Alcohol consumption as a function of marital status

In [1]:

```
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf
import statsmodels.stats.multicomp as multi
import matplotlib.pyplot as plt

%matplotlib inline
```

Data Cleaning:

Imputation, Subsetting, Recoding, etc.

In [2]:

```
nesarc = pd.read_csv('../nesarc_pds.csv', low_memory=False)

# drop people who report > 20 fl.oz. ethanol consumption per day
# these people are probably dead or bad at reporting consumption levels
nesarc['oz_ethanol'] = nesarc.ETOTLCA2.apply(lambda x: 0.0 if not x.strip() else
 float(x.strip()))
nesarc.drop(['ETOTLCA2'], axis=1, inplace=True)
nesarc.drop(nesarc[nesarc.oz_ethanol>20].index, axis=0, inplace=True)
```

In [3]:

```
alcokids = nesarc.ix[:,['CHLD0', 'CHLD1_4', 'CHLD5_12', 'CHLD13_15', 'CHLD16_17'
, 'CHLD0_17', 'ADULTCH', 'S1Q5A',
                        'S2AQ1', 'S2AQ2', 'S2AQ3', 'CONSUMER', 'oz_ethanol',
                        'S2BQ1A1', 'S2BQ1A2', 'S2BQ1A3', 'S2BQ1A4', 'S2BQ1A5',
'S2BQ1A6', 'S2BQ1A19',
                        'S2BQ2D', 'S2BQ3B', 'S2CQ1', 'ALCABDEP12DX', 'ALCABDEPP
12DX',
                        'BUILDTYP', 'AGE', 'SEX', 'MARITAL', 'S1Q6A', 'S1Q12A',
 'S1Q16',
                        'S1Q24FT', 'S1Q24IN', 'S1Q24LB',
                        'S4CQ3A3']]
del nesarc
```

In [4]:

```
alcokids['abuse_events'] = alcokids.S2BQ3B.apply(lambda x: 0 if not x.strip() el
se int(x.strip()))
alcokids['abuse_events'] = alcokids.abuse_events.apply(lambda x: 0 if x == 99 el
se x)
alcokids.drop(['S2BQ3B'], axis=1, inplace=True)

alcokids['ADULTCH'] = alcokids['ADULTCH'].apply(lambda x: x==1)

# insomnia
alcokids['insomnia'] = alcokids['S4CQ3A3'].apply(lambda x: x.strip()=='1')
alcokids.drop(['S4CQ3A3'], axis=1, inplace=True)

# merge height fields into a single height field (in inches)
alcokids['height'] = alcokids.S1Q24IN + alcokids.S1Q24FT*12
# rename weight
alcokids['weight'] = alcokids.S1Q24LB
# add bmi
alcokids['bmi'] = alcokids.weight / alcokids.height**2 * 703
alcokids.drop(['S1Q24FT', 'S1Q24IN', 'S1Q24LB'], axis=1, inplace=True)


#alcokids['CONSUMER'] = alcokids['CONSUMER'].apply(lambda x: x==1)

# recode marital status
mstat = {1:'married', 2:'cohabiting', 3:'widowed', 4:'divorced', 5:'separated',
6:'single'}
alcokids['marital_status'] = alcokids.MARITAL.apply(lambda x: mstat[x])
alcokids.drop(['MARITAL'], axis=1, inplace=True)
```

In [5]:

```
def is_alcoholic(fields):
    if fields['CONSUMER']==1:
        return False
    else:
        a = sum(fields[['S2BQ1A1', 'S2BQ1A2', 'S2BQ1A3', 'S2BQ1A4', 'S2BQ1A5', '
S2BQ1A6', 'S2BQ1A19',
                        'S2CQ1']] == '1')>2
        b = int(fields['abuse_events']>3)
        c = sum(fields[['ALCABDEP12DX', 'ALCABDEPP12DX']]>0)
        return a+b+c > 1

alcokids['alcoholic'] = alcokids.apply(is_alcoholic, axis=1)
alcokids.drop(['S2BQ1A1', 'S2BQ1A2', 'S2BQ1A3', 'S2BQ1A4', 'S2BQ1A5', 'S2BQ1A6',
 'S2BQ1A19', 'S2CQ1',
               'abuse_events', 'ALCABDEP12DX', 'ALCABDEPP12DX'], axis=1, inplace
=True)
```

