Statistical Techniques II

EXST7015

Logistic Regression

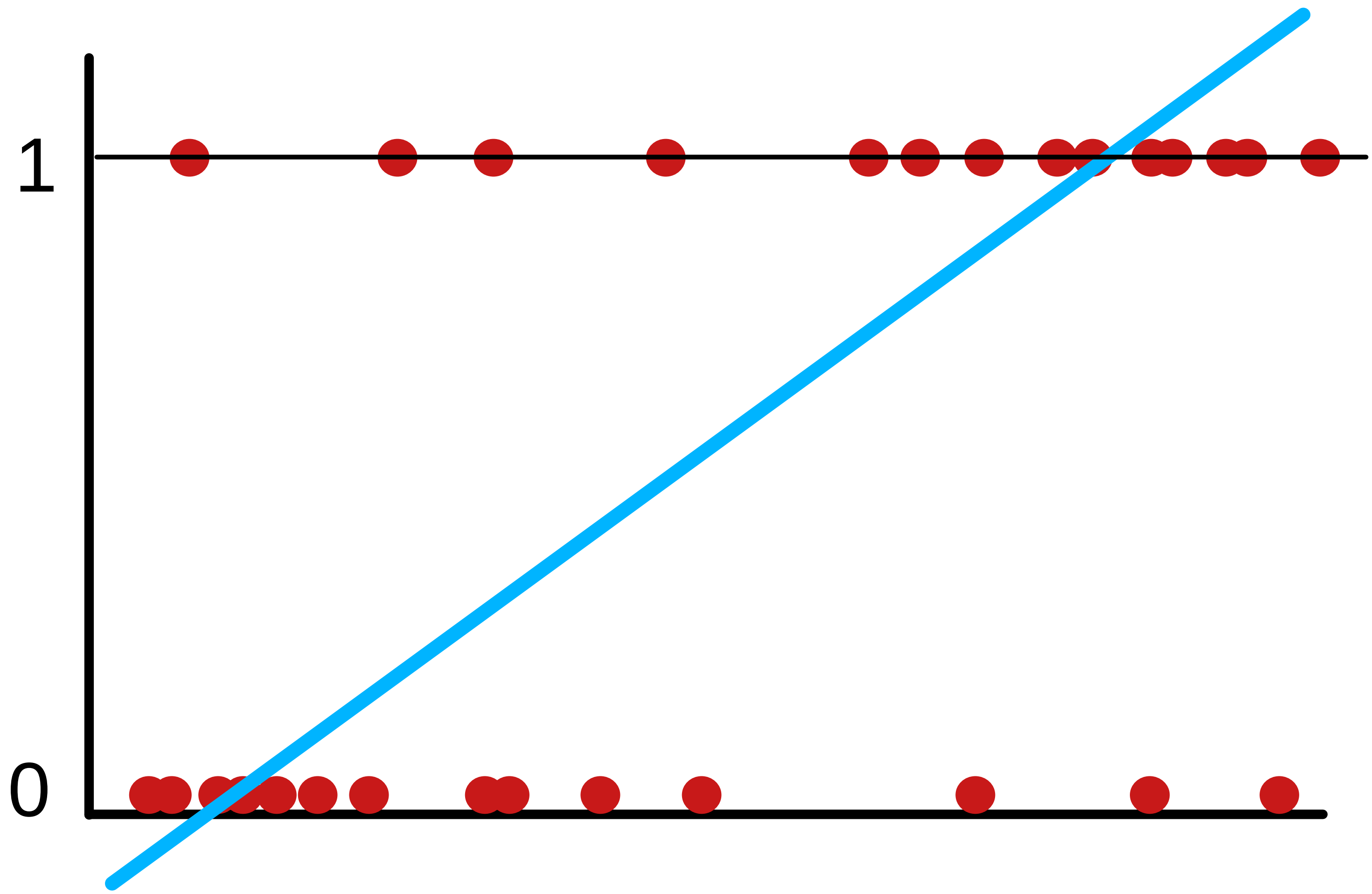
Regression on an indicator variable

- **What is an indicator variable? It is a variable with either the value 0 or 1.**
  - **When we get to ANOVA we will see that class variables (categorical, group) are usually reparameterized in parametric analyses as the value 0 or 1.**
  - **With only two levels (True-False, Up-Down, Male-Female, Marked-Unmarked, Heads-Tails) the values are easily re-coded.**
  - **If there are more levels (say t levels) then we need t-1 indicator variables.**

