

Problem Set 2
ANSWERS

1. If this output provides a direct estimate of θ and its standard error, as well as a direct estimate of all but one of the β 's, how must X be defined?

d) X is mlit + father's literacy

Since $\theta = \beta_3 - \beta_4 = 0$, we can substitute $\beta_3 = \theta + \beta_4$ into the original model. This gives us $\beta_0 + \beta_1\text{age} + \beta_2\text{rural} + (\theta + \beta_4)\text{mother's literacy} + \beta_4\text{father's literacy} + \beta_5\text{female} + u$. We gather terms and get $\beta_0 + \beta_1\text{age} + \beta_2\text{rural} + \theta\text{mother's literacy} + \beta_4(\text{father's literacy} + \text{mother's literacy}) + \beta_5\text{female} + u$, so X must be the sum of mother's and father's literacy

2. If this output provides a direct estimate of θ and its standard error, as well as a direct estimate of all but one of the β 's, what is the estimate of θ ?

a) -.49

As seen above, θ is now the coefficient on mother's literacy.

3. Consider the probability density function for a t-distribution. As degrees of freedom increase, the symmetric cutoff points that leave only 5% of the probability in the tails of the distribution:

b) move closer together (i.e. closer to zero)

As degrees of freedom go to infinity, the t-distribution becomes less and less spread out (and becomes identical to the standard normal distribution).

4. Suppose you want to test for whether you can predict house price based on assessments, without knowing house characteristics. You estimate $\text{price} = \beta_0 + \beta_1\text{assess} + \beta_2\text{lotsize} + \beta_3\text{sqrft} + \beta_4\text{bdrms} + u$ using a sample of 125 houses. Then your best approach to answering the question is to:

d) estimate $\text{price} = \beta_0 + \beta_1\text{assess} + u$ and calculate $((R^2_1 - R^2_2)/3) / ((1 - R^2_1)/(120))$, where R^2_1 is the R^2 from the first regression and R^2_2 is the R^2 from this second regression. If it is > 2.68 you definitely need the characteristics.

This is an F-test for the exclusion restrictions $H_0: \beta_2=0, \beta_3=0, \beta_4=0$, which is what you want to do. The 5% critical value with 3, 120 df is 2.68, so this means rejecting the null at the 5% level.

5. Suppose you want to estimate a model that will allow you to test the effect of being a pot smoker (pot=1) and a drinker (drink=1) on the number of sexual partners (partners) an individual has had. The best model to estimate would be:

c) $\text{partners} = \beta_0 + \beta_1\text{pot} + \beta_2\text{drink} + \beta_3\text{pot}*\text{drink} + u$

With this model, β_3 would give you the *additional* effect of being both a pot smoker and a drinker.

β_1 would estimate the effect of pot smoking only and β_2 would be drinking only. While $\beta_1 + \beta_2$ would be the total effect with model a), if there is an additional effect of being both, only c) will capture it. In any case, all other choices are nested in model c), so it would let you test if a simpler one is sufficient.

Part II. Stata Problems.

1. a) . desc

```
Contains data from C:\My Documents\ECON 20\Other Data\driving.dta
obs:          12,818
```

```
vars:          10
size:         563,992 (99.4% of memory free)
```

variable name	storage type	display format	value label	variable label
gradecat	float	%9.0g	q7	how were your grades past 12 months
female	float	%9.0g		=1 if female
ageyrs	float	%9.0g		age in years
white	float	%9.0g		self-identified as white
black	float	%9.0g		self-identified as black
hispanic	float	%9.0g		self-identified as hispanic, or mixed hispanic
other	float	%9.0g		self-identified as native, asian, hawaiian or mixed
belt pct	float	%9.0g		% of time wear a seatbelt
drivedrunk	float	%9.0g		have driven while drinking in last 30 days
ridedrunk	float	%9.0g		rode with drinking driver in last 30 days

Sorted by:

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gradecat	12818	2.180527	.9167696	1	5
female	12818	.5147449	.499802	0	1
ageyrs	12807	16.20419	1.20896	12	18
white	12650	.4811858	.4996656	0	1
black	12650	.1914625	.3934677	0	1
hispanic	12650	.2447431	.4299518	0	1
other	12650	.0826087	.2753007	0	1
belt pct	12728	78.04211	23.61128	20	100
drivedrunk	12551	.131384	.3378332	0	1
ridedrunk	12788	.322568	.4674773	0	1

```
. tab gradecat, gen(gradedum)
```

how were your grades past 12 months	Freq.	Percent	Cum.
mostly a's	3156	24.62	24.62
mostly b's	5282	41.21	65.83
mostly c's	3482	27.16	92.99
mostly d's	706	5.51	98.50
mostly f's	192	1.50	100.00
Total	12818	100.00	

```
. gen nodrive=1-drivedrunk
(267 missing values generated)
```

```
. gen noride=1-ridedrunk
(30 missing values generated)
```

b). reg beltpct female ageyrs black hispanic other nodrive noride gradedum1-gradedum4

