slides originally by
Dr. Richard Burns,
modified by
Dr. Stephanie Schwartz

REGRESSION WITH QUALITATIVE VALUES

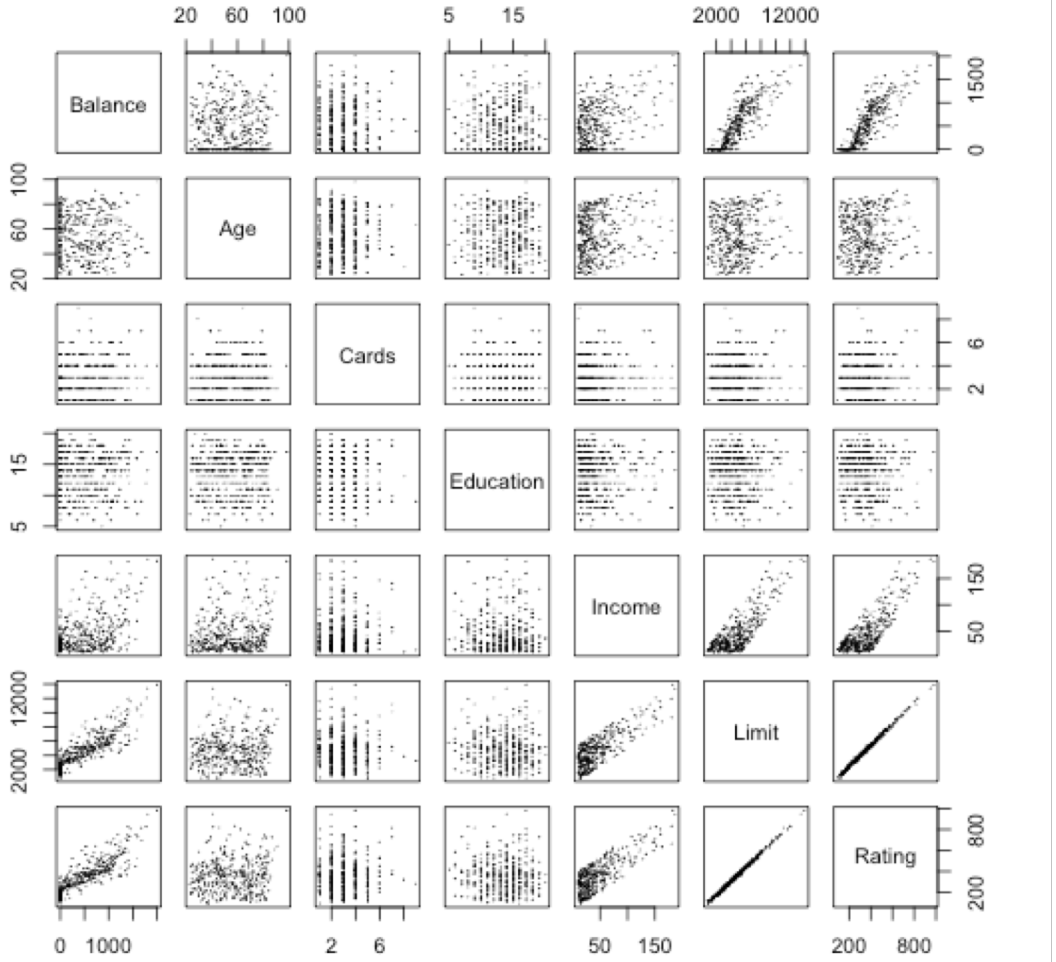
CSCI 452: Data Mining

Qualitative Predictors

- So far have assumed that all variables in linear regression model are *quantitative*.
- How to deal with *qualitative* variables?

Credit Dataset

- Response:
  - Balance (individual's average credit card debt)
- Quantitative Predictors:
  - Age (years)
  - Cards (number of credit cards)
  - Education (years of education)
  - Income (in thousands of dollars)
  - Limit (credit limit)
  - Rating (credit rating)
- Qualitative Predictors:
  - Gender {Male, Female}
  - Student {Yes, No}
  - Married {Yes, No}
  - Ethnicity {Caucasian, African American, Asian}




```
pairs(~Balance+Age+Cards+Education+Income+Limit+Rating, data=Credit, cex=.05)
```