Regression on an indicator variable *(continued)*

- **But indicator variables are usually independent variables. ANOVA is all about indicator variables as independent variables.**
- **Regression on an indicator variable is different**
  - **Basically it is a simple linear regression where the <u>dependent variable</u> has a value of either 0 or 1.**
  - **This is called a binary response.**

Regression on an indicator variable (*continued*)

- 

Regression on an indicator variable *(continued)*

- This is a "primitive" version of regression on an indicator variable.
- The predicted value (Yhat) is interpreted as the probability of getting a 1.
- However, this line will go below zero and will go above 1. This makes the properties somewhat undesirable.

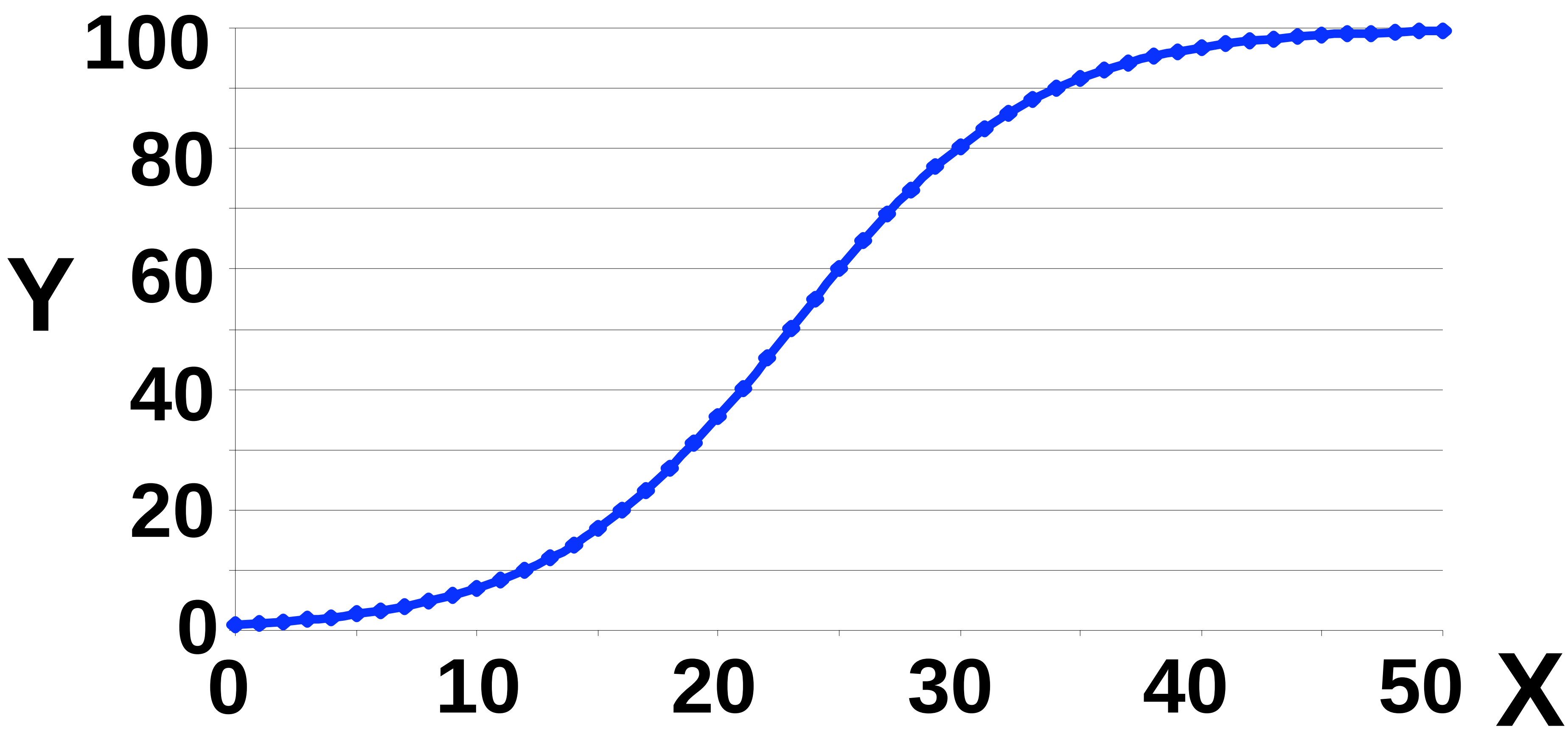
Regression on an indicator variable *(continued)*

