Statistical Learning

Lecture 02a

ANU - RSFAS

Last Updated: Tue Mar 1 13:31:29 2022

Qualitative covariates - Linear Models - Review Continued

- Some predictors are not **quantitative** but are **qualitative**, taking a discrete set of values.
- These are also called **categorical predictors** or **factor variables**.
- For our **Credit Card Balance Data** there are four qualitative variables.
  - gender, student status,marital status, ethnicity

```r
library(ISLR)
library(faraway)
summary(Credit[, 1:6])
```

```
##        ID            Income           Limit           Rating     
##  Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0  
##  1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2  
##  Median :200.5   Median : 33.12   Median : 4622   Median :344.0  
##  Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9  
##  3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2  
##  Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0  
##      Cards            Age       
##  Min.   :1.000   Min.   :23.00  
##  1st Qu.:2.000   1st Qu.:41.75  
##  Median :3.000   Median :56.00  
##  Mean   :2.958   Mean   :55.67  
##  3rd Qu.:4.000   3rd Qu.:70.00  
##  Max.   :9.000   Max.   :98.00  
```

```r
summary(Credit[, -c(1:6)])
```

```
##    Education        Gender     Student   Married              Ethnicity
##  Min.   : 5.00   Male :193   No :360   No :155   African American: 99
##  1st Qu.:11.00   Female:207  Yes: 40   Yes:245   Asian           :102
##  Median :14.00                                   Caucasian       :199
##  Mean   :13.45
##  3rd Qu.:16.00
##  Max.   :20.00
##     Balance
##  Min.   :   0.00
##  1st Qu.:  68.75
##  Median : 459.50
##  Mean   : 520.01
##  3rd Qu.: 863.00
##  Max.   :1999.00
```

- Example: investigate the differences in **credit card balance** quantitative between males and females, ignoring the other variables. We create a new variable

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

- This leads to the following model:

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

```r
mod <- lm(Balance ~ Gender, data=Credit)
sumary(mod)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  529.536     31.988 16.5541   <2e-16
## Gender Male  -19.733     46.051 -0.4285   0.6685
##
## n = 400, p = 2, Residual SE = 460.22995, R-Squared = 0
```

- This leads to the following model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \begin{cases} 529.536 & \text{if } i^{th} \text{ person is female} \\ 529.536 + -19.733 & \text{if } i^{th} \text{ person is male} \end{cases}$$

Qualitative Predictors with More than Two Levels

- With more than two levels, we create additional dummy variables. For example, for the **ethnicity variable** we create two dummy variables. The first could be:

$$x_{i1} = \begin{cases} 1 & \text{if } i^{th} \text{ person is Asian} \\ 0 & \text{if } i^{th} \text{ person is no Asian} \end{cases}$$

and the second could be:

$$x_{i2} = \begin{cases} 1 & \text{if } i^{th} \text{ person is Caucasian} \\ 0 & \text{if } i^{th} \text{ person is no Caucasian} \end{cases}$$

- Then both of these variables can be used in the regression equation, in order to obtain the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i^{th} \text{ person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i^{th} \text{ person is Caucasion} \\ \beta_0 + \epsilon_i & \text{if } i^{th} \text{ person is AA} \end{cases}$$

- In Statistics , we say that the factor **Ethnicity** has three levels.
- Here **African American** is the baseline. When on group is set to **zero** and the others are compared to it this is called **Treatment Coding**.

- What does our model matrix $\boldsymbol{X}$ look like?

```
X <- model.matrix(~Ethnicity, data=Credit)
head(X)
```

```
##   (Intercept) EthnicityAsian EthnicityCaucasian
## 1           1              0                  1
## 2           1              1                  0
## 3           1              1                  0
## 4           1              1                  0
## 5           1              0                  1
## 6           1              0                  1
```

```r
mod2 <- lm(Balance ~ Ethnicity, data=Credit)
sumary(mod2)
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          531.000     46.319 11.4641   <2e-16
## EthnicityAsian       -18.686     65.021 -0.2874   0.7740
## EthnicityCaucasian   -12.503     56.681 -0.2206   0.8255
##
## n = 400, p = 3, Residual SE = 460.86508, R-Squared = 0
```

- If we want to test whether the factor is important, we can use an F-test (as we saw for testing generally whether all the variables are important).

