Data Mining with Decision Trees

Wei-Yin Loh

Department of Statistics

University of Wisconsin–Madison

CASE STUDY:

Equitability of treatment

in Army judicial proceedings

The problem

In the early 1990s, concern was expressed that minorities are disproportionately represented in the Army's justice system

The facts

1. Proportion of minority offenders in the Army justice system significantly exceeds proportion of minorities in the Army
2. This overrepresentation is even more pronounced in the civilian sector
3. But the Army is a selective environment where recruits must meet certain entry requirements, with the expectation that this would result in a pattern of offenses generally matched across ethnic groups

Purpose of study

Provide an understanding of the conditions which characterize involvement in the judicial process, which may provide insights to remedy the problem of overrepresentation

Study objectives

1. Assess whether minority soldiers were treated as equitably as White soldiers using official court-martial data
2. Identify any specific factors in the data which could imply nonequitable treatment

Study scope

Army-wide court-martial cases over six years (1987-92) dealing with enlisted personnel, but excluding NCOs and limited to Black and White personnel — 12,177 total cases

Study effort

US Army Concepts Analysis Agency with participation of W.-Y. Loh

Study sponsor

Office of the Deputy Chief of Staff for Personnel

Reference: US Army's Center for Strategy and Force Evaluation Study Report CAA-SR-93-14, December 1993

Main assumptions/limitations of the study

- Focus is on the Army's formal judicial process, the court-martial
- Focus does not include other factors which may exist pre-trial, such as: enforcement activities and aspects of individual behavior, which may fall along racial lines
- Data to characterize pre-trial conditions are not available on an authoritative or systematic basis

	FY 87	FY 88	FY 89	FY 90	FY 91	FY 92
Enlisted strength (percent)						
Total	666,000	654,600	652,000	623,500	585,100	511,336
White	62%	61%	60%	59%	59%	58.4%
Black	29%	31%	32%	32%	32%	31.5%
Other	9%	8%	8%	9%	9%	10.1%
White/Black	2.1	2.0	1.9	1.8	1.8	1.9
Enlisted offenders (percent)						
Total	2,693	2,669	2,548	2,401	1,830	1,770
White	52%	52%	49%	47%	47%	43%
Black	44%	43%	46%	48%	48%	51%
Other	4%	5%	5%	5%	5%	6%
White/Black	1.2	1.2	1.1	1.0	1.0	0.8

Methodology

Task 1. Data acquisition and consolidation

Task 2. Factor identification — PROCESS and SOLDIER variables

Task 3. Data analysis

1. Factor-pair analysis — cross-tabs of RACE vs. PROCESS variables
2. Factor-set analysis — discriminant and decision tree analyses

Task 4. Assessment of differences in treatment

Issues for analysis

1. What factors should be used to characterize the court-martial proceedings to facilitate recognition of any differences in treatment?
2. Are there differences in the treatment of offenders, by race, in the court-martial proceedings?
3. Are there factors in the data which could imply nonequitable treatment?

PROCESS factors (variables)

1. Number of charges
2. Time faced on charges (months)
3. Nature of highest charge — crime involving substances, property, persons, general order, or military order
4. Plea to charges — guilty or not guilty
5. Pre-trial agreement — present or not
6. Trial type — general, bad conduct, or special court-martial
7. Trial board type — military judge, officers, or officers & enlisted personnel
8. Length of confinement (months)
9. Nature of discharge — none, bad conduct, or dishonorable
10. Reduction in charges (percent)
11. Reduction in confinement (percent)

SOLDIER factors (variables)

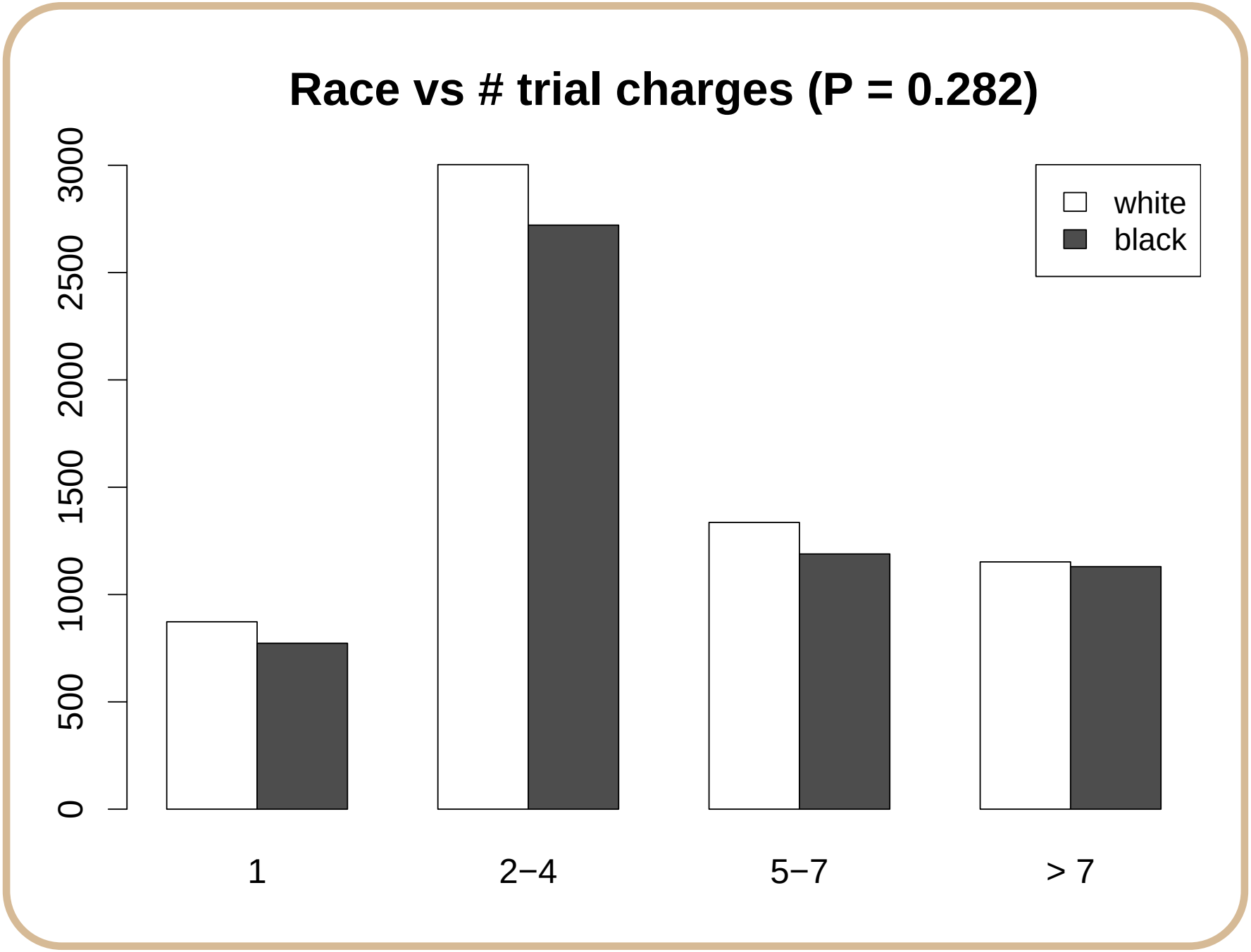
1. Race
2. Age
3. Gender
4. Education
5. Technical test score
6. Service time