- **The Logistic Model is a rather complicated model, it is not linear and cannot be fitted with PROC REG or PROC GLM.**
- **This equation is often used as a "growth" model.**
- **One version is given below.**

$$Y_i = \frac{b_1}{1+\left(\frac{b_1 - b_2}{b_1}\right)e^{b_3 X_i}} + e_i$$

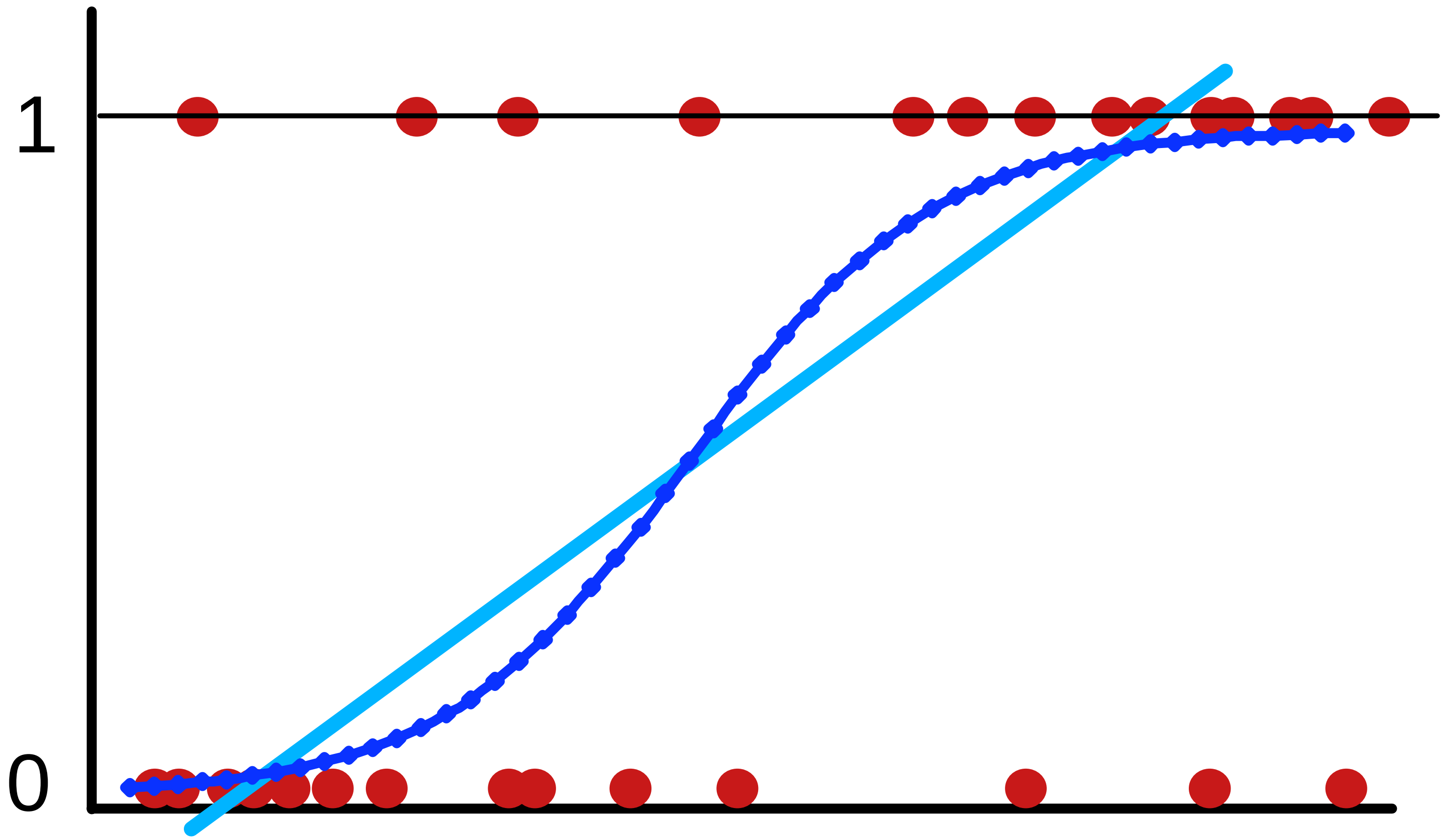
Regression on an indicator variable *(continued)*