Qualitative Predictors: Two Levels

- Levels (sometimes called factors): possible values of discrete variable
- Solution: create a dummy variable (or *indicator*) that takes on two possible numerical values
- *Credit* dataset, *Gender* variable: {Male, Female}
- Create new dummy variable:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Qualitative Predictors: Two Levels

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

… for now assuming that *Gender* is the only predictor in model …

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Simple Linear Regression Model

- Estimate coefficients B_0, B_1

Term zeros out for males

Qualitative Predictors: Two Levels

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

- Interpretation:
  - B_0: average credit card balance among males
  - $B_0 + B_1$: average credit card balance among females
  - B_1: average difference in credit card balance between females and males

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Qualitative Predictors: Two Levels

- Interpretation:
  - B_0: average credit card balance among males
  - $B_0 + B_1$: average credit card balance among females
  - B_1: average difference in credit card balance between females and males

$Balance = 509.80 + 19.73 * x_i$

- Average credit card debt for males is estimated to be $509.80.
- Females are estimated to carry $19.73 in additional debt, for a total of:
  $509.80+$19.73=$529.53

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Qualitative Predictors: Two Levels

```
> summary(lm.fit)

Call:
lm(formula = Credit.Balance ~ Gender_Female, data = Credit.2)

Residuals:
    Min      1Q  Median      3Q     Max
-529.54 -455.35  -60.17  334.71 1489.20

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    509.80      33.13  15.389   <2e-16 ***
Gender_Female   19.73      46.05   0.429    0.669
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Balance = 509.80 + 19.73 * x_i*

- p-value for the dummy variable is very high, indicating that there is no statistical difference in average credit card balance between the genders.

Qualitative Predictors: Two Levels

- Decision to code females as 1 and males as 0 is arbitrary.
  - It does alter the interpretation of the coefficients
- What would happen if we coded males as 1 and females as 0?

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is male} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is female} \end{cases}$$

Qualitative Predictors: Two Levels

- Interpretation:
  - B_0: average credit card balance among females
  - $B_0 + B_1$: average credit card balance among males
  - B_1: average difference in credit card balance between females and males

$Balance = 529.54 - 19.73 * x_i$

- Average credit card debt for females is estimated to be $529.54.
- Males are estimated to carry $19.73 in less debt, for a total of:
  $529.54-$19.73=$509.80

Same exact model!

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Qualitative Predictors: Two Levels

- Interpretation:
  - B_0: overall average credit card balance (ignoring gender)
  - B_1: amount that females are above the average, and males are below the average

$Balance = 519.67 + 9.865 * x_i$

- Average credit card debt, ignoring gender is $519.67.
- The average difference between males and females is:
  $9.865 * 2 = $19.73

Same exact model!

- It doesn't matter which coding scheme is used, as long as coefficients are correctly interpreted.

Qualitative Predictors: More than Two Levels

- Single dummy variable cannot represent all possible values for qualitative predictors with more than two levels
- *Solution:* create additional dummy variables
- For *Ethnicity* variable:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

Simple linear model, ignoring all other predictors….

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

Qualitative Predictors: More than Two Levels

- Interpretation:
  - B_0: average credit card balance for African Americans
  - B_1: difference in average balance between Asians and African Americans
  - B_2: difference in average balance between Caucasians and African Americans

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

Always one fewer dummy variable than number of levels.

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

Qualitative Predictors: Two Levels

- Interpretation:
  - B_0: average credit card balance for African Americans
  - B_1: difference in average balance between Asians and African Americans
  - B_2: difference in average balance between Caucasians and African Americans

$$Balance = 531.00 - 18.69^* x_{i1} - 12.50^* x_{i2}$$

Once again, arbitrary coding scheme.

- Estimated *balance* for African Americans is $531.00
- Asian category will have $18.69 in less debt than African American category
- Caucasian category will have $12.50 in less debt than African American category

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

Qualitative Predictors: Two Levels

Balance = 531.00 – 18.69 x_{i1} – 12.50* x_{i2}*

```
Call:
lm(formula = Credit.Balance ~ Ethnicity_Asian + Ethnicity_Caucasian,
    data = Credit.5)

Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          531.00      46.32  11.464   <2e-16 ***
Ethnicity_Asian      -18.69      65.02  -0.287    0.774
Ethnicity_Caucasian  -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- *p*-value for both dummy variables is very high, indicating that there is no statistical difference in average credit card balance between the ethnicity categories.