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: Balance
##            Df   Sum Sq Mean Sq F value Pr(>F)
## Ethnicity   2    18454    9227  0.0434 0.9575
## Residuals 397 84321458  212397
```

Relevel() - Change the Reference Level

```
Credit$Ethnicity <- relevel(Credit$Ethnicity, ref="Caucasian")
mod2 <- lm(Balance ~ Ethnicity, data=Credit)
sumary(mod2)
```

```
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 518.4975    32.6699 15.8708   <2e-16
## EthnicityAfrican American    12.5025    56.6810  0.2206   0.8255
## EthnicityAsian               -6.1838    56.1216 -0.1102   0.9123
##
## n = 400, p = 3, Residual SE = 460.86508, R-Squared = 0
```

```
confint(mod2)
```

```
##                                2.5 %  97.5 %
## (Intercept)                 454.2699 582.725
## EthnicityAfrican American   -98.9300 123.935
## EthnicityAsian             -116.5165 104.149
```

ANOVA

- If we want to test whether the factor is important, we can use an F-test (as we saw for testing generally whether all the variables are important).

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: Balance
##            Df   Sum Sq Mean Sq F value Pr(>F)
## Ethnicity   2    18454    9227  0.0434 0.9575
## Residuals 397 84321458  212397
```

Multiple Regression Again

```r
mod3 <- lm(Balance ~ Income + Gender + Ethnicity, data=Credit)
sumary(mod3)
```

```
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                260.81564   43.59638  5.9825 4.928e-09
## Income                       6.05422    0.58178 10.4065 < 2.2e-16
## Gender Male                -24.33958   40.96297 -0.5942    0.5527
## EthnicityAfrican American  -6.44694   50.36344 -0.1280    0.8982
## EthnicityAsian              -4.80970   49.84464 -0.0965    0.9232
##
## n = 400, p = 5, Residual SE = 409.21795, R-Squared = 0.22
```

Multiple Regression Again

```
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: Balance
##            Df   Sum Sq  Mean Sq  F value Pr(>F)
## Income      1 18131167 18131167 108.2721 <2e-16 ***
## Gender      1    59022    59022   0.3525 0.5531
## Ethnicity   2     3286     1643   0.0098 0.9902
## Residuals 395 66146436   167459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Regression - ANOVA - Order Matters

```r
mod3 <- lm(Balance ~ Ethnicity + Income + Gender, data=Credit)
sumary(mod3)
```

```
##                              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 260.81564   43.59638  5.9825 4.928e-09
## EthnicityAfrican American  -6.44694   50.36344 -0.1280    0.8982
## EthnicityAsian              -4.80970   49.84464 -0.0965    0.9232
## Income                       6.05422    0.58178 10.4065 < 2.2e-16
## Gender Male                -24.33958   40.96297 -0.5942    0.5527
##
## n = 400, p = 5, Residual SE = 409.21795, R-Squared = 0.22
```

Multiple Regression - ANOVA - Order Matters

```
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: Balance
##            Df   Sum Sq  Mean Sq  F value Pr(>F)
## Ethnicity   2    18454     9227   0.0551 0.9464
## Income      1 18115899 18115899 108.1809 <2e-16 ***
## Gender      1    59122    59122   0.3531 0.5527
## Residuals 395 66146436   167459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Classification

- Qualitative variables take values in an unordered set $\mathcal{C}$, such as:

$$\text{eye colour} \in \{\text{brown, blue, green}\}$$
$$\text{email} \in \{\text{spam, ham}\}$$

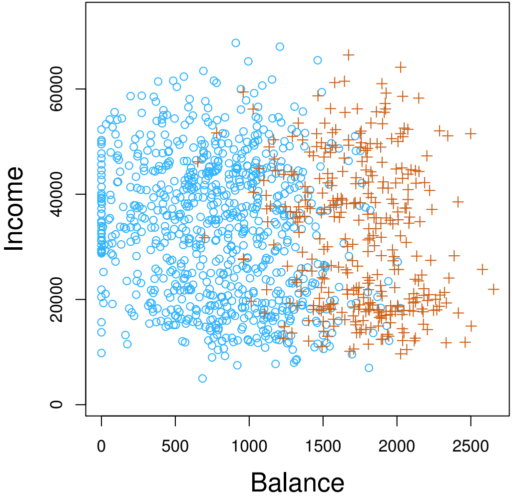
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.

