ECONOMETRICS

Collection of Exercises
(with Solutions)

PRELIMINARY AND INCOMPLETE
DO NOT SHARE WITHOUT PERMISSION

Marko Mlikota

https://markomlikota.github.io

marko.mlikota@graduateinstitute.ch

This version: 2025-09-11

Contents

1 Probability Theory

1.1 Covariance of Conditional Expectation

Let X and Y be two scalar random variables (RVs) with finite variance. Show that

(a) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \mathbb{E}[Y|X])$

(b) $\mathrm{Cov}(X, Y - \mathbb{E}[Y|X]) = 0$

1.2 Transformation of RVs

Suppose that $X \sim N(0,1)$, and define $Y = g(X) = exp\{X\}$.

(a) Write a small program that (i) generates 500 draws x_i from $N(0,1)$; (ii) plots a histogram of these 500 draws; (iii) overlays a plot of the pdf of the random variable X.

(b) Find the pdf of Y. Be sure to specify also the domain of Y.
Hint: first find the cdf of Y using that of X. Once you have $F_Y(y)$, take derivatives to find $f_Y(y)$, the pdf of Y.

(c) Write a small program that (i) converts the 500 x_i's obtained in (a) into y_i's by applying the transformation $y_i = g(x_i)$; (ii) plots a histgram of the y_i draws; (iii) overlays a plot of the pdf of the random variable Y derived in (b).

Exercise 1.1: Solution

Let X and Y be two scalar random variables (RVs) with finite variance. Show that

(a) $\text{Cov}(X, Y) = \text{Cov}(X, \mathbb{E}[Y|X])$

(b) $\text{Cov}(X, Y - \mathbb{E}[Y|X]) = 0$

Solution:

(a) We know that for any two RVs $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. Importantly, $\mathbb{E}[Y|X]$ is also a RV: the RV that is equal to the expectation of the RV Y given the (realization of the) RV X. Hence,

$$\begin{aligned}
\text{Cov}(X, \mathbb{E}[Y|X]) &= \mathbb{E}[X\mathbb{E}[Y|X|] - \mathbb{E}[X]\mathbb{E}[\mathbb{E}[Y|X|] \\
&= \mathbb{E}[X\mathbb{E}[Y|X|] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[\mathbb{E}(XY|X)] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \text{Cov}(X, Y) \ ,
\end{aligned}$$

where the second and fourth equalities follow by the Law of Iterated Expectations and the third equality is obtained by taking X, which is a constant for an expectation given X, into that expectation.

(b) Because we can write $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ and the expectation is a linear operator, we have

$$\begin{aligned}
\text{Cov}(X, Y + Z) &= \mathbb{E}[X(Y + Z)] - \mathbb{E}[X]\mathbb{E}[Y + Z] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] \\
&= \text{Cov}(X, Y) + \text{Cov}(X, Z) \ .
\end{aligned}$$

Using this together with the result from (a) gives

$$\text{Cov}(X, Y - \mathbb{E}[Y|X]) = \text{Cov}(X, Y) - \text{Cov}(X, \mathbb{E}[Y|X]) = 0 \ .$$

Exercise 1.2: Solution

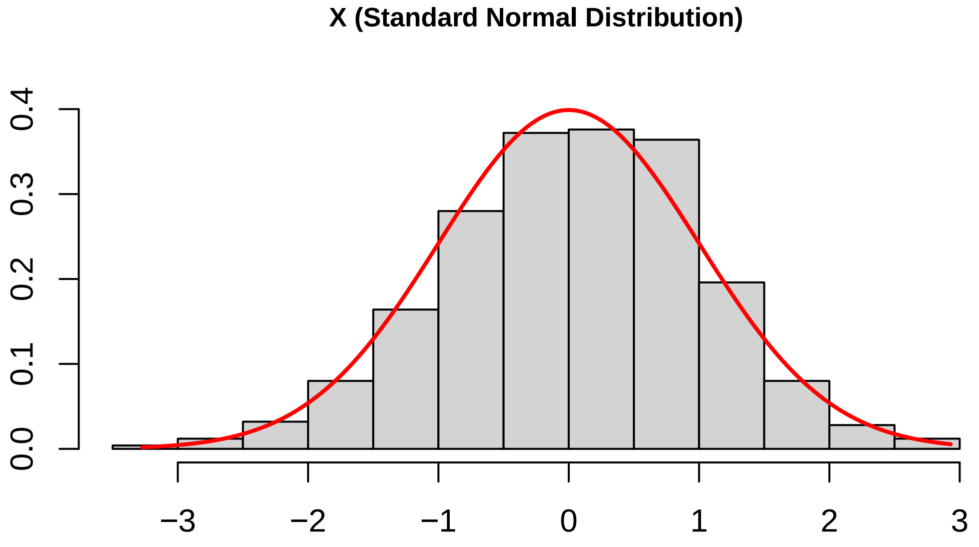
Suppose that $X \sim N(0,1)$, and define $Y = g(X) = exp\{X\}$.

1. Write a small program that (i) generates 500 draws x_i from $N(0,1)$; (ii) plots a histogram of these 500 draws; (iii) overlays a plot of the pdf of the random variable X.

Solution:

```r
#clear and set seed
rm(list = ls())
set.seed(2024)

#generate sample size and random vector
n <- 500
x <- rnorm(n)

#plot histogram
hist(x, #what variable to plot
     freq = FALSE, #absolute vs. relative frequency
     main = "X (Standard Normal Distribution)", #title
     cex.main = 0.8, #title size
     xlab = " ", #x-axis label
     ylab = " ", #y-axis label
     ylim = c(0, 0.4)) #y-axis scale

#plot normal density:
#(i) generate vector of sorted x random draws
x_draws <- seq(min(x), #smallest x draw
                max(x), #largest x draw
                length.out = 500) #total number x draws

#(ii) compute std. normal pdf values
density <- dnorm(x_draws, #points at which to compute the std. norm pdf values
                 mean = 0, #std. norm mean
                 sd = 1) #std. norm s.d.

#overlay line plot of std. normal pdf
lines(x_draws, #x values
      density, #y values
      col = "red", #line colour
      lwd = 2) #line width
```

Exercise 1.2: Solution



2. Find the pdf of Y. Be sure to specify also the domain of Y.

Hint: first find the cdf of Y using that of X. Once you have $F_Y(y)$, take derivatives to find $f_Y(y)$, the pdf of Y.

Solution: We know $F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[exp\{X\} \leq y] = \mathbb{P}[X \leq ln(y)] = F_X(ln(y))$, and therefore,

$$\begin{aligned}
f_Y(y) &= \frac{\partial F_Y(y)}{\partial y} \\
&= \frac{\partial F_X(ln(y))}{\partial y} \\
&= \frac{1}{y} f_X(ln(y)) \\
&= \frac{1}{y}(2\pi)^{-1/2} exp\left\{-\frac{1}{2} ln(y)^2\right\} \\
&= \frac{1}{y\sqrt{2\pi}} exp\left\{-\frac{1}{2} ln(2y)\right\} \ .
\end{aligned}$$

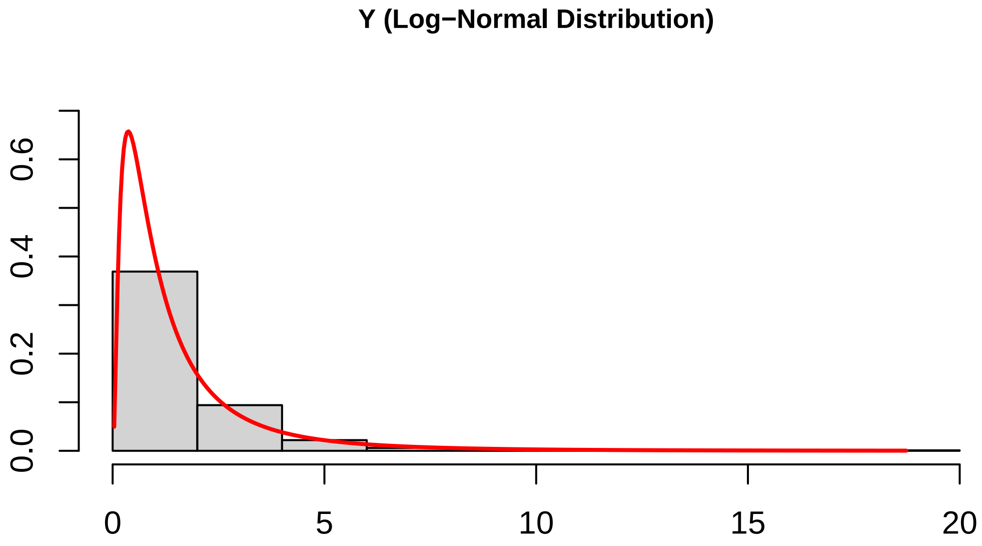
Note that this is a log-Normal distribution.

3. Write a small program that (i) converts the 500 x_i's obtained in
   (a) into y_i's by applying the transformation $y_i = g(x_i)$; (ii) plots a histgram of the y_i draws; (iii) overlays a plot of the pdf of the random variable Y derived in (b).

```
#generate y_i = g(x_i)
y <- exp(x)

#plot histogram
hist(y, #what variable to plot
     freq = FALSE, #absolute vs. relative frequency
     main = "Y (Log-Normal Distribution)", #title
     cex.main = 0.8, #title size
```

Exercise 1.2: Solution

```
     xlab = " ", #x-axis label
     ylab = " ", #y-axis label
     ylim = c(0, 0.7)) #y-axis scale

#plot normal density:
#(i) generate vector of sorted y random draws
y_draws <- seq(min(y), #smallest y draw
               max(y), #largest y draw
               length.out = 500) #total number y draws

#(ii) compute log-normal pdf values
density_y <- dlnorm(y_draws, #points at which to compute the log-norm pdf
                    meanlog = 0, #underlying normal mean
                    sdlog = 1) #underlying normal s.d.

#overlay line plot of log-normal pdf
lines(y_draws, #x values
      density_y, #y values
      col = "red", #line colour
      lwd = 2) #line width
```



2 Statistical Inference

2.1 Conditional Mean as Best Estimator under Quaratic Loss

Show that

$$\mathbb{E}[Y \mid X] \in \arg\min_{f(x)} \mathbb{E}[(Y - f(X))^2] \ ,$$

i.e. under the quadratic loss function, the best estimator of some random variable Y given a random variable X is the conditional mean $\mathbb{E}[Y|X]$.
Hint: First show that $g(x) \in \arg\min_{f(x)} \mathbb{E}[(Y - f(X))^2|X]$ *implies* $g(x) \in \arg\min_{f(x)} \mathbb{E}[(Y - f(X))^2]$. *Then show that the conditional mean solves the former by taking first-order conditions (FOCs).*

2.2 Inference in Location-Scale Model: Basics & Finite Sample-Calculations

Suppose you have data on the height of n female adults living in Switzerland – $\{x_i\}_{i=1}^n$ – whereby the observations in your sample are independent. Based on that, you want to estimate the average height of female adults in the whole population (i.e. the whole of Switzerland). Let this parameter of interest be denoted by θ. You can write your observations as

$$x_i = \theta + u_i \ , \quad \text{with} \quad \mathbb{E}[u_i|\theta] = 0 \ ,$$

i.e. the height of an individual i, x_i, is given by the true average height θ plus some noise u_i around it. Note that this is just another way of writing $\mathbb{E}[x_i|\theta] = \theta$.

(a) Find a point estimator for θ using the Least Squares (LS) estimation method, $\hat{\theta}$.

(b) What is the mean of $\hat{\theta}$? Is your $\hat{\theta}$ unbiased? Besides assuming $\mathbb{E}[x_i|\theta] = \theta$, is there any other assumption on the pdf of $x_i|\theta$ involved in finding this quantity? Are any assumptions regarding your sample $\{x_i\}_{i=1}^n$ involved?

(c) What is the variance of $\hat{\theta}$? Besides assuming $\mathbb{E}[x_i|\theta] = \theta$, is there any other assumption on the pdf of $x_i|\theta$ involved in finding this quantity? Are any assumptions regarding your sample $\{x_i\}_{i=1}^n$ involved?

(d) Suppose $u_i \sim U(-5,5)$, i.e. u_i is distributed Uniformly between -5 and 5. Using a statistical software of your choice, write a program that, given a choice of n and θ simulates a dataset $\{x_i\}_{i=1}^n$. Fix $\theta = 175$ and $n = 10$. Compute $\hat{\theta}$ using this simulated data. Is your estimate close to the true value of $\theta = 175$? What happens under a dataset with $n = 100$ observations? What happens if you take $n = 1000$?

(e) Now let's use the program you wrote to analyze the behavior of $\hat{\theta}$ in repeated sampling.

   (a) simulate $M = 100$ different datasets of size $n = 10$: $\{\{x_i^m\}_{i=1}^n\}_{m=1}^M$

   (b) for each dataset $\{x_i^m\}_{i=1}^n$, compue the LS-point estimator $\hat{\theta}^m$

   (c) plot a histogram of $\{\hat{\theta}^m\}_{m=1}^M$

   Comment on the histogram (distribution) of $\hat{\theta}$-values. Is it in line with your expectations, based on the calculations you did for the questions above?

(f) Redo the previous exercise using $n = 100$ as well as $n = 1000$. How does the histogram (distribution) of $\hat{\theta}$ change? Relate this to the theoretical analysis we conducted in class.

2.3 Inference in Location-Scale Model: LS Estimator & t-Test

Suppose $\mathbb{E}[X] = \theta$ and $\mathbb{V}[X] = \sigma^2$ for some known σ^2. Suppose we observe n i.i.d. observations of the random variable X, denoted by $\{x_i\}_{i=1}^n$.

(a) Define and derive the Ordinary Least Squares (OLS) estimator of θ, $\hat{\theta}_{OLS}$.

(b) What can you say about the (finite sample) distribution of $\hat{\theta}_{OLS}$? Is $\hat{\theta}_{OLS}$ unbiased?

(c) What is the asymptotic distribution of $\hat{\theta}_{OLS}$? Is $\hat{\theta}_{OLS}$ consistent?

(d) Set up the two-sided t-test with size $\alpha = 0.05$ for testing $\mathscr{H}_0 : \theta_0 = 0$ against $\mathscr{H}_1 : \theta_0 \neq 0$. More concretely, definining the test as

$$\varphi_t = \mathbf{1}\{T(X) < c_\alpha\} \ ,$$

define the test-statistic $T(X)$ and find the critical value c_α.
Hint: To find c_α, you need to use the definition of a size α-test and the distribution of $T(X)$ under $\mathscr{H}_0$. Thereby, you may not know the corresponding finite sample distribution, but you might know the asymptotic distribution, allowing you to construct an asymptotically valid test.

For the following exercises, suppose X follows a Normal distribution: $X \sim N(\theta, \sigma^2)$.

(e) Find the power function of your t-test, $\beta(\theta) = \mathbb{P}[\text{reject}|\theta] = \mathbb{P}[T(X) > c_\alpha|\theta]$.

(f) Find the critical values associated with $\alpha = 0.1$ and $\alpha = 0.01$ as well. Write a program that plots the power function $\beta(\theta)$ as a function of θ for the three different values of α. Set $\sigma = 2$.

How does the shape of $\beta(\theta)$ change across the different sizes α?

(g) Based on your t-test, find an expression for the 95% confidence interval for θ, $C(X)$. What happens with $C(X)$ as n increases? What happens with $C(X)$ as σ increases? Discuss.

(h) Set $\theta = 0$ and $\sigma = 2$ and simulate a sample of $n = 10$ draws x_i under $\mathcal{H}_0$. Find $C(x)$, the confidence interval associated with your draws $x = \{x_i\}_{i=1}^{10}$. Is θ in your confidence interval?

(i) If you were to repeatedly draw $\{x_i\}_{i=1}^{10}$ many times, how often would you expect the true θ to lie in $C(x)$? Find out the answer numerically as follows. Take $M = 1000$ and create an $M \times 1$ vector v. Then, for $m = 1 : M$,

(1) simulate $\{x_i^m\}_{i=1}^{10}$ from $N(\theta, \sigma^2)$, $\theta = 0$, $\sigma = 2$,

(2) construct $C(x^m)$,

(3) if $\theta \in C(x^m)$, then record a 1 as the mth entry of vc, otherwise record a zero.

To see how often (across M) $\theta \in C(x^m)$, compute the sum of elements in vc. Is the result in line with your expectations?

2.4 Inference in Location-Scale Model: ML Estimator & LR-Test

Suppose you have data on the height of n female adults living in Switzerland – $\{x_i\}_{i=1}^n$ – whereby the observations in your sample are independent. Based on that, you want to estimate the average height of female adults in the whole population (i.e. the whole of Switzerland). Let this parameter of interest be denoted by θ. You can write your observations as

$$x_i = \theta + u_i \ , \quad \text{with} \quad \mathbb{E}[u_i|\theta] = 0 \ ,$$

i.e. the height of an individual i, x_i, is given by the true average height θ plus some noise u_i around it. Note that this is just another way of writing $\mathbb{E}[x_i|\theta] = \theta$. In addition, you assume that this noise u_i is Normally distributed with some known variance σ^2: $u_i \sim N(0, \sigma^2)$. Note that this – combined with the equation for x_i above – is just another way of writing $x_i \sim N(\theta, \sigma^2)$.

(a) Define and derive the Maximum Likelihood (ML) estimator of θ, $\hat{\theta}_{ML}$.
Hint: You first need to derive the likelihood, i.e. the distribution of your data $\{X_i\}_{i=1}^n$ conditional on θ, $p(x|\theta)$, based on the distribution of a single observation X_i, $p(x_i|\theta)$.

(b) Set up a Likelihood Ratio (LR) test of size $\alpha = 0.05$ for testing $\mathcal{H}_0 : \theta = \theta_0$ against $\mathcal{H}_1 : \theta \neq \theta_0$, i.e. determine the test-statistic $T(X)$ and the corresponding critical value c_α.
Hint: Note that if $Y \sim N(\mu, v)$, then $(Y - \mu)/\sqrt{v} \sim N(0, 1)$ and $(Y - \mu)^2/v \sim \chi_1^2$.

(c) Suppose $\sigma^2 = 6$ and you observe $n = 4$ observations, $x_1 = 178$, $x_2 = 161$, $x_3 = 168$ and $x_4 = 172$. Based on this data, can you reject $\mathcal{H}_0 : \theta = \theta_0 = 175$ (i.e. that the average height of female adults in Switzerland is 175cm)?

(d) Now let's suppose you could only find the test-statistic for the LR test, $T(X)$, but not the critical value c_α. Do so numerically, i.e.

(1) For $m = 1 : M$, with $M = 1000$,

- draw a sample $\{x_i^m\}_{i=1}^n \sim N(\theta_0, \sigma^2)$, setting $\theta_0 = 175$, $\sigma^2 = 6$ and $n = 4$,
- compute $T(x^m)$.

Plot a histogram of $\{T(x^m)\}_{m=1}^M$. This is your numerical approximation of the distribution of $T(X)$ under $\mathscr{H}_0$.

(2) Sort your draws $\{T(x^m)\}_{m=1}^M$ from lowest to largest and take the $M(1-\alpha)$th draw. This is your numerical approximation of c_α, the $100(1-\alpha)$th quantile of the distribution of $T(X)$.

Is the value you get close to the true, analytically obtained c_α? What do you expect to happen if you take a larger value for M? Does your conclusion from the previous exercise change if you set up your test numerically as opposed to analytically?

(e) Based on your LR-test, find a (general) expression for the 95% confidence interval for θ_0, $C(X)$. How does that interval look like if you apply it to your particular data? Is $\theta_0 = 175$ in that interval? Explain why it should (not) be.

(f) Let's again suppose you were not able to analytically set up the LR test and, based on it, find $C(X)$. Find the confidence interval $C(x)$ for your sample numerically as follows. First, fix a grid $\mathcal{T}$ of values for θ_0, $\mathcal{T} = 160 : 0.1 : 180$, and create a vector vc of the same dimension as $\mathcal{T}$. Then, for each $\theta_0 \in \mathcal{T}$,

(1) repeat the numerical procedure from above to find $c_\alpha(\theta_0)$, the (numerical approximation of the) critical value for a size $\alpha = 0.05$ test for testing $\mathscr{H}_0 : \theta = \theta_0$.

(2) compute the LR-test-statistic $T(x;\theta_0)$ for your sample x. If $T(x;\theta_0) < c_\alpha(\theta_0)$, then $\theta_0 \in C(x)$ and you record a 1 in the corresponding entry in vc, otherwise $\theta_0 \notin C(x)$ and you record a 0.

Illustrate your $C(x)$ using a scatter plot: put $\mathcal{T}$ on the x-axis and, for each value $\theta \in \mathcal{T}$, have a one on the y-axis if θ is in $C(x)$ and a zero otherwise. How does your $C(x)$ compare to the one obtained analytically?

2.5 Bayesian Inference in Location-Scale Model

Suppose $X|\ \theta \sim N(\theta, \sigma^2)$ for some known σ^2. Suppose we observe n i.i.d. observations of the random variable X, $\{x_i\}_{i=1}^n$. Suppose also your prior (on where θ lies) is $\theta \sim N(0, \tau)$.

(a) Explain briefly the difference between the Bayesian paradigm and the frequentist/classical paradigm (of which OLS and ML are a part).

(b) Derive the posterior distribution $p(\theta|x) \propto p(x|\theta)p(\theta)$.
Hint: You first need to determine the distribution of $\{x_i\}_{i=1}^n$ *conditional on* θ, $p(x|\theta)$, *based on the distribution* $p(X|\theta)$.

(c) What happens with the posterior if τ changes? What happens if n or σ change?

(d) Define the Bayes estimator to be the posterior mean, $\hat{\theta}_B = \mathbb{E}[\theta|x]$, and forget about the posterior distribution for the next two exercises; we are back in the frequentist/classical paradigm and we want to evaluate the point-estimator $\hat{\theta}_B$, ignoring that it was derived under the Bayesian paradigm.
What is the distribution of $\hat{\theta}_B| \ \theta$? Is $\hat{\theta}_B$ unbiased? How does its variance compare with the variance of $\hat{\theta}_{OLS}$?

(e) Is $\hat{\theta}_B$ consistent? What is its asymptotic distribution, and how does it compare to that of $\hat{\theta}_{OLS}$?

(f) We are back in the Bayesian paradigm. Set $\theta = 0$ and $\sigma = 2$, and simulate a sample $\{x_i\}_{i=1}^{10}$. Then test $\mathscr{H}_0 : \theta \in [-0.5, 0.5]$ against $\mathscr{H}_1 : \theta \notin [-0.5, 0.5]$ by computing posterior odds. What do you conclude?

Exercise 2.1: Solution

Show that

$$\mathbb{E}[Y \mid X] \in \arg\min_{f(x)} \mathbb{E}[(Y - f(X))^2] \ ,$$

i.e. under the quadratic loss function, the best estimator of some random variable Y given a random variable X is the conditional mean $\mathbb{E}[Y|X]$.

Hint: First show that $g(x) \in \arg\min_{f(x)} \mathbb{E}[(Y - f(X))^2|X]$ *implies* $g(x) \in \arg\min_{f(x)} \mathbb{E}[(Y - f(X))^2]$. *Then show that the conditional mean solves the former by taking first-order conditions (FOCs).*

Solution: Note that:

$$\begin{aligned} g(x) \in \arg\min_{f(x)} \mathbb{E}[(Y - f(X))^2 \mid X] &\Rightarrow \mathbb{E}[(Y - g(X))^2 \mid X] \leq \mathbb{E}[(Y - f(X))^2 \mid X] \quad \forall\, f(X) \\ &\Rightarrow \mathbb{E}[\mathbb{E}[(Y - g(X))^2 \mid X]] \leq \mathbb{E}[\mathbb{E}[(Y - f(X))^2 \mid X]] \quad \forall\, f(X) \\ &\Rightarrow \mathbb{E}[(Y - g(X))^2] \leq \mathbb{E}[(Y - f(X))^2] \quad \forall\, f(X) \ . \end{aligned}$$

The first line writes out the definition of a minimum, the second applies an expectation operator to both sides, and the third line applies the Law of Iterated Expectations (LIE).

We can write out

$$\mathbb{E}[(Y - f(X))^2 \mid X] = \mathbb{E}[Y^2 \mid X] - 2f(X)\mathbb{E}[Y \mid X] + f(X)^2 \ .$$

Taking FOCs w.r.t. $f(X)$ gives $-2\mathbb{E}[Y \mid X] + 2f(X) = 0$, which yields $f(X) = \mathbb{E}[Y \mid X]$.

Exercise 2.2: Solution

Suppose you have data on the height of n female adults living in Switzerland – $\{x_i\}_{i=1}^n$ – whereby the observations in your sample are independent. Based on that, you want to estimate the average height of female adults in the whole population (i.e. the whole of Switzerland). Let this parameter of interest be denoted by θ. You can write your observations as

$$x_i = \theta + u_i \ , \quad \text{with} \quad \mathbb{E}[u_i|\theta] = 0 \ ,$$

i.e. the height of an individual i, x_i, is given by the true average height θ plus some noise u_i around it. Note that this is just another way of writing $\mathbb{E}[x_i|\theta] = \theta$.

1. Find a point estimator for θ using the Least Squares (LS) estimation method, $\hat{\theta}$.

Solution:

The LS estimator minimizes the total sum of squares, defined as

$$\sum_{i=1}^n (x_i - \mathbb{E}[x_i|\theta])^2 = \sum_{i=1}^n (x_i - \hat{\theta})^2 \ .$$

Since this function is globally concave, we can compute the LS estimator by taking the first order condition:

$$\begin{aligned}
\frac{\partial}{\partial \hat{\theta}} \sum_{i=1}^n (x_i - \hat{\theta})^2 &= -2 \sum_{i=1}^n (x_i - \hat{\theta}) \\
&= -2 \left(\sum_{i=1}^n x_i - \sum_{i=1}^n \hat{\theta} \right) \\
&= -2 \left(\sum_{i=1}^n x_i - n\hat{\theta} \right) = 0
\end{aligned}$$

We can then solve for $\hat{\theta}$ to find the LS estimator: $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X}$.

2. What is the mean of $\hat{\theta}$? Is your $\hat{\theta}$ unbiased? Besides assuming $\mathbb{E}[x_i|\theta] = \theta$, is there any other assumption on the pdf of $x_i|\theta$ involved in finding this quantity? Are any assumptions regarding your sample $\{x_i\}_{i=1}^n$ involved?

Exercise 2.2: Solution

Solution:

$$
\begin{aligned}
\mathbb{E}[\hat{\theta}|\theta] &= \mathbb{E}[\bar{X}|\theta] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i|\theta\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i|\theta\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i|\theta] \\
&= \frac{1}{n}\sum_{i=1}^{n}\theta \\
&= \frac{1}{n}n\theta = \theta
\end{aligned}
$$

$\hat{\theta}$ is therefore unbiased. We do not need any additional assumptions on the distribution of X, other than knowing the mean of X. We do not need to assume anything about our sample either.

3. What is the variance of $\hat{\theta}$? Besides assuming $\mathbb{E}[x_i|\theta] = \theta$, is there any other assumption on the pdf of $x_i|\theta$ involved in finding this quantity? Are any assumptions regarding your sample $\{x_i\}_{i=1}^{n}$ involved?

Solution:

$$
\begin{aligned}
\mathbb{V}[\hat{\theta}|\theta] &= \mathbb{V}[\bar{X}|\theta] \\
&= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} x_i|\theta\right] \\
&= \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} x_i|\theta\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left[x_i|\theta\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_U^2 \\
&= \frac{1}{n^2}n\sigma_U^2 \\
&= \frac{\sigma_U^2}{n}
\end{aligned}
$$

We only need to know the first and second moments of X (i.e. θ and σ_X^2), not its whole pdf. Also, in this case, we need to assume a random sample to compute the variance of the estimator $\hat{\theta}$. We made use of this assumption in the fourth passage above, by showing that the variance of the sum of random independent variables equals the sum of the variances.

4. Suppose $u_i \sim U(-5, 5)$, i.e. u_i is distributed Uniformly between -5 and 5. Using a statistical software

Exercise 2.2: Solution

of your choice, write a program that, given a choice of n and θ simulates a dataset $\{x_i\}_{i=1}^n$. Fix $\theta = 175$ and $n = 10$. Compute $\hat{\theta}$ using this simulated data. Is your estimate close to the true value of $\theta = 175$? What happens under a dataset with $n = 100$ observations? What happens if you take $n = 1000$?

Solution:

```
rm(list = ls())

# Set random number seed
# (it ensures that we always get the same quasi-random results, making our analysis replicable)
set.seed(2024)

# Set sample size
n <- 10

# Set population mean
theta <- 175

# Generate n-sized random sample from Uniform distribution on [-5,5] interval
u <- runif(n, -5, 5)

# Generate random sample for x_i
x <- theta + u

# Compute sample mean
mean(x)
```

```
## [1] 175.4093
```

The sample mean is rather close to the population mean.

```
# Set sample size to 100
n <- 100

# Draw Uniform random sample
u <- runif(n,-5,5)

# Generate random sample for x_i
x <- theta + u

# Compute mean
mean(x)
```

```
## [1] 174.847
```

The sample gets closer to the population mean.

Exercise 2.2: Solution

```r
# Repeat everything with sample size = 1000
n <- 1000
u <- runif(n,-5,5)
x <- theta + u
mean(x)
```

```
## [1] 175.0886
```

The sample mean gets even closer to the population mean. We can see that, as the sample size grows to infinity, the estimator $\hat{\theta}$ converges in probability to the population mean.

5. Now let's use the program you wrote to analyze the behavior of $\hat{\theta}$ in repeated sampling.
   (a) simulate $M = 100$ different datasets of size $n = 10$: $\{\{x_i^m\}_{i=1}^n\}_{m=1}^M$
   (b) for each dataset $\{x_i^m\}_{i=1}^n$, compue the LS-point estimator $\hat{\theta}^m$
   (c) plot a histogram of $\{\hat{\theta}^m\}_{m=1}^M$

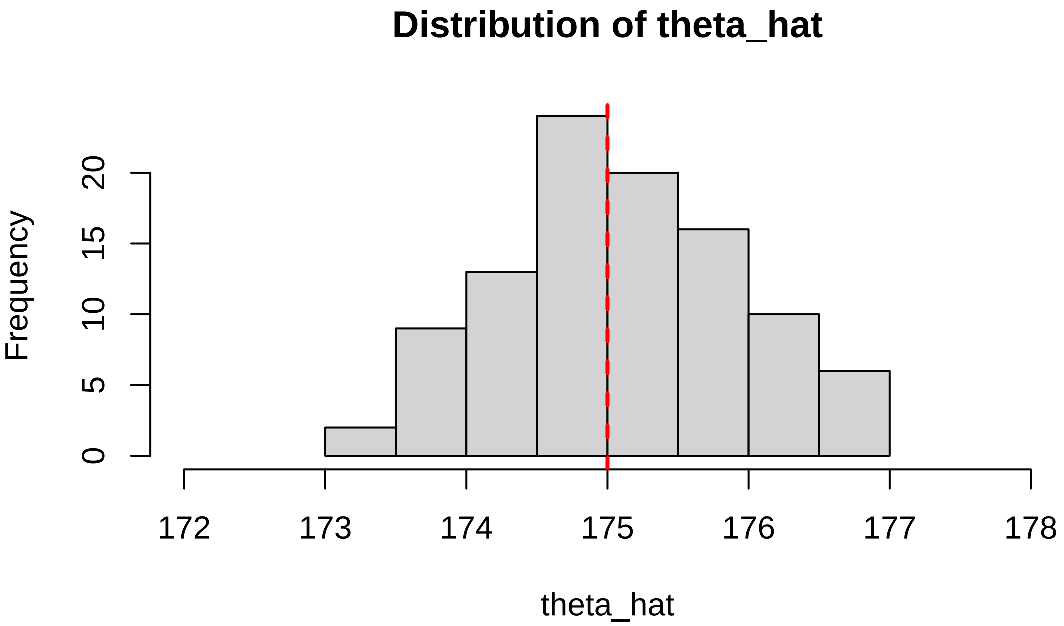
   Comment on the histogram (distribution) of $\hat{\theta}$-values. Is it in line with your expectations, based on the calculations you did for the questions above?

Solution:

```r
rm(list = ls())
set.seed(2024)

# Set the number of replications
M <- 100

# Set sample size for every replication
n <- 10

# Set population mean
theta <- 175

# Generate empty vector of length M
# (Here we store the replications)
v <- numeric(M)

# For each replication:
for (i in 1:M) {
  # Generate random sample of x_i
  u <- runif(n,-5,5)
  x <- theta + u
  # Compute sample mean and store it
  v[i] <- mean(x)
}

hist(v, main = "Distribution of theta_hat", xlab = "theta_hat", xlim = c(172,178))
```

Exercise 2.2: Solution

```r
abline(v = 175, col = "red", lwd = 2, lty = 2)
```

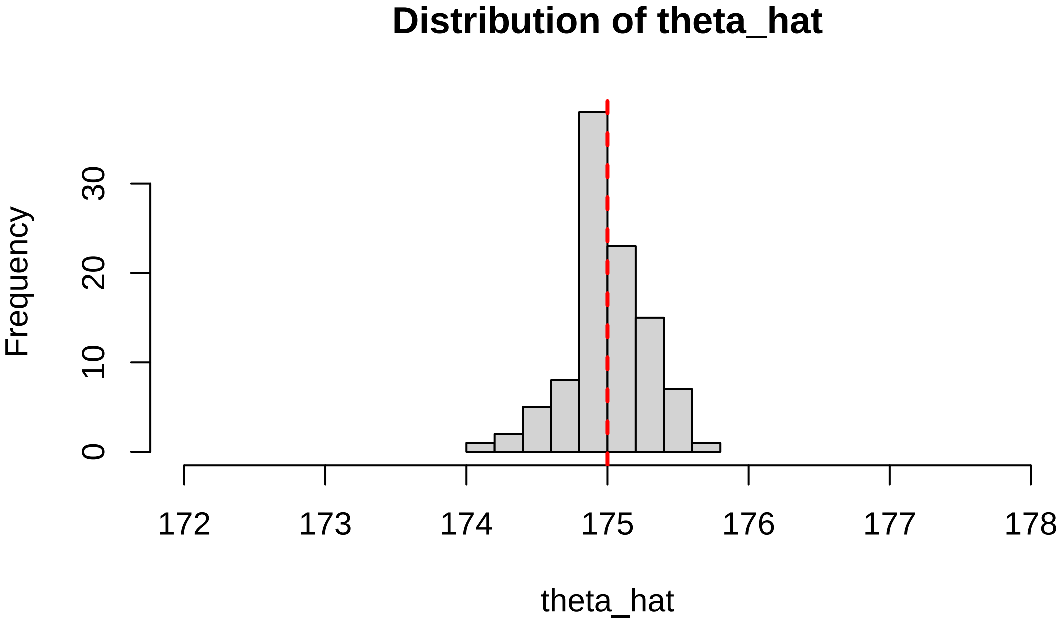


6. Redo the previous exercise using $n = 100$ as well as $n = 1000$. How does the histogram (distribution) of $\hat{\theta}$ change? Relate this to the theoretical analysis we conducted in class.

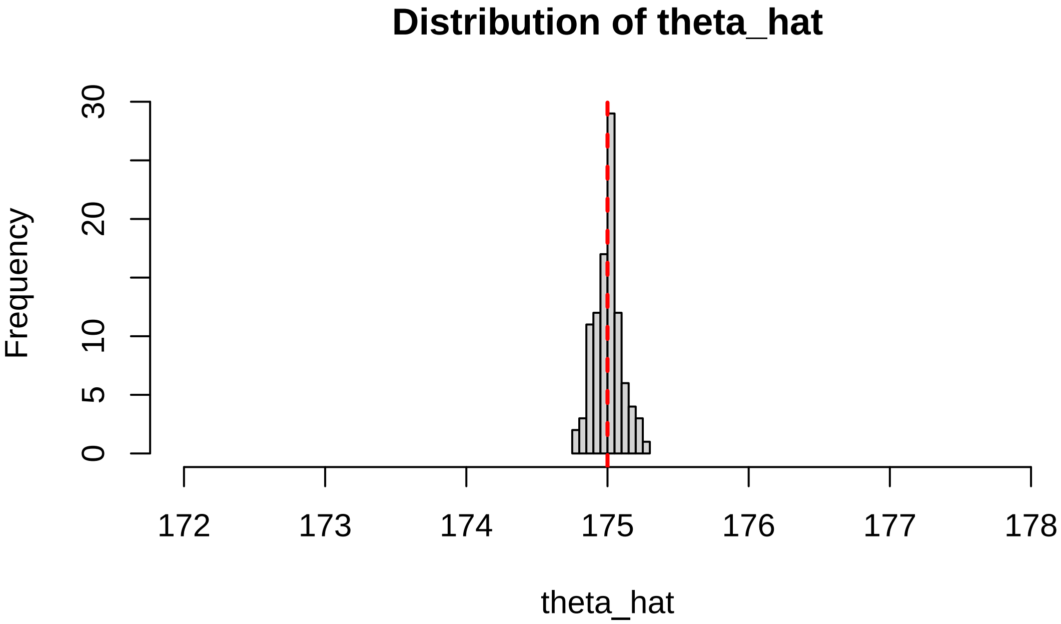
Solution:

```r
# Redo with sample size n = 100
n <- 100

for (i in 1:M) {
  u <- runif(n,-5,5)
  x <- theta + u
  v[i] <- mean(x)
}

hist(v, main = "Distribution of theta_hat", xlab = "theta_hat", xlim = c(172,178))
abline(v = 175, col = "red", lwd = 2, lty = 2)
```

Exercise 2.2: Solution



```r
# Redo with sample size n = 1000
n <- 1000

for (i in 1:M) {
  u <- runif(n,-5,5)
  x <- theta + u
  v[i] <- mean(x)
}

hist(v, main = "Distribution of theta_hat", xlab = "theta_hat", xlim = c(172,178))
abline(v = 175, col = "red", lwd = 2, lty = 2)
```



Exercise 2.2: Solution

As the sample size increases, the mean of the estimator $\hat{\theta}$ gets more tightly distributed around the true mean θ. As the sample size approaches infinity, the estimator eventually collapses on a spike corresponding to the population mean.

Exercise 2.3: Solution

Suppose $\mathbb{E}[X] = \theta$ and $\mathbb{V}[X] = \sigma^2$ for some known σ^2. Suppose we observe n i.i.d. observations of the random variable X, denoted by $\{x_i\}_{i=1}^n$.

1. Define and derive the Ordinary Least Squares (OLS) estimator of θ, $\hat{\theta}_{OLS}$.

Solution:

$$\hat{\theta}_{OLS} := \arg\min_{\theta \in \Theta} \sum_{i=1}^n (x_i - \theta)^2 = \arg\min_{\theta \in \Theta} (x - \theta)'(x - \theta) \ .$$

Using scalar-notation, we get the First Order Condition (FOC)

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta} \sum_{i=1}^n (x_i - \theta)^2 &= -\, 2 \sum_{i=1}^n (x_i - \theta) \\ &= -\, 2 \left(\sum_{i=1}^n x_i - \sum_{i=1}^n \theta \right) \\ &= -\, 2 (n\bar{X} - n\theta) \ , \end{aligned}$$

implying that

$$\hat{\theta}_{OLS} = \bar{X} \equiv \frac{1}{n} \sum_{i=1}^n x_i \ .$$

Alternatively, using vector notation:

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta} (x - \theta)'(x - \theta) &= -\, 2i'(x - \theta) \\ &= -\, 2x'i + 2\theta' i \\ &= -\, 2n\bar{X} + 2n\bar{\theta} \ , \end{aligned}$$

where ι is a vector of ones. This implies again $\hat{\theta}_{OLS} = \bar{X}$.

2. What can you say about the (finite sample) distribution of $\hat{\theta}_{OLS}$? Is $\hat{\theta}_{OLS}$ unbiased?

Solution: With the information provided – $\{x_i\}_{i=1}^n$ is an i.i.d. sample of n realizations ofthe RV X with $\mathbb{E}[X] = \theta$ and $\mathbb{V}[X] = \sigma^2$ –, we know the mean and variance of the OLS estimator:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_{OLS}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n x_i \right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n x_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[x_i] = \frac{1}{n} n\theta = \theta \ , \\ \mathbb{V}[\hat{\theta}_{OLS}] &= \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n x_i \right] = \frac{1}{n^2} \mathbb{V}\left[\sum_{i=1}^n x_i \right] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}[x_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \ . \end{aligned}$$

Because $\mathbb{E}[\hat{\theta}_{OLS}] = \theta$, $\hat{\theta}_{OLS}$ is indeed unbiased.

Unless we make a specific assumption on the distribution of the random sample $\{x_i\}_{i=1}^n$, we cannot determine the exact distribution of the estimator $\hat{\theta}$. For instance, if we assume that $x_i \sim N(\theta, \sigma^2)$, then also $\hat{\theta}$ is normally distributed with the above mean and variance.

3. What is the asymptotic distribution of $\hat{\theta}_{OLS}$? Is $\hat{\theta}_{OLS}$ consistent?

Solution: Given that the sample is i.i.d., we can invoke the WLLN, which states that sample averages

Exercise 2.3: Solution

converge in probability to the respective population mean:

$$\hat{\theta}_{OLS} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} \mathbb{E}[x_i] = \theta \ ,$$

proving that $\hat{\theta}_{OLS}$ is consistent.

Given that the sample is i.i.d., we can invoke the CLT, which states that the difference between the sample average and correspooonding population mean, when standardized by $\sqrt{n}$ converges to a Normal distribution:

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}[x_i]) \xrightarrow{d} N(0, \mathbb{V}[x_i]) \ ,$$

i.e.

$$\sqrt{n}(\hat{\theta}_{OLS} - \theta) \xrightarrow{d} N(0, \sigma^2) \ .$$

4. Set up the two-sided t-test with size $\alpha = 0.05$ for testing $\mathscr{H}_0 : \theta_0 = 0$ against $\mathscr{H}_1 : \theta_0 \neq 0$. More concretely, definining the test as

$$\varphi_t = \mathbf{1}\left\{T(X) < c_\alpha\right\} \ ,$$

define the test-statistic $T(X)$ and find the critical value c_α.

Hint: To find c_α, you need to use the definition of a size α-test and the distribution of $T(X)$ under $\mathscr{H}_0$. Thereby, you may not know the corresponding finite sample distribution, but you might know the asymptotic distribution, allowing you to construct an asymptotically valid test.

Solution:

The test-statistic of a two-sided t-test for testing whether the true value of θ is equal to $\theta_0 = 0$ is

$$T(x) = \left|\frac{\hat{\theta}_{OLS} - \theta_0}{\sigma/\sqrt{n}}\right| = \left|\frac{\hat{\theta}_{OLS}}{\sigma/\sqrt{n}}\right| \ .$$

Based on $\sqrt{n}(\hat{\theta}_{OLS} - \theta) \xrightarrow{d} N(0, \sigma^2)$, we can argue that $\hat{\theta}_{OLS} \overset{approx.}{\sim} N(\theta, \sigma^2/n)$ in finite samples. Similarly, under the null hypothesis that $\theta = \theta_0$,

$$\frac{\sqrt{n}}{\sigma}(\hat{\theta}_{OLS} - \theta_0) = \frac{\hat{\theta}_{OLS} - \theta_0}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1) \ ,$$

based on which we can argue that $\frac{\hat{\theta}_{OLS} - \theta_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$ already in finite samples.

(Alternatively, if $\theta = \tilde{\theta}$ is the true value, then we would get

$$\frac{\hat{\theta}_{OLS} - \theta_0}{\sigma/\sqrt{n}} \xrightarrow{d} N(\tilde{\theta} - \theta_0, 1)$$

instead.)

The size of the test is given by the probability of false rejection, i.e. the probability that $\varphi = 0$ when the true

Exercise 2.3: Solution

θ is indeed equal to $\theta_0 = 0$:

$$\begin{aligned}
\alpha =&\beta(\theta_0)\\
=&\mathbb{P}[\varphi = 0|\theta = \theta_0]\\
=&1 - \mathbb{P}[\varphi = 1|\theta = \theta_0]\\
=&1 - \mathbb{P}\left[|T(X)| \leq c_\alpha\right]\\
=&1 - \mathbb{P}\left[-c_\alpha \leq \frac{\hat{\theta}_{OLS} - \theta_0}{\sigma/\sqrt{n}} \leq c_\alpha \mid \theta = \theta_0\right]\\
=&1 - \mathbb{P}[-c_\alpha \leq Z \leq c_\alpha]\\
=&1 - [\Phi(c_\alpha) - \Phi(-c_\alpha)]\\
=&2(1 - \Phi(c_\alpha)) \ ,
\end{aligned}$$

where Φ is the standard Normal cdf. Under $\alpha = 0.05$, we get the critical value $c_\alpha = 1.96$.

4. Find the power function of your t-test, $\beta(\theta) = \mathbb{P}[\text{reject}|\theta] = \mathbb{P}[T(X) > c_\alpha|\theta]$.

