this equation, when $X = 10$ the average value for $Y$ is 90; when $X = 12$ the average value for $Y$ is 100; and, when $X = 14$ the average value for $Y$ is 110.

In a "simple linear regression" model, we assume that

$$E(Y|x) = \alpha + \beta x.$$

However, in practice we do not know what $\alpha$ and $\beta$ are; we must estimate $\alpha$ and $\beta$ from data. [[We also make assumptions about how $Y$ may differ from its conditional expectation. I will describe these assumptions later, in the context of the more general "multiple linear regression" model.]] Of course, a simple linear regression model is usually only an approximation to reality.

**Discussion question**. Why is a simple linear regression model usually only an approximation to reality?

#### Multiple linear regression

*Statistical formulation.* We have a continuous response variable $Y$ and several explanatory variables $X_1, X_2, \ldots, X_k$. Typically most of the explanatory variables are continuous, but some may be dichotomous. [[This is not restrictive. For example, if we wanted to include a categorical explanatory variable with three categories, we could replace it by two dichotomous explanatory variables $Z_1$ and $Z_2$ using the following strategy. Let $Z_1 = 1$ for the first category and $Z_1 = 0$ for the other two categories; let $Z_2 = 1$ for the second category and $Z_2 = 0$ for the other two categories. In general, a categorical variable with $m$ categories can be replaced by $(m-1)$ dichotomous variables.]] Values of the explanatory variables are denoted generically by $x_1, x_2, \ldots, x_k$ and for subject $i$ specifically by $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$.

In a "multiple linear regression" model, we assume that

$$E(Y|x_1, x_2, \ldots, x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Note that the simple linear regression model is just a special case with $k = 1$ (and a minor change in notation). We also need to state how $Y$ may differ from its conditional expectation. We do this by saying that

$$Y_i = E(Y|x_{1,i}, x_{2,i}, \ldots, x_{k,i}) + \epsilon_i,$$

where $Y_i$ is the actual response for subject $i$ and $\epsilon_i$ is the difference between the actual response and the expected response for someone with the values $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$ on $X_1, X_2, \ldots, X_k$. Hence, we have

$$Y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \epsilon_i.$$

We refer to the $\epsilon_i$ as "error terms". Given the $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$, we assume that the $\epsilon_i$ are independent and identically distributed normal random variables with mean 0 and (unknown) variance $\sigma^2$. In particular, we do not want the error terms for different subjects to be correlated, nor do we want the variance of $\epsilon_i$ to depend on $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$.

*Interpretations of regression coefficients.* We call $\alpha$ the intercept coefficient. If $X_1 = X_2 = \cdots = X_k = 0$ is meaningful, then $\alpha$ is the expected value of $Y$ when $X_1 = X_2 = \cdots = X_k = 0$. We refer to $\beta_1, \beta_2, \ldots, \beta_k$ as partial slope coefficients. Assuming that $X_1, X_2, \ldots, X_k$ are not mathematically related (i.e., we do not have, for example, $X_2 = X_1^2$ or $X_3 = X_1 \times X_2$), $\beta_j$ represents the change in the expected value of $Y$ corresponding to a one-unit increase in $x_j$ when $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ are fixed.

*Estimating the regression coefficients.* We may estimate $\alpha$ and $\beta_1, \beta_2, \ldots, \beta_k$ by the principle of "least squares". Specifically, we find the values of $a$ and $b_1, b_2, \ldots, b_k$ that minimize the sum of squared deviations

$$\sum_{i=1}^{n} (y_i - \{a + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_k x_{k,i}\})^2.$$

That is, we find the values of $a$ and $b_1, b_2, \ldots, b_k$ for which the approximation

$$y_i \approx a + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_k x_{k,i}$$

is most reasonable. Let these optimal values be denoted by $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$. They are our estimates of $\alpha$ and $\beta_1, \beta_2, \ldots, \beta_k$.

