ECON 3201
Econometrics for Economics and Finance

Assignment 4
Due December 8, 2025

Parambir Singh Atwal

Instructions

Answer all questions. Please complete your assignment using Quarto. Submit both your Quarto file and pdf/word output file. Append any written answers to a single PDF file. That is, **submit only one PDF file.**

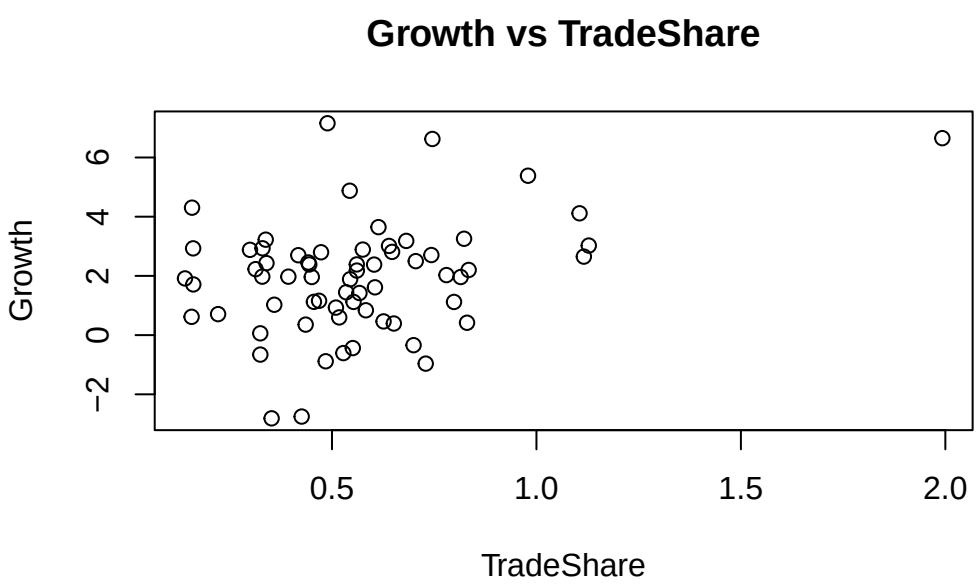
For this assignment, it would be helpful to create a new folder, e.g. **assignment_4**. Then, set your working directory to this folder. Next, save the file "a4.RData" to this directory. To use this data, type `load("a4.RData")` (This step is already done at the top of the quarto file). Once the data is loaded, the necessary datasets will be available for use, i.e., you will not have to use the `data()` command.

Questions

1. For this question, use the `Growth` data. This data frame contains data on average growth rates from 1960 to 1995 for 65 countries, along with variables that are potentially related to growth.

   (a) Construct a scatterplot of average growth rate (*Growth*) on the average trade share (*TradeShare*). Does there appear to be a relationship between the variables?

   Ans. 1.(a) The correlation achieved, i.e., Cor(x,y) = 0.3516, indicates a positive relationship between trade share and growth rate, but it is relatively weak with considerable scatter.




```
[1] 0.351682
```

(b) Estimate the following model using ordinary least squares

$$Growth = \beta_0 + \beta_1 TradeShare + u$$

by:

(i) Constructing the estimators manually.

```
[1] 0.6402653
```

```
[1] 2.306434
```

(ii) The `lm()` command. Did you get the same values as in part (i)?

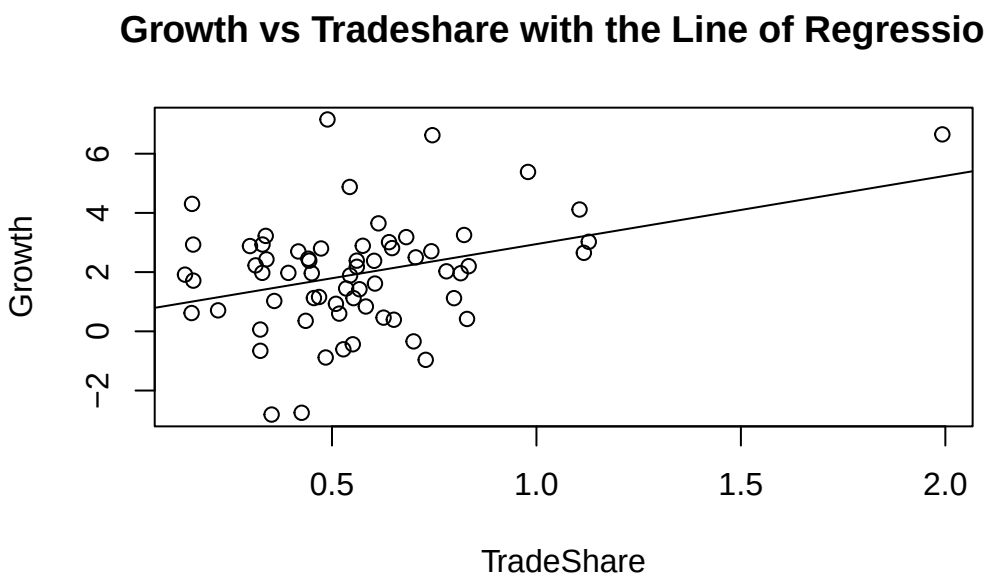
```
(Intercept)  tradeshare
  0.6402653   2.3064337
```

Yes, I received the same values as in part (i) after rounding-off the tradeshare to 6 decimal places.