Solution:

Power (by defining the true value of theta as $\tilde{\theta}$):

$$\begin{aligned}
\beta(\tilde{\theta}) =&\mathbb{P}[\varphi = 0|\theta = \tilde{\theta}] =\\
=&1 - \mathbb{P}\left[-c_\alpha \leq \frac{(\hat{\theta}_{OLS} - \theta_0)}{\sigma/\sqrt{n}} \leq c_\alpha \mid \theta = \tilde{\theta}\right] =\\
=&1 - \mathbb{P}\left[-c_\alpha \leq \frac{(\hat{\theta}_{OLS} - \tilde{\theta} + \tilde{\theta} - \theta_0)}{\sigma/\sqrt{n}} \leq c_\alpha \mid \theta = \tilde{\theta}\right] =\\
=&1 - \mathbb{P}\left[-c_\alpha - \frac{(\tilde{\theta} - \theta_0)}{\sigma/\sqrt{n}} \leq Z \leq c_\alpha - \frac{(\tilde{\theta} - \theta_0)}{\sigma/\sqrt{n}}\right] =\\
=&1 - \left[\Phi\left(c - \frac{\tilde{\theta} - \theta_0}{\sigma/\sqrt{n}}\right) - \Phi\left(-c - \frac{\tilde{\theta} - \theta_0}{\sqrt{n}}\right)\right]
\end{aligned}$$

5. Find the critical values associated with $\alpha = 0.1$ and $\alpha = 0.01$ as well. Write a program that plots the power function $\beta(\theta)$ as a function of θ for the three different values of α. Set $\sigma = 2$. How does the shape of $\beta(\theta)$ change across the different sizes α?

Solution:

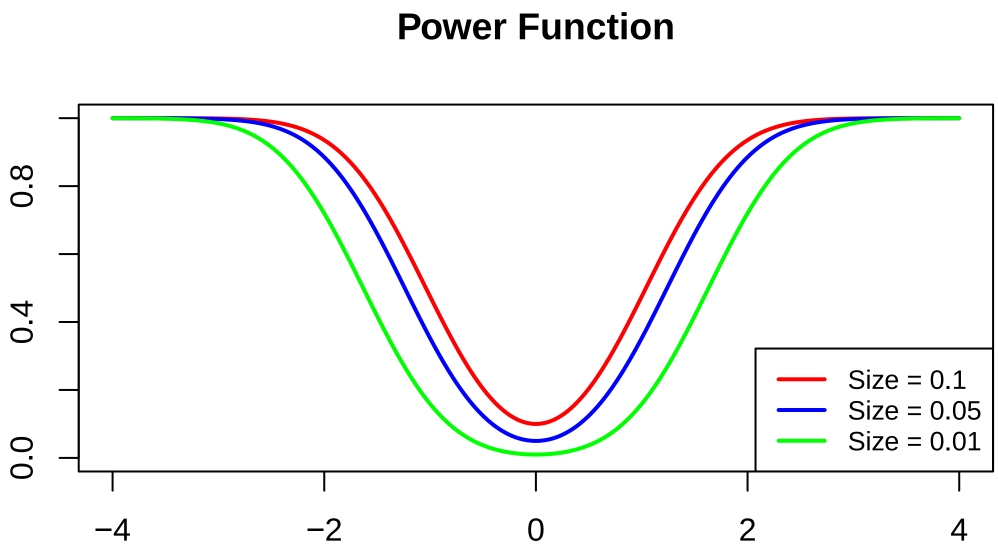
```r
rm(list = ls())

#domain of graph, i.e. variable theta_tilde, lying on x-axis
theta_tilde <- seq(-4,4,length.out = 1000)

#mean of x (theta_0)
theta_0 <- 0

#standard deviation of x (sigma)
```

Exercise 2.3: Solution

```
sigma <- 2

#sample size
n <- 10

#helper variable entering the power function with the critical values
vHelp <- (theta_tilde - theta_0)/(sigma/sqrt(n))

#compute sizes by substituting the vector
#c+(theta_tilde-theta_0)/(sigma/sqrt(n))
#into a standard normal cdf (see exercise solution above)

#power with alpha = 0.1
size01 <- 1-(pnorm(1.645+vHelp, mean = 0, sd = 1) - pnorm(-1.645+vHelp, mean = 0, sd = 1))

#power with alpha = 0.05
size005 <- 1-(pnorm(1.96+vHelp, mean = 0, sd = 1) - pnorm(-1.96+vHelp, mean = 0, sd = 1))

#power with alpha = 0.01
size001 <- 1-(pnorm(2.58+vHelp, mean = 0, sd = 1) - pnorm(-2.58+vHelp, mean = 0, sd = 1))

#plot the line for size = 0.1
plot(theta_tilde,
     size01,
     type = "l",
     col = "red",
     lwd = 2,
     xlab = " ",
     ylab = " ",
     ylim = c(0, 1),
     main = "Power Function")
#add line for size = 0.05
lines(theta_tilde,
      size005,
      col = "blue",
      lwd = 2)
#add line for size = 0.01
lines(theta_tilde,
      size001,
      col = "green",
      lwd = 2)
#add legend
legend("bottomright",
       legend = c("Size = 0.1", "Size = 0.05", "Size = 0.01"),
       col = c("red", "blue", "green"),
```

Exercise 2.3: Solution

```
       lty = 1,
       lwd = 2,
       cex = 0.8)
```



6. Based on your t-test, find an expression for the 95% confidence interval for θ, $C(X)$. What happens with $C(X)$ as n increases? What happens with $C(X)$ as σ increases? Discuss.

Solution: We know this 95% CI is the set of values for θ_0 which we could accept given our estimate $\hat{\theta}_{OLS}$, i.e. it is the set of θ_0s that satisfy

$$-c_{0.05} \leq \frac{\hat{\theta}_{OLS} - \theta_0}{\sigma/\sqrt{n}} \leq c_{0.05} \ .$$

This yields

$$C(X) := \{\theta_0 \in \Theta | \varphi(x; \theta_0) = 1\} = \left[\hat{\theta}_{OLS} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\theta}_{OLS} + 1.96\frac{\sigma}{\sqrt{n}}\right] \ .$$

The width of the CI is increasing in σ and decreasing in n. Both parameters influence the standard error of the estimator $\hat{\theta}$. The higher the population variance σ^2, the higher the uncertainty around the point estimate of θ, *ceteris paribus*. On the other hand, the higher our sample size n, the more precise is our estimator $\hat{\theta}$ and the narrower the resulting CI.

7. Set $\theta_0 = 0$ and $\sigma = 2$ and simulate a sample of $n = 10$ draws x_i under $\mathscr{H}_0$. Find $C(x)$, the confidence interval associated with your draws $x = \{x_i\}_{i=1}^{10}$. Is θ_0 in your confidence interval?

Solution:

```
rm(list = ls())
set.seed(2024)

#generate random sample
n <- 10
x0 <- rnorm(n, mean=0, sd=2)
```

Exercise 2.3: Solution

```
#create grid of total values of theta_0 to test
t <- seq(-5, 5, by = 0.1)
#here we store the values of theta_0 that lie in the CI
vc <- numeric(length(t))
#here we store critical values
c005lb <- numeric(length(t)) #lower bound, i.e. 0.25 percentile
c005ub <- numeric(length(t)) #upper bound, i.e. 0.95 percentile
#select total number of random draws
M <- 1000

#for every value of theta_0 in the grid draw M random
#samples and use them to compute the critical value
for (j in seq_along(t)) {
  #each point on the grid is theta_0
  theta_0 <- t[j]
  #here we store our test statistics
  #computed from M different random samples
  v <- numeric(length(M))
  #compute the test statistic for the specific point
  #of the grid t
  test_stat0 <- (mean(x0)-theta_0)/(2/sqrt(n))

  #get critical value for each point by simulating
  #M samples and computing the distributions
  for (i in 1:M) {
    n <- 10
    x <- rnorm(n, mean=t[j], sd=2)
    test_stat <- (mean(x)-theta_0)/(2/sqrt(n))
    v[i] <- test_stat
  }
  c005lb[j] <- quantile(v, probs=0.025) #lower bound
  c005ub[j] <- quantile(v, probs=0.975) #upper bound

  #select or reject t-stat according to critical value
  if (test_stat0 >= c005lb[j] && test_stat0 <= c005ub[j]) {
    vc[j] <- 1
  }
  else {
    vc[j] <- 0
  }
}

#generate bounds of simulated CI as highest and lowest points
#over the grid by which the test statistic does not reject H_0
```

Exercise 2.3: Solution

```r
lower_bound <- t[min(which(vc == 1))]
upper_bound <- t[max(which(vc == 1))]

#display the bounds of the simulatedCI
CI95 <- c(lower_bound, upper_bound)

print(CI95)
```

```
## [1] -0.8  1.6
```

Our 95% CI will have a lower bound of -0.8 and an upper bound of 1.6. Clearly $\theta_0 = 0$ lies in the interval.

8. If you were to repeatedly draw $\{x_i\}_{i=1}^{10}$ many times, how often would you expect θ_0 to lie in $C(x)$? Find out the answer numerically as follows. Take $M = 1000$ and create an $M \times 1$ vector v. Then, for $m = 1 : M$,

- simulate $\{x_i^m\}_{i=1}^{10}$ from $N(\theta_0, \sigma^2)$, $\theta_0 = 0$, $\sigma = 2$,
- construct $C(x^m)$,
- if $\theta_0 \in C(x^m)$, then record a 1 as the mth entry of vc, otherwise record a zero.

To see how often (across M) $\theta_0 \in C(x^m)$, compute the sum of elements in vc. Is the result in line with your expectations?

Solution:

```r
rm(list = ls())
set.seed(2024)

#total number of CI simulated
K <- 1:1000
#here we store results for every CI simulated
f <- numeric(length(K))

#simulations of 1000 samples and CI
for (k in seq_along(K)){

#generate random sample
n <- 10
x0 <- rnorm(n, mean=0, sd=2)

#create grid of total values of theta_0 to test
t <- seq(-5, 5, by = 0.1)
#here we store the values of theta_0 that lie in the CI
vc <- numeric(length(t))
#here we store critical values
c005lb <- numeric(length(t)) #lower bound, i.e. 0.25 percentile
c005ub <- numeric(length(t)) #upper bound, i.e. 0.95 percentile
#select total number of random draws
```

Exercise 2.3: Solution

```
M <- 100

#for every value of theta_0 in the grid draw M random
#samples and use them to compute the critical value
for (j in seq_along(t)) {
  #each point on the grid is theta_0
  theta_0 <- t[j]
  #here we store our test statistics
  #computed from M different random samples
  v <- numeric(length(M))
  #compute the test statistic for the specific point
  #of the grid t
  test_stat0 <- (mean(x0)-theta_0)/(2/sqrt(n))

  #get critical value for each point by simulating
  #M samples and computing the distributions
  for (i in 1:M) {
    n <- 10
    x <- rnorm(n, mean=t[j], sd=2)
    test_stat <- (mean(x)-theta_0)/(2/sqrt(n))
    v[i] <- test_stat
  }
  c005lb[j] <- quantile(v, probs=0.025) #lower bound
  c005ub[j] <- quantile(v, probs=0.975) #upper bound

  #select or reject t-stat according to critical value
  if (test_stat0 >= c005lb[j] && test_stat0 <= c005ub[j]) {
    vc[j] <- 1
  }
  else {
    vc[j] <- 0
  }
}

#generate bounds of simulated CI as highest and lowest points
#over the grid by which the test statistic does not reject H_0
lower_bound <- t[min(which(vc == 1))]
upper_bound <- t[max(which(vc == 1))]

#if a simulated CI contains the true mean, store
#the entry 1 in the vector f, otherwise store 0
if (0 >= lower_bound && 0 <= upper_bound) {
  f[k] <- 1
}
else {
```

Exercise 2.3: Solution

```
  f[k] <- 0
}
}

#define and display the percentage of simulated
#CI containing the true mean
coverage_numerical <- sum(f)/max(K)

print(coverage_numerical)
```

```
## [1] 0.939
```

Here we numerically simulated a 94% confidence interval, which is fairly close to the size obtained analytically. By increasing further the number of simulations, we obtain a more accurate CI.

Exercise 2.4: Solution

Suppose you have data on the height of n female adults living in Switzerland – $\{x_i\}_{i=1}^n$ – whereby the observations in your sample are independent. Based on that, you want to estimate the average height of female adults in the whole population (i.e. the whole of Switzerland). Let this parameter of interest be denoted by θ. You can write your observations as

$$x_i = \theta + u_i \ , \quad \text{with} \quad \mathbb{E}[u_i|\theta] = 0 \ ,$$

i.e. the height of an individual i, x_i, is given by the true average height θ plus some noise u_i around it. Note that this is just another way of writing $\mathbb{E}[x_i|\theta] = \theta$. In addition, you assume that this noise u_i is Normally distributed with some known variance σ^2: $u_i \sim N(0, \sigma^2)$. Note that this – combined with the equation for x_i above – is just another way of writing $x_i \sim N(\theta, \sigma^2)$.

1. Define and derive the Maximum Likelihood (ML) estimator of θ, $\hat{\theta}_{ML}$.

Hint: You first need to derive the likelihood, i.e. the distribution of your data $\{X_i\}_{i=1}^n$ conditional on θ, $p(x|\theta)$, based on the distribution of a single observation X_i, $p(x_i|\theta)$.

Solution:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta|x) = \arg\max_{\theta \in \Theta} \prod_{i=1}^n p(x_i|\theta)$$

Where $p(x_i|\theta)$ is the pdf we assume has generated the data. In this case, it's Normal. We can therefore substitute it into the Likelihood function to obtain:

$$\mathcal{L}(\theta|x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\left(\frac{x_i - \theta}{\sigma}\right)^2\} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i - \theta)^2}{\sigma^2}\}$$

Take logs to have the log-Likelihood function:

$$l(\theta|x_i) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2$$

Take the first derivative w.r.t. θ and set it equal to zero:

$$\frac{\partial l(\theta|x)}{\partial \theta} = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \theta) = 0$$

Solve for θ to obtain:

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^n x_i \equiv \bar{X}$$

2. Set up a Likelihood Ratio (LR) test of size $\alpha = 0.05$ for testing $\mathcal{H}_0 : \theta = \theta_0$ against $\mathcal{H}_1 : \theta \neq \theta_0$, i.e. determine the test-statistic $T(X)$ and the corresponding critical value c_α.

Hint: Note that if $Y \sim N(\mu, v)$, then $(Y - \mu)/\sqrt{v} \sim N(0, 1)$ and $(Y - \mu)^2/v \sim \chi^2_1$.

Solution:

The likelihood ratio test statistic is given by:

Exercise 2.4: Solution

$$\begin{aligned}
T(X) = \frac{p(x|\hat{\theta}_{ML})}{p(x|\theta_0)} &= \frac{((2\pi\sigma^2)^{-n/2})\exp\{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\hat{\theta}_{ML})^2}{\sigma^2}\}}{((2\pi\sigma^2)^{-n/2})\exp\{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\theta_0)^2}{\sigma^2}\}} = \\
&= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left[(x_i-\hat{\theta}_{ML})^2-(x_i-\theta_0)^2\right]\} = \\
&= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left[x_i^2-2x_i\hat{\theta}+\hat{\theta}^2-x_i^2+2x_i\theta_0-\theta_0^2\right]\} = \\
&= \exp\{-\frac{1}{2\sigma^2}\left[-2\hat{\theta}\underbrace{\sum_{i=1}^n x_i}_{n\hat{\theta}} + n\hat{\theta}^2 + 2\theta_0\underbrace{\sum_{i=1}^n x_i}_{n\hat{\theta}} - n\theta_0^2\right]\} \\
&= \exp\{-\frac{1}{2\sigma^2}\left[-2n\hat{\theta}^2+n\hat{\theta}^2+2n\hat{\theta}\theta_0-n\theta_0^2\right]\} = \\
&= \exp\{-\frac{1}{2\sigma^2/n}\left[-\hat{\theta}^2+2\hat{\theta}\theta_0-\theta_0^2\right]\} = \\
&= \exp\{\frac{1}{2\sigma^2/n}(\hat{\theta}-\theta_0)^2\}
\end{aligned}$$

We accept the null hypothesis if:

$$\begin{aligned}
\varphi_{LR} = & \ \mathbf{1}\{T(x) < c\} = \\
= & \ \mathbf{1}\left\{\exp\{\frac{1}{2\sigma^2/n}(\hat{\theta}-\theta_0)^2\} < c\right\} = \\
= & \ \mathbf{1}\left\{\frac{1}{\sigma^2/n}(\hat{\theta}-\theta_0)^2 < \tilde{c}\right\}
\end{aligned}$$

Where $\tilde{c} = 2\log(c)$.

To find $\tilde{c}$ (or, equivalently, c) we set the probability of rejection given that the null is true (i.e. the probability of false rejection) equal to our desired test-size α:

$$\begin{aligned}
\alpha &\equiv \mathbb{P}\left[\text{reject}\mathscr{H}_0|\mathscr{H}_0\text{is true}\right] = \\
&= \mathbb{P}\left[\varphi(x)=0|\theta_0\right] = \\
&= \mathbb{P}\left[T(x) > c|\theta_0\right] = \\
&= \mathbb{P}\left[\exp\{\frac{1}{2\sigma^2/n}(\hat{\theta}-\theta_0)^2\} \geq c|\theta_0\right] = \\
&= \mathbb{P}\left[\frac{1}{\sigma^2/n}(\hat{\theta}-\theta_0)^2 \geq \tilde{c}|\theta_0\right]
\end{aligned}$$

Note that the estimator for theta is normally distributed centered about the true mean: $\hat{\theta} \sim N(\theta, \sigma^2/n)$. Hence, under the null hypothesis: $\hat{\theta} \sim N(\theta_0, \sigma^2/n)$. This in turn implies that:

Exercise 2.4: Solution

$$\frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{and} \quad \frac{(\hat{\theta}-\theta_0)^2}{\sigma^2/n} \sim \chi^2_1$$

Therefore: $\alpha = \mathbb{P}[W \geq \tilde{c}]$, where $W \sim \chi^2_1$. Rearranging, we get $\mathbb{P}[W \leq \tilde{c}] = 1-\alpha$. Thus, $\tilde{c}$ is the $100(1-\alpha)$-th percentile of χ^2_1. The 95th percentile of a chi-squared distribution with 1 degree of freedom is 3.84. Hence, for a size $\alpha = 5\%$-test, we take the critical value $\tilde{c}_{\alpha=0.05} = 3.84$.

3. Suppose $\sigma^2 = 6$ and you observe $n = 4$ observations, $x_1 = 178$, $x_2 = 161$, $x_3 = 168$ and $x_4 = 172$. Based on this data, can you reject $\mathcal{H}_0 : \theta = \theta_0 = 175$ (i.e. that the average height of female adults in Switzerland is 175cm)?

Solution: First compute your ML estimator:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{178+161+168+172}{4} = 169.75$$

Compute the test statistic by plugging in the numbers and compare it to the $\alpha = 0.05$ critical value of a chi-squared distribution with 1 degree of freedom:

$$T_{LR}(X) = \frac{(169.75-175)^2}{6/4} = 18.375 > 3.84 = \tilde{c}_{0.05}$$

Since the test statistic is larger than the 5% critical value, we reject the null hypothesis at the 5% significance level. (Actually, we are able to reject the null hypothesis also at the 1% critical value, which has critical value $\tilde{c}_{0.01} = 6.63$).

4. Now let's suppose you could only find the test-statistic for the LR test, $T(X)$, but not the critical value c_α. Do so numerically, i.e.
   (a) For $m = 1 : M$, with $M = 1000$,
      - draw a sample $\{x_i^m\}_{i=1}^n \sim N(\theta_0, \sigma^2)$, setting $\theta_0 = 175$, $\sigma^2 = 6$ and $n = 4$,
      - compute $T(x^m)$.

      Plot a histogram of $\{T(x^m)\}_{m=1}^M$. This is your numerical approximation of the distribution of $T(X)$ under $\mathcal{H}_0$.

   (b) Sort your draws $\{T(x^m)\}_{m=1}^M$ from lowest to largest and take the $M(1-\alpha)$th draw. This is your numerical approximation of c_α, the $100(1-\alpha)$th quantile of the distribution of $T(X)$.

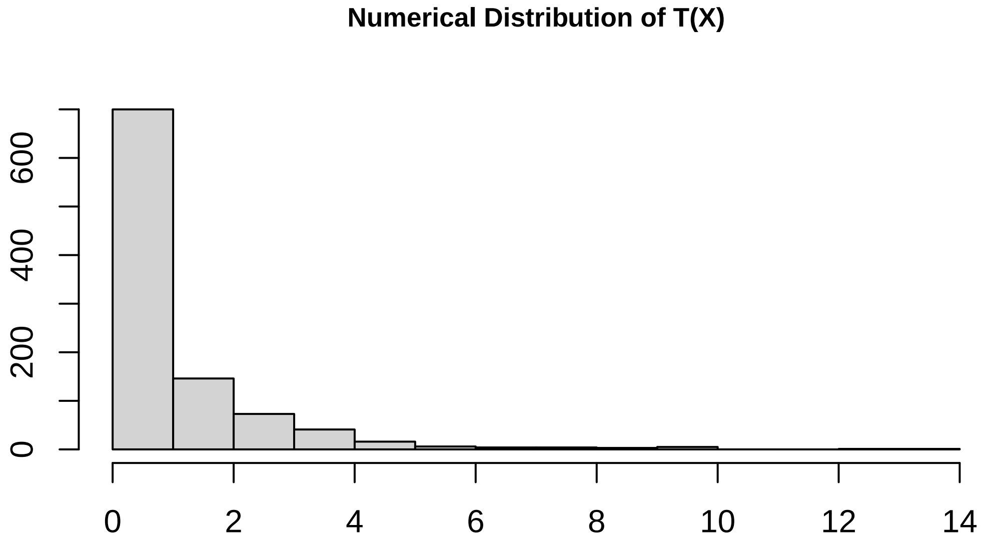
Is the value you get close to the true, analytically obtained c_α? What do you expect to happen if you take a larger value for M? Does your conclusion from the previous exercise change if you set up your test numerically as opposed to analytically?

Solution:

```r
rm(list = ls())
set.seed(2024)

M <- 1000 #total simulations
v <- numeric(length(M)) #here we store our draws

n <- 4 #set sample size
```

Exercise 2.4: Solution

```
theta_0 <- 175 #set mean under H_0
stdev <- sqrt(6) #set variance of distribution

for (i in 1:M) {
  x <- rnorm(n, mean = theta_0, sd = stdev) #draw from pdf characterised above
  theta_hat <- mean(x) #compute ML estimator
  test_stat <- ((theta_hat-theta_0)/(stdev/sqrt(n)))^2 #compute LR test statistic
  v[i] <- test_stat #save the computed T(x)
}

#plot histogram
hist(v,
     main = "Numerical Distribution of T(X)",
     cex.main = 0.8,
     xlab = " ",
     ylab = " ")
```


Numerical Distribution of T(X)

```
#find critical value at 5% significance
c005 <- quantile(v, probs = 0.95)

print(c005)
```

```
##      95%
## 3.627377
```

The numerically computed $c_{0.05}$ is 3.627. The one computed analytically is equal to $ c_{0.05} = 3.84$. The two are relatively close, but by increasing the total number of simulations M the distance between the two decreases. Overall, our results from the previous exercise would not change, as we would still reject the null hypothesis at 5%.

Exercise 2.4: Solution

5. Based on your LR-test, find a (general) expression for the 95% confidence interval for θ_0, $C(X)$. How does that interval look like if you apply it to your particular data? Is $\theta_0 = 175$ in that interval? Explain why it should (not) be.

Solution:

In the case of the LR test, the acceptance region is given by:

$$T(X) = \left(\frac{\hat{\theta}_{ML} - \theta_0}{\sigma/\sqrt{n}}\right)^2 \leq \tilde{c}_{0.05} = 3.84$$

With the following algebraic manipulations we can obtain:

$$\begin{aligned}
\left(\frac{\hat{\theta}_{ML} - \theta_0}{\sigma/\sqrt{n}}\right)^2 &\leq 3.84 \\
-\sqrt{3.84} \leq& \frac{\hat{\theta}_{ML} - \theta_0}{\sigma/\sqrt{n}} \leq \sqrt{3.84} \\
-\sigma\sqrt{3.84} \leq& \hat{\theta}_{ML} - \theta_0 \leq \sigma\sqrt{3.84} \\
-\frac{\sigma}{\sqrt{n}}\sqrt{3.84} \leq& \hat{\theta}_{ML} - \theta_0 \leq \frac{\sigma}{\sqrt{n}}\sqrt{3.84}
\end{aligned}$$

By computing $\sqrt{c_{0.05}} = \sqrt{3.84} = 1.96$ we are able to obtain the following CI:

$$C(X) = \left[\hat{\theta}_{ML} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\theta}_{ML} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

By plugging in the actual numbers $\hat{\theta} = 169.75$, $\sigma = \sqrt{6}$ and $n = 4$ we obtain:

$$C(X) = [167.35, 172.15]$$

Clearly, $\theta_0 = 175$ does not lie in the CI. It should not be in there, as by definition an $(1-\alpha)100$ CI is the set of values of θ_0 by which we are not rejecting the null hypothesis that $\theta = \theta_0$ at the α significance level. Since we rejected $\mathscr{H}_0$ that $\theta = 175$, it does not lie in the interval.

6. Let's again suppose you were not able to analytically set up the LR test and, based on it, find $C(X)$. Find the confidence interval $C(x)$ for your sample numerically as follows. First, fix a grid $\mathcal{T}$ of values for θ_0, $\mathcal{T} = 160 : 0.1 : 180$, and create a vector vc of the same dimension as $\mathcal{T}$. Then, for each $\theta_0 \in \mathcal{T}$,
   (a) repeat the numerical procedure from above to find $c_\alpha(\theta_0)$, the (numerical approximation of the) critical value for a size $\alpha = 0.05$ test for testing $\mathscr{H}_0 : \theta = \theta_0$.
   (b) compute the LR-test-statistic $T(x; \theta_0)$ for your sample x. If $T(x;\theta_0) < c_\alpha(\theta_0)$, then $\theta_0 \in C(x)$ and you record a 1 in the corresponding entry in vc, otherwise $\theta_0 \notin C(x)$ and you record a 0.

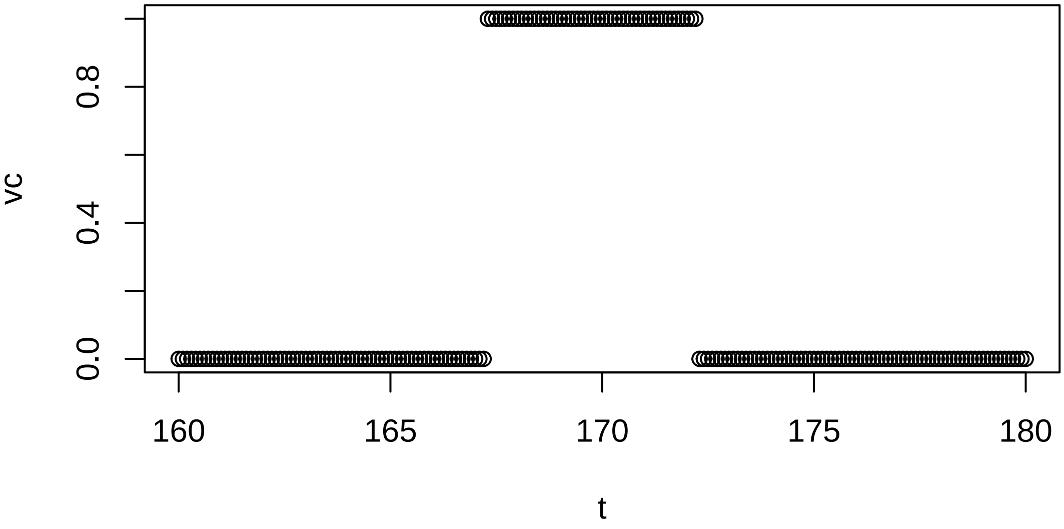
Illustrate your $C(x)$ using a scatter plot: put $\mathcal{T}$ on the x-axis and, for each value $\theta \in \mathcal{T}$, have a one on the y-axis if θ is in $C(x)$ and a zero otherwise. How does your $C(x)$ compare to the one obtained analytically?

Solution:

Exercise 2.4: Solution

```r
#compute our mean from previous exercise
theta_hat <- 169.75

#create grid of total values of theta_0 to test
t <- seq(160, 180, by = 0.1)
#here we store the values of theta_0 that lie in the CI
vc <- numeric(length(t))
#here we store critical values
c005 <- numeric(length(t))

#select total number of random draws
M <- 1000

#for every value of theta_0 in the grid draw M random
#samples and use them to compute the critical value
for (j in seq_along(t)) {
  #each point on the grid is theta_0
  theta_0 <- t[j]
  #here we store our test statistics
  #computed from M different random samples
  v <- numeric(length(M))
  #compute the test statistic for the specific point
  #of the grid t
  test_stat0 <- ((theta_hat-theta_0)/(stdev/sqrt(n)))^2

  #get critical value for each point by simulating
  #M samples and computing the distributions
  for (i in 1:M) {
    x <- rnorm(n, mean=t[j], sd=stdev)
    test_stat <- ((mean(x)-theta_0)/(stdev/sqrt(n)))^2
    v[i] <- test_stat
  }
  c005[j] <- quantile(v, probs=0.95)

  #select or reject t-stat according to critical value
  if ( test_stat0 <= c005[j]) {
    vc[j] <- 1
  }
  else {
    vc[j] <- 0
  }
}

#plot the scatter representing the CI
plot(t, vc)
```

Exercise 2.4: Solution



```
#show bounds of the computed CI
CI95 <- c(t[which(vc == 1)[1]], t[which(vc == 1)[length(which(vc == 1))]])
print(CI95)
```

```
## [1] 167.3 172.2
```

The confidence interval computed numerically is $C(X) = [167.4, 172.1]$, which is very close to the one that we found analytically. (Note that if we used a finer grid, for instance increasing by 0.05, our result would have been even more precise).

Exercise 2.5: Solution

Suppose $X|\theta \sim N(\theta, \sigma^2)$ for some known σ^2. Suppose we observe n i.i.d. observations of the random variable X, $\{x_i\}_{i=1}^n$. Suppose also your prior (on where θ lies) is $\theta \sim N(0, \tau)$.

1. Explain briefly the difference between the Bayesian paradigm and the frequentist/classical paradigm (of which OLS and ML are a part).

Solution:

Classical/Frequentist paradigm: fixed θ and and random X. Bayesian paradigm: fixed X and random θ. In the Bayesian estimation setting, considering the parameter that is the object of inference as a random variable allows us to introduce our prior beliefs of the distribution of such random variable. Imposing a prior on the distribution of θ introduces a bias in the finite-sample distribution of our estimator $\hat{\theta}$, but also decreases its variance. Hence, Bayesian estimation tries to take advantage of the bias-variance trade-off.

2. Derive the posterior distribution $p(\theta|x) \propto p(x|\theta)p(\theta)$.

Hint: You first need to determine the distribution of $\{x_i\}_{i=1}^n$ conditional on θ, $p(x|\theta)$, based on the distribution $p(X|\theta)$.

Solution:

Likelihood function:

$$p(x|\theta) \propto \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\}$$

Prior distribution:

$$p(\theta) \propto \exp\{-\frac{1}{2\tau}\theta^2\}$$

Posterior distribution:

$$\begin{aligned} p(\theta|x) \propto p(x|\theta)p(\theta) &\propto \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2 - \frac{1}{2\tau}\theta^2\} \\ &\propto \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2) - \frac{1}{2\tau}\theta^2\} \\ &\propto \exp\{-\frac{1}{2\sigma^2}\left(-2\theta\sum_{i=1}^n x_i + n\theta^2\right) - \frac{1}{2\tau}\theta^2\} \\ &\propto \exp\{\frac{\theta}{2\sigma^2}\sum_{i=1}^n x_i - \frac{\theta^2}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau}\right)\} \end{aligned}$$

Where the proportionality sign allows to eliminate all parts that depend solely on x_i.

The full distribution of the posterior is therefore normal:

$$p(\theta|x) = (\sqrt{2\pi\bar{V}})^{-1}\exp\{-\frac{1}{2\bar{V}}\bar{\theta}\}$$

With the parameters of the posterior variance $\bar{V}$ and mean $\bar{\theta}$ defined below.

Exercise 2.5: Solution

The distribution of the parameter θ conditional on X will be:

$$\theta|X \sim \mathcal{N}(\bar{\theta}, \bar{V})$$

With variance:

$$\bar{V} = \left(\frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau}}\right) = \left(\frac{n}{\sigma^2} + \frac{1}{\tau}\right)^{-1}$$

And mean:

$$\bar{\theta} = \left(\frac{n}{\sigma^2} + \frac{1}{\tau}\right)^{-1} \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i = \bar{V} \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i$$

Where the posterior mean can also be written as an average of the prior mean ($\underline{\theta} = 0$) and of the likelihood ($\hat{\theta}_{ML} = \bar{X}$) mean:

$$\begin{aligned} \bar{\theta}_B =& \frac{1}{\frac{1}{\mathbb{V}_{ML}} + \frac{1}{\tau}} \left[\frac{1}{\mathbb{V}_{ML}} \hat{\theta}_{ML} + \frac{1}{\tau} \times \theta_{prior}\right] = \\ =& \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau}} \left[\frac{n}{\sigma^2} \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{\tau} \times 0\right] = \end{aligned}$$

3. What happens with the posterior if τ changes? What happens if n or σ change?

Solution:

An increase in the prior variance τ and in the variance of X, i.e. σ^2, makes the posterior distribution flatter and lower. Intuitively, a higher degree of uncertainty on the distribution of θ makes the posterior more dispersed. On the other hand, as n rises, also the posterior increases. The intuition is that, the larger the sample size, the grater the available "information" we have to estimate θ.

4. Define the Bayes estimator to be the posterior mean, $\hat{\theta}_B = \mathbb{E}[\theta|x]$, and forget about the posterior distribution for the next two exercises; we are back in the Frequentist/Classical paradigm and we want to evaluate the point-estimator $\hat{\theta}_B$, ignoring that it was derived under the Bayesian paradigm.

What is the distribution of $\hat{\theta}_B|\theta$? Is $\hat{\theta}_B$ unbiased? How does its variance compare with the variance of $\hat{\theta}_{OLS}$?

Solution:

Considering $\bar{\theta} = \hat{\theta}_B$ as an estimator in the classical paradigm, it will be normally distributed conditionally on the true parameter θ:

$$\hat{\theta}_B|\theta \sim \mathcal{N}\left(\left(\frac{1}{\sigma^2} + \frac{1}{n\tau}\right)^{-1} \frac{\theta}{\sigma^2}, \left(\frac{1}{\sigma^2} + \frac{1}{n\tau}\right)^{-2} \frac{1}{\sigma^2 n}\right)$$

The Bayes estimator is biased, as its expected value differs from the "true" θ:

Exercise 2.5: Solution

$$\mathbb{E}[\hat{\theta}_B|\theta] = \mathbb{E}[\bar{\theta}|\theta] = \mathbb{E}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau}\right)^{-1}\frac{1}{\sigma}\sum_{i=1}^{n}x_i\right] =$$

$$= \mathbb{E}\left[\left(\frac{1}{\sigma^2}+\frac{1}{n\tau}\right)^{-1}\frac{1}{\sigma^2}\underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i}_{\bar{X}}\right] = \left(\frac{1}{\sigma^2}+\frac{1}{n\tau}\right)^{-1}\frac{1}{\sigma^2}\underbrace{\mathbb{E}[\bar{X}]}_{\theta} \neq \theta$$

And its variance is lower compared to the variance of the ML estimator:

$$\mathbb{V}[\hat{\theta}_B|\theta] = \mathbb{V}\left[\underbrace{\left(\frac{1}{\sigma^2}+\frac{1}{n\tau}\right)\frac{1}{\sigma^2}}_{<1}\hat{\theta}_{ML}\right] = \left(\frac{1}{\sigma^2}+\frac{1}{n\tau}\right)^{-2}\frac{1}{\sigma^4}\underbrace{\mathbb{V}[\hat{\theta}_{ML}]}_{\sigma^2/n}$$

$$= \left(\frac{1}{\sigma^2}+\frac{1}{n\tau}\right)^{-2}\frac{1}{\sigma^2 n} < \mathbb{V}[\hat{\theta}_{ML}]$$

We can therefore distinguish two extreme cases:

$$\mathbb{V}[\hat{\theta}_B|\theta] \to \mathbb{V}[\hat{\theta}_{ML}] \qquad \text{as} \quad \tau \to +\infty$$

$$\mathbb{V}[\hat{\theta}_B|\theta] \to 0 \qquad \text{as} \quad \tau \to 0$$

As the prior variance goes to infinity, the variance of the Bayes estimator approaches the variance of the ML estimator. As $\tau \to 0$, the variance of the Bayes estimator decreases to zero. Therefore, the parameter τ embodies the bias-variance trade-off.

5. Is $\hat{\theta}_B$ consistent? What is its asymptotic distribution, and how does it compare to that of $\hat{\theta}_{OLS}$?

Solution:

As the sample $\{x_i\}_{i=1}^n$ consists of independent and identically distributed observations, the WLLN and CLT guarantee that the Bayes estimator is consistent and asymptotically normal, with a large-sample distribution centered around the true value of θ. For fixed τ, as $n \to +\infty$ the bias disappears and the Bayes estimator approaches the OLS/ML estimator in large samples.

$$\hat{\theta}_B = \left(\frac{1}{\sigma^2}+\frac{1}{n\tau}\right)^{-1}\frac{1}{\sigma^2}\hat{\theta}_{ML} \to \theta_{ML} \qquad \text{as} \quad n \to +\infty$$

Since we know that the ML estimator is consistent, then also the Bayes estimator is consistent:

$$\hat{\theta}_{ML} \xrightarrow{p} \theta. \qquad \text{Hence,} \quad \hat{\theta}_B = \underbrace{\left(\frac{1}{\sigma^2}+\frac{1}{n\tau}\right)^{-1}\frac{1}{\sigma^2}}_{\to 1}\underbrace{\hat{\theta}_{ML}}_{\xrightarrow{p}\theta} \xrightarrow{p} \theta.$$

6. We are back in the Bayesian paradigm. Set $\theta = 0$ and $\sigma = 2$, and simulate a sample $\{x_i\}_{i=1}^{10}$. Then test

Exercise 2.5: Solution

$\mathscr{H}_0 : \theta \in [-0.5, 0.5]$ against $\mathscr{H}_1 : \theta \notin [-0.5, 0.5]$ by computing posterior odds. What do you conclude?

Solution:

```
rm(list = ls())
set.seed(2024)

n <- 10
x <- rnorm(n,mean=0,sd=2) #simulate random sample
sigma <- 2
tau <- 3 #assume arbitrary value for prior variance to allow integration
sum_x_i <- sum(x) #sum of x_i
theta_B <- sum_x_i*(tau/(n*tau+sigma^2)) #posterior mean
V_B <- (sigma^2*tau)/(n * tau + sigma^2) #posterior variance

#integrate the posterior over the H_0 interval (-0.5,0.5)
#by taking the difference of the cdf evaluated at 0.5 and -0.5

#compute P(theta lies in H_0)
p_theta_0 <- pnorm(0.5, mean=theta_B, sd=sqrt(V_B)) - pnorm(-0.5, mean=theta_B, sd=sqrt(V_B))

#compute P(theta lies in H_1)
p_theta_1 <- 1-p_theta_0

#compute posterior odds
posterior_odds <- p_theta_0/p_theta_1

print(posterior_odds)
```

```
## [1] 1.204794
```

Under these assumptions, we accept the null hypothesis $\mathscr{H}$ that $\theta \in [-0.5, 0.5]$.

3 Least Squares Estimation of the Linear Regression Model

3.1 Applied Linear Regression I

Remark: dataset `dat_CPS08.csv` required.

You can find the data set for this question and a description of the variables on Moodle. The data spreadsheet contains four variables: average hourly earnings *ahe*, age *age*, gender $female$, and education *bachelor*. To answer the questions, use the asymptotic distribution of the OLS/ML estimator to conduct hypothesis tests or generate 95% confidence intervals. In your quantitative statements, be mindful of the units (e.g. dollars vs. percentages).

(a) Do males on average earn more than females? Do individuals with a college degree earn on average more than individuals without? How large are the wage differentials?

(b) Run a regression of earnings on age, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

(c) Run a regression of the logarithm of earnings on age, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

(d) Run a regression of the logarithm of earnings on gender, education, and the logarithm of age. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

(e) Run a regression of the logarithm of earnings on age, age^2, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

(f) Plot the regression relation (the so-called age-earnings profile) between *age* (on the x-axis) and log *ahe* (on the y-axis) for the age range 20-65 using the estimates from (e) for males with a bachelor degree. At what age does the age-earnings profile peak?

(g) Is the effect of age on earnings different for males than for females? Specify and estimate a regression that you can use to answer this question. You can suppose that the relationship between age and log-earnings is linear for both males and females.
Hint: construct a covariate as the interaction $female * age$.

3.2 Applied Linear Regression II

Remark: dataset `dat_CPS08.csv` *required.*

You can find the data set for this question and a description of the variables on Moodle. The data spreadsheet contains four variables: average hourly earnings ahe, age age, gender $female$, and education $bachelor$. To answer the questions, use the asymptotic distribution of the OLS/ML estimator to conduct hypothesis tests or generate 95% confidence intervals.

(a) Run a regression of the logarithm of earnings on age, age^2, $female$, and $bachelor$. Based on your results, what are the predicted log-earnings of a 30 year old female with a bachelor degree? Note that you can write the quantity of interest as

$$\mathbb{E}[y_i|age = 30, female = 1, bachelor = 1] = \tilde{x}_i'\beta \;, \quad \text{where} \quad \tilde{x}_i = [1, 30, 30^2, 1, 1]' \;.$$

(b) Using a t-test and a significance level of $\alpha = 0.05$, can you reject the null hypoothesis that the expected hourly earnings of a 30 year old female with a bachelor degree are equal to 20 dollars per hour (i.e. that the expected log-earnings are equal to $\ln 20 \approx 2.99$)?
Hint: Note that we can write $\mathscr{H}_0 : \tilde{x}_i'\beta = 2.99$, with $\tilde{x}_i$ as defined above. Based on the (asymptotic) distribution of β, you can find that of $\tilde{x}_i'\beta$, which allows you to construct a t-test for that quantity.

(c) Using your t-test, construct a 95%-confidence interval for the expected log-earnings of a 30 year old female with a bachelor degree.

(d) Redo exercises (a) and (c) as a function of age. Concretely, plot the regression relation (the so-called age-earnings profile) between expected age (on the x-axis) and log ahe (on the y-axis) for the age range 20-65 for females with a bachelor degree, i.e. plot

$$\mathbb{E}[\log\ ahe \mid age,\ male,\ bachelor]$$

as a function of age. Also, overlay confidence bands around the age-earnings profile by plotting the 95% confidence interval for the above quantity as a function of age.

(e) Can you interpret the coefficient in front of $bachelor$ as the causal effect of obtaining a bachelor degree on earnings? Discuss.

(f) By virtue of including both age and age^2, the regression you interpreted so far assumes a non-linear relationship between age and log-earnings, and this relationship is assumed to be the same for males and females. Keeping the assumption of such a non-linear relationship between age and log-earnings, test whether this relationship is different for males and females. *Hint: construct two covariates as the interactions $female * age$ and $female * age^2$, and test whether they are jointly (!) significantly different from zero.*

3.3 Interpretation of Linear Regressions

This problem asks you to investigate some results from the paper Alesina, Giuliano & Nunn (2013): "On the Origins of Gender Roles: Women and the Plough," *Quarterly Journal of Economics,* 128(2), 469-530. We will focus on Tables 3 and 4. For some context, here is the abstract of the paper:

The study examines the historical origins of existing cross-cultural differences in beliefs and values regarding the appropriate role of women in society. We test the hypothesis that traditional agricultural practices influenced the historical gender division of labor and the evolution of gender norms. We find that, consistent with existing hypotheses, the descendants of societies that traditionally practiced plough agriculture today have less equal gender norms, measured using reported gender-role attitudes and female participation in the workplace, politics, and entrepreneurial activities. Our results hold looking across countries, across districts within countries, and across ethnicities within districts. To test for the importance of cultural persistence, we examine the children of immigrants living in Europe and the United States. We find that even among these individuals, all born and raised in the same country, those with a heritage of traditional plough use exhibit less equal beliefs about gender roles today.

For now, we focus on Table 3 (see Fig. 3.1), which contains 8 different regressions along columns. The regressions are ran using observations for different countries. You can see the description of the variables used in the regressions in the notes at the bottom of the table.

(a) What is an R^2 (R-squared)? What does the value of 0.22 in regression 1 indicate?

(b) The coefficient for the variable "Traditional plough use" in regression 1 is equal to -14.895. How do you interpret this number?

(c) The standard error corresponding to the coefficient mentioned in the previous exercise is given in parentheses below the coefficient. It is equal to 3.318. How do you interpret this number?

(d) Relative to regression 1, regression 2 adds "continent fixed effects", i.e. a dummy variable for each continent, which shows a 1 if country i is in that particular continent and a 0 otherwise. What does it mean to include such covariates in the regression?

Now focus on Table 4 (see Figs. 3.2 and 3.3), which adds two more covariates to each regression from Table 3: the logarithm of income in the year 2000 as well as the squared logarithm of income in the year 2000.

(e) Based on the results in regression 1, what is the expected change (in percentage points) in female labor force participation (in the year 2000) if income (in the year 2000) increases by 5%? Does that effect depend on the level of income in 2000?

(f) Can we credibly interpret the effect of “Traditional plough use” on the share of political positions held by women (in the year 2000) from regression 6 as causal? Discuss.