[[Besides having geometric appeal, the least squares principle is equivalent to maximum likelihood in this setting. That is, $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are maximum likelihood estimates. Differential calculus (more specifically, setting to zero the first-order partial derivatives of the sum of squared deviations with respect to $a$ and $b_1, b_2, \ldots, b_k$) yields the so-called "normal equations"

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} e_i x_{1,i} = \sum_{i=1}^{n} e_i x_{2,i} = \cdots = \sum_{i=1}^{n} e_i x_{k,i} = 0,$$

where

$$e_i := y_i - \{\hat{\alpha} + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \hat{\beta}_k x_{k,i}\}.$$

With $k = 1$ the normal equations can be converted to simple pencil-and-paper formulas for the least squares estimates; see Lecture 12 at {www.richardcharnigo.net/STA580F08}. With $k > 1$ the normal equations can be solved with matrix operations; such operations are implemented automatically in PROC REG of SAS and the REGDIAG macro of Fernandez.]]

*Fitted model and fitted values.* For generic $x_1, x_2, \ldots, x_k$, we define

$$\hat{y} := \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k.$$

For the values $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$ specific to subject $i$, we define

$$\hat{y}_i := \hat{\alpha} + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \hat{\beta}_k x_{k,i}.$$

We refer to the full equation

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

with generic $x_1, x_2, \ldots, x_k$ as a "fitted model". When we replace generic $x_1, x_2, \ldots, x_k$ by specific numbers, $\hat{y}$ becomes a number that we call a "fitted value" or "predicted value". We call $\hat{y}_i$ a "fitted value" or "predicted value" for subject $i$.

*Quantifying explanatory ability.* We might try to quantify explanatory ability of the multiple linear regression model by computing

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

which is called the "residual sum of squares". This quantity will be small when the fitted values for the subjects are close to the actual responses, which is to say that the explanatory variables are accounting for much of the variability in the response. The residual sum of squares will be large when the fitted values for the subjects are not close to the actual responses, which is to say that the explanatory variables are failing to account for much of the variability in the response.

Unfortunately, the residual sum of squares depends on the measurement scale. For example, changing from inches to feet would cause the residual sum of squares to shrink to $(1/12)^2 = 1/144$ its original value. Hence, we need to consider the residual sum of squares in relation to another quantity.

One possibility is

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \tag{1}$$

where the denominator of (1) is called the "total sum of squares". In fact, one can show that

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \bar{y})^2, \tag{2}$$

where the first quantity on the left side of (2) is called the "regression sum of squares". [[Put $e_i := y_i - \hat{y}_i$. Relation (2) holds because the cross product term $2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ can be expressed as

$$2\hat{\alpha}\sum_{i=1}^n e_i + 2\hat{\beta}_1\sum_{i=1}^n e_i x_{1,i} + 2\hat{\beta}_2\sum_{i=1}^n e_i x_{2,i} + \cdots + 2\hat{\beta}_k\sum_{i=1}^n e_i x_{k,i} - 2\bar{y}\sum_{i=1}^n e_i,$$

each of whose sums equals 0 by virtue of the aforementioned normal equations.]] The regression sum of squares reflects the improvement in approximating the actual responses by fitted values when we take into account the explanatory variables instead of ignoring them, in which case the fitted value for each subject would be the sample mean $\bar{y}$.

Let us define

$$R^2 := \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Using equation (2), we see that $R^2$ equals 1 minus expression (1). We often use $R^2$ rather than expression (1) for quantifying explanatory ability since $R^2$ may be interpreted as the fraction of variability in the response accounted for by the multiple linear regression model (or by the explanatory variables in the multiple linear regression model). We must be clear, however, that this fraction of variability pertains to the data from which we fit the model, not to the underlying population. In particular, we can almost always make $R^2$ bigger by adding more explanatory variables!