Multiple Quantitative and Qualitative Predictors

- Not a problem
  - Use as many dummy variables as needed
- R creates dummy variables automatically for the qualitative predictors

In conclusion…

- Pros of Linear Regression Model:
  - Provides nice interpretable results
  - Works well on many real-world problems
- Cons of Linear Regression Model:
  - Assumes linear relationship between *response* and *predictors:*
    - Change in the response Y due to a one-unit change in X_i is constant
  - Assumes additive relationship (unless you add interaction terms)
    - Effect of changes in a predictor X_i on response Y is independent of the values of the other predictors

Logistic Regression

- In standard linear regression, the *response* is a *continuous variable*:

$$y = f(x_1, x_2, .., x_p) + \varepsilon = \beta_0 + B_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

- In logistic regression, the *response* is *qualitative*

Can a Qualitative Response be Converted into a Quantitative Response?

- *Iris Dataset:*
  - *Qualitative Response:* {Setosa, Virginica, Versicolor}
- Encode a quantitative response:

$$Y = \begin{cases} 1 & \text{if Setosa} \\ 2 & \text{if Virginica} \\ 3 & \text{if Versicolor} \end{cases}$$

- Then fit a linear regression model using least squares

Can a Qualitative Response be Converted into a Quantitative Response?

- What's the problem?
- Encoding implies an ordering of the *Iris* classes
  - Difference between Setosa and Virginica is same as difference between Viginica and Versicolor
  - Difference between Setosa and Versicolor is greatest

$$Y = \begin{cases} 1 & \text{if Setosa} \\ 2 & \text{if Virginica} \\ 3 & \text{if Versicolor} \end{cases}$$

Can a Qualitative Response be Converted into a Quantitative Response?

$$Y = \begin{cases} 1 & \text{if Versicolor} \\ 2 & \text{if Setosa} \\ 3 & \text{if Virginica} \end{cases}$$

$$Y = \begin{cases} 1 & \text{if Setosa} \\ 2 & \text{if Virginica} \\ 3 & \text{if Versicolor} \end{cases}$$

- Two different encodings
- Two different linear models will be produced
- Would lead to different predictions for the same test instance

Holds for qualitative values without a natural ordering

Can a Qualitative Response be Converted into a Quantitative Response?

- Another example:
  - *Response*: {Mild, Moderate, Severe}

$$Y = \begin{cases} 1 & \text{if Mild} \\ 2 & \text{if Moderate} \\ 3 & \text{if Severe} \end{cases}$$

Encoding is fine if the gap between Mild and Moderate, is about the same as Moderate to Severe.

Can a Qualitative Response be Converted into a Quantitative Response?

- In general, no natural way to convert a *qualitative* response with more than two levels into a *quantitative* response.
- Binary response:

$$Y = \begin{cases} 0 & \text{if Default on Loan} \\ 1 & \text{if No Default on Loan} \end{cases}$$

- Predict *Default* if $y_i < 0.5$
- Else predict *NoDefault*

- Predicted values may lie outside range of [0,1]
- Predicted values are not probabilities

Logistic Model

- Logistic Regression models the *probability* that *Y* belongs to a particular category
- For *Default* dataset:
  - Probability of *Default* given *Balance*:

$$\Pr(default = Yes \mid balance)$$

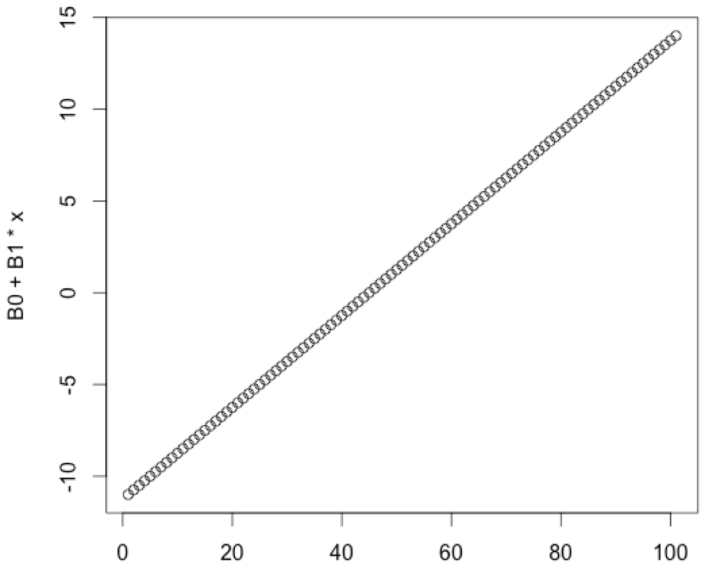
  - Values of *p(balance)* will range between 0 and 1.

