# IIA-3 Econometrics: Supervision 3

Emre Usenmez

2024-12-08

{Very grateful to Dr Oleg Kitov and Dr Clive Lawson for the very informative stylized answers to previous iterations of the supervision questions.}

## SUPPLEMENTARY QUESTIONS

### QUESTION 1

**In a study of the Cobb-Douglass production function, a researcher suspects that the parameters are subject to change over time. Data on output, $Y$, labor input, $X_1$, and capital stock, $X_2$, are available for years 1929 to 1967. $T$ represents the time trend. The results obtained are as follows (t-values in parantheses):**

$$\begin{aligned}
\text{Full Sample:} \quad \widehat{logY} &= -3.02 + 1.34\ logX_1 + 0.29\ logX_2 + 0.0052\ T \\
&\quad (-6.65) \quad (14.68) \quad (4.89) \quad (2.34) \\
&\quad R^2 = 0.99535 \quad \hat{\sigma} = 0.03274 \\
\text{1929-48:} \quad \widehat{logY} &= -3.22 + 1.36\ logX_1 + 0.32\ logX_2 + 0.0051\ T \\
&\quad (-4.63) \quad (4.95) \quad (1.36) \quad (1.40) \\
&\quad R^2 = 0.97853 \quad \hat{\sigma} = 0.04449 \\
\text{1949-67:} \quad \widehat{logY} &= -1.56 + 1.02\ logX_1 + 0.33\ logX_2 + 0.0095\ T \\
&\quad (-2.21) \quad (7.58) \quad (2.33) \quad (1.85) \\
&\quad R^2 = 0.99565 \quad \hat{\sigma} = 0.0135
\end{aligned}$$

**a) Conduct a test of the hypothesis that the four regression coefficients are jointly the same in both sub-periods, against the alternative that they differ.**

**Answer:** This question is effectively testing if there is a structural break from the start of 1949 which may be caused by different intercept, different slope coefficient, or both. Suppose the coefficients of the full sample regression are $\beta$s, and coefficients of the 1929-48 regression are $\psi$s, and coefficients of the 1949-67 are $\gamma$s. Our hypothesis is therefore:

$$\mathbb{H}_0 : (\psi_0 = \gamma_0) \cap (\psi_1 = \gamma_1) \cap (\psi_2 = \gamma_2) \cap (\psi_3 = \gamma_3)$$

We can test this in two ways. First is to reparameterize the model and then run an $F$-test, and the second is to run a Chow test.

First Approach:

We can create a new coefficient $\delta = \psi - \gamma$ whereby our model becomes: $\widehat{logY} = \delta_0 + \delta_1\ logX_1 + \delta_2\ logX_2 + \delta_3 T + \varepsilon$ for which the hypothesis becomes:

$$\mathbb{H}_0 : (\delta_0 = 0) \cap (\delta_1 = 0) \cap (\delta_2 = 0) \cap (\delta_3 = 0)$$

We would then use the $F$-test for this join hypothesis.

Second Approach:

An alternative appraoch is to use *Chow Test.*[1] This test assumes that:

- The error terms in the subperiod regressions are normally distributed with the same, i.e. homoskedastic, variance $\sigma^2$. That is, $u_{1929-48t} \sim N(0, \sigma^2)$ and $u_{1949-67t} \sim N(0, \sigma^2)$.
- The two error terms $u_{1929-48t}$ and $u_{1949-67t}$ are independently distributed.

The Chow test is an $F$-ratio which means we will need the $RSS$ of both unrestricted and restricted models. Here, *the full sample model is the restricted model* since that is the model we have by imposing the restrictions that all $\psi_j = \gamma_j$ for $j = 0, \ldots, 3$. The $RSS$ of the unrestricted model, on the other hand - and this is a key insight - is the combination of the two sub-sample $RSS$s.

**Why $RSS$ and not $R^2$ form of $F$-Test**

Note that the Chow test uses $RSS$ and that there is no simple $R^2$ form of the $F$-test if separate regressions have been estimated for each group. This is because the $TSS$s are not the same as $\bar{Y}$ is not the same in both samples.

The steps to carry out a Chow test is as follows:

1. Obtain the restricted model's residual sum of squares, $RSS_R$, by estimating the regression for the full sample model with $(n - k - 1)$ degrees of freedom, where $n = n_1 + n_2$ with $n_1$ being the sample size of the first sub-sample, and $n_2$ being he sample size of the second sub-sample, and where $k$ is the number of regressors in that model.
2. Estimate the first sub-sample model to obtain its residual sum of squares, $RSS_1$ with $n_1 - k - 1$ degrees of freedom.
3. Do the same for the second sub-sample model to obtain $RSS_2$ with $n_2 - k - 1$ degrees of freedom.
4. Add the two $RSS$s to compute the unrestricted model's residual sum of squares: $RSS_{UR} = RSS_1 + RSS_2$.
5. Compute the $F$-ratio:

$$F = \frac{\dfrac{RSS_R - RSS_{UR}}{k+1}}{\dfrac{RSS_{UR}}{n - 2(k+1)}} = \frac{\dfrac{\text{improvement in fit}}{\text{extra degrees of freedom used up}}}{\dfrac{\text{residual sum of squares remaining}}{\text{degrees of freedom remaining}}}$$

   $\hookrightarrow$ Because we are splitting the sample into two, we are estimating k regressors for each sample plus their intercepts, so we are using up $(k + 1)$ degrees of freedom twice.
6. Compare the $F$-ratio to the critical $F$ value with $\big((k+1), n - 2(k+1)\big)$ degrees of freedom and fail to reject the null hypothesis of *parameter stability*, i.e. no structural change, if $F$-ratio does not exceed the critical value at the chosen significance level.

[1] Chow, C Gregory (1960) *Tests of Equality Between Sets of Coefficients in Two Linear Regressions*, Econometrica, 28(3) 591:605

Accordingly, we first need to calculate the $RSS$s. For that, recall that $RSS = \hat{\sigma}^2(n-k-1)$ where $n = 39, n_1 = 20, n_2 = 19$ because the dates are inclusive. Therefore:

$$\begin{aligned} RSS_R &= 0.03274^2 \times 35 && = 0.03751677 \\ RSS_{UR} &= RSS_1 + RSS_2 \\ &= 0.04449^2 \times 16 + 0.0135^2 \times 15 \\ &= 0.03440351 \end{aligned}$$

With these we can now calculate our $F$-ratio:

$$F = \frac{\frac{0.03751677 - 0.04304995}{4}}{\frac{RSS_{UR}}{n-2(k+1)}} = \frac{\frac{0.03751677 - 0.03440351}{4}}{\frac{0.03440351}{31}} = 0.7013157$$

The F-statistic for $\alpha = 0.05$ is 2.678667 and for $\alpha = 0.01$ is 3.992811, and thus we fail to reject the null hypothesis of parameter stability at either of the $\psi$ values.

```
qf(p=c(0.05, 0.01), df1=4, df2=31, lower.tail = FALSE)
```

```
## [1] 2.678667 3.992811
```

**Why $RSS$ equals $\hat{\sigma}^2(n-k-1)$?**

Consider how we estimate the error variance, $\sigma^2$. First notice that $\sigma^2 = \mathbb{E}(u^2)$, so an unbiased estimator of $\sigma^2$ is $\frac{1}{n}\sum_{i=1}^n u_i^2$. However, since we do not observe the errors $u_i$ this is not a true estimator. What we have, though, is the estimates of the errors $u_i$ which are the OLS residuals $\hat{u}_i$. If we replace the errors with the OLS residuals then we have

$$\sigma^2 = \frac{\sum_{i=1}^n u_i^2}{n} = \frac{RSS}{n}$$

which is a true estimator because it gives a computable rule for any sample of data on $X$s and $Y$. However, this is biased because it does not account for the restrictions that must be satisfied by the OLS residuals. These restrictions are given by the two OLS first order conditions:

$$\sum_{i=1}^n \hat{u}_i = 0, \qquad \sum_{i=1}^n X_i \hat{u}_i = 0$$

for a simple regression with one regressor. In a way, if we know $n-k-1$ residuals, we can always get the other remaining residuals by using the restrictions implied by the first order conditions. Therefore there are only $n-k-1$ degrees of freedom in the OLS residuals, as opposed to $n$ degrees of freedom in the errors.
The unbiased estimator of the error variance is therefore:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1} = \frac{RSS}{n-k-1}$$

where in simple regression with one regressor, $k = 1$.

**(b) It is believed that irrespective of whether the form of the relationship has changed over the two periods (i.e. whether the coefficients of the equation have changed), there has still been a structural break. Test this hypothesis and use this result to comment on the assumptions made in part (a). What limitations are there to these methods for testing for stability over the whole period.**

**Answer:** What this question is effectively asking is that whether the error variances of the two subperiod regressions are the same. Recall from part (a) one of the two key assumptions of the Chow test is that the variances of the errors are homoskedastic. Therefore the hypothesis is:

$$\mathbb{H}_0 : \sigma_\psi^2 = \sigma_\gamma^2 \qquad \text{or} \qquad \mathbb{H}_0 : \frac{\sigma_\psi^2}{\sigma_\gamma^2} = 1$$

$$\mathbb{H}_1 : \sigma_\psi^2 \neq \sigma_\gamma^2 \qquad \text{or} \qquad \mathbb{H}_0 : \frac{\sigma_\psi^2}{\sigma_\gamma^2} \neq 1$$

If the variances of the two subpopulations are the same, i.e. $\sigma_\psi^2 = \sigma_\gamma^2$, as assumed by the Chow test, then the ratios of the ratios of estimated error variance to population error variance has an $F$ distribution with $(n_1 - k - 1)$ and $(n_2 - k - 1)$ degrees of freedom in the numerator and denominator, respectively. That is,

$$\frac{\frac{\hat{\sigma}_\psi^2}{\sigma_\psi^2}}{\frac{\hat{\sigma}_\gamma^2}{\sigma_\gamma^2}} \sim F_{(n_1-k-1),(n_2-k-1)}.$$

Notice that if $\sigma_\psi^2 = \sigma_\gamma^2$ then this ratio and thus the $F$-test becomes:

$$F = \frac{\hat{\sigma}_\psi^2}{\hat{\sigma}_\gamma^2}$$

where by convention the larger estimated variance is in the numerator.