[[To verify the last assertion, suppose that we add $X_{k+1}$ to a multiple linear regression model already containing $X_1, X_2, \ldots, X_k$. If $\hat{\beta}_{k+1} = 0$, then the minimal sum of squared deviations is the same as before, and hence so is $R^2$. If $\hat{\beta}_{k+1} \neq 0$, then the minimal sum of squared deviations must be smaller than before, and hence $R^2$ must be larger than before; if it were otherwise, we would obtain a logical contradiction since setting $b_{k+1} := 0$ in the sum of squared deviations would allow us to attain a smaller value than the "minimum"!]]

**Discussion questions**. Recall the bias-variance tradeoff from Lecture 2. If we have a very high $R^2$ because a lot of explanatory variables are in the model, is this a high-bias or a high-variance situation?

What might we try if we wanted to quantify explanatory ability so that our assessment better reflected the underlying population rather than the data from which we fit the model?

*Our illustrative example.* The data set in {FEV.xls} contains a continuous response variable, $FEV$ (forced expiratory volume), and four explanatory variables: $AGE$ (age), $HGT$ (height), $SEX$ ($1 =$ male, $0 =$ female), and

$SMOKE$ (1 = smoker, 0 = nonsmoker). Note that this data set, referenced on page 157 of Rosner (2005), pertains to children and adolescents.

Ultimately, we wish to find a model from which we can effectively predict $FEV$ and to quantify the accuracy of our predictions from that model. For now, however, let us examine some of the output I obtained by applying Fernandez's REGDIAG macro to a training subset, available from my web page as {fevtrain.sas7bdat}. This output is recorded in {RCLib.fevtrain27.rtf}.

*Page 24*. Here we have the correlation for each pair of variables in our data set. Recall from your introductory statistics course that the correlation measures the strength of the linear relationship between two variables and must be between $-1$ and 1. A positive correlation suggests that the two variables have a tendency to increase or decrease together, while a negative correlation suggests that the two variables have a tendency to move in opposite directions. [[See, for example, Lecture 11 at {www.richardcharnigo.net/STA580F08}.]]

**Discussion question**. The correlation between $FEV$ and $SMOKE$ is 0.2962, a positive number. Assuming no gross errors in the training subset, does this suggest that pulmonary function is better among smoking children and adolescents than among non-smoking children and adolescents?

*Page 25, Analysis of Variance box*. Here are some features to note.

• DF (degrees of freedom) column: The entry in the Model row is the number of partial slope coefficients in the model. The entry in the Total row is $(n - 1)$, where $n$ is the size of the data set used to fit the model. In this case, $n = 327$ because I allocated 50% of the 654 cases in {FEV.xls} to the

training subset. The entry in the Error row is the difference between the entry in the Total row and the entry in the Model row.
• Sum of Squares column: We have the regression sum of squares, the residual sum of squares, and the total sum of squares.
• Mean Square column: We have the regression sum of squares divided by the corresponding degrees of freedom and the residual sum of squares divided by the corresponding degrees of freedom.
• F value column: The f statistic is the regression mean square divided by the residual mean square. If we were interested in hypothesis testing, this is what we would use to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$.
• Pr > F column: This is the p-value corresponding to the f statistic.

*Page 25, Untitled middle box.* Here are some features to note.
• Root MSE: This is the square root of the residual mean square. We may interpret it as an estimate of $\sigma$, the standard deviation of the $\epsilon_i$ in the multiple linear regression model.
• Dependent Mean: This is just the sample mean, $\bar{y}$.
• Coeff Var: The coefficient of variation is the root MSE expressed as a percentage of the sample mean.
• R-Square: This is $R^2$.
• Adj R-Sq: This is a variant of $R^2$ that reduces the optimism of $R^2$ for complicated models. Adjusted $R^2$ will always be lower than $R^2$, much lower if $R^2$ has been inflated through the addition of extraneous explanatory variables. In analogy to

$$R^2 = 1 - \text{Res SS/Tot SS},$$

we have

$$\text{Adjusted } R^2 = 1 - \text{Res MS/Tot MS},$$

where total mean square is the total sum of squares divided by $(n - 1)$.

