Video Games and the Risk of Reckless Driving

An Application to Car Insurance

Kyle Dotterrer, Sam Marciano, Andrew Shipman, Lucy (Ao) Zhang

Context and Previous Studies

“36% of American adolescents game regularly with 80% of the players being boys” (Cummings & Vandewater 2007)

- Exceeds time spent watching films and TV (Fischer et. al 2007)

Previous research has a found a significant, positive relationship between video game playing and risky driving behavior

- Beullens, Roe & Van den Bulck (2011)
- Fischer et. al (2007)
- Hull, Draghici, & Sargent (2012)

Basic Stats: Data Organization

The data consists of 1389 observations, where each observation responds to 13 different fields.

Besides the “age” field, each field has one of the following responses:

“yes/always”

unsafe	videog	male	race	age	ed	sens	rebel	resp	dem	sprt	school	church
0	1	1	1	17.08333	0	0	0	-1	-1	1	0	0
0	0	0	1	16.41667	-1	0	-1	1	1	1	1	0
0	1	1	1	14.91667	-1	-1	0	0	0	1	0	0
0	1	0	1	15.91667	1	0	1	1	0	1	-1	0

Basic Stats: Field Breakdown & Proportions

	unsafe		videog		male		race		sprt		church	
yes/always	363	26.13%	666	47.95%	1002	72.14%	1154	83.08%	807	58.10%	721	51.91%
no/never	1026	73.87%	723	52.05%	387	27.86%	235	16.92%	582	41.90%	668	48.09%

	ed		sens		rebel		resp		dem		school	
yes/always	525	37.80%	674	48.52%	374	26.93%	382	27.50%	355	25.56%	363	26.13%
maybe/sometimes	446	32.11%	386	27.79%	551	39.67%	498	35.85%	592	42.62%	623	44.85%
no/never	418	30.09%	329	23.69%	464	33.41%	509	36.65%	442	31.82%	403	29.01%

Legend

Ed = parents' education (low/average/high)
Sens = sensation seeker
Rebel = rebel in nature
Resp = responsible
Dem = demanding parents
School = academic performance (poor/average/high)

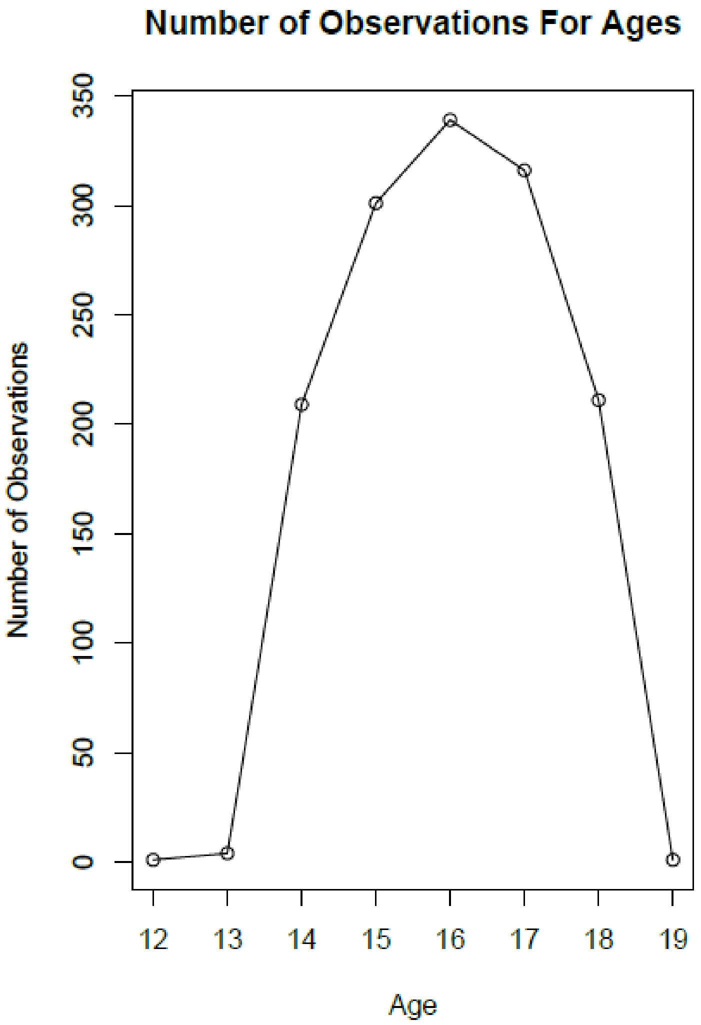
Unsafe = unsafe driving
Videog = video gaming or not
Male = bpy/girl
Race = white/non-white
Sprt = play sport or not
Church = go to church or not

Basic Stats: Age

Mean age: 16.472

Minimum age: 12 yrs, 10 mths

Maximum age: 19 yrs, 6 mths



Logistic Regression

How do we perform regression analysis when response variable is binary?