Accordingly, our $F$-statistic is:

$$F = \frac{0.04449^2}{0.0135^2} = 10.86069.$$

Since this is a two-tailed test, but we are putting the higher variance in the numerator, then we can treat it as one-sided test with alternative hypothesis is greater than 1, The critical values for $\alpha = 0.05$ with $(16, 15)$ degrees of freedom is 2.384875, and for $\alpha = 0.01$ it is 3.485246.

Thus at both $\alpha$ levels we can reject the null hypothesis and conclude that the subperiod variances are not the same at $\alpha = 0.01$. This means, the assumption of Chow test does not hold and we shouldn't use the Chow test, at least not in this form. There are modifications to Chow test that can be utilized but that is beyond this class.

Another point regarding the Chow test to bear in mind is that it is sensitive to the choice of the time at which we divide the subperiods. The $F$ values would be different if the cut-off point was 1947 or 1949.

Finally, the Chow test will tell us only if the two regressions are different but not whether the difference is due to the intercepts, the slopes, or both. We can use dummy variables for that, though.

```
qf(c(0.05, 0.01), df1 = 16, df2 = 15, lower.tail = FALSE)
```

```
## [1] 2.384875 3.485246
```

## QUESTION 2

**The following demand for money function was estimated from 60 observations, for which $\sum(M-\bar{M})^2 = 45600$:**

$$\hat{M}_t = 284 + 0.56Y_t - 0.43M_{t-1} \qquad R^2 = 0.841$$

**When a further 8 observations became available, the equation was re-estimated. The pooled data had $\sum(M-\bar{M})^2 = 50100$, and the re-estimated equation $R^2 = 0.818$. Carry out a Chow test for predictive failure. What do you conclude from your results? Explain carefully the role of the dummy variable in this test.**

**Answer:** The approach to the Chow test is similar to Question 1 whereby the $RSS$ for restricted model is the pooled data. However, notice we are not running a two separate regressions on the subsamples. So we follow the first approach from Question 1(a) by taking the differences of the model:

$$\mathbb{H}_0 : (\delta_0 = 0) \cap (\delta_1 = 0) \cap (\delta_2 = 0)$$

where $\delta$s are the coefficients of model $D_t$ which is the difference between the initial model and the model after the new observations.

Accordingly, we first need to derive the respective $RSS$s using the identity $R^2 = 1 - \frac{RSS}{TSS}$ or $RSS = (1-R^2)TSS$ where $TSS = \sum(M-\bar{M})^2$:

$$RSS_R = (1 - 0.841) \times 45,600 = 7,250.4$$
$$RSS_{UR} = (1 - 0.818) \times 50,100 = 9,118.2$$

We can then calculate the Chow test whereby

$$F = \frac{\frac{9118.2 - 7250.4}{8}}{\frac{7250.4}{57}} = 1.835495$$

The $F$ critical values for $\alpha = 0.05$ and for $\alpha = 0.01$ with $(8, 57)$ degrees of freedom are 2.105599 and 2.840694 respectively. Accordingly we cannot reject the null hypothesis and conclude that there is no predictive failure.

```
qf(c(0.05, 0.01), df1 = 8, df2 = 57, lower.tail = FALSE)
```

```
## [1] 2.105599 2.840694
```

## QUESTION 3

**(a) Use the dataset sup3.xls to import the following variables for the period 1955-1990:**

**W = Wages and Salaries (£ million), CSO code: CFAJ_AU**

**P = Implied Deflator for Consumers Expenditure, CSO code: GIEF_AU**

**E = Employees in Employment, CSO code: BCAD_AU**

**WF = Workforce, CSO code: DYDB_AU**

**U = Unemployed, CSO code: BCAB_AU**

Load the libraries:

```
libraries <- c("haven",       # to import/export SPSS, STATA, SAS files
               "readxl",      # to import/export Excel files
               "tidyverse",   # for tidy data
               "Statamarkdown", # for using STATA commands in R
               "kableExtra", # for creating nice tables in R
               "rstatix")     # converts stats functions to a tidyverse-friendly format

# lapply(libraries, library, character.only=TRUE) will load the libraries
```

Load the data:

```
salaries_df <- read_excel("../Data/sup3.xls", sheet = 1)
```

Briefly examine the data frame

```
# Yuo can use any of the following to examine data frame (df):
# `dim()`: for its dimensions, by row and column
# `str()`: for its structure
# `summary()`: for summary statistics on its columns
# `colnames()`: for the name of each column
# `head()`: for the first 6 rows of the data frame
# `tail()`: for the last 6 rows of the data frame
# `View()`: for a spreadsheet-like display of the entire data frame

summary(salaries_df)
```

```
      Year         CFAJ_AU          GIEF_AU          BCAD_AU
 Min.   :1955   Min.   : 10210   Min.   : 12.20   Min.   :21067
 1st Qu.:1964   1st Qu.: 17430   1st Qu.: 15.55   1st Qu.:21890
 Median :1972   Median : 35890   Median : 24.75   Median :22466
 Mean   :1972   Mean   : 76269   Mean   : 47.06   Mean   :22285
 3rd Qu.:1981   3rd Qu.:127400   3rd Qu.: 81.03   3rd Qu.:22740
 Max.   :1990   Max.   :271394   Max.   :127.90   Max.   :23257
    DYDB_AU         BCAB_AU
 Min.   :24180   Min.   : 210.0
 1st Qu.:25211   1st Qu.: 406.0
 Median :25634   Median : 626.5
 Mean   :26005   Mean   :1175.9
 3rd Qu.:26692   3rd Qu.:1602.8
 Max.   :28437   Max.   :3229.0
```

We can change the column names to something more memorable:

```
colnames(salaries_df) <- c("Year", "W", "P", "E", "WF", "U")
```

---

**(b) Use the data set to estimate the following wage equation by the OLS method:**

$$\Delta lnW_t = \beta_0 + \beta_1 lnP + \beta_2 lnP_{t-1} + \beta_3 lnE_t + \beta_4 lnE_{t-1} + \beta_5 lnUR_t + \beta_6 lnUR_{t-1} + \varepsilon_t$$

**where $\beta_i$ are constants, $\Delta$ denotes the first difference operator and $UR_t = \frac{U_t}{WF_t}$.**

**Verify that your estimate of $\beta_1$ is .91170. Also show that the above is a generalized version of the following hypothesis:**

$$\Delta ln\Big(\frac{W}{EP}\Big)_t = \beta_0 + \varepsilon.$$

**How would you explain the inclusion of the $UR$ terms?**

**Answer:**

```
#Transform the data to get it ready for the regression
salaries_df<- salaries_df %>%
  mutate(lnW = log(W),
         DlnW = lnW - lag(lnW,1),
         lnP = log(P),
         lag_lnP = lag(lnP,1),
         lnE = log(E),
         lag_lnE = lag(lnE,1),
         lnUR = log(U/WF),
         lag_lnUR = lag(lnUR,1))

#Let's put these in a new dataframe to keep them neat:
salaries_new_df <- data.frame(Year = salaries_df$Year,
         lnW = salaries_df$lnW,
         DlnW = salaries_df$DlnW,
         lnP = salaries_df$lnP,
         lag_lnP = salaries_df$lag_lnP,
         lnE = salaries_df$lnE,
         lag_lnE= salaries_df$lag_lnE,
         lnUR = salaries_df$lnUR,
         lag_lnUR = salaries_df$lag_lnUR,
         stringsAsFactors = FALSE)
```

Since we are taking the differences those columns should have any value for 1955:

```
#Look at total number of `NA` values per column:
colSums(is.na(salaries_new_df))
```

```
    Year      lnW     DlnW      lnP  lag_lnP      lnE  lag_lnE     lnUR
       0        0        1        0        1        0        1        0
lag_lnUR
       1
```

Thus we can adjust our sample size to take account of this:

```
n_unr <- nrow(salaries_new_df) - sum(apply(is.na(salaries_df), 1, any))
# Note, if we replace 1 with 2 within apply function it will give us 4, the total number of NAs.
n_unr
```

```
[1] 35
```

Thus our sample size is 35.

```
# Run the regression
lm_Q3b <- lm(DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR, data = salaries_new_df)
summary(lm_Q3b)
```

```
Call:
lm(formula = DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR,
    data = salaries_new_df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052235 -0.010767  0.001441  0.007637  0.046109

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.87957    1.96811   0.447   0.6584
lnP          0.91170    0.10094   9.032 8.67e-10 ***
lag_lnP     -0.89866    0.10072  -8.923 1.12e-09 ***
lnE          0.62385    0.38833   1.607   0.1194
lag_lnE     -0.71830    0.35308  -2.034   0.0515 .
lnUR        -0.03547    0.02731  -1.299   0.2046
lag_lnUR     0.01994    0.02391   0.834   0.4114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02058 on 28 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8324,    Adjusted R-squared:  0.7965
F-statistic: 23.18 on 6 and 28 DF,  p-value: 1.175e-09
```

We can therefore see that only lnP and lag_lnP coefficients are significant. However, $F$-stat is pretty high which suggests presence of multicollinearity. We can look at the variance inflation factor, VIF:

```
car::vif(lm_Q3b)
```

```
       lnP    lag_lnP        lnE    lag_lnE       lnUR   lag_lnUR
560.948084 539.661549   8.784345   7.269653  40.557237  32.997711
```

where we see each of the $VIF$ value is greater than 5 suggesting serious collinearity.

