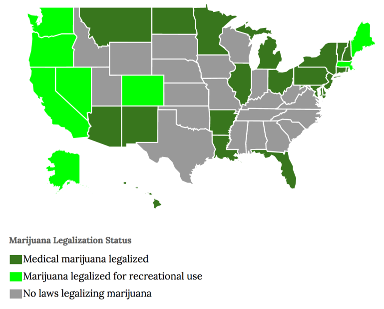
THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Logistic Regression 2

Lecture 23

STA 371G

Should pot be legal?



(Map source)

- The General Social Survey is an annual survey of attitudes and behaviors that has been conducted since the 1970s
- Let's use the GSS to examine the question of whether Americans think pot should be legalized
- An increasing number of states have done so already!

Response variable:

- **legal**: Answer to "Do you think the use of marijuana should be made legal or not?"

Response variable:

- **legal**: Answer to "Do you think the use of marijuana should be made legal or not?" This is binary (yes/no), so we'll need to use logistic regression.

Response variable:

- **legal**: Answer to “Do you think the use of marijuana should be made legal or not?” This is binary (yes/no), so we’ll need to use logistic regression.

Predictor variables:

- **year**: The year of the survey (1975-2014)
- **age**: The age of the respondent
- **schooling**: Number of years of schooling (e.g., 12 = HS degree, 16 = bachelor’s)
- **philosophy**: Political philosophy (on the spectrum of liberal to conservative)

Let's start by building a model using only the year variable:

```
model1 <- glm(legal ~ year, data=pot, family=binomial)
summary(model1)
```

```
Call:
glm(formula = legal ~ year, family = binomial, data = pot)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1202  -0.8596  -0.7330   1.3005   1.8827

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -75.022369   2.368408  -31.68   <2e-16 ***
year          0.037183   0.001187   31.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34665  on 28335  degrees of freedom
Residual deviance: 33646  on 28334  degrees of freedom
AIC: 33650

Number of Fisher Scoring iterations: 4
```

Evaluating model fit: Take 1

Our baseline prediction percentage is 69.9% (this is how many cases we'd predict correctly if we just predicted legal $= 0$ for everyone).

Evaluating model fit: Take 1

Our baseline prediction percentage is 69.9% (this is how many cases we'd predict correctly if we just predicted legal = 0 for everyone).

How well do we do by using the model?

```
predicted.legal <- (predict(model1, type='response') >= 0.5)
actual.legal <- (pot$legal == 1)
sum(predicted.legal == actual.legal) / nrow(pot)

[1] 0.6990401
```

Evaluating model fit: Take 1

Our baseline prediction percentage is 69.9% (this is how many cases we'd predict correctly if we just predicted legal = 0 for everyone).

How well do we do by using the model?

```
predicted.legal <- (predict(model1, type='response') >= 0.5)
actual.legal <- (pot$legal == 1)
sum(predicted.legal == actual.legal) / nrow(pot)

[1] 0.6990401
```

No better than a naive model that just predicts the same for everyone!

Evaluating model fit: Take 2

Let's also try computing McFadden's pseudo-R^2:

$$\text{Pseudo-}R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}} = 1 - \frac{33645.96}{34664.87} = 0.03$$

Evaluating model fit: Take 2

Let's also try computing McFadden's pseudo-R^2:

$$\text{Pseudo-}R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}} = 1 - \frac{33645.96}{34664.87} = 0.03$$

Both metrics show us that year does not help us predict attitude towards legalization very well (but we wouldn't expect it to — why not?)