- **Logistic Curve**



Regression on an indicator variable *(continued)*

- **Wouldn't it be nice if we could fit this model to our indicator variable?**



Logistic Regression

- Enter Logistic Regression.
- The big, nonlinear Logistic model can be much simplified if we know that b_1=1 and that b_2=0.
- This is basically what logistic regression does, it fits a logistic curve that goes from 0 to 1 asymptotically (so it never really reaches either one).
- Lets look at an example.

Odds

- What are the odds that you will get an A in this course?
- Interesting question, but first, what are exactly are "odds"?
  - Odds are an expression of the likelihood of some event happens compared to the likelihood that it does not happen.
  - If the odds on a horse in a race are 30 to 1, is that horse likely to win? Or loose?
- If the odds of an event happening are 50:50, what does that mean? What if the odds are 1:1? How is that different?

Odds *(continued)*

- **We will work with a number that is in some ways a little simpler, the odds ratio.**
- **This is the ratio of the probability of an event occurring and the probability of the same event not occurring.**
- **Since these add to 1, if the probability of an event is "p", then the probability of not occurring is "1-p".**
- **So "50:50" has an odds of 1.0, and the odds for 1:1 are also 1.0. They have the same odds.**

Odds *(continued)*

- If the odds are 1, then the likelihood of something happening is exactly equal to the likelihood that it will not happen. That is $p = 1 - p$.
- To simplify our concepts we will think of the odds as the ratio of two probabilities. The probability that some event happens (success) will be equal to p. The probability of failure will be $1-p$. The odds are calculated as $p/(1-p)$.

Odds *(continued)*

- **What if the odds are 2? This means that p is twice as large as (1-p), so success is twice as likely as failure. If the odds are 10 the probability of success is 10 times more likely than failure.**
- **If the odds are 0.5 or 0.1, then the probability of <u>failure</u> is twice as likely or ten times more likely than the probability of success.**

Odds *(continued)*

- **Now, what are the odds that you will get an A in this course?**
- *Disclaimer 1: The use of this example calculated from past grades in no way implies any promise or commitment about the distribution of future grades.*
- *Disclaimer 2: Although I will discuss only the number of A's, the non-A's are not all B's, there have been C's and D's. Sorry.*

Odds *(continued)*

- Now, what are the odds that you will get an A in this course?
- I have had 423 students take the course since I have been giving two exams in this course (previously 3).
- Of those 423, there have been 212 A's, we will call this "success".
- The odds of getting an A then are almost 50:50, and the odds are just about 1.

Odds *(continued)*

- **Interesting, but we can carry this one step further. What are the odds of getting an A if you had a 70 on the first exam? Or an 80? Or a 90?**
- **We will do Logistic regression to determine this. We will use SAS PROC Logistic. The structure is very similar to PROC REG.**
-

Logistic regression

```
proc logistic data=grades DESCENDING;
   TITLE1 'Logistic regression';
   model Grade_A = exam1;  run;
```

- The output is rather different because the analysis is not a least squares regression. You will be responsible only for interpreting the values that I have put in blue.

Logistic regression *(continued)*

- PROC LOGISTIC output

```
The LOGISTIC Procedure
                Model Information
Data Set                        WORK.GRADES
Response Variable               Grade_A
Number of Response Levels       2
Number of Observations          423
Link Function                   Logit
Optimization Technique          Fisher's scoring
```

- A logit is the natural logarithm of the odds. That is "$\log_e(p / (1-p))$".

Logistic regression *(continued)*

- PROC LOGISTIC output (continued)

```
            Response Profile

Ordered                             Total
  Value      Grade_A            Frequency

      1      TRUE                     212
      2      FALSE                    211
```

Logistic regression *(continued)*

- PROC LOGISTIC output (continued)

```
            Model Convergence Status
 Convergence criterion (GCONV=1E-8) satisfied.
          Model Fit Statistics
                                     Intercept
                     Intercept          and
Criterion              Only          Covariates
AIC                  588.400            425.407
SC                   592.448            433.502
-2 Log L             586.400            421.407
```

- These statistics are used to compare 2 models and we will not cover them here.