- Outcome variable follows bernoulli distribution

OLS with binary response = linear probability model

- Heteroskedasticity
- Errors not normally distributed
- Linearity (lack of constraints on outcome variable)

The solution? Logistic regression

$$\text{logit}(p) = \ln[\, p/(1-p) \,] = a + bx1$$

Output of logistic regression is *probability* of success outcome

Log odds of success = linear function of predictors

Link function = logit transformation (why not probit?)

Logistic Regression

The outputs of our regression (our coefficients) are log odds

Estimated regression equation (solve for *p*)

$$p = e^{(a + bx)} / [1 + e^{(a + bx)}]$$

estimating probability of a getting a 1 ('success'), or *p*, for any linear combination of values of the predictor variables

Domain -> $(-\infty, \infty)$ Range -> $[0, 1]$

Logistic Regression: Model & Results

glm(formula = unsafe ~ videog + age, family = "binomial", data = df)

	Estimate	Standard Error	Z Value	P(> |Z|)
(Intercept)	-10.98438	0.93244	-11.78	< 2e-16 ***
videog1	0.58732	0.13273	4.425	9.65e-06 ***
age	0.5765	0.05537	10.412	< 2e-16 ***

- Unit increase in age -> log odds of unsafe driving behavior increase by 0.58
- Playing video games -> log odds of unsafe driving behavior increase by 0.59

Predicted Probabilities

- Hold age constant at mean
- Compute predicted probability of unsafe driving behavior

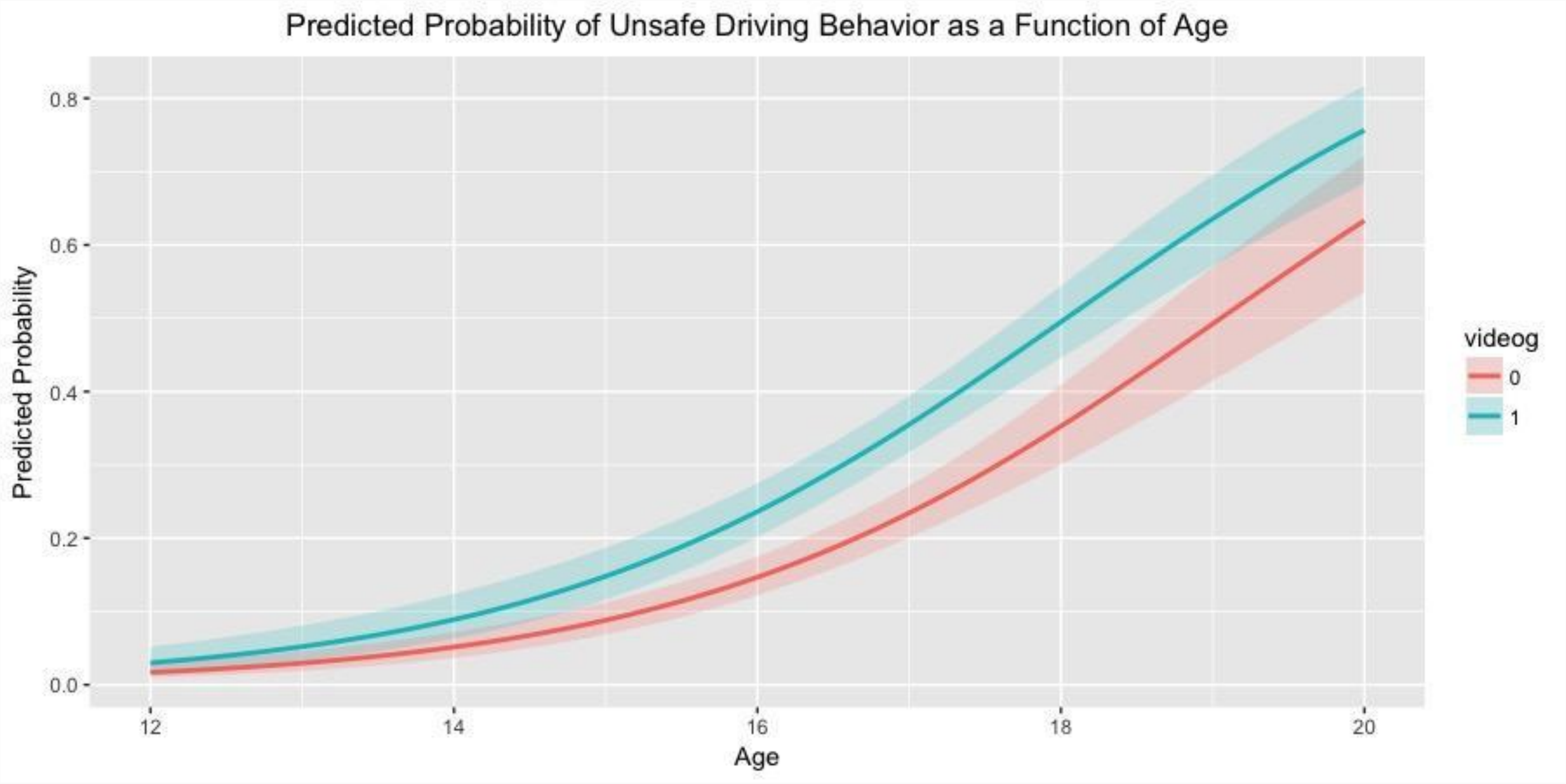
	Odds Ratio	2.50%	97.50%
Intercept	1.70E-05	2.64E-06	0.0001
videog	1.80E+00	1.39E+00	2.3363
age	1.78E+00	1.39E+00	1.9871