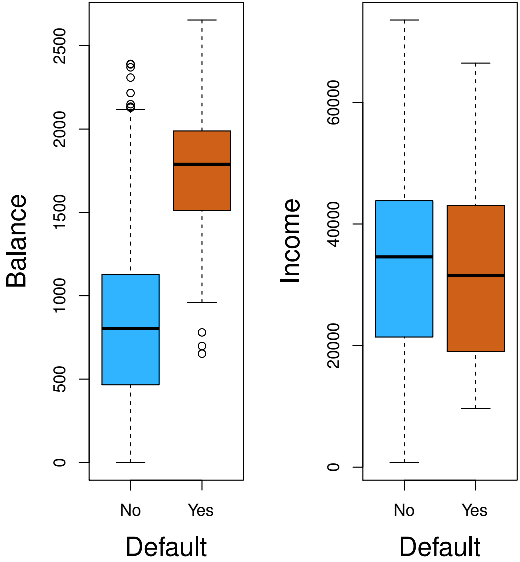
- Often we are more interested in estimating the probabilities that X belongs to each category in $\mathcal{C}$.

For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

Example: Credit Card Defualt

- The annual incomes and monthly credit card balances of a number of individuals.





Balance
0
500
1000
1500
2000
2500
No
Yes
Default
Income
0
20000
40000
60000
No
Yes
Default

Can we use Linear Regression?

- Suppose for the **Default** classification task that we code:

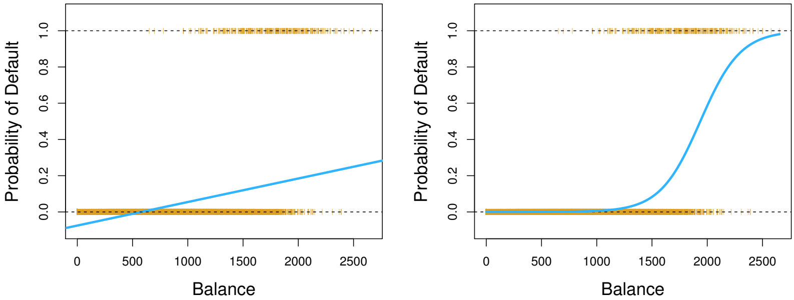
$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

- Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?

- As the population model: $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.

- However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

Linear versus Logistic Regression



- The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Linear Regression continued

- Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if Stroke} \\ 2 & \text{if Drug Overdose} \\ 3 & \text{if Epileptic Seizure} \end{cases}$$

- This coding suggests an ordering, and in fact implies that the difference between stroke and drug overdose is the same as between drug overdose and epileptic seizure.

- Linear regression is not appropriate here. Multiclass (Multinomial Logistic Regression} or Discriminant Analysis are more appropriate.

Logistic Regression - Back to CC Default

Let's write $p(X) = Pr(Y = 1|X)$ for short and consider using **balance** to predict **default**. Logistic regression uses the form

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

- No matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.
- A bit of rearrangement gives:

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

- This monotone transformation is called the log odds or logit transformation of $p(X)$.

Generalized Linear Models

- More generally, logistic regression falls into the class of generalized linear models.
- The basic idea is that we have data $\{y_1, x_1\}, \ldots, \{y_n, x_n\}$ where Y is believed to come from a particular distribution.

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_Y(y; \theta)$$

- Then we want to relate the $E[Y] = f(X)$.
- Suppose, usually, $E[Y] = \beta_0 + \beta_1 X$. However the $E[Y]$ may have a restriction on it. Has to be between 0 and 1, positive, etc.
- So we have a link function to link the mean to the linear predictor.

- For the **default data**, $Y = 0$ or 1. This suggests:

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$$

- Note: $E[Y] = Pr(Y = 1) = \theta$.
- Now let's relate that to a function of a covariate:

$$\theta = p(X) = g^{-1}(\beta_0 + \beta_1 X)$$

- Or

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

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

- For logistic regression, $g(X)$ is the log odds.

Estimation of GLMs

- Estimation is typically done through maximum likelihood estimation (MLE).

