Multivariate Analysis (Slides 12)

- In today's class we consider logistic regression.
- The basic idea will be motivated, along with the interpretation of the model and some software output.
- You may have come across logistic regression in other classes, but here we consider it as a classification technique.
- Similar to LDA and QDA, logistic regression is a parametric classification technique making distributional assumptions over the data (which may or may not be appropriate) so as to allow a probabilistic quantification of class assignment.

Resting Pulse Data

- We are motivated in studying how resting pulse rate depends on whether the patient smokes and their weight.
- The data consists of 92 subject recordings of $RestingPulse$ (either $1 = Low$, or $0 = High$), whether the subject $Smokes$ ($1 = Yes$, $0 = No$), and the subject's $Weight$ (in lbs).
- Logistic regression will be used to study the relationship between resting pulse, smoking and weight.

Logistic Regression

- Let $P(Low)$ be the probability that a patient has a low resting pulse.
- We want to establish how this value depends on the value for $Smokes$ and $Weight$.
- To ensure that $P(Low)$ falls within $[0, 1]$, we consider the following model:

$$\text{logit}(P(Low)) = \log\left(\frac{P(Low)}{1 - P(Low)}\right) = \alpha + \beta_S Smokes + \beta_W Weight$$

- Or equivalently,

$$\frac{P(Low)}{1 - P(Low)} = \exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)$$

- Or,

$$P(Low) = \frac{\exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)}{1 + \exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)}$$

Variants

- The logistic model is not the only choice that guarantees $P(Low) \in [0, 1]$.
- For example, Economists often use *probit* models.
- **Probit:**

$$\Phi^{-1}\left(P(Low)\right) = \alpha + \beta_S Smokes + \beta_W Weight$$

  Here $\Phi^{-1}(z) = w$ if the probability of a Normal$(0, 1)$ random variable is less than w is equal to z; that is to say, it is the inverse CDF of a Normal$(0, 1)$.