ANOVA

In [6]:

```
# simple ANOVA
model1 = smf.ols(formula='oz_ethanol ~ C(marital_status)', data=alcokids).fit()
model1.summary()
```

Out[6]:

OLS Regression Results

Dep. Variable:	oz_ethanol	**R-squared:**	0.012
Model:	OLS	**Adj. R-squared:**	0.012
Method:	Least Squares	**F-statistic:**	105.5
Date:	Wed, 17 Feb 2016	**Prob (F-statistic):**	4.25e-111
Time:	20:53:13	**Log-Likelihood:**	-61489.
No. Observations:	43072	**AIC:**	1.230e+05
Df Residuals:	43066	**BIC:**	1.230e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.5365	0.028	19.266	0.000	0.482 0.591
C(marital_status) [T.divorced]	-0.1181	0.031	-3.802	0.000	-0.179 -0.057
C(marital_status)[T.married]	-0.2715	0.029	-9.455	0.000	-0.328 -0.215
C(marital_status) [T.separated]	-0.0965	0.038	-2.508	0.012	-0.172 -0.021
C(marital_status)[T.single]	-0.0813	0.030	-2.745	0.006	-0.139 -0.023
C(marital_status) [T.widowed]	-0.4112	0.032	-12.913	0.000	-0.474 -0.349

Omnibus:	55343.177	**Durbin-Watson:**	2.005
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	10426093.013
Skew:	7.211	**Prob(JB):**	0.00
Kurtosis:	77.843	**Cond. No.**	16.6

In [7]:

```
# post hoc paired comparisons via Tukey's HSD
mc1 = multi.MultiComparison(alcokids.oz_ethanol, alcokids.marital_status).tukeyh
sd()
mc1.summary()
```

Out[7]:

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
cohabiting	divorced	-0.1181	-0.2065	-0.0296	True
cohabiting	married	-0.2715	-0.3533	-0.1897	True
cohabiting	separated	-0.0965	-0.2062	0.0131	False
cohabiting	single	-0.0813	-0.1658	0.0031	False
cohabiting	widowed	-0.4112	-0.5019	-0.3204	True
divorced	married	-0.1535	-0.1974	-0.1095	True
divorced	separated	0.0215	-0.0637	0.1067	False
divorced	single	0.0367	-0.012	0.0854	False
divorced	widowed	-0.2931	-0.352	-0.2342	True
married	separated	0.175	0.0967	0.2532	True
married	single	0.1902	0.155	0.2253	True
married	widowed	-0.1397	-0.188	-0.0914	True
separated	single	0.0152	-0.0658	0.0962	False
separated	widowed	-0.3146	-0.4022	-0.2271	True
single	widowed	-0.3298	-0.3825	-0.2772	True

Discussion

Here, I am examining mean alcohol consumption (measured in ounces of ethanol) as a function of marital status (married, cohabiting, divorced, separated, widowed, & single).

My null hypothesis (H_0) is that there is no difference in means across different marital statuses. The alternate hypothesis (H_A) is that there is a difference in means.

I performed ANOVA, which suggested that there is a statistically significant difference in means (with p-values ranging from 0.000 to 0.012). However, with Tukey's HSD test, it was revealed that in only some cases should the H_0 be rejected.

For example, there is a significant difference in mean alcohol consumption between divorced and married people (in which case H_0 would be rejected), but there is no statistically significant difference between single and cohabiting people or between divorced and separated people (in which cases we would fail to reject H_0).