Suggested Answer s
Problem Set 3

1. Stock-Watson Chapter 5 Exercises 5.7.

(a) The t-statistic is $\frac{3.2}{1.5} = 2.13$ with a p-value of 0.03; since the p-value is less than 0.05, the null hypothesis is rejected at the 5% level.

(b) $3.2 \pm 1.96 \times 1.5 = 3.2 \pm 2.94$

(c) Yes. If Y and X are independent, then $\beta_1 = 0$; but this null hypothesis was rejected at the 5% level in part (a).

(d) β_1 would be rejected at the 5% level in 5% of the samples; 95% of the confidence intervals would contain the value $\beta_1 = 0$.

2. Stock-Watson Chapter 5 Exercises 5.8.

. (a) $43.2 \pm 2.05 \times 10.2$ or 43.2 ± 20.91, where 2.05 is the 5% two-sided critical value from the t_{28} distribution.

(b) The t-statistic is $t^{act} = \frac{61.5-55}{7.4} = 0.88$, which is less (in absolute value) than the critical value of 20.5. Thus, the null hypothesis is not rejected at the 5% level.

(c) The one sided 5% critical value is 1.70; t^{act} is less than this critical value, so that the null hypothesis is not rejected at the 5% level.

3. You have obtained measurements of height in inches of 29 female and 81 male students (*Studenth*) at your university. A regression of the height on a constant and a binary variable (*BFemme*), which takes a value of one for females and is zero otherwise, yields the following result:

$$\hat{Studenth} = \underset{(0.3)}{71.0} - \underset{(0.57)}{4.84} \times BFemme, \quad R^2 = 0.40, SER = 2.0$$

a. What is the interpretation of the intercept? What is the interpretation of the slope? How tall are females, on average?

<u>Answer</u>: The intercept gives you the average height of males, which is 71 inches in this sample. The slope tells you by how much shorter females are, on average (almost 5 inches). The average height of females is therefore approximately 66 inches.

b. Test the hypothesis that females, on average, are shorter than males, at the 1% level.

Answer: The *t*-statistic for the difference in means is -8.49. For a one-sided test, the critical value is –2.33. Hence the difference is statistically significant.

c. Is it likely that the error term is homoskedastic here?

Answer: It is safer to assume that the variances for males and females are different. In the underlying *sample* the standard deviation for females was smaller.

4. You have obtained a sub-sample of 1744 individuals from the Current Population Survey (CPS) and are interested in the relationship between weekly earnings and age. The regression, using heteroskedasticity-robust standard errors, yielded the following result:

$$\widehat{Earn} = \underset{(20.24)}{239.16} + \underset{(0.57)}{5.20} \times Age\ ,\ R^2 = 0.05,\quad SER = 287.21.,$$

where *Earn* and *Age* are measured in dollars and years respectively.

a. Is the relationship between *Age* and *Earn* statistically significant?

Answer: The *t*-statistic on the slope is 9.12, which is above the critical value from the standard normal distribution for any reasonable level of significance.

b. The variance of the error term and the variance of the dependent variable are related. Given the distribution of earnings, do you think it is plausible that the distribution of errors is normal?

Answer: Since the earnings distribution is highly skewed, it is not reasonable to assume that the error distribution is normal.

c. Construct a 95% confidence interval for both the slope and the intercept.

Answer: The confidence interval for the slope is (4.08,6.32). The confidence interval for the intercept is (199.49,278.83).

5. Shortly before you are making a group presentation on the testscore/student-teacher ratio results, you realize that one of your peers forgot to type all the relevant information on one of your slides. Here is what you see:

$$\widehat{TestScore} = 698.9 - STR \qquad R^2 = 0.051, SER = 18.6$$

$$(9.47) \quad (0.48)$$

In addition, your group member explains that he ran the regression in a standard spreadsheet program, and that, as a result, the standard errors in parenthesis are homoskedasticity-only standard errors.

a. Find the value for the slope coefficient.

<u>Answer</u>: The relationship between the slope coefficient and the regression R^2 is $R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} \Leftrightarrow \hat{\beta}_1^2 = R^2 \times \frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i^2}$

Given the information above, you need to find the TSS ($= \sum_{i=1}^{n} y_i^2$) and $\sum_{i=1}^{n} x_i^2$.