The question is also asking for us to show that the model we have is a generalized version of the hypothesis that the change in average real wage is independent of all other variables:

$$\begin{aligned}
\Delta ln\left(\frac{W}{EP}\right)_t &= \beta_0 + \varepsilon \\
\Delta ln(W)_t - \Delta ln(E)_t - \Delta ln(P)_t &= \beta_0 + \varepsilon \\
\Delta ln(W)_t &= \beta_0 + \Delta ln(P)_t + \Delta ln(E)_t + \varepsilon \\
\Delta ln(W)_t &= \beta_0 + \beta_1 ln(P)_t + \beta_2 ln(P)_{t-1} + \beta_3 ln(E)_t + \beta_4 ln(E)_{t-1} + \varepsilon
\end{aligned}$$

To this we also include $UR$ because we want to control for the unemployment whereby avoid omitted variable bias. If $\Delta lnUR_t$ has effect on $\Delta lnW_t$ and correlated with $\Delta lnP_t$ and/or $\Delta lnE_t$, then not controlling for $UR$ would lead to biased estimators.

---

**(c) Append the variable $IP$ (strength of Income Policy Index) in Table 1 to your data set (use the `generate` command in Stata to generate a new variable, or copy and paste the data from the question sheet) and run the above regression including $IP$ as an additional regressor and give both a statistical and an economic interpretation of your results.**

**Answer:** Once we append $IP$, the model we will be testing is:

$$\Delta lnW_t = \beta_0 + \beta_1 lnP + \beta_2 lnP_{t-1} + \beta_3 lnE_t + \beta_4 lnE_{t-1} + \beta_5 lnUR_t + \beta_6 lnUR_{t-1} + \beta_7 IP_t + \varepsilon_t$$

```
#Append IP with data from the question

salaries_new_df <- salaries_new_df %>%
  mutate(IP = 0)
salaries_new_df$IP[salaries_new_df$Year == 1962] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1965] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1966] <- 1.5
salaries_new_df$IP[salaries_new_df$Year == 1967] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1968] <- 1.5
salaries_new_df$IP[salaries_new_df$Year == 1969] <- 1.75
salaries_new_df$IP[salaries_new_df$Year == 1975] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1976] <- 3.0
salaries_new_df$IP[salaries_new_df$Year == 1977] <- 4.5
salaries_new_df$IP[salaries_new_df$Year == 1978] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1979] <- 1.0

# Now the number of regressors increased to 7:
k_3c <- 7

#Run the regression
lm_Q3c <- lm(DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR + IP, data = salaries_new_df)
summary(lm_Q3c)
```

```
Call:
lm(formula = DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR +
    IP, data = salaries_new_df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.024935 -0.009426 -0.001330  0.006936  0.037838

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.008500   1.700503  -1.181  0.24785
lnP          0.951886   0.080394  11.840 3.37e-12 ***
lag_lnP     -0.961580   0.081034 -11.866 3.20e-12 ***
lnE          1.098018   0.327043   3.357  0.00235 **
lag_lnE     -0.887196   0.282076  -3.145  0.00401 **
lnUR        -0.003416   0.022894  -0.149  0.88250
lag_lnUR     0.012673   0.018987   0.667  0.51016
IP          -0.014722   0.003492  -4.216  0.00025 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01627 on 27 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.899, Adjusted R-squared:  0.8728
F-statistic: 34.31 on 7 and 27 DF,  p-value: 7.754e-12
```

With the introduction of IP, we now see that all are statistically significant except for $UR$ and its lag, as well as the intercept. Thus we can see that by adding a required variable, $IP$, we reduce the standard error for the whole equation, which in turn, increases the $t$-values of the other coefficients to rise; except for $UR$ which, as we will see, is because $UR$ has no effect.

$R^2$ is not helpful here since it will always increase when a new variable is added. Adjusted $R^2$ also increases from 0.7965 to 0.8728. However, a more useful indicated here is the $F$-statistic which tests that all slope coefficients are zero. If this is higher, then the equation is an improvement. With the introduction of IP, we see an increase from 23.18 to 34.31.

We can also check the significance of $IP$ by looking at its $t$-value which is $-4.216$. Since the $p$-value is 0.00025 we know it is statistically significant and we don't need to calculate the $t$-statistic. As long as $p$-value is less than 0.05 it is significant at the 5 level.

---

**(d) Test the following set of restrictions jointly by re-specifying your preferred equation:**

$$\beta_1 + \beta_2 = 0; \quad \beta_3 + \beta_4 = 0; \quad \beta_5 + \beta_6 = 0$$

**Verify your results using the Wald Test (the `test` command in Stata). Interpret your results.**

**Answer:**

We can do this in two different ways. First is to use the `test` command in Stata, or `linearHypothesis` command from the `car` package in R for the Wald Test. The second way is to impose the constraint on the equation and then compare the constrained and unconstrained equations, testing the restriction on the constrained equation.

```
# Approach 1:

car::linearHypothesis(lm_Q3c, c("lnP=-lag_lnP", "lnE=-lag_lnE", "lnUR=-lag_lnUR"))
```

```
Linear hypothesis test

Hypothesis:
lnP  + lag_lnP = 0
lnE  + lag_lnE = 0
lnUR  + lag_lnUR = 0

Model 1: restricted model
Model 2: DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR + IP

  Res.Df       RSS Df  Sum of Sq      F Pr(>F)
1     30 0.0077553
2     27 0.0071515  3 0.00060376 0.7598 0.5265
```

which gives us the $F$-statistic as 0.7598.

In the second approach, the restricted model we will test is:

$$\Delta lnW_t = \beta_0 + \beta_1 \Delta lnP_t + \beta_3 \Delta lnE_t + \beta_5 \Delta lnUR_t + \beta_7 IP_t + \Delta\varepsilon_t$$

**Why restricted model is a *dynamic* regression model?**

Suppose we have the following model:

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

This model is known as *level form*. Since this level form holds true for every time period, this can also be written as their lagged values:

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + u_{t-1}$$

If we subtract the latter from the former, we will get what is called *first difference form*:

$$\Delta Y_t = \beta_1 \Delta X_t + \Delta u_t$$

where the *first difference operator*, $\Delta$, tells us to take successive differences of the variables in the equation.
So, for example, if $X$ and $Y$ in the level form represent the logarithms of income and consumption expenditure, then $\Delta X$ and $\Delta Y$ represent changes in the logs of income and consumption expenditure, respectively. So, if we are interested in the relationships between the variables in their growth form, then the difference form may be more appropriate. This is because a change in the log of a variable is a relative change, or a percentage change if multiplied by 100.
If the error term in the level form satisfies the OLS assumptions, especially the assumption of no autocorrelation, then the error term $v_t = \Delta u_t = u_t - u_{t-1}$ is autocorrelated. To see this we need to take the expectation and variance of $v_t$:

$$\mathbb{E}(v_t) = \mathbb{E}(u_t - u_{t-1}) = \mathbb{E}(u) - \mathbb{E}(u_{t-1}) = 0$$

since $\mathbb{E}(u) = 0$ for each t.

$$\begin{aligned} Var(v_t) &= Var(u_t - u_{t-1}) \\ &= Var(u_t) + Var(-u_{t-1}) \\ &= \sigma^2 + \sigma^2 \\ &= 2\sigma^2 \end{aligned}$$

thus $v_t$ is homoskedastic. However we also need to look at the covariance of $v_t$ and $v_{t-1}$:

$$\begin{aligned} Cov(v_t, v_{t-1}) &= \mathbb{E}(v_t v_{t-1}) - \mathbb{E}(v_t)\mathbb{E}(v_{t-1}) \\ &= \mathbb{E}\big((u_t - u_{t-1})(u_{t-1} - u_{t-2})\big) - 0 \\ &= \mathbb{E}(u_t u_{t-1} - u_t u_{t-2} - u_{t-1}^2 + u_{t-1}u_{t-2}) \\ &= \mathbb{E}(-u_{t-1}^2) \\ &= -\sigma^2 \end{aligned}$$

which is nonzero. Therefore, although the $u$'s are not autocorrelated the $v$'s are.
Now to see why the restrictions in the question are expressed in the first difference form, consider a general distributed-lag model with a finite lag of $k$ time periods:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_{t-1} + \cdots + \beta_{k+1} X_{t-k} + u_t$$