3.4 Control Variables in Linear Regressions

Suppose you are interested in estimating the effect of fertilizer on crop yields. Let $y_i > 0$ denote crop yields in USD per acre (realized in one agricultural season), and let $x_i^* > 0$ denote the amount of fertilizer applied (in liters per square meter). The unit of observation i refers to a plot of land of size one acre. Suppose y_i is determined by the following linear function:

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + u_i \ ,$$

where $r_i \in \{0, 1\}$ is an indicator for whether a plot of land is of high quality, and $g_i > 0$ is the precipation (rainfall) (measured in liters per cubic meter).

(a) Simulate a dataset of size $n = 100$ using the following Data Generating Process (DGP):

1. $u_i \sim N(0, 5)$[1]
2. $g_i \sim \text{Gamma}(2, 2)$[2]
3. $r_i = 1$ and $r_i = 0$ with equal probability
4. $x_i^*|(r_i = 1) \sim \text{Gamma}(3, 1)$ and $x_i^*|(r_i = 0) \sim \text{Gamma}(7, 1)$
5. Generate y_i by the equation above, using $\beta_0 = 400$, $\beta_1 = 5$, $\beta_2 = 200$ and $\beta_3 = 10$.

In addition, simulate two further variables: $n_i \sim N(10, 3)$ and $b_i \sim N(5 + \sqrt{x_i^*}, 3)$.

(b) Using your simulated data, run the following five regressions. For each of them, report your estimate of β_1, compare it to the true value, report its standard error, and discuss your results more generally.

1. regress y_i on x_i^* and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \text{error}_i \ .$$

2. regress y_i on x_i^*, r_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \text{error}_i \ .$$

[1] The first parameter denotes the mean, the second the variance (not the standard deviation!).

[2] The first parameter denotes the shape, the second the scale. See the following wikipedia article.

3. regress y_i on x_i^*, r_i, g_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + \text{error}_i \ .$$

4. regress y_i on x_i^*, r_i, n_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_4 n_i + \text{error}_i \ .$$

5. regress y_i on x_i^*, r_i, b_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_4 b_i + \text{error}_i \ .$$

(c) Repeat the previous questions for $M = 100$ different samples of size $n = 100$. (Concretely, simulate one dataset, run all five regressions and store their output of interest, and proceed in that way $M = 100$ times.) Show histograms of the estimators of β_1 under the five different regressions. (No need to compute its standard error.) Comment on your results.

(d) Repeat your analysis (for $M = 100$ repeated samples) by changing the following elements (one at a time) in the DGP:

- Let $x_i^*|(r_i = 1) = x_i^*|(r_i = 0) \sim \text{Gamma}(5, 1)$.
- Let $\beta_2 = 0$.
- Let $r_i = 1$ with probability 0.1.
- Let $\beta_3 = 50$.

You may restrict yourself to the first three regressions.

Exercise 3.1: Solution

You can find the data set for this question and a description of the variables on Moodle. The data spreadsheet contains four variables: average hourly earnings *ahe*, age *age*, gender *female*, and education *bachelor*. To answer the questions, use the asymptotic distribution of the OLS/ML estimator to conduct hypothesis tests or generate 95% confidence intervals. In your quantitative statements, be mindful of the units (e.g. dollars vs. percentages).

1. Do males on average earn more than females? Do individuals with a college degree earn on average more than individuals without? How large are the wage differentials?

Solution:

```r
rm(list=ls())
data <- read.csv("../ProblemsDataPapers/dat_CPS08.csv", header = TRUE)

mandata <- data[ which(data$female==0), ]
womandata <- data[ which(data$female==1), ]

coldata <- data[ which(data$bachelor==1), ]
nocoldata <- data[ which(data$bachelor==0), ]

mean(mandata$ahe)
```

```
## [1] 20.11387
```

```r
mean(womandata$ahe)
```

```
## [1] 17.48396
```

```r
mean(mandata$ahe) - mean(womandata$ahe)#difference btw genders
```

```
## [1] 2.629912
```

```r
mean(coldata$ahe)
```

```
## [1] 22.90834
```

```r
mean(nocoldata$ahe)
```

```
## [1] 15.33174
```

```r
mean(coldata$ahe)-mean(nocoldata$ahe) #difference btw degrees
```

```
## [1] 7.576594
```

On average, males earn more than females by \$ 2.63 and individuals with a college degree earn more than individuals without by \$ 7.58. To see whether the differences are significant (i.e. whether indeed we can reject the null hypothesis that the true averages in the population are the same across two groups), we can use the t-test, which is derived in the following. The analysis talks about earnings of males vs females, but the analogous applies for individuals with a college degree vs without.

Let $x_i \sim N\left(\mathbb{E}[x_i], \mathbb{V}[x_i]\right)$ be earnings of males. We thus have the finite-sample distribution of the estimator for the mean:

Exercise 3.1: Solution

$$\bar{X} \equiv \frac{1}{n_x} \sum_{i=1}^{n_x} x_i \sim N\left(\mu_x, v_x\right) \ , \quad \mu_x = \mathbb{E}[x_i] \ , \ v_x = \frac{1}{n_x} \mathbb{V}[x_i] \ , \ n_x = \# \text{ of males in sample}$$

Hence:

$$\bar{X} \overset{approx.}{\sim} N(\mu_X, \hat{v}_x) \ , \quad \hat{v}_x = \frac{1}{n_x} \widehat{\mathbb{V}[x_i]} \ , \ \widehat{\mathbb{V}[x_i]} = \frac{1}{n_x} \sum_{i=1}^{n_x} (x_i - \bar{x})^2 \ .$$

Analogously, we have $\bar{Y} \equiv \frac{1}{n_y} \sum_{i=1}^{n_y} y_i \sim N\left(\mu_y, v_y\right)$ for the earnings of females, which we can denote by y_i.

Using this result, we could test $\mathscr{H}_0 : \mu_x = \mu_{x,0}$ (some specific value, e.g. 0, or 20) against $\mathscr{H}_1 : \mu_X \neq \mu_{x,0}$. For example, we could do that using the t-test (two-sided):

$$t_x = \frac{\bar{X} - \mu_{x,0}}{\sqrt{\frac{1}{n_x} \widehat{\mathbb{V}[x_i]}}} \overset{approx.}{\sim} N(0,1) \qquad \text{(becomes exact as } \ n_X \to \infty) \ ,$$

and we could compute p-values as $2(1 - \Phi(t_X))$, where Φ is the cdf of $N(0,1)$.

However, here we want to test whether the average difference in earnings between males and females in the relevant population, $\mu_X - \mu_Y = \mathbb{E}[x_i] - \mathbb{E}[y_i]$, is zero. Thus we proceed with analogous steps, just using the distribution of the difference in sample means, $\bar{X} - \bar{Y}$, instead of that of a single sample mean, e.g. $\bar{X}$:

$$\bar{X} - \bar{Y} \sim N(\mu_{x-y}, v_{x-y}) \ , \quad \mu_{x-y} = \mu_X - \mu_Y \ , \ v_{x-y} = \frac{1}{n_x} \mathbb{V}[x_i] + \frac{1}{n_y} \mathbb{V}[y_i] \ .$$

In turn, we get

$$t_{x-y} = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\hat{v}_{x-y}}} \overset{approx.}{\sim} N(0,1) \ ,$$

where $\hat{v}_{x-y} = \frac{1}{n_x} \widehat{\mathbb{V}[x_i]} + \frac{1}{n_y} \widehat{\mathbb{V}[y_i]}$.

```
# Conduct t-tests manually:

diff1 <- mean(mandata$ahe) - mean(womandata$ahe)
varDiff1 <- var(mandata$ahe)/length(mandata$ahe) + var(womandata$ahe)/length(womandata$ahe)
tStat1 <- diff1 / sqrt(varDiff1)
pVal1 <- 2*(1 - pnorm(tStat1))
tStat1
```

```
## [1] 11.61044
```

```
pVal1
```

```
## [1] 0
```

```
diff2 <- mean(coldata$ahe) - mean(nocoldata$ahe)
varDiff2 <- var(coldata$ahe)/length(coldata$ahe) + var(nocoldata$ahe)/length(nocoldata$ahe)
tStat2 <- diff2 / sqrt(varDiff2)
pVal2 <- 2*(1 - pnorm(tStat2))
```

Exercise 3.1: Solution

```
tStat2
```

```
## [1] 34.88667
```

```
pVal2
```

```
## [1] 0
```

```
# Conduct t-tests using built-in function in R:

t.test(mandata$ahe,womandata$ahe)
```

```
##
##  Welch Two Sample t-test
##
## data:  mandata$ahe and womandata$ahe
## t = 11.61, df = 7599.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.185884 3.073939
## sample estimates:
## mean of x mean of y
##  20.11387  17.48396
```

```
t.test(coldata$ahe,nocoldata$ahe)
```

```
##
##  Welch Two Sample t-test
##
## data:  coldata$ahe and nocoldata$ahe
## t = 34.887, df = 6599.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  7.150856 8.002332
## sample estimates:
## mean of x mean of y
##  22.90834  15.33174
```

From above, we can see the averages are significantly different both for different genders and for different education levels.

2. Run a regression of earnings on age, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```
# Compute the OLS estimator manually:
# Generate constant term
data$const <- rep(1)
# Generate y vector
y <- data$ahe
```

Exercise 3.1: Solution

```
# Generate X matrix
X <- cbind(data$const, data$age, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##            [,1]
## [1,] -0.6356977
## [2,]  0.5852144
## [3,] -3.6640258
## [4,]  8.0830009
```

```
# Use the built-in lm() function
myOLSb <- lm(ahe ~ age+female+bachelor, data=data)
summary(myOLSb)
```

```
##
## Call:
## lm(formula = ahe ~ age + female + bachelor, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.139  -5.773  -1.509   4.112  57.414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6357     1.0854  -0.586    0.558
## age           0.5852     0.0362  16.165   <2e-16 ***
## female       -3.6640     0.2107 -17.391   <2e-16 ***
## bachelor      8.0830     0.2088  38.709   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.072 on 7707 degrees of freedom
## Multiple R-squared:  0.1998, Adjusted R-squared:  0.1995
## F-statistic: 641.5 on 3 and 7707 DF,  p-value: < 2.2e-16
```

The model is:

$$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_k x_{i,k} + u_i$$

The marginal effects are given by:

$$\frac{\partial y_i}{\partial x_{i,c}} = \beta_c \qquad \text{for any covariate } \; c = 1, 2, ..., k$$

Exercise 3.1: Solution

This means that $\frac{\Delta y_i}{\Delta x_{i,c}} \approx \beta_c$, or

$$\underbrace{\Delta y_i}_{\text{level change in } y_i} \approx \beta_c \underbrace{\Delta x_{i,c}}_{\text{level change in } x_i}$$

for small Δx_{i_c}, i.e. small changes in x_{i_c}. Therefore, $x_{i,c}$ is going up by 1 *unit* ($\Delta x_{i,c} = 1$) increases $y_{i,c}$ by β_c *units*. The relationship/effect is independent of the level of $x_{i,c}$. For example, age going up by 1 **year* increases earnings by 0.59 *dollars*, no matter whether age increases from 28 to 29 or from 37 to 38.

Sidenote: the relationship is not only positive in our sample, but it is significantly different from zero, i.e. we can reject the null hypothesis that the true relationship/effect is zero at any common significance level, as indicated by the tiny p-value for a t-test, shown by the lm function.

3. Run a regression of the logarithm of earnings on age, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```
# Compute the OLS estimator manually:
# Generate logs
data$log_ahe <- log(data$ahe)
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$age, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##             [,1]
## [1,]  1.87634048
## [2,]  0.02732698
## [3,] -0.18592385
## [4,]  0.42812744
```

```
# Use the built-in lm() function
myOLSc <- lm(log(ahe) ~ age+female+bachelor, data=data)
summary(myOLSc)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34755 -0.27810  0.01842  0.30954  1.66410
##
## Coefficients:
```

Exercise 3.1: Solution

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.876340   0.056160   33.41   <2e-16 ***
## age          0.027327   0.001873   14.59   <2e-16 ***
## female      -0.185924   0.010901  -17.06   <2e-16 ***
## bachelor     0.428127   0.010804   39.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7707 degrees of freedom
## Multiple R-squared:  0.2007, Adjusted R-squared:  0.2003 
## F-statistic: 644.9 on 3 and 7707 DF,  p-value: < 2.2e-16
```

Still $\frac{\partial y_i}{\partial x_{i,c}} = \beta_c$, as we use the same linear regression model. However, y_i is equal to the log of the variable we are actually interested in, average hourly earnings: $y_i = log(\tilde{y}_i)$, or $\tilde{y}_i = \exp\{y_i\}$. This means:

$$\frac{\partial \tilde{y}_i}{\partial x_{i,c}} = \frac{\partial \tilde{y}_i}{\partial y_i}\frac{\partial y_i}{\partial x_{i,c}} = \tilde{y}_i \beta_c \ ,$$

and therefore

$$\frac{\Delta \tilde{y}_i}{\Delta x_{i,c}} \approx \tilde{y}_i \beta_c \quad \iff \quad \underbrace{\frac{\Delta \tilde{y}_i}{\tilde{y}_i}}_{\text{percentage change in } \tilde{y}_i} \approx \beta_c \quad \underbrace{\Delta x_{i,c}}_{\text{dis. change in } x_{i,c}} \ .$$

Hence, $x_{i,c}$ going up by 1 *unit* ($\Delta x_{i,c} = 1$) increases $\tilde{y}_i$ by β_c *percent* ($\frac{\Delta \tilde{y}_i}{\tilde{y}_i} = \beta_c$). For example, age going up by 1 *year* increases earnings by 2.73 *percent*. Again, this percentage effect is independent of the level of $x_{i,c}$. However, note that this means that the absolute effect (in units of $\tilde{y}_i$) depends on the level of y_i. Given that (based on our regression results) an older individual earns on average (everything else equal) more than a younger individual, the expected level change in earnings measured in dollars is higher for an individual going from 37 to 38 years than from 28 to 29. If age increases from 28 to 29 (37 to 38), wage increases by approximately \$0.77 (\$1.01).

4. Run a regression of the logarithm of earnings on gender, education, and the logarithm of age. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```r
# Compute the OLS estimator manually:
# Generate logs
data$log_age <- log(data$age)
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$log_age, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y
```

Exercise 3.1: Solution

```
beta_hat
```

```
##             [,1]
## [1,] -0.03452566
## [2,]  0.80390509
## [3,] -0.18588958
## [4,]  0.42825412
```

```
# Use the built-in lm() function
myOLSd <- lm(log(ahe) ~ log(age)+female+bachelor, data=data)
summary(myOLSd)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34852 -0.27913  0.02117  0.30921  1.66325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03453    0.18622  -0.185    0.853
## log(age)     0.80391    0.05496  14.626   <2e-16 ***
## female      -0.18589    0.01090 -17.054   <2e-16 ***
## bachelor     0.42825    0.01080  39.641   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7707 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.2005
## F-statistic: 645.3 on 3 and 7707 DF,  p-value: < 2.2e-16
```

Still $\frac{\partial y_i}{\partial x_{i,c}} = \beta_c$. However, if $x_{i,c} = \log(\tilde{x}_{i,c})$, then :

$$\frac{\partial y_i}{\partial \tilde{x}_{i,c}} = \frac{\partial y_i}{\partial x_{i,c}} \frac{\partial x_{i,c}}{\partial \tilde{x}_{i,c}} = \beta_c \frac{1}{\tilde{x}_{i,c}}$$

This means that:

$$\frac{\Delta y_i}{\Delta \tilde{x}_{i,c}} \approx \beta_c \frac{1}{\tilde{x}_{i,c}} \quad \Longleftrightarrow \quad \underbrace{\Delta y_i}_{\text{level change in } y_i} = \beta_c \underbrace{\frac{\Delta \tilde{x}_{i,c}}{\tilde{x}_{i,c}}}_{\text{percentage change in } \tilde{x}_{i,c}}$$

If also $y_i = \log(\tilde{y}_i)$, then, putting the two pieces from the last and present exercises together:

$$\frac{\partial \tilde{y}_i}{\partial \tilde{x}_{i,c}} = \frac{\partial \tilde{y}_i}{\partial y_i} \frac{\partial y_i}{\partial x_{i,c}} \frac{\partial x_{i,c}}{\partial \tilde{x}_{i,c}} = \tilde{y}_i \beta_c \frac{1}{\tilde{x}_{i,c}} .$$

Exercise 3.1: Solution

This means that:

$$\frac{\partial \tilde{y}_i}{\partial \tilde{x}_{i,c}} \approx \tilde{y}_i \beta_c \frac{1}{\tilde{x}_{i,c}} \iff \underbrace{\frac{\Delta \tilde{y}_i}{y_i}}_{\text{percentage change in } \tilde{y}_i} \approx \beta_c \underbrace{\frac{\Delta \tilde{x}_{i,c}}{\tilde{x}_{i,c}}}_{\text{percentage change in } \tilde{x}_{i,c}} .$$

Hence, if $\tilde{x}_{i,c}$ increases by 1 *percent*, $\tilde{y}_i$ increases by β_c *percent*. In our case, age increasing by 1% increases $\tilde{y}_i$ by 0.8%. Therefore, when age increases from 28 to 29 (37 to 38), this is an increase of 3.57% (2.70%), and so the wage is expected to increase by 2.87% (2.17%).

5. Run a regression of the logarithm of earnings on age, age^2, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```r
# Compute the OLS estimator manually:
# Generate squares
data$agesq <- data$age^2
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$age, data$agesq, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##                [,1]
## [1,]  1.0854297723
## [2,]  0.0813724732
## [3,] -0.0009148162
## [4,] -0.1858687321
## [5,]  0.4283779959
```

```r
# Use the built-in lm() function
myOLSe <- lm(log(ahe) ~ age+I(age^2)+female+bachelor, data=data)
summary(myOLSe)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + I(age^2) + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34922 -0.27960  0.02046  0.30927  1.66268
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

Exercise 3.1: Solution

```
## (Intercept)  1.0854298  0.6382725   1.701   0.0891 .
## age          0.0813725  0.0434864   1.871   0.0614 .
## I(age^2)    -0.0009148  0.0007354  -1.244   0.2135
## female      -0.1858687  0.0109006 -17.051   <2e-16 ***
## bachelor     0.4283780  0.0108057  39.644   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7706 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.2004
## F-statistic: 484.1 on 4 and 7706 DF,  p-value: < 2.2e-16
```

Still, $\frac{\partial y_i}{\partial x_{i,c}} = \beta_c$ for any covariate c.

Now, we have the case where one of the covariates is age and one is age squared. Denoting the former covariate by $x_{i,c}$, the latter by $x_{i,d}$ and denoting age by $\tilde{x}_i$, we have $x_{i,c} = \tilde{x}_i$ and $x_{i,d} = \tilde{x}^2$. As a result:

$$\frac{\partial y_i}{\partial \tilde{x}_i} = \frac{\partial y_i}{\partial x_{i,c}} \frac{\partial x_{i,c}}{\partial \tilde{x}_i} + \frac{\partial y_i}{\partial x_{i,d}} \frac{\partial x_{i,c}}{\partial \tilde{x}_i} = \beta_c \times 1 + \beta_d \times 2\tilde{x}_i = \beta_c + 2\beta_d \tilde{x}_i \; .$$

If also $y_i = \log(\tilde{y}_i)$, following the derivations above, we have:

$$\frac{\Delta \tilde{y}_i}{\tilde{y}_i} \approx (\beta_c + 2\beta_d \tilde{x}_i)\Delta \tilde{x}_i \; ,$$

i.e., when $\tilde{x}_i$ increases by one *unit*, from $\tilde{x}_i$ to $\tilde{x}_i + 1$, then $\tilde{y}_i$ increases by $(\beta_c + \beta_d \times 2\tilde{x}_i)$ *percent*. Note that this effect depends on the level of $\tilde{x}_i$.

Therefore, when age increases from 28 to 29 (from 37 to 38) earnings are expected to increase by approximately $(\hat{\beta}_c + 2\hat{\beta}_d 28)1 = 3\%$ $((\hat{\beta}_c + 2\hat{\beta}_d 37)1 = 1.3\%)$.

6. Plot the regression relation (the so-called age-earnings profile) between *age* (on the x-axis) and log *ahe* (on the y-axis) for the age range 20-65 using the estimates from (e) for males with a bachelor degree. At what age does the age-earnings profile peak?

Solution:

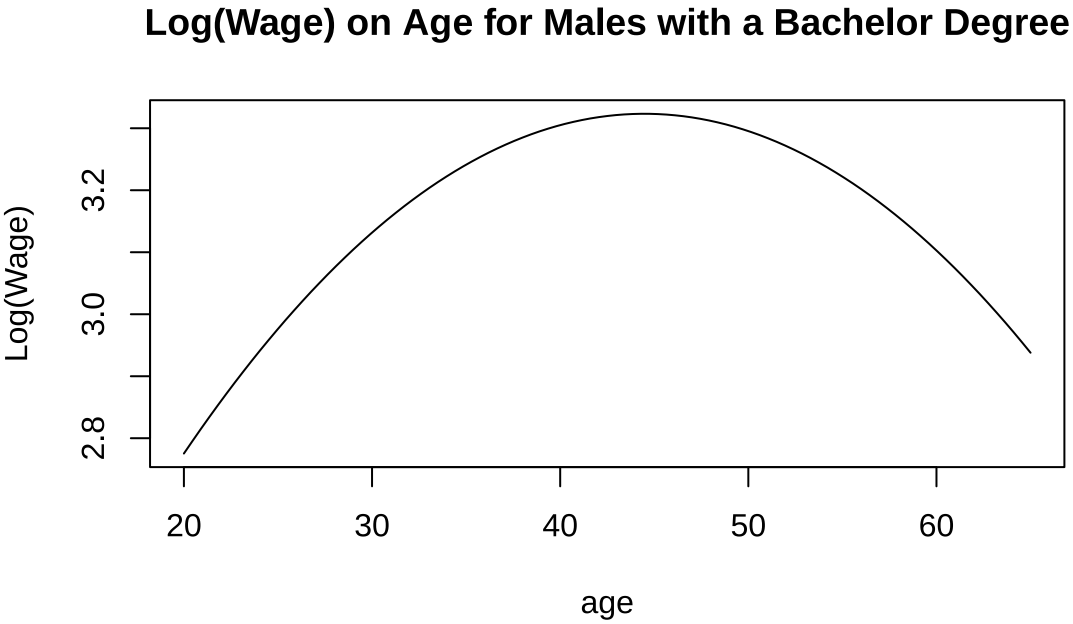
We take the estimated regression model for males (i.e. $female_i = 0$) with a bachelor degree (i.e. $bachelor_i = 1$):

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 x_i^2 + \hat{\beta}_5$$

Where $\hat{y}_i$ are the estimated log average hourly earnings, depending on the age x_i and the estimated coefficients $\hat{\beta}_j$ of the $\hat{\beta}$ vector. We construct a line plot of this function, for all the values of age between 20 to 65 on the x-axis.

```r
beta_e <- myOLSe$coefficients
curve(beta_e[1]+beta_e[2]*x+beta_e[3]*x^2+beta_e[4]*0+beta_e[5]*1,from=20,to=65,
xlab="age", ylab="Log(Wage)",
main="Log(Wage) on Age for Males with a Bachelor Degree")
```

Exercise 3.1: Solution



Since this is a globally concave function, we can find its maximum by taking the first order condition and solving for x_i:

$$\frac{\partial \hat{y}_i}{\partial x_i} = \hat{\beta}_2 + 2\hat{\beta}_3 x_i = 0 \iff x_i^{\max} = -\frac{\hat{\beta}_2}{2\hat{\beta}_3}$$

```
peak_age <- -beta_e[2]/(2*beta_e[3])
peak_age
```

```
##      age
## 44.47477
```

We can see the relation between log(wage) and age is reverse U-shaped and peaks at age 44-45.

7. Is the effect of age on earnings different for males than for females? Specify and estimate a regression that you can use to answer this question. You can suppose that the relationship between age and log-earnings is linear for both males and females.

Hint: construct a covariate as the interaction $female * age$.

Solution:

Recall the model is $y_i = x_i'\beta + u_i$. If $x_{i,c} = \text{age}$ and $x_{i,d} = \text{age} \times \text{female} = \text{age} \times \mathbf{1}\{\text{i is female}\}$, then:

$$\frac{\partial y_i}{\partial \text{age}_i} = \beta_c + \beta_d \times \mathbf{1}\{\text{i is female}\}$$

If y_i denotes the log of earnings $\tilde{y}_i$, following the derivations from the previous exercises, we get

$$\frac{\Delta \tilde{y}_i}{\tilde{y}_i} \approx (\beta_c + \beta_d \, \mathbf{1}\{\text{i is female}\})\Delta \text{age}_i$$

i.e. for males ageing by one year, earnings change by β_c percent, while for females the effect is $\beta_c + \beta_d$. Testing $\mathcal{H}_0 : \beta_d = 0$ directly tests whether the effect of age on (log-)earnings is different for males and females.

Exercise 3.1: Solution

Let's build the test statistic for a t-test. We have

$$\mathbb{E}[\hat{\beta}] = \beta, \qquad \mathbb{V}[\hat{\beta}] = \sigma^2 \mathbb{E}[X'X]^{-1} = \frac{\sigma^2}{n}\mathbb{E}[x_i x_i']^{-1}$$

(where we used the fact that $X'X = \sum_{i=1}^n x_i x_i'$) and therefore

$$\mathbb{E}[\hat{\beta}_d] = \beta_d \quad \text{and} \quad \mathbb{V}[\hat{\beta}_d] = \left(\sigma^2 \mathbb{E}[X'X]^{-1}\right)_{dd} \; .$$

(Note that the same expression for $\mathbb{V}[\hat{\beta}]$ is obtained when approximating it via the asymptotic distribution of $\hat{\beta}$.) We can estimate $\mathbb{V}[\hat{\beta}]$ and therefore $\mathbb{V}[\hat{\beta}_d]$ by

$$\hat{\mathbb{V}}[\hat{\beta}] = \frac{\hat{\sigma}^2}{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} = \hat{\sigma}^2\left(\sum_{i=1}^n x_i x_i'\right)^{-1} = \hat{\sigma}^2\left(X'X\right)^{-1} \; .$$

We thus have the test statistic:

$$t = \frac{\hat{\beta}_d - 0}{\sqrt{\hat{\mathbb{V}}[\hat{\beta}_d]}} = \frac{\hat{\beta}_d}{\sqrt{\hat{\sigma}^2\left[\left(X'X\right)^{-1}\right]_{dd}}} \xrightarrow{d} N(0,1) \; .$$

We can compute the p-value as usual as : $p = 2(1 - \Phi(t))$.

```
# Compute the OLS estimator manually:
# Generate interaction terms
data$age_female <- data$age*data$female
data$agesq_female <- data$agesq*data$female
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$age, data$age_female, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##             [,1]
## [1,]  1.66019772
## [2,]  0.03463648
## [3,] -0.01676664
## [4,]  0.30980214
## [5,]  0.42667287
```

```
# Compute test statistic manually
# Generate predicted values
y_hat <- X %*% beta_hat
# Generate residuals
u_hat <- y - y_hat
# Compute sigma_hat
```

Exercise 3.1: Solution

```r
sig_hat <- (t(u_hat) %*% u_hat)/(nrow(X)-ncol(X))
# Compute (X'X)^{-1}
XXinv <- solve( t(X) %*% X )
# Compute s.e. of beta_hat for the interaction term
se_bet_d <- sqrt(sig_hat * XXinv[3,3])
# Compute test statistic
t_stat <- beta_hat[3]/se_bet_d

t_stat
```

```
##           [,1]
## [1,] -4.442041
```

```r
# Use the built-in lm() function
myOLSh <- lm(log(ahe) ~ age+I(age*female)+female+bachelor,
data=data)
summary(myOLSh)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + I(age * female) + female + bachelor,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32731 -0.27812  0.01866  0.30528  1.68288
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.660198   0.074256  22.358  < 2e-16 ***
## age              0.034636   0.002492  13.902  < 2e-16 ***
## I(age * female) -0.016767   0.003775  -4.442 9.04e-06 ***
## female           0.309802   0.112129   2.763  0.00574 **
## bachelor         0.426673   0.010796  39.521  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4688 on 7706 degrees of freedom
## Multiple R-squared:  0.2027, Adjusted R-squared:  0.2023
## F-statistic: 489.8 on 4 and 7706 DF,  p-value: < 2.2e-16
```

We obtain a t-statistic of -4.44, which means that the negative effect of $female * age$ is significantly different from zero, i.e. the effect of age on earnings is higher for males than for females and this difference is statistically significant.

Exercise 3.2: Solution

You can find the data set for this question and a description of the variables on Moodle. The data spreadsheet contains four variables: average hourly earnings *ahe*, age *age*, gender *female*, and education *bachelor*. To answer the questions, use the asymptotic distribution of the OLS/ML estimator to conduct hypothesis tests or generate 95% confidence intervals.

1. Run a regression of the logarithm of earnings on age, age^2, $female$, and $bachelor$. Based on your results, what are the predicted log-earnings of a 30 year old female with a bachelor degree? Note that you can write the quantity of interest as

$$\mathbb{E}[y_i|age = 30, female = 1, bachelor = 1] = \tilde{x}_i'\beta \ , \quad \text{where} \quad \tilde{x}_i = [1, 30, 30^2, 1, 1]' \ .$$

Solution:

```
# Import data
rm(list = ls())
data <- read.csv("../ProblemsDataPapers/dat_CPS08.csv", header = TRUE)

# Compute OLS manually
y <- log(data$ahe)
X <- cbind(rep(1), data$age, data$age^2, data$female, data$bachelor)
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta_hat
```

```
##                [,1]
## [1,]  1.0854297723
## [2,]  0.0813724732
## [3,] -0.0009148162
## [4,] -0.1858687321
## [5,]  0.4283779959
```

```
# Use built-in function
myOLSa <- lm(log(ahe) ~ age + I(age^2) + female + bachelor, data = data)
summary(myOLSa)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + I(age^2) + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34922 -0.27960  0.02046  0.30927  1.66268
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0854298  0.6382725   1.701   0.0891 .
## age          0.0813725  0.0434864   1.871   0.0614 .
## I(age^2)    -0.0009148  0.0007354  -1.244   0.2135
## female      -0.1858687  0.0109006 -17.051   <2e-16 ***
```

Exercise 3.2: Solution

```
## bachelor     0.4283780  0.0108057  39.644   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7706 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.2004
## F-statistic: 484.1 on 4 and 7706 DF,  p-value: < 2.2e-16
```

```
# Compute predicted value
x_fem30 <- c(1, 30, 900, 1, 1)
y_fem30 <- t(x_fem30) %*% beta_hat
y_fem30
```

```
##          [,1]
## [1,] 2.945779
```

The log average hourly wage for 30-year old females with a bachelor's degree is 2.95. This amounts to \$19.1.

2. Using a t-test and a significance level of $\alpha = 0.05$, can you reject the null hypothesis that the expected hourly earnings of a 30 year old female with a bachelor degree are equal to 20 dollars per hour (i.e. that the expected log-earnings are equal to $\ln 20 \approx 2.99$)?

Hint: Note that we can write $\mathcal{H}_0 : \tilde{x}_i'\beta = 2.99$, with $\tilde{x}_i$ as defined above. Based on the (asymptotic) distribution of β, you can find that of $\tilde{x}_i'\beta$, which allows you to construct a t-test for that quantity.

Solution:

To test $\mathcal{H}_0 : \hat{y}_{fem30} = 2.99$ against the alternative $\mathcal{H}_1 : \hat{y}_{fem30} \neq 2.99$, we use the test statistic:

$$T(X) = \left| \frac{\hat{y}_{fem30} - 2.99}{SE(\hat{y}_{fem30})} \right| \sim N(0,1)$$

The standard error of a predicted value in a linear regression model is given by:

$$SE(\hat{y}_i) = SE(\tilde{x}_i'\hat{\beta}) = \sqrt{\widehat{\mathbb{V}[\tilde{x}_i'\hat{\beta}]}} = \sqrt{\tilde{x}_i'\widehat{var[\hat{\beta}]}\tilde{x}_i} = \sqrt{\tilde{x}_i'\hat{\sigma}^2(X'X)^{-1}\tilde{x}_i}\ ,$$

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

```
# Construct prediction error for y_fem30
y_hat <- X %*% beta_hat # Predicted values
u_hat <- y - y_hat # Sample residuals
sigma2 <- (t(u_hat) %*% u_hat)/(nrow(X)-ncol(X)) # Variance of error term
SE_y_fem30 <- sqrt(t(x_fem30) %*% solve(t(X) %*% X) %*% x_fem30) * sqrt(sigma2)

# Construct test statistic
t_stat <- abs((y_fem30-2.99)/SE_y_fem30)
t_stat
```

```
##          [,1]
```

Exercise 3.2: Solution

```
## [1,] 3.953751
```

The value of the test statistic is above the critical value of 1.96, so we reject the null hypothesis at the 5% significance level that the average hourly wage for women with a bachelor's degree is equal $20.

3. Using your t-test, construct a 95%-confidence interval for the expected log-earnings of a 30 year old female with a bachelor degree.

Solution:

Given the t-test with size $\alpha = 0.05$:

$$\varphi(X) = \mathbf{1}\left\{\left|\frac{\hat{y}_{fem30} - y}{SE(\hat{y}_{fem30})}\right| \leq 1.96\right\}$$

We can construct the 95% CI by solving for all the values of y that allow us to accept the null hypothesis $\mathcal{H}_0 : y = y_{fem30}$:

$$\begin{aligned}
\left|\frac{y - \hat{y}_{fem30}}{SE(\hat{y}_{fem30})}\right| &\leq 1.96 \\
-1.96 \leq \frac{y - \hat{y}_{fem30}}{SE((\hat{y}_{fem30}))} &\leq 1.96 \\
-1.96 \times SE(\hat{y}_{fem30}) \leq y - \hat{y}_{fem30} &\leq 1.96 \times SE(\hat{y}_{fem30}) \\
\hat{y}_{fem30} - 1.96 \times SE(\hat{y}_{fem30}) \leq y &\leq \hat{y}_{fem30} + 1.96 \times SE(\hat{y}_{fem30})
\end{aligned}$$

```
# 95% CI for 30-y females with bachelor
CI_fem30 <- c(y_fem30-1.96*SE_y_fem30 , y_fem30+1.96*SE_y_fem30)
CI_fem30
```

```
## [1] 2.923857 2.967701
```

We thus have that: $CI_{95\%}^{fem30} = [2.924, 2.968] = [\$18.62, \$19.45]$.

4. Redo exercises (1) and (3) as a function of age. Concretely, plot the regression relation (the so-called age-earnings profile) between expected *age* (on the x-axis) and log *ahe* (on the y-axis) for the age range 20-65 for females with a bachelor degree, i.e. plot

$$\mathbb{E}[\log\ ahe \mid age,\ male,\ bachelor]$$

as a function of *age*. Also, overlay confidence bands around the age-earnings profile by plotting the 95% confidence interval for the above quantity as a function of *age*.

Solution:

So far we have considered point- and interval-estimation of the log average hourly earnings for 30-year old females with a bachelor, i.e.

Exercise 3.2: Solution

$$\hat{y}_i = \tilde{x}_i' \hat{\beta}$$

for a specific $\tilde{x}_i = [1, 30, 30^2, 1, 1]$. Now we want to conduct our estimation for many different ages. To do this, we consider the point estimate

$$\hat{y}_i = \tilde{x}_i' \hat{\beta}, \qquad \tilde{x}_i \equiv [1, age, age^2, 1, 1]'$$

for many different values of age, and the confidence interval

$$[\hat{y}_i - 1.96 \times SE(\hat{y}_i) \ , \ \hat{y}_i + 1.96 \times SE(\hat{y}_i)] \ ,$$

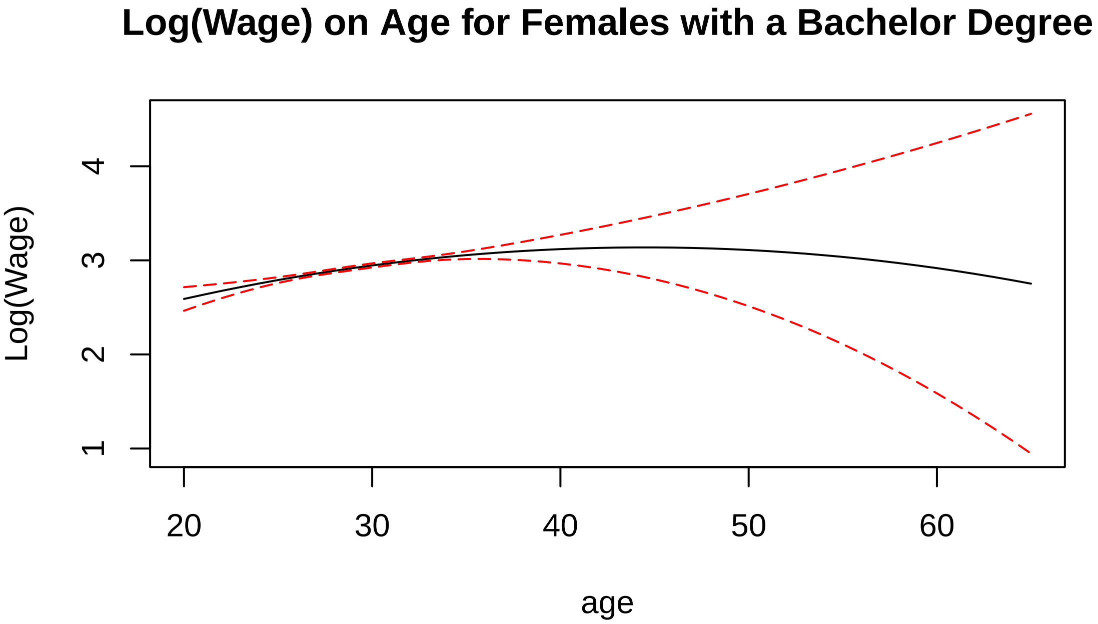
whereby

$$SE(\hat{y}_i) = \sqrt{\tilde{x}_i' \hat{\sigma}^2 (X'X)^{-1} \tilde{x}_i}$$

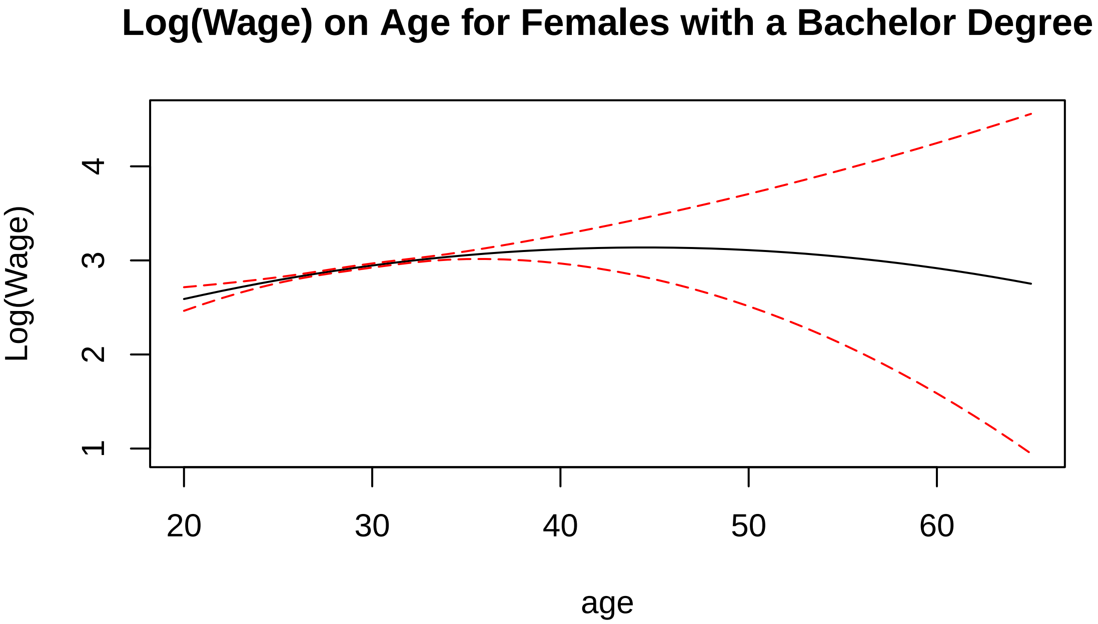
for $x_i = [1, age, age^2, 1, 1]'$, as defined above. Hence, both the point estimate and the confidence interval will be a function of age, and we are predicting the wage of females with a bachelor for all ages from 20 to 65 years.

```r
# Compute prediction manually
age <- seq(20,65,by=1) # Generate ages from 20 to 65
x_ages <- cbind(rep(1, length(age)), age, age^2,
                    rep(1, length(age)), rep(1, length(age)))
y_ages <- x_ages %*% beta_hat # Compute point estimate
# Compute SE
se_ages <- sqrt(diag(x_ages%*%solve(t(X)%*%X)%*%t(x_ages)))*sqrt(sigma2)
y_ages_LB <- y_ages-1.96*se_ages
y_ages_UB <- y_ages+1.96*se_ages

# Plot the predition with 95% CI
plot(age,y_ages,xlab="age", ylab="Log(Wage)", type="l",
ylim=c(min(y_ages_LB),max(y_ages_UB)),
main="Log(Wage) on Age for Females with a Bachelor Degree")
lines(age,y_ages_LB,lty=5,col="red")
lines(age,y_ages_UB,lty=5,col="red")
```

Exercise 3.2: Solution



```
# Could also have used the predict() function
data1 <- data.frame(age=age,female=rep(1,length(age)),
                    bachelor=rep(1,length(age)))
y_ages <- predict(myOLSa, newdata = data1, interval = 'confidence')
plot(age,y_ages[,1],xlab="age", ylab="Log(Wage)", type="l",
ylim=c(min(y_ages[,2]),max(y_ages[,3])),
main="Log(Wage) on Age for Females with a Bachelor Degree")
lines(age,y_ages[,2],lty=5,col="red")
lines(age,y_ages[,3],lty=5,col="red")
```



5. Can you interpret the coefficient in front of *bachelor* as the causal effect of obtaining a bachelor degree on earnings? Discuss.

Exercise 3.2: Solution

Solution:

No because omitted variables like ability are correlated with the regressors. Therefore, A3 fails and we have omitted variable bias, which leads to inconsistent estimators. More precisely, if our estimated model is

$$y_i = x_i'\beta + u_i, \qquad x_i = [1, age_i, age_i^2, female_i, bachelor_i]' \ ,$$

but the true Data Generating Process (DGP) is

$$y_i = x_i'\beta + z_i'\delta + v_i, \qquad z_i = [ability_i] \ ,$$

then the error term in our regression is $u_i = z_i'\delta + v_i$, and it is correlated with our regressors:

$$\mathbb{E}[x_i'u_i] = \mathbb{E}[x_i'(z_i'\delta)] = \mathbb{E}[x_i'z_i']\delta \neq 0 \ ,$$

because ability is correlated with whether or not someone gets a bachelor's degree. To solve this issue, we could try to find an IV to isolate the exogenous variation in education (e.g. quarter of birth as in Angrist øKrueger, 1991).

6. By virtue of including both age and age^2, the regression you interpreted so far assumes a non-linear relationship between age and log-earnings, and this relationship is assumed to be the same for males and females. Keeping the assumption of such a non-linear relationship between age and log-earnings, test whether this relationship is different for males and females.

Hint: construct two covariates as the interactions $female * age$ and $female * age^2$, and test whether they are jointly (!) significantly different from zero.

