Problem Set 4 Practice Problem Solutions only

1. US states differ in the generosity of their welfare programs. We here wish to analyze which factors play a role in the level of benefits across different states. The data set TANF2.dta contains data from each of 49 states. The variables in the data set are given in the following table:

Table 3
DATA DESCRIPTION, FILE: TANF2.dta

Variable	Definition
tanfreal	State's real maximum benefit for single parent with three kids.
black	Percentage of state's population who are African Americans.
blue	Dummy variable, equals 1 if state voted Democratic in 2004 presidential election.
mdinc	State's median income.
west	= 1 if state is in West = 0 otherwise
south	= 1 if state is in South. = 0 otherwise.
midwest	= 1 if state is in Midwest = 0 otherwise
northeast	= 1 if state is in Northeast =0 otherwise

Use data set TANF2.dta to examine whether Midwest states differ in their welfare programs from other states. To do this, we will use the following regression model:

$$tanfreal = \beta_0 + \beta_1\, black + \beta_2\, blue + \beta_3\, midwest + \beta_4\, (black{*}midwest) + \beta_5\, (blue{*}midwest) + u$$

Here, *black*midwest* is the product of the regressors *black* and *midwest* and so forth.

(a) Write the null hypothesis to test whether there is a difference between the welfare programs of Midwest states and all other states, explain.

H_0: $\beta_3 = \beta_4 = \beta_5 = 0$

Under H_0, we have $\beta_0 + \beta_1\, black + \beta_2\, blue + u$

That is, the expected level of benefits does not depend on whether the state is in the Midwest or not.

(b) Construct new set of interaction regressors in STATA. Estimate the model above. Write your answer as a regression equation with standard errors in parenthesis underneath each coefficient. Perform the test for the null hypothesis in part (a) with a robust F-test. What is your conclusion?

The fitted model is:

$$tam\hat{f}real = 347.53 - 522.03 \times black + 31.76 \times blue + 141.42 \times midwest - 1420.53 \times [black \cdot midwest] - 204.14 \times [blue \cdot midwest]$$

For non-Midwest states(*Midwest*=0), the fitted expected benefits level therefore is:

$\widehat{tamfreal} = 347.53 - 522.03 \times black + 31.76 \times blue$

So a larger black population has a negative effect on benefit levels while Democratic states tend to give higher levels.

For Midwest states (*Midwest*=1)

$\widehat{tamfreal} = 488.95 - 1942.56 \times black - 172.38 \times blue$

Again, the effect of size of black population is negative, but much larger impact than for non-Midwest states. Midwest Democratic states tend to give lower benefits. With the command **test**, we test H_0 in STATA and obtain $F_{obs} = 26.68$ which is from $F_{3,\infty}$ distribution. STATA reports a p-value of 0.00. This is based on a $F_{3,43}$ distribution which is close to the $F_{3,\infty}$ one. So we reject the hypothesis and conclude that Midwestern states have significantly different welfare programs compared with non-Midwest ones.

(c) Introduce a new variable *nonmidwest* = 1 – *midwest.* That is *nonmidwest* = 1 if a state is not in the Midwest and zero otherwise. Consider the following alternative regression model:

tanfreal = γ_1 *nonmidwest* + γ_2 *(black*nonmidwest)* + γ_3 *(blue*nonmidwest)* + γ_4 *midwest*
+ γ_5 *(black*midwest)* + γ_6 *(blue*midwest)* + u

Write up the hypothesis of no differences in welfare programs in terms of $\gamma_1 .... \gamma_6$
What is the relationship between the parameters $\gamma_1 .... \gamma_6$ in this new model and $\beta_1 .... \beta_6$ in the previous model? Estimate the model in STATA and write the result in usual regression equation form with standard errors in parentheses underneath coefficients.

The hypothesis of no differences in welfare programs in terms of $\gamma_1, \cdots, \gamma_6$ amounts to $\gamma_1 = \gamma_{4,} \gamma_2 = \gamma_5$ and $\gamma_3 = \gamma_6$. We can rewrite the model as

tanfreal = γ_1 *(1-midwest)* + γ_2 *(black -black*midwest)* + γ_3 *(blue-blue*midwest)* + γ_4 *midwest*
+ γ_5 *(black*midwest)* + γ_6 *(blue*midwest)* + u
= γ_1 + γ_2 *black* + γ_3 *blue* + (γ_4- γ_1)* *midwest*
+ ($\gamma_5 - \gamma_2$)* *(black*midwest)* + ($\gamma_6 - \gamma_3$ *)(blue*midwest)* + u