(c) Use the estimated parameters to predict the growth rate for a country with a trade share of 0.5 and for another with trade share equal to 1.0.

```
[1] 1.793482 2.946699
```

(d) Plot the estimated regression function from (b) along with your scatterplot from (a).



2. For this question use the dataset `Earnings_and_Height`, which contains data on earnings, height, and other characteristics of a random sample of U.S. workers. In this question, you will investigate the relationship between earnings and height.

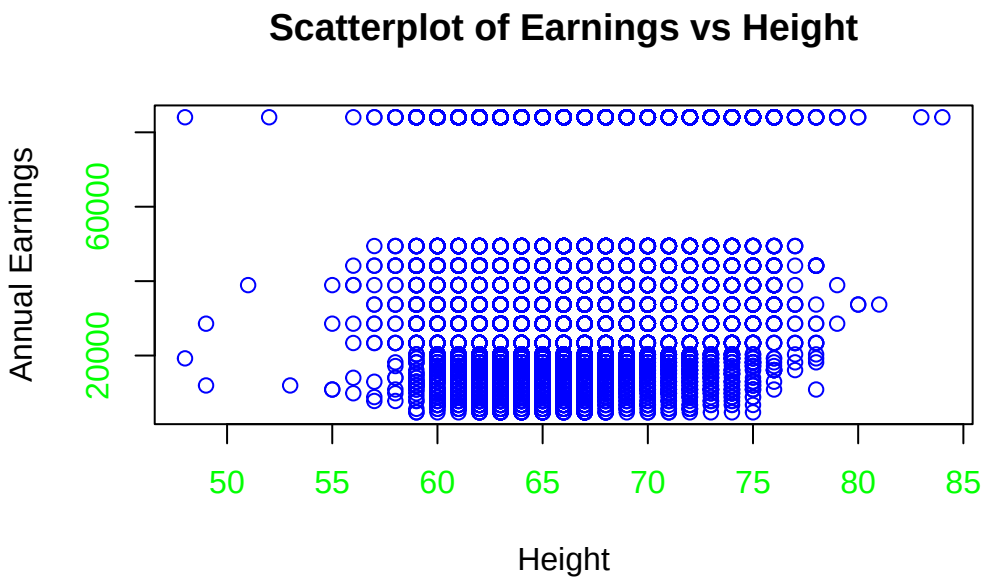
   (a) What are mean, median, and mode values of `height` in the sample?

   ```
   [1] 66.96335
   ```

   ```
   [1] 67
   ```

   ```
   [1] 66
   ```

   (b) Construct a scatterplot of annual earnings (`Earnings`) on height (`Height`).



(c) Run a regression of `Earnings` on `Height`.

(i) What is the estimated slope?

```
  height
707.6716
```

(ii) What is the estimated intercept?

```
(Intercept)
  -512.7336
```

(iii) What is the R^2?

```
[1] 0.0108753
```

3. The sample covariance of two random variables X and Y in a sample of size n is given by

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

The sample correlation is then given by

$$r_{XY} = \frac{s_{XY}}{s_X s_Y},$$

where s_X and s_Y are the sample standard deviations of X and Y, respectively.

(a) Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r^2_{XY}$.

We know that $R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$, and $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$\Rightarrow \hat{Y}_i - \bar{Y} = \hat{\beta}_1(X_i - \bar{X}) \Rightarrow R^2 = \frac{\sum_{i=1}^{n}(\hat{\beta}_1(X_i - \bar{X}))^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \Rightarrow R^2 = \hat{\beta}_1^2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

$\Rightarrow R^2 = \hat{\beta}_1^2 \frac{(n-1)s_X^2}{(n-1)s_Y^2} = \hat{\beta}_1^2 \frac{s_X^2}{s_Y^2}$

$\Rightarrow \hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$ {since, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$}

$\Rightarrow R^2 = \frac{s_{XY}^2}{s_X^4} \frac{s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2}$

$\Rightarrow R^2 = r^2_{XY}$ {since, $r^2_{XY} = \frac{s_{XY}^2}{s_X^2 s_Y^2}$}

(b) Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

Now we know that, $\Rightarrow R^2 = r^2_{XY}$

The regression of Y on X: $R^2_{Y|X} = r^2_{XY}$, and the regression of X on Y: $R^2_{X|Y} = r^2_{YX} = r^2_{XY}$

{since, the sample correlation r_{xy} is symmetric over X and Y $\Rightarrow r_{XY} = r_{YX}$}

$\Rightarrow R^2_{Y|X} = R^2_{X|Y}$

Hence, proved.

