Problem Set 8 – Answers

Text book problems:

7.4

a. variance of regression s^2=RSS/(n-k) where k is the number of parameters we estimate. Hence in this case s^2=880/(25-4)=41.9
b. s^2=1220/(14-4)=122.

7.12

a. Marginal propensity to consume is 0.93, i.e. for each additional dollar of disposable income consumption expenditure increases by 94 cents.
b. H_0: B_2=1
   H_1: $B_2 \neq 1$
   test statistic $t=(B_2-1)/s_{b2}$. Standard error of the estimate of B_2 is s_{b2}=0.93/249.06=0.00373 t=(0.93-1)/0.00373=18.7. The critical region for a two sided test at 5% level of confidence with 73 degrees of freedom is (-∞ ,–2) and (2,∞). The test statistic is in the critical region and hence we reject the null hypothesis that B_2=1.
c. The higher the interest rate the greater the incentive for saving rather than consumption. I would expect the coefficient to be negative.
d. The critical values as found in part b are –2 and 2. The t-statistic –3.09 falls into the rejection region.

Computer Exercises:

Please note that, as I have mentioned in class, there is no single "correct" answer. The model below is one among several reasonable models. We may have different assessments about which variables matter.

1.My regression problem sets 6 and 7 involved regressing scores on missed classes. There was a negative coefficient. The interpretation I gave to this result was that attending classes increases your expected score, because what is being taught in the class is valuable. Clearly, I would like to believe that interpretation but someone may object that perhaps the reason for a negative coefficient is that students who do not come to class are irresponsible and also do not study. Hence, the negative coefficient does not reflect the effect of a valuable class instead it is the effect of not studying. In order to defend my interpretation I ran a multiple regression where I control for hours studied. Fortunately, the coefficient on missed classes is still negative. Hence, holding hours of studying constant missing classes still lowers your expected score. Unfortunately, all coefficients in this regression are insignificant. It may have to do with the low number of observations or with the fact that we do not know the data generating process that gives rise to outcomes on the midterm.

```
. regress score  miss   hours_e hours_c

  Source |       SS       df       MS                  Number of obs =      27
---------+------------------------------               F(  3,    23) =    0.48
   Model |  187.326887     3  62.4422956               Prob > F      =  0.6967
Residual |  2967.33978    23  129.014773               R-squared     =  0.0594
---------+------------------------------               Adj R-squared = -0.0633
   Total |  3154.66667    26  121.333333               Root MSE      =  11.358
```

```
------------------------------------------------------------------------------
   score |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    miss |  -1.180324    1.99836     -0.591   0.561      -5.314246    2.953599
 hours_e |  -.6803851   1.041752     -0.653   0.520      -2.835413    1.474643
 hours_c |  -.6480525   1.505952     -0.430   0.671      -3.763351    2.467246
   _cons |   61.79923   6.213077      9.947   0.000        48.9465    74.65196
------------------------------------------------------------------------------
```

2.
(a) Fatalities per 1000 licensed drivers is our dependent variable. The more people drive, the higher the chance they will have accidents, so I have included miles driven per capita. I would expect more accidents to occur when people exceed the speed limit, so I include percent exceeding 55 in 55 mph areas. My guess is that more accidents will occur in densely populated areas. There is no reason for the effect to be linear, so I could include population density and population density squared. An alternative is to use log population density, which is what I do. (Here I am assuming a constant impact on the dependent variable of percentage increases in population density.) Spending on highway law enforcement should reduce the number of accidents. Finally, I know that insurance companies charge higher rates for young drivers, so I suspect states with a larger proportion of drivers under 20 should have higher fatality rates.

Note: Once we use percentage of drivers exceeding a 55 mph speed limit, we shouldn't also use the percentage exceeding 65 miles per hour. The two variables are measuring similar things, and we will have problems with multi-collinearity (the correlation between the two variables is actually 0.65).

Determinants of Automobile Fatalities

Dependent Variable: Fatalities per thousand licensed drivers

Variable	Coefficient (t-statistic)
Annual vehicle miles per capita	0.0239 (3.297)*
Percent exceeding 55 in 55 speed limit areas	-0.0000547 (-0.066)
Log population density	-0.0184 (-2.431)*
Percent under 20	0.0014 (0.153)
Spending on law enforcement per capita ($)	-0.0008 (-1.585)
Constant	0.1249 (1.21)
Number of observations	50
Mean of dependent variable	0.25
R^2, Adjusted R^2	0.47, 0.41

Note: T-stats are in parentheses, * = significant at 5%, two-tail.

(c) As you can see, only some of my conjectures are born out by the data. Only the coefficients on the miles driven per capita and population density are statistically significant (in each case the null that the coefficient is zero is rejected at 5%, two-tail). An extra mile driven per capita is associated with an increase of 0.0239 in the number of fatalities per thousand drivers in 1993. The elasticity at the sample means is (0.0239)(9.39/0.25) = 0.898. Thus, a 1% increase in the miles driven per capita is associated with an increase of 0.898% in fatalities per driver. This is a large effect, though it is not in the least bit surprising.

The log of population density enters negatively (to my surprise) and is statistically significant. It appears the more densely populated areas have fewer accidents. A 1% increase in population density is associated with a decline of 0.00018 in the number of accidents per thousand licensed drivers. To find the elasticity we need the percentage change in the dependent variable from a 1% change in the explanatory variable: (-0.00018 /0.25)*100 = -0.072; a 1% increase in population density is associated with a decline in fatalities per driver of 0.07%. It is possible that drivers are more careful in densely populated areas.

None of the other variables is statistically significant, so we do not have much to comment on. I suspect these results suffer from omitted variable bias. In particular, it is hard to believe that the percentage of young drivers has nothing to do with automobile fatalities. In the case of spending on highway law enforcement per capita, if we did a one-tail tail test at 10% (test the null that the coefficient is zero against the alternative that it is negative), we would have statistical significance. An increase of one dollar per capita in law enforcement is associated with a decrease in fatalities per thousand licensed drivers of 0.00082. The elasticity here works out to: (-0.00082)(28.52/0.25) = -0.09. So a 1% increase in highway law enforcement expenditures per capita is associated with a decline of 0.09% in fatalities per licensed driver.

(e) The model "explains" 47% of the variation in fatalities per driver across states, and null that all the slopes are jointly zero is clearly rejected (the p-value associated with the F-statistic is almost zero).

(f) There is some support (not very strong)for the argument that spending more on law enforcement will reduce fatalities. It would also be interesting to figure out why the less densely populated areas have higher fatality rates (maybe because speed limits are higher in rural areas?).