# Chapter 5 Lab Answer Key

## Preparation

```r
require(knitr)
require(haven)
require(AER)
require(car)

opts_chunk$set(echo = TRUE)
options(digits = 3)
```

### (a) Estimate a model where GDP per capita (measured in 1000s of dollars *GDPpc1000*) explains fertility. What is the coefficient on *GDPpc1000*?

```r
load("Chapter5_Lab_Fertility.RData")

reg1 <- lm(fertility ~ GDPpc1000, data = dta)
summary (reg1)
```

```
##
## Call:
## lm(formula = fertility ~ GDPpc1000, data = dta)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.357 -1.559 -0.179  1.521  8.625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.49898    0.02207   203.8   <2e-16 ***
## GDPpc1000   -0.07278    0.00157   -46.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.78 on 8170 degrees of freedom
##   (4018 observations deleted due to missingness)
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.209
## F-statistic: 2.16e+03 on 1 and 8170 DF,  p-value: <2e-16
```

The estimated coefficient for GDPpc1000 is -0.073, meaning that an increase of one unit of GDPpc1000 will decrease fertility by 0.073 unit. The coefficient is statistically significant.

### (b) Add female life expectancy to the above model. What happens to the coefficient on on GDPpc1000? Explain in terms of omitted variable bias by calculating the expected value of the coefficient on GDP in the above model. Be careful that the sample size stays the same on your various models (for example, re-estimate a model with GDP as the only independent variable but exclude observations for which female life expectancy is missing, as those observations are omitted when you include female life expectancy in the model).

```r
reg2 <- lm(fertility ~ GDPpc1000 + female_lifeexp, data = dta)
summary (reg2)
```

```
##
## Call:
## lm(formula = fertility ~ GDPpc1000 + female_lifeexp, data = dta)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.780 -0.633 -0.090  0.567  4.116
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.48179    0.06878  196.00  < 2e-16 ***
## GDPpc1000      -0.00517    0.00102   -5.07  4.1e-07 ***
## female_lifeexp -0.14196    0.00107 -132.75  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 8147 degrees of freedom
##   (4040 observations deleted due to missingness)
## Multiple R-squared:  0.75,   Adjusted R-squared:  0.75
## F-statistic: 1.22e+04 on 2 and 8147 DF,  p-value: <2e-16
```

```r
reg.1b <- lm(fertility ~ GDPpc1000, data = dta[!is.na(dta$female_lifeexp),])
summary(reg.1b)
```

```
##
## Call:
## lm(formula = fertility ~ GDPpc1000, data = dta[!is.na(dta$female_lifeexp),
##     ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.359 -1.561 -0.179  1.522  8.622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.50092    0.02211   203.5   <2e-16 ***
## GDPpc1000   -0.07278    0.00157   -46.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.78 on 8148 degrees of freedom
##   (2372 observations deleted due to missingness)
## Multiple R-squared:  0.208,  Adjusted R-squared:  0.208
## F-statistic: 2.14e+03 on 1 and 8148 DF,  p-value: <2e-16
```

```r
reg_aux <- lm(female_lifeexp ~ GDPpc1000, data = dta[!is.na(dta$fertility),])
#summary(reg_aux)

expected_beta1 = reg2$coefficients[2] + reg2$coefficients[3] * reg_aux$coefficient[2]
expected_beta1
```

```
## GDPpc1000 
##   -0.0728
```

When variable female_lifeexp is included in the model, the coefficient on GDPpc1000 becomes -0.005, compared with -0.073 in the model where female_lifeexp is not included. Based on the multivariate model and the auxiliary regression, we obtain the expected value of beta_1 hat is -0.0728, which is consistent with the results from reg.1b (where female_lifeexp is excluded). This is omitted variable bias.