Solution:

To test the joint hypothesis we use the Wald test. We rely on the asymptotic distribution of the OLS estimator:

$$\hat{\beta} \overset{approx.}{\sim} N\left(\beta, \frac{1}{N}\hat{\mathbb{V}}[\beta]\right), \qquad \mathbb{V}[\beta] = \hat{\sigma}^2\hat{Q}^{-1}$$

Under the null hypothesis that $\hat{\beta}_4 = \hat{\beta}_5 = 0$, the linear combination $g(\hat{\beta})$ of the estimator $\hat{\beta}$ will be asymptotically distributed as:

$$g(\hat{\beta}) \overset{approx.}{\sim} N\left(g(\beta), \frac{1}{N}\hat{\mathbb{V}}[g(\beta)]\right) \ ,$$

where

$$\hat{\mathbb{V}}[g(\hat{\beta})] \approx \frac{1}{N}G(\hat{\beta})\hat{\sigma}^2\hat{Q}^{-1}G(\hat{\beta})'$$

and

Exercise 3.2: Solution

$$g(\hat{\beta}) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \beta - \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \text{and } G(\hat{\beta}) = \frac{\partial g(\hat{\beta})}{\partial \beta'} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

The Wald test statistic will therefore be chi-squared distributed with degrees of freedom equal to 2, the number of tested restrictions:

$$T_W = N g(\hat{\beta})' [G(\hat{\beta}) \hat{\sigma}^2 \hat{Q}^{-1} G(\hat{\beta})']^{-1} g(\hat{\beta}) \sim \chi^2_m$$

```
# Run OLS manually with interaction terms
X <- cbind(rep(1), data$age, data$age^2,
           data$age*data$female, data$age^2*data$female,
           data$female, data$bachelor)
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
sigma2 <- (t(y-X%*%beta_hat) %*% (y-X%*%beta_hat))/(nrow(X)-ncol(X))

# Wald test statistic for joint hypothesis
G <- matrix(c(0, 0, 0, 1, 0, 0, 0,
              0, 0, 0, 0, 1, 0, 0),
              nrow = 2, byrow = TRUE) # Matrix of restrictions
W <- (t(G%*%beta_hat)%*%solve(G%*%solve(t(X) %*% X)%*%t(G))%*%(G%*%beta_hat))/sigma2
W
```

```
##          [,1]
## [1,] 20.74468
```

```
# Implement this by using the car package
library(car)
myOLSb <- lm(log(ahe) ~ age * female + I(age^2) * female + female + bachelor, data = data)
summary(myOLSb)
```

```
## 
## Call:
## lm(formula = log(ahe) ~ age * female + I(age^2) * female + female + 
##     bachelor, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.32985 -0.27819  0.02009  0.30440  1.68064 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       0.315174   0.853681   0.369   0.7120    
## age               0.126467   0.058118   2.176   0.0296 *  
## female            1.417720   1.283361   1.105   0.2693    
## I(age^2)         -0.001553   0.000982  -1.582   0.1138    
## bachelor          0.426972   0.010797  39.544   <2e-16 ***
```

Exercise 3.2: Solution

```
## age:female      -0.092385   0.087482  -1.056   0.2910
## female:I(age^2)  0.001278   0.001480   0.864   0.3877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4688 on 7704 degrees of freedom
## Multiple R-squared:  0.203,  Adjusted R-squared:  0.2023
## F-statistic:   327 on 6 and 7704 DF,  p-value: < 2.2e-16
```

```
linearHypothesis(myOLSb, c("age:female = 0", "female:I(age^2) = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age:female = 0
## female:I(age^2) = 0
##
## Model 1: restricted model
## Model 2: log(ahe) ~ age * female + I(age^2) * female + female + bachelor
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   7706 1697.7
## 2   7704 1693.1  2    4.5591 10.372 3.173e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Wald test statistic equals 20.745. Since the test statistic is higher than the critical value, we reject the null hypothesis that the non-linear relationship between age and log-earnings is the same for males and females at the 1% significance level. (The F-test statistic, which equals the Wald test statistic divided by the number of degrees of freedom, is equal to $20.745/2 = 10.372$. It is distributed as a Chi-squared RV with one degree of freedom, and – unsurprisingly – it leads to the same conclusion about our hypothesis as the Wald test.)

Exercise 3.3: Solution

This problem asks you to investigate some results from the paper Alesina, Giuliano & Nunn (2013): "On the Origins of Gender Roles: Women and the Plough," *Quarterly Journal of Economics,* 128(2), 469-530. We will focus on Tables 3 and 4. For some context, here is the abstract of the paper:

The study examines the historical origins of existing cross-cultural differences in beliefs and values regarding the appropriate role of women in society. We test the hypothesis that traditional agricultural practices influenced the historical gender division of labor and the evolution of gender norms. We find that, consistent with existing hypotheses, the descendants of societies that traditionally practiced plough agriculture today have less equal gender norms, measured using reported gender-role attitudes and female participation in the workplace, politics, and entrepreneurial activities. Our results hold looking across countries, across districts within countries, and across ethnicities within districts. To test for the importance of cultural persistence, we examine the children of immigrants living in Europe and the United States. We find that even among these individuals, all born and raised in the same country, those with a heritage of traditional plough use exhibit less equal beliefs about gender roles today.

For now, we focus on Table 3 (see Fig. 1), which contains 8 different regressions along columns. The regressions are ran using observations for different countries. You can see the description of the variables used in the regressions in the notes at the bottom of the table.

(a) What is an R^2 (R-squared)? What does the value of 0.22 in regression 1 indicate?

Solution: The R^2 is a statistic between 0 and 1 representing the portion of the sum of squares (variance) of the dependent variable y that is explained by the explanatory variables X under a particular model (here the linear regression model). More precisely, we can decompose the dependent variable $y_i = \hat{y}_i + \hat{u}_i$ into the predicted value $\hat{y}_i$ and the sample residual term $\hat{u}_i$. In matrix notation, this is

$$Y = \hat{Y} + \hat{U} \ .$$

Based on this, we define the total sum of squares (SST) of the dependent variable $Y'Y$, the explained sum of squares (SSE) $\hat{Y}'\hat{Y}$ and the sum of squared residuals (SSR) $\hat{U}'\hat{U}$ and one can show that

$$Y'Y = \hat{Y}'\hat{Y} + \hat{U}'\hat{U} \ .$$

The R^2 is then

$$R^2 = \frac{SSE}{SST} = \frac{\hat{y}'\hat{y}}{y'y} = 1 - \frac{SSR}{SST} = 1 - \frac{\hat{u}'\hat{u}}{y'y} \ .$$

An R^2 of 0.22 implies that 22% of the sum of squares of the dependent variable is

Exercise 3.3: Solution

explained by the explanatory variables.

(b) The coefficient for the variable "Traditional plough use" in regression 1 is equal to -14.895. How do you interpret this number?

Solution: To understand the effect, it is important to know the units of the variables *traditional plough use* and *female labor force participation in 2000*. The former is measured in percent, ranging from 0 to 1, the latter in percent, but ranging from 0 to 100. Hence, a 1 unit (or 100 percentage point (pp)) increase in the (estimated) proportion of citizens with ancestors that used the plough in pre-industrial agriculture is associated with a 15 pp decrease in the female labor force participation rate in 2000.

(c) The standard error corresponding to the coefficient mentioned in the previous exercise is given in parentheses below the coefficient. It is equal to 3.318. How do you interpret this number?

Solution: The standard error of an estimator $\hat{\beta}_j$ shows how precisely $\hat{\beta}_j$ estimates the true parameter β_j. More precisely, it shows the (estimated) standard deviation of $\hat{\beta}_j$ in repeated sampling. Even more precisely, it shows the number we would get if we were to repeatedly draw different samples from some underlying population, compute our estimate for each of them and compute the standard deviation of these different estimates.

Assuming that the estimator $\hat{\beta}_j$ is normally distributed (which we do based on the result that it is indeed Normally distributed asymptotically), the absolute value of our estimate $\hat{\beta}_j$ divided by its standard error, is equal to the t-statistic for testing whether the true value of β_j is equal to zero. In this case, we get

$$\left| \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right| = \left| \frac{-14.895}{3.318} \right| = 4.489 \ ,$$

which lets us conclude that the true β_j is significantly different from zero at all commonly used significance levels. (Similarly, we could test whether β_j is equal to any other number by subracting this number in the numerator.)

Under the same assumptions, we can construct a 95% CI for the true coefficient β_j based on the point estimate $\hat{\beta}_j$ and its standard error. We get

$$[14.895 - 1.96 \times 3.318 \ \ , \ 14.895 + 1.96 \times 3.318] = [8.268 \ \ , \ 21.522] \ .$$

(d) Relative to regression 1, regression 2 adds "continent fixed effects", i.e. a dummy

Exercise 3.3: Solution

variable for each continent, which shows a 1 if country i is in that particular continent and a 0 otherwise. What does it mean to include such covariates in the regression?

Solution: This means that we include a separate intercept for every continent, accounting for the fact that, for the same level of traditional plough usage, different continents tend to have different female labor force participation rates. To see this, suppose for simplicity that there are only two continents, A and B. Including continent dummies means estimating

$$y_i = \beta_0 + \beta_1 \, \mathbf{1}\{i \text{ is in cont. } B\} + z_i'\gamma + u_i \ ,$$

where z_i are other covariates. Because of multicollinearity, if we have an (unconditional) intercept β_0, we cannot include both continent dummies, but only one of them. Under this specification, the intercept for countries i in continent A is β_0, while that for countries in continent B is $\beta_0 + \beta_1$. (Alternatively, we could drop the constant and include both dummies, leading to the same conclusion that we have separate intercepts for each continent.)

Because we do not interact the variable *traditional plough use* with these continent dummies, we keep the assumption that the effect of traditional plough use on female labor force participation is the same for every continent. We only account for differences in levels across continents.

Now focus on Table 4 (see Figs. 2 and 3), which adds two more covariates to each regression from Table 3: the logarithm of income in the year 2000 as well as the squared logarithm of income in the year 2000.

(e) Based on the results in regression 1, what is the expected change (in pps) in female labor force participation (in the year 2000) if income (in the year 2000) increases by 5%? Does that effect depend on the level of income in 2000?

Solution: We are estimating a regression of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + z_i'\gamma + u_i \ ,$$

where x_i is the log of income in 2000 and z_i are other covariates. We can write $x_i = \log inc_i$. We get the following marginal effect of increasing x_i on y_i:

$$\frac{\partial y_i}{\partial x_i} = \beta_1 + 2\beta_2 x_i \ .$$

Exercise 3.3: Solution

Because $\partial x_i / \partial inc_i = 1/inc_i$, we get the following marginal effect of increasing inc_i on y_i:

$$\frac{\partial y_i}{\partial inc_i} = \frac{\partial y_i}{\partial x_i}\frac{\partial x_i}{\partial inc_i} = (\beta_1 + 2\beta_2 x_i)\frac{1}{inc_i} \quad \text{i.e.} \quad \Delta y_i \approx (\beta_1 + 2\beta_2 x_i)\frac{\Delta inc_i}{inc_i} \ .$$

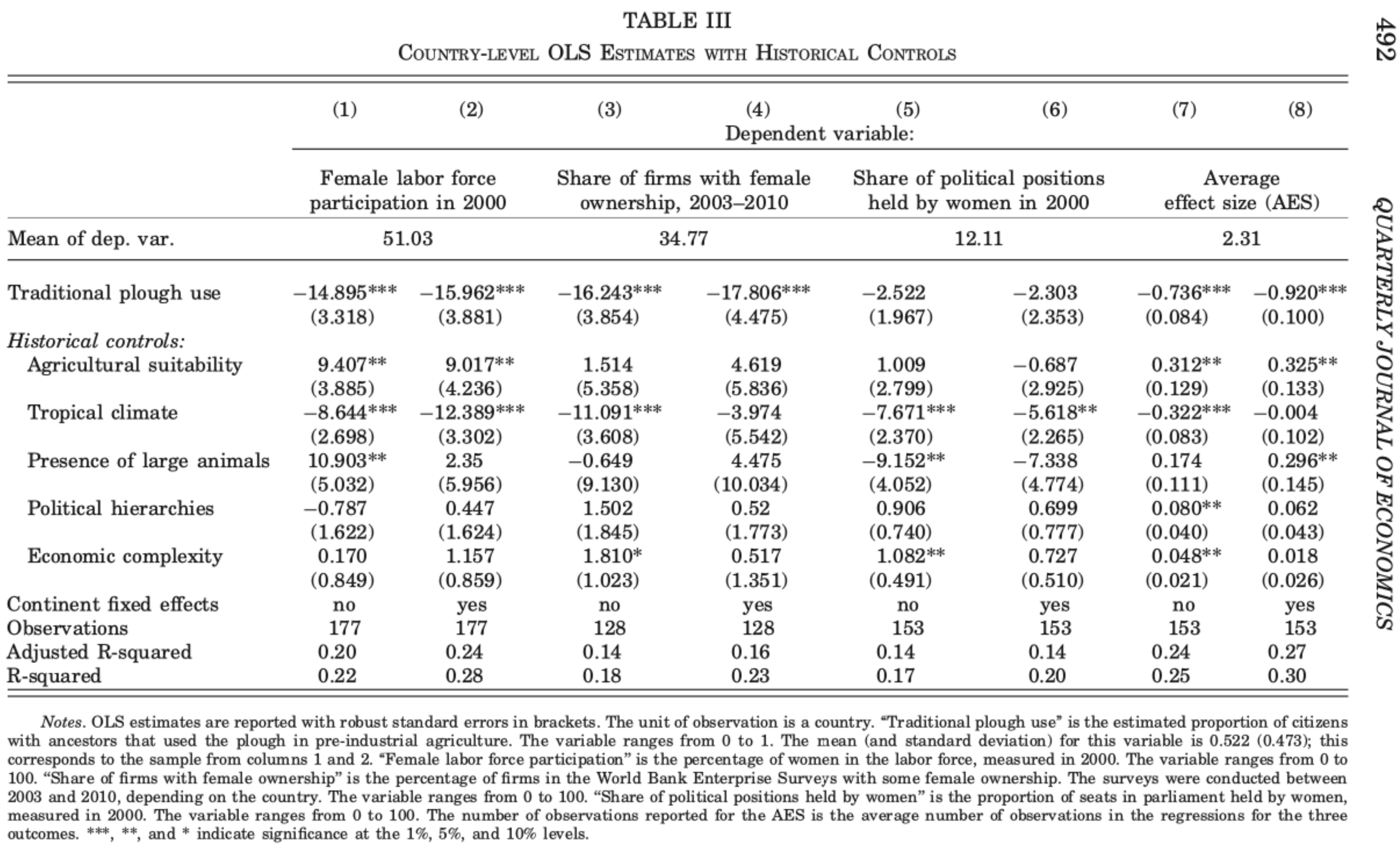
Therefore, a 5% increase in inc_i – $\Delta inc_i / inc_i = 0.05$ – is estimated to lead to a change in female labor force participation of $0.05(\hat{\beta}_1 + 2\hat{\beta}_2 x_i) = 0.05(34.61 + 2 \times 2.04 x_i)$ pps. This effect clearly depends on $x_i = \log inc_i$ and therefore inc_i, the level of income in 2000. It is stronger if the latter is higher.

(f) Can we credibly interpret the effect of "Traditional plough use" on the share of political positions held by women (in the year 2000) from regression 6 as causal? Discuss.

Solution: The estimator from the linear regression estimates the "true" causal effect of traditional plug use on female labor force participation only if the estimated model is correctly specified. More concretely, this requires that i) indeed female labor force participation y_i is a linear function of traditional plough use and the other included explanatory variables and ii) the resulting error term from such a linear model, $u_i = y_i - x_i'\beta$ (i.e. the part of y_i not accounted for linearly by these explanatory variables), is uncorrelated with these explanatory variables: $\mathbb{E}[x_i' u_i] = 0$. If the error term is correlated with at least one of the regressors, then our estimate of the effect is biased and inconsistent. This happens, for example, if we omitted some variable that affects female labor force participation y_i and is correlated with one of the regressors. It also happens if we measured traditional plough use with an error.

Taken together, the assumptions that the relationship between y_i and x_i is linear, that we did not leave out any relevant variables correlated with x_i and that there are no measurement errors appear quite restrictive. (In particular, as the authors acknowledge in the notes below the tables, the variable *traditional plough use* is estimated and therefore surely measured with error.) This prevents us to attach a causal interpretation to the estimate(s). Instead, we can take them as a partial correlation analysis that suggests that, even if various other explanatory variables are taken into account, there is still a significant (and economically meaningful) correlation between traditional plough use and female labor force participation.

Exercise 3.3: Solution



492 *QUARTERLY JOURNAL OF ECONOMICS*

TABLE III

COUNTRY-LEVEL OLS ESTIMATES WITH HISTORICAL CONTROLS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable:							
	Female labor force participation in 2000		Share of firms with female ownership, 2003–2010		Share of political positions held by women in 2000		Average effect size (AES)	
Mean of dep. var.	51.03		34.77		12.11		2.31	
Traditional plough use	−14.895***	−15.962***	−16.243***	−17.806***	−2.522	−2.303	−0.736***	−0.920***
	(3.318)	(3.881)	(3.854)	(4.475)	(1.967)	(2.353)	(0.084)	(0.100)
Historical controls:								
Agricultural suitability	9.407**	9.017**	1.514	4.619	1.009	−0.687	0.312**	0.325**
	(3.885)	(4.236)	(5.358)	(5.836)	(2.799)	(2.925)	(0.129)	(0.133)
Tropical climate	−8.644***	−12.389***	−11.091***	−3.974	−7.671***	−5.618**	−0.322***	−0.004
	(2.698)	(3.302)	(3.608)	(5.542)	(2.370)	(2.265)	(0.083)	(0.102)
Presence of large animals	10.903**	2.35	−0.649	4.475	−9.152**	−7.338	0.174	0.296**
	(5.032)	(5.956)	(9.130)	(10.034)	(4.052)	(4.774)	(0.111)	(0.145)
Political hierarchies	−0.787	0.447	1.502	0.52	0.906	0.699	0.080**	0.062
	(1.622)	(1.624)	(1.845)	(1.773)	(0.740)	(0.777)	(0.040)	(0.043)
Economic complexity	0.170	1.157	1.810*	0.517	1.082**	0.727	0.048**	0.018
	(0.849)	(0.859)	(1.023)	(1.351)	(0.491)	(0.510)	(0.021)	(0.026)
Continent fixed effects	no	yes	no	yes	no	yes	no	yes
Observations	177	177	128	128	153	153	153	153
Adjusted R-squared	0.20	0.24	0.14	0.16	0.14	0.14	0.24	0.27
R-squared	0.22	0.28	0.18	0.23	0.17	0.20	0.25	0.30

Notes. OLS estimates are reported with robust standard errors in brackets. The unit of observation is a country. "Traditional plough use" is the estimated proportion of citizens with ancestors that used the plough in pre-industrial agriculture. The variable ranges from 0 to 1. The mean (and standard deviation) for this variable is 0.522 (0.473); this corresponds to the sample from columns 1 and 2. "Female labor force participation" is the percentage of women in the labor force, measured in 2000. The variable ranges from 0 to 100. "Share of firms with female ownership" is the percentage of firms in the World Bank Enterprise Surveys with some female ownership. The surveys were conducted between 2003 and 2010, depending on the country. The variable ranges from 0 to 100. "Share of political positions held by women" is the proportion of seats in parliament held by women, measured in 2000. The variable ranges from 0 to 100. The number of observations reported for the AES is the average number of observations in the regressions for the three outcomes. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Figure 1: Table 3 from Alesina, Giuliano & Nunn (2013)

Exercise 3.3: Solution

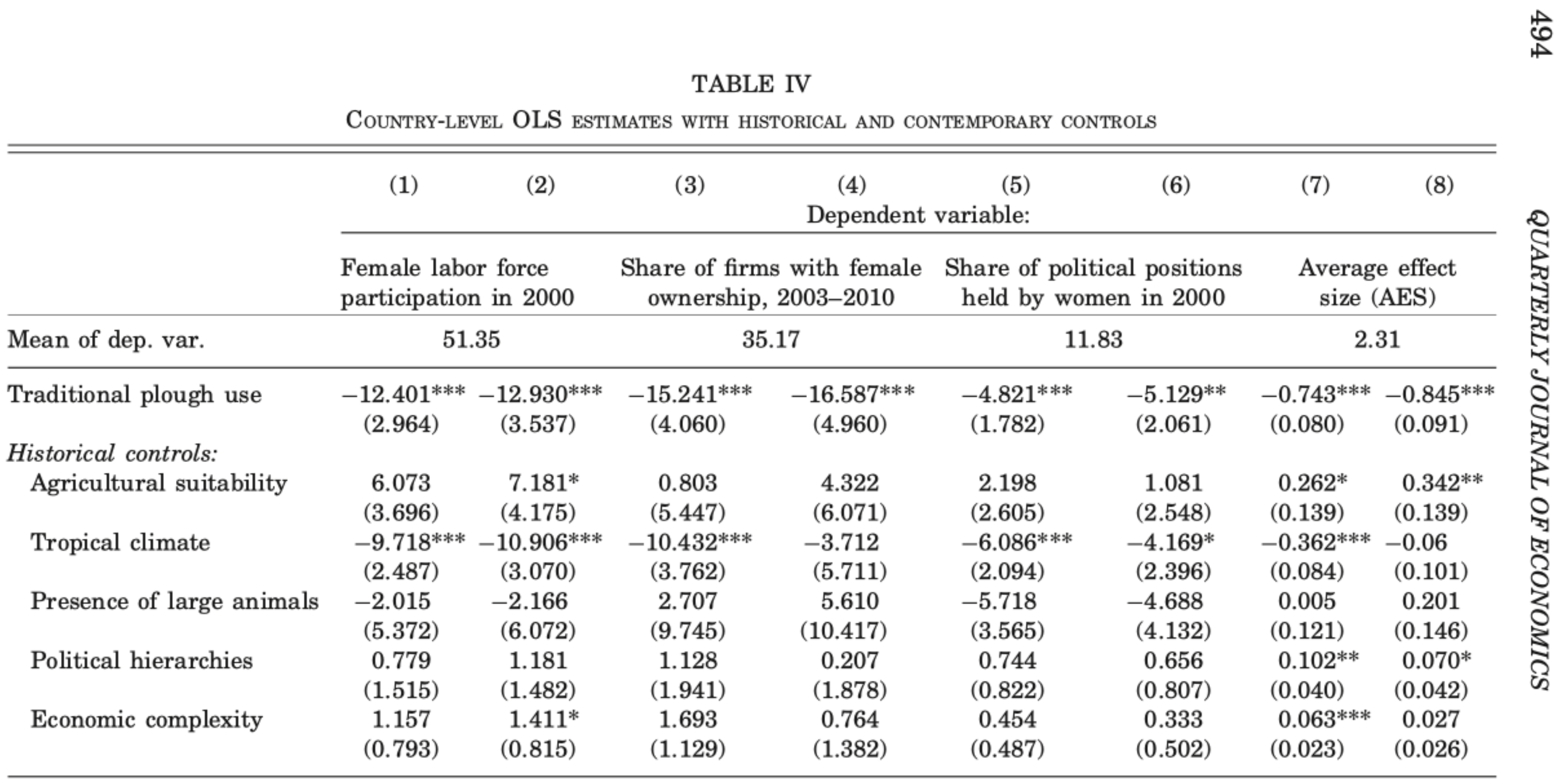


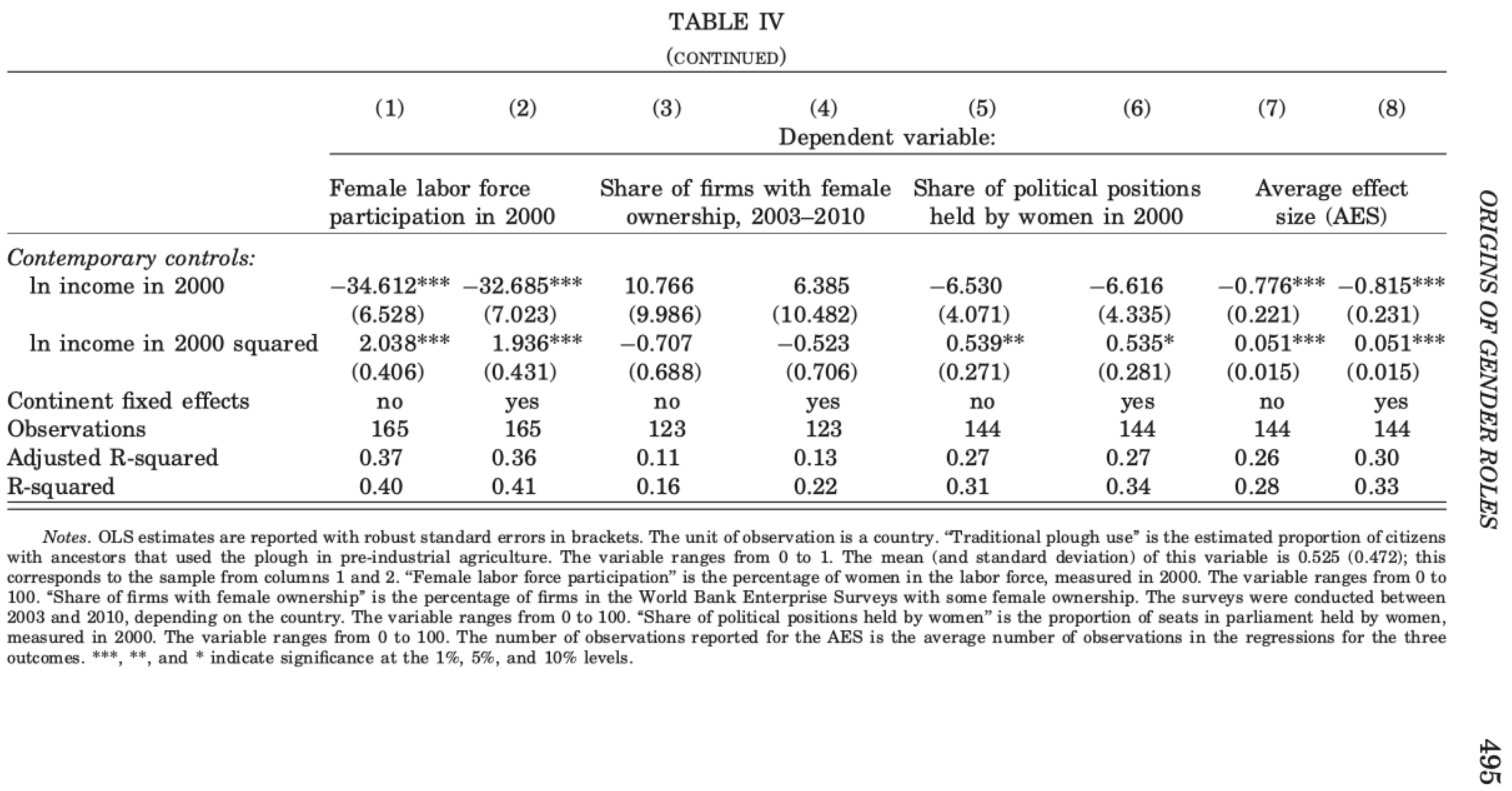
TABLE IV

COUNTRY-LEVEL OLS ESTIMATES WITH HISTORICAL AND CONTEMPORARY CONTROLS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable:							
	Female labor force participation in 2000		Share of firms with female ownership, 2003–2010		Share of political positions held by women in 2000		Average effect size (AES)	
Mean of dep. var.	51.35		35.17		11.83		2.31	
Traditional plough use	−12.401***	−12.930***	−15.241***	−16.587***	−4.821***	−5.129**	−0.743***	−0.845***
	(2.964)	(3.537)	(4.060)	(4.960)	(1.782)	(2.061)	(0.080)	(0.091)
Historical controls:								
Agricultural suitability	6.073	7.181*	0.803	4.322	2.198	1.081	0.262*	0.342**
	(3.696)	(4.175)	(5.447)	(6.071)	(2.605)	(2.548)	(0.139)	(0.139)
Tropical climate	−9.718***	−10.906***	−10.432***	−3.712	−6.086***	−4.169*	−0.362***	−0.06
	(2.487)	(3.070)	(3.762)	(5.711)	(2.094)	(2.396)	(0.084)	(0.101)
Presence of large animals	−2.015	−2.166	2.707	5.610	−5.718	−4.688	0.005	0.201
	(5.372)	(6.072)	(9.745)	(10.417)	(3.565)	(4.132)	(0.121)	(0.146)
Political hierarchies	0.779	1.181	1.128	0.207	0.744	0.656	0.102**	0.070*
	(1.515)	(1.482)	(1.941)	(1.878)	(0.822)	(0.807)	(0.040)	(0.042)
Economic complexity	1.157	1.411*	1.693	0.764	0.454	0.333	0.063***	0.027
	(0.793)	(0.815)	(1.129)	(1.382)	(0.487)	(0.502)	(0.023)	(0.026)

Figure 2: First Part of Table 4 from Alesina, Giuliano & Nunn (2013)

Exercise 3.3: Solution



ORIGINS OF GENDER ROLES 495

TABLE IV

(CONTINUED)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable:							
	Female labor force participation in 2000		Share of firms with female ownership, 2003–2010		Share of political positions held by women in 2000		Average effect size (AES)	
Contemporary controls:								
ln income in 2000	−34.612***	−32.685***	10.766	6.385	−6.530	−6.616	−0.776***	−0.815***
	(6.528)	(7.023)	(9.986)	(10.482)	(4.071)	(4.335)	(0.221)	(0.231)
ln income in 2000 squared	2.038***	1.936***	−0.707	−0.523	0.539**	0.535*	0.051***	0.051***
	(0.406)	(0.431)	(0.688)	(0.706)	(0.271)	(0.281)	(0.015)	(0.015)
Continent fixed effects	no	yes	no	yes	no	yes	no	yes
Observations	165	165	123	123	144	144	144	144
Adjusted R-squared	0.37	0.36	0.11	0.13	0.27	0.27	0.26	0.30
R-squared	0.40	0.41	0.16	0.22	0.31	0.34	0.28	0.33

Notes. OLS estimates are reported with robust standard errors in brackets. The unit of observation is a country. "Traditional plough use" is the estimated proportion of citizens with ancestors that used the plough in pre-industrial agriculture. The variable ranges from 0 to 1. The mean (and standard deviation) of this variable is 0.525 (0.472); this corresponds to the sample from columns 1 and 2. "Female labor force participation" is the percentage of women in the labor force, measured in 2000. The variable ranges from 0 to 100. "Share of firms with female ownership" is the percentage of firms in the World Bank Enterprise Surveys with some female ownership. The surveys were conducted between 2003 and 2010, depending on the country. The variable ranges from 0 to 100. "Share of political positions held by women" is the proportion of seats in parliament held by women, measured in 2000. The variable ranges from 0 to 100. The number of observations reported for the AES is the average number of observations in the regressions for the three outcomes. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Figure 3: Second Part of Table 4 from Alesina, Giuliano & Nunn (2013)

Exercise 3.4: Solution

Suppose you are interested in estimating the effect of fertilizer on crop yields. Let $y_i > 0$ denote crop yields in USD per acre (realized in one agricultural season), and let $x_i^* > 0$ denote the amount of fertilizer applied (in liters per square meter). The unit of observation i refers to a plot of land of size one acre. Suppose y_i is determined by the following linear function:

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + u_i \ ,$$

where $r_i \in \{0, 1\}$ is an indicator for whether a plot of land is of high quality, and $g_i >0 $ is the precipation (rainfall) (measured in liters per cubic meter).

(a) Simulate a dataset of size $n = 100$ using the following Data Generating Process (DGP):

1. $u_i \sim N(0, 5)$[1]
2. $g_i \sim \text{Gamma}(2, 2)$[2]
3. $r_i = 1$ and $r_i = 0$ with equal probability
4. $x_i^* | (r_i = 1) \sim \text{Gamma}(3, 1)$ and $x_i^* | (r_i = 0) \sim \text{Gamma}(7, 1)$
5. Generate y_i by the equation above, using $\beta_0 = 400$, $\beta_1 = 5$, $\beta_2 = 200$ and $\beta_3 = 10$.

In addition, simulate two further variables: $n_i \sim N(10, 3)$ and $b_i \sim N(5 + \sqrt{x_i^*}, 3)$.

Solution:

```r
rm(list = ls())
set.seed(2025)

n <- 100 # Sample size
ui <- rnorm(n,0,sqrt(5)) # Variable u_i
gi <- rgamma(n,shape=2,scale=2) # Variable g_i
ri <- rbinom(n,1,0.5) # Variable r_i
xi <- numeric(n) # Variable x_i
xi[ri == 1] <- rgamma(sum(ri == 1), shape = 3, scale = 1)
xi[ri == 0] <- rgamma(sum(ri == 0), shape = 7, scale = 1)
Betas <- c(400,5,200,10) # Vector of true params
X <- cbind(rep(1),xi,ri,gi)
yi <- X %*% Betas + ui # DGP of y_i

# Further variables
ni <- rnorm(n,10,sqrt(3))
bi <- rnorm(n,5+sqrt(xi),sqrt(3))
```

(b) Using your simulated data, run the following five regressions. For each of them, report your estimate of β_1, compare it to the true value, report its standard error, and discuss your results more generally.

[1] The first parameter denotes the mean, the second the variance (not the standard deviation!).

[2] The first parameter denotes the shape, the second the scale. See the following wikipedia article.

Exercise 3.4: Solution

1. regress y_i on x_i^* and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \text{error}_i \; .$$

2. regress y_i on x_i^*, r_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \text{error}_i \; .$$

3. regress y_i on x_i^*, r_i, g_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + \text{error}_i \; .$$

4. regress y_i on x_i^*, r_i, g_i, n_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_4 n_i + \text{error}_i \; .$$

5. regress y_i on x_i^*, r_i, g_i, b_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_4 b_i + \text{error}_i \; .$$

Solution:

```r
# 1
model1 <- lm(yi ~ xi)
summary(model1)
```

```
##
## Call:
## lm(formula = yi ~ xi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -155.828  -55.496    5.953   46.518  299.147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  657.668     16.846  39.041  < 2e-16 ***
## xi           -17.132      2.882  -5.945 4.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.73 on 98 degrees of freedom
## Multiple R-squared:  0.265,  Adjusted R-squared:  0.2575
## F-statistic: 35.34 on 1 and 98 DF,  p-value: 4.262e-08
```

Exercise 3.4: Solution

```r
summary(model1)$coefficients[2,c(1,2)]
```

```
##   Estimate Std. Error 
## -17.132144   2.881988
```

Our estimate of β_1 in Model 1 is biased and inconsistent. In fact, we are in the case of Omitted Variable Bias (OVB), where the relevant omitted covariate is r_i. Since r_i has a direct effect on the dependent variable (i.e., $\beta_2 \neq 0$) and r_i is also correlated with x_i^* (i.e., $\mathbb{E}[r_i x_i^*] \neq 0$), the residuals of our regression estimated via Model 1 are:

$$\begin{aligned} u_i^1 &= y_i - \beta_0 - \beta_1 x_i^* \\ &= \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + u_i - \beta_0 - \beta_1 x_i^* \\ &= u_i + \beta_2 r_i + \beta_3 g_i \end{aligned}$$

Clearly, we have that A3 is violated, as $\mathbb{E}[x_i^* u_i^1] = \mathbb{E}[x_i^*(u_i + \beta_2 r_i + \beta_3 g_i)] = \beta_2 \mathbb{E}[x_i^* r_i] \neq 0$ by assumption.

We can notice that our point estimate of the effect of x_i^* on y_i equals -17, which is very far from its true value of 5.

```r
#2
model2 <- lm(yi ~ xi+ri)

summary(model2)$coefficients[2,c(1,2)]
```

```
##   Estimate Std. Error 
##   4.842491   1.332335
```

Including the relevant covariate r_i allows to obtain a consistent estimator of β_i. In fact, by the same reasoning above, it can be shown that in Model 2 A3 holds, in that $\mathbb{E}[x_i^* u_i^2] = 0$, where u_i^2 are the residuals from this second regression.

From the regression output table, we can see that the point estimate of 4.9 is relatively close to its true value of 5.

At the same time, however, $\hat{\beta}_1$ estimated in Model 2 is inefficient, as there is an omitted variable g_i that has an effect on y_i. In fact, by defining the residuals of Model 2 as:

$$\begin{aligned} u_i^2 &= \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + u_i - \beta_0 - \beta_1 x_i^* - \beta_2 r_i \\ &= u_i + \beta_3 g_i \end{aligned}$$

We see that the sum of squared residuals of Model 2 is strictly larger than the SSR of the true DGP:

$$\mathbb{E}[(u^2)'(u^2)] = \mathbb{E}[(u_i + \beta_3 g_i)'(u_i + \beta_3 g_i)] = \mathbb{E}[u'u] + \beta_3^2 \mathbb{E}[g^2]$$

Since $\beta_3 \neq 0$ and $\mathbb{V}[g] \neq 0$ by assumption.

Exercise 3.4: Solution

We can see that the standard error of $\hat{\beta}_1$ in Model 2 is very large (6.15), which prevents us from rejecting the null hypothesis that $\beta_1 \neq 0$ at any significance level.

```
#3
model3 <- lm(yi ~ xi+ri+gi)

summary(model3)$coefficients[2,c(1,2)]
```

```
##   Estimate Std. Error
## 5.09327448 0.09846816
```

In Model 3, by adding also the explanatory variable g_i, we have a significant increase in precision of the estimator for β_1. Now the standard error of $\hat{\beta}_1$ is 0.8, which is almost 10 times smaller than in the previous model.

```
#4
model4 <- lm(yi ~ xi+ri+ni)

summary(model4)$coefficients[2,c(1,2)]
```

```
##   Estimate Std. Error
##   4.839456   1.340684
```

In Model 4, we are adding an irrelevant variable n_i which is uncorrelated with either of our regressors or with the explanatory variable. This is a bad control, as it increases the variance of the error term without providing any explanatory power to the model. In fact, we can see that in this case the residuals of the model become:

$$\begin{aligned}
u_i^4 &= \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + u_i - \beta_0 - \beta_1 x_i^* - \beta_2 r_i - \beta_4 n_i \\
&= u_i + \beta_3 g_i - \beta_4 n_i
\end{aligned}$$

Similarly to the case analysed for Model 2, also here ahe have a substantial increase in the SSR, thus a loss in efficiency, as long as $\hat{\beta}_4 \neq 0$ and $\mathbb{V}[n_i] \neq 0$.

We can see in the regression output table that the SE of $\hat{\beta}_1$ increases slightly to 1.34 from 1.32 in Model 2.

```
#5
model5 <- lm(yi ~ xi+ri+bi)

summary(model5)$coefficients[2,c(1,2)]
```

```
##   Estimate Std. Error
##   4.386392   1.423097
```

Also in Model 5 we are estimating a regression by adding an extra irrelevant covariate. In this case, however, we have that $\mathbb{E}[x_i^* b_i] \neq 0$. Since b_i is not correlated with the dependent variable, it does not bias our estimate of β_1. However, since it's correlated with x_i^*, it "reduces" the part of x_i^* that is available to estimate y_i.

In fact, by FWL theorem, the OLS estimator of $\hat{\beta}_j$ only employs the part of regressor x_j that is uncorrelated

Exercise 3.4: Solution

with other regressors to estimate the effect of variable x_j on y. Hence, the more collinear the regressors, the smaller the "information" that is used to identify $\hat{\beta}_j$.

This is why in this case b_i is a bad control.

(c) Repeat the previous questions for $M = 100$ different samples of size $n = 100$. (Concretely, simulate one dataset, run all five regressions and store their output of interest, and proceed in that way $M = 100$ times.) Show histograms of the estimators of β_1 under the five different regressions. (No need to compute its standard error.) Comment on your results.

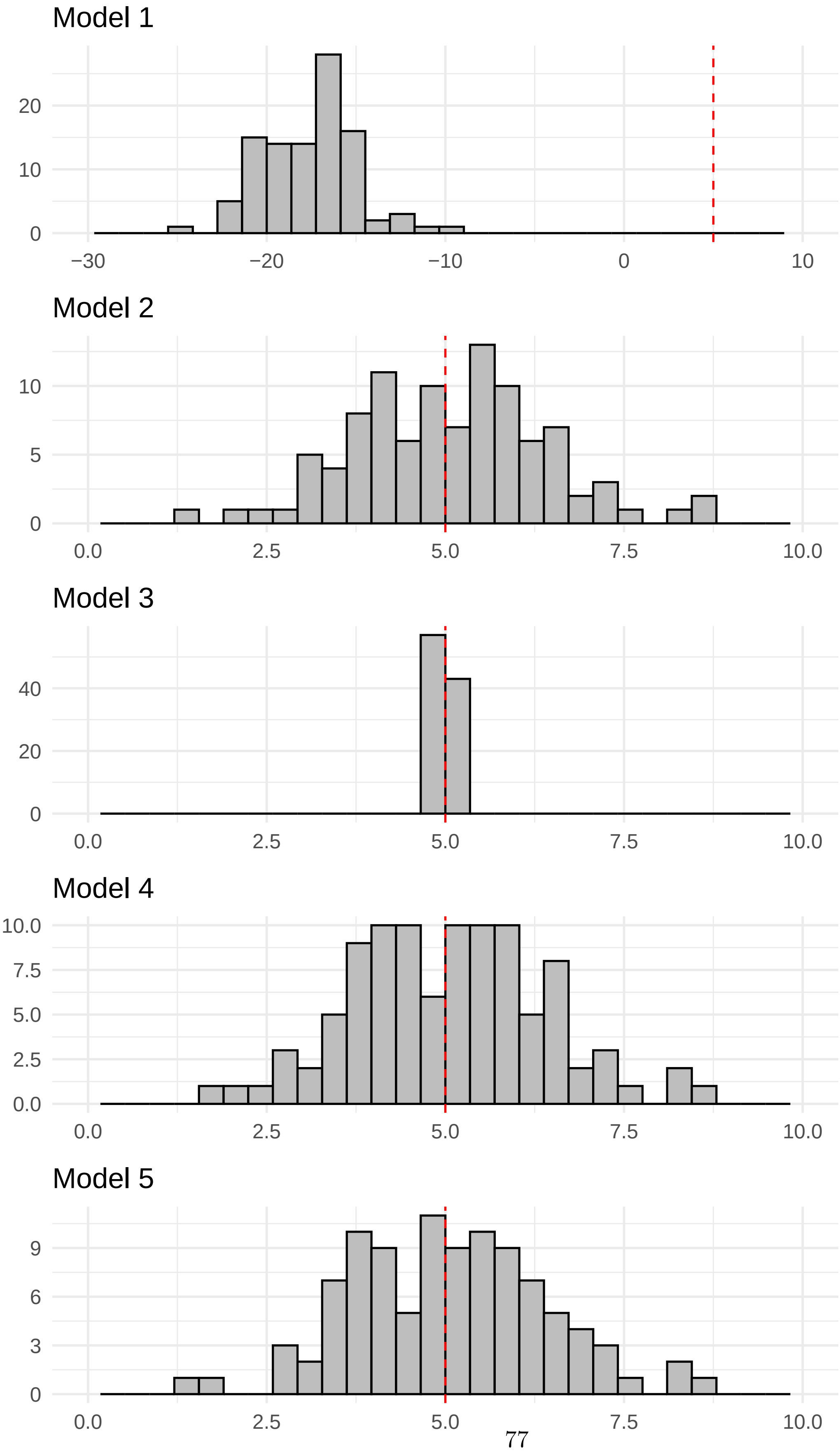
Solution:

```r
M <- 100
mBetahats <- matrix(numeric(5*M),nrow=M,ncol=5)
n <- 100

for (i in 1:M) {
# Generate random sample
ui <- rnorm(n,0,sqrt(5))
gi <- rgamma(n,shape=2,scale=2)
ri <- rbinom(n,1,0.5)
xi <- numeric(n)
xi[ri == 1] <- rgamma(sum(ri == 1), shape = 3, scale = 1)
xi[ri == 0] <- rgamma(sum(ri == 0), shape = 7, scale = 1)
Betas <- c(400,5,200,10)
X <- cbind(rep(1),xi,ri,gi)
yi <- X %*% Betas + ui
ni <- rnorm(n,10,sqrt(3))
bi <- rnorm(n,5+sqrt(xi),sqrt(3))

# Simulate 5 models and store beta1 estimates
model1 <- lm(yi ~ xi)
mBetahats[i,1] <- model1$coefficients[2]

model2 <- lm(yi ~ xi+ri)
mBetahats[i,2] <- model2$coefficients[2]

model3 <- lm(yi ~ xi+ri+gi)
mBetahats[i,3] <- model3$coefficients[2]

model4 <- lm(yi ~ xi+ri+ni)
mBetahats[i,4] <- model4$coefficients[2]

model5 <- lm(yi ~ xi+ri+bi)
mBetahats[i,5] <- model5$coefficients[2]
}
```

Exercise 3.4: Solution

```r
library(ggplot2)
library(patchwork)

# Function to create a histogram plot
create_plot <- function(data, title, x_limits) {
  ggplot(data.frame(x = data), aes(x)) +
    geom_histogram(bins = 30, fill = "gray", color = "black") +
    geom_vline(xintercept = Betas[2], color = "red", linetype = "dashed") +
    ggtitle(title) +
    scale_x_continuous(limits = x_limits) +  # Set x-axis limits
    theme_minimal(base_size = 12) +
    theme(
      panel.background = element_rect(fill = "white", color = "white"),
      plot.background = element_rect(fill = "white", color = "white"),
      axis.title.x = element_blank(),
      axis.title.y = element_blank()
    )
}

# Create individual plots
p1 <- create_plot(mBetahats[,1], "Model 1", x_limits = c(-30, 10))
p2 <- create_plot(mBetahats[,2], "Model 2", x_limits = c(0, 10))
p3 <- create_plot(mBetahats[,3], "Model 3", x_limits = c(0, 10))
p4 <- create_plot(mBetahats[,4], "Model 4", x_limits = c(0, 10))
p5 <- create_plot(mBetahats[,5], "Model 5", x_limits = c(0, 10))