age	videog	Predicted Probability
16.47222	0	0.184
16.47222	1	0.289

Predicted probability of unsafe driving increases from 0.184 to 0.289 in response to video game playing

Odds Ratios & Confidence Intervals

- Exponentiate regression coefficients
- Able to interpret as odds ratios
- Exponentiate confidence



Predicted Probability of Unsafe Driving Behavior as a Function of Age
Predicted Probability
0.8
0.6
0.4
0.2
0.0
12
14
16
18
20
Age
videog
0
1

Logistic Regression: Model Fit

No direct analog to linear regression R^2 in logistic regression

Use *deviance* instead of sum of squares

D = -2ln(likelihood of fitted model / likelihood of saturated model)

Two important deviances measures: null and model

Null = intercept only, no predictors

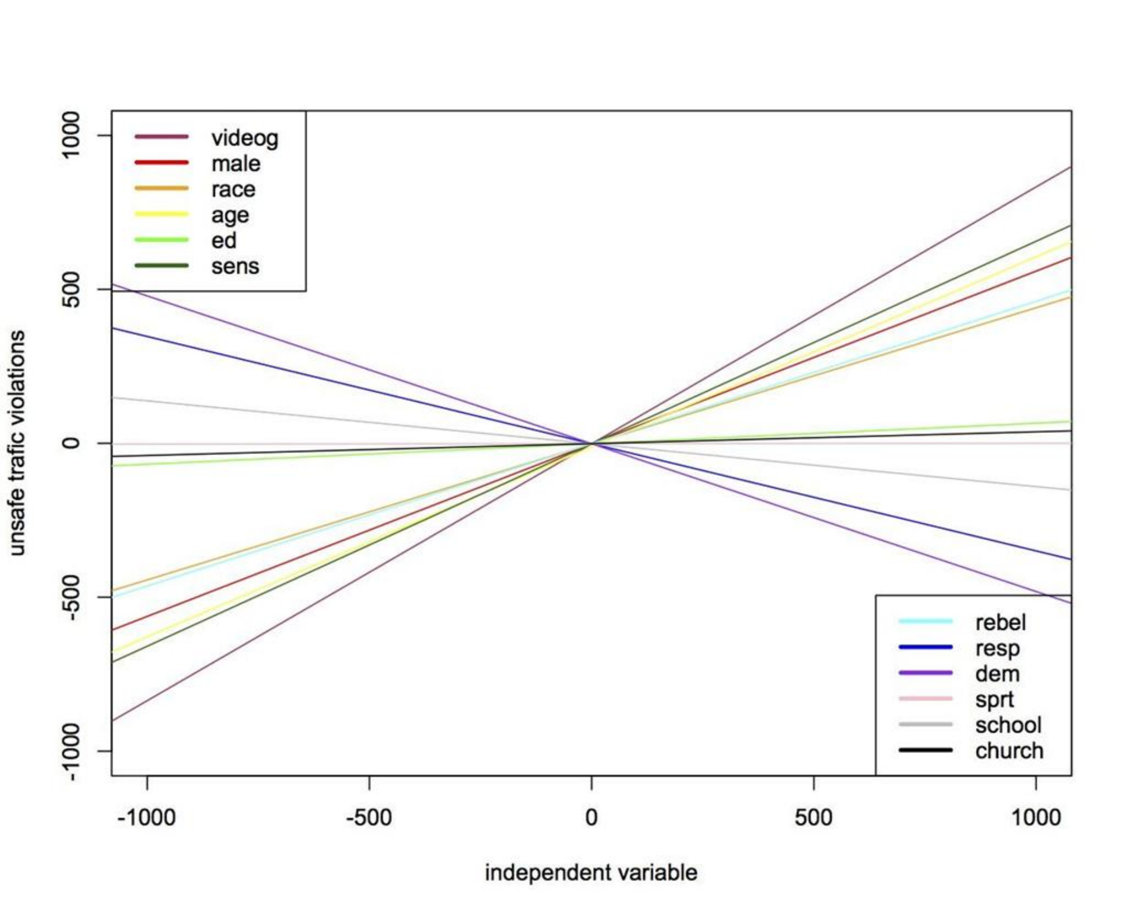
Model = at least one predictor

Linear Regression: Model Fit

R Code		Output
with(model, null.deviance - deviance)	→	168.6126
with(model, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))	→	2.433522e-37

Takeway*:* the addition of predictor variables *significantly* improved model fit!

Expanded Individual Regressions



Correlation Matrix

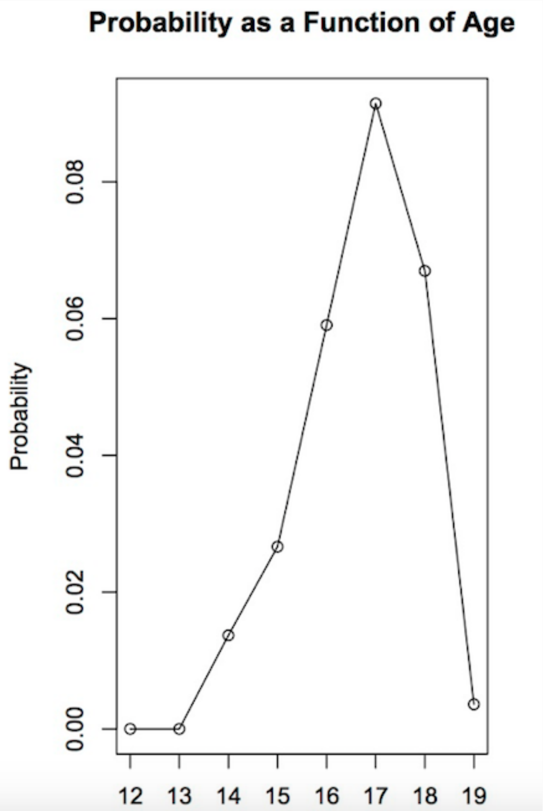
Legend

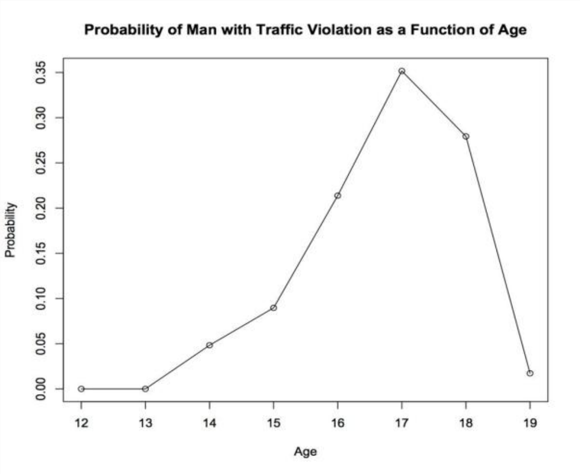
unsf = unsafe driving
vg = video gaming
male = boy/girl
race = white/nonwhite
age = years
ed = parents’ education
sns = sensation seeker
rbl = rebel in nature
resp = responsible
dem = demanding parents or not
sprt = play sports or not
sch = academic performance
ch = go to church or not