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

- This likelihood gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

- Most statistical packages can fit linear logistic regression models by maximum likelihood. In R we use the glm function to fit many different types of GLMs.

```r
library(ISLR)
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

```r
library(MASS)
```

```r
mod <- glm(default ~ balance, family=binomial, data=Default)
sumary(mod)
```

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.0651e+01  3.6116e-01 -29.492 < 2.2e-16
## balance      5.4989e-03  2.2037e-04  24.953 < 2.2e-16
##
## n = 10000 p = 2
## Deviance = 1596.45168 Null Deviance = 2920.64971 (Difference = 1324.19803)
```

Making Predictions

- What is our estimated probability of **default** for someone with a balance of $1000?

$$\begin{aligned}\hat{p}(X) &= \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X)} \\ &= \frac{exp(-10.6513 + 0.0055 \times 1000)}{1 + exp(-10.6513 + 0.0055 \times 1000)} = 0.006\end{aligned}$$

```
predict(mod, newdata=data.frame(balance=1000),
        type="response")
```

```
##           1
## 0.005752145
```

- What is our estimated probability of **default** for someone with a balance of $2000?

$$
\begin{aligned}
\hat{p}(X) &= \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X)} \\
&= \frac{exp(-10.6513 + 0.0055 \times 2000)}{1 + exp(-10.6513 + 0.0055 \times 2000)}
\end{aligned}
$$

```r
predict(mod, newdata=data.frame(balance=2000),
        type="response")
```

```
##         1
## 0.5857694
```

- Lets do it again, using **student** as the predictor.

```
mod2 <- glm(default ~ student, family=binomial, data=Default)
sumary(mod2)
```

```
##              Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) -3.504128   0.070713 -49.5542 < 2.2e-16
## studentYes   0.404887   0.115019   3.5202 0.0004313
##
## n = 10000 p = 2
## Deviance = 2908.68306 Null Deviance = 2920.64971 (Difference = 11.96665)
```

$$\widehat{Pr}(\text{default=Yes} \mid \text{student=Yes}) = \frac{exp(-3.5041 + 0.4049 \times 1)}{1 + exp(-3.5041 + 0.4049 \times 1)}$$

```
predict(mod2, newdata=data.frame(student="Yes"),
        type="response")
```

```
##          1
## 0.04313859
```

$$\widehat{Pr}(\text{default=Yes} \mid \text{student=No}) = \frac{exp(-3.5041 + 0.4049 \times 0)}{1 + exp(-3.5041 + 0.4049 \times 0)}$$

```
predict(mod2, newdata=data.frame(student="No"),
        type="response")
```

```
##          1
## 0.02919501
```

Logistic Regression with Several Variables

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

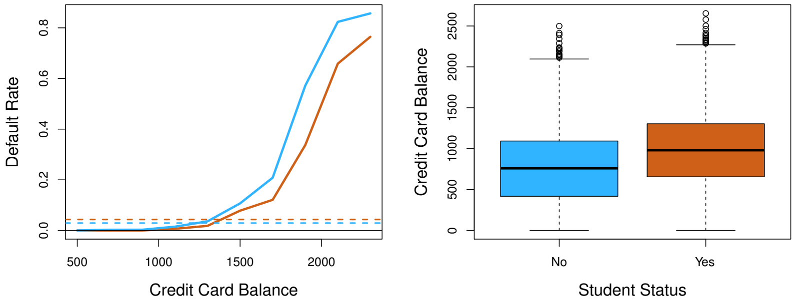
$$p(X) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}$$

```r
mod3 <- glm(default ~ balance + student, family=binomial, data=Default)
sumary(mod3)
```

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.0749e+01  3.6919e-01 -29.116 < 2.2e-16
## balance      5.7381e-03  2.3185e-04  24.750 < 2.2e-16
## studentYes  -7.1488e-01  1.4752e-01  -4.846  1.26e-06
##
## n = 10000 p = 3
## Deviance = 1571.68160 Null Deviance = 2920.64971 (Difference = 1348.96811)
```

- Why is coefficient for student negative, while it was positive before?

Confounding



- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

Another Binary Regression Models

- Another popular choice is the probit model.
- All we do is change the link function:

$$Pr(Y = 1|X) = p(X) = \Phi(\beta_0 + \beta_1 X)$$

Where $\Phi(\cdot)$ is the CDF of a standard normal distribution (so it ranges from 0 to 1).

- Rewriting we have:

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

- Actually there are other link functions:
  - complementary log-log: $\beta_0 + \beta_1 X = log(1 - log(p(X))$
  - cauchit: $\beta_0 + \beta_1 X = \tan^{-1}(\pi(p(X) - 1/2))$
- Note: π is the standard constant ≈ 3.14.

