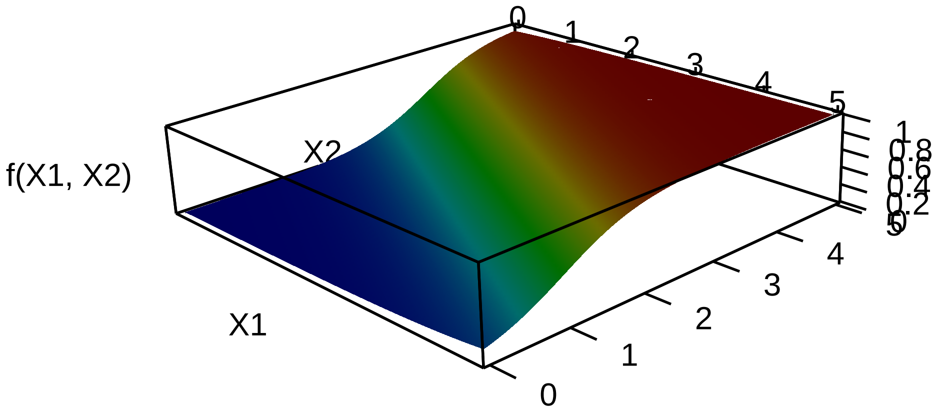
Multiple Logistic Regression

Logistic Regression with Multiple Explanatory Variables

We will now extend logistic regression to allow for p explanatory variables which may be either quantitative or categorical.

$$P(Y_i = 1 | X_{i1}, \dots, X_{ip}) = p(X_{i1}, \dots, X_{ip}) = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}$$

Illustration with $p = 2$ explanatory variables:



Running Example

This example is adapted from section 4.3 of ISLR. We have information on ten thousand customers; our goal is to predict which customers will default on their credit card debt.

```
head(Default, 4)
```

```
##   default student   balance   income
## 1      No      No  729.5265 44361.63
## 2      No     Yes  817.1804 12106.13
## 3      No      No 1073.5492 31767.14
## 4      No      No  529.2506 35704.49
```