# Combine the plots
p1 / p2 / p3 / p4 / p5
```

Exercise 3.4: Solution



Exercise 3.4: Solution

In the case of Model 1, the whole distribution of $\hat{\beta}_1$ is completely off its true value. We confirm that the estimator is biased and inconsistent.

Model 2: $\hat{\beta}_1$ is indeed consistent, although it has quite a high variance, as its distribution is rather spread out.

Model 3: the DGP of the model is properly identified. There the estimator for β is consistent and efficient. We see how the distribution of $\hat{\beta}_1$ is tightly clustered around its population mean.

Models 4 and 5: including an irrelevant regressor that is not correlated with other covariates increases the variance of the distribution of $\hat{\beta}$, although apparently not that much in our present case. On the other hand, the consistency property of the estimator remains unaffected as long as the correlation between y_i and n_i is zero.

(d) Repeat your analysis (for $M = 100$ repeated samples) by changing the following elements (one at a time) in the DGP:

- Let $x_i^*|(r_i = 1) = x_i^*|(r_i = 0) \sim \text{Gamma}(5, 1)$.
- Let $\beta_2 = 0$.
- Let $r_i = 1$ with probability 0.1.
- Let $\beta_3 = 50$.

You may restrict yourself to the first three regressions.

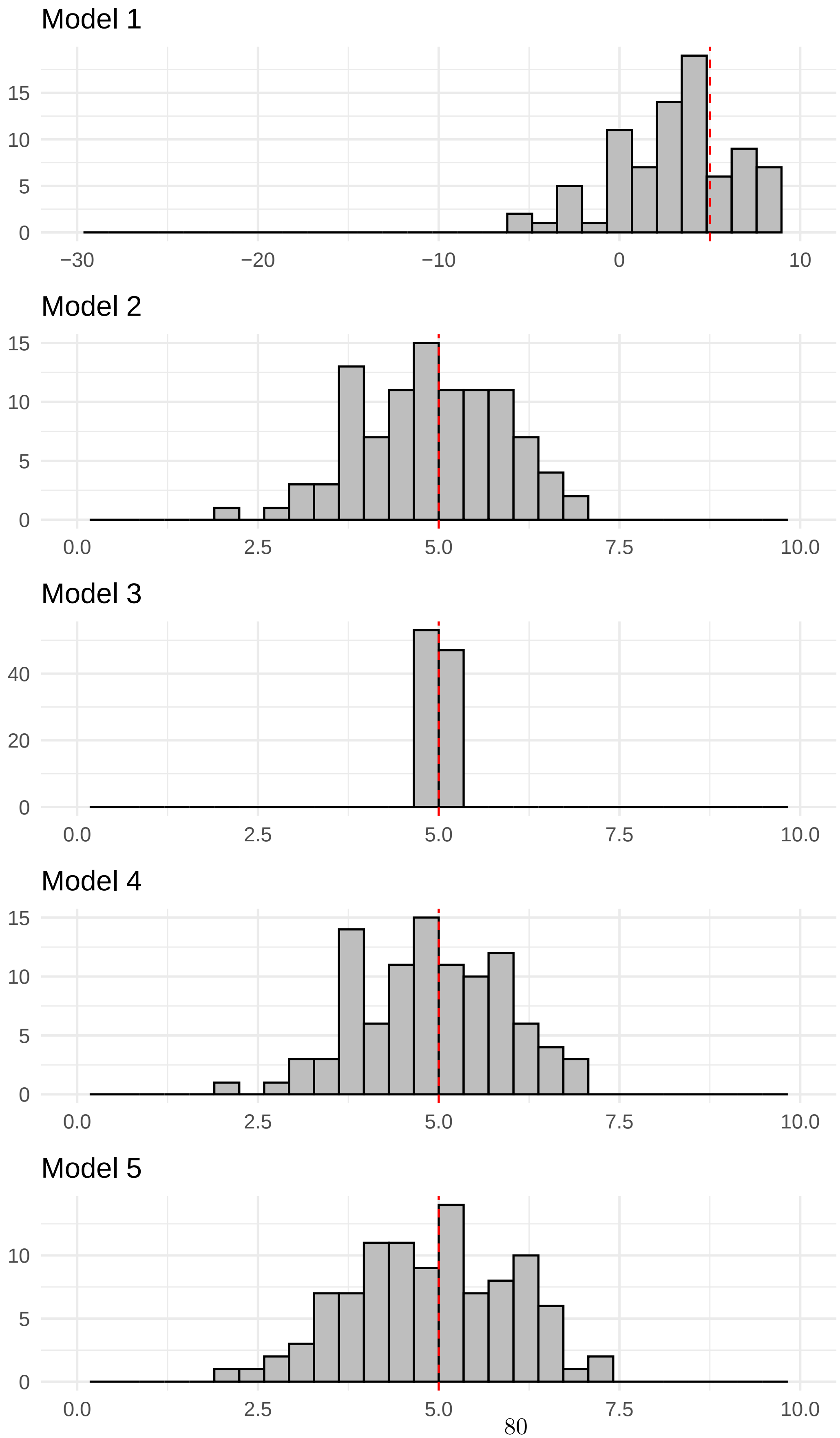
Solution:

```r
# Case 1
# xi | ri follows symmetric Gamma distr

M <- 100
mBetahats <- matrix(numeric(5*M),nrow=M,ncol=5)
n <- 100

for (i in 1:M) {
# Generate random sample
ui <- rnorm(n,0,sqrt(5))
gi <- rgamma(n,shape=2,scale=2)
ri <- rbinom(n,1,0.5)
xi <- numeric(n)
xi[ri == 1] <- rgamma(sum(ri == 1), shape = 5, scale = 1)
xi[ri == 0] <- rgamma(sum(ri == 0), shape = 5, scale = 1)
Betas <- c(400,5,200,10)
X <- cbind(rep(1),xi,ri,gi)
yi <- X %*% Betas + ui
ni <- rnorm(n,10,sqrt(3))
bi <- rnorm(n,5+sqrt(xi),sqrt(3))

# Simulate 5 models and store beta1 estimates
model1 <- lm(yi ~ xi)
```

Exercise 3.4: Solution

```
mBetahats[i,1] <- model1$coefficients[2]

model2 <- lm(yi ~ xi+ri)
mBetahats[i,2] <- model2$coefficients[2]

model3 <- lm(yi ~ xi+ri+gi)
mBetahats[i,3] <- model3$coefficients[2]

model4 <- lm(yi ~ xi+ri+ni)
mBetahats[i,4] <- model4$coefficients[2]

model5 <- lm(yi ~ xi+ri+bi)
mBetahats[i,5] <- model5$coefficients[2]
}

# Plots
p1 <- create_plot(mBetahats[,1], "Model 1", x_limits = c(-30, 10))
p2 <- create_plot(mBetahats[,2], "Model 2", x_limits = c(0, 10))
p3 <- create_plot(mBetahats[,3], "Model 3", x_limits = c(0, 10))
p4 <- create_plot(mBetahats[,4], "Model 4", x_limits = c(0, 10))
p5 <- create_plot(mBetahats[,5], "Model 5", x_limits = c(0, 10))

# Combine the plots
p1 / p2 / p3 / p4 / p5
```

Exercise 3.4: Solution



Model 1

Model 2

Model 3

Model 4

Model 5

Exercise 3.4: Solution

If $\mathbb{E}[x_i^*|r_i = 1] = \mathbb{E}[x_i^*|r_i = 0]$, then we can see that, by the law of iterated expectations, x_i^* is independent of r_i. As a consequence, the violation of A3 ceases to exist, since $\mathbb{E}[x_i^* r_i] = 0$ and thus $\mathbb{E}[x_i^* u_i] = 0$.

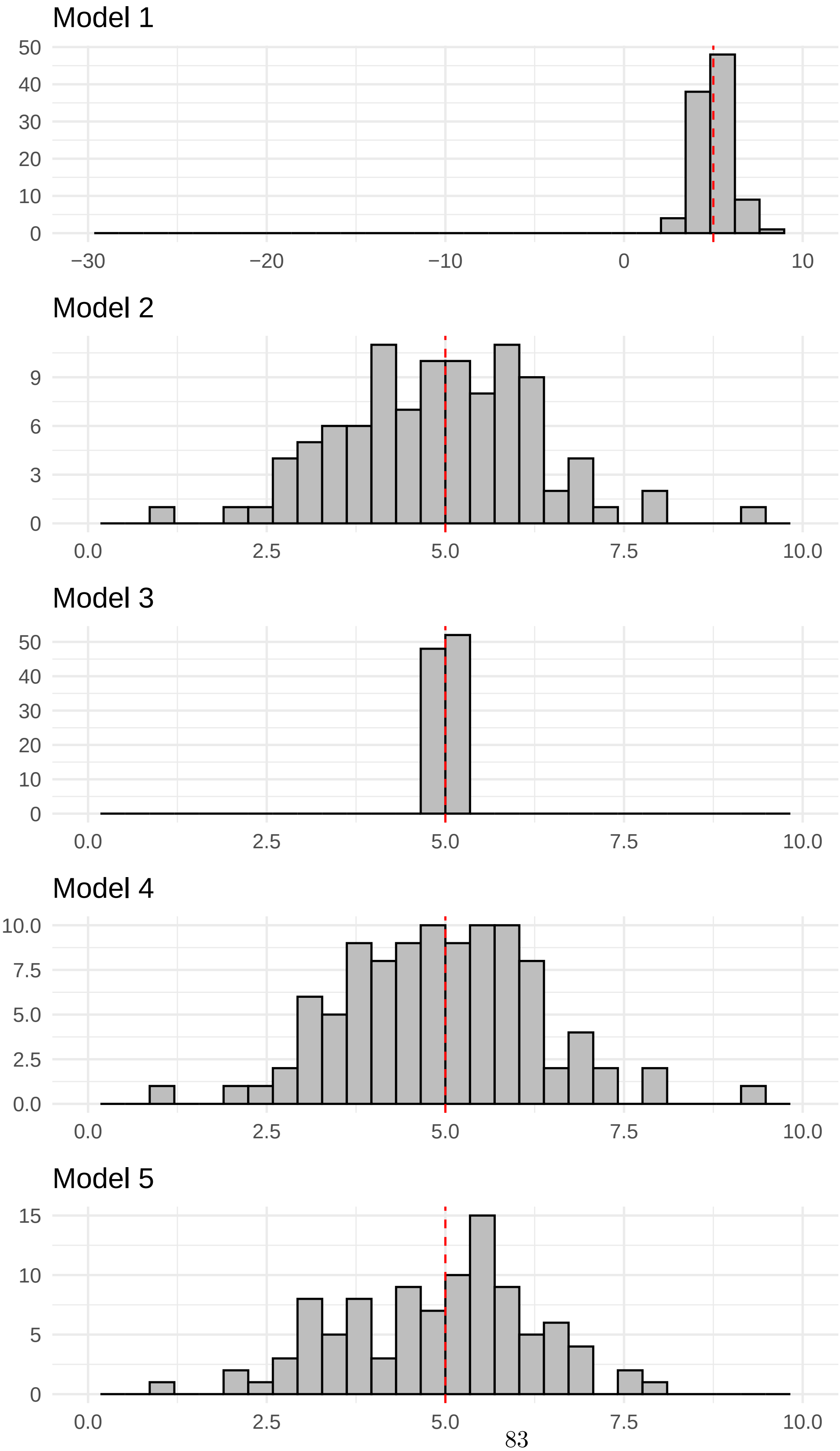
Now, the omission of r_i in Model 1 does not bias $\hat{\beta}_1$. However, it impacts the precision of the estimator. Since r_i still has a non-zero effect on y_i (i.e., $\beta_2 \neq 0$), when it's not included the standard error of the regression is higher.

```
# Case 2
# beta2 = 0

M <- 100
mBetahats <- matrix(numeric(5*M),nrow=M,ncol=5)
n <- 100

for (i in 1:M) {
# Generate random sample
ui <- rnorm(n,0,sqrt(5))
gi <- rgamma(n,shape=2,scale=2)
ri <- rbinom(n,1,0.5)
xi <- numeric(n)
xi[ri == 1] <- rgamma(sum(ri == 1), shape = 3, scale = 1)
xi[ri == 0] <- rgamma(sum(ri == 0), shape = 7, scale = 1)
Betas <- c(400,5,0,10)
X <- cbind(rep(1),xi,ri,gi)
yi <- X %*% Betas + ui
ni <- rnorm(n,10,sqrt(3))
bi <- rnorm(n,5+sqrt(xi),sqrt(3))

# Simulate 5 models and store beta1 estimates
model1 <- lm(yi ~ xi)
mBetahats[i,1] <- model1$coefficients[2]

model2 <- lm(yi ~ xi+ri)
mBetahats[i,2] <- model2$coefficients[2]

model3 <- lm(yi ~ xi+ri+gi)
mBetahats[i,3] <- model3$coefficients[2]

model4 <- lm(yi ~ xi+ri+ni)
mBetahats[i,4] <- model4$coefficients[2]

model5 <- lm(yi ~ xi+ri+bi)
mBetahats[i,5] <- model5$coefficients[2]
}

# Plots
```

Exercise 3.4: Solution

```r
p1 <- create_plot(mBetahats[,1], "Model 1", x_limits = c(-30, 10))
p2 <- create_plot(mBetahats[,2], "Model 2", x_limits = c(0, 10))
p3 <- create_plot(mBetahats[,3], "Model 3", x_limits = c(0, 10))
p4 <- create_plot(mBetahats[,4], "Model 4", x_limits = c(0, 10))
p5 <- create_plot(mBetahats[,5], "Model 5", x_limits = c(0, 10))

# Combine the plots
p1 / p2 / p3 / p4 / p5
```

Exercise 3.4: Solution


Model 1

Model 2

Model 3

Model 4

Model 5

Exercise 3.4: Solution

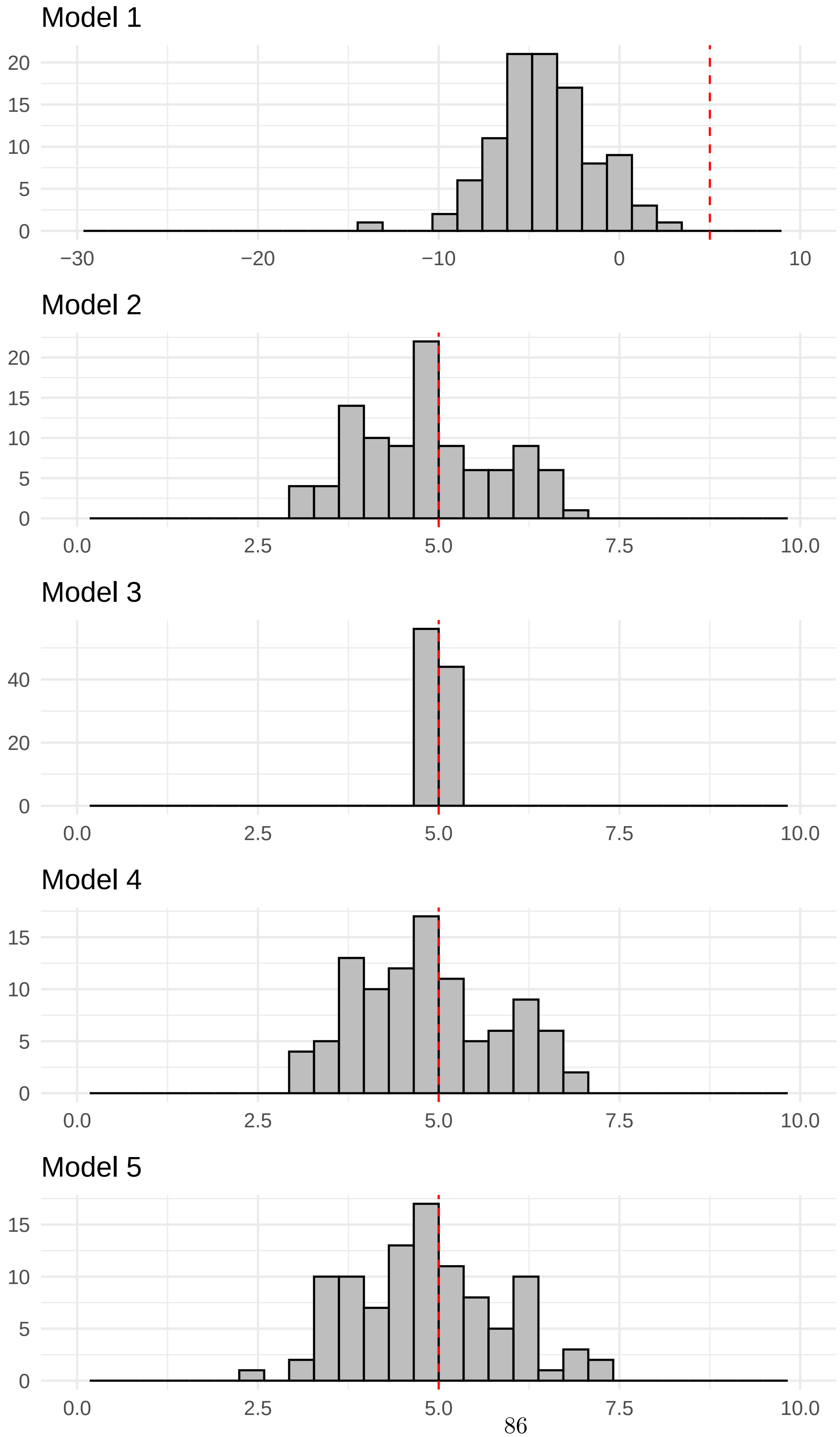
Instead, if we assume that in the real DGP r_i does not have an effect on y_i, then r_i becomes a bad control. Hence, its inclusion in Model 2 and subsequently has a negative effect on the efficiency of $\hat{\beta}_1$. By the same argument exposed above, $\mathbb{E}[x_i^* r_i]$ implies that the part of x_i^* used to estimate $\hat{\beta}_1$ is reduced.

```r
# Case 3
# ri ~ Bin(0.1)

M <- 100
mBetahats <- matrix(numeric(5*M),nrow=M,ncol=5)
n <- 100

for (i in 1:M) {
# Generate random sample
ui <- rnorm(n,0,sqrt(5))
gi <- rgamma(n,shape=2,scale=2)
ri <- rbinom(n,1,0.1)
xi <- numeric(n)
xi[ri == 1] <- rgamma(sum(ri == 1), shape = 3, scale = 1)
xi[ri == 0] <- rgamma(sum(ri == 0), shape = 7, scale = 1)
Betas <- c(400,5,200,10)
X <- cbind(rep(1),xi,ri,gi)
yi <- X %*% Betas + ui
ni <- rnorm(n,10,sqrt(3))
bi <- rnorm(n,5+sqrt(xi),sqrt(3))

# Simulate 5 models and store beta1 estimates
model1 <- lm(yi ~ xi)
mBetahats[i,1] <- model1$coefficients[2]

model2 <- lm(yi ~ xi+ri)
mBetahats[i,2] <- model2$coefficients[2]

model3 <- lm(yi ~ xi+ri+gi)
mBetahats[i,3] <- model3$coefficients[2]

model4 <- lm(yi ~ xi+ri+ni)
mBetahats[i,4] <- model4$coefficients[2]

model5 <- lm(yi ~ xi+ri+bi)
mBetahats[i,5] <- model5$coefficients[2]
}

# Plots
p1 <- create_plot(mBetahats[,1], "Model 1", x_limits = c(-30, 10))
p2 <- create_plot(mBetahats[,2], "Model 2", x_limits = c(0, 10))
p3 <- create_plot(mBetahats[,3], "Model 3", x_limits = c(0, 10))
```

Exercise 3.4: Solution

```r
p4 <- create_plot(mBetahats[,4], "Model 4", x_limits = c(0, 10))
p5 <- create_plot(mBetahats[,5], "Model 5", x_limits = c(0, 10))

# Combine the plots
p1 / p2 / p3 / p4 / p5
```

Exercise 3.4: Solution


Model 1

Model 2

Model 3

Model 4

Model 5

Exercise 3.4: Solution

Notice that a Bernoulli-distributed random variable $r_i \sim Ber(p)$, where $p \in (0,1)$ is the probability that such r.v. takes value 1, has variance $p(1-p)$. Such variance is indeed maximised when $p = 0.5$.

If $p \neq 0.5$ the variance of $r_i = 1$ decreases. The lower the variance of r_i, the lower $\mathbb{E}[x_i^* r_i]$, hence the lower the bias in $\hat{\beta}_1$ when r_i is omitted from the regression. In the limit case when either $p = 0$ or $p = 1$, r_i has no variance and therefore its omission no longer causes OVB.

We can see by comparing the distribution of $\hat{\beta}_1$ in this case with the distribution plotted in the previous exercise that indeed the bias, while still present, is indeed reduced.

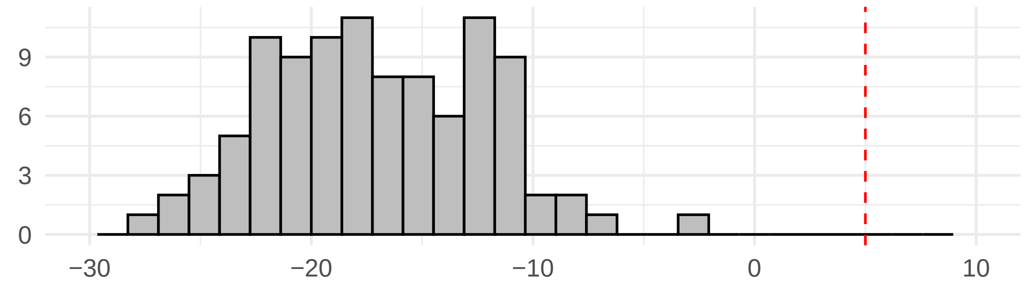
```
# Case 4
# beta3 = 50

M <- 100
mBetahats <- matrix(numeric(5*M),nrow=M,ncol=5)
n <- 100

for (i in 1:M) {
# Generate random sample
ui <- rnorm(n,0,sqrt(5))
gi <- rgamma(n,shape=2,scale=2)
ri <- rbinom(n,1,0.5)
xi <- numeric(n)
xi[ri == 1] <- rgamma(sum(ri == 1), shape = 3, scale = 1)
xi[ri == 0] <- rgamma(sum(ri == 0), shape = 7, scale = 1)
Betas <- c(400,5,200,50)
X <- cbind(rep(1),xi,ri,gi)
yi <- X %*% Betas + ui
ni <- rnorm(n,10,sqrt(3))
bi <- rnorm(n,5+sqrt(xi),sqrt(3))

# Simulate 5 models and store beta1 estimates
model1 <- lm(yi ~ xi)
mBetahats[i,1] <- model1$coefficients[2]

model2 <- lm(yi ~ xi+ri)
mBetahats[i,2] <- model2$coefficients[2]

model3 <- lm(yi ~ xi+ri+gi)
mBetahats[i,3] <- model3$coefficients[2]

model4 <- lm(yi ~ xi+ri+ni)
mBetahats[i,4] <- model4$coefficients[2]

model5 <- lm(yi ~ xi+ri+bi)
mBetahats[i,5] <- model5$coefficients[2]
}
```

Exercise 3.4: Solution

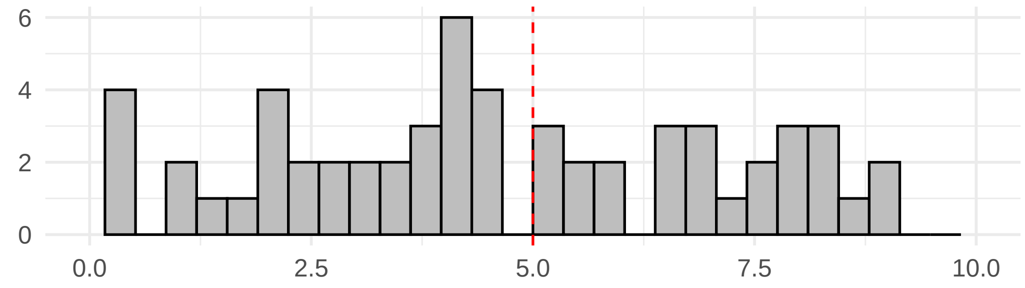
```r
# Plots
p1 <- create_plot(mBetahats[,1], "Model 1", x_limits = c(-30, 10))
p2 <- create_plot(mBetahats[,2], "Model 2", x_limits = c(0, 10))
p3 <- create_plot(mBetahats[,3], "Model 3", x_limits = c(0, 10))
p4 <- create_plot(mBetahats[,4], "Model 4", x_limits = c(0, 10))
p5 <- create_plot(mBetahats[,5], "Model 5", x_limits = c(0, 10))