The coefficient $\beta_1$ is known as the *short-run multiplier*, or *impact multiplier*, because it gives the change in the mean value of $Y$ following a unit change in $X$ in the same time period. If the change in $X$ is maintained at the same level thereafter, then $(\beta_1 + \beta_2)$ gives the change in the mean value of $Y$ in the next period, $(\beta_1 + \beta_2 + \beta_3)$ in the following period, so on. These partial sums are called *interim multipliers* or *intermediate multipliers*. After $k$ time periods we obtain:

$$\sum_{i=1}^{k+1} \beta_i = \beta_1 + \beta_1 + \cdots + \beta_{k+1} = \beta$$

which is known as the *long-run distributed-lag multiplier* or *total distributed-lag multiplier*.
In this question, what we are testing is if that total distributed-lag multiplier is different from 0 for first differences form.

```
#Approach 2

#Modify the data frame to add the constrained versions:
salaries_new_df <- salaries_new_df %>%
  mutate(DlnP = lnP - lag_lnP,
         DlnE = lnE - lag_lnE,
         DlnUR = lnUR - lag_lnUR)

#Estimate the constrained model:
lm_Q3d <- lm(DlnW ~ DlnP + DlnE + DlnUR + IP, data = salaries_new_df)
summary(lm_Q3d)
```

```
Call:
lm(formula = DlnW ~ DlnP + DlnE + DlnUR + IP, data = salaries_new_df)

Residuals:
      Min        1Q    Median        3Q       Max
```

```
-0.029306 -0.009883  0.000174  0.007155  0.036716

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.033841   0.004885   6.928 1.08e-07 ***
DlnP         0.975417   0.064695  15.077 1.53e-15 ***
DlnE         0.961982   0.259391   3.709 0.000845 ***
DlnUR       -0.008418   0.017647  -0.477 0.636796
IP          -0.012631   0.003071  -4.114 0.000279 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01608 on 30 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8904,    Adjusted R-squared:  0.8758
F-statistic: 60.94 on 4 and 30 DF,  p-value: 5.662e-14
```

We will then need to calculate the $F$-statistic:

$$F = \frac{\frac{RSS_R - RSS_{UR}}{df_1}}{\frac{RSS_{UR}}{df_2}}$$

which means we need to obtain the $RSS$s from both regressions which can be done either by `deviance` function in R or by calculating the $RSS$ manually:

```r
#RSS unrestricted
# Manual calculation: sum(resid(lm_Q3c)^2)
RSS_unr_Q3d <- deviance(lm_Q3c)

#RSS restricted
#Manual calculation: sum(resid(lm_Q3d)^2)
RSS_res_Q3d <- deviance(lm_Q3d)

# Difference between the two models (ie number of restrictions):
q_3d <- 3

#F-statistic with (3,27) df:
Fstat_Q3d <- ((RSS_res_Q3d - RSS_unr_Q3d)/q_3d)/(RSS_unr_Q3d/(n_unr-k_3c-1))

#F Critical value:
Crit_Value_Q3d <- qf(0.05, df1=q_3d, df2=n_unr-k_3c-1,lower.tail = FALSE)

tbl_Q3d <- as.table(c("RSS_unr"=RSS_unr_Q3d,
           "RSS_res"=RSS_res_Q3d,
           "F-Stat"=Fstat_Q3d,
           "Critical_Value"=Crit_Value_Q3d,
           "df_1"=q_3d,
           "df_2"=n_unr-k_3c-1))
tbl_Q3d %>%
   kbl(caption = "F-Test for Supplementary Question 3(d)") %>%
  kable_classic(full_width=FALSE)
```

Table 1: F-Test for Supplementary Question 3(d)

| Var1 | Freq |
|---|---|
| RSS_unr | 0.0071515 |
| RSS_res | 0.0077553 |
| F-Stat | 0.7598124 |
| Critical_Value | 2.9603513 |
| df_1 | 3.0000000 |
| df_2 | 27.0000000 |

We again obtain the same $F$-statistic of 0.7598 as in Approach 1, which is much lower than the critical value of 2.96, thus we cannot reject the null hypothesis. That is, the constraints seem to be correct.

---

**(e) Test the further restrictions that $\beta_5 = \beta_6 = 0$. (N.B. your unrestricted equation is always your currently preferred equation, so in this case you only need to test $\beta_5 = 0$ explicitly.)**

**Answer:** We can respecify the constraint and run the regression on this model. Our constrained is $\beta_5 = 0$. This is because in part (d) we have $\beta_5 + \beta_6 = 0$ thus if $\beta_5 = 0$ then $\beta_6 = 0$. a t-test or check the t-statistic of the regression summary:

```
summary(lm_Q3d)$coefficients["DlnUR", "t value"]
```

```
[1] -0.4770329
```

We see that the $t$-stat for $DlnUR$ is $-0.4770329$.

But the question asks this to be done by respecifying the model. Also note that this time our unrestricted model is the restricted model in (d) since we are imposing further restrictions. Accordingly the model we test is:

$$\Delta lnW_t = \beta_0 + \beta_1 \Delta lnP_t + \beta_3 \Delta lnE_t + \beta_7 IP_t + \Delta\varepsilon_t$$

```
#Estimate the constrained model:
lm_Q3e <- lm(DlnW ~ DlnP + DlnE + IP, data = salaries_new_df)
summary(lm_Q3e)
```

```
Call:
lm(formula = DlnW ~ DlnP + DlnE + IP, data = salaries_new_df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.028085 -0.010693  0.000627  0.007825  0.034508

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.033677   0.004812   6.999 7.43e-08 ***
DlnP         0.971549   0.063380  15.329 5.12e-16 ***
DlnE         1.051632   0.176547   5.957 1.39e-06 ***
IP          -0.012928   0.002969  -4.354 0.000135 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01588 on 31 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8896,    Adjusted R-squared:  0.8789
F-statistic: 83.26 on 3 and 31 DF,  p-value: 6.321e-15
```

As we did in part (d) we can obtain the $F$-statistic:

```
# New RSS unrestricted:
RSS_unr_Q3e <- RSS_res_Q3d

# Number of regressors in the new unrestricted model:
k_3e <- 4

#RSS restricted
#Manual calculation: sum(resid(lm_Q3e)^2)
RSS_res_Q3e <- deviance(lm_Q3e)

# Difference between the two models (ie number of restrictions):
q_3e <- 1

#F-statistic with (1,30) df:
Fstat_Q3e <- ((RSS_res_Q3e - RSS_unr_Q3e)/q_3e)/(RSS_unr_Q3e/(n_unr-k_3e-1))

#F Critical value:
Crit_Value_Q3e <- qf(0.05, df1=q_3e, df2=n_unr-k_3e-1,lower.tail = FALSE)

tbl_Q3e <- as.table(c("RSS_unr"=RSS_unr_Q3e,
           "RSS_res"=RSS_res_Q3e,
           "F-Stat"=Fstat_Q3e,
           "Critical_Value"=Crit_Value_Q3e,
           "df_1"=q_3e,
           "df_2"=n_unr-k_3e-1))
tbl_Q3e %>%
  kbl(caption = "F-Test for Supplementary Question 3(e)") %>%
  kable_classic(full_width=FALSE)
```

If we take the square of the t-statistic from the summary table, we should also get this F-statistic:

```
summary(lm_Q3d)$coefficients["DlnUR", "t value"]^2
```

```
[1] 0.2275603
```

which gives us the F-Stat as expected.

Table 2: F-Test for Supplementary Question 3(e)

| Var1 | Freq |
|---|---|
| RSS_unr | 0.0077553 |
| RSS_res | 0.0078141 |
| F-Stat | 0.2275603 |
| Critical_Value | 4.1708768 |
| df_1 | 1.0000000 |
| df_2 | 30.0000000 |

---

**(f) Interpret your preferred equation in light of your results to (c), (d), and (e).**

**Answer:** The equation in (e), $\Delta lnW_t = \beta_0 + \beta_1 \Delta lnP_t + \beta_3 \Delta lnE_t + \beta_7 IP_t + \varepsilon_t$, should be our new preferred equation since on all measures it is the best equation so far. This is good for the first hypothesis, i.e. the real wage resistance hypothesis, as this is exactly what we would expect if the hypothesis is correct.

---

**(g) Test the hypothesis that $\beta_3 = 1$ by respecifying your preferred equation in terms of average wages and salaries. (Your F-statistic for this test should be about 0.086)**

**Answer:** One way to answer this question is to run a $t$-test on the coefficient of $\Delta lnE$:

```
(summary(lm_Q3e)$coefficients["DlnE","Estimate"]-1)/summary(lm_Q3e)$coefficients["DlnE","Std. Error"]
```

```
[1] 0.2924517
```

We check against the critical $t$ value

```
qt(p=0.025, df=34, lower.tail = FALSE)
```

```
[1] 2.032245
```

Thus we cannot reject the null hypothesis.

