Statistical Relationship Between Annual Income and Drinking Alcohol on Multiple Days of the Month

Society often correlates regularly drinking alcohol with lower financial success. Whether it's the stereotype of the drunkard drinking during a weekday afternoon or personal anecdotes, it is often easy to point to a person who both drinks regularly and suffers from a low wage. To support this, people point to the fact that some studies show that habitual drinking is associated with decreased brain function in the short term and long term. With this decreased functionality, it seems logical to assume that career opportunities, especially in white collar jobs, would also diminish which in turn would lead to decreased financial returns. This study attempts to look at the relationship between average income and the number of days out of a month in which alcohol is consumed by collecting data on 16,000 respondents. The study accounts for multiple factors including but not limited to measured IQ, family size, race, sex, and age in order to better isolate the relationship between alcohol consumption and wages. While it is difficult to assign a causal relationship in which regularly drinking alcohol increases wages or vice versa, the two seem to increase together for light to moderate alcohol use.

Looking at the summary statistics of the dependent variable, average income, and the key independent variable, number of days respondents drank in a month, it's clear that the data is left-skewed and not normally distributed. The average number of days is 4.86 and the median is 2. With a data set of 16,000, it's likely that large outliers skewed the mean higher. For average income, the mean and median are both close to $17400, suggesting that this variable is closer to a normal distribution.

The correlation between number of drinking days and averaged income is positive, although residual values increase for more than 20 days. The average income was found by separating respondents by the number of days reported then taking the arithmetic average of each subset. This new set of averages was then regressed on the number of days. This single variable regression shows a R-squared value of 0.3679, which signifies that the number of days in which alcohol consumed can account for nearly 37% of the variation in the dataset. The t-statistic is also significantly large at 96.8, with a 95% confidence interval of [339.59, 353.63]. This single variable regression suggests that for each additional day spent drinking (regardless of number of drinks consumed), the average income of those who consumed for the same number of days increases by $346.62. As the graph shows, the residual values are relatively small for less than 20 days while there are increasingly larger residuals for more than 20 days. This discrepancy is likely due to other factors in the respondents' lives related to their demographics and type of work. For instance, those who drink a glass of wine every day for health reasons may do well financially while another may drink heavily every day due to an unhealthy environment and suffer financially for it. Due to limitations in the data, some demographics can't be adequately accounted for. In order to account for as many differences in the respondents as possible, the regression was expanded to account for the demographics of the respondents by including 10 additional variables. Additional variables like parents' education level, number of siblings, and religion were tested but their significance was deemed too low to include and they had a minimal effect on the r-squared value.

The most notable additions are number of times in a month the individual drank more than 6 drinks, whether they drank at all that month, their sex, and their measured IQ. All of these had t-statistics well above 3 (absolute value), signifying a high possibility of statistical

significance. The number of times a person drank more than six drinks in a day is positively correlated with average income when the person does this for less than 10 days. For values greater than 10 days, there is a negative effect on average income of -$596.53, suggesting that while occasional heavy drinking may signify the ability to drink without impacting financial sources, regularly drinking large amounts is, on average, detrimental to one's income. Unlike the case of the number of days of drinking (at any quantity), it seems much easier to draw a causal relationship between the amount of drinking and one's income. Whether a low income causes heavy drinking or heavy drinking leads to a lower income over time is still unclear. Whether or not the person drank at all appears to have a large effect with a coefficient of over $4400, t-statistic of 86.73, and standard deviation of only $50.78. The decrease in the coefficient on the variable "days" can likely be attributed to imperfect multicollinearity with the variable tracking whether a respondent drank at all, "drnkmo". While there is likely imperfect multicollinearity between the two variables, the magnitude of the t-statistics and standard deviations still make the regression model viable. These values suggest that while increasing the number of days is positive, the biggest effect comes from drinking at any quantity on any number of days. The next most important control variable, sex, shows that being female has a negative effect on income, with an estimated effect of -$349.79. With a t-statistic of -8.49 and standard deviation of $41.20, the value is statistically significant with a 95% confidence interval of -$430 to -$269. Having a large family also had a negative impact on income. For each additional member in the family, individual income dropped by $39.89 with a standard deviation of about $12. At the 5% significance level, the coefficient on family size is statistically significant even if the magnitude of the effect is relatively small in this data set. Finally, average IQ, tracked by the variable "afqtrev", had little effect on income (less than $10 per additional IQ point) but with a t-statistic

of 9.35 is very likely statistically significant. The remaining control variables, employment status, race, education level, region, age, health, and whether respondents lived in a rural area were less statistically significant. Of these, only region and education level were statistically significant at the 5% level. Surprisingly, race and age both had muted effects on average income in this model.

By adding in these additional control variables, the effect of the number of days in which alcohol was consumed decreased. For every additional day respondents drank, there was a $167 increase in wages. The study suggests that the effect is within $159.88 and $174.62 with 95% certainty. While the t-statistics decreased and the relative magnitude of the standard deviation increased for the main independent variable, the r-squared value for the model increased to .6397. This suggests that almost 64% of the variation in the average income can be explained by a combination of the variables used, nearly double of what the main independent variable covers on its own.