# Combine the plots
p1 / p2 / p3 / p4 / p5
```

Exercise 3.4: Solution

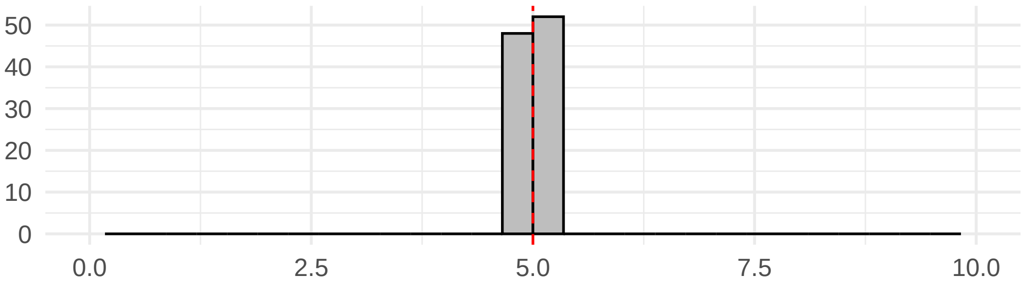
Model 1



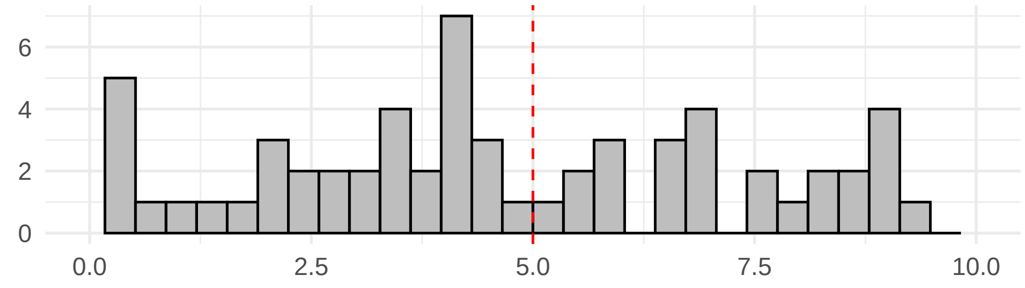
Model 2



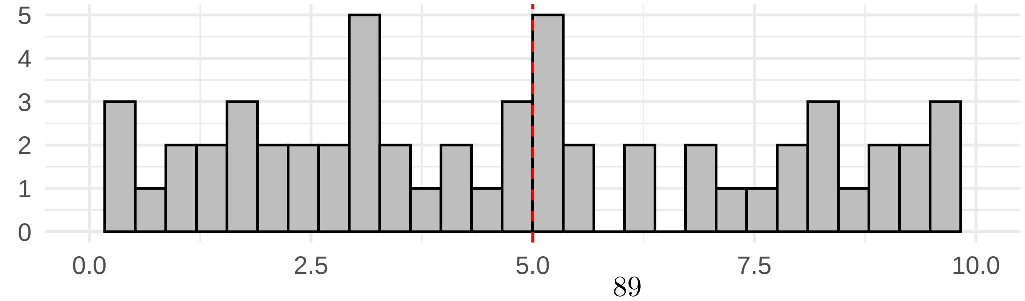
Model 3



Model 4



Model 5



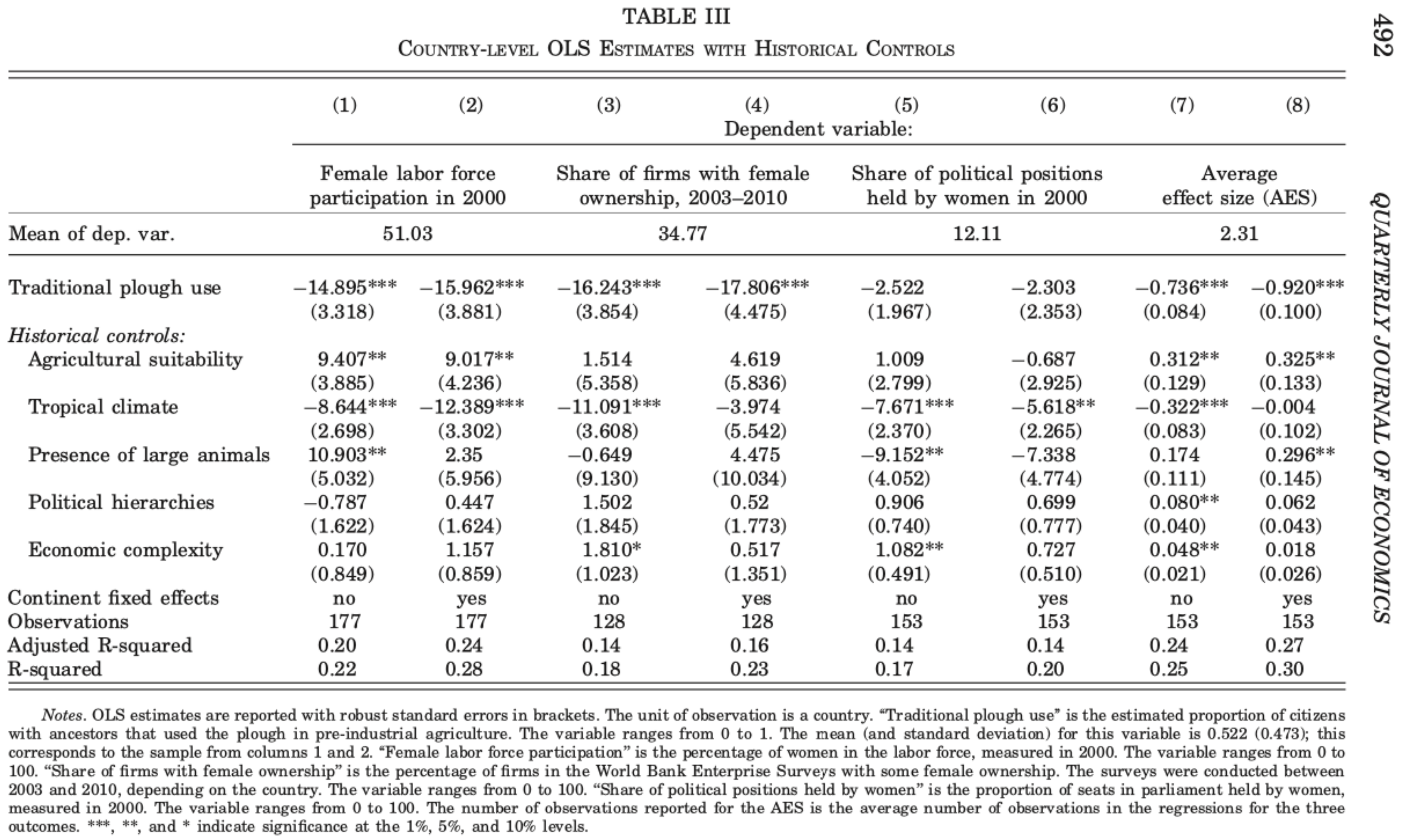
Exercise 3.4: Solution

Finally, by increasing the effect of g_i on y_i, also increases the loss in efficiency and precision of the OLS estimator for β_1 when g_i is omitted from the regression.

In fact, we can see that the loss in efficiency from Model 3 to Model 2 as defined in part (b) is:

$$\mathbb{E}[(u^2)'(u^2)] - \mathbb{E}[u'u] = \beta_3^2 \mathbb{E}[g_i^2]$$

Then, the larger β_3, the larger the efficiency loss.



492 QUARTERLY JOURNAL OF ECONOMICS

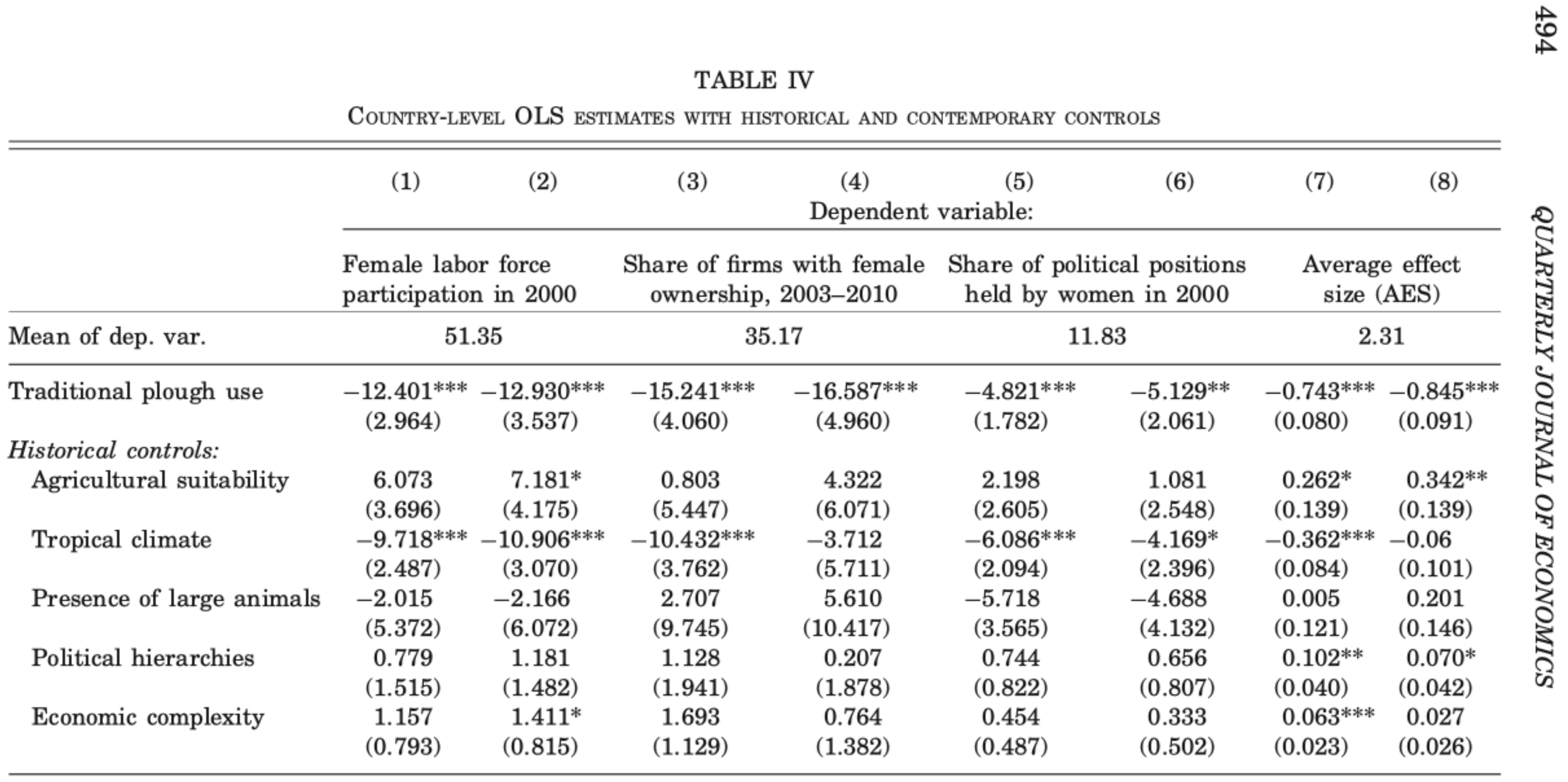
TABLE III

COUNTRY-LEVEL OLS ESTIMATES WITH HISTORICAL CONTROLS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable:							
	Female labor force participation in 2000		Share of firms with female ownership, 2003–2010		Share of political positions held by women in 2000		Average effect size (AES)	
Mean of dep. var.	51.03		34.77		12.11		2.31	
Traditional plough use	−14.895***	−15.962***	−16.243***	−17.806***	−2.522	−2.303	−0.736***	−0.920***
	(3.318)	(3.881)	(3.854)	(4.475)	(1.967)	(2.353)	(0.084)	(0.100)
Historical controls:								
Agricultural suitability	9.407**	9.017**	1.514	4.619	1.009	−0.687	0.312**	0.325**
	(3.885)	(4.236)	(5.358)	(5.836)	(2.799)	(2.925)	(0.129)	(0.133)
Tropical climate	−8.644***	−12.389***	−11.091***	−3.974	−7.671***	−5.618**	−0.322***	−0.004
	(2.698)	(3.302)	(3.608)	(5.542)	(2.370)	(2.265)	(0.083)	(0.102)
Presence of large animals	10.903**	2.35	−0.649	4.475	−9.152**	−7.338	0.174	0.296**
	(5.032)	(5.956)	(9.130)	(10.034)	(4.052)	(4.774)	(0.111)	(0.145)
Political hierarchies	−0.787	0.447	1.502	0.52	0.906	0.699	0.080**	0.062
	(1.622)	(1.624)	(1.845)	(1.773)	(0.740)	(0.777)	(0.040)	(0.043)
Economic complexity	0.170	1.157	1.810*	0.517	1.082**	0.727	0.048**	0.018
	(0.849)	(0.859)	(1.023)	(1.351)	(0.491)	(0.510)	(0.021)	(0.026)
Continent fixed effects	no	yes	no	yes	no	yes	no	yes
Observations	177	177	128	128	153	153	153	153
Adjusted R-squared	0.20	0.24	0.14	0.16	0.14	0.14	0.24	0.27
R-squared	0.22	0.28	0.18	0.23	0.17	0.20	0.25	0.30

Notes. OLS estimates are reported with robust standard errors in brackets. The unit of observation is a country. "Traditional plough use" is the estimated proportion of citizens with ancestors that used the plough in pre-industrial agriculture. The variable ranges from 0 to 1. The mean (and standard deviation) for this variable is 0.522 (0.473); this corresponds to the sample from columns 1 and 2. "Female labor force participation" is the percentage of women in the labor force, measured in 2000. The variable ranges from 0 to 100. "Share of firms with female ownership" is the percentage of firms in the World Bank Enterprise Surveys with some female ownership. The surveys were conducted between 2003 and 2010, depending on the country. The variable ranges from 0 to 100. "Share of political positions held by women" is the proportion of seats in parliament held by women, measured in 2000. The variable ranges from 0 to 100. The number of observations reported for the AES is the average number of observations in the regressions for the three outcomes. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Figure 3.1: Table 3 from Alesina, Giuliano & Nunn (2013)



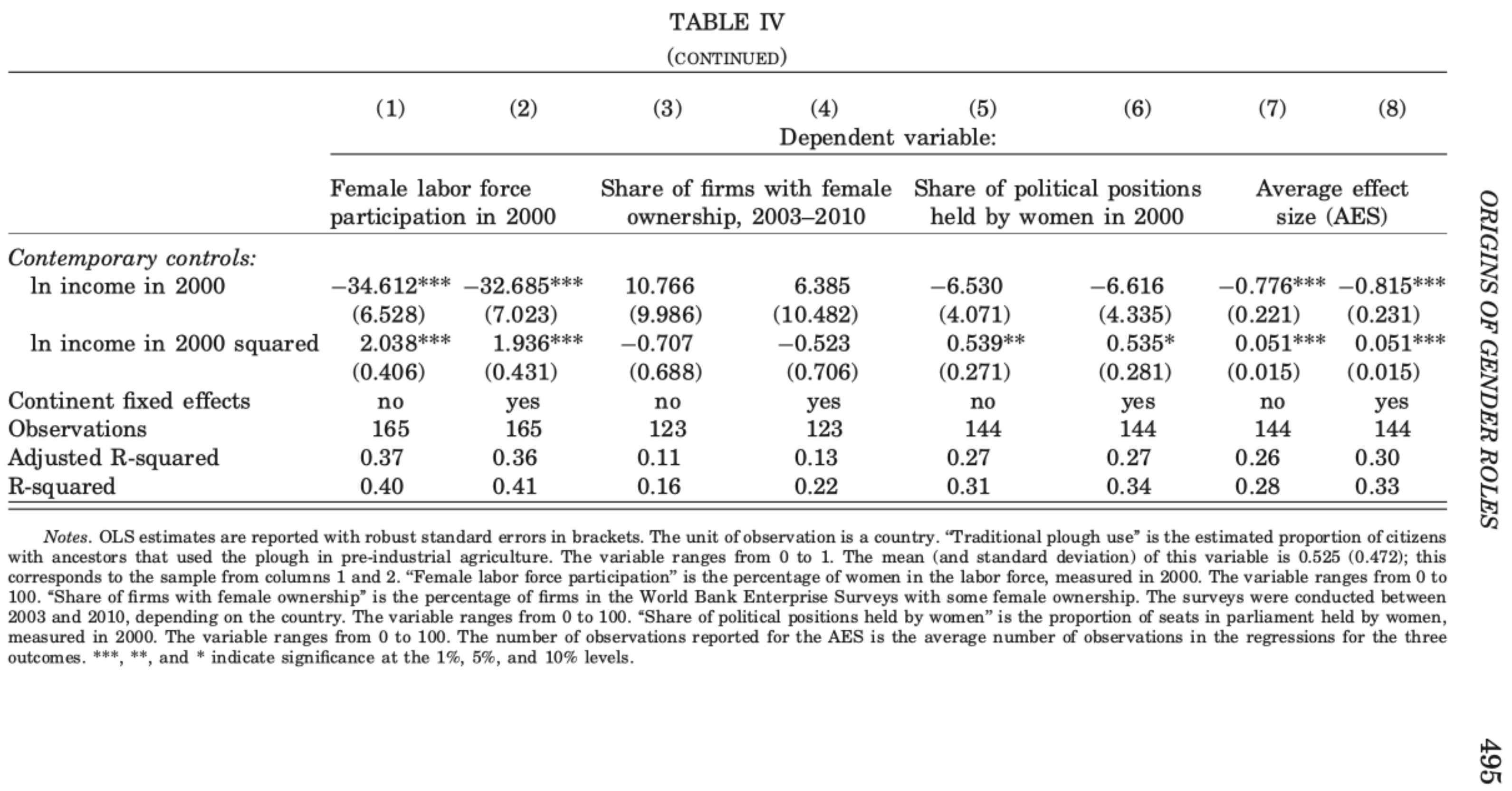
494 QUARTERLY JOURNAL OF ECONOMICS

TABLE IV

COUNTRY-LEVEL OLS ESTIMATES WITH HISTORICAL AND CONTEMPORARY CONTROLS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable:							
	Female labor force participation in 2000		Share of firms with female ownership, 2003–2010		Share of political positions held by women in 2000		Average effect size (AES)	
Mean of dep. var.	51.35		35.17		11.83		2.31	
Traditional plough use	−12.401***	−12.930***	−15.241***	−16.587***	−4.821***	−5.129**	−0.743***	−0.845***
	(2.964)	(3.537)	(4.060)	(4.960)	(1.782)	(2.061)	(0.080)	(0.091)
Historical controls:								
Agricultural suitability	6.073	7.181*	0.803	4.322	2.198	1.081	0.262*	0.342**
	(3.696)	(4.175)	(5.447)	(6.071)	(2.605)	(2.548)	(0.139)	(0.139)
Tropical climate	−9.718***	−10.906***	−10.432***	−3.712	−6.086***	−4.169*	−0.362***	−0.06
	(2.487)	(3.070)	(3.762)	(5.711)	(2.094)	(2.396)	(0.084)	(0.101)
Presence of large animals	−2.015	−2.166	2.707	5.610	−5.718	−4.688	0.005	0.201
	(5.372)	(6.072)	(9.745)	(10.417)	(3.565)	(4.132)	(0.121)	(0.146)
Political hierarchies	0.779	1.181	1.128	0.207	0.744	0.656	0.102**	0.070*
	(1.515)	(1.482)	(1.941)	(1.878)	(0.822)	(0.807)	(0.040)	(0.042)
Economic complexity	1.157	1.411*	1.693	0.764	0.454	0.333	0.063***	0.027
	(0.793)	(0.815)	(1.129)	(1.382)	(0.487)	(0.502)	(0.023)	(0.026)

Figure 3.2: First Part of Table 4 from Alesina, Giuliano & Nunn (2013)



ORIGINS OF GENDER ROLES 495

TABLE IV

(CONTINUED)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable:							
	Female labor force participation in 2000		Share of firms with female ownership, 2003–2010		Share of political positions held by women in 2000		Average effect size (AES)	
Contemporary controls:								
ln income in 2000	−34.612***	−32.685***	10.766	6.385	−6.530	−6.616	−0.776***	−0.815***
	(6.528)	(7.023)	(9.986)	(10.482)	(4.071)	(4.335)	(0.221)	(0.231)
ln income in 2000 squared	2.038***	1.936***	−0.707	−0.523	0.539**	0.535*	0.051***	0.051***
	(0.406)	(0.431)	(0.688)	(0.706)	(0.271)	(0.281)	(0.015)	(0.015)
Continent fixed effects	no	yes	no	yes	no	yes	no	yes
Observations	165	165	123	123	144	144	144	144
Adjusted R-squared	0.37	0.36	0.11	0.13	0.27	0.27	0.26	0.30
R-squared	0.40	0.41	0.16	0.22	0.31	0.34	0.28	0.33

Notes. OLS estimates are reported with robust standard errors in brackets. The unit of observation is a country. "Traditional plough use" is the estimated proportion of citizens with ancestors that used the plough in pre-industrial agriculture. The variable ranges from 0 to 1. The mean (and standard deviation) of this variable is 0.525 (0.472); this corresponds to the sample from columns 1 and 2. "Female labor force participation" is the percentage of women in the labor force, measured in 2000. The variable ranges from 0 to 100. "Share of firms with female ownership" is the percentage of firms in the World Bank Enterprise Surveys with some female ownership. The surveys were conducted between 2003 and 2010, depending on the country. The variable ranges from 0 to 100. "Share of political positions held by women" is the proportion of seats in parliament held by women, measured in 2000. The variable ranges from 0 to 100. The number of observations reported for the AES is the average number of observations in the regressions for the three outcomes. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Figure 3.3: Second Part of Table 4 from Alesina, Giuliano & Nunn (2013)

4 Maximum Likelihood Estimation: Linear Regressions & Beyond

4.1 ML Inference for Gamma-distributed RV

Suppose you observe a sample of n *i.i.d.* observations $\{x_i\}_{i=1}^{n}$ from the Gamma distribution $G(k, 1/\theta)$ with pdf

$$p(x) = \frac{\theta^k}{(k-1)!} x^{(k-1)} e^{-\theta x} \ , \quad x > 0 \ ,$$

where $k \geq 1$ is a known integer and $\theta > 0$ is the unknown parameter of interest.

(a) Derive the likelihood and the ML estimator of θ.

(b) Is the ML estimator of θ biased? If so, propose an unbiased estimator.

Hint: if $x \sim G(k, 1/\theta)$*, then* $\sum_{i=1}^{n} x_i \sim G(nk, 1/\theta)$*, i.e.* $Z = \sum_{i=1}^{n} x_i$ *has the pdf*

$$p_Z(z) = \frac{\theta^{nk}}{(nk-1)!} z^{(nk-1)} e^{-\theta z} \ .$$

Also, by virtue of being a pdf, $p_Z(z)$ *has to integrate to one for any* nk *(provided that* $nk \geq 1$ *and* nk *being an integer), so too for* $nk - 1$*:*

$$\int_0^\infty \frac{\theta^{nk-1}}{(nk-2)!} z^{(nk-2)} e^{-\theta z} dz = 1 \ .$$

(c) Derive the Cramer-Rao lower bound. Does it depend on θ?

(d) Does your bias-corrected estimator attain the CR lower bound?

4.2 ML Inference in Linear Regression Model

Remark: excercise continued in 7.1, 8.1.

Suppose you observe n i.i.d. observations of an outcome variable y_i and a k-dimensional vector of covariates x_i: $\{(y_i, x_i)\}_{i=1:n}$. Suppose further that you want to examine their relation using a linear regression model,

$$y_i = x_i'\beta + u_i \ , \quad u_i|x_i \sim N(0, \sigma^2) \ ,$$

or, in matrix notation,

$$Y = X\beta + U \ ,$$

where Y and U are $(n \times 1)$ and X is $(n \times k)$.

You are exclusively concerned with frequentist/classical estimation, i.e. you consider the parameters β and σ^2 to be fixed and the data $\{(y_i, x_i)\}_{i=1:n}$ to be random (i.e. your particular sample is randomly drawn from some population).

(a) Derive the (conditional) likelihood function $\mathcal{L}(\beta, \sigma^2|Y, X) \equiv p(Y|X, \beta, \sigma^2)$, and define the Maximum Likelihood (ML) estimators $\hat{\beta}$ and $\hat{\sigma}^2$.

(b) Derive the ML estimators $\hat{\beta}$ and $\hat{\sigma}^2$.

Hint: for a matrix B and vectors a and c, it holds that $\dfrac{\partial(a - Bc)'(a - Bc)}{\partial c} = -2B'(a - Bc)$.

(c) Find $\mathbb{E}[\hat{\beta}]$ and $\mathbb{V}[\hat{\beta}] = \mathbb{E}[\Big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\Big)\Big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\Big)']$. What else do you know about the finite sample distribution of $\hat{\beta}$?

Hint: use the Law of Iterated Expectations, computing these two quantities conditional on X first.

(d) Is your estimator $\hat{\beta}$ consistent? Use the expression that you derived for it in (b) and theorems to answer this question.

(e) What is the asymptotic distribution of $\hat{\beta}$? Use the expression that you derived for it in (b) and theorems to answer this question.

(f) Suppose you have $k = 4$ regressors (including the intercept: $x_{i1} = 1 \ \forall \ i$), and you want to test $\mathcal{H}_0 : \{ \ \beta_2 + 3\beta_3 = 7 \ , \ \log\beta_4 = 0 \ \}$ against the alternative $\mathcal{H}_1$, specifying that at least one of the two conditions in $\mathcal{H}_0$ is not true. Describe two approaches to conduct this test and their relative advantages/disadvantages.

(g) Construct a 95% confidence interval for $exp\{\beta_4\}$.

4.3 Probit vs. Logit

The probit model is a modeling approach for binary outcome variables $y_i \in \{0, 1\}$. It specifies a latent variable model,

$$y_i^* = x_i'\beta + u_i \quad , \quad u_i|x_i \sim N(0, 1) \ ,$$

and assumes we observe $y_i = \mathbf{1}\{y_i^* \geqslant 0\}$.

(a) Show that the assumption that the variance of $u_i|x_i$ is equal to one is without loss of generality. To answer the question, redo the calculations (likelihood derivation and calculation of partial effects) with $u_i|x_i \sim N(0, \sigma^2)$ instead, and use your knowledge on properties of partial effects under the probit model.

(b) The logit model is very similar to the probit model. The only difference is that it assumes that $u_i|x_i$ follows a standard logistic distribution instead of the standard Normal distribution. Its cdf is then

$$F(x) = 1/(1 + exp\{-x\})$$

instead of $\Phi(x)$. Redo the calculations (likelihood derivation and calculation of partial effects) for the logit model.

4.4 Inference in Probit Model

Suppose you have a dataset containing a shop's sales, which includes the date, some characteristics of the customer (like income, age), some characteristics of the transaction (like type of good sold, price, and whether cash or a card was used). You are interested in shedding light on the determinants of cash vs card payment.

(a) How could you use the probit model for your research question? What is your y_i variable? How can we interpret the underlying latent variable y_i^*?

(b) In your probit model, derive the effect of age increasing by 5 years on the probability of using cash. Does the effect depend on the current age of the customer? Does it depend on the values of the other variables?

(c) Could you use a standard linear regression, estimated via OLS, to answer your question?

(d) Derive the same effect as in (b) in your linear regression model. Does it depend on the current age of the customer? Does it depend on the values of the other variables?

(e) Based on your reasoning so far, for which customers would you expect the predicted effect under the linear regression to be close to the one under probit? For what type of customers will the two differ more? As a result, for what kind of research questions is the linear regression a good/bad specification?
Hint: Besides comparing (partial) effects under the two models, you might want to compare the functional form of $\mathbb{E}[y_i|x_i]$ *under the two models.*

4.5 Inference in Tobit Model

Suppose you are interested in relating air quality in different cities – measured by the concentration of carbon monoxide in the air, y_i – to possible determinants x_i. The measurement device used in your data cannot detect concentrations below a certain value, δ, but simply codes them as zero. For this purpose, you set up a Tobit model for observations y_i with a lower-censoring at δ:

$$\begin{aligned} y_i^* &= x_i'\beta + u_i \ , \quad u_i \sim N(0,\sigma^2) \ , \\ y_i &= y_i^* \ \mathbf{1}\left\{y_i^* > \delta\right\} \ . \end{aligned} \tag{4.1}$$

(a) Derive the probability of measuring a concentration of carbon monoxide of zero as a function of determinants x_i (and parameters β and the censoring point δ), $\mathbb{P}[y_i = 0|x_i]$.

(b) Derive the conditional mean $\mathbb{E}[y_i^*|x_i]$, i.e. the expected air quality (true concentration of carbon monoxide) for generic a city i with characteristics x_i.

(c) Derive the conditional mean $\mathbb{E}[y_i|x_i]$, i.e. the expected (measurable) concentration of carbon monoxide for generic a city i with characteristics x_i.
Hint: recall that for $z_i \sim N(0,1)$, $\mathbb{E}[z_i|z_i > -c] = \phi(c)/\Phi(c)$ (Inverse-Mills ratio).

(d) Suppose one of your variables in x_i is the cost of public transport as a fraction of the average hourly wage in the city, c_i. Using your result from the previous two exercises, derive the predicted effect of decreasing this ratio by 10 percentage points on air quality y_i^* and measured carbon monoxide concentration y_i.

(e) Instead, suppose you simply use a linear regression to relate y_i to x_i for the cities for whom the concentration was measured precisely, i.e. for cities $i \in \mathcal{U} \equiv \{i : y_i > \delta\}$:

$$y_i = x_i'\gamma + v_i \ , \quad i \in \mathcal{U} \ . \tag{4.2}$$

What is the effect of decreasing c_i on y_i in this specification? Presuming for a moment that γ and β are the same thing, for which cities is the predicted effect under the linear regression close to/far from the one under the above tobit model?

(f) (Bonus question) You are in fact not interested in relating y_i to x_i, but in relating the true air quality y_i^* – of which y_i is an imperfect measure – to x_i, i.e. you are interested in β, not γ. Supposing that Eq. (4.1) is the true model generating the data, can you use the OLS estimator for γ from Eq. (4.2) to consistently estimate β? Under which circumstances will OLS work better/worse?
Hint: For $i \in \mathcal{U}$, we simply have $y_i = y_i^* = x_i'\beta + u_i$. Also, for a generic random variable z_i,

$$\frac{1}{n_u}\sum_{i \in \mathcal{U}} z_i \xrightarrow{p} \mathbb{E}[z_i|y_i > \delta] = \mathbb{E}[z_i|y_i^* > \delta] = \mathbb{E}[z_i|u_i > \delta - x_i'\beta] \ ,$$

where $n_u = |\mathcal{U}|$ is the number of observations i in $\mathcal{U}$.

4.6 Inference in Duration Model

Suppose you observe a sample of n unemployed invidiuals. Let y_i denote the time (in weeks) that individual i spent in unemployment, and let $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ be the average unemployment span in your sample. You can assume that your observations are independent.

One can model y_i using an expontential distribution:

$$p(y_i|\ \lambda) = \lambda exp\left\{-\lambda y_i\right\} \quad , \quad \lambda > 0 \ .$$

The parameter λ is the job-finding rate. It tells you how many (acceptable) job offers per week an individual gets. For example, $\lambda = 3$ would tell you that (on average) an individual receives three offers every week, while $\lambda = 1/2$ would tell you that (on average), an individual receives an offer every two weeks. For now, we assume that this λ is the same for all individuals in the sample.

(a) The mean and variance of y_i are given by

$$\mathbb{E}[y_i] = \frac{1}{\lambda} \quad \text{and} \quad \mathbb{V}[y_i] = \frac{1}{\lambda^2} \ .$$

Interpret these expressions, relying on the two examples $\lambda = 3$ and $\lambda = 1/2$.

(b) Derive the log-likelihood $\ell(\lambda|Y)$ and find the Maximum Likelihood (ML) estimator

$$\hat{\lambda} \equiv \arg\max_{\lambda} \ell(\lambda|Y) \ .$$

(c) What is the probability limit of the average unemployment span in your sample, $\bar{y}$? Based on that result, is $\hat{\lambda}$ consistent?

(d) What is the approximate distribution of $\bar{y}$ for large n? Based on that result, what is the approximate distribution of $\hat{\lambda}$ for large n?

(e) Describe another, numerical approach to approximate the distribution of $\hat{\lambda}$.

Now let's make the job-finding rate heterogeneous. Specifically, suppose x_i denotes the number of applications per week that individual i sent out, and let

$$\lambda_i = \exp\left\{\alpha + \beta x_i\right\} \ .$$

Note that this implies that

$$\mathbb{E}[y_i|x_i] = \frac{1}{\lambda_i} = \exp\left\{-(\alpha + \beta x_i)\right\} \quad \text{and} \quad \mathbb{V}[y_i|x_i] = \frac{1}{\lambda_i^2} = \exp\left\{-2(\alpha + \beta x_i)\right\} \ . \qquad (4.3)$$

For simplicity, let's suppose you know α (so it's just a constant) and you only need to estimate β.

The ML estimator can be defined as

$$\hat{\beta} = \arg\min_{\beta} Q_n(\beta) \ , \quad Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} -(\alpha + \beta x_i) + y_i \exp\{\alpha + \beta x_i\} \ .$$

(f) How do you interpret the parameters α and β?

(g) How can you find $\hat{\beta}$? Derive the first-order condition associated with the above optimization problem.

(h) Is $\hat{\beta}$ a consistent estimator for β_0, the true value for β?
Hint: Note that $\mathbb{E}[y_i|x_i] = \exp\{-(\alpha + \beta_0 x_i)\}$*, and remember the Law of Iterated Expectations (LIE). Also, note that* $x_i \geq 0$ *and a function like* $\mathbb{E}[x_i exp\{x_i\beta\}]$ *is strictly increasing in* β*.*[1]

(i) Show that the asymptotic distribution of $\hat{\beta}$ is given by

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \mathbb{E}[x_i^2]^{-1}\right) \ .$$

Hint: To simplify notation, write λ_{i0} *for* $exp\{\alpha + \beta_0 x_i\}$*. Also, your formula for the asymptotic variance simplifies thanks to the LIE and the results in Eq.* (4.3)*.*

(j) Construct a hypothesis test with size $\alpha = 0.05$ for testing

$$\mathscr{H}_0 : \mathbb{E}[y_i|x_i = 10] = \frac{1}{2}\mathbb{E}[y_i|x_i = 5] \quad \text{vs.} \quad \mathscr{H}_0 : \mathbb{E}[y_i|x_i = 10] \neq \frac{1}{2}\mathbb{E}[y_i|x_i = 5] \ ,$$

i.e. testing whether an individual submitting 10 applications per week spends (in expectation) exactly half as long in unemployment than a person sending out only 5 applications per week. More concretely, definining the test as $\varphi = \mathbf{1}\{T(X) < c_\alpha\}$, define the test-statistic $T(X)$ and find the critical value c_α.

[1] Strictly speaking, this holds provided that $\mathbb{E}[x_i] > 0$.

Exercise 4.1: Solution

Suppose you observe a sample of n *i.i.d.* observations $\{x_i\}_{i=1}^n$ from the Gamma distribution $G(k, 1/\theta)$ with pdf

$$p(x) = \frac{\theta^k}{(k-1)!} x^{(k-1)} e^{-\theta x} \ , \quad x > 0 \ ,$$

where $k \geq 1$ is a known integer and $\theta > 0$ is the unknown parameter of interest.

(a) Derive the likelihood and the ML estimator of θ.

(b) Is the ML estimator of θ biased? If so, propose an unbiased estimator.

Hint: if $x \sim G(k, 1/\theta)$*, then* $\sum_{i=1}^n x_i \sim G(nk, 1/\theta)$*, i.e.* $Z = \sum_{i=1}^n x_i$ *has the pdf*

$$p_Z(z) = \frac{\theta^{nk}}{(nk-1)!} z^{(nk-1)} e^{-\theta z} \ .$$

Also, by virtue of being a pdf, $p_Z(z)$ *has to integrate to one for any* nk *(provided that* $nk \geq 1$ *and* nk *being an integer), so too for* $nk - 1$*:*

$$\int_0^\infty \frac{\theta^{nk-1}}{(nk-2)!} z^{(nk-2)} e^{-\theta z} dz = 1 \ .$$

(c) Derive the Cramer-Rao lower bound. Does it depend on θ?

(d) Does your bias-corrected estimator attain the CR lower bound?

Solution:

(a) Likelihood function:

$$\begin{aligned}
\mathcal{L}(\theta|X) &= \prod_{i=1}^n p(x_i|\theta) \\
&= \prod_{i=1}^n \frac{\theta^k}{(k-1)!} x_i^{(k-1)} \exp\{-\theta x_i\} \\
&= \left(\frac{\theta^k}{(k-1)!} \right)^n \prod_{i=1}^n x_i^{(k-1)} \exp\left\{ -\theta \sum_{i=1}^n x_i \right\} \ .
\end{aligned}$$

Log-likelihood:

$$l(\theta|X) = n\left(k \ln(\theta) - \ln((k-1)!)\right) + (k-1)\sum_{i=1}^n \ln(x_i) - \theta \sum_{i=1}^n x_i \ .$$

Exercise 4.1: Solution

The First-Order Condition (FOC) sets the score to zero:

$$s(\theta|X) \equiv \frac{\partial}{\partial\theta} l(\theta|X) = \frac{nk}{\theta} - \sum_{i=1}^{n} x_i = 0 \; .$$

We can solve for θ to obtain:

$$\hat{\theta}_{ML} = \frac{nk}{\sum\limits_{i=1}^{n} x_i} \; .$$

(b)

$$\mathbb{E}[\hat{\theta}_{ML}] = \mathbb{E}\left[\frac{nk}{\sum\limits_{i=1}^{n} x_i}\right] = nk\mathbb{E}\left[\frac{1}{Z}\right] \; ,$$

where $Z \equiv \sum\limits_{i=1}^{n} x_i$ is a RV with pdf

$$p_Z(z) = \frac{\theta^{nk}}{(nk-1)!} z^{(nk-1)} \exp\left\{-\theta z\right\} \; .$$

This allows to compute:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{Z}\right] &= \int_{\mathbb{R}^+} \frac{1}{z} \frac{\theta^{nk}}{(nk-1)!} z^{(nk-1)} \exp\left\{-\theta z\right\} dz \\
&= \frac{\theta^{nk}}{(nk-1)!} \int_{\mathbb{R}^+} z^{(nk-2)} \exp\left\{-\theta z\right\} dz \\
&= \frac{\theta^{nk}}{(nk-1)!} \frac{(nk-2)!}{\theta^{nk-1}} \\
&= \theta \frac{(nk-2)!}{(nk-1)!} \\
&= \frac{(nk-2)!}{(nk-1)(nk-2)!}\theta \\
&= \frac{\theta}{nk-1} \; ,
\end{aligned}$$

where we substituted the integral in brackets by using the hint that

$$\int_{\mathbb{R}^+} \frac{\theta^{(nk-1)}}{(nk-2)!} z^{(nk-2)} \exp\left\{-\theta z\right\} dz = 1 \implies \int_{\mathbb{R}^+} z^{(nk-2)} \exp\left\{-\theta z\right\} dz = \frac{(nk-2)!}{\theta^{(nk-1)}} \; .$$

Exercise 4.1: Solution

Plugging $\mathbb{E}[1/Z]$ into the expectation of the ML estimator yields

$$\mathbb{E}[\hat{\theta}_{ML}] = nk\mathbb{E}\left[\frac{1}{Z}\right] = \theta\frac{nk}{nk-1} \neq \theta \ ,$$

i.e. we see that the ML estimator for θ is biased. We can correct for the bias by multiplying $\hat{\theta}$ by $(nk-1)/(nk)$ to obtain an unbiased estimator $\hat{\theta}_u$:

$$\hat{\theta}_u = \frac{nk-1}{\sum\limits_{i=1}^{n} x_i} \ ,$$

with

$$\mathbb{E}[\hat{\theta}_u] = \frac{nk-1}{nk}\mathbb{E}\left[\hat{\theta}_{ML}\right] = \frac{nk-1}{nk}\frac{nk}{nk-1}\theta = \theta \ .$$

(c) Based on the score above – $s(\theta|X) = \frac{nk}{\theta} - \sum\limits_{i=1}^{n} x_i$ –, we compute the Hessian

$$H(\theta|X) = \frac{\partial}{\partial\theta}s(\theta|X) = -\frac{nk}{\theta^2}$$

and the Information Matrix

$$\mathcal{I}(\theta_0) = -\mathbb{E}\left[H(\theta_0)\right] = \frac{nk}{\theta_0^2} \ .$$

This gives us the Cramér-Rao Lower Bound

$$\text{CRLB} \equiv \mathcal{I}(\theta_0)^{-1} = \frac{\theta_0^2}{nk} \ .$$

Yes, the CRLB depends on the true value of θ, denoted by θ_0.

(d) No, it does not. In order to prove it, we need to compute the variance of $\hat{\theta}_u$,

$$\mathbb{V}[\hat{\theta}_u] = \mathbb{V}\left[\left(\frac{nk-1}{nk}\right)\hat{\theta}_{ML}\right] = \left(\frac{nk-1}{nk}\right)^2\mathbb{V}[\hat{\theta}_{ML}] \ .$$

The variance of $\hat{\theta}_{ML}$ is

$$\mathbb{V}[\hat{\theta}_{ML}] = \mathbb{E}[(\hat{\theta}_{ML})^2] - \mathbb{E}[\hat{\theta}_{ML}]^2 \ .$$

Exercise 4.1: Solution

We know that the latter term is given by

$$\mathbb{E}[\hat{\theta}_{ML}]^2 = \left(\frac{nk}{nk-1}\right)^2 \theta^2 \ .$$

We compute the former part analogously as we computed the latter part:

$$\mathbb{E}[(\hat{\theta}_{ML})^2] = \mathbb{E}\left[\left(\frac{nk}{\sum_{i=1}^{n} x_i}\right)^2\right] = \mathbb{E}\left[\left(\frac{nk}{Z}\right)^2\right] = (nk)^2 \mathbb{E}\left[\frac{1}{Z^2}\right] \ ,$$

and

$$\begin{aligned}\mathbb{E}\left[\frac{1}{Z^2}\right] &= \int_{\mathbb{R}^+} \frac{1}{z^2} \frac{\theta^{nk}}{(nk-1)!} z^{nk-1} \exp\{-\theta z\}\, dz \\ &= \int_{\mathbb{R}^+} \frac{\theta^{nk}}{(nk-1)!} z^{nk-3} \exp\{-\theta z\}\, dz \\ &= \frac{\theta^{nk}}{(nk-1)!} \int_{\mathbb{R}^+} z^{nk-3} \exp\{-\theta z\}\, dz \\ &= \frac{\theta^{nk}}{(nk-1)!} \frac{(nk-3)!}{\theta^{(nk-2)}} \\ &= \frac{\theta^2}{(nk-1)(nk-2)} \ ,\end{aligned}$$

where we substituted the integral once again by using the hint:

$$\int_{\mathbb{R}^+} \frac{\theta^{nk-2}}{(nk-3)!} z^{(nk-3)} \exp\{-\theta z\}\, dz = 1 \implies \int_{\mathbb{R}^+} z^{(nk-3)} \exp\{-\theta z\}\, dz = \frac{(nk-3)!}{\theta^{(nk-2)}} \ .$$

Using this result, we have

$$\mathbb{E}[(\hat{\theta}_{ML})^2] = \frac{(nk)^2}{(nk-1)(nk-2)} \theta^2 \ ,$$

and

$$\begin{aligned}\mathbb{V}[\hat{\theta}_{ML}] &= \frac{(nk)^2\theta^2}{(nk-1)(nk-2)} - \frac{(nk)^2\theta^2}{(nk-1)^2} \\ &= \theta^2 \frac{(nk)^2}{(nk-1)} \left(\frac{1}{nk-2} - \frac{1}{nk-1}\right) \ .\end{aligned}$$

Exercise 4.1: Solution

Finally, we can plug $\mathbb{V}[\hat{\theta}_{ML}]$ into the expression for $\mathbb{V}[\hat{\theta}_u]$ to obtain

$$\begin{aligned}\mathbb{V}[\hat{\theta}_u] =& \frac{(nk-1)^2}{(nk)^2}\theta^2\frac{(nk)^2}{(nk-1)}\left(\frac{1}{nk-2}-\frac{1}{nk-1}\right)\\ =&\theta^2\left(\frac{nk-1}{nk-2}-1\right)\\ =&\theta^2\left(\frac{nk-1-(nk-2)}{nk-2}\right)=\frac{\theta^2}{nk-2}\ .\end{aligned}$$

We can see that the bias-corrected estimator does not attain the Cramér-Rao Lower Bound:

$$\mathbb{V}[\hat{\theta}_u] = \frac{\theta^2}{nk-2} > \frac{\theta^2}{nk} = \text{CRLB}\ .$$

Exercise 4.2: Solution

Suppose you observe n i.i.d. observations of an outcome variable y_i and a k-dimensional vector of covariates x_i: $\{(y_i, x_i)\}_{i=1:n}$. Suppose further that you want to examine their relation using a linear regression model,

$$y_i = x_i'\beta + u_i \ , \quad u_i|x_i \sim N(0, \sigma^2) \ ,$$

or, in matrix notation,

$$Y = X\beta + U \ ,$$

where Y and U are $(n \times 1)$ and X is $(n \times k)$.

You are exclusively concerned with frequentist/classical estimation, i.e. you consider the parameters β and σ^2 to be fixed and the data $\{(y_i, x_i)\}_{i=1:n}$ to be random (i.e. your particular sample is randomly drawn from some population).

(a) Derive the (conditional) likelihood function $\mathcal{L}(\beta, \sigma^2|Y, X) \equiv p(Y|X, \beta, \sigma^2)$, and define the Maximum Likelihood (ML) estimators $\hat{\beta}$ and $\hat{\sigma}^2$.

Solution: We have

$$y_i \mid x_i, \beta, \sigma^2 \sim N(x_i'\beta, \sigma^2) \ ,$$

i.e.

$$p(y_i|x_i, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \right\} \ .$$

The likelihood function is defined as the joint pdf of all observations, which by independence is equal to the product of the marginal pdfs of the individual observations:

$$\mathcal{L}(\beta, \sigma^2|Y, X) \equiv p(Y|X, \beta, \sigma^2) = \prod_{i=1}^{n} p(y_i|x_i, \beta, \sigma^2) \ .$$

Algebraic manipulations allow us to write the likelihood out explicitly:

$$\begin{aligned}
\mathcal{L}(\beta, \sigma^2|Y, X) =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \right\} \\
=& (2\pi\sigma^2)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \right\} \\
=& (2\pi\sigma^2)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right\} \ .
\end{aligned}$$

Exercise 4.2: Solution

The log-likelihood is then

$$\ell(\beta, \sigma^2 | Y, X) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \ .$$

[3p] The ML estimator is defined as

$$(\hat{\beta}, \hat{\sigma}^2) := \arg\max_{\beta, \sigma^2} l(\beta, \sigma^2 | Y, X) \ . \quad \textbf{[1p]}$$

(b) Derive the ML estimators $\hat{\beta}$ and $\hat{\sigma}^2$.

Hint: for a matrix B and vectors a and c, it holds that $\frac{\partial (a - Bc)'(a - Bc)}{\partial c} = -2B'(a - Bc)$.

Solution:

Using the hint, we get the following First-Order Conditions (FOCs) for β and σ^2:

$$\frac{\partial}{\partial \beta} l(\beta, \sigma^2 | Y, X) = \frac{1}{\sigma^2} X'(Y - X\beta) = 0 \ ,$$
$$\frac{\partial}{\partial \sigma^2} l(\beta, \sigma^2 | Y, X) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta) = 0 \ .$$

[1p each] Rearrange the former to get:

$$X'Y - (X'X)\beta = 0 \quad \Rightarrow \quad (X'X)\beta = X'Y \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'Y \ .$$

Define $S = (Y - X\beta)'(Y - X\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2$ and rearrange the latter to get:

$$-n + \frac{1}{\sigma^2}S = 0 \quad \Rightarrow \quad n = \frac{1}{\sigma^2}S \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n}S \ .$$

[1p each]

(c) Find $\mathbb{E}[\hat{\beta}]$ and $\mathbb{V}[\hat{\beta}] = \mathbb{E}[\Big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\Big)\Big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\Big)']$. What else do you know about the finite sample distribution of $\hat{\beta}$?

Hint: use the Law of Iterated Expectations, computing these two quantities conditional on X first.

Solution:

First of all, note that $\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + U) = \beta + (X'X)^{-1}X'U$.

Exercise 4.2: Solution

The conditional expectation of $\hat{\beta}$ is

$$\begin{aligned}\mathbb{E}[\hat{\beta}|X] &= \beta + \mathbb{E}[(X'X)^{-1}X'U|X] \\ &= \beta + (X'X)^{-1}X'\mathbb{E}[U|X] \\ &= \beta \ .\end{aligned}$$

The conditional variance of $\hat{\beta}$ is

$$\begin{aligned}\mathbb{V}[\hat{\beta}|X] &= \mathbb{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] = \\ &= \mathbb{E}[((X'X)^{-1}X'U)((X'X)^{-1}X'U)'|X] = \\ &= \mathbb{E}[(X'X)^{-1}X'UU'X(X'X)^{-1}|X] = \\ &= (X'X)^{-1}X'\mathbb{E}[UU'|X]X(X'X)^{-1} = \\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} = \\ &= \sigma^2(X'X)^{-1} \ .\end{aligned}$$

By LIE, we have then

$$\mathbb{E}_X[\mathbb{E}[\hat{\beta}|X]] = \mathbb{E}[\hat{\beta}] = \beta \ ,$$

as well as

$$\mathbb{V}[\hat{\beta}] = \mathbb{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = \mathbb{E}[\mathbb{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X]] = \sigma^2\mathbb{E}[(X'X)^{-1}] \ .$$

Thus, we know that (in finite samples) $\hat{\beta}$ has mean β and variance $\sigma^2\mathbb{E}[(X'X)^{-1}]$. **[2p each]**

When conditioning on X, we know not only the mean and variance of $\hat{\beta}|X$, but we also know that $\hat{\beta}|X$ is normally distributed, as it is the (weighted) sum of Normally distributed Random Variables (RVs):

$$\hat{\beta}|X = \beta + (X'X)^{-1}X'U|X \sim N(\beta, \sigma^2(X'X)^{-1}) \ ,$$

as $U|X \sim N(0, \sigma^2 I)$ and the X-terms are just constants when we condition on X (and β is a constant under our frequentist perspective). **[2p]**

However, note that we do not know the unconditional distribution of $\hat{\beta}$ (i.e. we do not know whether it is Normally distributed or not in finite samples), we only know its mean and variance. This is because we do not know (we did not make any assumptions on) the distribution of X.

Exercise 4.2: Solution

(d) Is your estimator $\hat{\beta}$ consistent? Use the expression that you derived for it in (b) and theorems to answer this question.

Solution:

Since our sample is i.i.d., we can apply the WLLN and Slutsky's theorem to prove that $\hat{\beta}$ is consistent. We have

$$\hat{\beta} - \beta = \left(\frac{X'X}{n}\right)^{-1} \frac{X'U}{n} = \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \frac{1}{n}\sum_i x_i u_i \ .$$

By WLLN,

$$\frac{1}{n}\sum_i x_i x_i' \xrightarrow{p} \mathbb{E}[X'X] \equiv Q \ ,$$

and then by Slutsky,

$$\left(\frac{X'X}{n}\right)^{-1} = \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \xrightarrow{p} \mathbb{E}[(X'X)^{-1}] = Q^{-1} \ .[\mathbf{1p}]$$

Also by WLLN, we have

$$\frac{1}{n}X'U = \frac{1}{n}\sum_i x_i u_i \xrightarrow{p} \mathbb{E}[x_i u_i] = 0 \ . \quad [\mathbf{1p}]$$

Hence, putting the two pieces together (again using Slutsky's theorem), we get

$$\hat{\beta} - \beta \xrightarrow{p} Q^{-1}0 = 0 \ . \quad [\mathbf{1p}]$$

(e) What is the asymptotic distribution of $\hat{\beta}$? Use the expression that you derived for it in (b) and theorems to answer this question.

Solution:

Since our sample is i.i.d., we can apply the CLT to prove that $\hat{\beta}$ is asymptotically Normal. We have

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_i x_i u_i \ .$$

Exercise 4.2: Solution

By CLT,

$$\frac{1}{\sqrt{n}}\sum_i x_i u_i = \sqrt{n}\left(\frac{1}{n}\sum_i x_i u_i - \mathbb{E}[x_i u_i]\right) \xrightarrow{d} N(0, \mathbb{E}[(x_i u_i)(x_i u_i)']) \ ,$$

where we know that $\mathbb{E}[x_i u_i] = \mathbb{E}[\mathbb{E}[x_i u_i|x_i]] = \mathbb{E}[x_i \mathbb{E}[u_i|x_i]] = 0$ and

$$\mathbb{E}[(x_i u_i)(x_i u_i)'] = \mathbb{E}[x_i x_i' u_i^2] = \mathbb{E}[\mathbb{E}[x_i x_i' u_i^2|x_i]] = \mathbb{E}[x_i x_i' \mathbb{E}[u_i^2|x_i]] = \sigma^2 \mathbb{E}[x_i x_i'] = \sigma^2 Q \ .$$

[2p] From before, we know that $\left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \xrightarrow{p} Q^{-1}$. Putting the two pieces together using Slutsky's theorem, we get

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} Q^{-1} N(0, \sigma^2 Q) Q^{-1} = N(0, \sigma^2 Q^{-1}) \ . \quad \textbf{[2p]}$$

(f) Suppose you have $k = 4$ regressors (including the intercept: $x_{i1} = 1 \ \forall \ i$), and you want to test $\mathcal{H}_0 : \{ \ \beta_2 + 3\beta_3 = 7 \ , \ \log \beta_4 = 0 \ \}$ against the alternative $\mathcal{H}_1$, specifying that at least one of the two conditions in $\mathcal{H}_0$ is not true. Describe two approaches to conduct this test and their relative advantages/disadvantages.

Solution:

[max. 3p per approach, max. 6p overall]

Approach 1: Wald test. **[1p]** We are testing $\mathcal{H}_0 : g(\beta) = 0$, where

$$g(\beta) = \begin{bmatrix} 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \beta - \begin{bmatrix} 7 \\ 1 \end{bmatrix}$$

(because testing $\log(\beta_4) = 0$ is equivalent to testing $\beta_4 = 1$).[1] We have the test-statistic

$$T_W = n g(\hat{\beta})[G \hat{V} G]^{-1} g(\hat{\beta})' \to \chi^2_2 \ ,$$

where

$$G = \frac{\partial g(\beta)}{\partial \beta} = \begin{bmatrix} 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \ ,$$

[1] Alternatively, one can also use

$$g(\beta) = \begin{bmatrix} 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \beta + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} log \ \beta - \begin{bmatrix} 7 \\ 0 \end{bmatrix} \ .$$

Exercise 4.2: Solution

and

$$V = A\mathbb{V}[\hat{\beta}] = \sigma^2 Q^{-1} \ ,$$

with consistent estimator $\hat{V} = \hat{\sigma}^2\hat{Q}^{-1}$. The advantage of the Wald test is that we only need the unrestricted estimator $\hat{\beta}$, not an additional, unrestricted estimator that imposes $\mathscr{H}_0$ in the model. **[2p]**

Approach 2: Likelihood Ratio (LR) test. **[1p]** It uses the test statistic

$$T_{LR} = -2[\ell(\hat{\beta}|Y,X) - \ell(\bar{\beta}|Y,X)] \to \chi^2_2 \ ,$$

where $\ell(\beta|Y,X)$ is log-likelihood function defined above, $\hat{\beta}$ is estimator of unrestricted model,

$$y_i = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u_i \ , \quad u_i \sim N(0,\sigma^2) \ ,$$

and $\bar{\beta} = [\bar{\beta}_1, (7-3\bar{\beta}_3), \bar{\beta}_3, 1]$ is estimator of restricted model,

$$y_i = \beta_1 + (7-3\beta_3)x_2 + \beta_3 x_3 + 1x_4 + v_i \ , \quad v_i \sim N(0,\sigma^2) \ .$$

The advantage of the LR test is that it is the uniformly most powerful test. The disadvantage is that we need $\hat{\beta}$ as well as $\bar{\beta}$. **[2p]**

Approach 3: Lagrange Multiplier (LM) test. **[1p]** It uses the test-statistic

$$T_{LM} = s(\bar{\beta}|Y,X)' I(\bar{\beta})^{-1} s(\bar{\beta}|Y,X) \to \chi^2_2 \ ,$$

where

$$s(\bar{\beta}|Y,X) = \frac{\partial}{\partial\beta} l(\beta|Y,X)|_{\beta=\bar{\beta}} = \frac{1}{\sigma^2}\sum_i x_i(y_i - x_i'\bar{\beta})$$

is the score evaluated at $\bar{\beta}$, and

$$I(\bar{\beta}) = -H(\beta|Y,X) = -\frac{\partial}{\partial\beta} s(\beta|Y,X) = \frac{1}{\sigma^2}\sum_i x_i x_i' \ .$$

The advantage of this test is that we need only $\bar{\beta}$. **[2p]**

(g) Construct a 95% confidence interval for $exp\{\beta_4\}$.

Solution: There are several approaches to this question. The easiest approach is to find a CI for β_4 and then turn it into a CI for $exp\{\beta_4\}$, which works because the exponential is a monotonic function. We can find a CI for β_4 by using the finite sample distribution of $\hat{\beta}_4$. Because we do not know it, we approximate it using the asymptotic

Exercise 4.2: Solution

distribution (see exercise (f) above):

$$\hat{\beta}_4 \overset{approx.}{\sim} N\left(\beta_4, \hat{V}\right) \ , \quad \text{where } \hat{V} = \frac{1}{n}\hat{\sigma}^2 \left(\hat{Q}^{-1}\right)_{44} \ ,$$

and $\left(\hat{Q}^{-1}\right)_{44}$ denotes the element (4,4) in the matrix $\hat{Q}^{-1}$. (Alternatively, if we condition on X, then this distribution is exact already in finite samples; see exercise (c).) **[1p]** Given this Normal distribution, we can construct the 95% CI for β_4 as usual:

$$\text{CI}_{95\%}(\hat{\beta}_4) = \left[\hat{\beta}_4 - 1.96\sqrt{\hat{V}} \ , \ \hat{\beta}_4 + 1.96\sqrt{\hat{V}}\right] \ . \quad \textbf{[1p]}$$

We can then turn it into a CI for $exp\{\beta_4\}$ by taking the exponent of both bounds of this interval:

$$\text{CI}_{95\%}(exp\{\beta_4\}) = \left[exp\left\{\hat{\beta}_4 - 1.96\sqrt{\hat{V}}\right\} \ , \ exp\left\{\hat{\beta}_4 + 1.96\sqrt{\hat{V}}\right\}\right] \ . \quad \textbf{[2p]}$$

This works because if the probability that β_4 lies in $\text{CI}_{95\%}$ is 95%, then the probability that $exp\{\beta_4\}$ lies in $\text{CI}_{95\%}(exp\{\beta_4\})$ is 95%, too, i.e. we have a valid CI for $exp\{\beta_4\}$. Note that this method does take into account that $exp\{\beta_4\} > 0$, as we get a CI that lies on $\mathbb{R}_{++}$.

A related approach (which should give the same results...) is to conclude that – because $\hat{\beta}_4$ is Normal – $exp\{\hat{\beta}_4\}$ is log-Normal, and construct directly a CI for $exp\{\hat{\beta}_4\}$ based on the log-Normal distribution.

Another approach is to use the Delta method to find the asymptotic distribution of $exp\{\hat{\beta}_4\}$ and construct an asymptotically valid CI for $exp\{\beta_4\}$. We know

$$\sqrt{n}(\hat{\beta}_4 - \beta_4) \xrightarrow{d} N(0, V) \ , \quad V = A\mathbb{V}[\hat{\beta}_4] = \sigma^2(\mathbb{E}[x_i x_i']^{-1})_{(4,4)} \ .$$

Defining $g(\hat{\beta}_4) = \exp\left\{\hat{\beta}_4\right\}$, the Delta method tells us that

$$\sqrt{n}(g(\hat{\beta}_4) - g(\beta_4)) \xrightarrow{d} N(0, GVG') = N(0, G^2 V) \ ,$$

where

$$G = \frac{\partial g(\beta_4)}{\partial \beta_4} = \exp\{\beta_4\} \ .$$

Exercise 4.2: Solution

Based on this, we can approximate the finite sample distribution of $g(\hat{\beta}_4)$ as

$$g(\hat{\beta}_4) \overset{\text{approx.}}{\sim} N\left(g(\beta_4)\ ,\ \frac{1}{n}\hat{G}^2\hat{V}\right) = N\left(g(\beta_4)\ ,\ \frac{1}{n}\exp\left\{2\hat{\beta}_4\right\}\hat{\sigma}^2\left(\left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1}\right)_{(4,4)}\right)\ ,$$

[2p] and, given this Normal distribution, we can construct the 95% CI as usual:

$$\text{CI}_{95\%} = \left[\exp\left\{\hat{\beta}_4\right\} - 1.96\sqrt{n^{-1}\hat{G}^2\hat{V}}\ \ ,\ \ \exp\left\{\hat{\beta}_4\right\} + 1.96\sqrt{n^{-1}\hat{G}^2\hat{V}}\right]\ . \quad \textbf{[2p]}$$

Note that this approach is only valid asymptotically, and it does not restrict $g(\beta_4)$ to be positive. Hence, it is very well possible that one gets a negative lower bound of the CI (in any finite sample). That's why the approaches above are preferred.

[max. 4p for either of the two approaches, max. 2p for sketching the approach via log-Normal, max. 2p for discussing the relation of approaches, max. 4p overall]

Exercise 4.3: Solution

The probit model is a modeling approach for binary outcome variables $y_i \in \{0,1\}$. It specifies a latent variable model,

$$y_i^* = x_i'\beta + u_i \quad , \quad u_i|x_i \sim N(0,1) \ ,$$

and assumes we observe $y_i = \mathbf{1}\{y_i^* \geqslant 0\}$.

(a) Show that the assumption that the variance of $u_i|x_i$ is equal to one is without loss of generality. To answer the question, redo the calculations (likelihood derivation and calculation of partial effects) with $u_i|x_i \sim N(0,\sigma^2)$ instead, and use your knowledge on properties of partial effects under the probit model.

(b) The logit model is very similar to the probit model. The only difference is that it assumes that $u_i|x_i$ follows a standard logistic distribution instead of the standard Normal distribution. Its cdf is then

$$F(x) = 1/(1+exp\{-x\})$$

instead of $\Phi(x)$. Redo the calculations (likelihood derivation and calculation of partial effects) for the logit model.

Solution:

(a) We take the latent variable model

$$\tilde{y}_i^* = x_i'\tilde{\beta} + \tilde{u}_i \quad , \quad \tilde{u}_i|x_i \sim N(0,\sigma^2) \ ,$$

and we observe $y_i = 1$ when the latent variable $\tilde{y}_i^*$ lies above the threshold of zero, and zero otherwise:

$$y_i = \mathbf{1}\{\tilde{y}_i^* \geq 0\} \ .$$

We can scale this latent variable model by the constant $1/\sigma$ to obtain the transformed model

$$\frac{\tilde{y}_i^*}{\sigma} = \frac{x_i'\tilde{\beta} + \tilde{u}_i}{\sigma} \ , \quad \frac{\tilde{u}_i}{\sigma}|x_i \sim N(0,1) \ ,$$

and

$$y_i = \mathbf{1}\{\tilde{y}_i^* \geq 0\} = \mathbf{1}\left\{\frac{\tilde{y}_i^*}{\sigma} \geq 0\right\} \ .$$

In this model, the probability of $y_i = 1$ is given by:

$$\mathbb{P}[y_i = 1] = \mathbb{P}\left[\frac{\tilde{y}_i^*}{\sigma} \geq 0\right] = \mathbb{P}\left[\frac{x_i'\tilde{\beta} + \tilde{u}_i}{\sigma} \geq 0\right] = \mathbb{P}\left[\frac{\tilde{u}_i}{\sigma} \geq -\frac{x_i'\tilde{\beta}}{\sigma}\right] = 1 - \Phi\left(-\frac{x_i'\tilde{\beta}}{\sigma}\right) = \Phi\left(\frac{x_i'\tilde{\beta}}{\sigma}\right) \ .$$

Exercise 4.3: Solution

The observed binary outcome variable follows a Bernoulli distribution:

$$p(y_i|x_i, \tilde{\beta}, \sigma) = \begin{cases} \Phi\left(\frac{x_i'\tilde{\beta}}{\sigma}\right) & \text{if } y_i = 1 \\ \Phi\left(-\frac{x_i'\tilde{\beta}}{\sigma}\right) & \text{if } y_i = 0 \end{cases} .$$

We get the likelihood

$$\mathcal{L}(\tilde{\beta}, \sigma|Y, X) = \prod_{i=0}^{n} \Phi(x_i'\tilde{\beta}/\sigma)^{y_i} \Phi(-x_i'\tilde{\beta}/\sigma)^{(1-y_i)} .$$

Note that we are not able to estimate both parameters $\tilde{\beta}$ and σ, as the two are not jointly identified: if $(\hat{\tilde{\beta}}, \hat{\sigma})$ is a maximizer of the likelihood, then so is $(c\hat{\tilde{\beta}}, c\hat{\sigma})$ for any $c > 0$, i.e. there are infinitely many combinations of $\tilde{\beta}$ and σ that return the same value of the likelihood function. These combinations also return the same value of partial effects:

$$\frac{\partial}{\partial x}\mathbb{E}[y_i|x_i = x] = \frac{\partial}{\partial x}\mathbb{P}[y_i = 1|x_i = x] = \frac{\partial}{\partial x}\Phi\left(\frac{x_i'\tilde{\beta}}{\sigma}\right) = \frac{\tilde{\beta}}{\sigma}\phi\left(\frac{x_i'\tilde{\beta}}{\sigma}\right) ,$$

as multiplying both $\tilde{\beta}$ and σ by $c > 0$ does not change the partial effects.

To be able to estimate the model, we define $\beta \equiv \tilde{\beta}/\sigma$ and use the likelihood function

$$\mathcal{L}(\beta|Y, X) = \prod_{i=0}^{n} \Phi(x_i'\beta)^{y_i} \Phi(-x_i'\beta)^{(1-y_i)} .$$

This is the same likelihood as the one obtained under $\sigma = 1$. In other words, it is the likelihood obtained if we transform our latent variable model by dividing by σ (as we did above) and re-defining our latent variable as $y_i^* = \frac{\tilde{y}_i^*}{\sigma}$, our error term as $u_i = \frac{\tilde{u}_i}{\sigma}$ and our parameter-vector as $\beta = \tilde{\beta}/\sigma$:

$$\frac{y_i^*}{\sigma} = \frac{x_i'\beta + u_i}{\sigma} \;,\quad \frac{u_i}{\sigma}|x_i \sim N(0,1) \quad\Rightarrow\quad y_i^* = x_i'\beta + u_i \;,\quad u_i|x_i \sim N(0,1) .$$

(b) We have the following latent variable model:

$$y_i^* = x_i'\beta + u_i \;,\quad u_i|x_i \sim L(0,1)$$

Exercise 4.3: Solution

and binary outcome variable

$$y_i = \mathbf{1}\{y_i^* \geq 0\} \ .$$

The error term $u_i|x_i$ is distributed according to the standard logistic distribution, with cdf

$$F(x) = \frac{1}{1+\exp\{-x\}}$$

and pdf

$$f(x) = \frac{\partial F(x)}{\partial x} = \frac{-\exp\{-x\} \times (-1)}{(1+\exp\{-x\})^2} = \frac{\exp\{-x\}}{(1+\exp\{-x\})^2} \ .$$

From this, we get the following probability of $y_i = 1$:

$$\begin{aligned}
\mathbb{P}[y_i = 1] =& \mathbb{P}[y_i^* \geq 0] \\
=& \mathbb{P}[x_i'\beta + u_i \geq 0] \\
=& \mathbb{P}[u_i \geq -x_i'\beta] \\
=& 1 - \mathbb{P}[u_i \leq -x_i'\beta] \\
=& 1 - F(-x_i'\beta) \\
=& 1 - \frac{1}{1+\exp\{x_i'\beta\}} \\
=& \frac{\exp\{x_i'\beta\}}{1+\exp\{x_i'\beta\}} \ .
\end{aligned}$$

Moreover, we have that:

$$\mathbb{P}[y_i = 0] = 1 - \mathbb{P}[y_i = 1] = \frac{1}{1+\exp\{x_i'\beta\}} = F(-x_i'\beta) \ .$$

The binary outcome variable y_i follows therefore a Bernoulli distribution with probability of success $F(x_i'\beta)$:

$$p(y_i|x_i,\beta) = \begin{cases} \dfrac{\exp\{x_i'\beta\}}{1+\exp\{x_i'\beta\}} = F(x_i'\beta) & \text{if } y_i = 1 \\ \dfrac{1}{1+\exp\{x_i'\beta\}} = F(-x_i'\beta) & \text{if } y_i = 0 \end{cases} \ .$$

From this, we get the likelihood function

$$\mathcal{L}(\beta|Y,X) = \prod_{i=1}^n F(x_i'\beta)^{y_i} F(-x_i'\beta)^{(1-y_i)} = \prod_{i=1}^n \left(\frac{\exp\{x_i'\beta\}}{1+\exp\{x_i'\beta\}}\right)^{y_i} \left(\frac{1}{1+\exp\{x_i'\beta\}}\right)^{(1-y_i)} ,$$

Exercise 4.3: Solution

the log-likelihood

$$\begin{aligned}\ell(\beta|Y,X) &= \sum_{i=1}^{n} y_i(x_i'\beta - \log(1+\exp\{x_i'\beta\}) + (1-y_i)(-\log(1+\exp\{x_i'\beta\})) \\ &= \sum_{i=1}^{n} y_i x_i'\beta - \log(1+\exp\{x_i'\beta\})\ .\end{aligned}$$

We cannot find $\hat{\beta}_{logit} \equiv \arg\max_{\beta}\ \ell(\beta|Y,X)$ analytically, but have to use numerical optimisation methods.

For a given parameter β, the partial effects of a change in the exogenous variable x_i are given by:

$$\begin{aligned}\frac{\partial}{\partial x}\mathbb{E}[y_i|x_i=x] =& \frac{\partial}{\partial x}\mathbb{P}(y_i=1|x_i=x) \\ =& \beta\frac{\partial}{\partial x}F(x_i'\beta) \\ =& \beta f(x_i'\beta) \\ =& \beta\frac{\exp\{-x_i'\beta\}}{(1+\exp\{-x_i'\beta\})^2}\ .\end{aligned}$$

Exercise 4.4: Solution

Suppose you have a dataset containing a shop's sales, which includes the date, some characteristics of the customer (like income, age), some characteristics of the transaction (like type of good sold, price, and whether cash or a card was used). You are interested in shedding light on the determinants of cash vs card payment.

(a) How could you use the probit model for your research question? What is your y_i variable? How can we interpret the underlying latent variable y_i^*?

Solution: The observed outcome variable y_i is a binary variable taking value one when a payment is settled in cash and zero otherwise:

$$y_i = \mathbf{1}\{y_i^* > 0\}$$

The underlying latent (i.e. unobserved) variable y_i^* can be interpreted as the customer's willingness to pay in cash; if it is above a certain threshold (normalized to zero without loss of generality) the payment actually takes place in cash.

(b) In your probit model, derive the effect of age increasing by 5 years on the probability of using cash. Does the effect depend on the current age of the customer? Does it depend on the values of the other variables?

Solution: In the probit model

$$\mathbb{E}[y_i|x_i] = \Phi(x_i'\beta) \ .$$

When the explanatory variable changes from x_1 to x_2 by $\Delta x_i = x_2 - x_1$, the corresponding change in the dependent variable is given by:

$$\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_2] = \Phi(x_2'\beta) - \Phi(x_1'\beta) \ .$$

For small changes in the explanatory variable, this effect can be approximated by the first derivative:

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{i,c}} = \frac{\partial \Phi(x_i'\beta)}{\partial x_{i,c}} = \phi(x_i'\beta)\beta_c$$

for any covariate $c = 1, 2, ...k$. Therefore, for a discrete change in x_c by 5 units

Exercise 4.4: Solution

($\Delta x_c = 5$), the approximate expected change in the dependent variable is:

$$\Delta \mathbb{E}[y_i|x_i] = \Phi(x_2'\beta) - \Phi(x_1'\beta) \approx \Delta x_c \phi(x_i'\beta)\beta = 5 \times \phi(x_i'\beta)\beta_c \ ,$$

whereby x_1 is the vector of regressors that includes the current age of the customer and x_2 is the same vector of regressors (same values), but with age increased by 5 years.

Notice that the marginal effect of a change in the regressor $x_{i,c}$ depends not only on the level of $x_{i,c}$ itself, but also on the level of all other regressors in the vector of explanatory variables x_i.

(c) Could you use a standard linear regression, estimated via OLS, to answer your question?