However, the question actually ask us to respecify the equation. For this, we set the coefficient on $\Delta lnE$ to 1, i.e. $\beta_3 = 1$ and then take it over to the other side of the equation so that the model becomes:

$$\Delta lnW_t - \Delta lnE_t = \beta_0 + \beta_1 \Delta lnP_t + \beta_7 IP_t + \Delta\varepsilon_t$$

or

$$\Delta ln\left(\frac{W_t}{E_t}\right) = \beta_0 + \beta_1 \Delta lnP_t + \beta_7 IP_t + \Delta\varepsilon_t$$

We can run the regression on this:

```r
# First modify the data frame to reflect the new dependent variable:
salaries_new_df <- salaries_new_df %>%
  mutate("DlnWE" = DlnW-DlnE)

# Then run the regression:
lm_Q3g <- lm(DlnWE ~ DlnP + IP, salaries_new_df)
summary(lm_Q3g)
```

```
Call:
lm(formula = DlnWE ~ DlnP + IP, data = salaries_new_df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.029237 -0.010758  0.000967  0.007744  0.034665

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.033934   0.004663   7.278 2.85e-08 ***
DlnP         0.968616   0.061681  15.704  < 2e-16 ***
IP          -0.012897   0.002925  -4.410 0.000109 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01565 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8862,    Adjusted R-squared:  0.8791
F-statistic: 124.6 on 2 and 32 DF,  p-value: 7.943e-16
```

We can now derive the $F$-statistic as before:

```r
# New RSS unrestricted:
RSS_unr_Q3g <- RSS_res_Q3e

# Number of regressors in the new unrestricted model:
k_3g <- 3

#RSS restricted
#Manual calculation: sum(resid(lm_Q3g)^2)
RSS_res_Q3g <- deviance(lm_Q3g)

# Difference between the two models (ie number of restrictions):
q_3g <- 1

#F-statistic with (1,30) df:
Fstat_Q3g <- (RSS_res_Q3g - RSS_unr_Q3g/q_3g)/(RSS_unr_Q3g/(n_unr-k_3g-1))

#F Critical value:
Crit_Value_Q3g <- qf(0.05, df1=q_3g, df2=n_unr-k_3g-1,lower.tail = FALSE)

tbl_Q3g <- as.table(c("RSS_unr"=RSS_unr_Q3g,
           "RSS_res"=RSS_res_Q3g,
           "F-Stat"=Fstat_Q3g,
           "Critical_Value"=Crit_Value_Q3g,
```

Table 3: F-Test for Supplementary Question 3(g)

| Var1 | Freq |
|---|---|
| RSS_unr | 0.0078141 |
| RSS_res | 0.0078357 |
| F-Stat | 0.0855280 |
| Critical_Value | 4.1596151 |
| df_1 | 1.0000000 |
| df_2 | 31.0000000 |

```
            "df_1"=q_3g,
            "df_2"=n_unr-k_3g-1))
tbl_Q3g %>%
   kbl(caption = "F-Test for Supplementary Question 3(g)") %>%
  kable_classic(full_width=FALSE)
```

we can check that this is the square of the $t$-stat:

```
sqrt(Fstat_Q3g)
```

```
[1] 0.2924517
```

We cannot reject $\mathbb{H}_0$ and conclude that $\beta_3$ is not statistically different from 1.

---

**(h) In the light of your preferred specification for the total period does it appear that price inflation is fully passed on in wage claims?**

**Answer:** This question is basically asking if the coefficient of $\Delta lnP$ is equal to 1 or not. We can use the $t$-test for this:

```
(summary(lm_Q3g)$coefficients["DlnP","Estimate"]-1)/summary(lm_Q3g)$coefficients["DlnP","Std. Error"]
```

```
[1] -0.5088118
```

We check against the critical $t$ value

```
qt(p=0.025, df=34, lower.tail = FALSE)
```

```
[1] 2.032245
```

Thus we cannot reject the null hypothesis.

---

**(i) Interpret and comment upon your findings (especially in relation to the hypothesis suggested in (b)) and explain why your preferred equation is now given by:**

$$\Delta ln\left(\frac{W}{EP}\right)_t = \beta_0 + \beta_7 IP_t + \Delta\varepsilon$$

**Answer:** We have tested back as far as we can go and at each stage the real wage resistance hypothesis gives us the best equation. It is not clear what an hypothesis would have to do in order to be better.

---

**(j) Why might it be expected that a structural break took place around 1979/1980? Test the hypothesis that the structure of wage determination changed significantly from 1980 onwards using the method of dummy variables (NB this is a different test to those you used in Q1 and Q2). Discuss your results.**

**Answer:** The short answer is Thatcher. End to corporatism and wage bargaining etc.

To test this, rather than using the Chow Test, use the dummy variables method. For this, we define a variable, $M$ which is 0 for the period 1956-1979, and 1 for 1980 onwards. We then estimate the following:

$$\Delta lnW_t = \beta_0 + \beta_1 \Delta lnP_t + \beta_8 M_t + \beta_9 M \Delta lnP_t + \Delta\varepsilon$$

```
# First, add M column to the data frame:

salaries_new_df <- salaries_new_df %>%
  mutate(
    M = case_when(
      Year < 1980 ~ 0,
      Year >= 1980 ~ 1
    ),
    MDlnP = M * DlnP
  )

# Then run the regression:

lm_Q3j <- lm(DlnWE ~ DlnP + IP + M + MDlnP, salaries_new_df)
summary(lm_Q3j)
```

```
Call:
lm(formula = DlnWE ~ DlnP + IP + M + MDlnP, data = salaries_new_df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027717 -0.010469  0.001171  0.010470  0.027110

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.0354777  0.0052097   6.810 1.49e-07 ***
DlnP         1.0073834  0.0681582  14.780 2.59e-15 ***
IP          -0.0151997  0.0031867  -4.770 4.46e-05 ***
M           -0.0003343  0.0118536  -0.028    0.978
MDlnP       -0.1351737  0.1586255  -0.852    0.401
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01543 on 30 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8962,    Adjusted R-squared:  0.8824
F-statistic: 64.76 on 4 and 30 DF,  p-value: 2.523e-14
```

Notice that there is not a $MlnP$ term because if we include it we would have perfect multicollinearity. From the regression output we can see that neither $M$ nor $MDlnP$ are significant - see their t-stats, so we can conclude that there is not any evidence of structural break for these years.

---

**(k) Discuss the advantages of using the method of dummy variables over the Chow Test you used in part 1(a).**

**Answer:** With the dummy variable we find out which variable or variables have changed. Whereas Chow test tells us only that something has changed. That is, it will tell us if the two sub-regressions are different but not whether the difference is due to the intercepts, the slope, or both.

# FACULTY QUESTIONS

## QUESTION 1

**An investigator analysing the relationship between food expenditure, disposable income and prices across a random sample of 25 counties in the UK estimates the relationship**

$$\begin{array}{c} log(FOOD) = 4.7377 + 0.3506\ log(PDI) - 0.5086\ log(PRICE) \\ \quad\quad\quad\quad\quad (0.6805)\ \ (0.0899)\quad\quad\quad\quad (0.1010) \end{array}$$

**where figures in paranthesis are standard errors and where:**

**FOOD = Average household expenditure on food**

**PDI = Average personal dosposable income**

**PRICE = Average price of food deflated by a general price index**

**(i) Give an economic interpretation of the coefficients on log(PDI) and log(PRICE).**

**Answer:** The coefficient of $log(PDI)$ is income elasticity, and the coefficient of $log(PRICE)$ is the price elasticity.

---

**(ii) Test the hypothesis, using 5% significance level, that the coefficient of log(PRICE) is equal to zero against the alternative that it is nonzero.**

**Answer:** For this we would use a $t$-test for the hypothesis that

$$\mathbb{H}_0 : \beta_2 = 0 \text{vs } \mathbb{H}_A : \beta_2 \neq 0$$

The $t$-statistic is:

$$\frac{-0.5086 - 0}{0.1010} = -5.035644$$

which is significantly different than $t_{0.025,24} = -2.064$, and so we would reject the null hypothesis.

---

**(iii) Test the hypothesis using 5% significance level that the coefficient of log(INCOME) is equal to 1 against the alternative that it is significantly different from 1.**

**Answer:** For this we would use a $t$-test for the hypothesis that

$$\mathbb{H}_0 : \beta_1 = 1 \quad \text{vs} \quad \mathbb{H}_A : \beta_1 \neq 1$$

The $t$-statistic is:

$$\frac{0.3506 - 1}{0.0899} = -7.223582$$

which is significantly different than $t_{0.025,24} = -2.064$, and so we would reject the null hypothesis.

---

**(iv) You are now given the following extra information:**

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 0.52876$$
$$RSS = \sum_{i=1}^{n} e_i^2 = 0.46276$$

**Compute ESS and $R^2$ for the above regression**.

**Answer:**

$$ESS = TSS - RSS = 0.52876 - 0.46276 = 0.066$$
$$R^2 = \frac{ESS}{TSS} = \frac{0.066}{0.52876} = 0.1248203.$$

---

**(v) Test the joint hypothesis at the 5% level that the two slope coefficients are all equal to zero against the alternative that at least one slope coefficient is nonzero.**

**Answer:** For this we would use an $F$-test for the hypothesis that

$$\mathbb{H}_0 : (\beta_1 = 0) \cap (\beta_2 = 0)$$

The $F$-statistic is:

$$F = \frac{\frac{ESS}{df_1}}{\frac{RSS}{df_2}} = \frac{\frac{0.066}{2}}{\frac{0.46279}{22}} = 1.568746$$

or

$$F = \frac{\frac{R^2_{unr} - R^2_{res}}{df_1}}{\frac{1 - R^2_{unr}}{df_2}} = \frac{\frac{0.1248203}{2}}{\frac{1-0.1248203}{22}} = 1.568847$$

giving a marginal difference due to rounding errors. In either case the F-statistic is significantly different than $F_{0.05,2,22} = 0.0514$, and so we would reject the null hypothesis.