That is, $\beta_0 = \gamma_{1,} \beta_1 = \gamma_2, \beta_2 = \gamma_3, \beta_3 = \gamma_4 - \gamma_{1,} \beta_4 = \gamma_5 - \gamma_2$ and $\beta_5 = \gamma_6 - \gamma_3$. So the two models are just different parameterizations of the same equations. This relationship implies that the OLS estimates become

$$\hat{\gamma}_1 = \hat{\beta}_0 = 347.53, \hat{\gamma}_2 = \hat{\beta}_1 = -522.03, \hat{\gamma}_3 = \hat{\beta}_2 = 31.76, \hat{\gamma}_4 = \hat{\beta}_3 + \hat{\gamma}_1 = 488.95,$$
$$\hat{\gamma}_5 = \hat{\beta}_4 + \hat{\gamma}_2 = -1942.56, \hat{\gamma}_6 = \hat{\beta}_5 + \hat{\gamma}_3 = 235.91$$

(d) What happens if you include an intercept γ_0 in the model in part (c)? Explain.

By including an intercept term in the model, the model will suffer from multicollinearity since *Midwest+nonmidwest*=1. Thus, we cannot estimate the model by OLS since the model is overparameterized.

Do file for the question above:

```
clear all
set more off
cap log close

use "/Users/Downloads/tanf2.dta"

log using "/Users/W3412/ps5/q10.smcl", replace
gen blackmid = black*midwest
gen bluemid = blue*midwest
reg tanfreal black blue midwest blackmid bluemid, robust
test midwest blackmid bluemid
gen nonmidwest = 1 - midwest
gen blacknomid = black*nonmidwest
gen bluenomid = blue*nonmidwest
reg tanfreal nonmidwest blacknomid bluenomid midwest blackmid bluemid, robust noconstant
reg tanfreal nonmidwest blacknomid bluenomid midwest blackmid bluemid, robust
log close
```

2. [Practice question, not graded] SW Empirical Exercise E8.1 (use lead_mortality.dta)

Calculations are carried out using **lead_mortality.dta**

(a) The table shows the sample mean ($\bar{Y}$) and its standard error for lead and no-lead cities. The difference in the sample means is 0.022 with a standard error of 0.024. The estimate implies that cities with lead pipes have a larger infant mortality rate (by 0.02 deaths per 100 people in the population), but the standard error is large (0.024) and the difference is not statistically significant (t = 0.022/0.024 = 0.090).

	n	$\bar{Y}$	SE($\bar{Y}$)
Lead	117	0.403	0.014
No Lead	55	0.381	0.020
Difference		0.022	0.024

(b) The regression is

$$\widehat{Infrate} = \underset{(0.150)}{0.919} + \underset{(0.208)}{0.462}\times lead - \underset{(0.021)}{0.075}\times pH - \underset{(0.028)}{0.057}\times lead\times pH$$

(i) The first coefficient is the intercept, which shows the level of *Infrate* when *lead* = 0 and *pH* = 0. It dictates the level of the regression line.

The second coefficient and fourth coefficients measure the effect of *lead* on the infant mortality rate. Comparing 2 cities, one with lead pipes (*lead* = 1) and one without lead pipes (*lead* = 0), but the same of *pH*, the difference in predicted infant mortality rate is

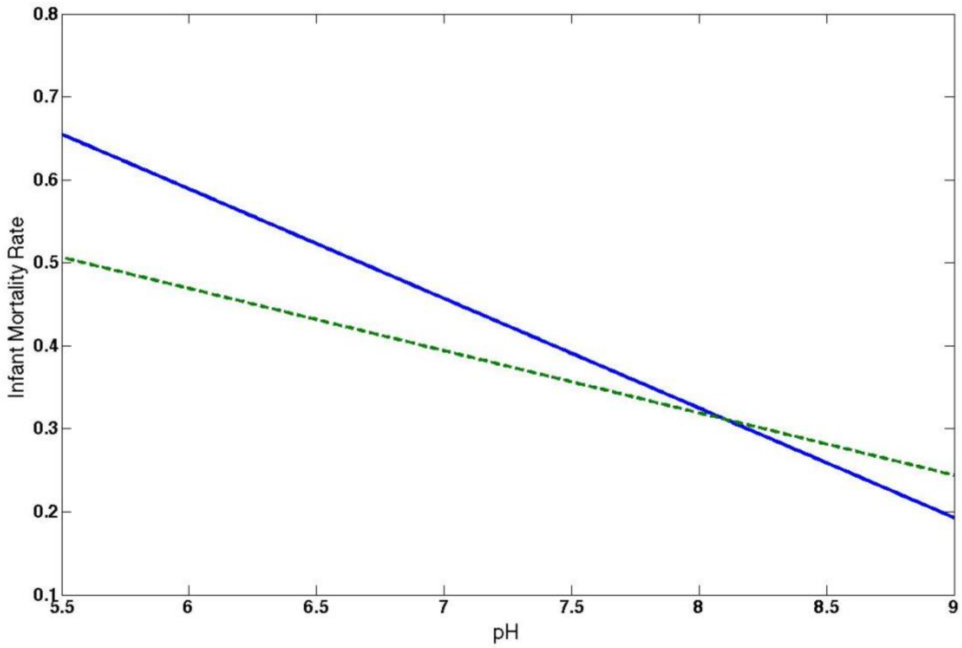
$$\widehat{Infrate}\ (lead = 1) - \widehat{Infrate}\ (lead = 0) = 0.462 - 0.057pH$$