Example 1: Two Quantitative Variables

Let's try using `balance` and `income` as explanatory variables.

```r
fit <- train(
  form = default ~ balance + income,
  data = Default,
  family = "binomial", # this is an argument to glm; response is 0 or 1, binomial
  method = "glm", # method for fit; "generalized linear model"
  trControl = trainControl(method = "none")
)
summary(fit)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

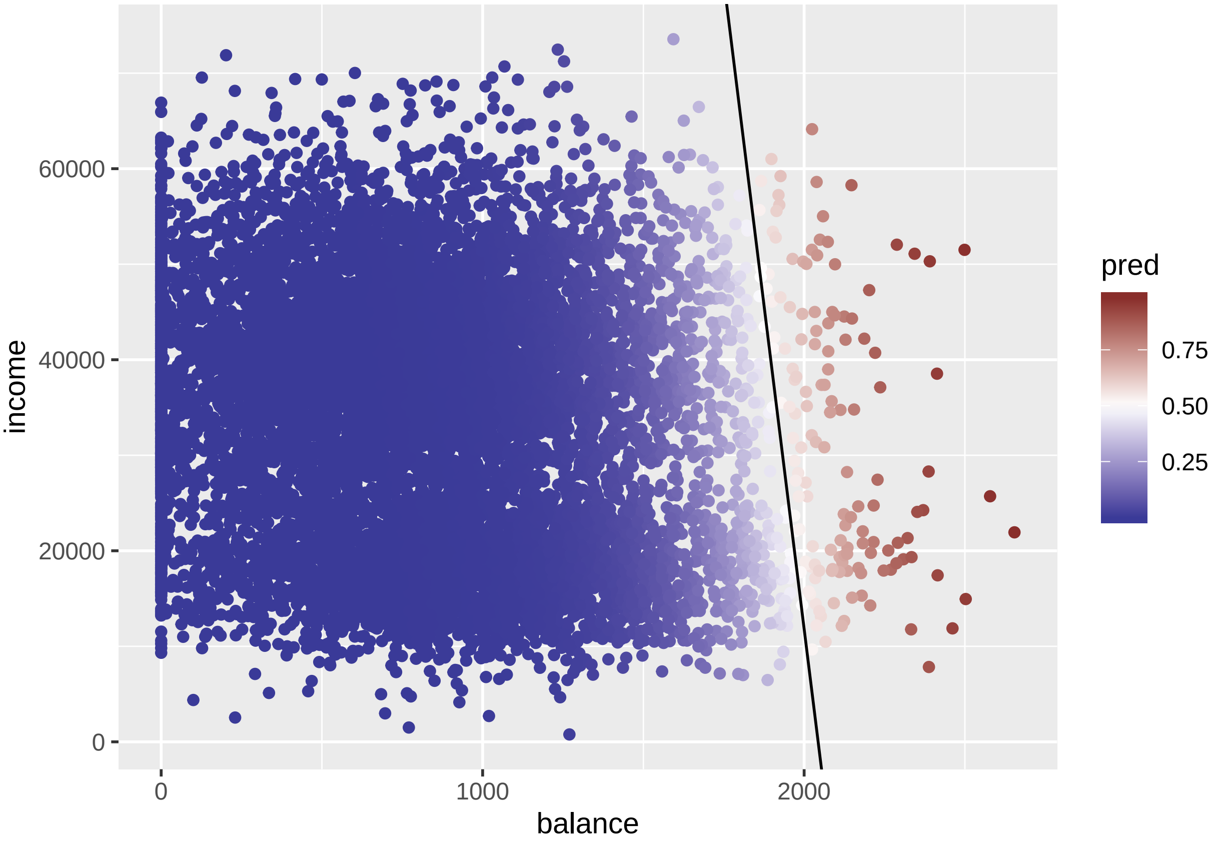
(a) What is the estimated equation for this model?

(b) What is the decision boundary?

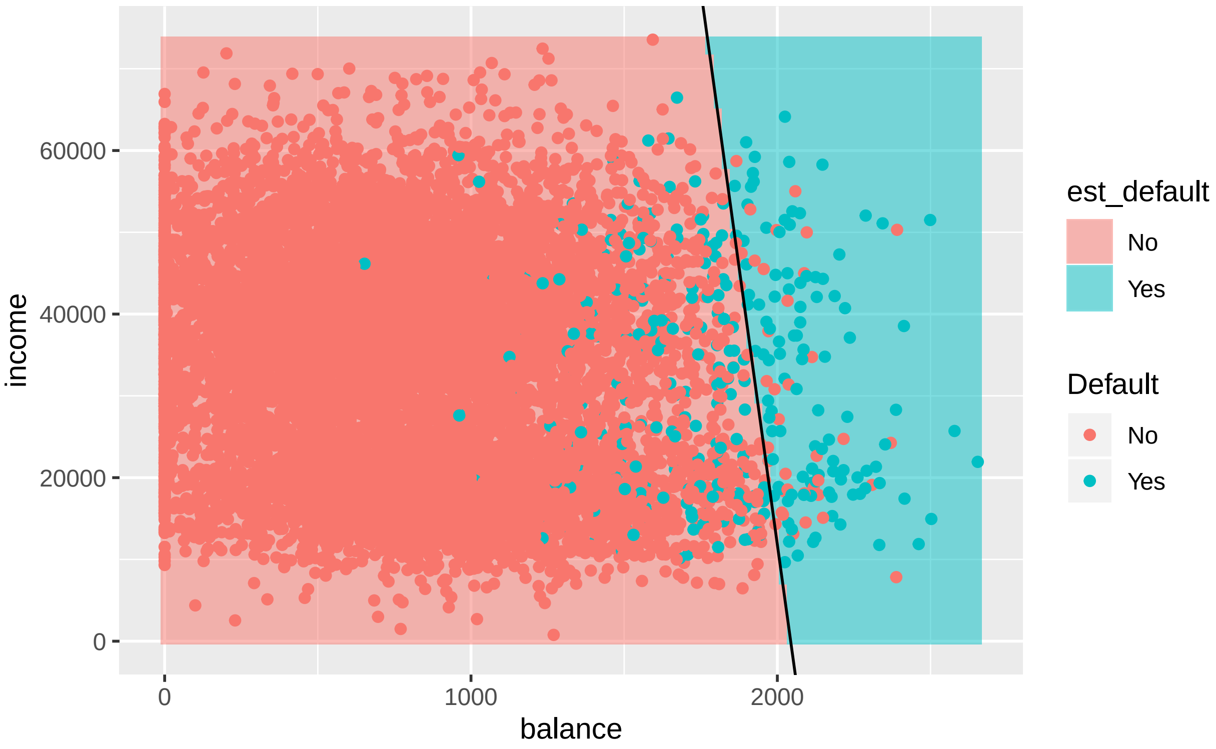
$$0.5 = \hat{P}(Y_i = 1 | X_{i1}, \ldots, X_{ip}) = \hat{f}_1(X_{i1}, \ldots, X_{ip}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}}}$$

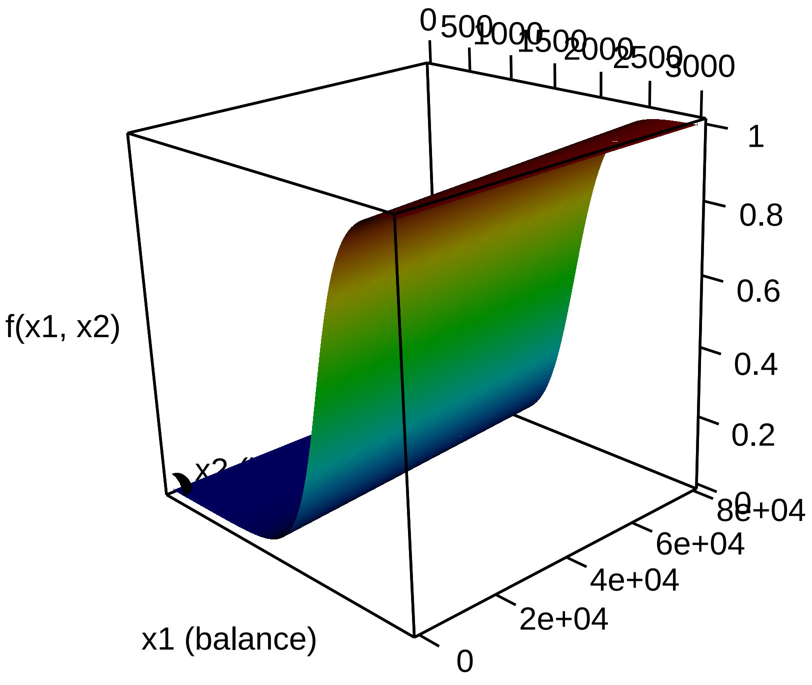
Plots