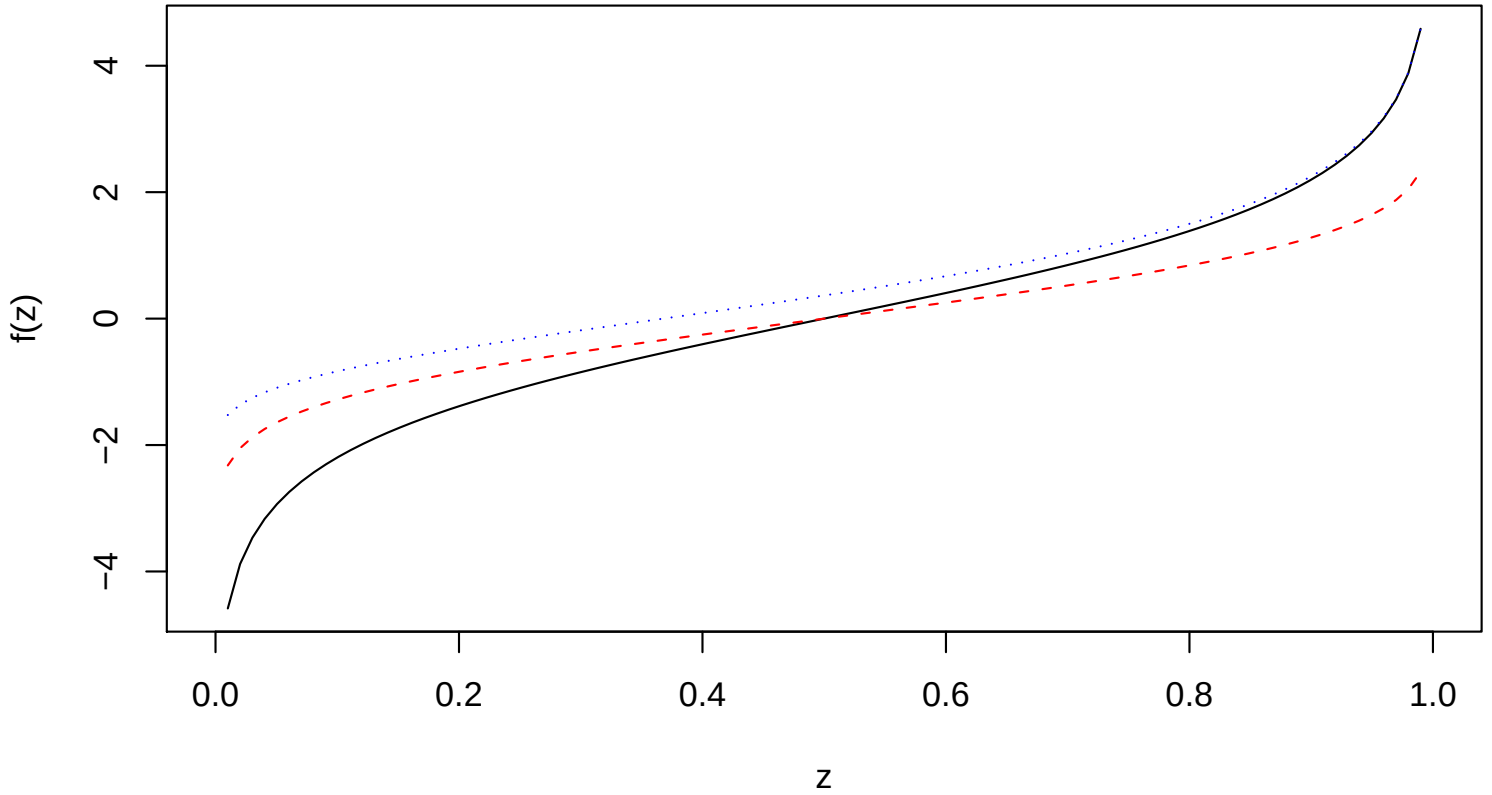
- **Gompit/Complementary log-log:**

$$-\log\left[-\log(P(Low))\right] = \alpha + \beta_S Smokes + \beta_W Weight.$$

- There are many alternative possibilities, but assigning the logit function of the probability to a linear expression of the covariates is common.

Plot of Functions

- A plot of the logit ($\log[z/(1-z)]$) —, probit ($\Phi^{-1}(z)$) – – – and gompit ($-\log[-\log(z)]$) $\cdots$ functions reveals some differences.



- These functions are known as link functions.

Behaviour near 0 or 1

- Compared to the probit function, the logit function tends to $-\infty$ faster as its argument approaches 0, and also tends to $+\infty$ faster as its argument approaches 1.
- If the class indicator in labeled as $\{0, 1\}$ and the probability is for assignment to class 1, then this means that for a given set of covariate values, the logistic model will give a larger probability than the probit model if both are below 0.5, and a smaller probability than the probit model if both are above 0.5.
- In other words, more evidence is required in the logistic model than the probit model to increase probability of assignment to a particular class.
- The gombit is an example of an a-symmetric link function, making it easier to assign membership to one class than the alternative.

Some Probability

- Suppose that there are n independent observations of blood pressure. For example,

$$\{Low, Low, High, Low, \ldots, High\}.$$

- What is the probability of observing this?
- We know that

$$P(Low) = \frac{\exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)}{1 + \exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)}.$$

- So,

$$P(High) = \frac{1}{1 + \exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)}.$$

- Also, due to independence, $P(Low, Low, High, Low, \ldots, High)$ can be simplified to $P(Low)P(Low)P(High)P(Low)\cdots P(High)$.

Some More Probability

- Furthermore, $P(Low)P(Low)P(High)P(Low)\cdots P(High)$, can be expressed as:

$$\prod_{i=1}^{n}\left[\frac{\exp\left(\alpha+\beta_S Smokes_i+\beta_W Weight_i\right)}{1+\exp\left(\alpha+\beta_S Smokes_i+\beta_W Weight_i\right)}\right]^{y_i}$$

$$\times\prod_{i=1}^{n}\left[\frac{1}{1+\exp\left(\alpha+\beta_S Smokes_i+\beta_W Weight_i\right)}\right]^{1-y_i}$$

- Here $y_i=1$ if the person has Low blood pressure and $y_i=0$ if the person has $High$ blood pressure.
- We could find the values of α, β_S and β_W that maximize this probability.
- These are known as the maximum likelihood estimates for the parameters, whilst the probability of the observed data is known as the likelihood.