Logistic regression *(continued)*

- PROC LOGISTIC output (continued)

```
The LOGISTIC Procedure

        Testing Global Null Hypothesis: BETA=0

Test                     Chi-Square       DF      Pr > ChiSq
Likelihood Ratio           164.9934        1          <.0001
Score                      132.7164        1          <.0001
Wald                        96.1179        1          <.0001

             Analysis of Maximum Likelihood Estimates
                                  Standard
Parameter    DF     Estimate         Error    Chi-Square    Pr > ChiSq
Intercept     1     -16.9098        1.7443       93.9760        <.0001
Exam1         1       0.1952        0.0199       96.1179        <.0001
```

Logistic regression *(continued)*

- PROC LOGISTIC output (continued)

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Exam1	1.216	1.169	1.264

Association of Predicted Probabilities and Observed Responses

Percent Concordant	82.8	Somers' D	0.681
Percent Discordant	14.7	Gamma	0.698
Percent Tied	2.4	Tau-a	0.341
Pairs	44732	c	0.841

Logistic regression *(continued)*

- Do how do we interpret this output?
- We are interested primarily in the slope and intercept, and the test of the slope.
  - Intercept = -16.91
  - Slope = 0.1952
  - Likelihood ratio test P-value <0.0001
- So we have a highly significant slope. The interpretation here is the same as for regression, though the test statistic is different.

Logistic regression *(continued)*

- Now, we have a slope and an intercept. Could we get predicted values as with regression? Absolutely!
- However, we entered a log-odds into the model, so the predicted values are log-odds.
- We can get these predicted values and convert to probabilities.

