Problem 1

In this question, consider models LINEAR, QUAD, and DUMMIES in the Stata output below. In these models, the dependent variable is price of a diamond ring and the pairs of numbers given are the OLS coefficients and their standard errors. The dummy d1 indicates above-average, d2 indicates average, and d3 indicates below-average quality of diamond.

```
. sum price lnprice size sizesq lnsize d1 d2 d3
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	48	500.0833	213.6428	223	1086
lnprice	48	6.134642	.3950927	5.407172	6.990256
size	48	2.041667	.5678752	1.2	3.5
sizesq	48	4.484167	2.632861	1.44	12.25
lnsize	48	.6792832	.2597902	.1823215	1.252763
d1	48	.2708333	.4490929	0	1
d2	48	.1041667	.3087093	0	1
d3	48	.625	.4892461	0	1

```
. est table LINEAR LINHET QUAD DUMMIES LOGLIN LOGLOG, b(%10.3f) se stats(N F r2 r2_a rmse rss)
```

Variable	LINEAR	LINHET	QUAD	DUMMIES	LOGLIN	LOGLOG
size	372.102	372.102	292.013	372.182	0.679	
	8.179	7.775	68.130	8.362	0.023	
sizesq			17.399			
			14.695			
d1				3.982		
				10.796		
d2				1.552		
				15.724		
lnsize						1.498
						0.038
_cons	-259.626	-259.626	-174.130	-261.027	4.749	5.117
	17.319	15.856	74.238	18.219	0.048	0.028
N	48	48	48	48	48	48
F	2069.991	2290.555	1044.740	662.090	906.175	1515.544
r2	0.978	0.978	0.979	0.978	0.952	0.971
r2_a	0.978	0.978	0.978	0.977	0.951	0.970
rmse	31.841	31.841	31.702	32.506	0.088	0.069
rss	46635.671	46635.671	45226.677	46491.431	0.354	0.216

```
t_.05,v for v = 48     v = 47     v = 46     v = 45     v = 44     v = 43
           1.6772242  1.6779267  1.6786604  1.6794274  1.68023      1.6810707
t_.025,v for v = 48    v = 47     v = 46     v = 45     v = 45     v = 43
           2.0106348  2.0117405  2.0128956  2.0141034  2.0141034    2.0166922
t_.01,v for v = 48     v = 47     v = 46     v = 45     v = 45     v = 43
           2.4065813  2.4083451  2.4101881  2.4121159  2.4121159    2.4162501
t_.005,v for v = 48    v = 47     v = 46     v = 45     v = 45     v = 43
           2.682204   2.6845556  2.6870135  2.689585   2.689585     2.6951021

F_.05,v1,v2 for  v1,v2=2,48  v1,v2=2,47  v1,v2=2,46  v1,v2=2,45  v1,v2=2,44  v1,v2=2,43
                 3.1907273   3.1950563   3.1995817   3.2043173   3.209278    3.2144803
F_.05,v1,v2 for  v1,v2=3,48  v1,v2=3,47  v1,v2=3,46  v1,v2=3,45  v1,v2=3,44  v1,v2=3,43
                 2.7980606   2.8023552   2.8068449   2.8115435   2.8164658   2.8216282
```

Part a. In model `QUAD`, what is the marginal effect at the mean on price of increasing size by one unit?

Part b. After controlling for the size of the diamond, what is the difference in price between a ring of average quality and a ring of below-average quality?

Part c. After controlling for the size of a diamond, what is the difference in price between a ring of above-average quality and a ring of average quality?

Part d. Are all of the regressors in model `DUMMIES` jointly significant at significance level 0.05? Perform an appropriate test. State clearly the null and alternative hypotheses of your test as well as your conclusion.

Part e. Are the dummy variables `d2` and `d2` in model `DUMMIES` jointly statistically significant at significance level 0.05? Perform an appropriate test. State clearly the null and alternative hypotheses of your test as well as your conclusion.

Part f. Do you see any problems in adding the variable `d3` as a regressor in the model `DUMMIES`? Explain.

Part g. Using a measure of model fit that controls for the size of the model, which of the three models best explains the data? Explain your answer.

Part h. Provide a meaningful interpretation of the effect of variable `size` on `price` in model `LOGLIN`.

Part i. Provide a meaningful interpretation of the effect of variable `size` on `price` in model `LOGLOG`.

Part j. Suppose we use model `LOGLIN`. Do you see any problems in using

$$\widehat{price} = \exp(4.749 + 0.679 \times size)$$

to predict price? Explain.

Problem 2

Consider the following regression that you are probably sick of seeing by now. Recall that variable `tv` is in units of $1000.