*Page 25, Parameter estimates box.* Here are some features to note.

• Parameter Estimate column: Here we have $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_k$.

• Standard Error column: Recognizing that the parameter estimates are realizations of random variables, because the subjects from whom we acquired data were randomly selected, we may ask what are the standard deviations of those random variables. The standard error column provides estimates of those standard deviations. For a more practical interpretation, we anticipate that each of the parameters $\alpha$, $\beta_1$, $\beta_2$, ..., $\beta_k$ is within two standard errors of the corresponding estimate.

• T value column: The t statistics are quotients of parameter estimates by their corresponding standard errors. If we were interested in hypothesis testing, this is what we would look at to test $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$, and so forth.

• Pr > |T| column: These are the p-values corresponding to the t statistics.

• Next three columns: Don't worry about these.

• VIF column: The explanation is deferred to Lecture 5.

• 95% Confidence Limits: These are 95% confidence intervals for $\alpha$, $\beta_1$, $\beta_2$, ..., $\beta_k$. For reasonably large $n$, the confidence intervals are roughly the estimates $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_k$ plus or minus twice their standard errors. [[The exact multiplier for the standard errors is the 0.975 quantile of a t distribution on $(n - k - 1)$ degrees of freedom.]]

*Page 26.* The residual for subject $i$ is simply the actual value minus the fitted value,

$$y_i - \hat{y}_i,$$

as we know since we have already encountered the residual sum of squares,

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

The predicted residual sum of squares (PRESS) is defined similarly except that $\hat{y}_i$ is determined from an auxiliary multiple linear regression model fitted without the data from subject $i$. The purpose of PRESS is to reduce the optimism for complicated models, and PRESS will typically be higher than the ordinary residual sum of squares. A "prediction $R^2$" may then be calculated via

$$\text{Prediction } R^2 = 1 - \text{PRESS/Tot SS}.$$

*Page 27, Title.* You may have noticed a running title, "R2 calculation based on response mean model: 0.84", and that the $R^2$ quoted there is higher than the $R^2$ quoted on page 25. Fernandez explains (page 211) that this is the result of averaging the actual responses for all subjects with the same values of the explanatory variables and then fitting a multiple linear regression model to the averages of the actual responses. My opinion is that the "response mean $R^2$" is misleading and should be ignored.

*Page 27, Fitted Model.* The fitted model appears near the top of the page, although $FEV$ should really be $\widehat{FEV}$,

$$\widehat{FEV} = -4.3386 + 0.0309SMOKE + 0.0797AGE + 0.0999HGT + 0.1575SEX.$$

*Page 27, Graphic.* The vertical positions of the blue dots are the actual responses $y_i$, whereas the horizontal positions are the fitted values $\hat{y}_i$. The red dots form a straight line relative to which the patterns in the blue dots can be assessed. In this instance, we see that there is some tendency for the multiple linear regression model to yield fitted values that are too small for subjects with the highest forced expiratory volumes. We know this because there are many more blue dots above the red dots than below as we approach the right edge of the graphic.

*Page 28, Graphic.* This graphic reflects response variability and residual variability. The dark part of the graphic represents the actual responses, placed in ascending order and centered about zero. The centering facilitates comparison with the corresponding residuals, which are represented by the light part of the graphic. The better the multiple linear regression model fits the data, the smaller the light part of the graphic will be. We can also see in this graphic that the residuals are predominantly positive for subjects with the highest forced expiratory volumes.

*Pages 29 and 30.* These pages present information on the 10% of subjects with the lowest fitted values and the 10% of subjects with the highest fitted values. The last four columns contain 95% confidence intervals and 95% prediction intervals. Each 95% confidence interval pertains to the mean response among all individuals in the population who are like the present subject with regard to the explanatory variables. Each 95% prediction interval pertains to the response for a randomly selected member of the population (not already in the sample) who is like the present subject with regard to the explanatory variables.