```r
df2 <- Default %>%
  mutate(
    pred = predict(fit, type = "prob")[["Yes"]]
  )

ggplot(data = df2, mapping = aes(x = balance, y = income, color = pred)) +
  geom_point() +
  geom_abline(intercept = 1.154e+01 / 2.081e-05, slope = - 5.647e-03 / 2.081e-05) +
  scale_color_gradient2(low = scales::muted("blue"), high = scales::muted("red"), midpoint = 0.5)
```



```r
max_balance <- max(Default$balance)
max_income <- max(Default$income)
background <- expand.grid(
    balance = seq(from = 0, to = max_balance, length = 101),
    income = seq(from = 0, to = max_income, length = 101))
background <- background %>%
  mutate(
    est_prob_default = predict(fit, newdata = background, type = "prob")[["Yes"]],
    est_default = predict(fit, newdata = background)
  )

ggplot() +
  geom_raster(data = background,
    mapping = aes(x = balance, y = income, fill = est_default), alpha = 0.5) +
  geom_point(data = Default, mapping = aes(x = balance, y = income, color = default)) +
  scale_color_discrete("Default") +
geom_abline(intercept = 1.154e+01 / 2.081e-05, slope = - 5.647e-03 / 2.081e-05)
```





Example 2: One Categorical Explanatory variable

```
fit <- train(
  form = default ~ student,
  data = Default,
  family = "binomial", # this is an argument to glm; response is 0 or 1, binomial
  method = "glm", # method for fit; "generalized linear model"
  trControl = trainControl(method = "none")
)
summary(fit)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2970  -0.2970  -0.2434  -0.2434   2.6585
##
## 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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 2908.7  on 9998  degrees of freedom
## AIC: 2912.7
##
## Number of Fisher Scoring iterations: 6
```

Similar to the use of categorical explanatory variables in linear models, R has created a new indicator variable for use in the regression:

$$X_{i1} = \text{studentYes}_i = \begin{cases} 1 \text{ if customer } i \text{ is a student} \\ 0 \text{ otherwise} \end{cases}$$

(a) What is the estimated equation for this model?

(b) What is the predicted probability of default for a non-student?