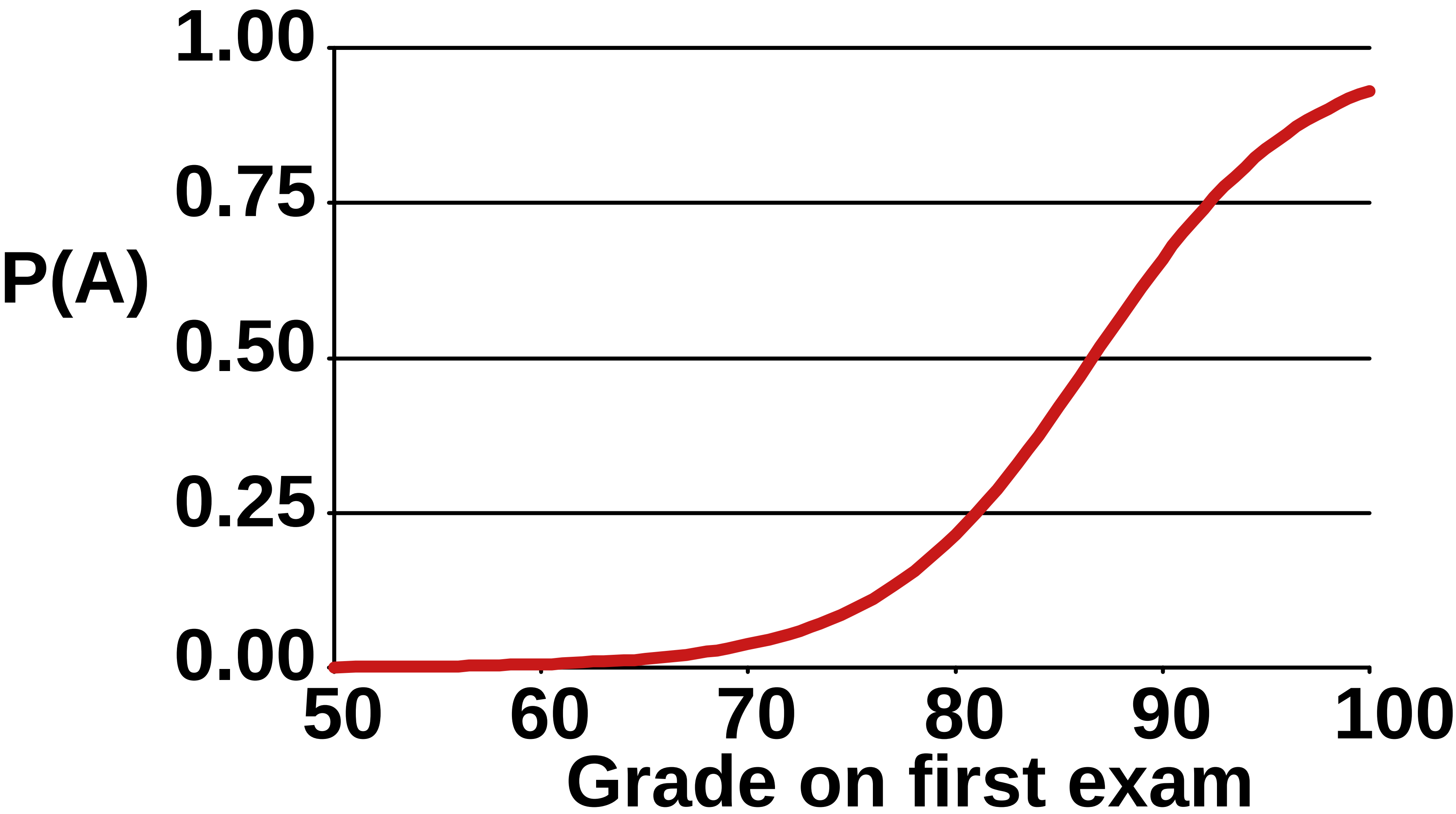
Logistic regression *(continued)*

- Calculate the probability that a person with an 83 will get an A in the course.
-
- $\text{Yhat}_i = -16.91 + 0.1952X_i$
- $= -16.91 + 0.1952(83) = -0.7086$
-
- This is a log odds, first take the antilog. $\exp(-0.7086) = 0.4923$
- So, $p/(1-p) = 0.4923$, then solving for p we get $p = 0.4923/(1+ 0.4923) = 0.3299$.

Logistic regression *(continued)*

- So the probability of getting an A with a grade of 83 on the first exam is 0.33, or we could figure that about 33% of students with a grade of 83 on the first exam will get an A.
- The predicted values, detransformed to probabilities, will form a logistic regression line going from 0 to 1.

Logistic regression (*continued*)



Logistic regression *(continued)*

- Interpretation of the slope.
- The analysis is log transformed, so to get the slope back on the original scale we need to take an antilog.
  - exp(0.1952) = 1.2156
  - As with other slopes, this is the change in Y per unit X, or the change in the odds for an increase of one point in the grade.
  - So the odds go up by 21 percent for each additional point in grade.

Logistic regression *(continued)*

- **Don't confuse the increase in odds with the increase in probability.**
- **The increase of 1.2156 is an odds ratio, and gives the increase in the odds (per unit change in X_i), not the ratio of the increase in the probability.**
- **See calculations next page.**

Logistic regression *(continued)*

Grade	Odds	Probability	Ratio of values of P	Odds ratio
70	0.03892	0.03746		
71	0.04731	0.04517	1.2058	1.21556
80	0.27412	0.21514		
81	0.33320	0.24993	1.1617	1.21556
90	1.93054	0.65877		
91	2.34668	0.70120	1.0644	1.21556

Logistic regression *(continued)*

- So the slope is an odds ratio, the proportional change in the odds (per unit change in X_i).
- SAS provides tests and confidence intervals for this value.
- The analysis as mentioned is NOT a least squares regression, it is actually a weighted maximum likelihood estimation carried out on the logit values.

Logistic regression *(continued)*

- **Multiple regression is perfectly feasible, and SAS also has the stepwise analyses and "best models" selection you are familiar with. The best model selection is not called RSquare, it is the SCORE option (based on chi square, not RSquare).**
- **CLASS variables can be included in the analysis. We will discuss these in ANOVA.**

Summary

- **Regression on an indicator variable is similar in concept to ordinary least squares regression, but differs considerably in the execution. The analysis is generally called Logistic Regression.**
- **The slope and intercept are used in regression in a way that is similar to ordinary least squares, but the value predicted is a log of the odds. This can be converted to probabilities.**

Summary *(continued)*

- **SAS provides statistics to evaluate the significance of the fit, and confidence intervals for the estimate of the slope.**
- **SAS provides other options for various selection techniques and the addition of class variables.**