Improving the model

Let’s add more predictors to the model:

- Years of schooling
- Age of respondent
- Political philosophy
- Gender

Interpreting the coefficients

Let's interpret the coefficients:

```
model2 <- glm(legal ~ year + age + schooling + philosophy
                      + gender, data=pot, family=binomial)
```

```
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -80.242      2.560 -31.343     0.00
year                              0.039      0.001  30.663     0.00
age                              -0.018      0.001 -20.956     0.00
schooling                         0.061      0.005  12.524     0.00
philosophyExtremely liberal       1.730      0.085  20.412     0.00
philosophyExtrmly conservative   -0.009      0.097  -0.089     0.93
philosophyLiberal                 1.414      0.055  25.645     0.00
philosophyModerate                0.605      0.046  13.047     0.00
philosophySlghtly conservative    0.372      0.053   6.954     0.00
philosophySlightly liberal        0.974      0.054  17.987     0.00
genderMale                       -0.016      0.030  -0.549     0.58
```

Interpreting the coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-80.242	2.560	-31.343	0.00
year	0.039	0.001	30.663	0.00
age	-0.018	0.001	-20.956	0.00
schooling	0.061	0.005	12.524	0.00
philosophyExtremely liberal	1.730	0.085	20.412	0.00
philosophyExtrmly conservative	-0.009	0.097	-0.089	0.93
philosophyLiberal	1.414	0.055	25.645	0.00
philosophyModerate	0.605	0.046	13.047	0.00
philosophySlghtly conservative	0.372	0.053	6.954	0.00
philosophySlightly liberal	0.974	0.054	17.987	0.00
genderMale	-0.016	0.030	-0.549	0.58

All else being equal, being a year older decreases the predicted *odds* that you will support marijuana legalization by 1.8% (since $e^{-0.018} = 0.982$ and $1 - 0.982 = 0.018$).

Evaluating model fit: Take 1

Recall that our baseline prediction percentage is 69.9% (this is how many cases we'd predict correctly if we just predicted legal $= 0$ for everyone).

Evaluating model fit: Take 1

Recall that our baseline prediction percentage is 69.9% (this is how many cases we'd predict correctly if we just predicted legal = 0 for everyone).

How well do we do by using the model?

```
predicted.legal <- (predict(model2, type='response') >= 0.5)
actual.legal <- (pot$legal == 1)
sum(predicted.legal == actual.legal) / nrow(pot)

[1] 0.721
```

Evaluating model fit: Take 2

$$\text{Pseudo-}R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}} = 1 - \frac{31593.96}{34664.87} = 0.09$$

Evaluating model fit: Take 2

$$\text{Pseudo-}R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}} = 1 - \frac{31593.96}{34664.87} = 0.09$$

Is it surprising that our measures of model fit are fairly low?

Testing the overall null hypothesis

Like with linear regression, there is an overall null hypothesis for the model that all coefficients (except the intercept) are 0 in the population.

Testing the overall null hypothesis

Like with linear regression, there is an overall null hypothesis for the model that all coefficients (except the intercept) are 0 in the population.

To test this, we can use a *likelihood-ratio test* (the likelihood measures how likely we are to see a particular set of data if a particular model is correct).

Testing the overall null hypothesis

Like with linear regression, there is an overall null hypothesis for the model that all coefficients (except the intercept) are 0 in the population.

To test this, we can use a *likelihood-ratio test* (the likelihood measures how likely we are to see a particular set of data if a particular model is correct).

We first have to define a null model (with no predictors), just like we did for stepwise regression:

```
null <- glm(legal ~ 1, data=pot, family=binomial)
```

Testing the overall null hypothesis

Now we can test our current model against the null model:

```
library(lmtest)
lrtest(null, model2)

Likelihood ratio test

Model 1: legal ~ 1
Model 2: legal ~ year + age + schooling + philosophy + gender
  #Df LogLik Df Chisq Pr(>Chisq)
1   1 -17332
2  11 -15797 10  3071     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing the overall null hypothesis

Now we can test our current model against the null model:

```
library(lmtest)
lrtest(null, model2)

Likelihood ratio test

Model 1: legal ~ 1
Model 2: legal ~ year + age + schooling + philosophy + gender
  #Df LogLik Df Chisq Pr(>Chisq)
1   1 -17332
2  11 -15797 10  3071     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since $p < 2 \times 10^{-16}$, we can reject the overall model null hypothesis (not surprising since we had many significant coefficients).