Solution: Yes, in principle OLS could be used to estimate the probability of a purchase being made in cash. The problem is that the predicted values of the dependent variable under OLS, $\hat{y}_i = x_i'\hat{\beta}_{OLS}$ would not be bounded between zero and one. Therefore, they cannot be interpreted as probabilities of a payment being settled in cash.

(d) Derive the same effect as in (b) in your linear regression model. Does it depend on the current age of the customer? Does it depend on the values of the other variables?

Solution: When estimating the OLS model $y_i = x_i'\beta + u_i$, the marginal effect of an increase in x would be:

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{i,c}} = \frac{\partial x_i'\beta}{\partial x_{i,c}} = \beta_c$$

For any covariate $c = 1, 2, ...k$. Hence, the marginal effect of a change in the regressor x_c is constant and does not depend on the level of $x_{i,c}$, as the OLS model is linear.

(e) Based on your reasoning so far, for which customers would you expect the predicted effect under the linear regression to be close to the one under probit? For what type of customers will the two differ more? As a result, for what kind of research questions is the linear regression a good/bad specification?
Hint: Besides comparing (partial) effects under the two models, you might want to compare the functional form of $\mathbb{E}[y_i|x_i]$ under the two models.

Exercise 4.4: Solution

Solution: The predicted effects under any model are simply given by the difference in the conditional expectation $\mathbb{E}[y_i|x_i]$ between $x_i = x_1$ and $x_i = x_2$ (e.g. age going up 5 years while all other covariates stay constant). In the probit model, this condictional expectation is a non-linear function of covariates x_i: $\mathbb{E}[y_i|x_i] = \Phi(x_i'\beta)$. By properties of Φ, it goes from 0 to 1, whereby it starts pretty flat for small values of $x_i'\beta$, increases gradually and becomes pretty linear around $x_i'\beta = 0$, after which it starts to flatten out. In contrast, under the linear regression, $\mathbb{E}[y_i|x_i] = x_i'\beta$ is always a linear function of covariates x_i.

As a result, the effects under the two models will be the closest when $\hat{y}_i = x_i'\hat{\beta}$ is around zero, i.e. when the predicted probability of cash-payment is around 50%. Hence, for the customers who are at the margin between cash- and card-payment, the marginal effect estimated with OLS will be closest to the marginal effect estimated under probit. Conversely, for the more extreme values of $x_i'\hat{\beta}$ (i.e. for the more "extreme" customers; the (predictedly) convinced card- or cash-users), the marginal effects estimated under OLS will differ the most from those estimated under probit. Hence, if our research question explores customers "at the margin", we could also use OLS, while to get sensible results for the customers "at the extremes", we should use probit.

Exercise 4.5: Solution

Suppose you are interested in relating air quality in different cities – measured by the concentration of carbon monoxide in the air, y_i – to possible determinants x_i. The measurement device used in your data cannot detect concentrations below a certain value, δ, but simply codes them as zero. For this purpose, you set up a Tobit model for observations y_i with a lower-censoring at δ:

$$\begin{aligned} y_i^* &= x_i'\beta + u_i \;, \quad u_i \sim N(0, \sigma^2) \;, \\ y_i &= y_i^* \, \mathbf{1}\{y_i^* > \delta\} \;. \end{aligned} \tag{1}$$

(a) Derive the probability of measuring a concentration of carbon monoxide of zero as a function of determinants x_i (and parameters β and the censoring point δ), $\mathbb{P}[y_i = 0|x_i]$.

Solution:

$$\begin{aligned} \mathbb{P}[y_i = 0] &= \mathbb{P}[y_i^* \leq \delta] \\ &= \mathbb{P}[x_i'\beta + u_i \leq \delta] \\ &= \mathbb{P}[u_i \leq \delta - x_i'\beta] \\ &= \Phi\left(\frac{\delta - x_i'\beta}{\sigma}\right) \end{aligned}$$

(b) Derive the conditional mean $\mathbb{E}[y_i^*|x_i]$, i.e. the expected air quality (true concentration of carbon monoxide) for generic a city i with characteristics x_i.

Solution:

$$\mathbb{E}[y_i^*|x_i] = \mathbb{E}[x_i'\beta + u_i|x_i] = x_i'\beta$$

(c) Derive the conditional mean $\mathbb{E}[y_i|x_i]$, i.e. the expected (measurable) concentration of carbon monoxide for generic a city i with characteristics x_i.
Hint: recall that for $z_i \sim N(0,1)$, $\mathbb{E}[z_i|z_i > -c] = \phi(c)/\Phi(c)$ (Inverse-Mills ratio).

Exercise 4.5: Solution

Solution: For simplicity, we omit the conditioning on x_i. We have

$$\begin{aligned}
\mathbb{E}[y_i] &= \mathbb{E}[y_i|y_i^* \leq \delta]\mathbb{P}[y_i^* \leq \delta] + \mathbb{E}[y_i|y_i^* > \delta]\mathbb{P}[y_i^* > \delta] \\
&= \mathbb{E}[y_i|y_i^* > \delta](1 - \mathbb{P}[y_i^* \leq \delta]) \\
&= \mathbb{E}[y_i|u_i > \delta - x_i'\beta](1 - \mathbb{P}[u_i \leq \delta - x_i'\beta]) \\
&= (x_i'\beta + \mathbb{E}[u_i|u_i > \delta - x_i'\beta])\,(1 - \mathbb{P}[u_i \leq \delta - x_i'\beta]) \\
&= \left(x_i'\beta + \mathbb{E}\left[z_i|z_i > \frac{\delta - x_i'\beta}{\sigma}\right]\right)\left(1 - \mathbb{P}\left[z_i \leq \frac{\delta - x_i'\beta}{\sigma}\right]\right) \\
&= \left(x_i'\beta + \sigma\frac{\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)}\right)\Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) \\
&= \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \sigma\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)\ ,
\end{aligned}$$

where $z_i = u_i/\sigma \sim N(0,1)$.

(d) Suppose one of your variables in x_i is the cost of public transport as a fraction of the average hourly wage in the city, c_i. Using your result from the previous two exercises, derive the predicted effect of decreasing this ratio by 10 percentage points on air quality y_i^* and measured carbon monoxide concentration y_i.

Solution: We can compute the predicted change in y_i when we change x_i from x_1 to x_2 by by taking the difference in the conditional expectation function:

$$\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_1] = \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \sigma\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)\ .$$

In our case, x_1 includes c_i and x_2 includes $c_i - 10$ (or $c_i - 0.1$ if c_i is measured as a decimal rather than in percent).

Sidenote: for small changes $\Delta x_i = x_2 - x_1$, the above is approximately equal to Δx_i (a vector of dimension k) multiplied by the first derivative of $\mathbb{E}[y_i|x_i]$ w.r.t. x_i (also a vector of dimension k),

$$\frac{\partial\mathbb{E}[y_i|x_i]}{\partial x_i} = \frac{\beta}{\sigma}\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \beta\Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) + \beta\phi'\left(\frac{x_i'\beta - \delta}{\sigma}\right)\ .$$

If we only change one of the covariates $c = 1, 2, ..., k$ in x_i, we get the marginal effect

$$\Delta\mathbb{E}[y_i|x_i] \approx \Delta x_{ic} \times \left(\frac{\beta_c}{\sigma}\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)\beta_c + \beta_c\phi'\left(\frac{x_i'\beta - \delta}{\sigma}\right)\right)\ ,$$

Exercise 4.5: Solution

whereby $\Delta x_{ic} = -10$ (or -0.1) in our case.

Notice that this marginal effect depends on the level of $x_{i,c}$, as well as on the levels of all the other elements of the vector of covariates x_i.

(e) Instead, suppose you simply use a linear regression to relate y_i to x_i for the cities for whom the concentration was measured precisely, i.e. for cities $i \in \mathcal{U} \equiv \{i : y_i > \delta\}$:

$$y_i = x_i'\gamma + v_i \ , \quad i \in \mathcal{U} \ . \tag{2}$$

What is the effect of decreasing c_i on y_i in this specification? Presuming for a moment that γ and β are the same thing, for which cities is the predicted effect under the linear regression close to/far from the one under the above tobit model?

Solution: In the case of a linear regression, we have

$$\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_1] = x_2'\gamma - x_1'\gamma = \Delta x_i'\gamma \ .$$

When we change a single regressor c, we get $\gamma_c \Delta x_{ic}$. (Sidenote: since the conditional mean is linear in x_i, the exact partial effect and the approximate one that uses first derivatives coincide.)

For ease of exposition, we compare this to the approximate marginal effect under Tobit (rather than the exact one),

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{i,c}} = \beta_c \left(\frac{1}{\sigma}\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) x_i'\beta + \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) + \phi'\left(\frac{x_i'\beta - \delta}{\sigma}\right)\right) \ .$$

Assuming that $\beta = \gamma$, then the two marginal effects will be closer when the term in brackets is close to one. By recalling the properties of Normal pdf's and cdf's, we know that:

- $\phi(z) \in (0,1)$ for $z \in \mathbb{R}$ and that $\phi(z) \to 0$ for $z \to \pm\infty$
- $\phi'(z) \to 0$ for $z \to \pm\infty$
- $\Phi(z) \to 1$ as $z \to +\infty$ and that $\Phi(z) \to 0$ as $z \to -\infty$

By putting the three pieces together, we know that the expression in brackets will be closer to 1 for the highest values of the distribution of x_i and it will be closer to zero for the lowest values of x_i. Therefore, we know that the marginal effects estimated under OLS will be the closest to those estimated under tobit for the most polluted

Exercise 4.5: Solution

cities (i.e. the cities farthest away from the censoring point of δ) and they will be the most distant from those estimated under tobit for the least polluted cities.

(f) (Bonus question) You are in fact not interested in relating y_i to x_i, but in relating the true air quality y_i^* – of which y_i is an imperfect measure – to x_i, i.e. you are interested in β, not γ. Supposing that Eq. (1) is the true model generating the data, can you use the OLS estimator for γ from Eq. (2) to consistently estimate β? Under which circumstances will OLS work better/worse?
Hint: For $i \in \mathcal{U}$, we simply have $y_i = y_i^* = x_i'\beta + u_i$. Also, for a generic random variable z_i,

$$\frac{1}{n_u}\sum_{i\in\mathcal{U}} z_i \xrightarrow{p} \mathbb{E}[z_i|y_i > \delta] = \mathbb{E}[z_i|y_i^* > \delta] = \mathbb{E}[z_i|u_i > \delta - x_i'\beta] \ ,$$

where $n_u = |\mathcal{U}|$ is the number of observations i in $\mathcal{U}$.

Solution: The OLS estimator is given by

$$\begin{aligned}
\hat{\gamma} &= \left[\frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i x_i'\right]^{-1} \frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i y_i \\
&= \left[\frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i x_i'\right]^{-1} \frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i(x_i'\beta + u_i) \\
&= \beta + \left[\frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i x_i'\right]^{-1} \frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i u_i \\
&= \beta + \left[\frac{1}{n}\sum_{i} x_i x_i' \, \mathbf{1}\{y_i^* > \delta\}\right]^{-1} \frac{1}{n}\sum_{i} x_i u_i \, \mathbf{1}\{y_i^* > \delta\} \\
&\xrightarrow{p} \beta + \mathbb{E}[x_i x_i | y_i^* > \delta]^{-1}\mathbb{E}[x_i u_i | y_i^* > \delta] \\
&= \beta + \mathbb{E}[x_i x_i | u_i > \delta - x_i'\beta]^{-1}\mathbb{E}[x_i \mathbb{E}[u_i | u_i > \delta - x_i'\beta, x_i]] \\
&= \beta + \mathbb{E}[x_i x_i | u_i > \delta - x_i'\beta]^{-1}\mathbb{E}\left(x_i \sigma \frac{\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta-\delta}{\sigma}\right)}\right) \\
&\neq \beta \ .
\end{aligned}$$

Exercise 4.5: Solution

The OLS estimator is not consistent. By the LIE, the asymptotic bias equals

$$\begin{aligned}\mathbb{E}[x_i x_i | y_i^* > \delta]^{-1}\mathbb{E}[x_i u_i | y_i^* > \delta] &= \mathbb{E}[x_i x_i | u_i > \delta - x_i'\beta]^{-1}\mathbb{E}[x_i \mathbb{E}[u_i | u_i > \delta - x_i'\beta, x_i]] \\ &= \mathbb{E}[x_i x_i | u_i > \delta - x_i'\beta]^{-1}\mathbb{E}\left(x_i \sigma \frac{\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)} \right) \,.\end{aligned}$$

Its size depends on the magnitude of the ratio $\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) / \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)$. By the properties of ϕ and Φ (see solution to previous exercise), the asymptotic bias is expected to be high (low) if the "average" $x_i'\beta$ is close to (far above of) the censoring point δ as then $x'\beta - \delta$ is close to (far above of) zero and, therefore, $\phi(x'\beta - \delta)$ is high (low). In our application, we get a low bias if we have lots of polluted cities in our sample. If the average $x_i'\beta$ is far below δ, it's not immediately clear what happens, as not only the numerator $\phi(x'\beta - \delta)$ is close to zero, but also the denominator, $\Phi(x'\beta - \delta)$.

Exercise 4.6: Solution

Suppose you observe a sample of n unemployed invidiuals. Let y_i denote the time (in weeks) that individual i spent in unemployment, and let $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$ be the average unemployment span in your sample. You can assume that your observations are independent.

One can model y_i using an expontential distribution:

$$p(y_i|\ \lambda) = \lambda exp\left\{-\lambda y_i\right\} \quad , \quad \lambda > 0 \ .$$

The parameter λ is the job-finding rate. It tells you how many (acceptable) job offers per week an individual gets. For example, $\lambda = 3$ would tell you that (on average) an individual receives three offers every week, while $\lambda = 1/2$ would tell you that (on average), an individual receives an offer every two weeks. For now, we assume that this λ is the same for all individuals in the sample.

(a) (4 points) The mean and variance of y_i are given by

$$\mathbb{E}[y_i] = \frac{1}{\lambda} \quad \text{and} \quad \mathbb{V}[y_i] = \frac{1}{\lambda^2} \ .$$

Interpret these expressions, relying on the two examples $\lambda = 3$ and $\lambda = 1/2$.

Solution: The expected time spent in unemployment is the inverse of the rate at which (acceptable) job offers arrive. For example, if $\lambda = 3$, the individual receives 3 job offers per week, which means that they are expected to be unemployed for 1/3 of a week. If $\lambda = 1/2$, they receive one job offer every five weeks, which means that they are expected to be unemployed for 2 weeks.

There is more uncertainty around this mean estimate (i.e. the variance of the actually spent time in unemployment is higher) if λ is high. For $\lambda = 3$, we get a variance of 1/9, which means that the number of weeks an individual spends in unemployment is rather concentrated around the mean value of 3. In contrast, for $\lambda = 1/2$, the variance is 4, which means that the time an individual spends in unemployment can be considerably shorter or longer than the mean estimate of 2 weeks.

(b) (3 points) Derive the log-likelihood $\ell(\lambda|Y)$ and find the Maximum Likelihood (ML) estimator

$$\hat{\lambda} \equiv \arg\max_{\lambda} \ell(\lambda|Y) \ .$$

Exercise 4.6: Solution

Solution: Since the sample is i.i.d., we can set up the likelihood function:

$$p(Y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda y_i\} = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} y_i\right\} .$$

The log-likelihood is then:

$$l(\lambda|Y) = \log p(Y|\lambda) = n\log(\lambda) - \lambda \sum_{i=1}^{n} y_i .$$

Take the FOC to find the maximum w.r.t. λ:

$$\frac{\partial l(\lambda|Y)}{\partial \lambda} = n\frac{1}{\lambda} - \sum_{i=1}^{n} y_i = 0 \quad \rightarrow \quad \hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^{-1} = \bar{Y}^{-1} .$$

(c) (3 points) What is the probability limit of the average unemployment span in your sample, $\bar{y}$? Based on that result, is $\hat{\lambda}$ consistent?

Solution: Given i.i.d.-ness, we can invoke the WLLN:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \xrightarrow{p} \mathbb{E}[y_i] = \frac{1}{\lambda} .$$

Since the function $g(x) = x^{-1}$ is continuous, by Slutsky's theorem we have that:

$$\hat{\lambda} = \bar{Y}^{-1} \xrightarrow{p} \left(\frac{1}{\lambda}\right)^{-1} = \lambda .$$

Thus, $\hat{\lambda}$ is consistent.

(d) (4 points) What is the approximate distribution of $\bar{y}$ for large n? Based on that result, what is the approximate distribution of $\hat{\lambda}$ for large n?

Solution: By i.i.d-ness, we can invoke the CLT:

$$\sqrt{n}\left(\bar{Y} - \mathbb{E}[y_i]\right) \xrightarrow{d} N(0, \mathbb{V}[y_i]) .$$

That is, for n large, $\bar{Y}$ follows the approximate distribution:

$$\bar{Y} \overset{approx.}{\sim} N\left(\mathbb{E}[y_i], \frac{1}{n}\mathbb{V}[y_i]\right) ,$$

Exercise 4.6: Solution

where $\mathbb{E}[y_i] = \frac{1}{\lambda}$ and $\mathbb{V}[y_i] = \frac{1}{\lambda^2}$.

Thus, $\hat{\lambda}$ is approximately distributed as the inverse of a Normal with mean $\frac{1}{\lambda}$ and variance $\frac{1}{n\lambda^2}$. (Note that $\hat{\lambda}$ cannot be Normally distributed, as it has to be positive, whereas the Normal distribution goes from $-\infty$ to ∞.)

(e) (3 points) Describe another, numerical approach to approximate the distribution of $\hat{\lambda}$.

Solution: The finite sample distribution of $\hat{\lambda}$ can be numerically approximated by bootstrapping. This consists in randomly drawing with replacement M samples of size n from our original sample and using them to compute M estimates $\{\hat{\lambda}_m\}_{m=1}^M$, which approximate the distribution of $\hat{\lambda}$ in repeated sampling.

More formally, given our i.i.d. sample $\{y_i\}_{i=1}^n$, for $m = 1 : M$ (and M large),

- draw n observations with replacement from the original sample, yielding a sample $\{y_i^m\}_{i=1}^n$
- compute $\hat{\lambda}_m = (\bar{Y}_m)^{-1}$ based on this sample

Then, the set $\{\hat{\lambda}_m\}_{m=1}^M$ approximates the distribution of $\hat{\lambda}$.

Now let's make the job-finding rate heterogeneous. Specifically, suppose x_i denotes the number of applications per week that individual i sent out, and let

$$\lambda_i = \exp\{\alpha + \beta x_i\} \ .$$

Note that this implies that

$$\mathbb{E}[y_i|x_i] = \frac{1}{\lambda_i} = \exp\{-(\alpha + \beta x_i)\} \quad \text{and} \quad \mathbb{V}[y_i|x_i] = \frac{1}{\lambda_i^2} = \exp\{-2(\alpha + \beta x_i)\} \ . \tag{1}$$

For simplicity, let's suppose you know α (so it's just a constant) and you only need to estimate β. The ML estimator can be defined as

$$\hat{\beta} = \arg\min_{\beta} Q_n(\beta) \ , \quad Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} -(\alpha + \beta x_i) + y_i \exp\{\alpha + \beta x_i\} \ .$$

(f) (5 points) How do you interpret the parameters α and β?

Solution: <u>Parameter α</u>: Assume that $x_i = 0$, then $\lambda_i = \exp\{\alpha\}$. That is, α is the log

Exercise 4.6: Solution

of the job-finding rate for individuals who do not send out any applications. Higher (lower) values of α mean that such individuals have it easier (harder) to find a job.

Parameter β: Tells you how much the job-finding rate changes as individuals send out more or less applications. Concretely, we have

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_i} = -\beta \exp\left\{-(\alpha + \beta x_i)\right\} = -\beta \mathbb{E}[y_i|x_i] \ ,$$

which implies

$$\frac{\Delta \mathbb{E}[y_i|x_i]}{\Delta x_i} \approx -\beta \mathbb{E}[y_i|x_i] \quad \Leftrightarrow \quad \frac{\Delta \mathbb{E}[y_i|x_i]}{\mathbb{E}[y_i|x_i]} \approx -\beta \Delta x_i$$

for small Δx_i. Therefore β tells us the approximate percentage reduction of the expected weeks spent in unemployment when the individual sends out one more application per week.

(Clearly, the exact calculation is:

$$\mathbb{E}[y_i|x_i + 1] - \mathbb{E}[y_i|x_i] = \exp\left\{-(\alpha + \beta x_i + 1)\right\} - \exp\left\{-(\alpha + \beta x_i)\right\}$$

but it is not very helpful for intuition).

(g) (4 points) How can you find $\hat{\beta}$? Derive the first-order condition associated with the above optimization problem.

Solution: Take the FOC:

$$\begin{aligned} S_n(\beta) = Q_n^{(1)}(\beta) = \frac{\partial \mathcal{Q}(\beta)}{\partial \beta} &= \frac{1}{n}\sum_{i=1}^{n} -x_i + y_i \exp\left\{\alpha + \beta x_i\right\} x_i \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i (y_i \exp\left\{\alpha + \beta x_i\right\} - 1) = 0 \ . \end{aligned}$$

We cannot solve this for β. Hence, we need to use a numerical algorithm to obtain the $\hat{\beta}$ that minimizes $Q_n(\beta)$.

(h) (6 points) Is $\hat{\beta}$ a consistent estimator for β_0, the true value for β?
Hint: Note that $\mathbb{E}[y_i|x_i] = \exp\left\{-(\alpha + \beta_0 x_i)\right\}$, and remember the Law of Iterated Expectations (LIE). Also, note that $x_i \geq 0$ and a function like $\mathbb{E}[x_i exp\left\{x_i \beta\right\}]$ is strictly increasing in β.[1]

[1] Strictly speaking, this holds provided that $\mathbb{E}[x_i] > 0$.

Exercise 4.6: Solution

Solution: As $\hat{\beta}$ is not analytically available, we can verify consistency using extremum estimation theory (its simplified version seen in class). To verify consistency, i.e. $\hat{\beta} \xrightarrow{p} \beta_0$, we need to check that the following four conditions are satisfied:

1) Let $\beta \in \mathcal{B} = [-c, c]$, for some large c. Then $\mathcal{B}$, the parameter space we consider, is compact.

2) The function $Q_n(\beta)$ converges in probability to the true $Q(\beta)$:

$$\begin{aligned} Q_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n} - \log(\lambda_i) + y_i\lambda_i \\ &\xrightarrow{p} \mathbb{E}[-log(\lambda_i) + y_i\lambda_i] \\ &= \mathbb{E}[-(\alpha + \beta x_i) + y_i \exp\{\alpha + \beta x_i\}] = Q(\beta) \end{aligned}$$

3) $Q(\beta)$ is a continuous function of β.

4) To show that $Q(\beta)$ is uniquely minimised at β_0, some work is required. By the LIE, we have that:

$$\begin{aligned} Q(\beta) &= \mathbb{E}[-(\alpha + \beta x_i) + y_i \exp\{\alpha + \beta x_i\}] \\ &= \mathbb{E}\left[\mathbb{E}[-(\alpha + \beta x_i) + y_i \exp\{\alpha + \beta x_i\} | x_i]\right] \\ &= \mathbb{E}\left[-\alpha - \beta x_i + \mathbb{E}[y_i | x_i] \exp\{\alpha + \beta x_i\}\right] \\ &= \mathbb{E}[-\alpha - \beta x_i + \exp\{(\beta - \beta_0)x_i\}] \\ &= -\alpha - \beta\mathbb{E}[x_i] + \mathbb{E}[\exp\{x_i(\beta - \beta_0)\}] \; . \end{aligned}$$

Therefore, the FOC is:

$$\frac{\partial Q(\beta)}{\partial \beta} = -\mathbb{E}[x_i] + \mathbb{E}[x_i \exp\{x_i(\beta - \beta_0)\}] = 0 \; .$$

Note that this FOC is satisfied at $\beta = \beta_0$. This solution is unique (i.e. $Q(\beta)$ attains a global minimum at β_0) because:

- For β very low (as $\beta \to -\infty$), the second term goes to zero, meaning that the overall expression is $-\mathbb{E}[x_i]$, which is smaller than zero.

- For β very high (as $\beta \to +\infty$), the second term goes to $+\infty$, meaning that the overall expression is larger than zero,

- In-between, the function is strictly increasing in β.

Exercise 4.6: Solution

Hence, there can only be a single point at which $\dfrac{\partial Q(\beta)}{\partial \beta} = 0$, which is $\beta = \beta_0$.

(i) (6 points) Show that the asymptotic distribution of $\hat{\beta}$ is given by

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \mathbb{E}[x_i^2]^{-1}\right) \ .$$

Hint: To simplify notation, write λ_{i0} for exp$\{\alpha + \beta_0 x_i\}$. Also, your formula for the asymptotic variance simplifies thanks to the LIE and the results in Eq. (1)*.*

Solution: Again, as $\hat{\beta}$ is not analytically available, we can use extremum estimation theory to find the asymptotic distribution of $\hat{\beta}$. For this, we need to verify three more conditions:

1) We assume $\beta_0 \in int(\mathscr{B})$ (which is given, provided that β_0 is some finite value).

2) Using the result from the exercise (g),

$$\begin{aligned}
\sqrt{n}Q_n^{(1)}(\beta) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^n x_i(y_i \exp \alpha + \beta x_i - 1) \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^n x_i(y_i\lambda_0 - 1) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n x_i(y_i\lambda_0 - 1) - \mathbb{E}[x_i(y_i\lambda_{i,0} - 1)]\right) \ .
\end{aligned}$$

The second term in brackets can be added because $\mathbb{E}[x_i(y_i\lambda_{i,0} - 1)] = 0$ by LIE, owing to the fact that $\mathbb{E}[y_i|x_i] = \lambda_{i,0}^{-1}$.

By CLT, then:

$$\sqrt{n}Q_n^{(1)} \xrightarrow{d} N(0, M) \ , \quad M = \mathbb{V}[x_i(y_i\lambda_{i,0} - 1)] \ .$$

Thereby, we can simplify

$$\begin{aligned}
M = \mathbb{V}[x_i(y_i\lambda_{i,0} - 1)] &= \mathbb{E}[x_i^2(y_i\lambda_{i,0} - 1)^2] \\
&= \mathbb{E}[x_i^2(y_i^2\lambda_{i,0}^2 + 1 - 2y_i\lambda_{i,0})] \\
&= \mathbb{E}[x_i^2] + \mathbb{E}[x_i^2 y_i^2 \lambda_{i,0}^2] - 2\mathbb{E}[x_i^2 y_i \lambda_{i,0}] \\
&= \mathbb{E}[x_i^2] + 2\mathbb{E}[x_i^2] - 2\mathbb{E}[x_i^2] \\
&= \mathbb{E}[x_i^2] \ ,
\end{aligned}$$

Exercise 4.6: Solution

whereby we apply the LIE to the terms $\mathbb{E}[x_i^2 y_i^2 \lambda_{i,0}^2]$ and $2\mathbb{E}[x_i^2 y_i \lambda_{i,0}]$ and use the facts that

$$\begin{aligned}
&\mathbb{E}[y_i|x_i] = \lambda_{i,0}^{-1} \ , \\
&\mathbb{E}[y_i|x_i] = \mathbb{V}[y_i|x_i] + \mathbb{E}[y_i|x_i]^2 = \lambda_{i,0}^{-2} + (\lambda_{i,0}^2)^{-1} = 2\lambda_{i,0}^{-2} \ .
\end{aligned}$$

3)

$$\begin{aligned}
Q_n^{(2)}(\beta) = \frac{1}{n}\sum_{i=1}^n x_i^2 y_i \exp\{\alpha + \beta x_i\} &= \frac{1}{n}\sum_{i=1}^n x_i^2 y_i \lambda_i \\
&\xrightarrow{p} \mathbb{E}[x_i^2 y_i \lambda_{i,0}] = \mathbb{E}[x_i^2] \equiv H
\end{aligned}$$

Combining the three together, we obtain that:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, H^{-1}MH^{-1}) \ ,$$

where $H^{-1}MH^{-1} = \mathbb{E}[x_i^2]^{-1}\mathbb{E}[x_i^2]\mathbb{E}[x_i^2]^{-1} = \mathbb{E}[x_i^2]$.

(j) (4 points) Construct a hypothesis test with size $\alpha = 0.05$ for testing

$$\mathscr{H}_0 : \mathbb{E}[y_i|x_i = 10] = \frac{1}{2}\mathbb{E}[y_i|x_i = 5] \quad \text{vs.} \quad \mathscr{H}_0 : \mathbb{E}[y_i|x_i = 10] \neq \frac{1}{2}\mathbb{E}[y_i|x_i = 5] \ ,$$

i.e. testing whether an individual submitting 10 applications per week spends (in expectation) exactly half as long in unemployment than a person sending out only 5 applications per week. More concretely, definining the test as $\varphi = \mathbf{1}\{T(X) < c_\alpha\}$, define the test-statistic $T(X)$ and find the critical value c_α.

Solution: One can apply several tests here: t-test, Likelihood Ratio (LR) test, Wald test. The following solution is based on the t-test, which requires rewriting our hy-

Exercise 4.6: Solution

pothesis:

$$\begin{aligned}
\mathbb{E}[y_i|x_i = 10] &= \frac{1}{2}\mathbb{E}[y_i|x_i = 5] \\
\Leftrightarrow \quad \exp\{-(\alpha + 10\beta)\} &= \frac{1}{2}\exp -(\alpha + 5\beta) \\
\Leftrightarrow \quad -\alpha - 10\beta &= \log\left(\frac{1}{2}\right) - \alpha - 5\beta \\
\Leftrightarrow \quad -5\beta &= -\log(2) \\
\Leftrightarrow \quad \beta &= \frac{\log(2)}{5}
\end{aligned}$$

Under $\mathscr{H}_0 : \beta = \beta_0 = \log(2)/5$, we have that

$$\hat{\beta} \overset{approx.}{\sim} N\left(\beta_0, \frac{1}{n}\hat{\mathbb{E}[x_i^2]}^{-1}\right) ,$$

where $\hat{\mathbb{E}[x_i^2]} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$. Therefore,

$$\frac{\hat{\beta} - \beta_0}{(n\hat{\mathbb{E}[x_i^2]})^{-1/2}} \overset{approx.}{\sim} N(0, 1) ,$$

and we can set up a t-test $\varphi_i = \{T(X) < c_\alpha\}$, where:

$$T(X) = \left|\frac{\hat{\beta} - \beta_0}{\widehat{(n\mathbb{E}[x_i^2])}^{-1/2}}\right| = \left|\frac{\hat{\beta} - (\log(2)/5)}{\widehat{(n\mathbb{E}[x_i^2])}^{-1/2}}\right| ,$$

and where c_α is the 97.5-th quantile of the standard Normal distribution.

As mentioned, one can also apply the Wald test or the LR test. Thereby, the LR test would compare the likelihood of the unrestricted model (where β is estimated) to the likelihood of the restricted model (where $\beta = \log(2)/5$ is imposed). In contrast, the Wald test does not (necessarily) require simplifying $\mathscr{H}_0$ to arrive at $\beta = \log(2)/5$. Instead, it suffices to write $\mathscr{H}_0$ as

$$\mathscr{H}_0 : g(\beta) = 0 , \quad g(\beta) = \exp\{-(\alpha + 10\beta)\} - \frac{1}{2}\exp -(\alpha + 5\beta) .$$

(This is equivalent to $g(\beta) = \beta - \log(2)/5$.)

5 Bayesian Estimation: Linear Regressions & Beyond

5.1 Bayesian Inference in Linear Regression Model: Basics

Remark: excercise continued in 5.2.

Suppose you have a sample of n observations, $Z \equiv \{y_i, x_i\}_{i=1}^n$, and you assume thse observations have been generated by the linear regression

$$y_i = x_i'\beta + u_i \quad , \quad u_i|x_i \sim N(0, \sigma^2) \; .$$

In all what follows, we condition on σ^2, i.e. we assume you know σ^2, and same for X. As a result, for example, we write the likelihood as $p(Y|\beta)$ instead of $p(Y|X, \beta, \sigma^2)$, and we denote all (random) data by Y instead of Z.

(Also, note that exercises (a), (b) and (f) could be copied one-to-one from the script. However, try first very hard to solve them yourself before consulting the script.)

(a) Derive the (conditional) likelihood $p(Y|\beta)$. Note that it tells you the distribution of your data Y given the parameter β under your supposed model.

(b) Suppose you have the following prior belief where the true value of β lies (i.e. an ex-ante belief, before having seen the data Z),

$$\beta \sim N\left(0, \frac{\sigma^2}{\tau} I\right) \; ,$$

and you would like to update it to the posterior belief (i.e. an ex-post belief, after having seen the data Z (through the lens of the model you assumed has generated the data)). Derive this posterior $p(\beta|Y)$, i.e. use Bayes' formula to update your prior belief $p(\beta)$ to your posterior belief $p(\beta|Y)$.

(c) Now let's get some things straight regarding the difference between the Bayesian and frequentist/classical paradigm. We are in the the Bayesian paradigm whenever we treat the parameter β as a RV (and condition on data Y). This allows us to compute objects like

$$\mathbb{E}[\beta]\ , \quad \bar{\beta} \equiv \mathbb{E}[\beta|Y]\ , \quad \text{or} \quad \mathbb{V}[\beta|Y]\ \ .$$

We are in the frequentist/classical paradigm when we treat β as a fixed (but unknown) parameter (and the data as random). This allows us to compute objects like

$$\mathbb{E}[y_i|x_i,\beta]\ , \quad \mathbb{E}[\hat{\beta}|X,\beta]\ , \quad \mathbb{V}[\hat{\beta}|X,\beta]\ , \quad \mathbb{E}[\bar{\beta}|X,\beta]\ , \quad \text{or} \quad \mathbb{V}[\bar{\beta}|X,\beta]\ ,$$

where $\hat{\beta} = (X'X)^{-1}X'Y$. Compute these objects to make sure you understood. In particular, compare the mean and variance of $\hat{\beta}$ (the OLS/ML estimator) with those of $\bar{\beta}$ (the posterior mean a.k.a. the Bayes estimator). Are they biased? Which one has a higher variance? What happens with $\bar{\beta}$ and its mean and variance if $\tau \to \infty$ and if $\tau \to 0$?

5.2 Bayesian Inference in Linear Regression Model: Hyperparameter-Selection & Prediction

Remark: excercise continued in 5.1.

(a) Simulate a dataset of $n = 100$ observations, proceeding as follows:

  (a) let $x_{1i} = 1\ \forall\ i$ and draw $x_{2i}, x_{3i}, x_{4i}, x_{5i} \sim N(0,1)$ for $i = 1 : n$,

  (b) draw $u_i \sim N(0,\sigma^2)$ for $i = 1 : n$, with $\sigma^2 = 10$,

  (c) let $\beta_0 = [3, 8, -4, 0, 0]'$ (the true value of β),

  (d) for $i = 1 : n$, compute $y_i = x_i'\beta_0 + u_i$, with $x_i = [x_{1i}, ..., x_{5i}]'$.

  Split this dataset into two: the first 80 observations are your actual dataset, which you will use in the subsequent estimations, while the remaining 20 observations are left out so that we can analyze the properties of your estimator out-of-sample. Store each of these two separately.

(b) Note that your prior and posterior are in fact conditional on the hyperparameter τ: $p(\beta|\tau)$ and $p(\beta|Y,\tau)$. The same holds for the posterior mean, $\bar{\beta}(\tau)$. Use your dataset (i.e. the 80 observations) to compute $\bar{\beta}(\tau)$ for different values of τ, using $\sigma^2 = 10$. In particular, create a single figure that shows each $\bar{\beta}_j(\tau)$ for $j = 1 : 5$ as a function of τ, putting τ on the x-axis and increasing it from $\tau = 0$ to a value large enough that all $\bar{\beta}_j$ are roughly zero.

(c) Now you would like to find an optimal value for τ, τ^*, that allows you to best predict y_i for the 20 observations i which are out-of-sample. For this purpose, derive the marginal data density (MDD) $p(Y|\tau)$ and its natural logarithm. Then, use an optimization command of your choice to find

$$\tau_1^* = \arg\max_{\tau}\ \log\ p(Y|\tau)\ .$$

(d) Compute two other τ^*'s using approximations of the log MDD: the Bayesian/Schwarz Information Criterion (BIC/SIC),

$$\tau_2^* = \arg\max_{\tau} \ \log \ p(Y|\bar{\beta}(\tau)) - \frac{k}{2}\log \ n \ ,$$

and the Akaike Information Criterion (AIC),

$$\tau_3^* = \arg\max_{\tau} \ 2\log \ p(Y|\bar{\beta}(\tau)) - 2k \ .$$

Thereby, $k = 5$ is your number of regressors, $n = 80$ is your sample size, and $p(Y|\bar{\beta}(\tau))$ is the likelihood, $p(Y|\beta)$, evaluated at $\beta = \bar{\beta}(\tau)$.

(e) For each value of $\tau^* \in \{\tau_1^*, \tau_2^*, \tau_3^*\}$, compute $\bar{\beta}(\tau^*)$, predict the outcome variable y_i of the 20 out-of-sample observations i and calculate the out-of-sample mean squared error (MSE),

$$\frac{1}{20}\sum_{i=1}^{20}(y_i - \hat{y}_i^*)^2 \ , \quad \text{where } \hat{y}_i^* = x_i'\bar{\beta}(\tau^*) \ .$$

Compare the three MSE_1, MSE_2 and MSE_3. Which value of τ^* results in the lowest MSE?

5.3 Bayesian Inference in Linear Regression Model: Numerical Posterior (MNIW) Computations

Bayesian inference on both β as well as σ^2 in a linear regression model is done standardly by assuming the following prior distribution:

$$p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2) \ ,$$

with

$$\beta|\sigma^2 \sim N(\underline{\beta}, \sigma^2\underline{V}) \ , \quad \sigma^2 \sim IG(\underline{\nu}, \underline{s}^2) \ ,$$

i.e. we break up the joint prior of (β, σ^2) into a conditional prior for $\beta|\sigma^2$ and a marginal prior for σ^2. Thereby, $\underline{\beta}$, $\underline{V}$, $\underline{\nu}$ and $\underline{s}^2$ are hyperparameters that define the exact shape of the prior distribution and that have to specified by the researcher. As shown in the lecture notes, this results in the following posterior, which is also broken up into a conditional and a marginal:

$$p(\beta, \sigma^2|Y) = p(\beta|Y, \sigma^2)p(\sigma^2|Y) \ ,$$

with

$$\beta|Y, \sigma^2 \sim N(\bar{\beta}, \sigma^2\bar{V}) \ , \quad \sigma^2|Y \sim IG(\bar{\nu}, \bar{s}^2) \ .$$

Thereby,

$$\bar{V} = [\underline{V}^{-1} + X'X]^{-1} \;, \qquad \bar{\beta} = \bar{V}[X'Y + \underline{V}^{-1}\underline{\beta}] \;,$$

$$\bar{\nu} = \underline{\nu} + n \;, \qquad \bar{s}^2 = \underline{s}^2 + Y'Y + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - \bar{\beta}'\bar{V}^{-1}\bar{\beta} \;.$$

are parameters that define the exact shape of the prior distribution and that are obtained using the hyperparameters $\underline{\beta}$, $\underline{V}$, $\underline{\nu}$ and $\underline{s}^2$ and the data (Y, X).

(a) Let's take a Ridge-prior for $\beta|\sigma^2$, i.e. take $\bar{\beta} = [0,0,0,0,0]'$ and $\underline{V} = \frac{1}{\tau^*}I$ for a τ^* of your choice in $[0, 10]$. Also, let's take $\underline{\nu} = 4$ and $\underline{s}^2 = 10$. Draw $M = 1000$ values from the prior $p(\beta, \sigma^2) = p(\beta_1, \beta_2, ..., \beta_5, \sigma^2)$, denoted by $\{\beta_{(m)}, \sigma^2_{(m)}\}_{m=1}^M$. To obtain a draw $\beta_{(m)}, \sigma^2_{(m)}$, draw first

$$\sigma^2_{(m)} \sim IG(\underline{\nu}, \underline{s}^2) \;,$$

and then, conditional on this $\sigma^2_{(m)}$, draw

$$\beta_{(m)}|\sigma^2_{(m)} \sim N(\underline{\beta}, \sigma^2_{(m)}\underline{V}) \;.$$

Plot the marginal prior of each of your six parameters $\{\beta_1, ..., \beta_5, \sigma^2\}$ by plotting a histogram of your draws: e.g. for β_1 we have

$$p(\beta_1) \approx \text{histogram of } \; \{\beta_{1,(m)}\}_{m=1}^M \;.$$

Also, compute the prior means and variances of each parameter by approximating them with the means and variances across your draws: e.g., for β_1 we have

$$\mathbb{E}[\beta_1] \approx \frac{1}{M}\sum_{m=1}^M \beta_{1,(m)} \equiv \mu_1 \;, \quad \mathbb{V}[\beta_1] \approx \frac{1}{M}\sum_{m=1}^M (\beta_{1,(m)} - \mu_1)^2 \;.$$

Finally, illustrate the joint prior of β_2 and σ^2 by showing a scatterplot of their draws.

Note that I define the Inverse Gamma distribution $IG(\nu, s^2)$ to have pdf

$$f(x) \propto x^{-(\nu+2)/2} exp\left\{-\frac{s^2}{2x}\right\} \;,$$

whereas some other people and software commands define the Inverse Gamma distribution as $IG(\alpha, \beta)$ with pdf proportional to

$$f(x) \propto x^{-(\alpha+1)} exp\left\{-\beta/x\right\} \;.$$

Check which definition applies in your software of choice. If it's the latter, you simply take ν and s^2 from above and you transform them into $\beta = s^2/2$ and $\alpha = \nu/2$.

(b) Show the prior distribution of β_2/β_3 by plotting a histogram of your draws $\{\beta_{2,(m)}/\beta_{3,(m)}\}_{m=1}^{M}$. Also, use your draws to compute the prior mean and variance of β_2/β_3:

$$\mathbb{E}[\beta_2/\beta_3] \approx \frac{1}{M}\sum_{m=1}^{M} \beta_{2,(m)}/\beta_{3,(m)} \equiv \mu \ , \quad \mathbb{V}[\beta_2/\beta_3] \approx \frac{1}{M}\sum_{m=1}^{M} (\beta_{2,(m)}/\beta_{3,(m)} - \mu)^2 \ .$$

(c) Simulate a dataset of $n = 100$ observations, proceeding as follows:

(a) let $x_{1i} = 1 \ \forall \ i$ and draw $x_{2i}, x_{3i}, x_{4i}, x_{5i} \sim N(0,1)$ for $i = 1 : n$,

(b) draw $u_i \sim N(0, \sigma^2)$ for $i = 1 : n$, with $\sigma^2 = 10$,

(c) let $\beta_0 = [3, 8, -4, 0, 0]'$ (the true value of β),

(d) for $i = 1 : n$, compute $y_i = x_i'\beta_0 + u_i$, with $x_i = [x_{1i}, ..., x_{5i}]'$.

(d) Use your dataset of $n = 100$ observations to compute the posterior of (β, σ^2). Specifically, compute $\bar{\beta}$, $\bar{V}$, $\bar{\nu}$ and $\bar{s}^2$ using your dataset, and repeat exercises (a) and (b): draw from the posterior $p(\beta, \sigma^2|Y) = p(\beta_1, \beta_2, ..., \beta_5, \sigma^2|Y)$,and use your draws to plot marginal posteriors, to compute means and variances of each parameter, to plot the joint posterior of β_2 and σ^2, and to show the posterior distribution of β_2/β_3.

5.4 Bayesian Inference in Linear Regression Model: Lasso/Ridge Estimators as Posterior Modes

The linear regression model

$$y_i = x_i'\beta + u_i \quad , \quad u_i|x_i \sim N(0, \sigma^2) \tag{5.1}$$

and the prior $\beta| \ \sigma^2 \sim N\left(0, \sigma^2 \frac{1}{\lambda} I\right)$ lead to the posterior $\beta|Y, \sigma^2 \sim N(\bar{\beta}, \sigma^2 \bar{V})$ with

$$\bar{V} = [X'X + \lambda I]^{-1} \ , \quad \bar{\beta} = [X'X + \lambda I]^{-1} X'Y \ .$$

(a) The lecture notes claim that

$$\bar{\beta} = \arg\min_{\beta} \ (Y - X\beta)'(Y - X\beta) + \lambda \sum_{j=1}^{k} \beta_j^2 \ , \tag{5.2}$$

i.e. the posterior mean under the above prior equals the Ridge estimator, which minimizes the sum of squared residuals plus the penalty-term $\lambda \sum_{j=1}^{k} \beta_j^2$. Let's show that. Specifically, derive the likelihood of the linear regression in Eq. (5.1) and multiply it by the prior to get the posterior up to proportionality:

$$p(\beta|Y) = \frac{p(Y|\beta)p(\beta)}{p(Y)} \propto p(Y|\beta)p(\beta) \; .$$

Because this posterior is a Normal distribution, you know that $\bar{\beta}$ is not only the posterior mean, but also the posterior mode, i.e.

$$\bar{\beta} = \arg\max_{\beta} \; p(\beta|Y) = \arg\max_{\beta} \; p(Y|\beta)p(\beta) \; .$$

Use this result to show the claim from Eq. (5.2).
Hint: note that the pdf of a multivariate normal distribution $X \sim N(\mu, \Sigma)$ is

$$f(x) = (2\pi)^{-k/2}|\Sigma|^{-1/2}exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \; ,$$

where k is the dimension (length) of the vector X. Also, note that $\sum_{j=1}^{k}\beta_j^2 = \beta'\beta$.

(b) Now use analogous steps as above to show that the posterior mode under the prior

$$p(\beta) = \prod_{j=1}^{k}\frac{1}{2}\lambda \; exp\left\{-\lambda|\beta_j|\right\}$$

solves

$$\min_{\beta} \; (Y - X\beta)'(Y - X\beta) + \tilde{\lambda}\sum_{j=1}^{k}|\beta_j| \; ,$$

for $\tilde{\lambda} = 2\lambda\sigma^2$ (a one-to-one transformation of λ). We call this posterior mode the Lasso estimator. Note that it cannot be obtained analytically.

(c) One can use the two Ridge- and Laplace-priors above not only to obtain the Ridge/Lasso estimator of β in the linear regression model, but also in many other models, e.g. the Probit model. Again, use analogous steps as above to show that the posterior mode under the Probit model and the Ridge-prior solves

$$\max_{\beta} \; \log \; p(Y|X,\beta) - \tilde{\lambda}\sum_{j=1}^{k}\beta_j^2 \; ,$$

where $\tilde{\lambda} = \lambda/2$ is a simple one-to-one transformation of λ, and $\log \; p(Y|X,\beta)$ is the log-likelihood under the Probit model. Note that you do not have to derive this likelihood! Also, show that the posterior mode under the Probit model and the Laplace-prior solves

$$\max_{\beta} \; \log \; p(Y|X,\beta) - \lambda \sum_{j=1}^{k} |\beta_j| \; .$$

5.5 Inference in Linear Regression Model: OLS vs. ML vs. Bayes

Consider the linear regression model

$$y_i = x_i'\beta + u_i \; ,$$

with $\mathbb{E}[x_i u_i] = 0$ and $\mathbb{V}[u_i|x_i] = \sigma^2$. Suppose you observe n i.i.d. observations: $\{(y_i, x_i)\}_{i=1:n}$.

(a) Derive the OLS estimator $\hat{\beta}_{OLS}$. What do you know about the finite sample distribution of $\hat{\beta}_{OLS}|X$? Is $\hat{\beta}_{OLS}$ unbiased? Is it consistent?

(b) Assuming $u_i|x_i \sim N(0, \sigma^2)$, derive the ML estimator $\hat{\beta}_{MLE}$. What are the reasons why one might want to make this assumption on the distribution of $u_i|x_i$, in general (for the linear regression model and beyond)?

(c) Assuming in addition $\beta|\sigma^2 \sim N(0, \tau\sigma^2 I)$, derive the (conditional) posterior distribution $p(\beta|Y, \sigma^2)$. Define the Bayes estimator $\hat{\beta}_B$ to be the posterior mean. What is the (finite sample) distribution of $\hat{\beta}_B|X$ under the Bayesian paradigm? What is its distribution under the frequentist paradigm? Is $\hat{\beta}_B$ unbiased? Is it consistent? How could you choose τ and what does this choice signify?

(d) How would you estimate σ^2 under OLS vs. MLE vs. the Bayesian paradigm? You do not need to derive the estimators.

Exercise 5.1: Solution

Suppose you have a sample of n observations, $Z \equiv \{y_i, x_i\}_{i=1}^n$, and you assume thse observations have been generated by the linear regression

$$y_i = x_i'\beta + u_i \ , \quad u_i|x_i \sim N(0, \sigma^2) \ . \tag{1}$$

In all what follows, we condition on σ^2, i.e. we assume you know σ^2, and same for X. As a result, for example, we write the likelihood as $p(Y|\beta)$ instead of $p(Y|X, \beta, \sigma^2)$, and we denote all (random) data by Y instead of Z.

(Also, note that exercises (a), (b) and (f) could be copied one-to-one from the script. However, try first very hard to solve them yourself before consulting the script.)

1. Derive the (conditional) likelihood $p(Y|\beta)$. Note that it tells you the distribution of your data Y given the parameter β under your supposed model.

Solution

$$y_i = x_i'\beta + u_i, \quad u_i \sim N(0, \sigma^2) \quad \rightarrow y_i - x_i'\beta = u_i|x_i \sim N(0, \sigma^2)$$

Treat σ^2 as a known quantity. Drop the conditioning on σ^2 and x_i to ease the notation. The likelihood function is:

$$\begin{aligned} p(Y|\beta) =& \prod_{i=1}^n f(y_i|\beta) = \\ =& \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \right\} = \\ =& (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i'\beta)^2 \right\} = \\ =& (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right\} \end{aligned}$$

2. Suppose you have the following prior belief where the true value of β lies (i.e. an ex-ante belief, before having seen the data Z),

$$\beta \sim N\left(0, \sigma^2 \frac{1}{\tau} I\right) \ ,$$

and you would like to update it to the posterior belief (i.e. an ex-post belief, after having seen the data Z (through the lens of the model you assumed has generated the data)). Derive this posterior $p(\beta|Y)$, i.e. use Bayes' formula to update your prior belief $p(\beta)$ to your posterior belief $p(\beta|Y)$.

Solution

Exercise 5.1: Solution

$$\begin{aligned} p(\beta|Y) &\propto p(Y|\beta)p(\beta) \\ &\propto \exp\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\}\exp\{-\frac{1}{2\sigma^2}\beta'\tau I\beta\} \\ &\propto \exp\{-\frac{1}{2\sigma^2}[-\beta'X'Y-2Y'X\beta+\beta'X'X\beta+\beta'\tau I\beta]\} \\ &\propto \exp\{-\frac{1}{2\sigma^2}[\beta'[X'X+\tau I]\beta-2Y'X\beta]\} \end{aligned}$$

Thus:

$$\beta|Y \sim N(\bar{\beta}, \sigma^2\bar{V})$$

Where:

$$\bar{\beta} = [\tau I + X'X]^{-1}X'Y \qquad \bar{V} = [\tau I + X'X]^{-1}$$

3. Now let's get some things straight regarding the difference between the Bayesian and frequentist/classical paradigm. We are in the the Bayesian paradigm whenever we treat the parameter β as a RV (and condition on data Y). This allows us to compute objects like

$$\mathbb{E}[\beta]\ , \quad \bar{\beta} \equiv \mathbb{E}[\beta|Y]\ , \quad \text{or} \quad \mathbb{V}[\beta|Y]\ .$$

We are in the frequentist/classical paradigm when we treat β as a fixed (but unknown) parameter (and the data as random). This allows us to compute objects like

$$\mathbb{E}[y_i|x_i,\beta]\ , \quad \mathbb{E}[\hat{\beta}|X,\beta]\ , \quad \mathbb{V}[\hat{\beta}|X,\beta]\ , \quad \mathbb{E}[\bar{\beta}|X,\beta]\ , \quad \text{or} \quad \mathbb{V}[\bar{\beta}|X,\beta]\ ,$$

where $\hat{\beta} = (X'X)^{-1}X'Y$. Compute these objects to make sure you understood. In particular, compare the mean and variance of $\hat{\beta}$ (the OLS/ML estimator) with those of $\bar{\beta}$ (the posterior mean a.k.a. the Bayes estimator). Are they biased? Which one has a higher variance? What happens with $\bar{\beta}$ and its mean and variance if $\tau \to \infty$ and if $\tau \to 0$?

Solution

Bayesian paradigm.

Object 1 (this is simply the prior mean):

$$\mathbb{E}[\beta] = \underline{\beta} = 0$$

Object 2:

$$\bar{\beta} \equiv \mathbb{E}[\beta|Y] = [\underline{V}^{-1} + X'X]^{-1}[X'Y + \underline{V}^{-1}\underline{\beta}] = [\tau I + X'X]^{-1}X'Y$$

Object 3:

$$\mathbb{V}[\beta|Y] = \sigma^2\bar{V} = \sigma^2[\tau I + X'X]^{-1}$$

Classical/Frequentist paradigm.

Exercise 5.1: Solution

Object 1:

$$\mathbb{E}[y_i|x_i,\beta] = x_i'\beta$$

Object 2:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|X,\beta] &= \mathbb{E}[(X'X)^{-1}X'Y|X,\beta] = \\
&= \mathbb{E}[(X'X)^{-1}X'(X\beta+U)|X,\beta] = \\
&= \mathbb{E}[(X'X)^{-1}X'X\beta|X,\beta] + \mathbb{E}[(X'X)^{-1}X'U|X,\beta] = \\
&= \beta + 0 = \beta.
\end{aligned}$$

Object 3:

$$\begin{aligned}
\mathbb{V}[\hat{\beta}|X,\beta] &= \mathbb{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X,\beta] = \\
&= \mathbb{E}[((X'X)^{-1}X'U)((X'X)^{-1}X'U)'|X,\beta] = \\
&= \mathbb{E}[(X'X)^{-1}X'UU'X(X'X)^{-1}|X,\beta] = \\
&= (X'X)^{-1}X'\mathbb{E}[UU'|X,\beta]X(X'X)^{-1} = \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

Object 4:

$$\begin{aligned}
\mathbb{E}[\bar{\beta}|X,\beta] &= \mathbb{E}[[\tau I + X'X]^{-1}X'Y|X,\beta] = \\
&= \mathbb{E}[[\tau I + X'X]^{-1}X'(X\beta+U)|X,\beta] = \\
&= \mathbb{E}[[\tau I + X'X]^{-1}X'X\beta|X,\beta] + \mathbb{E}[[\tau I + X'X]^{-1}X'U|X,\beta] = \\
&= [\tau I + X'X]^{-1}[\tau I + X'X - \tau I]\beta + [\tau I + X'X]^{-1}\mathbb{E}[X'U|X,\beta] = \\
&= \beta - (\tau I + X'X)^{-1}\tau\beta + 0
\end{aligned}$$

From the computation of this last object it is clear that the Bayes estimator is biased towards the prior mean (in this case zero) for any finite prior variance.

The variance of the Bayes estimator is:

$$\begin{aligned}
\mathbb{V}[\bar{\beta}|X,\beta] &= \mathbb{E}[(\bar{\beta}-\beta)(\bar{\beta}-\beta)'|X,\beta] = \\
&= \mathbb{E}[[(\tau I + X'X)^{-1}X'U][(\tau I + X'X)^{-1}X'U]'|X,\beta] = \\
&= \mathbb{E}[(\tau I + X'X)^{-1}X'UU'X(\tau I + X'X)^{-1}|X,\beta] = \\
&= \sigma^2(\tau I + X'X)^{-1}X'X(\tau I + X'X)^{-1} < \sigma^2(X'X)^{-1} = \mathbb{V}[\hat{\beta}_{ML}|X,\beta]
\end{aligned}$$

For any finite prior variance, the variance of the Bayes estimator is strictly lower than the variance of the

Exercise 5.1: Solution

OLS/ML estimator, as:

$$|(\tau I + X'X)^{-1}X'X(\tau I + X'X)^{-1}| < |(X'X)^{-1}| \qquad \forall \;\; \tau > 0$$

(And we know that because of (1) and (2) and in-between the variance is strictly decreasing in τ).

$$\mathbb{V}[\bar{\beta}|\beta] \to \sigma^2(X'X)^{-1} = \mathbb{V}[\hat{\beta}_{ML}] \qquad \text{as} \;\; \tau \to 0 \tag{1}$$

$$\mathbb{V}[\bar{\beta}|\beta] \to 0 \qquad \text{as} \;\; \tau \to +\infty \tag{2}$$

Exercise 5.2: Solution

1. Simulate a dataset of $n = 100$ observations, proceeding as follows:
   (a) let $x_{1i} = 1 \, \forall \, i$ and draw $x_{2i}, x_{3i}, x_{4i}, x_{5i} \sim N(0, 1)$ for $i = 1 : n$,
   (b) draw $u_i \sim N(0, \sigma^2)$ for $i = 1 : n$, with $\sigma^2 = 10$,
   (c) let $\beta_0 = [3, 8, -4, 0, 0]'$ (the true value of β),
   (d) for $i = 1 : n$, compute $y_i = x_i' \beta_0 + u_i$, with $x_i = [x_{1i}, ..., x_{5i}]'$.

   Split this dataset into two: the first 80 observations are your actual dataset, which you will use in the subsequent estimations, while the remaining 20 observations are left out so that we can analyze the properties of your estimator out-of-sample. Store each of these two separately.

Solution

