Chapter 7 Lab Answer Key

Tianwei Liu

Preparation

```
require(knitr)
require(haven)
require(car)
require(AER)
require(lm.beta)  ## install.packages("lm.beta")

opts_chunk$set(echo = TRUE)
options(digits = 6)
```

(a) Estimate a model for women predicting wages in 1996 as a function of height in 1985 and 1981, siblings and esteem from 1980. Use standardized coefficients and report results. How do the t statistics compare to t statistics in an unstandardized model? (You don't need to report the unstandardized results.)

```
reg1 <- lm(scale(wage96) ~ scale(height85) + scale(height81) + scale(siblings) + scale(esteem80), data =
summary(reg1)
```

```
##
## Call:
## lm(formula = scale(wage96) ~ scale(height85) + scale(height81) +
##     scale(siblings) + scale(esteem80), data = dta[dta$male ==
##     0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -0.55  -0.21  -0.10   0.04  43.39
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.00174    0.01835   -0.09   0.9246
## scale(height85)  0.02811    0.03978    0.71   0.4799
## scale(height81) -0.00947    0.04034   -0.23   0.8144
## scale(siblings) -0.05066    0.01901   -2.66   0.0077 **
## scale(esteem80)  0.05817    0.01914    3.04   0.0024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 3097 degrees of freedom
##   (3181 observations deleted due to missingness)
## Multiple R-squared:  0.00668,    Adjusted R-squared:  0.0054
## F-statistic: 5.21 on 4 and 3097 DF,  p-value: 0.000353
```

- Coefficients on height85 and height81 are not statistically significant at conventional level alpha = 95% as the p-values associated with the two coefficients are greater than 0.05. Therefore, we fail to reject the null that heights have an effect on wages.

- Coefficients on siblings is -0.051, and the coefficient is statistically significant because the p-value associated with it is 0.0077, smaller than level of significance 0.01. This means that an increase of a standard deviation in number of siblings is associated with 0.051 standard deviation decrease in wages.
- Similarly, coefficients in siblings is 0.058, and the coefficient is statistically significant as the p-value associated with it is 0.0024, smaller than 0.01 level of significance. This means that an increase of a std in self-esteem is associated with an increase of 0.058 std increase in wages.

```
reg1_unstd <- lm(wage96 ~ height85 + height81 + siblings + esteem80, data = dta[dta$male == 0,])
summary(reg1_unstd)
```

```
##
## Call:
## lm(formula = wage96 ~ height85 + height81 + siblings + esteem80,
##     data = dta[dta$male == 0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -19.3   -7.3   -3.5    1.3 1516.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -15.766     16.374   -0.96   0.3357
## height85       0.354      0.500    0.71   0.4799
## height81      -0.121      0.516   -0.23   0.8144
## siblings      -0.662      0.249   -2.66   0.0077 **
## esteem80       0.711      0.234    3.04   0.0024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.6 on 3097 degrees of freedom
##   (3181 observations deleted due to missingness)
## Multiple R-squared:  0.00668,    Adjusted R-squared:  0.0054
## F-statistic: 5.21 on 4 and 3097 DF,  p-value: 0.000353
```

The two models differ only in the estimates coefficients and standard errors, becasue the variables are standardized in one but not in the other. t-stats and p-values associated with the coefficients are the same across two models.

(b) Add several covariates (your choice) to the above model, including dummy variables for race/ethnicity. Use standardized coefficients as appropriate to each variable and briefly discuss effects, focusing on effect of race relative to the esteem variable (as an example of a continuous covariate).

```
# Estimate regression models using scale command for continuous variables
dta$esteemblack = dta$esteem80 * dta$black
reg2 <- lm(scale(wage96) ~ scale(height85) + scale(height81) + scale(siblings) + scale(esteem80) + black
summary(reg2)
```

```
##
## Call:
## lm(formula = scale(wage96) ~ scale(height85) + scale(height81) +
##     scale(siblings) + scale(esteem80) + black + esteemblack,
```

```
##     data = dta[dta$male == 0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -0.57  -0.21  -0.10   0.04  43.43
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.00801    0.02208    0.36   0.7168
## scale(height85)  0.02841    0.03979    0.71   0.4753
## scale(height81) -0.00999    0.04035   -0.25   0.8045
## scale(siblings) -0.04745    0.01946   -2.44   0.0148 *
## scale(esteem80)  0.06716    0.02282    2.94   0.0033 **
## black            0.19620    0.34439    0.57   0.5689
## esteemblack     -0.00971    0.01445   -0.67   0.5016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 3095 degrees of freedom
##   (3181 observations deleted due to missingness)
## Multiple R-squared:  0.00704,    Adjusted R-squared:  0.00511
## F-statistic: 3.66 on 6 and 3095 DF,  p-value: 0.00128
```

Compared with the model in the previous question, this model includes black as a dummy independent variable and an interaction term between self-esteem and black. Since coefficients on black and on esteemblack are not statistically significant, there is not a significant of being black on wages nor the differential effect of self-esteem for blacks.