### (c) Add female labor participation to the above model and briefly discuss the results.

```
reg3 <- lm(fertility ~ GDPpc1000 + female_lifeexp + laborpart_female, data = dta)
summary (reg3)
```

```
## 
## Call:
## lm(formula = fertility ~ GDPpc1000 + female_lifeexp + laborpart_female, 
##     data = dta)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -3.906 -0.462 -0.012  0.418  2.786 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      14.674731   0.158802   92.41   <2e-16 ***
## GDPpc1000         0.009038   0.000955    9.47   <2e-16 ***
## female_lifeexp   -0.154559   0.002058  -75.10   <2e-16 ***
## laborpart_female -0.014505   0.000975  -14.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.722 on 2765 degrees of freedom
##   (9421 observations deleted due to missingness)
## Multiple R-squared:  0.725,  Adjusted R-squared:  0.725 
## F-statistic: 2.43e+03 on 3 and 2765 DF,  p-value: <2e-16
```

The coefficient on GDPpc1000 is 0.009. An increase of a unit in GDPpc1000 will increase fertility by 0.009 unit. The coefficient on female_lifeexp is -0.155. An one-unit increase in female_lifeexp will decrease fertility by 0.155 unit. The coefficient on laborpart_female is -0.015. An one-unit increase in laborpart_female will decrease fertility by 0.015 unit. All three coefficients are significant.

### (d) Is there multicollinearity in the model from the previous part? Provide evidence.

```
# To test for multicollinearity we look at VIF (variance inflation factor)
vif (reg3)
```

```
##        GDPpc1000   female_lifeexp laborpart_female 
##             1.51             1.42             1.08
```

VIFs are not very high, suggesting there are not severe multicollinearity problem in this model.

**(e) Create a variable called *female_lifeexp_noisy* that is *female_lifeexp* with some random error included; use a normally distributed random error with a standard deviation of 15. In other words, *female_lifeexp_noisy* does not equal the actual life expectancy for women, but is a "noisy" measure of the quantity. Estimate the model using the variable *female_lifeexp_noisy* and discuss any changes in coefficient on the female life expectancy variable. Relate the changes to theoretical expectations about measurement error discussed in Chapter 5.**

```
set.seed(1)
dta$female_lifeexp_noisy = dta$female_lifeexp + rnorm(length(dta$female_lifeexp), sd = 15)
```

```
reg4 <- lm(fertility ~ GDPpc1000 + female_lifeexp_noisy + laborpart_female, data = dta)
summary (reg4)
```

```
##
## Call:
## lm(formula = fertility ~ GDPpc1000 + female_lifeexp_noisy + laborpart_female,
##     data = dta)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.409 -0.807 -0.219  0.529  4.654
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.15346    0.13055   39.48  < 2e-16 ***
## GDPpc1000            -0.02367    0.00137  -17.30  < 2e-16 ***
## female_lifeexp_noisy -0.02513    0.00141  -17.82  < 2e-16 ***
## laborpart_female     -0.00847    0.00160   -5.28  1.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.19 on 2765 degrees of freedom
##   (9421 observations deleted due to missingness)
## Multiple R-squared:  0.25,   Adjusted R-squared:  0.249
## F-statistic:  308 on 3 and 2765 DF,  p-value: <2e-16
```

In the previous model (without noise), the estimated coefficient on female_lifeexp is -0.155. After adding noise, the coefficient on female_lifeexp_noisy becomes -0.25. This is becasue measurement errors causes attenuation bias which makes estimated coefficients closer to 0.

**(f) (Go back to non-noisy data.) Estimate a model with standardized coefficients. Which variable seems to have the largest effect?**

```
reg5 <- lm(scale(fertility) ~ scale(GDPpc1000) + scale(female_lifeexp) + scale(laborpart_female), data =
summary (reg5)
```

```
##
## Call:
## lm(formula = scale(fertility) ~ scale(GDPpc1000) + scale(female_lifeexp) +
##     scale(laborpart_female), data = dta)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.9206 -0.2272 -0.0057  0.2056  1.3699
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.11608    0.01047  -11.09   <2e-16 ***
## scale(GDPpc1000)         0.06359    0.00672    9.47   <2e-16 ***
## scale(female_lifeexp)   -0.92620    0.01233  -75.10   <2e-16 ***
## scale(laborpart_female) -0.10559    0.00709  -14.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.355 on 2765 degrees of freedom
##   (9421 observations deleted due to missingness)
## Multiple R-squared:  0.725,  Adjusted R-squared:  0.725
## F-statistic: 2.43e+03 on 3 and 2765 DF,  p-value: <2e-16
```

female_lifeexp seems to have the largest effect.