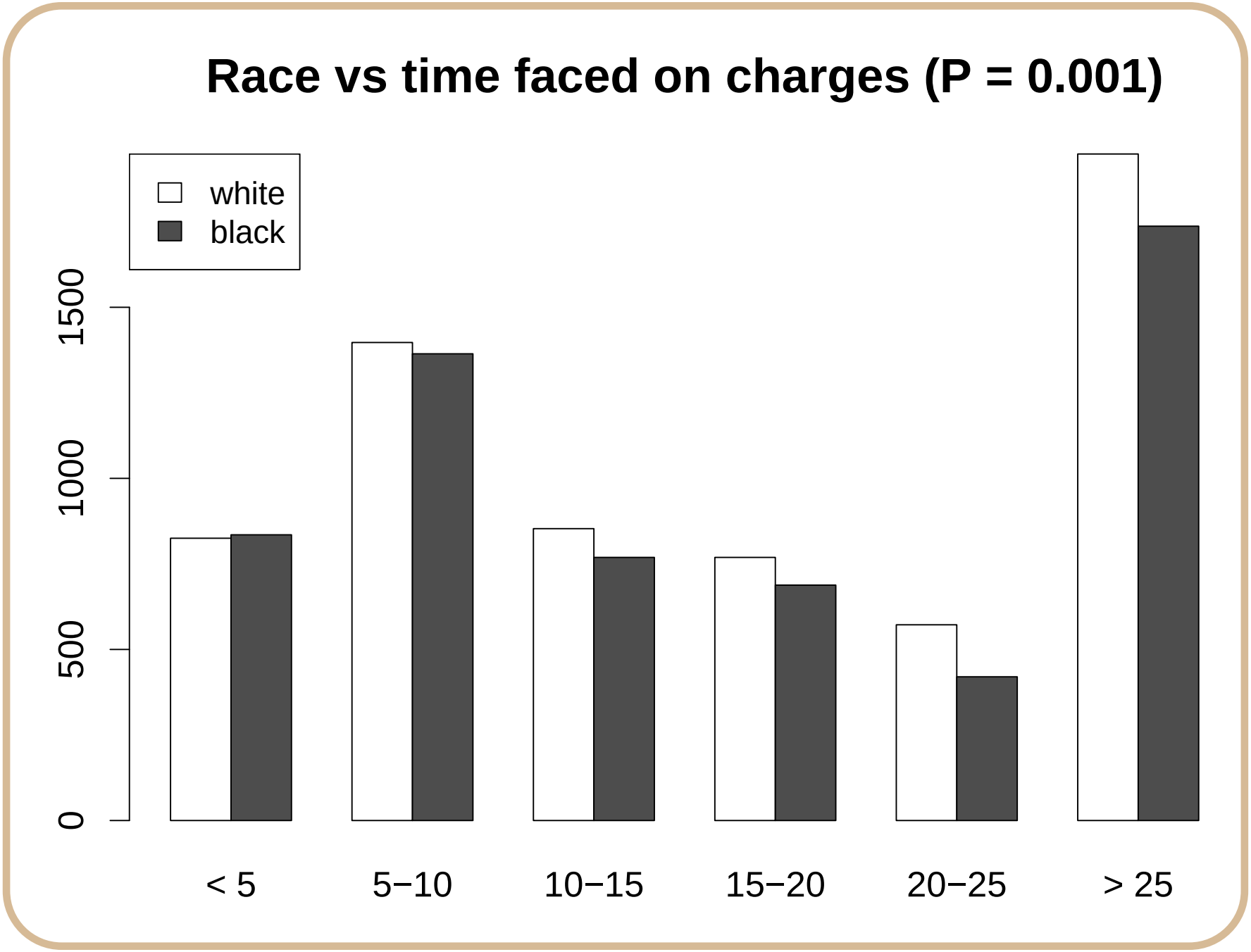
Factor-pair analysis

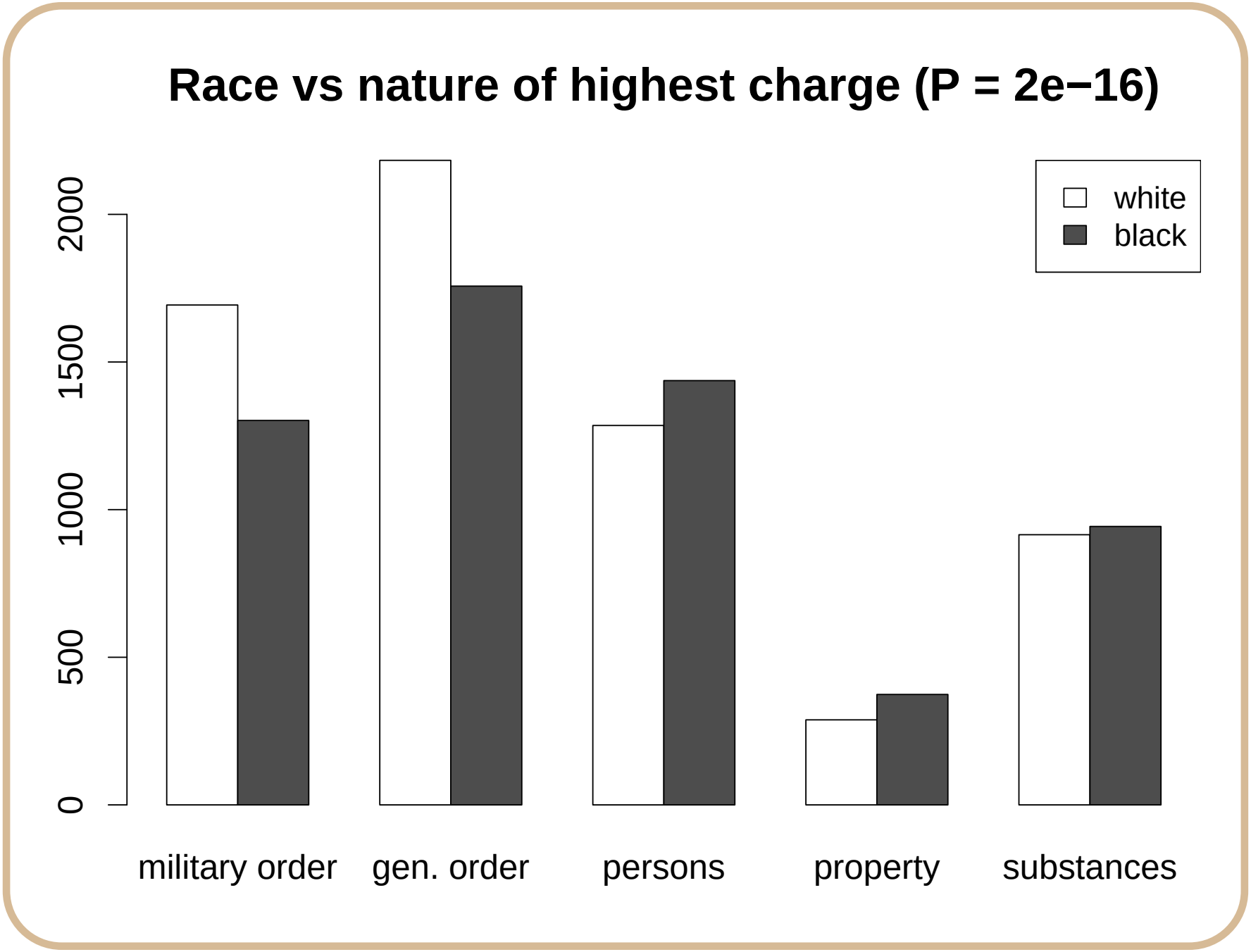
Multiyear assessment: Pairings of RACE with each PROCESS variable for the multiyear period FY 87-92 (two-way tables)

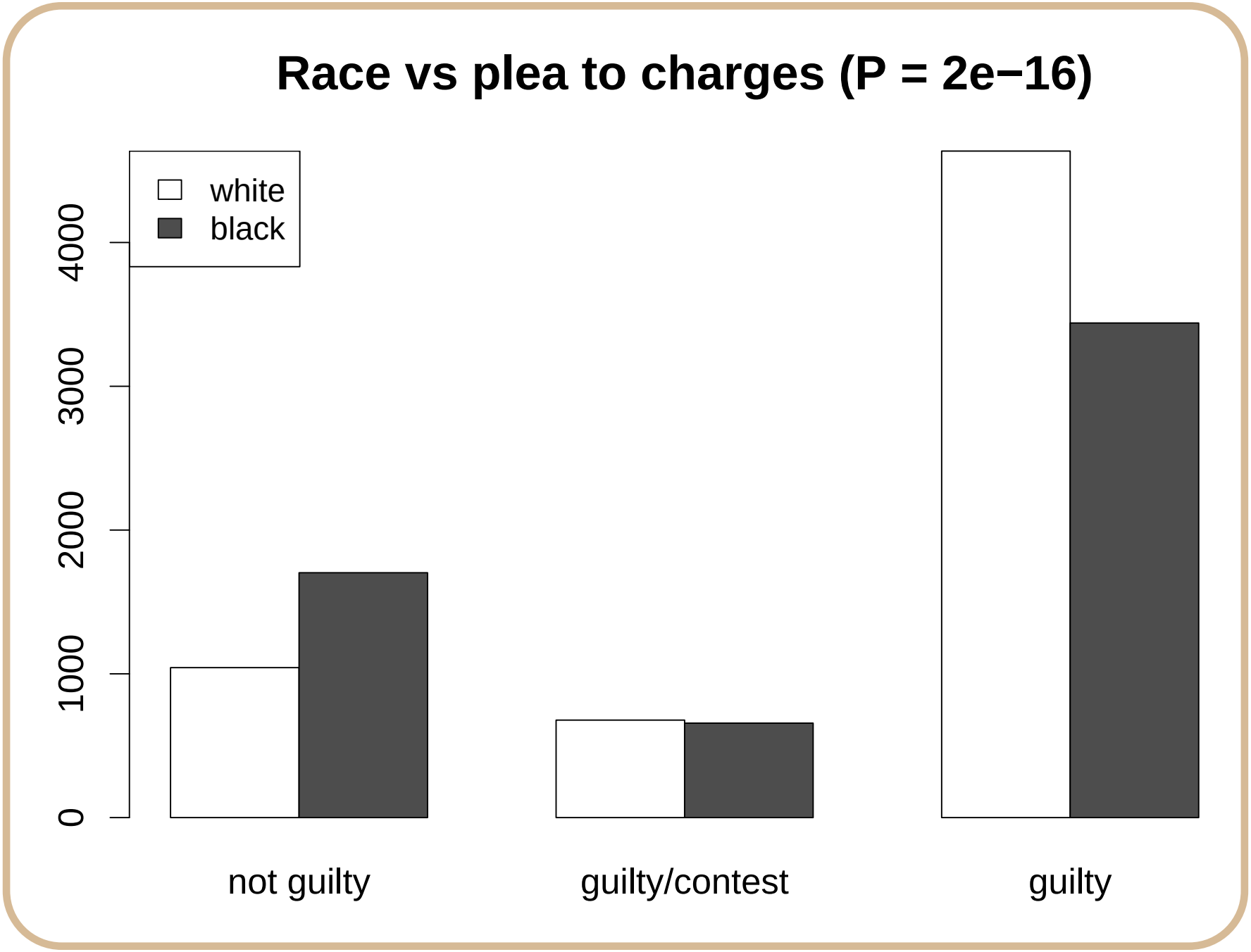
Multiyear assessment with controls: RACE paired with PROCESS variables, using each SOLDIER variable as control (three-way tables)