— — — — — — — — — — — — — — — — — — — — — — — — — —

## QUESTION 2

**Download the dataset `wage2.dta` Use the STATA commands `des` and `sum` to understand the structure of the data and the meaning of the variable labels. Now answer the following questions relating to performance on the IQ test for this sample of working age women. The IQ test was taken as an adult after the woman had completed her formal education.**

Load the data in R, though the equivalent commands produce output that is not as neat as Stata here are the commands for obtaining summary and structure of the data frame. I keep them in the "comment" format to save paper when printing:

```
wage2_df <- read_dta("../Data/wage2.dta")
# summary(wage2_df)
# str(wage2_df)

# We should also check for any NA data entries per column
colSums(is.na(wage2_df))
```

```
   wage   hours      IQ     KWW    educ   exper  tenure     age married   black
      0       0       0       0       0       0       0       0       0       0
  south   urban    sibs brthord   meduc   feduc   lwage
      0       0       0      83      78     194       0
```

**(a) Run a regression of IQ test score on parents' education. What do you conclude?**

**Answer:** We need to estimate the following regression:

$$IQ = \beta_0 + \beta_1 feduc_i + \beta_2 meduc_i + u_i, \quad i = 1, \ldots, n.$$

The regression would give us:

```
lm_FQ2a <- lm(IQ ~ feduc + meduc, data = wage2_df)
summary(lm_FQ2a)
```

```
Call:
lm(formula = IQ ~ feduc + meduc, data = wage2_df)

Residuals:
    Min      1Q  Median      3Q     Max
-51.161  -8.161   0.023   9.839  39.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.2109     2.0864  38.445  < 2e-16 ***
feduc         1.0168     0.1884   5.396 9.24e-08 ***
meduc         1.0624     0.2202   4.825 1.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.67 on 719 degrees of freedom
  (213 observations deleted due to missingness)
Multiple R-squared:  0.1465,    Adjusted R-squared:  0.1442
F-statistic: 61.72 on 2 and 719 DF,  p-value: < 2.2e-16
```

We can see that the coefficients for both parents' education are positive, individually close to unity, and statistically significant given the high $t$-values and very low $p$-values. We can also see that the $F$-statistic of 61.72 is quite high which means the regression coefficients are jointly significant. On the other hand, we can see that the goodness-of-fit is relatively low at $R^2 = 0.147$.

---

**(b) Suppose I want to know whether the only way parents' education increases their daughter's test score is through the daughter's own education. How would you test that hypothesis?**

**Answer:** This question is effectively asking us to test whether parents' education level has no effect on daughter's IQ beyond its influence on the child's education level. Therefore, we need to estimate a regression on $IQ$ on the daughter's own education level, in addition to father's and mother's education levels, and subsequently test the joint significance of the parents' education. Thus we have a restricted and unrestricted model:

$$\text{Unestricted Model: } IQ = \beta_0 + \beta_1 feduc_i + \beta_2 meduc_i + \beta_3 educ_i + u_i,$$
$$\text{Restricted Model: } IQ = \gamma_0 + \gamma_1 educ_i + v_i.$$

The joint hypothesis for the $F$-test is:

$$\mathbb{H}_0 : (\beta_1 = 0) \cap (\beta_2 = 0)$$
$$\mathbb{H}_A : (\beta_1 \neq 0) \cup (\beta_2 \neq 0)$$

The $F$-test statistic has $F_{q,(n-k-1)}$ distribution and is obtained either via $RSS$ or $R^2$:

$$F = \frac{\frac{RSS_{res} - RSS_{unr}}{q}}{\frac{RSS_{unr}}{(n-k-1)}} = \frac{\frac{(R^2_{unr} - R^2_{res})}{q}}{\frac{1-R^2_{unr}}{(n-k-1)}} \sim F_{q,(n-k-1)}$$

We can calculate this in R as follows:

```
# Before obtaining estimates we need to remove the rows with NAs:
 wage2_nona_df <- wage2_df %>%
  filter(!is.na(feduc), !is.na(meduc))

# Obtain the estimates
lm_FQ2b_unr <- lm(IQ ~ feduc + meduc + educ, data = wage2_df)
lm_FQ2b_res <- lm(IQ ~ educ, data = wage2_nona_df)

#RSS unrestricted
# Manual calculation: sum(resid(lm_FQ2b_unr)^2)
RSS_unr_FQ2b <- deviance(lm_FQ2b_unr)

#RSS restricted
#Manual calculation: sum(resid(lm_FQ2b_res)^2)
RSS_res_FQ2b <- deviance(lm_FQ2b_res)

# or to use R-squareds:
Rsq_unr <- summary(lm_FQ2b_unr)$r.squared
Rsq_res <- summary(lm_FQ2b_res)$r.squared

# Difference between the two models (ie number of restrictions):
q_FQ2b <- 2

# Sample size - since we have NAs in both columns but in different rows we need to remove them:
n_FQ2b <- nrow(wage2_nona_df)

# number of regressors in unrestricted model:
k_FQ2b <- 3

#F-statistic:
Fstat_FQ2b_withRSS <- ((RSS_res_FQ2b - RSS_unr_FQ2b)/q_FQ2b)/(RSS_unr_FQ2b/(n_FQ2b-k_FQ2b-1))
Fstat_FQ2b_withRsq <- ((Rsq_unr-Rsq_res)/q_FQ2b)/((1-Rsq_unr)/(n_FQ2b-k_FQ2b-1))

#F Critical value:
Crit_Value_FQ2b <- qf(0.95, df1=q_FQ2b, df2=n_FQ2b-k_FQ2b-1,lower.tail = TRUE)

tbl_FQ2b <- as.table(c("RSS_unr"=round(RSS_unr_FQ2b,4),
           "RSS_res"=round(RSS_res_FQ2b,4),
           "R-squared_unr" = round(Rsq_unr,4),
           "R-squared_res" = round(Rsq_res,4),
           "F-Stat (using RSS)"=round(Fstat_FQ2b_withRSS, 4),
```

Table 4: F-Test for Faculty Question 2(b)

| Var1 | Freq |
|---|---|
| RSS_unr | 107986.9064 |
| RSS_res | 111952.2591 |
| R-squared_unr | 0.3136 |
| R-squared_res | 0.2884 |
| F-Stat (using RSS) | 13.1827 |
| F-Stat (using R-sq) | 13.1827 |
| Critical_Value | 3.0083 |
| df_1 | 2.0000 |
| df_2 | 718.0000 |

```
            "F-Stat (using R-sq)"=round(Fstat_FQ2b_withRsq, 4),
            "Critical_Value"=round(Crit_Value_FQ2b, 4),
            "df_1"=round(q_FQ2b, 4),
            "df_2"=round(n_FQ2b-k_FQ2b-1,4)))

tbl_FQ2b %>%
  kbl(caption = "F-Test for Faculty Question 2(b)") %>%
  kable_classic(full_width=FALSE)
```

Which gives us an F-statistic of 13.18272. Comparing this to the critical values at 95 we can reject the null.

The same test can be done using `linearHypothesis()` function from the `car` package:

```
# A quick way to do the F-test:
car::linearHypothesis(lm_FQ2b_unr, c("feduc=0","meduc=0"))
```

```
Linear hypothesis test

Hypothesis:
feduc = 0
meduc = 0

Model 1: restricted model
Model 2: IQ ~ feduc + meduc + educ

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    720 111952
2    718 107987  2    3965.4 13.183 2.385e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which gives us an F-statistic of 13.183. Comparing this to the critical values at 95 and 99 we can reject the null.

```
qf(0.99,2,718, lower.tail = TRUE)
```

```
[1] 4.634834
```

```r
qf(0.95,2,718, lower.tail = TRUE)
```

```
[1] 3.008266
```

---

**(c) What do you conclude from your test of part (b)?**

**Answer:** In either approaches, we can conclude that the parents' education affects daughter's IQ beyond their influence on the daughter's own education. Perhaps there is a direct genetic effect on IQ.

---

**(d) Is there any policy recommendation from your results?**

**Answer:** Notice from the coefficients of the unrestricted model that mother's education has twice the partial effect of the father's education and its statistical significance is higher. One could suggest that the emphasis of the policy should be on raising educational level on women.

---

**(e) Suppose I wish to know whether the impact of own and parents' education on IQ test score varies by race. Conduct such a test of hypothesis and report your finding.**

**Answer (though not necessarily a robust way to answer the question)** One approach is to regress $IQ$ on father's education, mother's education, and own education for two subsets separately, using a dummy variable for *black*, and then test the hypothesis that the coefficents in the two regressions are the same.

To conduct this test we need three regressions: a full sample regression which we already obtained in part (b), a regression for the sub-sample of non-black individuals, and a regression for the sub-sample of black individuals:

$$black = 0 : IQ = \beta_0^{b0} + \beta_1^{b0} educ_i + \beta_2^{b0} meduc_i + \beta_3^{b0} educ_i + u_i$$
$$black = 1 : IQ = \beta_0^{b1} + \beta_1^{b1} educ_i + \beta_2^{b1} meduc_i + \beta_3^{b1} educ_i + u_i$$

The hypothesis we are testing is whether the coefficients are equal or not:

$$\mathbb{H}_0 : (\beta_0^{b0} = \beta_0^{b1}) \cap (\beta_1^{b0} = \beta_1^{b1}) \cap (\beta_2^{b0} = \beta_2^{[}b1]) \cap (\beta_3^{b0} = \beta_3^{b1})$$
$$\mathbb{H}_A : (\beta_0^{b0} \neq \beta_0^{b1}) \cup (\beta_1^{b0} \neq \beta_1^{b1}) \cup (\beta_2^{b0} \neq \beta_2^{b1}) \cup (\beta_3^{b0} \neq \beta_3^{b1})$$

To test this we use the Chow test with the following $F$-test statistic:

$$F = \frac{\frac{\left(RSS - (RSS^{b0} + RSS^{b1})\right)}{k+1}}{\frac{(RSS^{b0} + RSS^{b1})}{n-2(k+1)}} \sim F_{(k+1,(n-2(k+1)))}$$

The estimation are as follows:

```
#First divide data into two based on the dummy variable `black`:
wage2_nona_b1_df <- wage2_nona_df %>%
  filter(black==1)
wage2_nona_b0_df <- wage2_nona_df %>%
  filter(black==0)

# Estimate the regressions:
lm_FQ2e_b0 <- lm(IQ ~ feduc + meduc + educ, data = wage2_nona_b0_df)
lm_FQ2e_b1 <- lm(IQ ~ feduc + meduc + educ, data = wage2_nona_b1_df)

#Conduct the Chow test:
RSS_FQ2e_b0 <- deviance(lm_FQ2e_b0)
RSS_FQ2e_b1 <- deviance(lm_FQ2e_b1)

# For degrees of freedom:
k_FQ2e <- 3

# Chow Test F-statistic:
F_Chow_FQ2e <- ((RSS_unr_FQ2b-(RSS_FQ2e_b0 + RSS_FQ2e_b1))/(k_FQ2e+1))/((RSS_FQ2e_b0 + RSS_FQ2e_b1)/(n_

# F Critical value:
Crit_Value_FQ2e <- qf(0.95, df1=k_FQ2e+1, df2=n_FQ2b-2*(k_FQ2e+1),lower.tail = TRUE)

tbl_FQ2e <- as.table(c(
  "RSS" = round(RSS_unr_FQ2b,4),
  "RSS_b0" = round(RSS_FQ2e_b0, 4),
  "RSS_b1" = round(RSS_FQ2e_b1, 4),
  "Chow F-Stat" = round(F_Chow_FQ2e,4),
  "Crit. Value" = round(Crit_Value_FQ2e, 4),
  "df1" = k_FQ2e+1,
  "df2" = n_FQ2b-2*(k_FQ2e+1)
))

tbl_FQ2e %>%
   kbl(caption = "Chow Test for Faculty Question 2(e)") %>%
  kable_classic(full_width=FALSE)
```

The $F$-stat of 17.1274 is much higher than the critical value of 2.3844 at 95 with $(4, 714)$ degrees of freedom. We can therefore reject the null hypothesis and conclude that the coefficients for the black and non-black individuals are different. Accordingly, from the data we can ascertain that race may have an effect on how one's own and parental education influences IQ.

Table 5: Chow Test for Faculty Question 2(e)

| Var1 | Freq |
|---|---|
| RSS | 107986.9064 |
| RSS_b0 | 86164.2032 |
| RSS_b1 | 12368.3212 |
| Chow F-Stat | 17.1274 |
| Crit. Value | 2.3844 |
| df1 | 4.0000 |
| df2 | 714.0000 |

However, notice that the Chow test here is testing the equality of all coefficients including the intercept. But in this question we are interested only in the three slope coefficients.

**Answer: A more robust approach**

A more appropriate approach would be to test the hypothesis of race difference with interacting variables as in the following model:

$$IQ = \beta_0 + \beta_1 feduc_i + \beta_2 meduc_i + \beta_3 educ_i + \beta_4 (feduc_i \times black_i) + \beta_5 (meduc_i \times black_i) + \beta_6 (educ_i \times black_i) + u_i$$

Notice that the additional variables are the interactions of the educational variables with the dummy variable. The coefficients $\beta_4, \beta_5$ and $\beta_6$ show the additional partial effect of father's education, mother's education, and own education for black individuals as compared to non-black individuals. Therefore, in order to test whether the impact of own and parents' education on IQ test score varies by race, we have the following hypothesis:

$$\mathbb{H}_0 : (\beta_4 = 0) \cap (\beta_5 = 0) \cap (\beta_6 = 0)$$
$$\mathbb{H}_A : (\beta_4 \neq 0) \cup (\beta_5 \neq 0) \cup (\beta_6 \neq 0)$$

We therefore create a new set of variables with these interacting terms in the dataframe and then run the regression:

```
# create the new variables
wage2_nona_df <- wage2_nona_df %>%
  mutate(
    educb = educ * black,
    feducb = feduc * black,
    meducb = meduc * black
  )

# run the regression

lm_FQ2e <- lm(IQ ~ feduc + meduc + educ + feducb + meducb + educb, data = wage2_nona_df)
```

We would now conduct an $F$-test where this model is the unrestricted regression, and the unrestricted model in 2(b) as the restricted regression here:

```
#RSS unrestricted
# Manual calculation: sum(resid(lm_FQ2e_unr)^2)
RSS_unr_FQ2e <- deviance(lm_FQ2e)
```

```r
#RSS restricted
#Manual calculation: sum(resid(lm_FQ2b_unr)^2)
RSS_res_FQ2e <- deviance(lm_FQ2b_unr)

# or to use R-squareds:
Rsq_unr_FQ2e <- summary(lm_FQ2e)$r.squared
Rsq_res_FQ2e <- summary(lm_FQ2b_unr)$r.squared

# Difference between the two models (ie number of restrictions):
q_FQ2e <- 3

# Sample size - since we have NAs in both columns but in different rows we need to remove them:
n_FQ2e <- nrow(wage2_nona_df)

# number of regressors in unrestricted model:
k_FQ2e <- 6

#F-statistic:
Fstat_FQ2e_withRSS <- ((RSS_res_FQ2e - RSS_unr_FQ2e)/q_FQ2e)/(RSS_unr_FQ2e/(n_FQ2e-k_FQ2e-1))
Fstat_FQ2e_withRsq <- ((Rsq_unr_FQ2e-Rsq_res_FQ2e)/q_FQ2e)/((1-Rsq_unr_FQ2e)/(n_FQ2e-k_FQ2e-1))

#F Critical value:
Crit_Value_FQ2e <- qf(0.95, df1=q_FQ2e, df2=n_FQ2e-k_FQ2e-1,lower.tail = TRUE)

tbl_FQ2b <- as.table(c("n" = n_FQ2e,
          "RSS_unr"=round(RSS_unr_FQ2e,4),
          "RSS_res"=round(RSS_res_FQ2e,4),
          "R-squared_unr" = round(Rsq_unr_FQ2e,4),
          "R-squared_res" = round(Rsq_res_FQ2e,4),
          "F-Stat (using RSS)"=round(Fstat_FQ2e_withRSS, 4),
          "F-Stat (using R-sq)"=round(Fstat_FQ2e_withRsq, 4),
          "Critical_Value @ 95%"=round(Crit_Value_FQ2e, 4),
          "df_1"=round(q_FQ2e, 4),
          "df_2"=round(n_FQ2e-k_FQ2e-1,4)))

tbl_FQ2b %>%
  kbl(caption = "F-Test for Faculty Question 2(e)") %>%
  kable_classic(full_width=FALSE)
```

The test statistic of 22.2773 is greater than the critical value of 2.6174 so we reject the null hypothesis and conclude that race has an effect on the way education affects IQ test score.

As before, we could run an F-test in R instead of calculating it manually:

```r
car::linearHypothesis(lm_FQ2e, c("feducb=0", "meducb=0", "educb=0"))
```

```
Linear hypothesis test

Hypothesis:
feducb = 0
meducb = 0
educb = 0
```

Table 6: F-Test for Faculty Question 2(e)

| Var1 | Freq |
|---|---|
| n | 722.0000 |
| RSS_unr | 98756.0542 |
| RSS_res | 107986.9064 |
| R-squared_unr | 0.3723 |
| R-squared_res | 0.3136 |
| F-Stat (using RSS) | 22.2773 |
| F-Stat (using R-sq) | 22.2773 |
| Critical_Value @ 95% | 2.6174 |
| df_1 | 3.0000 |
| df_2 | 715.0000 |

```
Model 1: restricted model
Model 2: IQ ~ feduc + meduc + educ + feducb + meducb + educb

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    718 107987
2    715  98756  3    9230.9 22.277 8.476e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

which yields the same result.

— — — — — — — — — — — — — — — — — — — — — — — — — —

## QUESTION 3

**Let us revisit the following problem you did in the previous supervision sheet. Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + U_i$ for an i.i.d. sample with $N = 1000$ observations. Suppose $U_i \sim i.i.d.(0, \sigma^2)$ and the $X_i$ are i.i.d. for $i = 1, 2, \ldots, 1000$ and that $X_i$ is independent of $U_i$. Let $\hat{\beta}_i$ denote the OLS estimator of $\beta_1$ and consider another estimator $\tilde{\beta}_1$ of $\beta_1$ constructed in the following way:**

$$\tilde{\beta}_1 = \frac{Y_3 - Y_2}{X_3 - X_2}$$

**You can assume that $X_i$ are continuously distributed and that $X_3 - X_2$ never takes the value of 0.**

**Is $\tilde{\beta}_1$ a consistent estimator of $\beta_1$? Why?**

**Answer:** First observe that $\tilde{\beta}_1$ does not change as more data are added, and so it remains a random variable as $n \to \infty$. That is, it does not converge to $\beta_1$.