```r
predict(fit, newdata = data.frame(student = "No"), type = "prob")[["Yes"]]
```

```
## [1] 0.02919501
```

```r
# compare to...
exp(-3.50413) / (1 + exp(-3.50413))
```

```
## [1] 0.02919495
```

```r
Default %>%
  filter(student == "No") %>%
  summarize(prop_default = mean(default == "Yes"))
```

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

(c) What are the estimated odds of default for a non-student?

(d) What is the predicted probability of default for a student?

```r
predict(fit, newdata = data.frame(student = "Yes"), type = "prob")[["Yes"]]
```

```
## [1] 0.04313859
```

```r
# compare to...
exp(-3.50413 + 0.40489) / (1 + exp(-3.50413 + 0.40489))
```

```
## [1] 0.04313862
```

```r
Default %>%
  filter(student == "Yes") %>%
  summarize(prop_default = mean(default == "Yes"))
```

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

(e) What are the estimated odds of default for a student?

(f) What is the interpretation of $\hat{\beta}_1$?

```
exp(0.40489)
```

```
## [1] 1.499138
```

(g) Does someone's student status have a statistically significant association with whether or not they default?

Note about decision boundaries

- In this example, our predicted class is 0 for all values of x_i!
- In general, a decision boundary need not exist; this often happens with categorical explanatory variables.

Example 3: All 3 Explanatory Variables

```
fit <- train(
  form = default ~ student + balance + income,
  data = Default,
  family = "binomial", # this is an argument to glm; response is 0 or 1, binomial
  method = "glm", # method for fit; "generalized linear model"
  trControl = trainControl(method = "none")
)
summary(fit)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

(a) What is the estimated equation for this model?

(b) Does an individual's student status have a statistically significant association with whether or not they default? Compare your estimate to that from example 2.

(c) Does an individual's income have a statistically significant association with whether or not they default? Compare to your result from example 1.

(d) What is the estimated equation for non-students?

(e) What is the estimated equation for students?

(f) What is the interpretation of the coefficient for `studentYes`? Note that $e^{-0.6468} \approx 0.523719$.

(g) What is the interpretation of the coefficient for `balance`? Note that $e^{0.005737} \approx 1.005753$. Is it helpful to consider that $e^{(0.005737*100)} \approx 1.775$?

(h) Tests about more than one coefficient

This is a little artificial, but to demonstrate the code let's consider a test of the hypotheses that neither of the `student` and the `income` variables are related to the probability that a person will default. We will:

- fit a reduced model (similar to what we would do in a `lm` context)
- call `anova` to compare the reduced and full model
  - Unlike `anova` comparisons with linear models, we need to specify a `test` argument to `anova`. A common option is `test = "LRT"` (for likelihood ratio test). This is a large-sample approximate test procedure (see Stat 343).

```r
fit_reduced <- train(
  form = default ~ balance,
  data = Default,
  family = "binomial", # this is an argument to glm; response is 0 or 1, binomial
  method = "glm", # method for fit; "generalized linear model"
  trControl = trainControl(method = "none")
)
anova(fit_reduced$finalModel, fit$finalModel, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: .outcome ~ balance
## Model 2: .outcome ~ studentYes + balance + income
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9998     1596.5
## 2      9996     1571.5  2   24.907 3.904e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ethical Considerations

In the U.S., there is a history of discrimination against demographic groups in granting loans. The Equal Credit Opportunity Act was passed in 1974, and "makes it unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction, on the basis of race, color, religion, national origin, sex, marital status, or age (provided the applicant has the capacity to contract)" (https://en.wikipedia.org/wiki/Equal_Credit_Opportunity_Act).

Our model uses the covariates `balance`, `income`, and `student` to predict probability of loan default, which are allowed by the law. However, it's a fact that some of the covariates in our model (like `balance` and `income`) are correlated with protected characteristics like race, sex, or marital status. At a population level, the model we've developed would deem women and people of color creditworthy at lower rates than other groups.