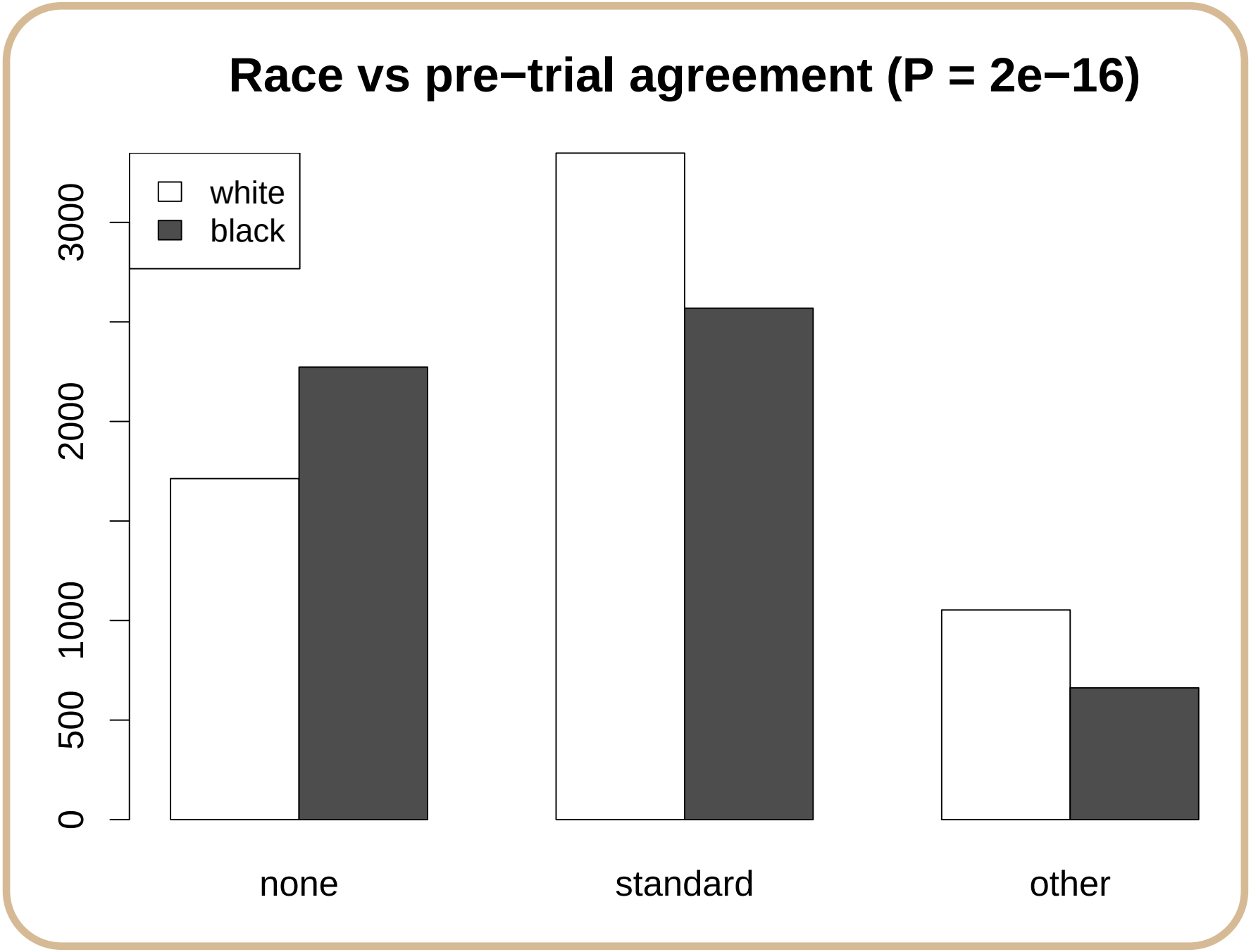
Year-by-year assessment: RACE paired with PROCESS variables, for each year in the period and with each SOLDIER variable as control (two and three-way tables)

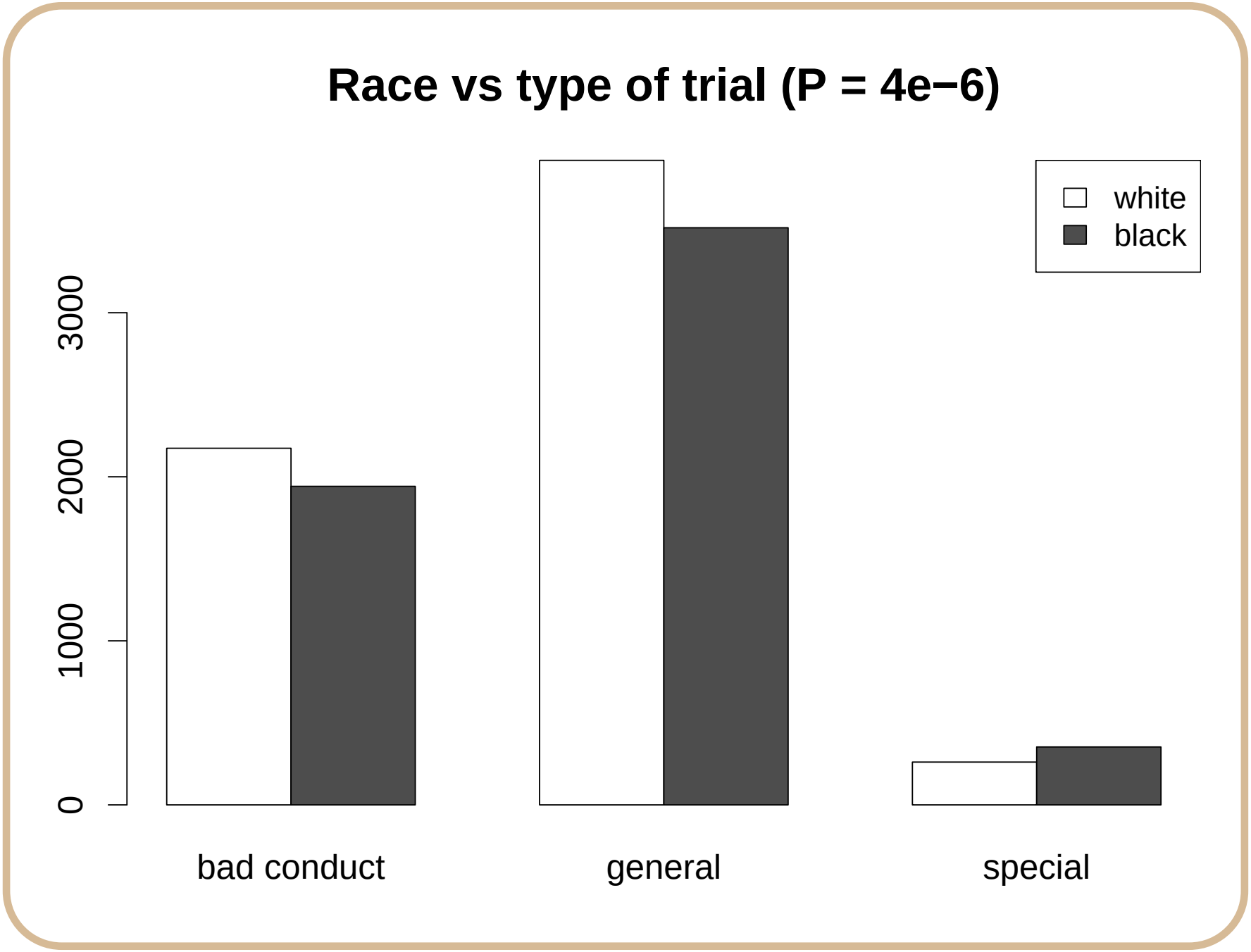

Race vs # trial charges (P = 0.282)
0
500
1000
1500
2000
2500
3000
white
black
1
2−4
5−7
> 7