Logistic Model

- $p(balance) > 0.5$
  - Predict *Default=Yes*
- … or for a conservative company
  - Lower the threshold and predict *Default=Yes* if $p(balance) > 0.1$

Linear Model vs. Logistic Model

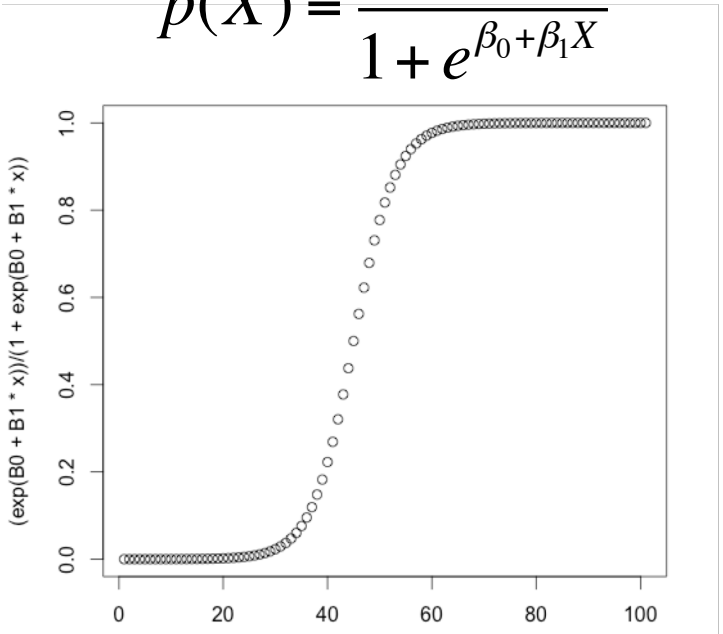
$$p(X) = \beta_0 + \beta_1 X$$




```
linear <- plot(B0 + B1*x)
```

```
x <- -50:50
B0 <- 1.5
B1 <- .250
```

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$




```
logistic <- plot((exp(B0 + B1*x))/(1+exp(B0 + B1*x)))
```

Linear Model vs. Logistic Model

- With linear model, can always predict $p(X) < 0$ for some values of X and $p(X) > 1$ for others
  - (unless X has limited range)
- Logistic Function: outputs between 0 and 1 for all values of X
  - (many functions meet this criteria, logistic regression uses the logistic function on next slide)

Logistic Function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

- Values of the odds close to 0 indicate very low probabilities.
  - On average, 1 in 5 people with an odds of $1/4$ will default. $\frac{0.2}{1-0.2} = \frac{1}{4}$, so $p(default) = 0.2$
- High odds indicate very high probabilities.

odds

… algebra …

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

log-odds (logit)

… algebra …

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

"Logit is linear in X."

Interpretation

- Linear:
  - "B_1 *gives the average change in Y with a one-unit increase in X.*"
  - "*Relationship is constant (straight line).*"
- Logistic:
  - "*Increasing X by one unit changes the log odds by* B_1, *(multiplies the odds by* e^{B1}*).*"
  - "*Relationship is not a straight line. The amount that Y changes depends on the current value of X.*"

Estimating the Regression Coefficients

- Usually the method of maximum likelihood is used (instead of least squares)
  - Reasoning is beyond the scope of this course
- R calculates "best" coefficients automatically for us

“Default” Dataset

- Simulated toy dataset
- 10,000 observations
- 4 variables
  - *Default*: {Yes, No} – whether customer defaulted on their debt
  - *Student*: {Yes, No} – whether customer is a student
  - *Balance*: average CC balance
  - *Income:* customer income

"Default" dataset

"Since B_1=0.0055, an increase in *balance* is associated with an increase in the probability of *default*."