Some R Output

- The estimates for α, β_S and β_W for the resting pulse data are:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.98717    1.67931  -1.183   0.2367
SmokesYes   -1.19297    0.55298  -2.157   0.0310 *
Weight       0.02502    0.01226   2.042   0.0412 *
```

What does this tell us about the effect of weight or the smoking indicator upon resting pulse?

Interpreting the R Output

- The `Estimate` column gives us the estimates of α, β_S and β_W in our logistic model. Whether or not they are positive or negative tells us how a change in a covariate value will influence the probability of assignment.
- Here the 1 category is Low, whilst the 0 category is $High$.
- As the coefficient for $SmokesYes$ is negative, this means that if the individual smoked, they would have smaller probability of being assigned to the 1 category, or in other words, are more likely to have $High$ resting pulse.
- The coefficient for $Weight$ is positive, meaning that the more somebody weighs, the more likely they will belong to the 1 category, *i.e.*, have low resting pulse.
- In making interpretations like these, you have to be aware of how the data is being coded (and even more careful if you have categorical and not simply binary covariates).

Interpreting the R Output

- Let us consider what the output tells us for an individual who smokes, and who weighs 190 lbs.
- The logistic model was shown to give:

$$P(Low) = \frac{\exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)}{1 + \exp\left(\alpha + \beta_S Smokes + \beta_W Weight\right)}$$

- Hence for our person this becomes:

$$P(Low) = \frac{\exp\left(-1.99 - 1.19 \times 1 + 0.03 \times 190\right)}{1 + \exp\left(-1.99 - 1.19 \times 1 + 0.03 \times 190\right)}$$

- Your calculator can then confirm this is approximately 0.93.
- If our person didn't smoke then this would be approximately 0.98.

Standard Errors

- The standard errors of the coefficients record the uncertainty in the resulting estimate.
- These tend to decrease with sample size.
- They can be used to construct 95% confidence intervals for the coefficients:

$$\text{(Coefficient Estimate)} \pm 2\text{(Standard Error of Coefficient)}.$$

- This yields,

$$\beta_S \in (-2.30, -0.09) \text{ and } \beta_W \in (0.001, 0.050).$$

- These can be used to find 95% confidence intervals for $\exp(\beta_S)$ and $\exp(\beta_W)$.

$$\exp(\beta_S) \in (0.10, 0.90) \text{ and } \exp(\beta_W) \in (1.00, 1.05).$$

Tests

- We may wish to test certain hypotheses, such as:

$$H_0 : \beta_S = 0 \text{ versus } H_1 : \beta_S \neq 0$$

  or alternatively,

$$H_0 : \beta_W = 0 \text{ versus } H_1 : \beta_W \neq 0$$

- To do this we can compute

$$\frac{(\text{Coefficient Estimate}) - 0}{(\text{Standard Error of Estimate})}.$$

- Under H_0 this quantity is approximately Normal$(0, 1)$.
- We can use this fact to determine if the observed value is surprisingly large or not.
- We can then determine if we have sufficient evidence to reject the null hypothesis (H_0) in favour of the alternative, or not.

Interpreting the R Output

- If the data had been standardized, then the magnitude of the coefficient would indicate the importance of the variable in predicting the class.
- As it is, the remaining output tells us even more.
- If we assume that the true coefficient β of a variable was 0, and that the actual estimate generated $\hat{\beta}$ was random and subject to a draw from a Normal distribution with mean 0 and standard deviation the estimated standard error $S_{\hat{\beta}}$, then:

$$\frac{\hat{\beta}}{S_{\hat{\beta}}} \sim \mathcal{N}(0,1)$$

- The column `z value` gives us $\hat{\beta}/S_{\hat{\beta}}$, and then the column `Pr(>|z|)` tells us the probability of observing a `z value` at least that far from 0 under the $\mathcal{N}(0,1)$ distributional assumption.

Interpreting the R Output

- If `Pr(>|z|)` is small, then under the hypothesis that the true coefficient is 0 and that the estimate is drawn from a Normal distribution with mean 0 and standard deviation as given by the standard error, we have observed an unlikely outcome.
- If the `Pr(>|z|)` value is very small we may start to question the appropriateness of our original hypothesis, as we are faced with the dilemma of justifying such a rare event.
- A (too) often used rule is to reject the hypothesis if `Pr(>|z|)`< 0.05.
- Note that `Pr(>|z|)` is NOT the probability that the true coefficient is 0, it is simply the probability of having observed such a coefficient estimate under the assumption that the true coefficient is 0. This often made mistake arises because of confusion over conditional probability.
- Generally $P(\text{Hypothesis}|\text{Data}) \neq P(\text{Data}|\text{Hypothesis})$.
- The use of a prior and Bayesian statistics allows talk of $P(H|D)$.

Interpreting the R Output

- For the resting pulse data, we find that, if we work at the 0.05 significant level:
- The estimate for the intercept α is not so extreme as to reject the hypothesis its true value is 0. So we might consider a simpler model in which it is not included.
- The estimate for β_S is extreme under the hypothesis that its true value is 0, so we may wish to reject such an hypothesis. If so, we may consider the `SmokesYes` indicator as significant in predicting resting pulse.
- The estimate for β_W is extreme under the hypothesis that its true value is 0, so we may wish to reject such an hypothesis. If so, we may consider the `Weight` value as significant in predicting resting pulse.

Example: Low Birth Weight

- In 1986, a study at Baystate Medical Center, Springfield, MA, was completed to determine what factors influence the birth weight of newborns.
- The following factors were recorded for each case:
  - 'low' indicator of birth weight less than 2.5kg
  - 'age' mother's age in years
  - 'lwt' mother's weight in pounds (at last menstrual period)
  - 'race' mother's race ('1' = white, '2' = black, '3' = other)
  - 'smoke' smoking status during pregnancy
  - 'ptl' number of previous premature labours
  - 'ht' history of hypertension
  - 'ui' presence of uterine irritability
  - 'ftv' number of physician visits during the first trimester (0,1,2+)

Results: Low Birth Weight

- The following coefficients were obtained:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.82302    1.24471   0.661  0.50848
age         -0.03723    0.03870  -0.962  0.33602
lwt         -0.01565    0.00708  -2.211  0.02705 *
race2        1.19241    0.53597   2.225  0.02609 *
race3        0.74069    0.46174   1.604  0.10869
smoke        0.75553    0.42502   1.778  0.07546 .
ptl          1.34376    0.48062   2.796  0.00518 **
ht           1.91317    0.72074   2.654  0.00794 **
ui           0.68019    0.46434   1.465  0.14296
ftv1        -0.43638    0.47939  -0.910  0.36268
ftv2         0.17901    0.45638   0.392  0.69488
```