The third and fourth coefficients measure the effect of *pH* on the infant mortality rate. Comparing 2 cities, one with a *pH* = 6 and the other with *pH* = 5, but the same of *lead*, the difference in predicted infant mortality rate is

$$\widehat{Infrate}\ (pH = 6) - \widehat{Infrate}\ (pH = 5) = -0.075 - 0.057 \times lead$$

so the difference is -0.075 for cities without lead pipes and –0.132 for cities with lead pipes.

(ii)



Solid: Cities with lead pipes
Dashed: Cities without lead pipes

The infant mortality rate is higher for cities with lead pipes, but the difference declines as the pH level increases. For example:

The 10th percentile of pH is 6.4. At this level, the difference in infant mortality rates is

$$\widehat{Infrate}(lead=1, pH=6.4)-\widehat{Infrate}(lead=0, pH6.4)=0.462-0.057\times 6.4=0.097$$

The 50th percentile of pH is 7.5. At this level, the difference in infant mortality rates is

$$\widehat{Infrate}(lead=1, pH=7.5)-\widehat{Infrate}(lead=0, pH=7.5)=0.462-0.057\times 7.5=0.035$$

The 90th percentile of pH is 8.2. At this level, the difference in infant mortality rates is

$$\widehat{Infrate}(lead=1, pH=8.2)-\widehat{Infrate}(lead=0, pH=8.2)=0.462-0.057\times 8.2=-0.01$$

(iii) The *F*-statistic for the coefficient on *lead* and the interaction term is *F* = 3.94, which has a *p*-value of 0.02, so the coefficients are jointly statistically significantly different from zero at the 5% but not the 1% significance level.

(iv) The interaction term has a *t* statistic of *t* = –2.02, so the coefficient is significant at the 5% but not the 1% significance level.

(v) The mean of pH is 7.5. At this level, the difference in infant mortality rates is

$$\widehat{Infrate}(lead=1, pH=7.5)-\widehat{Infrate}(lead=0, pH=7.5)=0.462-0.057\times 7.5=0.035$$

The standard deviation of pH is 0.69, so that the mean plus 1 standard devation is 8.19 and the mean minus 1 standard deviation is 6.81. The infant mortality rates at the pH levels are:

$$\widehat{Infrate}(lead=1, pH=8.19)-\widehat{Infrate}(lead=0, pH=8.19)=0.462-0.057\times 8.19=-0.005$$
$$\widehat{Infrate}(lead=1, pH=6.81)-\widehat{Infrate}(lead=0, pH=6.81)=0.462-0.057\times 6.81=0.074$$

(vi) Write the regression as

$$Infrate = \beta_0 + \beta_1 lead + \beta_2 pH + \beta_3 lead\times pH + u$$

so the effect of *lead* on *Infrate* is $\beta_1 + \beta_3 pH$. Thus, we want to construct a 95% confidence interval for $\beta_1 + 6.5\beta_3$. Using method 2 of Section 7.3, add and subtract $6.5\beta_3 lead$ to the regression to obtain:

$$Infrate = \beta_0 + (\beta_1 + 6.5\beta_3)lead + \beta_2 pH + \beta_3(lead\times pH - 0.65 lead) + u$$

or

The estimated regression is

$$\begin{array}{llll} \widehat{Infrate} = 0.919 & +\ 0.092\times lead & -\ 0.075\times pH & -\ 0.057\times lead\times(pH-6.5) \\ \quad\quad\quad\ (0.150) & \ (0.033) & \ (0.021) & \ (0.028) \end{array}$$

and the 95% confidence interval for the coefficient on lead (which is $\beta_1 + 6.5\beta_3$) is 0.027 to 0.157.

(c) There are several demographic variables in the dataset. You should add these and see if the conclusions from (b) change in an important way.