Since *p*-value of *Balance* coefficient is tiny, it is statistically significant that there is an association between *Balance* and the probability of *Default*.

```
Call:
glm(formula = default ~ balance, family = binomial, data = Default)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A one-unit increase in *balance* is associated with an increase in the log odds of *default* by 0.0055 units.

“Default” dataset

- Making Predictions:

$$\hat{p}(1000) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576 = 0.576\%$$

$$\hat{p}(2000) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586 = 58.6\%$$

“Default” dataset

- Do students have a higher chance of default?

Multiple Logistic Regression

- How to prediction a *binary response* using multiple *predictors?*
- Can generalize the logistic function to *p* predictors:

$$p(X) = \frac{e^{\beta_0+\beta_1 X+\ldots+\beta_p X}}{1+e^{\beta_0+\beta_1 X+\ldots+\beta_p X}} \qquad \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Can use maximum likelihood to estimate the *p+1* coefficients.

“Default” dataset

- Full model using all three predictor variables:
  - *RESPONSE*
    - *Default*: {Yes, No} – whether customer defaulted on their debt
  - *PREDICTORS*
    - *(DUMMY VARIABLE USED) Student*: {Yes, No} – whether customer is a student
    - *Balance*: average CC balance
    - *Income:* customer income

Full Model

```
> glm.fit <- glm(default ~ balance+income+student, data=
family=binomial)
> summary(glm.fit)

Call:
glm(formula = default ~ balance + income + student, fam
    data = Default)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *p*-values associated with *student* and *balance* are very small, indicating that each of these variables is associated with the probability of *Default*.
- Coefficient for the dummy variable *student* is negative, indicating that students are less likely to default than nonstudents.

Model Comparison

```
Call:
glm(formula = default ~ student, family = binomial, data = Default)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

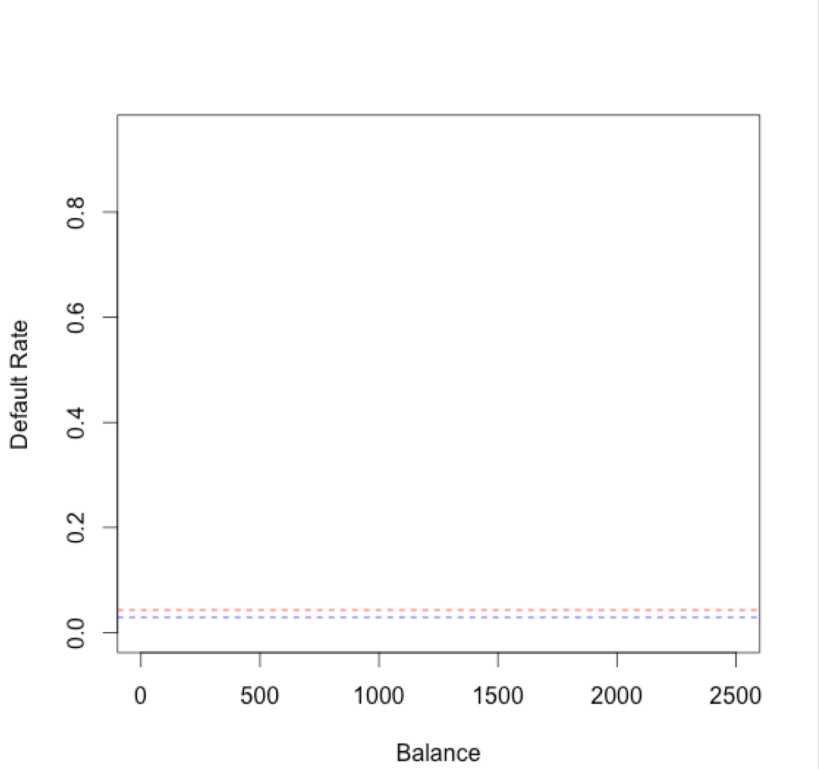
```
Call:
glm(formula = default ~ balance + income + student, family = binomi

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conflicting Results?

Conflicting Results?

- How is it possible for student status to be associated with an *increase* in probability of default when it was the only predictor, and now a *decrease* in probability of default once income is also factored in?