```
mod.probit <- glm(default ~ balance,
                  family=binomial(link="probit"), data=Default)
sumary(mod.probit)
```

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -5.35390527  0.16974731 -31.540 < 2.2e-16
## balance      0.00271069  0.00010884  24.905 < 2.2e-16
##
## n = 10000 p = 2
## Deviance = 1605.51267 Null Deviance = 2920.64971 (Difference = 1315.13704)
```

```r
mod.cloglog <- glm(default ~ balance,
                   family=binomial(link="cloglog"),
                   data=Default)
mod.cauchit <- glm(default ~ balance,
                   family=binomial(link="cauchit"),
                   data=Default)

x <- seq(500, 3000, by=1)
```

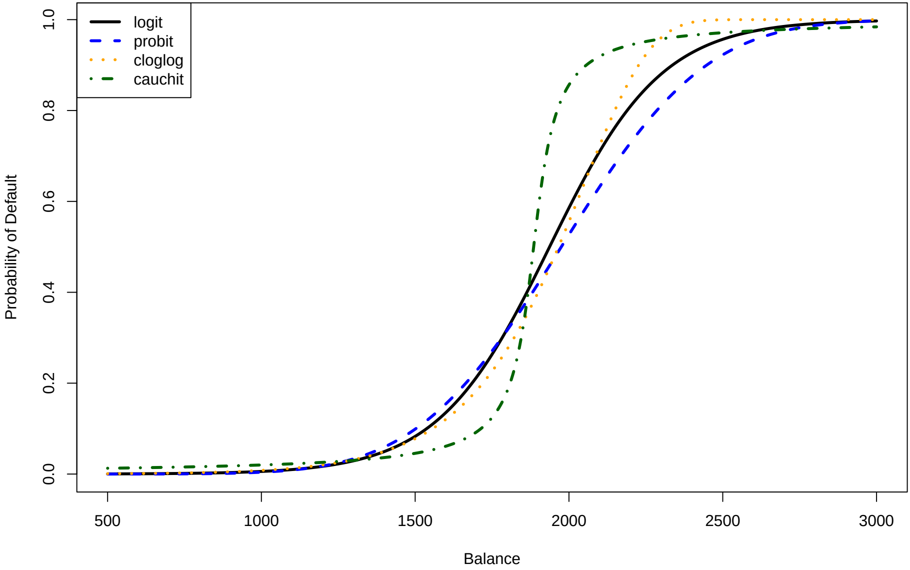
```r
plot(x, predict(mod, data.frame(balance=x),
                type="response"), type="l", lwd=3,
     ylab="Probability of Default", xlab="Balance")

lines(x, predict(mod.probit, data.frame(balance=x),
                 type="response"), lty=2, lwd=3,
      col="blue", ylab="Probability of Default",
      xlab="Balance")

lines(x, predict(mod.cloglog, data.frame(balance=x),
                 type="response"), lty=3, lwd=3,
      col="orange", ylab="Probability of Default",
      xlab="Balance")

lines(x, predict(mod.cauchit, data.frame(balance=x),
                 type="response"), lty=4, lwd=3,
      col="dark green", ylab="Probability of Default",
      xlab="Balance")

legend("topleft", c("logit", "probit", "cloglog", "cauchit"),
       col=c("black", "blue", "orange", "dark green"),
       lty=c(1,2,3,4), lwd=3)
```


logit
probit
cloglog
cauchit
Probability of Default
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
2500
3000
Balance

For GLMs

- If we want to test whether the factor is important, we can use an χ^2-test ("chi-squared test").
- Let's look at modeling Gender (Y) by using Ethnicity (X). Not that interesting but as an example.

```r
mod3 <- glm(Gender ~ Ethnicity, data=Credit, family="binomial")
sumary(mod3)
```

```
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.050262   0.141821 -0.3544   0.7230
## EthnicityAfrican American  0.030059   0.246011  0.1222   0.9028
## EthnicityAsian             -0.106924   0.244073 -0.4381   0.6613
##
## n = 400 p = 3
## Deviance = 553.75391 Null Deviance = 554.02764 (Difference = 0.27374)
```

```r
anova(mod3, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Gender
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        399     554.03
## Ethnicity  2  0.27374       397     553.75   0.8721
```