3. [Practice question, not graded] SW Empirical Exercise E8.2 (use CSP2015.dta)

Calculations for this exercise are carried out in the STATA file **CSP2015.do.**

This table contains the results from seven regressions that are referenced in these answers.

Data from 2012

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
	Dependent Variable							
	AHE	ln(*AHE*)	ln(*AHE*)	ln(*AHE*)	ln(*AHE*)	ln(*AHE*)	ln(*AHE*)	ln(*AHE*)
Age	0.510** (0.040)	0.026** (0.002)		0.104* (0.046)	0.104* (0.046)	0.02 (0.06)	0.037 (0.065)	-0.027 (0.073)
Age^2				–0.0013 (0.00077)	–0.0013 (0.00077)	0.0001 (0.0010)	–0.0002 (0.0011)	0.0009 (0.0012)
ln(*Age*)			0.75** (0.06)					
Female × *Age*						0.193* (0.092)		0.174* (0.093)
Female × Age^2						-0.0034* (0.0016)		-0.003 (0.0016)
Bachelor × *Age*							0.128 (0.091)	0.106 (0.928)
Bachelor × Age^2							-0.0021 (0.0015)	-0.0017 (0.0015)
Female	–3.81** (0.22)	–0.19** (0.01)	–0.19** (0.01)	–0.19** (0.01)	–0.24** (0.02)	-2.95* (1.36)	–0.24** (0.02)	-2.67 (1.36)
Bachelor	8.32** (0.22)	0.44** (0.01)	0.44** (0.01)	0.44** (0.01)	0.40** (0.02)	0.40** (0.01)	-1.53 (1.34)	-1.22 (1.35)
Female × *Bachelor*					0.090** (0.022)	0.089** (0.023)	0.090** (0.023)	0.089** (0.023)
Intercept	1.87 (1.18)	1.94** (0.06)	0.15 (0.20)	0.79 (0.67)	0.80 (0.67)	1.99* (0.88)	1.81 (0.95)	2.72* (1.07)
***F*-statistic and *p*-values on joint hypotheses**								
F-stat. on terms involving *Age*				86.0 (0.00)	86.7 (0.00)	45.4 (0.00)	43.74 (0.00	30.47 (0.00)
Interaction terms with *Age* and Age^2						4.14 (0.02)	1.30 (0.27)	2.71 (0.03)
SER	9.68	0.48	0.48	0.48	0.48	0.48	0.48	0.48
$\bar{R}^2$	0.18	0.20	0.20	0.20	0.20	0.20	0.20	0.20

Significant at the *5% and **1% significance level.

(a) The regression results for this question are shown in column (1) of the table. If *Age* increases from 25 to 26, earnings are predicted to increase by \$0.510 per hour. If *Age* increases from 33 to 34, earnings are predicted to increase by \$0.510 per hour. These values are the same because the regression is a linear function relating *AHE* and *Age*.

(b) The regression results for this question are shown in column (2) of the table. If *Age* increases from 25 to 26, ln(*AHE*) is predicted to increase by 0.026, so earnings are predicted to increase by 2.6%. If *Age* increases from 34 to 35, ln(*AHE*) is predicted to increase by 0.026, o earnings are predicted to increase by 2.6%. These values, in percentage terms, are the same because the regression is a linear function relating ln(*AHE*) and *Age*.

(c) The regression results for this question are shown in column (3) of the table. If *Age* increases from 25 to 26, then ln(*Age*) has increased by ln(26) – ln(25) = 0.0392 (or 3.92%). The predicted increase in ln(*AHE*) is 0.75 × (.0392) = 0.029. This means that earnings are predicted to increase by 2.9%. If *Age* increases from 34 to 35, then ln(*Age*) has increased by ln(35) – ln(34) = .0290 (or 2.90%). The predicted increase in ln(*AHE*) is 0.75 × (0.0290) = 0.021. This means that earnings are predicted to increase by 2.1%.

(d) The regression results for this question are shown in column (4) of the table. When *Age* increases from 25 to 26, the predicted change in ln(*AHE*) is

$$(0.104 \times 26 - 0.0013 \times 26^2) - (0.104 \times 25 - 0.0013 \times 25^2) = 0.036.$$

This means that earnings are predicted to increase by 3.6%.
When *Age* increases from 34 to 35, the predicted change in ln(*AHE*) is

$$(0.104 \times 35 - 0.0013 \times 35^2) - (0.104 \times 34 - 0.0013 \times 34^2) = 0.012.$$

This means that earnings are predicted to increase by 1.2%.