(c) Show that

$$\hat{\beta}_1 = r_{XY}\left(\frac{s_Y}{s_X}\right)$$

Again, from part (a), we have: $\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$

$\Rightarrow \hat{\beta}_1 = \frac{s_{XY}}{s_X s_X}\frac{s_Y}{s_Y} = \frac{s_{XY}}{s_X s_Y}\frac{s_Y}{s_X} \Rightarrow \hat{\beta}_1 = r_{XY}\left(\frac{s_Y}{s_X}\right)$

4. Use the `hprice1` data set to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

(a) Write out the results in equation form.

```
(Intercept)       sqrft       bdrms
-19.3149958   0.1284362  15.1981910
```

Equation form is as follows (I have not added u as we are making this equation after estimation):

$\widehat{price} = -19.3149958 + 0.1284362 \times sqrft + 15.1981910 \times bdrms$

(b) What is the estimated increase in the price for a house with one more bedroom, holding square footage constant?

```
   bdrms
15.19819
```

(c) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size?

```
   sqrft
17.98107
```

```
   bdrms
15.19819
```

```
   sqrft
33.17926
```

5. Use the `birthweight` dataset to answer this questions.

(a) Regress `birthweight` on `smoker`. What is the estimated effect of smoking on birth weight?

```
(Intercept)      smoker
  3432.0600   -253.2284
```

```
   smoker
-253.2284
```

(b) Regress `birthweight` on `smoker`, `alcohol`, and `nprevist`.

```
Call:
lm(formula = birthweight ~ smoker + alcohol + nprevist, data = birthweight)

Coefficients:
(Intercept)       smoker      alcohol     nprevist
    3051.25      -217.58       -30.49        34.07
```

(i) Explain why the exclusion of `alcohol` and `nprevist` could lead to omitted variable bias in the regression estimated in (a).

Ans (i) If alcohol consumption and prenatal visits are correlated with both smoking and birth weight, excluding them could bias the estimate of smoking's effect. E.g., smokers might also drink more or have fewer prenatal visits, both affecting birth weight.

(ii) Does the regression in (a) suffer from omitted variable bias?

Ans (ii) Yes, comparing the smoking coefficient in (a): efined: $\hat{\beta_{smoker}}^{(a)} = -253.2284$, with that in (b) in (b) $\hat{\beta_{smoker}}^{(b)} = -217.58$, they are significantly different after including alcohol and nprevist, suggesting omitted variable bias.

(iii)Jane smoked during pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birthweight of Jane's child.

```
[1] 3106.23
```

```
       1
3106.228
```

(c) An alternative way to control for prenatal visits is to use the binary variables `tripre0` through `tripre3`, which indicate whether the the mother had no prenatal visits (`tripre0`), first prenatal visit in the first trimester (`tripre1`), first prenatal visit in the second trimester (`tripre2`), and the first prenatal visit in the third trimester (`tripre3`). Regress `birthweight` on `smoker, alcohol, tripre0, tripre2, tripre3`.

```
Call:
lm(formula = birthweight ~ smoker + alcohol + tripre0 + tripre2 +
    tripre3, data = birthweight)

Coefficients:
(Intercept)       smoker      alcohol      tripre0      tripre2      tripre3
     3454.5       -228.8        -15.1       -698.0       -100.8       -137.0
```

Regression Equation (estimated) = $\widehat{birthweight} = 3454.5 - 228.8smoker - 15.1Alcohol - 698tripre0 - 100.8tripre2 - 137.0tripre3$

(i) Why is `tripre1` excluded from the regression? What would happen if you include it in the regression?

The variables tripre0 through tripre3 represent mutually exclusive and exhaustive categories. Each observation belongs to exactly one group, so the sum of these four binary indicators is always equal to 1 for every individual. Consequently, if all four are included in a regression that also contains a constant term, the resulting design matrix will be

singular, leading to perfect collinearity. This prevents unique estimation of the model parameters. To avoid this, we have to exclude one category and use it as the baseline or reference group for comparison.

(ii) What does the estimated coefficient on `tripre0` measure? Interpret its value.

The coefficient indicates that mothers who did not attend prenatal visits give birth to infants with a lower predicted weight, with the magnitude of this decrease estimated to be 697.97 units.

The End. Thank you Professor…