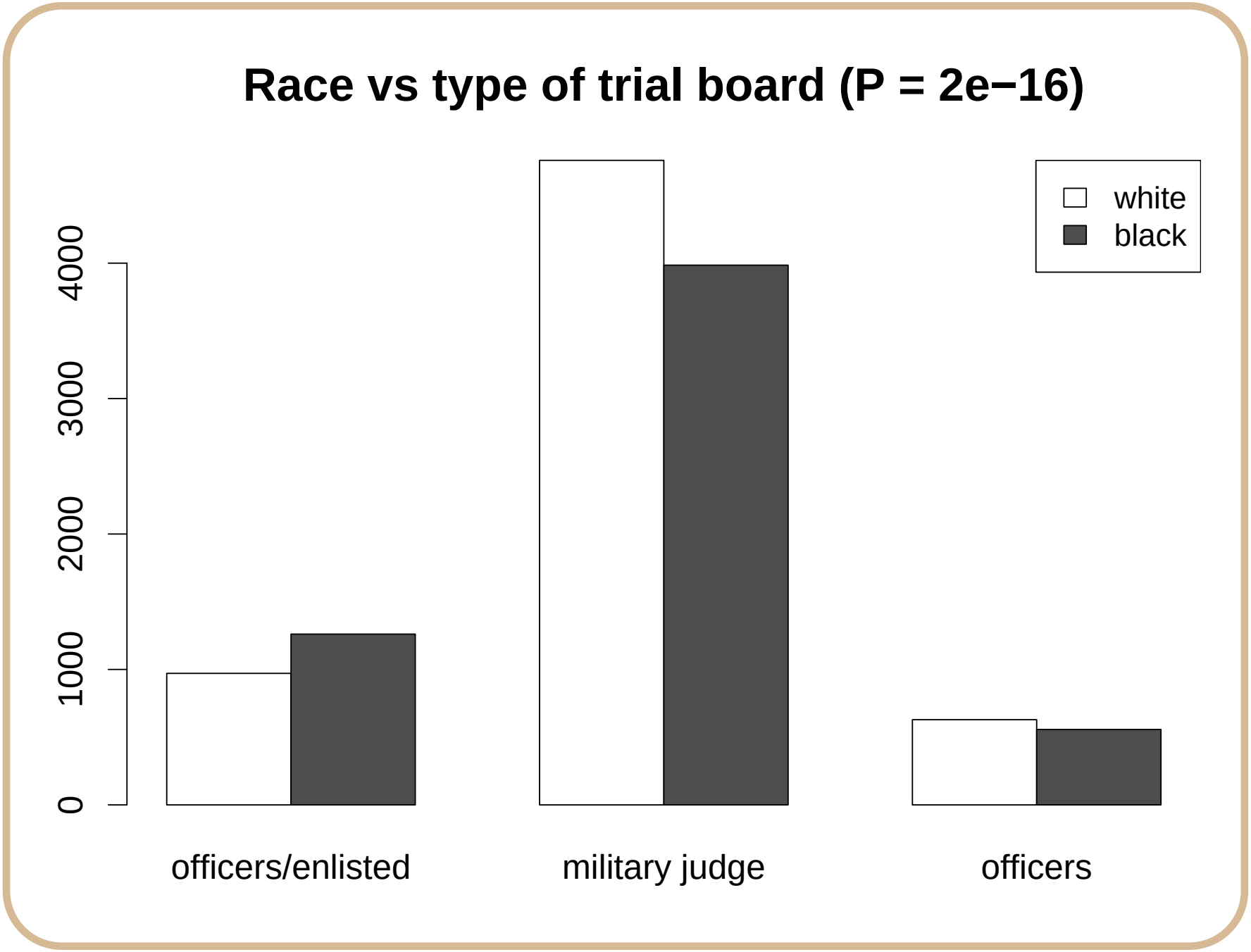

Race vs time faced on charges (P = 0.001)
white
black
0
500
1000
1500
< 5
5−10
10−15
15−20
20−25
> 25

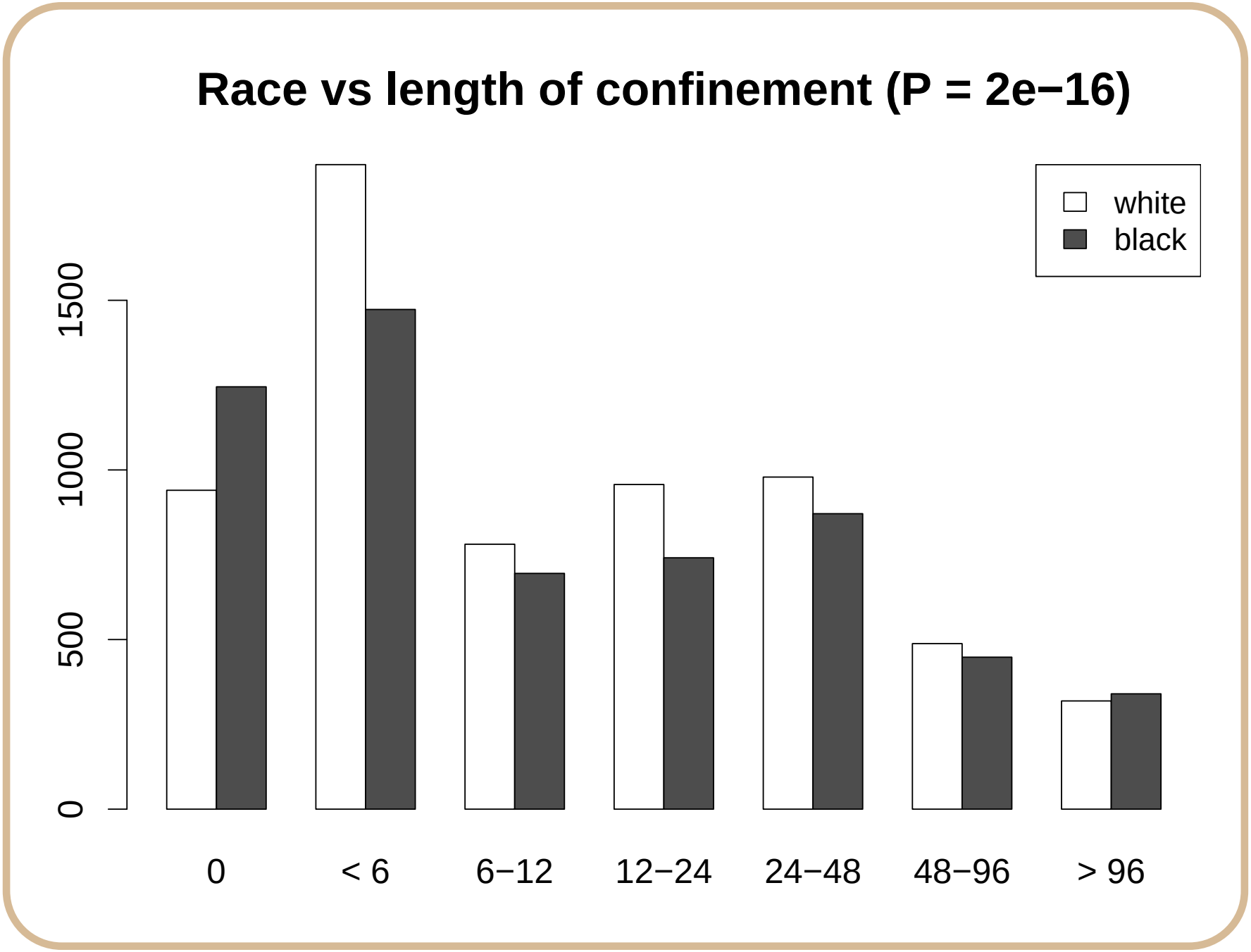

Race vs nature of highest charge (P = 2e−16)
white
black
0
500
1000
1500
2000
military order
gen. order
persons
property
substances