(e) The regressions differ in their choice of one of the regressors. They can be compared on the basis of the $\bar{R}^2$. The regression in (3) has a (marginally) higher $\bar{R}^2$, so it is preferred.

(f) The regression in (4) adds the variable Age^2 to regression (2). The coefficient on Age^2 is not statistically significant ($t = -1.72$) and the estimated coefficient is very close to zero. This suggests that (2) is preferred to (4), the regressions are so similar that either may be used.

(g) The regressions differ in their choice of the regressors (ln(*Age*) in (3) and *Age* and Age^2 in (4)). They can be compared on the basis of the $\bar{R}^2$. The regression in (4) has a (marginally) higher $\bar{R}^2$, so it is preferred.

4. [Practice question, not graded] SW Empirical Exercises 9.1 (use CSP2015.dta)

(a) (1) Omitted variables: There is the potential for omitted variable bias when a variable is excluded from the regression that (i) has an effect on ln(*AHE*) and (ii) is correlated with a variable that is included in the regression. There are several candidates. The most important is a worker's *Ability*. Higher ability workers will, on average, have higher earnings and are more likely to go to college. Leaving *Ability* out of the regression may lead to omitted variable bias, particularly for the estimated effect of education on earnings. Also omitted from the regression is *Occupation*. Two workers with the same education (a BA for example) may have different occupations (accountant versus 3rd grade teacher) and have different earnings. To the extent that occupation choice is correlated with gender, this will lead to omitted variable bias. *Occupation* choice could also be correlated with *Age*. Because the data are a cross section, older workers entered the labor force before younger workers (35 year-olds in the sample were born in 1974, while 25 year-olds were born in 1984), and their occupation reflects, in part, the state of the labor market when they entered the labor force.

(2) Misspecification of the functional form: This was investigated carefully in exercise 8.2. There does appear to be a nonlinear effect of *Age* on earnings, which is adequately captured by the polynomial regression with interaction terms.

(3) Errors-in-variables: *Age* is included in the regression as a “proxy” for experience. Workers with more experience are expected to earn more because their productivity increases with experience. But *Age* is an imperfect measure of experience. (One worker might start his career at age 22, while another might start at age 25. Or, one worker might take a year off to start a family, while another might not). There is also potential measurement error in *AHE* as these data are collected by retrospective survey in which workers in March 2013 are asked about their average earnings in 2012.

(4) Sample selection: The data are full-time workers only, so there is potential for sample-selection bias.

(5) Simultaneous causality: This is unlikely to be a problem. It is unlikely that *AHE* affects *Age* or gender.

(6) Inconsistency of OLS standard errors: Heteroskedastic robust standard errors were used in the analysis, so that heteroskedasticity is not a concern. The data are collected, at least approximately, using i.i.d. sampling, so that correlation across the errors is unlikely to be a problem.

(b) Results for 1992 are shown in the table above. Using results from (8), several conclusions were reached in E8.2(l) using the data from 2012. These are summarized in the table below, and are followed by a similar table for the 1992 data.

Results using (8) from the 2012 Data

Gender, Education	**Predicted Value of ln(*AHE*) at Age**			**Predicted Increase in ln(*AHE*) (Percent per year)**	
	25	**32**	**34**	**25 to 32**	**32 to 34**
Females, High School	2.36	2.52	2.53	2.3	0.4
Males, High School	2.60	2.78	2.84	2.5	3.3
Females, BA	2.81	3.03	3.02	3.1	-0.4
Males, BA	2.96	3.19	3.24	3.3	2.6

Results using (8) from the 1992 Data

Gender, Education	**Predicted Value of ln(*AHE*) at Age**			**Predicted Increase in ln(*AHE*) (Percent per year)**	
	25	**32**	**34**	**25 to 32**	**32 to 34**
Females, High School	2.47	2.61	2.59	1.9	–0.6
Males, High School	2.61	2.84	2.88	3.3	2.1
Females, BA	2.83	3.06	3.06	3.2	0.1
Males, BA	2.89	3.21	3.27	4.5	2.8

Based on the 2012 data E81.2 (l) concluded: Earnings for those with a college education are higher than those with a high school degree, and earnings of the college educated increase more rapidly early in their careers (age 25–34). Earnings for men are higher than those of women, and earnings of men increase more rapidly early in their careers (age 25–34). For all categories of workers (men/women, high school/college) earnings increase more rapidly from age 25–32 than from 32–34. While the percentage increase in women's earning is similar to the percentage increase for men from age 25-32, women's earning tend to stagnate from age 32-34, while men's continues to increase.

All of these conclusions continue to hold for the 1992 data (although the precise values for the differences change somewhat.)