For consistency, the variance of this estimator also needs to converge to 0 as $n \to \infty$. So, let's calculate its variance. For this, recall that the regressors $X_i$ are not fixed but are assumed to random and drawn

from some distribution. This means, $X_i$ have an expectation and a variance. Importantly, this also means $\mathbb{E}(X_i) \neq X_i$ but $E(X_i|X_i = x) = x$, or, alternatively, $\mathbb{E}(X_i|X_i) = X_i$. That is, conditional on $X_i$ taking a realization $x_i$, the expection of $X_i$ is $x$. Therefore, we will need to rely on *the law of iterated expectations* to obtain the variance of this estimator.

For this, we need to follow these steps:

1. Express the estimator in terms of $X_1, X_2$ and $U_1, U_2$,
2. Obtain the variance of this estimator
3. Take the conditional expectation of this variance
4. Utilize the law of iterated expectations to derive the unconditional variance

**Step 1: Express $\tilde{\beta}_1$ without $Y_2, Y_3$**

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{Y_3 - Y_2}{X_3 - X_2} \\
&= \frac{\beta_0 + \beta_1 X_3 + U_3 - \beta_0 - \beta_1 X_2 - U_2}{X_3 - X_2} \\
&= \frac{\beta_1(X_3 - X_2) + (U_3 - U_2)}{X_3 - X_2} \\
&= \beta_1 + \frac{U_3 - U_2}{X_3 - X_2}.
\end{aligned}$$

**Step 2: Take variance of this expression**

$$\begin{aligned}
Var(\tilde{\beta}_1) &= Var(\beta_1 + \frac{U_3 - U_2}{X_3 - X_2}) \\
&= Var(\frac{U_3 - U_2}{X_3 - X_2}) \quad \text{because } \beta_1 \text{ is a constant parameter} \\
&= \mathbb{E}\left[\Big(\frac{U_3 - U_2}{X_3 - X_2}\Big)^2\right] - \left[\mathbb{E}\Big(\frac{U_3 - U_2}{X_3 - X_2}\Big)\right]^2 \\
&= \mathbb{E}\left[\Big(\frac{U_3 - U_2}{X_3 - X_2}\Big)^2\right] - \left[\Big(\frac{1}{X_3 - X_2}\Big)\big(\mathbb{E}(U_3) - \mathbb{E}(U_2)\big)\right]^2 \\
&= \mathbb{E}\left[\Big(\frac{U_3 - U_2}{X_3 - X_2}\Big)^2\right] - 0 \quad \text{since } \mathbb{E}(U_i) = 0,\ i = 2, 3
\end{aligned}$$

**Step 3: Take the conditional expectation of this expression**

$$\begin{aligned}
\mathbb{E}[Var(\tilde{\beta}_1|X_2,X_3)] &= \mathbb{E}\left[\Big(\frac{U_3-U_2}{X_3-X_2}\Big)^2\middle|X_2,X_3\right] \\
&= \mathbb{E}\left[\Big(\frac{U_3^2+U_2^2-2U_3U_2}{(X_3-X_2)^2}\Big)\middle|X_2,X_3\right] \\
&= \mathbb{E}\Big(\frac{1}{(X_3-X_2)^2}\middle|X_2,X_3X_2,X_3\Big) \\
&\quad\times\left[\mathbb{E}\big(U_3^2|X_2,X_3\big)+\mathbb{E}\big(U_2^2|X_2,X_3\big)-\mathbb{E}\big(2U_3U_2|X_2,X_3\big)\right] \\
&= \mathbb{E}\Big(\frac{1}{(X_3-X_2)^2}\middle|X_2,X_3X_2,X_3\Big)(\sigma^2+\sigma^2-0) \\
&\quad \text{since } Var(u_i|\vec{X})=\sigma^2,\quad i=1,\ldots,n \\
&\quad \text{and } Cov(u_i,u_j)=\mathbb{E}(u_iu_j)-\mathbb{E}(u_i)\mathbb{E}(u_j)=\mathbb{E}(u_iu_j)=0 \\
&= \frac{2\sigma^2}{(X_3-X_2)^2} \quad \text{since } \mathbb{E}\Big(\frac{1}{(X_3-X_2)^2}\middle|X_2,X_3X_2,X_3\Big)=\frac{1}{(X_3-X_2)^2}.
\end{aligned}$$

**Step 4: Utilize law of iterated expectations to derive unconditional variance**

$$\begin{aligned}
Var(\tilde{\beta}_1) &= \mathbb{E}\big[\mathbb{E}(\tilde{\beta}_1)|X_2,X_3\big] \\
&= \mathbb{E}\left(\mathbb{E}\left[\Big(\frac{U_3-U_2}{X_3-X_2}\Big)^2\middle|X_2,X_3\right]\right) \\
&= \mathbb{E}\Big(\frac{2\sigma^2}{(X_3-X_2)^2}\Big) \\
&= 2\sigma^2\mathbb{E}\Big(\frac{1}{(X_3-X_2)^2}\Big).
\end{aligned}$$

$\hookrightarrow$ Note that we cannot simplify this any further since $\mathbb{E}\left(\frac{1}{(X_3-X_2)^2}\right) \neq \frac{1}{(X_3-X_2)^2}$ unlike $\mathbb{E}\left(\frac{1}{(X_3-X_2)^2}\middle|X_2,X_3\right)=\frac{1}{(X_3-X_2)^2}$. This is because $X_i$ are random variables which means $\mathbb{E}(X_i)\neq X_i$ but $\mathbb{E}(X_i|X_i=x)=x$, or alternatively $\mathbb{E}(X_i|X_i)=X_i$, as discussed at the top of this answer.

Now we can check how this variance behaves when $n$ tends to infinity. We can again see that this variance does not change as more data are added, and consequently does not converge to 0 as $n\to\infty$. Accordingly, this estimator is not consistent.

## STATA - FACULTY QUESTIONS

Load the data in STATA:

```
* Change the working directory to access the data file:
    quietly cd ..
* Load the data:
    use Data/wage2
* Run the command for describe and summarize:
  * des
  * summ

* part (a) regression:
  * regress IQ feduc meduc

* part (b) regressions:
  * regress IQ feduc meduc educ
  * regress IQ educ if ((feduc!=.) & (meduc!=.))

* part (b) alternative way:
  * regress IQ feduc meduc educ
  * test feduc meduc

* part (e) regressions for not so robust answer:
  * reg IQ feduc meduc educ if black == 0
  * reg IQ feduc meduc educ if black == 1

* part (e) regressions for more robust answer:
  * generate the new variables:
      gen educb = educ * black
      gen feducb = feduc * black
      gen meducb = meduc * black
  * reg IQ feduc meduc educ feducb meducb educb
  * test feducb meducb educb

* put all of these into a neat table using the `STATA` command `esttab` from the `estout` package:
  * first install the package:
      ssc install estout, replace
  * then run the following regressions together to build the table:
      eststo: quietly reg IQ feduc meduc
      eststo: quietly reg IQ feduc meduc educ
      eststo: quietly reg IQ educ if ((feduc!=.) & (meduc!=.))
      eststo: quietly reg IQ feduc meduc educ if black == 0
      eststo: quietly reg IQ feduc meduc educ if black == 1
      eststo: quietly reg IQ feduc meduc educ feducb meducb educb
      esttab, r2 ar2 scalars(F rss rmse)
```

```
(194 missing values generated)

(78 missing values generated)

checking estout consistency and verifying not already installed...
all files already exist and are up to date.
```

```
(est1 stored)

(est2 stored)

(est3 stored)

(est4 stored)

(est5 stored)

(est6 stored)


--------------------------------------------------------------------------------
> --------------------------
                      (1)             (2)             (3)             (4)
>     (5)             (6)
                       IQ              IQ              IQ              IQ
>      IQ              IQ
--------------------------------------------------------------------------------
> --------------------------
feduc                1.017***        0.363*                          0.187
>   0.703            0.185
                    (5.40)          (2.06)                          (1.07)
>  (1.09)           (1.04)

meduc                1.062***        0.612**                         0.399*
>   0.745            0.421*
                    (4.82)          (3.05)                          (1.98)
>  (1.09)           (2.05)

educ                                 3.030***        3.547***        3.000***
>   2.478*           3.060***
                                    (13.22)         (17.08)         (13.53)
>  (2.26)          (13.80)

feducb
>                    0.545

>                   (0.97)

meducb
>                    0.329

>                   (0.55)

educb
>                   -1.573***

>                  (-3.62)

_cons                80.21***        50.38***        53.66***        56.06***
>   42.45**          55.00***
                    (38.45)         (17.19)         (18.67)         (19.14)
```

```
>  (3.41)         (19.12)
--------------------------------------------------------------------------------
> -------------------------
N                     722             722             722             657
>      65             722
R-sq                0.147           0.314           0.288           0.300
>   0.225           0.372
adj. R-sq           0.144           0.311           0.287           0.297
>   0.187           0.367
F                   61.72           109.4           291.8           93.29
>   5.903           70.68
rss               134273.9        107986.9        111952.3        86164.2
> 12368.3         98756.1
rmse                13.67           12.26           12.47           11.49
>   14.24           11.75
--------------------------------------------------------------------------------
> -------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```