Source	SS	df	MS			
Model	723447.809	11	65767.9827	Number of obs = 12276		
Residual	6029925.67	12264	491.676914	F(11, 12264) = 133.76		
				Prob > F = 0.0000		
				R-squared = 0.1071		
				Adj R-squared = 0.1063		
				Root MSE = 22.174		
Total	6753373.48	12275	550.172992			

beltpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	2.537805	.4108782	6.18	0.000	1.732419	3.343191
ageyrs	.5709901	.1686835	3.38	0.001	.240344	.9016363
black	-2.480124	.5493792	-4.51	0.000	-3.556994	-1.403254
hispanic	2.546094	.5031856	5.06	0.000	1.559771	3.532417
other	-.6226304	.7598386	-0.82	0.413	-2.112034	.8667728
nodrive	8.841244	.6539211	-13.52	0.000	-10.12303	-7.559456
noride	8.417635	.4671284	-18.02	0.000	-9.333281	-7.50199
gradedum1	18.04048	1.75731	10.27	0.000	14.59587	21.48508
gradedum2	16.30865	1.732964	9.41	0.000	12.91177	19.70554
gradedum3	10.53914	1.743745	6.04	0.000	7.121121	13.95715
gradedum4	6.654881	1.902838	3.50	0.000	2.925019	10.38474
_cons	39.68469	3.266785	12.15	0.000	33.28128	46.0881

Note that *female*, *ageyrs*, *black*, *hispanic* and *other* are demographics, *nodrive* and *noride* are other car behaviors and *gradedum1-gradedum4* are school behaviors. By giving you the “comparison” group, I was telling you which groups to omit when including dummy variables. I probably should have made not driving drunk and not riding drunk the omitted groups to save making new dummy variables, but it gives you some good practice and makes you think about it instead of just tossing in the available ones.

c) From the above we can see that compared to whites, blacks are significantly less likely to wear seatbelts, while Hispanics are significantly more likely. In both cases the size of the effect is about 2.5 percentage points. People who ride with drunk drivers are about 8.4 percentage points less likely to wear seatbelts, and people who drive drunk are almost 9 percentage points less likely. Both are significant. Being a better student seems to significantly increase the probability of wearing a seatbelt. Compared to those who usually get F’s, each category is increasingly more likely to wear seatbelts. Those who usually get A’s are 18 percentage points more likely than those who usually get F’s.

d) Again based on the above, females are significantly more likely to wear a seatbelt. If we had a male dummy, the coefficient would have been the negative of the female dummy, so -2.54.

```
. gen femdrive=female*drivedrunk
(267 missing values generated)
```

```
. gen femride=female*ridedrunk
(30 missing values generated)
```

```
. reg beltpct ageyrs female other black hispanic drivedrunk ridedrunk gradedum1-
gradedum4 femdriv
> e femride
```

Source	SS	df	MS			
				Number of obs = 12276		
				F(13, 12262) = 114.80		

Model		732741.165	13	56364.705	Prob > F	=	0.0000
Residual		6020632.31	12262	490.99921	R-squared	=	0.1085

Total		6753373.48	12275	550.172992	Adj R-squared	=	0.1076
					Root MSE	=	22.159

belt_pct		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

ageyrs		.5882737	.1686248	3.49	0.000	.2577425	.9188049
female		1.340923	.4946625	2.71	0.007	.3713068	2.31054
other		-.5383448	.7595619	-0.71	0.478	-2.027206	.9505161
black		-2.397099	.5494766	-4.36	0.000	-3.47416	-1.320038
hispanic		2.54908	.5028553	5.07	0.000	1.563405	3.534756
nodrive		8.856183	.8478753	-10.45	0.000	-10.51815	-7.194214
noride		10.28738	.6807742	-15.11	0.000	-11.62181	-8.952959
gradedum1		17.97457	1.756189	10.23	0.000	14.53216	21.41697
gradedum2		16.19452	1.731977	9.35	0.000	12.79958	19.58947
gradedum3		10.46947	1.742621	6.01	0.000	7.053662	13.88529
gradedum4		6.571023	1.901626	3.46	0.001	2.843537	10.29851
femdrive		.8148789	1.313548	0.62	0.535	-1.759882	3.389639
femride		3.529946	.927051	3.81	0.000	1.71278	5.347112
_cons		38.2029	3.291576	11.61	0.000	31.7509	44.65491

Note that I wrote this question thinking I was including *drivedrunk* and *ridedrunk*. Since *nodrive* and *noride* are included, we have to think about the negative of the coefficients in order to interpret in terms of driving and riding with drunks. The same was true in part c). Sorry about that.

Only the interaction of *female* and *ridedrunk* is significantly different from zero. The other interaction is only significant at the 54% level. For males, the impact of riding with a drunk is estimated from the main effect so they are 10.3 percentage points less likely to wear a seatbelt. For females it is the sum of the main effect and the interaction, or $-10.3 + 3.5$. So, females that ride with a drunk driver are 6.8 percentage points less likely to wear seatbelts. Finally, the derivative of seatbelt wearing with respect to being female is $1.34 + .8*\textit{drivedrunk} + 3.53*\textit{ridedrunk}$. Note this implies that the positive effect of being female is increased by bad behaviors. What do you make of this finding?