```
. regress sales tv
```

Source	SS	df	MS		
Model	3.3146e+09	1	3.3146e+09	Number of obs	= 200
Residual	2.1025e+09	198	10618841.6	F(1, 198)	= 312.14
				Prob > F	= 0.0000
Total	5.4171e+09	199	27221853	R-squared	= 0.6119
				Adj R-squared	= 0.6099
				Root MSE	= 3258.7

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tv	47.53664	2.690607	17.67	0.000	42.23072	52.84256
_cons	7032.594	457.8429	15.36	0.000	6129.719	7935.468

Degrees of freedom:	200	199	198	197	196	195	194	193
t_.05:	1.6525	1.6525	1.6526	1.6526	1.6527	1.6527	1.6527	1.6528
t_.025:	1.9719	1.9720	1.9720	1.9721	1.9721	1.9722	1.9723	1.9723
t_.01:	2.3451	2.3452	2.3453	2.3454	2.3455	2.3456	2.3457	2.3458
t_.005:	2.6006	2.6008	2.6009	2.6010	2.6011	2.6013	2.6014	2.6015

Part a. Predict the actual sales when tv advertising equals $100,000.

Part b. A statistician states that a 95 percent confidence interval for actual sales given tv advertising equals $100,000 will have width of at least 10,000 units. Is she correct? Explain your answer. (This is tricky.)

Part c. Consider the model below that accounts for region of advertising, which is captured by dummy variables `region1` and `region2`.

```
. regress sales tv radio newspaper tvbynews region1 region2
```

Source	SS	df	MS		
Model	4.8988e+09	6	816466409	Number of obs	= 200
Residual	518350292	193	2685752.81	F(6, 193)	= 304.00
				Prob > F	= 0.0000
Total	5.4171e+09	199	27221853	R-squared	= 0.9043
				Adj R-squared	= 0.9013
				Root MSE	= 1638.8

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tv	38.80747	2.31232	16.78	0.000	34.24681	43.36813
radio	187.3695	8.701045	21.53	0.000	170.2081	204.5308
newspaper	-32.16059	10.46367	-3.07	0.002	-52.79842	-11.52276
tvbynews	.2010003	.0568861	3.53	0.001	.088802	.3131985
region1	-404.474	346.3489	-1.17	0.244	-1087.589	278.6409
region2	-308.8007	275.7715	-1.12	0.264	-852.7135	235.1121
_cons	4246.044	493.7597	8.60	0.000	3272.187	5219.902

How does the regression change if we replace `region1` with `region3`?

Problem 3

```
. regress mpg hp curbwt torque disp
```

Source	SS	df	MS		
Model	6955.79742	4	1738.94935	Number of obs	= 330
Residual	2764.22219	325	8.50529904	F(4, 325)	= 204.45
				Prob > F	= 0.0000
Total	9720.0196	329	29.5441325	R-squared	= 0.7156
				Adj R-squared	= 0.7121
				Root MSE	= 2.9164

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hp	-.0432345	.0042664	-10.13	0.000	-.0516277	-.0348412
curbwt	-.0025332	.0004105	-6.17	0.000	-.0033408	-.0017256
torque	.0142477	.0035139	4.05	0.000	.0073348	.0211606
disp	-.8329362	.3037788	-2.74	0.006	-1.430557	-.2353152
_cons	44.40531	1.119392	39.67	0.000	42.20314	46.60748

In the regression above, do you think multicollinearity is a problem? Explain.

Problem 4

A regression of `wage` (hourly wage) on an intercept and an indicator variable `gender` (equal to 1 if female and equal to 0 if male) leads to an estimate $\widehat{wage} = 20 - 4 \times gender$. What are average wages for men and for women in the sample?