- What are the effects of the covariates upon birth weight?
- Note the treatment of categorical covariates.

Interactions

- Including interactions adds possible products of the covariates to the model.
- In the pulse data this would then be:

$$\log\left(\frac{P(Low)}{1-P(Low)}\right) = \alpha + \beta_S Smokes + \beta_W Weight + \beta_{SW} Smokes \times Weight$$

- Interactions allow for the effect of one covariate to be altered depending of the value of another covariate.
- In the above, the interaction would be reasonable if the effect that weight has upon resting pulse would differ depending on whether or not the individual smoked. For example, the weight of the individual may be considered to be very important in predicting low pulse if the individual smoked, but of little importance if the individual did not smoke.
- When appropriate, interactions can greatly increase the model's predictive ability.

R Output with interactions

- The estimates for α, β_S, β_W and β_{SW} are:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.9769778  2.1876955  -0.904    0.366
SmokesYes       -1.2187425  3.5901342  -0.339    0.734
Weight           0.0249468  0.0160917   1.550    0.121
SmokesYes:Weight 0.0001804  0.0248266   0.007    0.994
```

What does this tell us about whether an interaction is important for the pulse data?

Goodness Of Fit

- Additional output from the birth weight regression is as follows:

```
    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 195.48  on 178  degrees of freedom
AIC: 217.48
```

Akaike's Information Criterion

- Fitting a model can be considered as trying to optimize two opposing goals:
  - Model fit: make the fit as close as possible to the data.
  - Model complexity: make the model simple.
- The Akaike information criterion (AIC) defines a measure that combines both aspects:

$$AIC = -2\log(Likelihood) + 2p$$

- Here p is the number of model parameters.
- Hence the model with minimal AIC offers a compromise between model fit and model complexity.
- AIC can be used to compare nested models.
- AIC is only relevant when there are at least two competing models.

Deviance

- Consider the full (saturated) model in which as many parameters are estimated as there are data points.
- In effect we simply determine a number of indicator terms for whether a data point was in class '1' or not.
- Then the likelihood is $L(Data|I_1, \ldots, I_n) = 1$.
- The log-likelihood is $l(Data|I_1, \ldots, I_n) = 0$.

Deviance

- If there are multiple data points sharing the same covariate vector but with different classifications, then instead we have:

$$L(Data|\pi_1, \ldots, \pi_M) = \propto \prod_{i=1}^{M} \pi_i^{r_i} (1 - \pi_i)^{n_i - r_i}$$

- Here M is the number of distinct covariate vectors, n_i is the number of data points sharing covariate vector i, and r_i is the number of data points sharing covariate vector i and assigned to class '1'. Then:

$$l(Data|\pi_1, \ldots, \pi_M) = const + \sum_{i=1}^{M} r_i \log(\pi_i) + (n_i - r_i) \log(1 - \pi_i)$$

- In the latter case the log-likelihood is maximised when $\pi_i = r_i/n_i$ (assign $0 \log(0)$ to be 0 in case either r_i or n_i is 0).
- In either the distinct or non-distinct covariate case, let L_{max} be the maximum value of L.

Deviance

- L_{max} represents the largest possible value for the likelihood under any model.
- Now assume that we propose a simpler model based on a fewer number of parameters relating to the covariate and/or interaction terms, and let L_{mod} be the maximum value of the likelihood under this model.
- The deviance for the model is defined to be:

$$Dev = 2[\log(L_{max}) - \log(L_{mod})]$$

- Intuitively, the closer L_{mod} is to L_{max}, and so the smaller the deviance, the better the proposed model is at fitting the data.