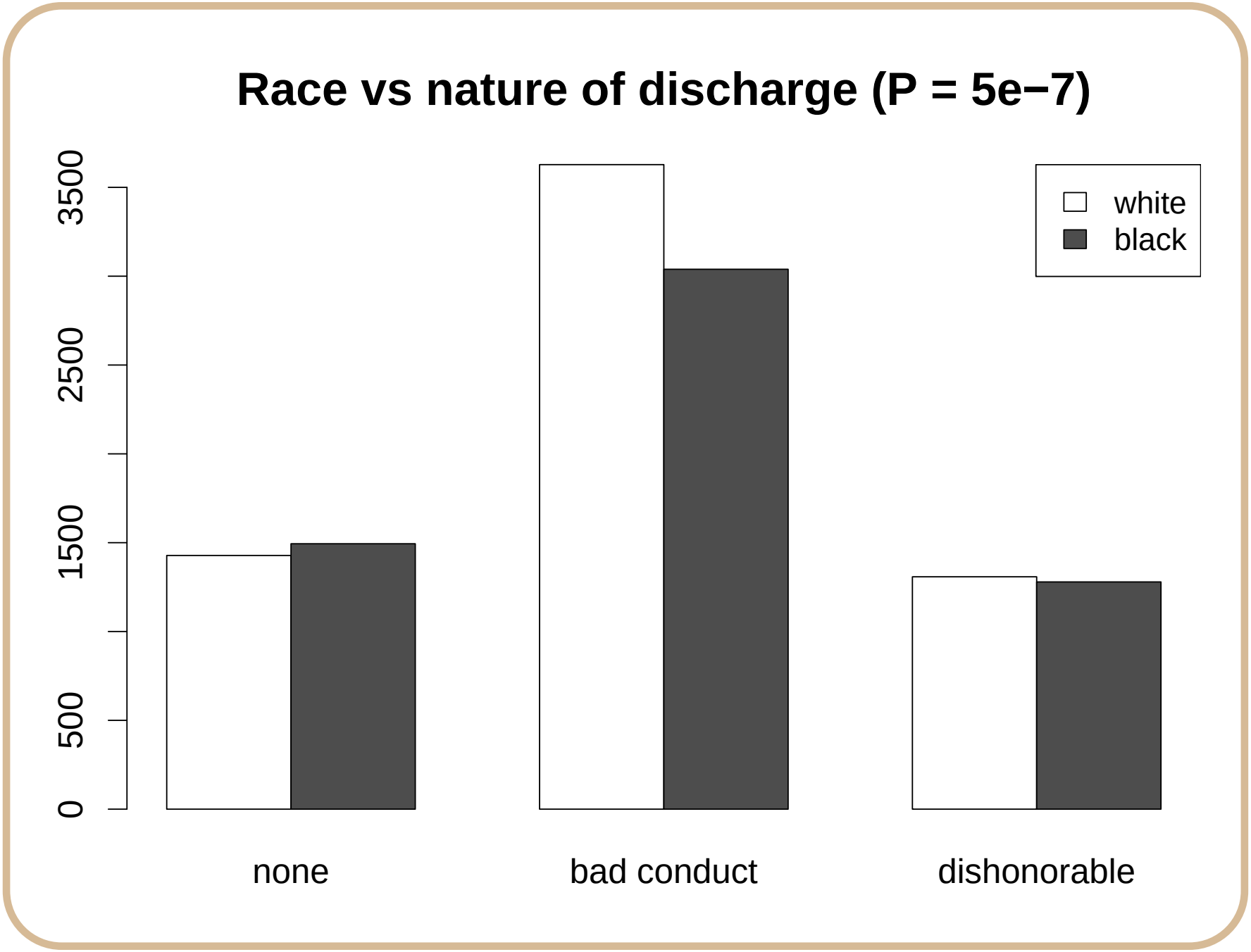

Race vs plea to charges (P = 2e−16)
white
black
0
1000
2000
3000
4000
not guilty
guilty/contest
guilty

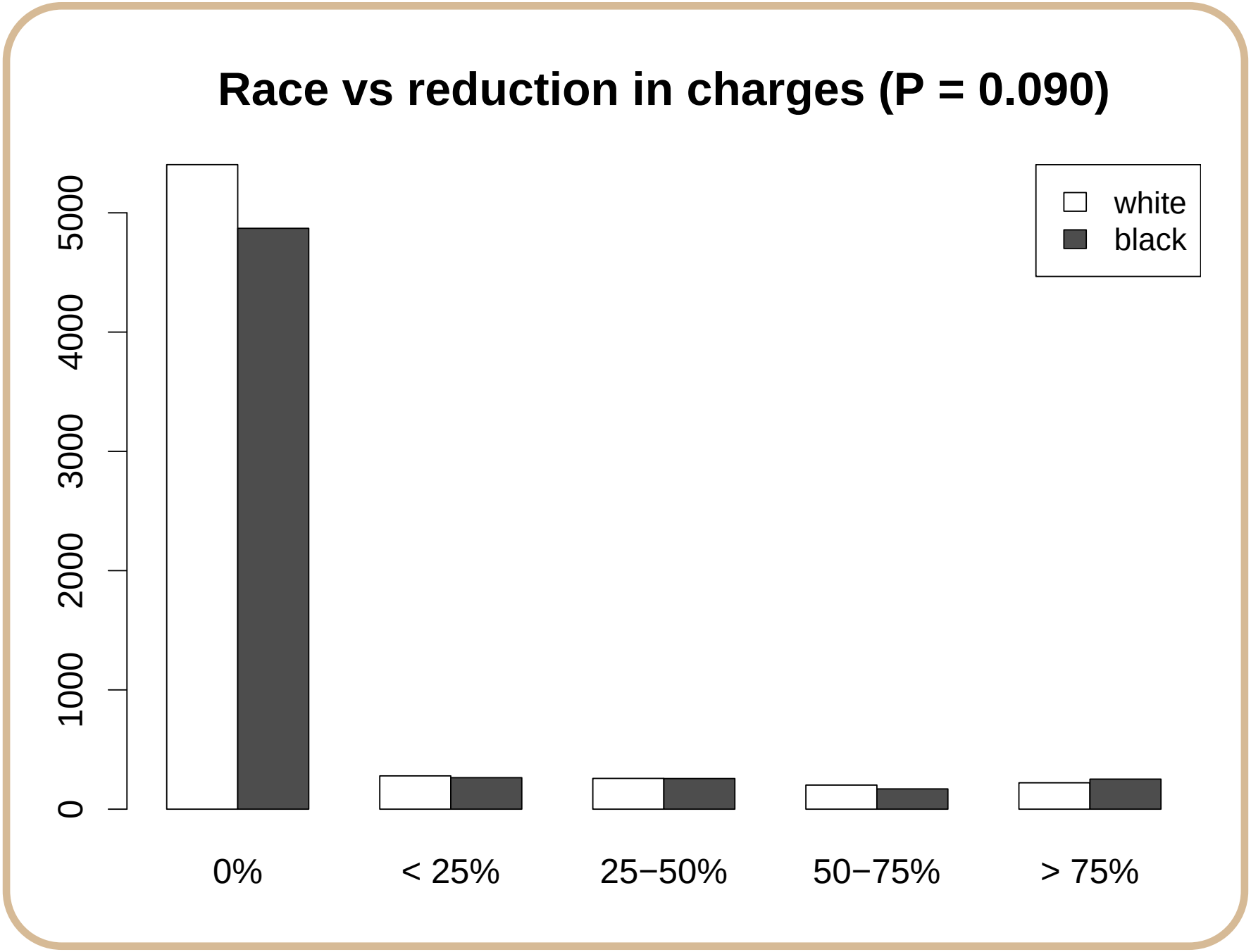

Race vs pre−trial agreement (P = 2e−16)
white
black
0
500
1000
2000
3000
none
standard
other

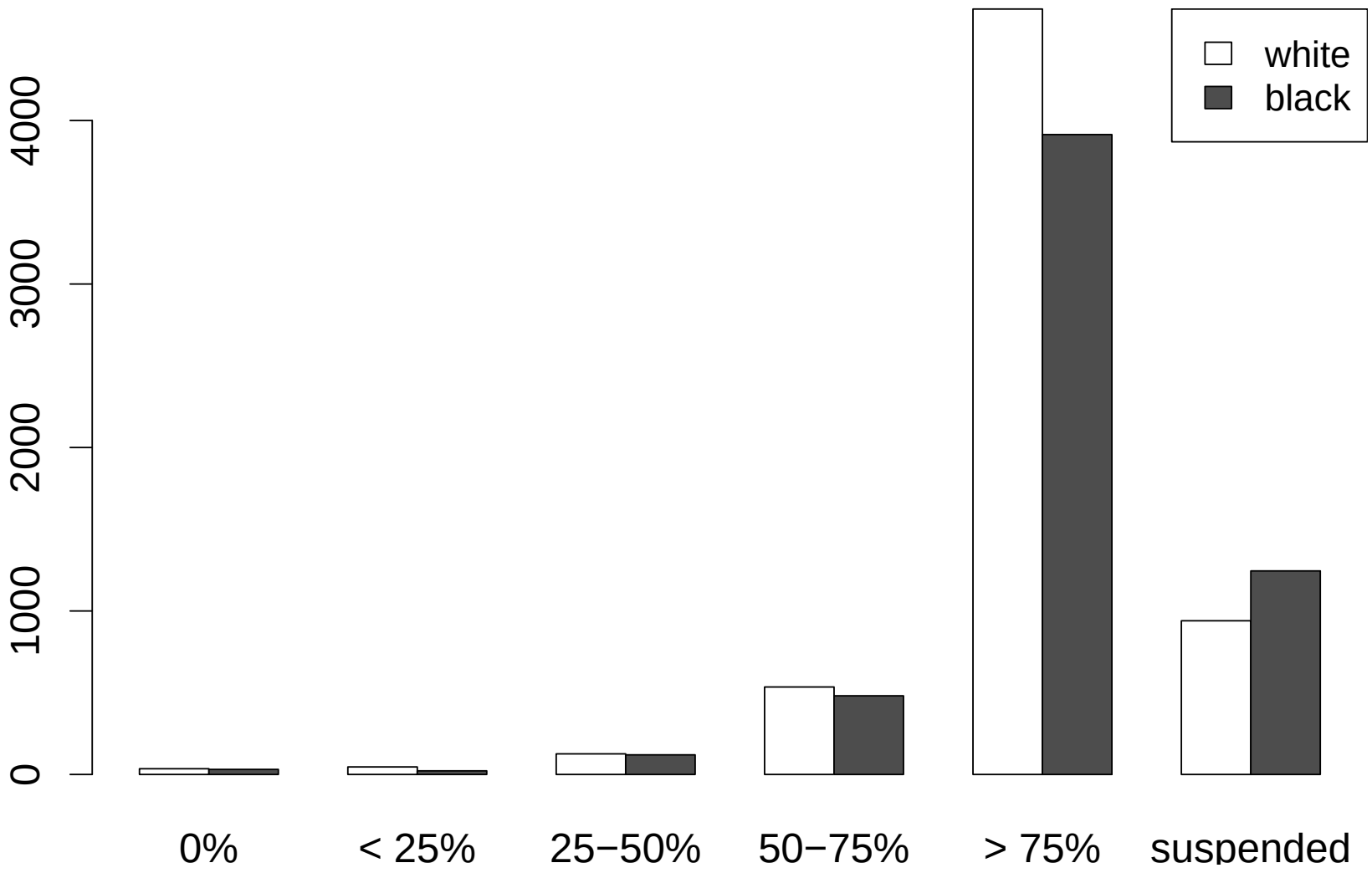

Race vs type of trial (P = 4e−6)
white
black
0
1000
2000
3000
bad conduct
general
special