### (g) Test the null hypothesis that the effect of labor participation for women is the same as the female literacy rate.

```
reg6 <- lm(fertility ~ GDPpc1000 +female_lifeexp +laborpart_female +female_litrate, data = dta) # Unres

## Prepare data for the restricted model
dta$female_labor_plus_literacy = dta$laborpart_female + dta$female_litrate
reg6b <- lm(fertility ~ GDPpc1000 +female_lifeexp +female_labor_plus_literacy, data = dta)  # restricte

F.stat = ((summary(reg6)$r.squared - summary(reg6b)$r.squared)/1 )/((1-summary(reg6)$r.squared)/(summar
F.stat
```

```
## [1] 0.504
```

```
qf(1-0.05, df1=1, df2= summary(reg6)$df[2])
```

```
## [1] 3.86
```

The F-stat we obtained is 0.504 while the critical value is 3.86. Therefore, we fail to reject the null and that the effect of labor participation for women is the same as the female literacy rate.

### (h) Test the null hypothesis that both the effect of labor participation for women and the effect of female literacy rate are both zero.

```
## Unrestricted model is reg6
# Now we need to create a restricted model by forcing conditions
reg7 <- lm(fertility ~ GDPpc1000 +female_lifeexp, data = dta[(!is.na(dta$laborpart_female) & !is.na(dta

F.stat1 = ((summary(reg6)$r.squared - summary(reg7)$r.squared)/2 )/((1-summary(reg6)$r.squared)/(summar
F.stat1
```

```
## [1] 70.5
```

```
qf(1-0.05, df1=1, df2= summary(reg6)$df[2])
```

```
## [1] 3.86
```

Now the F-stat is 70.5 which is way above the critical value 3.86. Therefore, we reject the null that both the effect of labor participation for women and the effect of female literacy rate are both zero.

**Bonus: Add male labor participation and male life expectancy to your model. Discuss changes in the model and results.**

```
reg8 <- lm(formula = fertility ~ GDPpc1000 + female_lifeexp + laborpart_female + female_litrate + labor
summary (reg8)
```

```
##
## Call:
## lm(formula = fertility ~ GDPpc1000 + female_lifeexp + laborpart_female +
##     female_litrate + laborpart_male + male_lifeexp, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3277 -0.3327  0.0107  0.3373  1.7721
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.03248    0.66509   16.59  < 2e-16 ***
## GDPpc1000        -0.00544    0.00703   -0.77  0.44004
## female_lifeexp   -0.08371    0.02323   -3.60  0.00042 ***
## laborpart_female -0.00669    0.00341   -1.96  0.05156 .
## female_litrate   -0.04212    0.00379  -11.10  < 2e-16 ***
## laborpart_male    0.00449    0.00382    1.18  0.24168
## male_lifeexp      0.02056    0.02413    0.85  0.39552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.63 on 155 degrees of freedom
##   (12028 observations deleted due to missingness)
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.815
## F-statistic:  119 on 6 and 155 DF,  p-value: <2e-16
```

```
vif (reg8)
```

```
##        GDPpc1000   female_lifeexp laborpart_female   female_litrate
##             1.38            13.47             1.32             2.46
##   laborpart_male     male_lifeexp
##             1.03            12.23
```

Now, we test for multicolinearity and noticed that female_lifeexp and male_lifeexp have a high VIF therefore high multicolinearity. That's also why the standard error of these two variables are significantly higher, because multicolinearity yields higher variances.