Deviance

- It can be shown that, under the hypothesis the proposed model is true, the deviance is approximately distributed as a χ^2_{M-k-1} distribution, where k is the number of parameters of the proposed model (or χ^2_{n-k-1} if no repeated covariate vectors).
- Hence, if the deviance is greater than the 95% point of the appropriate chi-square distribution, we might claim that, under the hypothesis the proposed model is true, we have observed a rather rare event.
- We might then begin to question the validity of the original hypothesis (though remember we can not assign any probability to it holding true).

Null Deviance

- The simplest model is the one in which any data point is assigned a common probability of class assignment regardless of its covariate scores. In this case we have one parameter to determine the probability of class assignment and it is estimated as the proportion of data in class '1'.
- The deviance for this proposed model is known as the 'null deviance'
- Intuitively, if none of the covariates actually influence class assignment, the model deviance will not be much smaller than the null deviance.
- We have the following relationship:

$$L_{null} \leq L_{mod} \leq L_{max}$$

- The deviance (or residual deviance) is $2[\log(L_{max}) - \log(L_{mod})]$.
- The null deviance is $2[\log(L_{max}) - \log(L_{null})]$.

Deviance Continued

- For n reasonably large and M small we can interpret residual deviance as a measure of fit.
- In `R` we would calculate

$$\texttt{> 1-pchisq(}resid.deviance\texttt{,}deg.freedom\texttt{)}$$

- If the resulting value is very small, then a rare event has occured under the assumption that our proposed model is correct.

Deviance Continued

- To test if the model is explaining the variation in the data we can compute the p-value from a goodness-of-fit test:

```
> 1 - pchisq(deviance(res), df.residual(res))
[1] 0.1755189
```

- The smaller the value the greater the evidence that there is a significant difference between the full (saturated) model and the model of interest. A cut-off of 0.05 is usually assumed.

- Often, however, there will be a significant difference, and so checking the the difference of deviance against null deviance against a χ^2_p-distribution will indicate whether the model of interest is better than the null model with no covariates.

Additional Tests

- There are of course many different possibilities for assessing the appropriateness of the model.
- For example, the Pearson Test computes the statistic:

$$\sum_{i=1}^{n} \frac{(y_i - P(y_i = 1))^2}{P(y_i = 1)}.$$

- This is reminiscent of the χ^2 test that you may have seen before:

$$\sum_{i} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}.$$

- The Pearson statistic will be small if 1's have high probabilities and the 0's have small probabilities.
- It is approximately χ^2 distributed.

Classification

- As far as this course is concerned logistic regression can be considered as a classification technique for when there are two groups.
- A new observation is classified as being of one group or another depending on whether the predicted probability falls above or below the threshold of 0.5.
- In this sense logistic regression is very similar to linear discriminant analysis.

Comparison to LDA

- In LDA the decision boundary between class k and class l is given by:

$$\log \frac{P(k|\mathbf{x})}{P(l|\mathbf{x})} = \log \frac{\pi_k}{\pi_l} + \log \frac{f(\mathbf{x}|k)}{f(\mathbf{x}|l)} = 0$$

- In logistic regression the model, by assumption, is:

$$\log \frac{P(k|\mathbf{x})}{P(l|\mathbf{x})} = \beta_0 + \beta^T \mathbf{x}$$

- Hence these models have the same form.
- The difference lies in the way the linear coefficients are estimated.