Race vs type of trial board (P = 2e−16)
white
black
0
1000
2000
3000
4000
officers/enlisted
military judge
officers


Race vs length of confinement (P = 2e−16)
white
black
0
500
1000
1500
0
< 6
6−12
12−24
24−48
48−96
> 96


Race vs nature of discharge (P = 5e−7)
3500
2500
1500
500
0
white
black
none
bad conduct
dishonorable


Race vs reduction in charges (P = 0.090)
white
black
0
1000
2000
3000
4000
5000
0%
< 25%
25−50%
50−75%
> 75%


Race vs reduction of confinement (P = 2e−16)
white
black
0
1000
2000
3000
4000
0%
< 25%
25−50%
50−75%
> 75%
suspended

Results of factor-pair analysis

Multiyear assessment (base case). Largest difference observed for PLEA TO CHARGES — Black pleaded not guilty twice as often as White

Multiyear assessment with controls. When controlled for SOLDIER factors, difference patterns did not generally depart from trends in the base case. Largest difference was PRETRIAL AGREEMENT, when controlled for SCORE.

Year-by-year assessment. Trends similar to base case

Conclusion. Base case results taken to be representative for factor-pair assessment

Linear discriminant analysis

1. Nonnumeric variables were excluded
2. A linear discriminant model was fitted to each numeric variable for each year to determine significance (α = 0.001 and discriminatory power $>$ 0.5%)
3. A multiple linear discriminant model was fitted to the four statistically significant variables — AGE, SERVICE, EDUCATION, SCORE

Linear discriminant results

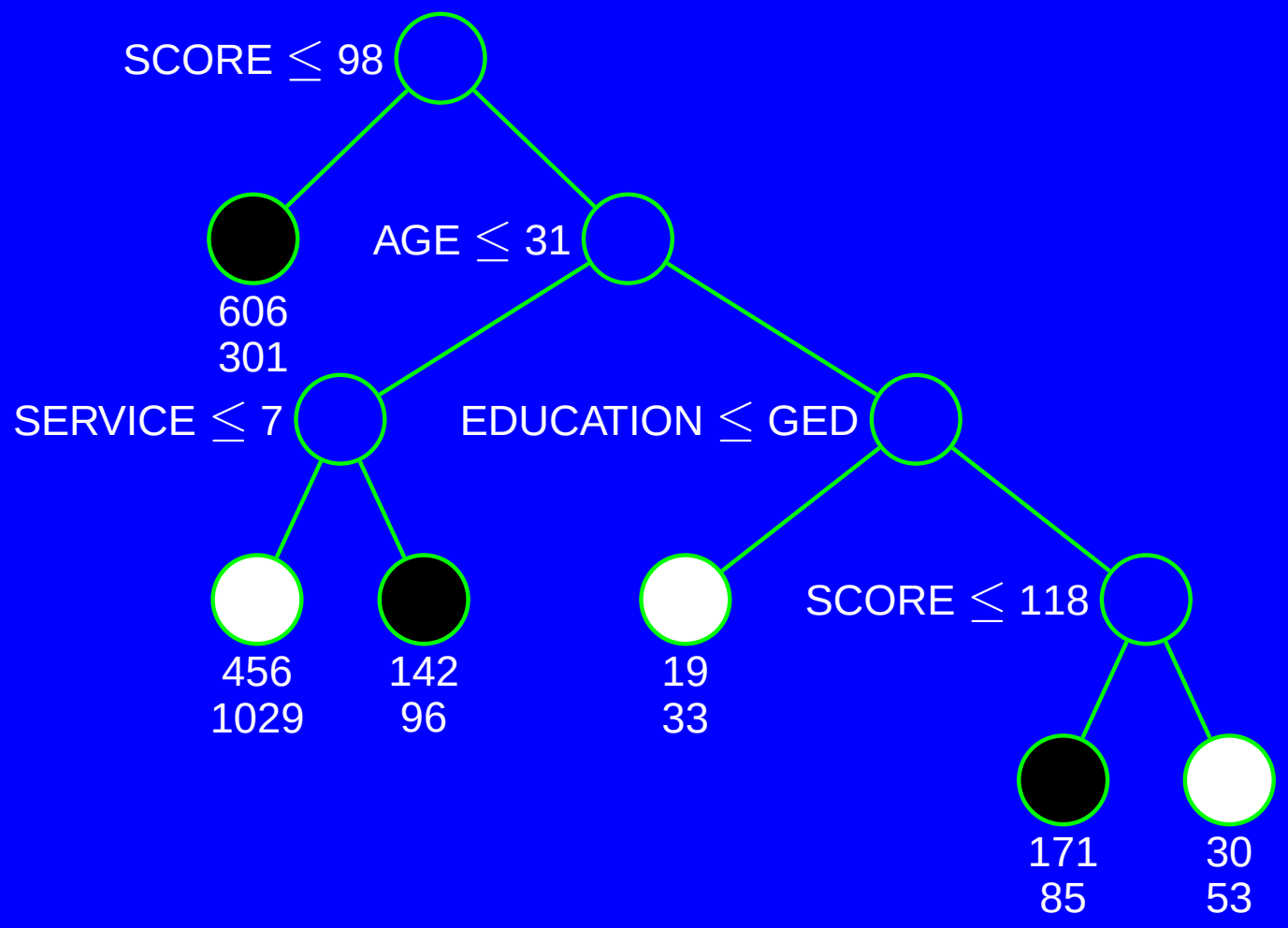
Variable	Coeff	Std. coeff	Rank
SCORE	.0558	.8808	1
AGE	-.0888	-.5271	2
EDUCATION	-.1528	-.2195	3
SERVICE	-.0197	-.0957	4

1. White offender has higher SCORE
2. Black offender is older
3. Black offender has more education
4. Black offender has longer service
5. Discriminatory power of model is 9%

Weaknesses of analyses

1. Factor-pairings by cross-tabs (even controlling for third variable) ignores multivariate nature of data
   - Simpson's paradox
2. Stepwise linear discriminant modeling also ignores multivariate nature of data
   - PROCESS variables eliminated in first stage
3. P-values are hard to interpret here (they are absent from the final report)
   - large number of tests
   - sample or population