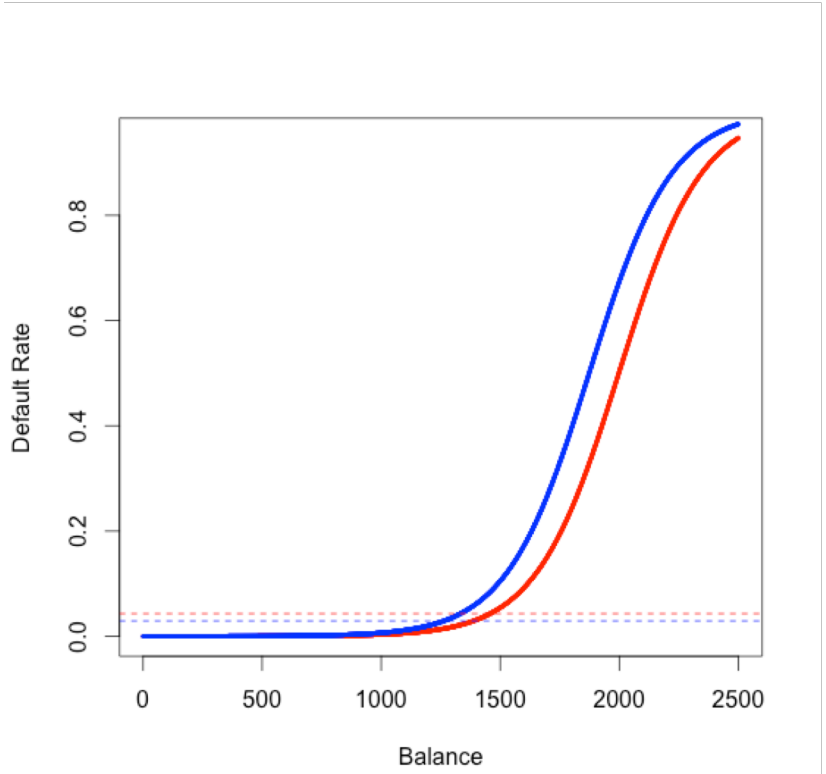
Interpretation



Default rate of students/non-students, averaged over all values of balance

Red = student
Blue = not a student

Interpretation

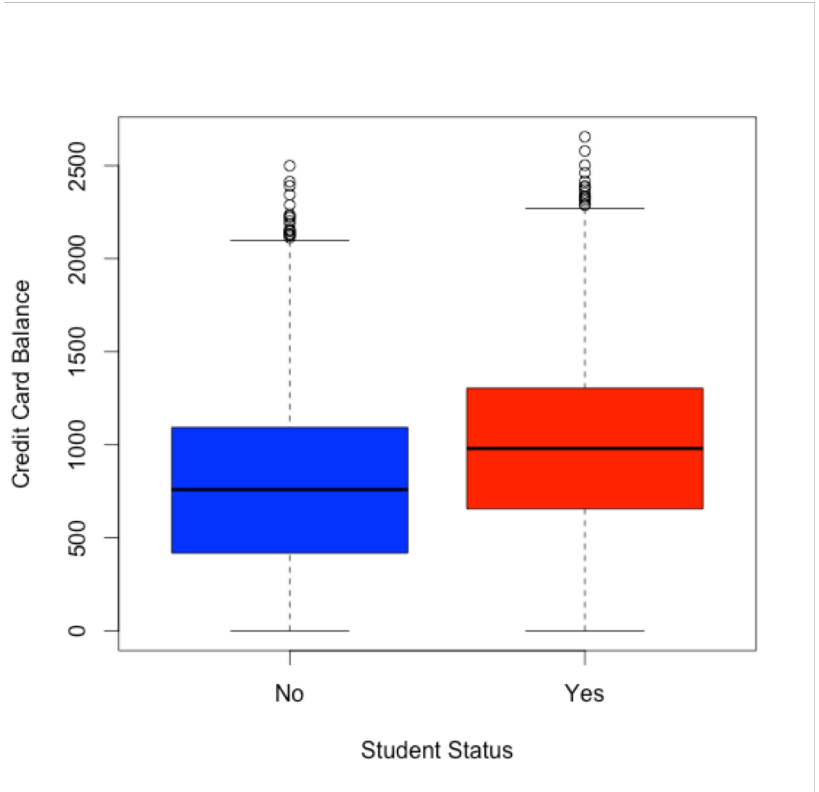


Default rate of students/non-students, as a function of balance value

Red = student
Blue = not a student

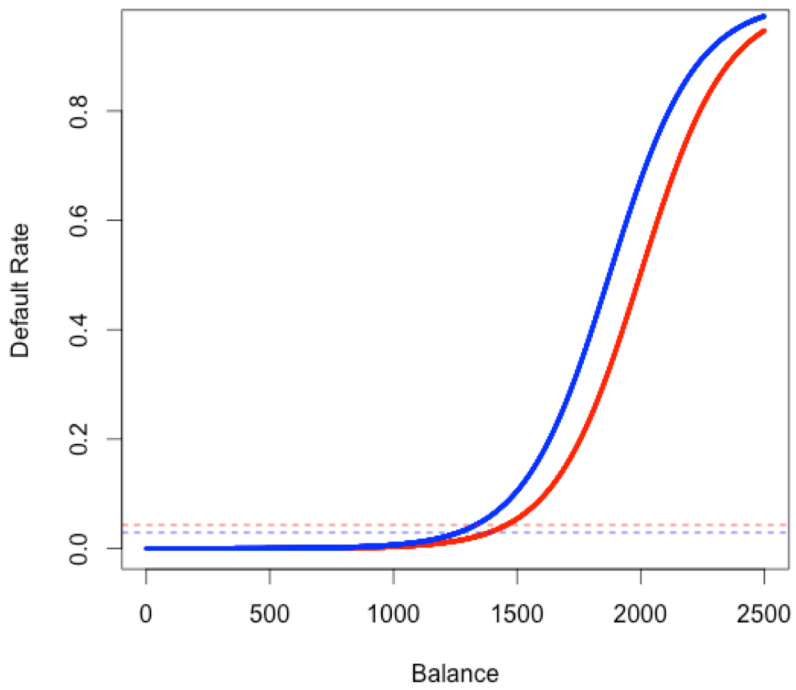
A student is less likely to default than a non-student, for a fixed value of balance.

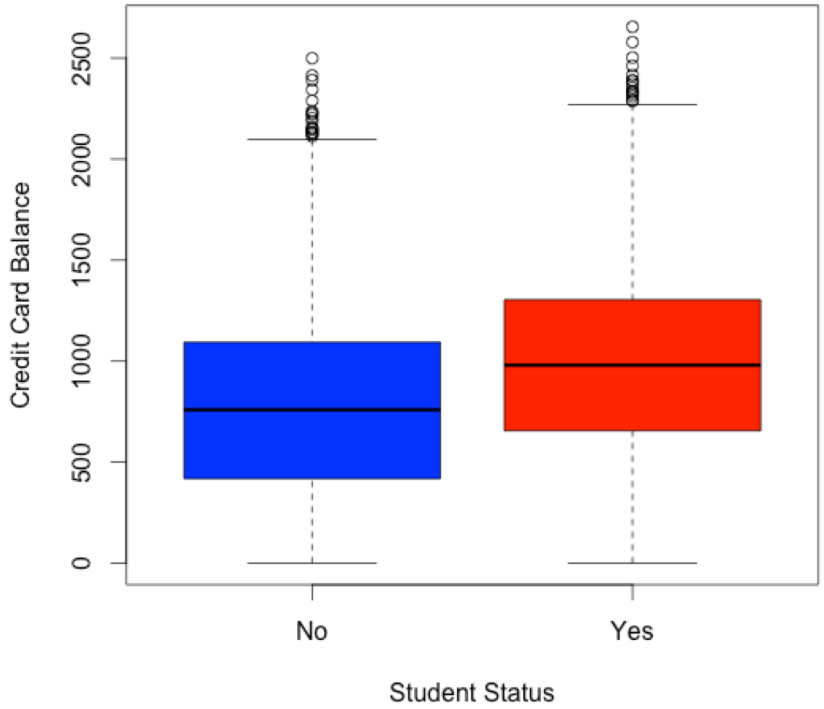
Interpretation



Variables *Student* and *Balance* are correlated.

- Students tend to hold higher levels of debt, which is associated with a higher probability of default.





- A student is riskier for default than a non-student if no information about the student's CC balance is available.
- But, that student is less risky than a non-student with the same CC balance.

Interpretation is Key

- In linear regression, results obtained using one *predictor* may be vastly different compared to when *multiple predictors* are used
  - Especially when there is correlation among the predictors

Logistic Regression: >2 Response Classes

- Yes, Two-class logistic regression models have multiple-class extensions.
- Yes, they are implemented in R.
- However, we'll use other data mining and statistical techniques instead.

References

- *Data Mining and Business Analytics in R*, 1st edition, Ledolter
- *An Introduction to Statistical Learning*, 1st edition, James et al.
- *Discovering Knowledge in Data*, 2nd edition, Larose et al.