(c) Models with wages often have logged variables in order to be able to provide results in percentage terms rather than absolute dollars. Estimate a log-linear model for women. To keep things simple, use only *height85*, *height81*, *esteem80*, *black* and *siblings* as the covariates.

```
dta$wage96.NoNA = dta$wage96
dta$wage96.NoNA[dta$wage96==0] = NA ## Clean for wages == 0 because logs do not work with non-positive
reg3 <- lm(log(wage96.NoNA) ~ height85 + height81 + siblings + esteem80 + black, data = dta[dta$male ==
summary(reg3)
```

```
##
## Call:
## lm(formula = log(wage96.NoNA) ~ height85 + height81 + siblings +
##     esteem80 + black, data = dta[dta$male == 0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.004 -0.380  0.042  0.443  5.051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.12558    0.36314   -0.35     0.73
## height85     0.01376    0.01125    1.22     0.22
## height81     0.00688    0.01160    0.59     0.55
## siblings    -0.03121    0.00562   -5.56  3.0e-08 ***
```

```
## esteem80     0.05108    0.00517    9.89  < 2e-16 ***
## black       -0.16770    0.03169   -5.29  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.777 on 3014 degrees of freedom
##   (3263 observations deleted due to missingness)
## Multiple R-squared:  0.0664, Adjusted R-squared:  0.0648
## F-statistic: 42.9 on 5 and 3014 DF,  p-value: <2e-16
```

- Coefficients on two height variables are not statistically significant, so there appears to be no effect of height on wages.
- The coefficient on siblings is -0.0312, meaning that one more siblings is associated with 3.12% decrease in wages, and this coefficient is statistically significant because the p-value is smaller than 0.001 level of significance.
- The coefficient on self-esteem is 0.051 which is significant as the p-value associated with it is smaller then 0.001 level of significance, meaning that an unit increase in self-esteem is associated with 5.1% increase in wages.
- The coefficient on black is -0.168. This coefficient is statistically significant as the p-value associated with it is smaller than 0.001 level of significance. This coefficient suggests that black people, holding all else equal, earn 16.8% less than non-blacks.

(d) Starting with the above model (for women only), create a model in which the effect of siblings is potentially non-linear via a quadratic equation. Discuss the results and note the effect of siblings in general term and for specific cases when *siblings* equals 1 and when *siblings* equals 5. (For fun, estimate the same model for men. You don't need to report or discuss those results.)

```
reg4 <- lm(log(wage96.NoNA) ~ height85 + height81 + esteem80 + black + siblings + I(siblings^2), data =
summary(reg4)
```

```
##
## Call:
## lm(formula = log(wage96.NoNA) ~ height85 + height81 + esteem80 +
##     black + siblings + I(siblings^2), data = dta[dta$male ==
##     0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.975 -0.384  0.046  0.446  4.987
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.06965    0.36354   -0.19    0.848
## height85       0.01348    0.01124    1.20    0.230
## height81       0.00728    0.01159    0.63    0.530
## esteem80       0.05105    0.00516    9.89  < 2e-16 ***
## black         -0.16698    0.03167   -5.27  1.4e-07 ***
## siblings      -0.06512    0.01484   -4.39  1.2e-05 ***
## I(siblings^2)  0.00307    0.00124    2.47    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.776 on 3013 degrees of freedom
##   (3263 observations deleted due to missingness)
## Multiple R-squared:  0.0683, Adjusted R-squared:  0.0664
## F-statistic: 36.8 on 6 and 3013 DF,  p-value: <2e-16
```

- In general terms, the effect of siblings is equal to -0.065 * siblings + 0.0031 * siblingsˆ2
- When siblings == 1, the effect = -0.0619, meaning that having one sibling will likely decrease wage by 6.5%.
- When siblings == 5, the effect = -0.2475, meaning that having five siblings will likely decrease wage by 24.75%.

```
reg4_male <- lm(log(wage96.NoNA) ~ height85 + height81 + esteem80 + black + siblings + I(siblingsˆ2), da
summary(reg4_male)
```

```
##
## Call:
## lm(formula = log(wage96.NoNA) ~ height85 + height81 + esteem80 +
##     black + siblings + I(siblingsˆ2), data = dta[dta$male ==
##     1, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.876 -0.343  0.020  0.394  3.873
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.045077   0.311169   -0.14   0.8848
## height85       0.002286   0.008350    0.27   0.7843
## height81       0.022629   0.008093    2.80   0.0052 **
## esteem80       0.041024   0.004429    9.26   <2e-16 ***
## black         -0.312412   0.027960  -11.17   <2e-16 ***
## siblings      -0.033545   0.012930   -2.59   0.0095 **
## I(siblingsˆ2)  0.000881   0.001099    0.80   0.4224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.696 on 3265 degrees of freedom
##   (3131 observations deleted due to missingness)
## Multiple R-squared:  0.0949, Adjusted R-squared:  0.0932
## F-statistic: 57.1 on 6 and 3265 DF,  p-value: <2e-16
```