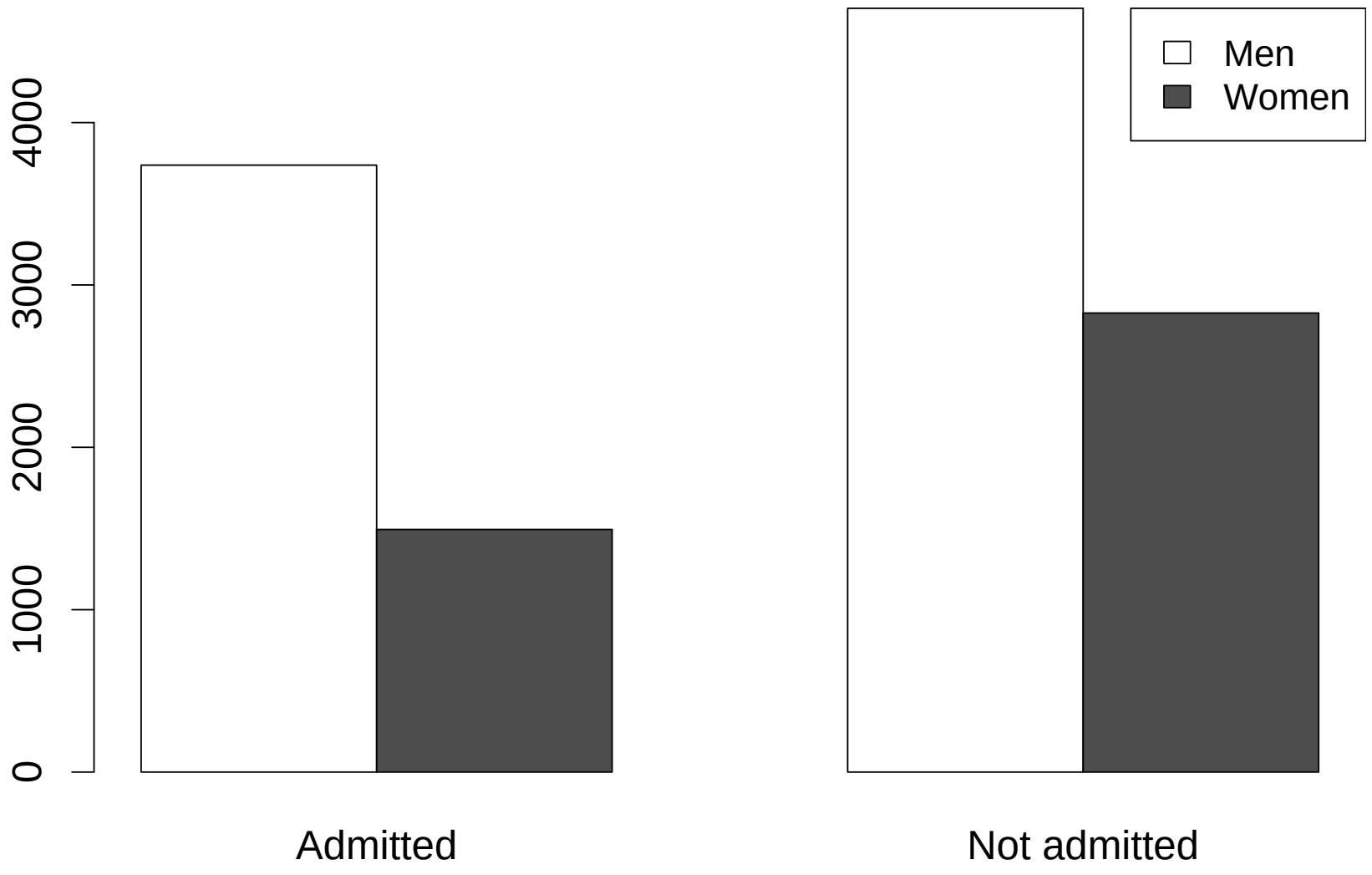
Classification tree to predict RACE



Comparison of tree and LDA models

1. Both use the same variables and rank them in the same order
2. Tree model has 65% prediction accuracy, LDA has 9% discriminatory power (nearest-neighbor and all-factor LDA have similar accuracy as tree model)
3. Tree model is easier to interpret
4. Conclusions from tree model:
   (a) SOLDIER variables more important than PROCESS variables for predicting RACE
   (b) After controlling for SOLDIER variables, PROCESS variables have no predictive power

Berkeley 1973 graduate admissions



Berkeley admissions by major department

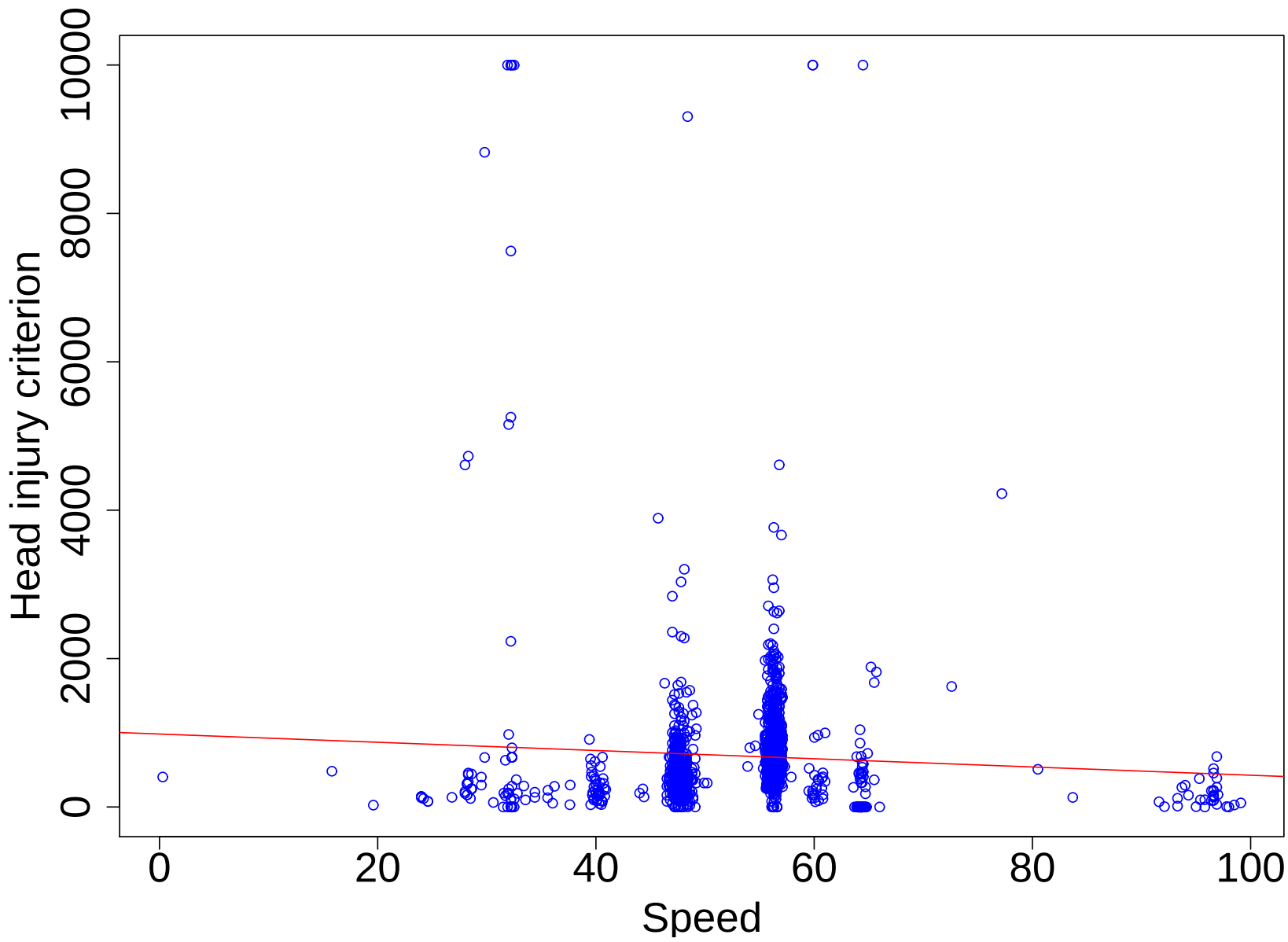
	Men		Women	
Major	#Applied	%Admitted	#Applied	%Admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Simpson's paradox in regression:

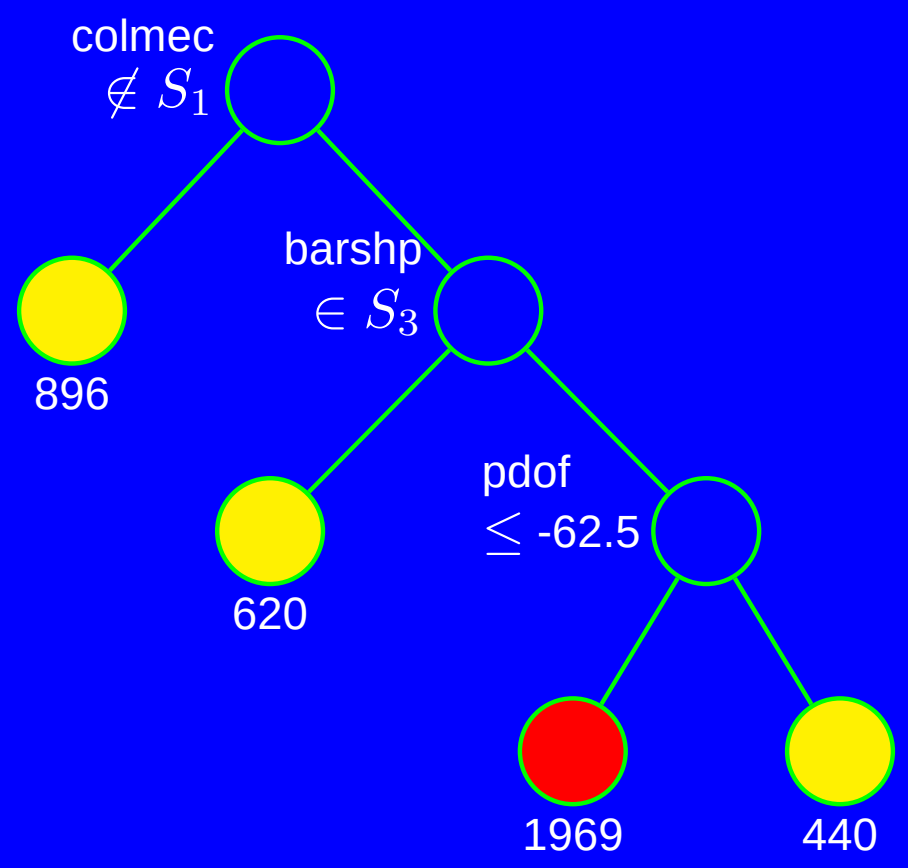
Vehicle crash test data

- National Highway Transportation Safety Administration (NHTSA) has been crash-testing vehicles since 1972
- 1,789 vehicles tested as of 2004
- Dependent variable is head injury criterion (HIC)
- $0 < \text{HIC} < 10{,}000$
- Threshold for severe head injury is HIC = 1000
- Twenty-five predictor variables give information on the vehicles, dummies, and test conditions

Name	Description	Name	Description
HIC	Head injury criterion	make	Car manufacturer (62)
year	Car model year	mkmodel	Car model (464)
body	Car body type (18)	transm	Transmission type (7)
engine	Engine type (15)	engdsp	Engine displacement (liters)
vehtwt	Vehicle weight (kg)	colmec	Collapse mechanism (11)
vehwid	Vehicle width (mm)	modind	Modification indicator (5)
vehspd	Vehicle speed (km/h)	crbang	Crabbed angle
tksurf	Track surface (5)	pdof	Principal direction of force
tkcond	Track condition (6)	impang	Impact angle
occtyp	Occupant type (10)	dumsiz	Dummy size (6)
seposn	Seat position (5)	barrig	Barrier rigidity (2)
barshp	Barrier shape (14)	belts	Seat belt type (3)
airbag	Airbag present (2)	knee	Knee restraint present (2)