The TSS is relatively easy to find: the *SER* is 18.6, and hence the *SSR* is 144,315.5. (Recall that $SER = s_{\hat{u}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2} = \sqrt{\frac{SSR}{n-2}}$). This allows you to calculate the *TSS*, which is 152,109.6. (Recall that $R^2 = 1 - \frac{SSR}{TSS} \Leftrightarrow TSS = \frac{SSR}{1-R^2}$).

To find $\sum_{i=1}^{n} x_i^2$, note that the homoskedasticity-only standard error for the slope is $s_{\hat{\beta}_1} = \frac{s_{\hat{u}}}{\sqrt{\sum_{i=1}^{n} x_i^2}} \Leftrightarrow \sum_{i=1}^{n} x_i^2 = \left(\frac{SER}{s_{\hat{\beta}_1}}\right)^2$. Hence, $\sum_{i=1}^{n} x_i^2 = 38.7^2 = 1{,}499.6$.

Inserting these results into the above formula, you get $\hat{\beta}_1^2 = 0.051 \times \frac{152{,}109.6}{1{,}499.6} = 5.20 \Leftrightarrow \hat{\beta}_1 = -2.28$ (luckily for you, your group member entered the negative sign in front of the slope).

b. Calculate the *t*-statistic for the slope and the intercept. Test the hypothesis that the intercept and the slope are different from zero.

Answer: The *t*-statistics are 73.82 and 4.75 respectively. Hence you can reject the two null hypothesis at any reasonable level of significance.

c. Should you be concerned that your group member only gave you the result for the homoskedasticity-only standard error formula, instead of using the heteroskedasticity-robust standard errors?

Answer: There is no theory that suggests the homoskedasticity in the error terms in this case. Given the serious consequences for using homoskedasticity only standard errors in the presence of heteroskedasticity, you should definitely use the heteroskedasticity robust standard errors for inference.

6. The neoclassical growth model predicts that for identical savings rates and population growth rates, countries should converge to the per capita income level. This is referred to as the convergence hypothesis. One way to test for the presence of convergence is to compare the growth rates over time to the initial starting level.

a. The results of the regression for 104 countries were as follows:

$$\hat{g}6090 = \underset{(0.004)}{0.019} - \underset{(0.0073)}{0.0006} \times RelProd_{60}, \quad R^2 = 0.00007,\ SER = 0.016$$

where $g6090$ is the average annual growth rate of GDP per worker for the 1960-1990 sample period, and $RelProd_{60}$ is GDP per worker relative to the United States in 1960. Numbers in parenthesis are heteroskedasticity robust standard errors.

Using the OLS estimator with homoskedasticity-only standard errors, the results changed as follows:

$$\hat{g}6090 = \underset{(0.002)}{0.019} - \underset{(0.0068)}{0.0006} \times RelProd_{60},\ R^2 = 0.00007,\ SER = 0.016$$

Why didn't the estimated coefficients change? Given that the standard error of the slope is now smaller, can you reject the null hypothesis of no beta convergence? Are the results in the second equation more reliable than the results in the first equation? Explain.

Answer: Using homoskedasticity-only standard errors has no effect on the OLS estimator. The *t*- statistic remains small and is certainly below the critical value. The results are less reliable since there is no reason to believe that the error variance is homoskedastic.

b. You decide to restrict yourself to the 24 OECD countries in the sample. This changes your regression output as follows (numbers in parenthesis are heteroskedasticity robust standard errors):

$$\hat{g}6090 = \underset{(0.004)}{0.048} - \underset{(0.0063)}{0.0404}\, RelProd_{60}\ ,\ R^2 = 0.82\ ,\ SER = 0.0046$$

Test for evidence of convergence now. If your conclusion is different than in (a), speculate why this is the case.

Answer: The *t*-statistic for the slope is 6.41. At face value, there is strong evidence for convergence. Neoclassical growth theory does not predict unconditional convergence. Instead it only predicts convergence if the savings rates and population growth rates are identical. It stands to reason that these are much more similar between OECD countries than between the countries of the world.

c. The authors of your textbook have informed you that unless you have more than 100 observations, it may not be plausible to assume that the distribution of your OLS estimators is normal. What are the implications here for testing the significance of your theory?