(e) Estimate a model for women only in which the dependent variable is log of wages and the independent variables are *momed79*, *daded79*, *height85*, *height81*, *black* and *hispanic*. Test the null hypothesis that the effect of mother's education is the same as the effect of father's education. Report unrestricted and unrestricted models and then show the calculation of the F-statistic and explain the results.

```
## Unrestricted model
reg5_unres <- lm(log(wage96.NoNA)~momed79 + daded79 + height85 + height81 + black + hispanic, data = dta
summary(reg5_unres)
```

```
##
```

```
## Call:
## lm(formula = log(wage96.NoNA) ~ momed79 + daded79 + height85 +
##     height81 + black + hispanic, data = dta[dta$male == 0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.968 -0.382  0.028  0.426  5.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52559    0.39155    1.34    0.180
## momed79      0.02755    0.00672    4.10  4.2e-05 ***
## daded79      0.03273    0.00523    6.26  4.6e-10 ***
## height85     0.01490    0.01264    1.18    0.238
## height81     0.00167    0.01310    0.13    0.898
## black       -0.08170    0.03718   -2.20    0.028 *
## hispanic     0.20800    0.04780    4.35  1.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.774 on 2577 degrees of freedom
##   (3699 observations deleted due to missingness)
## Multiple R-squared:  0.0615, Adjusted R-squared:  0.0593
## F-statistic: 28.1 on 6 and 2577 DF,  p-value: <2e-16
```

- Both of our variable of interest are statistically significant as the p-values are below 0.001 level of significance. Therefore, we reject the null and claim that mom's education level and dat's education level have effects on wages. An one-unit increase in mother's education level is associated with 2.8% increase in wages; an one-unit increase in father's education level is associated with roughly 3.3% increase in wages.

```
## For the restricted model, our hypothesis is that mother's education has the same effect as father's
dta$parentsedu <- dta$momed79 + dta$daded79
reg5_res <- lm(log(wage96.NoNA)~ parentsedu + height85 + height81 + black + hispanic, data = dta[dta$ma
summary(reg5_res)
```

```
##
## Call:
## lm(formula = log(wage96.NoNA) ~ parentsedu + height85 + height81 +
##     black + hispanic, data = dta[dta$male == 0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.974 -0.384  0.030  0.425  5.125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52648    0.39149    1.34    0.179
## parentsedu   0.03054    0.00269   11.34  < 2e-16 ***
## height85     0.01477    0.01263    1.17    0.242
## height81     0.00165    0.01310    0.13    0.900
## black       -0.08312    0.03706   -2.24    0.025 *
## hispanic     0.21129    0.04731    4.47  8.3e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.774 on 2578 degrees of freedom
##   (3699 observations deleted due to missingness)
## Multiple R-squared:  0.0614, Adjusted R-squared:  0.0596
## F-statistic: 33.7 on 5 and 2578 DF,  p-value: <2e-16
```

- In the restricted model, parentsedu is statistically significant with a very small p-value, well below the 0.001 threshold. An one-unit increase in parents education level is associated with a 3.1% increase in wages.

```
#calculate F stat
F.stat.top = ((summary(reg5_unres)$r.squared - summary(reg5_res)$r.squared)/1)
F.stat.bottom = ((1-summary(reg5_unres)$r.squared)/(summary(reg5_unres)$df[2]))
F.stat = F.stat.top/F.stat.bottom
F.stat
```

```
## [1] 0.237572
```

```
qf(1-0.05, df1=1, df2= summary(reg5_unres)$df[2])
```

```
## [1] 3.84507
```

Since the F-statistic is smaller than the critical value, we fail to reject the null that the effect of mother's education is the same as the effect of father's education on wages. Therefore, these two effects are different.

BONUS: Based on the previous model, test the null hypothesis that both **height85* and *height81* equal zero. Report unrestricted and unrestricted models and then show the calculation of the F-statistic and explain the results.

```
reg6_res <- lm(log(wage96.NoNA)~momed79 + daded79 + black + hispanic, data = dta[dta$male ==0,])
F.stat2.top = ((summary(reg5_unres)$r.squared - summary(reg6_res)$r.squared)/2)
F.stat2.bottom = ((1-summary(reg5_unres)$r.squared)/(summary(reg5_unres)$df[2]))
F.stat2 = F.stat2.top/F.stat2.bottom
F.stat2
```

```
## [1] 8.03429
```

```
qf(1-0.05, df1=2, df2= summary(reg5_unres)$df[2])
```

```
## [1] 2.99922
```

As the F-stat we calcualted is larger than the critical value, we reject the null that both height85 and height81 are equal to zero. We claim that at least one of these two coefficients are non-zero.