Head injury criterion
0
2000
4000
6000
8000
10000
0
20
40
60
80
100
Speed

Regression tree for predicting HIC

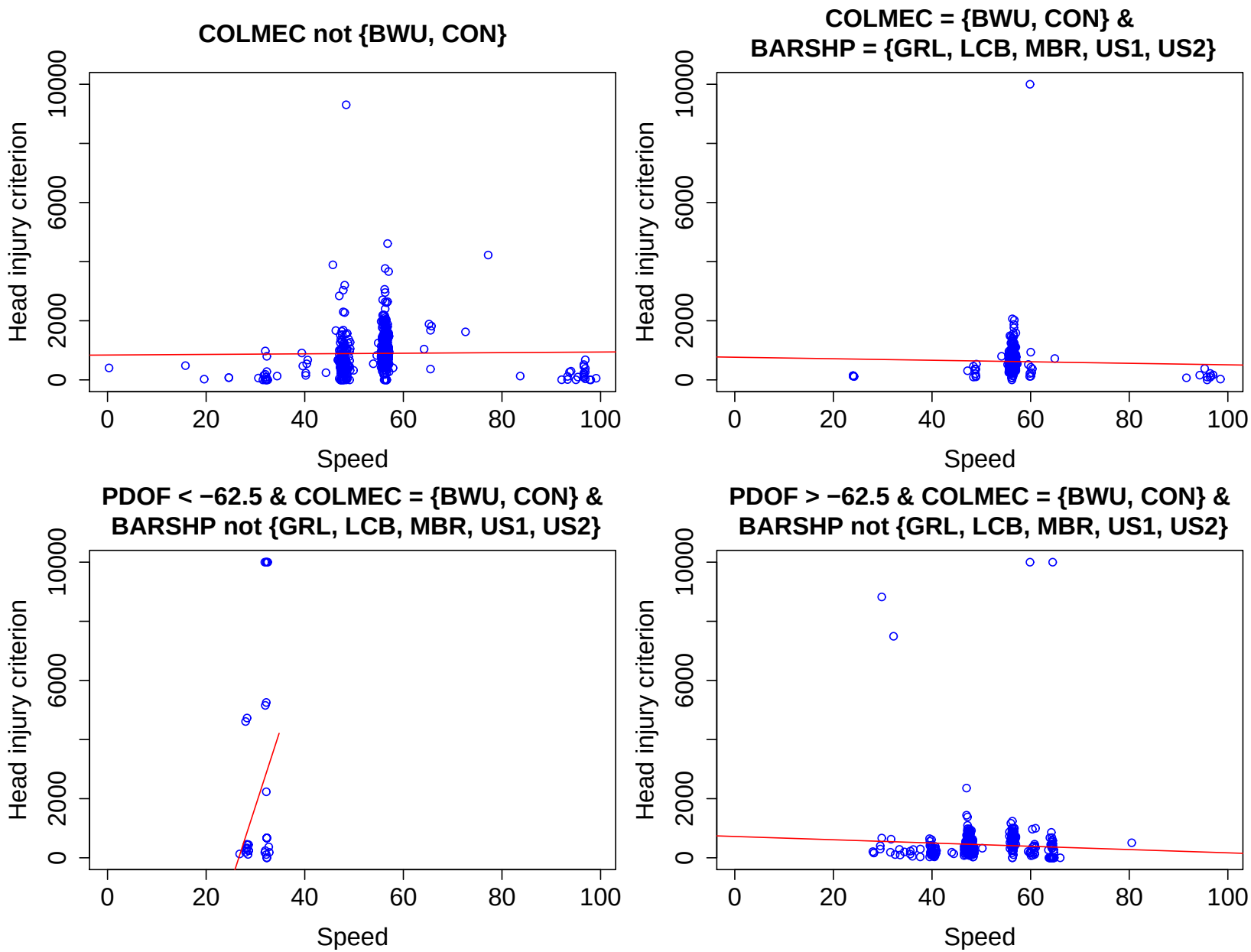


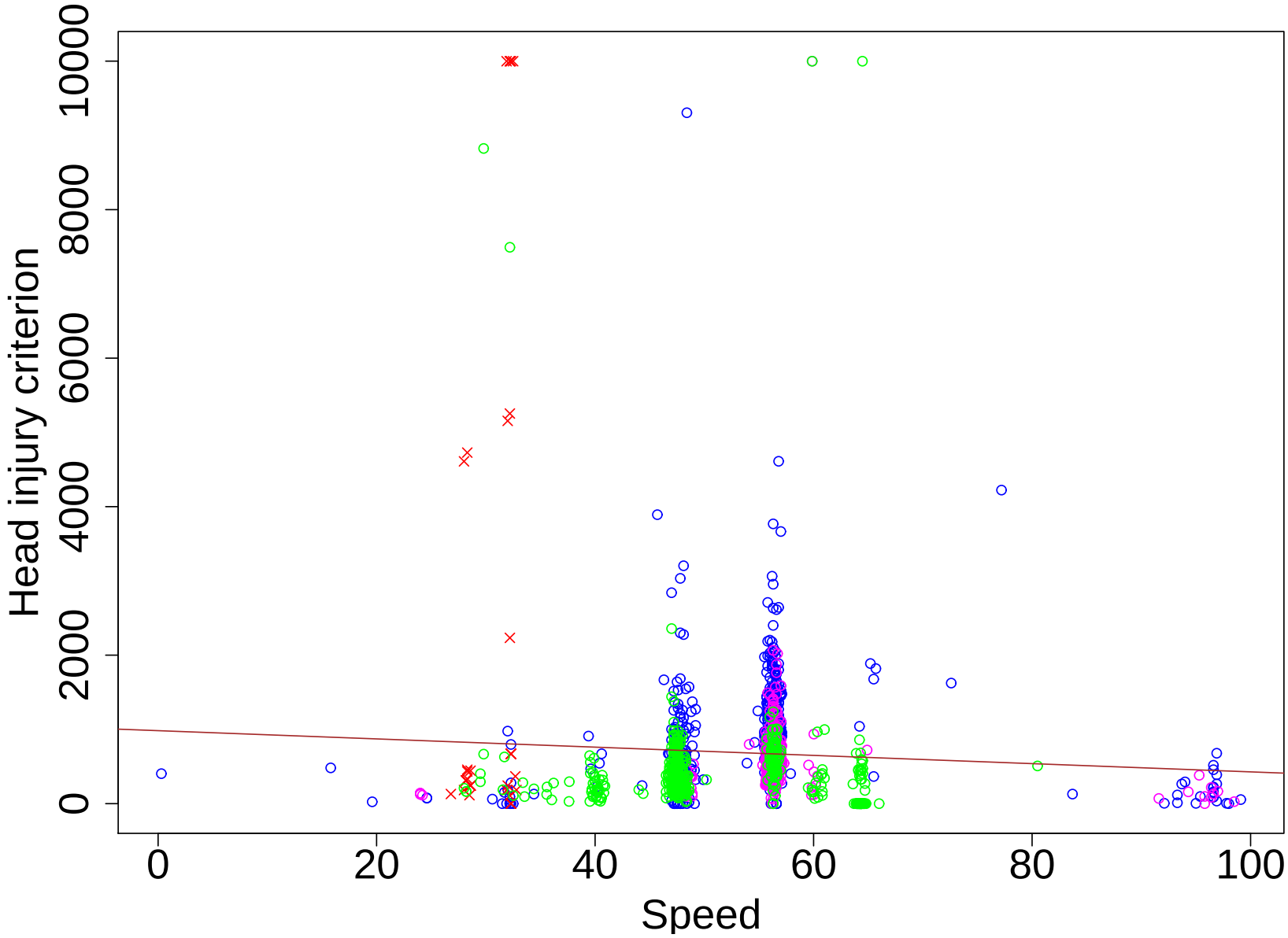
S_1 = {Behind wheel unit, Convoluted tube}

S_3 = {Guard rail, Load cell barrier, Median barrier, US1, US2}

Number beneath node is sample mean HIC

Red node indicates statistically significant positive slope for Speed


COLMEC not {BWU, CON}
COLMEC = {BWU, CON} &
BARSHP = {GRL, LCB, MBR, US1, US2}
PDOF < −62.5 & COLMEC = {BWU, CON} &
BARSHP not {GRL, LCB, MBR, US1, US2}
PDOF > −62.5 & COLMEC = {BWU, CON} &
BARSHP not {GRL, LCB, MBR, US1, US2}
Head injury criterion
Speed
0
2000
6000
10000
0
20
40
60
80
100


Head injury criterion
0
2000
4000
6000
8000
10000
0
20
40
60
80
100
Speed

Tree algorithm in a nutshell

1. Recursively partition the data and sample space
2. Fit a model to each partition
3. Prune tree to generate a nested sequence of submodels
4. Select a submodel using an independent test sample or by cross-validation

GUIDE software and documentation

`www.stat.wisc.edu/~loh/guide.html`

Acknowledgment

Research and development of GUIDE is supported by Army Research Office