Answer: Since there are less than 30 observations, the distribution of the *t*-statistic is unknown. You should therefore not conduct statistical inference.

7. You have collected data from Major League Baseball (MLB) to find the determinants of winning. You have a general idea that both good pitching and strong

hitting are needed to do well. However, you do not know how much each of these contributes separately. To investigate this problem, you collect data for all MLB during 1999 season. Your strategy is to first regress the winning percentage on pitching quality ("Team ERA"), second to regress the same variable on some measure of hitting ("OPS – On-base Plus Slugging percentage"), and third to regress the winning percentage on both.

Summary of the Distribution of Winning Percentage, On Base plus Slugging percentage, and Team Earned Run Average for MLB in 1999

	Average	**Standard deviation**	**Percentile**						
			10%	**25%**	**40%**	**50% (median)**	**60%**	**75%**	**90%**
Team ERA	4.71	0.53	3.84	4.35	4.72	4.78	4.91	5.06	5.25
OPS	0.778	0.034	0.720	0.754	0.769	0.780	0.790	0.798	0.820
Winning Percentage	0.50	0.08	0.40	0.43	0.46	0.48	0.49	0.59	0.60

The results are as follows:

$$\widehat{Winpct} = 0.94 - 0.100 \times teamera, \quad R^2 = 0.49,\ SER = 0.06.$$

$$\widehat{Winpct} = -0.68 + 1.513 \times ops, \quad R^2 = 0.45,\ SER = 0.06.$$

$$\widehat{Winpct} = -0.19 - 0.099 \times teamera + 1.490 \times ops, \quad R^2 = 0.92,\ SER = 0.02.$$

(a) Interpret the multiple regression. What is the effect of a one point increase in team ERA? Given that the Atlanta Braves had the most wins that year, wining 103 games out of 162, do you find this effect important? Next analyze the importance and statistical significance for the OPS coefficient. (The Minnesota Twins had the minimum OPS of 0.712, while the Texas Rangers had the maximum with 0.840.) Since the intercept is negative, and since winning percentages must lie between zero and one, should you rerun the regression through the origin?

Answer: A single point increase in team ERA lowers the winning percentage by approximately 10 percent. A 0.1 increase in OPS results roughly in an increase of 15 percent. Given that there are no observations close to the origin, you should not interpret the intercept. The multiple regression explains 92 percent of the variation in winning percentage. The Atlanta Braves only won 63.6 percent of their games. Given that this represents the best record during that season, a 10 percentage point drop is important. Although the intercept cannot be interpreted, it anchors the regression at a certain level and should therefore not be omitted.

(b) What are some of the omitted variables in your analysis? Are they likely to affect the coefficient on *Team ERA* and *OPS* given the size of the R^2 and their potential correlation with the included variables?

Answer: The quality of the management and coaching comes to mind, although both may be reflected in the performance statistics, as are salaries. There are other aspects of baseball performance that are missing, such as the fielding percentage of the team.

8. Stock –Watson Chapter 6 Empirical Exercise E6.3.

(a)

Variable	Mean	Standard Deviation	Units
growth	1.86	1.82	Percentage Points
rgdp60	3131	2523	$1960
tradeshare	0.542	0.229	unit free
yearsschool	3.95	2.55	years
rev_coups	0.170	0.225	coups per year
assasinations	0.281	0.494	assasinations per year
oil	0	0	0–1 indicator variable

(b) Estimated Regression (in table format):

Regressor	Coefficient
tradeshare	1.34
	(0.88)
yearsschool	0.56**
	(0.13)
rev_coups	–2.15*
	(0.87)
assasinations	0.32
	(0.38)
rgdp60	–0.00046**
	(0.00012)
intercept	0.626
	(0.869)
SER	1.59
R^2	0.29
$\bar{R}^2$	0.23

The coefficient on *Rev_Coups* is −2.15. An additional coup in a five year period, reduces the average year growth rate by (2.15/5) = 0.43% over this 25 year period. This means the GPD in 1995 is expected to be approximately .43×25 = 10.75% lower. This is a large effect.

(c) The 95% confidence interval is $1.34 \pm 1.96 \times 0.88$ or -0.42 to 3.10. The coefficient is not statistically significant at the 5% level.

(d) The F-statistic is 8.18 which is larger than 1% critical value of 3.32.