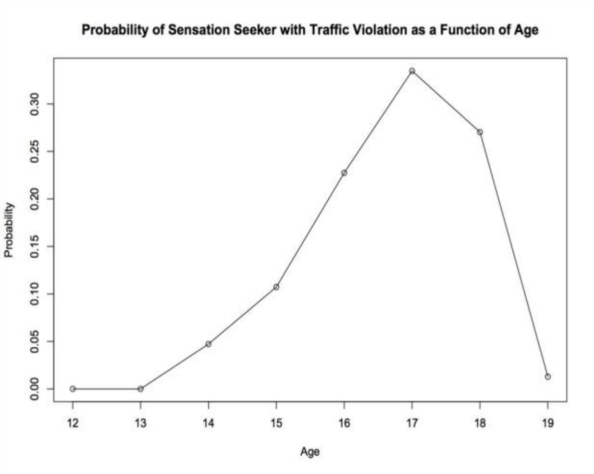
	unsf	vg	male	race	age	ed	sns	rbl	resp	dem	sprt	sch	ch
ch	0.01	-0.19	-0.02	-0.03	-0.1	0.01	-0.13	-0.09	0.14	0.14	0.05	0.1	
sch	-0.05	-0.2	-0.09	0.13	-0.05	0.18	-0.24	-0.29	0.21	0.2	0.12		0.1
sprt	0	-0.01	0.09	0.03	-0.06	0.07	-0.04	-0.01	0.04	0.08		0.12	0.05
dem	-0.15	-0.24	-0.09	0.02	-0.22	-0.02	-0.27	-0.3	0.48		0.08	0.2	0.14
resp	-0.12	-0.22	-0.06	0.05	-0.16	0.01	-0.27	-0.32		0.48	0.04	0.21	0.14
rbl	0.16	0.26	0.14	-0.07	0.07	-0.07	0.43		-0.32	-0.3	-0.01	-0.29	-0.09
sns	0.21	0.3	0.03	0.03	0.14	-0.05		0.43	-0.27	-0.27	-0.04	-0.24	-0.13
ed	0.02	-0.04	0.01	0.2	0.09		-0.05	-0.07	0.01	-0.02	0.07	0.18	0.01
age	0.32	0.21	0.07	0.08		0.09	0.14	0.07	-0.16	-0.22	-0.06	-0.05	-0.1
race	0.07	-0.03	0.04		0.08	0.2	0.03	-0.07	0.05	0.02	0.03	0.13	-0.03
male	0.1	0.22		0.04	0.07	0.01	0.03	0.14	-0.06	-0.09	0.09	-0.09	-0.02
vg	0.18		0.22	-0.03	0.21	-0.04	0.3	0.26	-0.22	-0.24	-0.01	-0.2	-0.19
unsf		0.18	0.1	0.07	0.32	0.02	0.21	0.16	-0.12	-0.15	0	-0.05	0.01

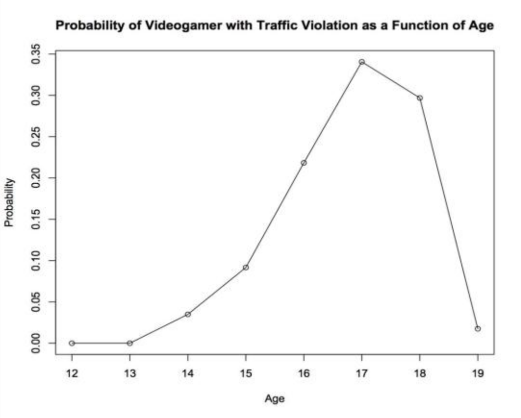
Multicollinearity assumption

Probabilities









Regression With Additional Controls

Checked the sensitivity of our previous regression by including controls for age, gender and sensation seeking

	Estimate	Standard Error	Z Value	P(>Z)
(Intercept)	-11.119	0.953	-11.671	< 2e-16 ***
videog	0.296	0.142	2.082	0.0374 *
gender	0.388	0.16	2.428	0.0152 *
age	0.565	0.0563	10.043	< 2e-16 ***
sensation	0.537	0.0936	5.739	9.54e-09 ***

Conclusions and Implications

Limitations

- Selection bias; response bias (self-reporting)
- Non time-series data

Risk Policy

Male, sensation-seeking video gamers will be charged the highest premium. Video gamers, independent of other variables, will not be charged as high of a premium as other independent fields (e.g. age and sensation-seekers).