In conclusion, the effect of drinking on average income is likely positive. Those that consumed alcohol at all during the month had markedly higher incomes on average and those that consumed for multiple days had additional benefits to their income. There is no clear causal effect, however. It is possible that a high income gives individuals the freedom to drink more regularly, that drinking with superiors after work leads to higher income through nepotism, or any number of other causal relationships. Likely, more than one reason is true at the same time for different groups in the data. While there are limitations in the demographic data, the relationship between income and the number of days in which alcohol is consumed appears consistently positive across all sampled demographics.

Summary Data

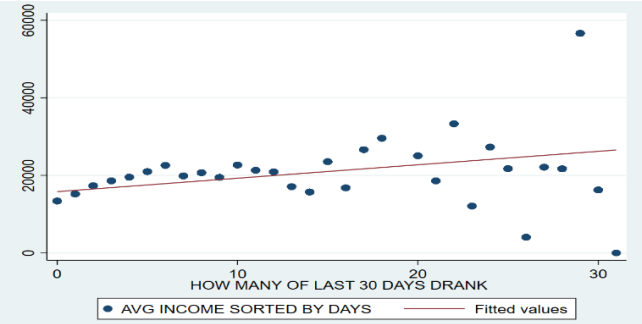
Variable	n	Mean	S.D.	Min	.25	Mdn	.75	Max
					Quantiles			
avgincome	16000	17486.87	3974.04	0.00	13410.91	17343.70	20694.85	56626.55
days	16000	4.86	6.97	0.00	0.00	2.00	6.00	31.00
drnkmo	16000	0.66	0.47	0.00	0.00	1.00	1.00	1.00
drnk6m	16000	0.81	1.48	0.00	0.00	0.00	1.00	6.00
empst	16000	1.39	0.75	1.00	1.00	1.00	1.00	4.00
race	16000	2.38	0.76	1.00	2.00	3.00	3.00	3.00
urbrur	15991	0.80	0.40	0.00	1.00	1.00	1.00	1.00
higrad	16000	12.92	2.41	0.00	12.00	12.00	14.00	20.00
sex	16000	1.52	0.50	1.00	1.00	2.00	2.00	2.00
famsz	16000	3.13	1.64	1.00	2.00	3.00	4.00	15.00
health	15751	0.06	0.23	0.00	0.00	0.00	0.00	1.00
afqtrev	16000	40.03	28.81	1.00	14.00	35.00	64.00	99.00
region	16000	2.61	0.99	1.00	2.00	3.00	3.00	4.00
logage	16000	3.40	0.11	3.18	3.30	3.40	3.47	3.61

Single Variable Regression: Average Income vs. Drinking Days

Source	SS	df	MS		
				Number of obs	= 16,000
				F(1, 15998)	= 9370.81
Model	9.3333e+10	1	9.3333e+10	Prob > F	= 0.0000
Residual	1.5934e+11	15,998	9959957.3	R-squared	= 0.3694
				Adj R-squared	= 0.3693
Total	2.5267e+11	15,999	15793000.6	Root MSE	= 3155.9

avgincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
days	346.6158	3.580634	96.80	0.000	339.5973	353.6342
_cons	15801.5	30.42398	519.38	0.000	15741.86	15861.13

Single Variable Scatter Plot and Linear Fit: avgincome vs days



Multiple Variable Regression

Source	SS	df	MS	Number of obs	=	15,742
				F(23, 15718)	=	1213.18
Model	1.5933e+11	23	6.9273e+09	Prob > F	=	0.0000
Residual	8.9750e+10	15,718	5710032.87	R-squared	=	0.6397
				Adj R-squared	=	0.6391
Total	2.4908e+11	15,741	15823559.8	Root MSE	=	2389.6

avgincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
days	167.2487	3.75987	44.48	0.000	159.8789	174.6185
drnk6m						
1	194.6186	72.08053	2.70	0.007	53.3325	335.9047
2	1116.709	68.28071	16.35	0.000	982.871	1250.547
3	1347.176	94.05448	14.32	0.000	1162.818	1531.533
4	1364.259	137.6061	9.91	0.000	1094.535	1633.983
5	1230.922	199.1462	6.18	0.000	840.5722	1621.271
6	-596.5393	123.5855	-4.83	0.000	-838.781	-354.2975
drnkmo	4404.114	50.78092	86.73	0.000	4304.577	4503.65
empst						
2	-29.66815	83.33756	-0.36	0.722	-193.0193	133.683
3	-121.5041	57.24164	-2.12	0.034	-233.7043	-9.303913
4	3356.528	2392.472	1.40	0.161	-1332.993	8046.049
race						
2	100.3384	63.33152	1.58	0.113	-23.79869	224.4754
3	109.2804	58.94646	1.85	0.064	-6.261444	224.8222
urbrur	147.4492	50.05105	2.95	0.003	49.34343	245.5551
higrad	21.24956	10.72375	1.98	0.048	.229776	42.26935
sex	-349.7916	41.20854	-8.49	0.000	-430.5651	-269.0182
famsz	-39.89663	12.34609	-3.23	0.001	-64.09638	-15.69688
health	-101.6165	83.73631	-1.21	0.225	-265.7493	62.51628
afqtrev	9.265976	.9913294	9.35	0.000	7.322856	11.2091
region						
2	151.1541	60.45153	2.50	0.012	32.66219	269.6461
3	117.3567	56.33397	2.08	0.037	6.935687	227.7778
4	152.7647	63.80518	2.39	0.017	27.69918	277.8301
logage	-261.617	175.5317	-1.49	0.136	-605.6793	82.44533
_cons	14144.78	608.6428	23.24	0.000	12951.77	15337.79