Conclusion

Video gaming important factor, even once controlled for additional variables, but not the only significant indicator for driving violations

References

http://stats.idre.ucla.edu/r/dae/logit-regression/

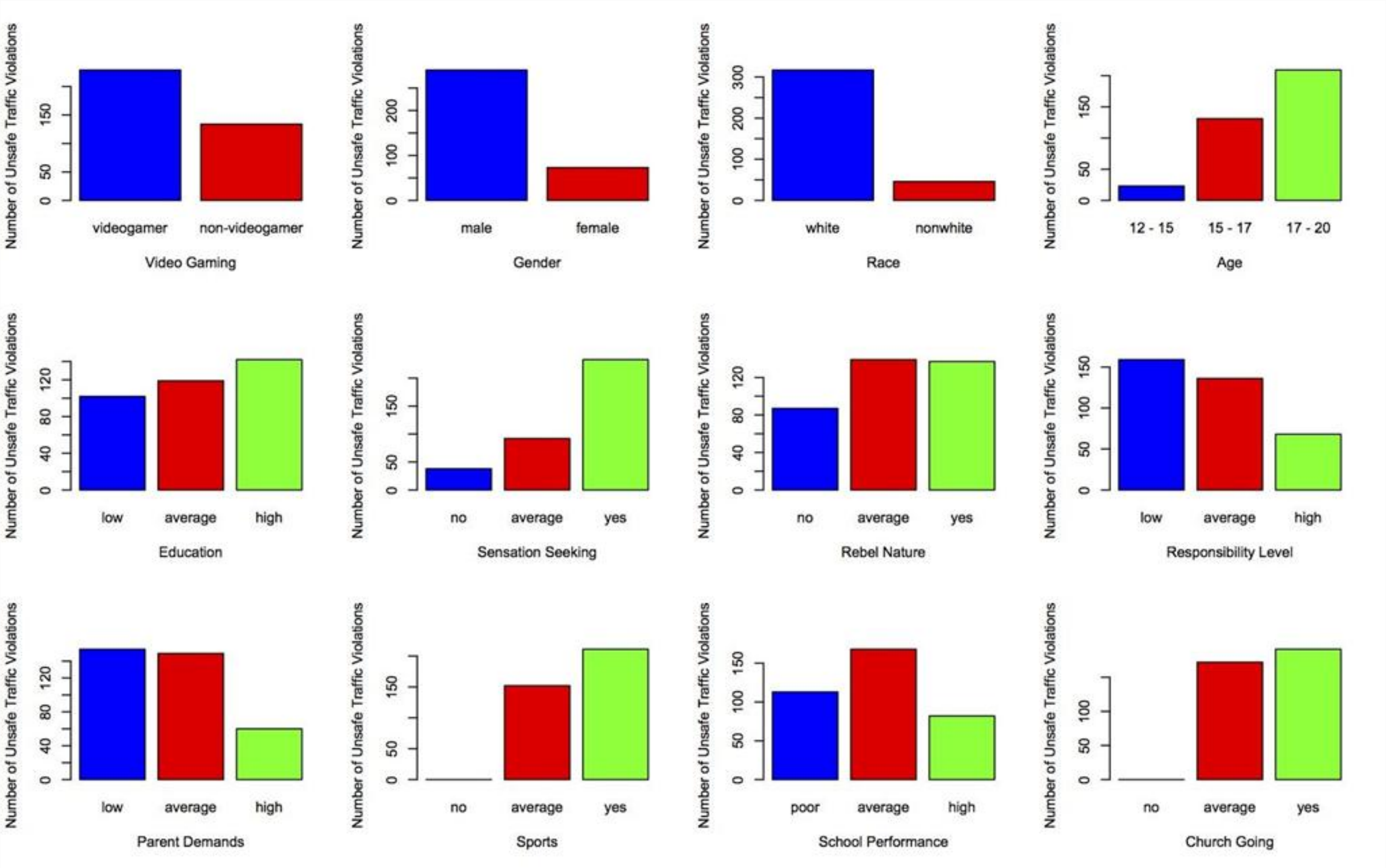
http://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

http://www.apa.org/pubs/journals/releases/ppm-1-4-244.pdf

http://ac.els-cdn.com/S0001457510001995/1-s2.0-S0001457510001995-main.pdf?_tid=279d7620-00ed-11e7-abfd-00000aab0f26&acdnat=1488640603_3ebdca5615003b0a45bb76b1395eed7f

http://www.apa.org/pubs/journals/releases/xap-13122.pdf

Appendix: Conditional Basic Stats Graphically



Appendix: Probability of Traffic Violation by Age