```
rm(list = ls())
set.seed(2024)

#sample size
n <- 100
#explanatory variables
x2 <- rnorm(n)
x3 <- rnorm(n)
x4 <- rnorm(n)
x5 <- rnorm(n)
X <- cbind(1, x2, x3, x4, x5)
#disturbance term
U <- rnorm(n, 0, sqrt(10))
#true parameters
beta0 <- c(3, 8, -4, 0, 0)
#generate y
Y <- X %*% beta0 + U

#training sample: first 80 obs
X1 <- X[1:80, ]
Y1 <- Y[1:80]

#testing sample: fast 20 obs
X2 <- X[(n-19):n, ]
Y2 <- Y[(n-19):n]
```

2. Note that your prior and posterior are in fact conditional on the hyperparameter τ: $p(\beta|\tau)$ and $p(\beta|Y, \tau)$. The same holds for the posterior mean, $\bar{\beta}(\tau)$. Use your dataset (i.e. the 80 observations) to compute $\bar{\beta}(\tau)$ for different values of τ, using $\sigma^2 = 10$. In particular, create a single figure that shows each $\bar{\beta}_j(\tau)$ for $j = 1 : 5$ as a function of τ, putting τ on the x-axis and increasing it from $\tau = 0$ to a value large enough that all $\bar{\beta}_j$ are roughly zero.

Solution

```
#generate functions for the beta_J's
fGetPostMean <- function(tau, X, Y) {
  #compute (X'X + tau*I)
```

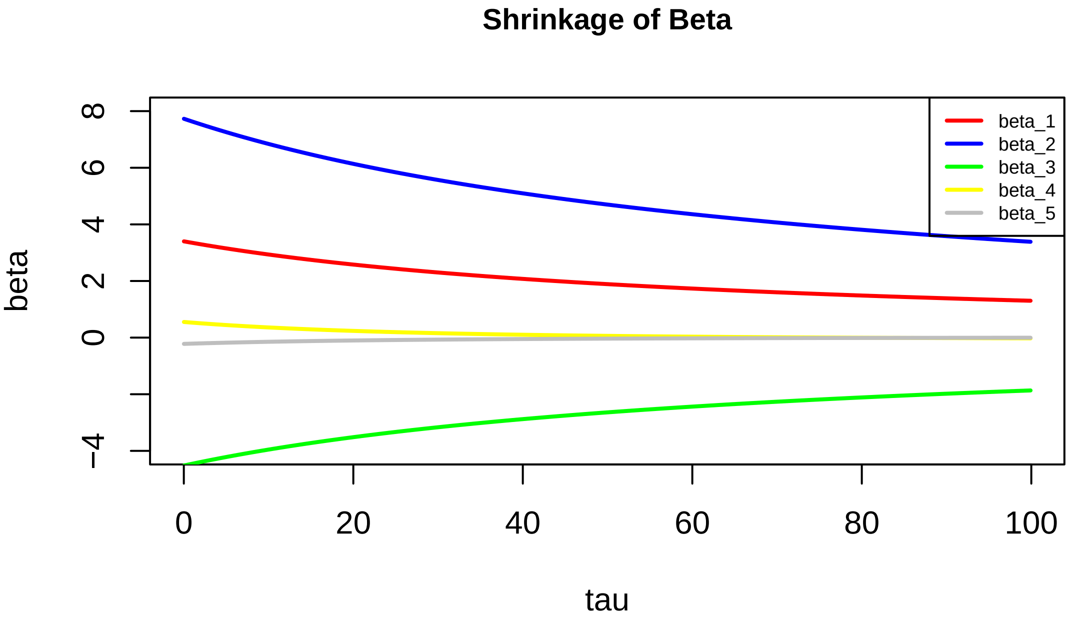
Exercise 5.2: Solution

```
  XtX <- t(X) %*% X
  tauI <- tau * diag(ncol(X))
  XtX_tauI <- XtX + tauI
  #invert (X'X + tau*I)
  inv_XtX_tauI <- solve(XtX_tauI)
  #compute (X'Y)
  XtY <- t(X) %*% Y
  #compute beta = (X'X + tau*I)^(-1) %*% (X'Y)
  beta <- inv_XtX_tauI %*% XtY
  #return beta_1 (first element of the beta vector)
  return(beta)
}

#define grid over which plot betas
tau_grid <- seq(0.01, 100, by = 0.1)
beta_1_values <- numeric(length(tau_grid))
beta_2_values <- numeric(length(tau_grid))
beta_3_values <- numeric(length(tau_grid))
beta_4_values <- numeric(length(tau_grid))
beta_5_values <- numeric(length(tau_grid))

#evaluate beta_j over the grid
for (i in 1:length(tau_grid)) {
  beta_bar <- fGetPostMean(tau_grid[i], X1, Y1)
  beta_1_values[i] <- beta_bar[1]
  beta_2_values[i] <- beta_bar[2]
  beta_3_values[i] <- beta_bar[3]
  beta_4_values[i] <- beta_bar[4]
  beta_5_values[i] <- beta_bar[5]
}

#plot beta_1
plot(tau_grid,
     beta_1_values,
     type = "l",
     col = 'red',
     lwd = 2,
     xlab = "tau",
     ylab = "beta",
     ylim = c(-4, 8),
     main = "Shrinkage of Beta",
     cex.main = 0.9)
#add line for beta_2
lines(tau_grid,
```

Exercise 5.2: Solution

```
      beta_2_values,
      col = "blue",
      lwd = 2)
#add line for beta_3
lines(tau_grid,
      beta_3_values,
      col = "green",
      lwd = 2)
#add line for beta_4
lines(tau_grid,
      beta_4_values,
      col = "yellow",
      lwd = 2)
#add line for beta_5
lines(tau_grid,
      beta_5_values,
      col = "grey",
      lwd = 2)
#add legend
legend("topright",
       legend = c("beta_1",
                  "beta_2",
                  "beta_3",
                  "beta_4",
                  "beta_5"),
       col = c("red",
               "blue",
               "green",
               "yellow",
               "grey"),
       lty = 1,
       lwd = 2,
       cex = 0.6)
```

Exercise 5.2: Solution



3. Now you would like to find an optimal value for τ, τ^*, that allows you to best predict y_i for the 20 observations i which are out-of-sample. For this purpose, derive the marginal data density (MDD) $p(Y|\tau)$ and its natural logarithm. Then, use an optimization command of your choice to find

$$\tau_1^* = \arg\max_{\tau}\ \log\ p(Y|\tau)\ .$$

Solution

Apply the Bayes formula to derive the marginal data density:

$$\begin{aligned}
p(Y) &= \frac{p(Y|\beta)p(\beta)}{p(\beta|Y)} = \\
&= \frac{(2\pi)^{-n/2}\exp\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\}(2\pi)^{k/2}\tau^{k/2}\exp\{-\frac{1}{2\sigma^2}\beta'\tau I\beta\}}{(2\pi)^{-k/2}|\bar{V}|^{-1/2}\exp\{-\frac{1}{2\sigma^2}(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\}} = \\
&= \frac{(2\pi)^{-n/2}\exp\{-\frac{1}{2\sigma^2}Y'Y\}\tau^{k/2}}{|\tau I+X'X|^{1/2}\exp\{-\frac{1}{2\sigma^2}\bar{\beta}'[\tau I+X'X]\bar{\beta}\}}
\end{aligned}$$

Take the log of the expression above to obtain:

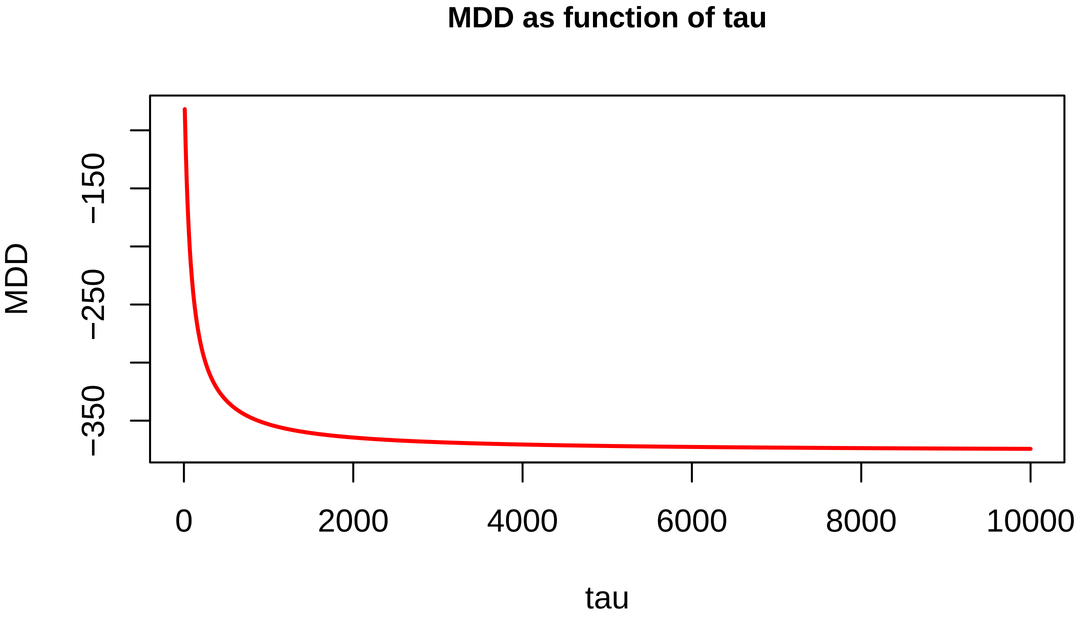
$$\begin{aligned}
\log p(Y) &= c - \frac{1}{2\sigma^2}Y'Y + \frac{k}{2}\log(\tau) - \frac{1}{2}\log(|\tau I+X'X|) + \frac{1}{2\sigma^2}\bar{\beta}'|\tau I+X'X|\bar{\beta} = \\
&= c - \frac{1}{2\sigma^2}[Y'Y - Y'X\bar{V}X'Y] - \frac{1}{2}\log(\tau^{-k}|\tau I+X'X|) = \\
&= c - \frac{1}{2\sigma^2}[Y'Y - Y'X[\tau I+X'X]^{-1}X'Y] - \frac{1}{2}\log(|\tau^{-1}X'X+I|)
\end{aligned}$$

Where $c = -\frac{n}{2}\log(2\pi)$. We can therefore optimise the function:

Exercise 5.2: Solution

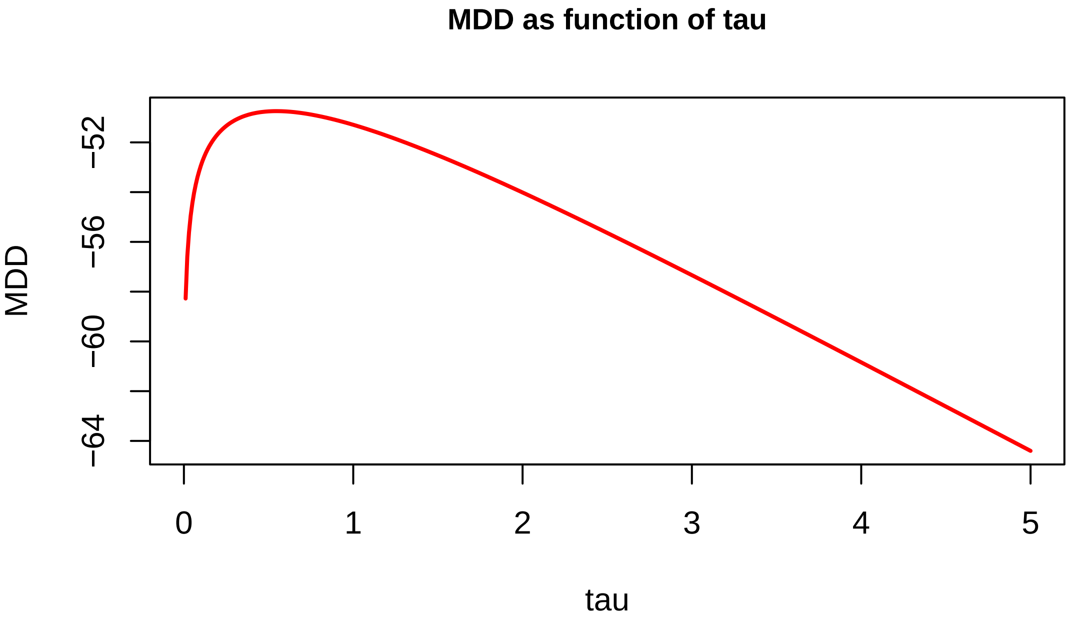
```
# Construct function for MDD:

fGetLogMDD <- function(tau, Y, X, sigmasq) {
    #compute components
  XtX <- t(X) %*% X
  XtY <- t(X) %*% Y
  YtY <- t(Y) %*% Y
  #compute [tau I + X'X]^(-1)
  tauI_XtX_inv <- solve(tau * diag(ncol(X)) + XtX)
  #compute Y'X[tau I + X'X]^(-1)X'Y
  term1 <- t(Y) %*% X %*% tauI_XtX_inv %*% XtY
  #compute det of (\tau^{-1} X'X + I)
  det_val <- det((1 / tau) * XtX + diag(ncol(X)))
  #compute log of determinant
  log_det_val <- log(det_val)
  #put all together
  logMDD = - 1/(2*sigmasq) * (YtY - term1) - (1/2)*log_det_val
  return(-logMDD)
}



# Visualize log MDD:

# coarse grid for tau:
tau_grid2 <- seq(0, 10000, by = 10)
mdd_values <- numeric(length(tau_grid2))

for (i in 1:length(tau_grid2)) {
  mdd_values[i] <- -fGetLogMDD(tau = tau_grid2[i],
                         X = X1,
                         Y = Y1,
                         sigmasq = 10)
}

plot(tau_grid2,
     mdd_values,
     type = "l",
     col = 'red',
     lwd = 2,
     xlab = "tau",
     ylab = "MDD",
     main = "MDD as function of tau",
     cex.main = 0.9)
```

Exercise 5.2: Solution



```r
#finer grid for tau:
tau_grid2 <- seq(0, 5, by = 0.01)
mdd_values <- numeric(length(tau_grid2))

for (i in 1:length(tau_grid2)) {
  mdd_values[i] <- -fGetLogMDD(tau = tau_grid2[i],
                         X = X1,
                         Y = Y1,
                         sigmasq = 10)
}

plot(tau_grid2,
     mdd_values,
     type = "l",
     col = 'red',
     lwd = 2,
     xlab = "tau",
     ylab = "MDD",
     main = "MDD as function of tau",
     cex.main = 0.9)
```

Exercise 5.2: Solution



```
# Optimize fGetMDD:

# function for constrained optimization:

helpfun <- function(tau){ fGetLogMDD(tau,Y1,X1,10) }
o <- optimize(helpfun,interval=c(0,100))
tau_star1 <- o$minimum
tau_star1
```

```
## [1] 0.5470491
```

```
# function for unconstrained optimization:
init_val <- 0.1
tau_star <- optim(init_val,
                   fGetLogMDD,
                   Y = Y1,
                   X = X1,
                   sigmasq = 10,
                   method = "BFGS")
tau_star$par
```

```
## [1] 0.5470436
```

4. Compute two other τ^*'s using approximations of the log MDD: the Bayesian/Schwarz Information Criterion (BIC/SIC),

$$\tau_2^* = \arg\max_{\tau} \; \log \; p(Y|\bar{\beta}(\tau)) - \frac{k}{2} \log \; n \; ,$$

and the Akaike Information Criterion (AIC),

$$\tau_3^* = \arg\max_{\tau} \; 2 \log \; p(Y|\bar{\beta}(\tau)) - 2k \; .$$

Exercise 5.2: Solution

Thereby, $k = 5$ is your number of regressors, $n = 80$ is your sample size, and $p(Y|\bar{\beta}(\tau))$ is the likelihood, $p(Y|\beta)$, evaluated at $\beta = \bar{\beta}(\tau)$.

Solution

```r
#function to get Log Likelihood
fGetLL <- function(beta, X, Y, sigmasq){
  #compute RSS
  residuals <- Y - X %*% beta
  RSS <- t(residuals) %*% residuals
  #compute exponent
  exponent <- -1 / (2*sigmasq) * as.numeric(RSS)
  #term1
  term1 <- -(n/2) * log(2*pi*sigmasq)
  #compute LL
  return(term1 + exponent)
}

#function to return BIC
fGetBIC <- function(tau, X, Y, sigmasq){
  beta_bar <- fGetPostMean(tau, X, Y)
  loglik <- fGetLL(beta_bar, X, Y, sigmasq)
  k <- length(beta_bar)
  n <- length(Y)
  BIC <- loglik - (k/2) * log(n)
  return(-1*BIC)
}

# Optimize BIC:

# function for constrained optimization:

helpfun <- function(tau){ fGetBIC(tau,X1,Y1,10) }
o <- optimize(helpfun,interval=c(0,100))
tau_star2 <- o$minimum

# function for unconstrained optimization:

init_val <- 1
tau_star <- optim(init_val,
                  fGetBIC,
                  Y = Y1,
                  X = X1,
                  sigmasq = 10,
                  method = "BFGS",
                  control = list(reltol = 1e-12))
```

Exercise 5.2: Solution

```r
tau_star$par
```

```
## [1] 2.378113e-08
```

```r
# Optimize AIC:

fGetAIC <- function(tau, X, Y, sigmasq){
  beta_bar <- fGetPostMean(tau, X, Y)
  loglik <- fGetLL(beta_bar, X, Y, sigmasq)
  k <- length(beta_bar)
  n <- length(Y)
  AIC <- 2*loglik - 2*k
  return(-1*AIC)
}

# function for constrained optimization:

helpfun <- function(tau){ fGetAIC(tau,X1,Y1,10) }
o <- optimize(helpfun,interval=c(0,100))
tau_star3 <- o$minimum
tau_star3
```

```
## [1] 4.627768e-05
```

```r
# function for unconstrained optimization:

init_val <- 1
tau_star <- optim(init_val,
                  fGetAIC,
                  Y = Y1,
                  X = X1,
                  sigmasq = 10,
                  method = "BFGS",
                  control = list(reltol = 1e-12))

tau_star$par
```

```
## [1] 2.321357e-08
```

Since both AIC and BIC, expressed as functions of τ, are proportional to the log likelihood:

$$\text{BIC}(\tau) \propto \text{AIC}(\tau) \propto p(Y|\bar{\beta}(\tau))$$

And the log likelihood is maximised by definition at $\hat{\beta}_{ML} = (X'X)^{-1}X'Y$, the optimal value of τ that maximises $p(Y|\bar{\beta}(\tau))$ is $\tau^* = 0$, as we showed before that $\bar{\beta}(\tau) \to \hat{\beta}_{ML}$ as $\tau \to 0$. (The likelihood function derived exercise 1 is defined for any vector β, in particular for any $\bar{\beta}(\tau)$, i.e. for any $\tau \geq 0$.)

We thus have, both for AIC and BIC:

Exercise 5.2: Solution

$$\tau_2^* = \tau_3^* = 0$$

AIC and BIC are not suited for the model selection we would like to perform here; they are only useful when we compare models with different numbers of covariates.

5. For each value of $\tau^* \in \{\tau_1^*, \tau_2^*, \tau_3^*\}$, compute $\bar{\beta}(\tau^*)$, predict the outcome variable y_i of the 20 out-of-sample observations i and calculate the out-of-sample mean squared error (MSE),

$$\frac{1}{20}\sum_{i=1}^{20}(y_i - \hat{y}_i^*)^2 \ , \quad \text{where } \hat{y}_i^* = x_i'\bar{\beta}(\tau^*) \ .$$

Compare the three MSE_1, MSE_2 and MSE_3. Which value of τ^* results in the lowest MSE?

Solution

```
#compute betas for different tau
beta_logMDD <- fGetPostMean(tau_star1,X1,Y1)
beta_BIC <- fGetPostMean(tau_star2,X1,Y1)
beta_AIC <- fGetPostMean(tau_star3,X1,Y1)

#fit Y over three testing samples
Y_logMDD <- X2 %*% beta_logMDD
Y_BIC <- X2 %*% beta_BIC
Y_AIC <- X2 %*% beta_AIC

#compute MSE
MSE_logMDD <- mean((Y2-Y_logMDD)^2)
MSE_BIC <- mean((Y2-Y_BIC)^2)
MSE_AIC <- mean((Y2-Y_AIC)^2)

MSE_logMDD
```

```
## [1] 3.982616
```

```
MSE_BIC
```

```
## [1] 3.999888
```

```
MSE_AIC
```

```
## [1] 3.999888
```

We get a slightly better MSE with the optimal tau from logMDD than with the other two, and add that we could in fact also find the τ that minimises MSE:

```
# generate function for MSE
fGetMSE <- function(tau){
  beta_tau <- fGetPostMean(tau,X1,Y1)
  Y_tau <- X2 %*% beta_tau
  MSE_tau <- mean((Y2-Y_tau)^2)
```

Exercise 5.2: Solution

```
  return(MSE_tau)
  }

# Minimise

init_val <- 1
tau_min_MSE <- optim(init_val,
                     fGetMSE,
                     method = "BFGS",
                     control = list(reltol = 1e-12))
tau_min_MSE$par
```

```
## [1] 1.398999
```

```
fGetMSE(tau_min_MSE$par)
```

```
## [1] 3.972685
```

Exercise 5.3: Solution

Bayesian inference on both β as well as σ^2 in a linear regression model is done standardly by assuming the following prior distribution:

$$p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2) \ ,$$

with

$$\beta|\sigma^2 \sim N(\underline{\beta}, \sigma^2\underline{V}) \ , \quad \sigma^2 \sim IG(\underline{\nu}, \underline{s}^2) \ ,$$

i.e. we break up the joint prior of (β, σ^2) into a conditional prior for $\beta|\sigma^2$ and a marginal prior for σ^2. Thereby, $\underline{\beta}$, $\underline{V}$, $\underline{\nu}$ and $\underline{s}^2$ are hyperparameters that define the exact shape of the prior distribution and that have to specified by the researcher. As shown in the lecture notes, this results in the following posterior, which is also broken up into a conditional and a marginal:

$$p(\beta, \sigma^2|Y) = p(\beta|Y, \sigma^2)p(\sigma^2|Y) \ ,$$

with

$$\beta|Y, \sigma^2 \sim N(\bar{\beta}, \sigma^2\bar{V}) \ , \quad \sigma^2|Y \sim IG(\bar{\nu}, \bar{s}^2) \ .$$

Thereby,

$$\bar{V} = [\underline{V}^{-1} + X'X]^{-1} \ , \qquad \bar{\beta} = \bar{V}[X'Y + \underline{V}^{-1}\underline{\beta}] \ ,$$

$$\bar{\nu} = \underline{\nu} + n \ , \qquad \bar{s}^2 = \underline{s}^2 + Y'Y + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - \bar{\beta}'\bar{V}^{-1}\bar{\beta} \ .$$

are parameters that define the exact shape of the prior distribution and that are obtained using the hyperparameters $\underline{\beta}$, $\underline{V}$, $\underline{\nu}$ and $\underline{s}^2$ and the data (Y, X).

1. Let's take a Ridge-prior for $\beta|\sigma^2$, i.e. take $\bar{\beta} = [0, 0, 0, 0, 0]'$ and $\underline{V} = \frac{1}{\tau^*}I$ for a τ^* of your choice in $[0, 10]$. Also, let's take $\underline{\nu} = 4$ and $\underline{s}^2 = 10$. Draw $M = 1000$ values from the prior $p(\beta, \sigma^2) = p(\beta_1, \beta_2, ..., \beta_5, \sigma^2)$, denoted by $\{\beta_{(m)}, \sigma^2_{(m)}\}_{m=1}^M$. To obtain a draw $\beta_{(m)}, \sigma^2_{(m)}$, draw first

$$\sigma^2_{(m)} \sim IG(\underline{\nu}, \underline{s}^2) \ ,$$

and then, conditional on this $\sigma^2_{(m)}$, draw

$$\beta_{(m)}|\sigma^2_{(m)} \sim N(\underline{\beta}, \sigma^2_{(m)}\underline{V}) \ .$$

Plot the marginal prior of each of your six parameters $\{\beta_1, ..., \beta_5, \sigma^2\}$ by plotting a histogram of your draws: e.g. for β_1 we have

$$p(\beta_1) \approx \text{histogram of } \{\beta_{1,(m)}\}_{m=1}^M \ .$$

Also, compute the prior means and variances of each parameter by approximating them with the means and variances across your draws: e.g., for β_1 we have

$$\mathbb{E}[\beta_1] \approx \frac{1}{M}\sum_{m=1}^M \beta_{1,(m)} \equiv \mu_1 \ , \quad \mathbb{V}[\beta_1] \approx \frac{1}{M}\sum_{m=1}^M (\beta_{1,(m)} - \mu_1)^2 \ .$$

Finally, illustrate the joint prior of β_2 and σ^2 by showing a scatterplot of their draws.

Exercise 5.3: Solution

Note that I define the Inverse Gamma distribution $IG(\nu, s^2)$ to have pdf

$$f(x) \propto x^{-(\nu+2)/2} exp\left\{-\frac{s^2}{2x}\right\} ,$$

whereas some other people and software commands define the Inverse Gamma distribution as $IG(\alpha, \beta)$ with pdf proportional to

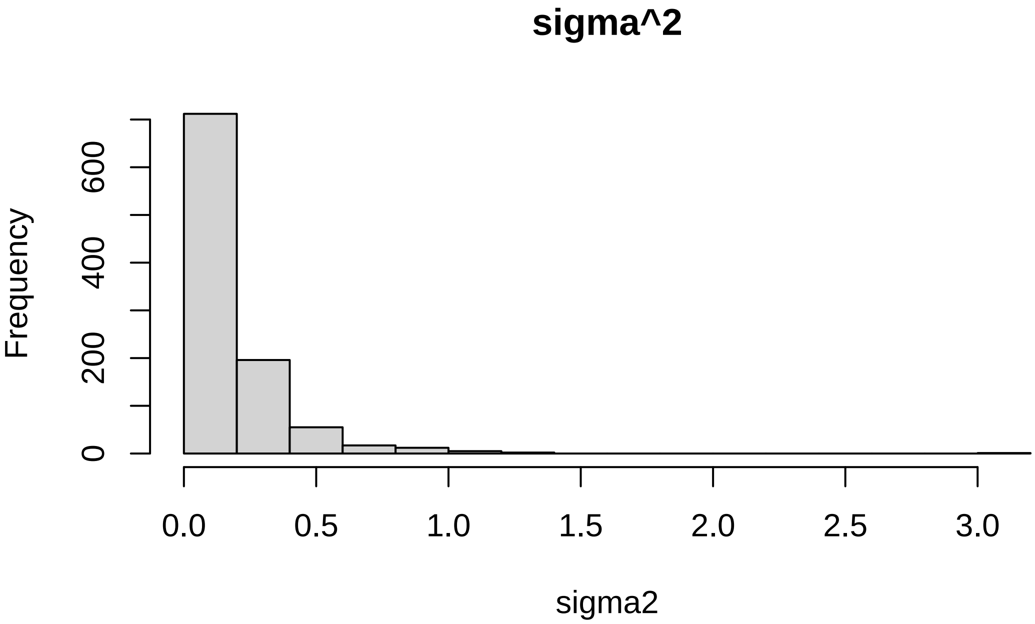
$$f(x) \propto x^{-(\alpha+1)} exp\left\{-\beta/x\right\} .$$

Check which definition applies in your software of choice. If it's the latter, you simply take ν and s^2 from above and you transform them into $\beta = s^2/2$ and $\alpha = \nu/2$.

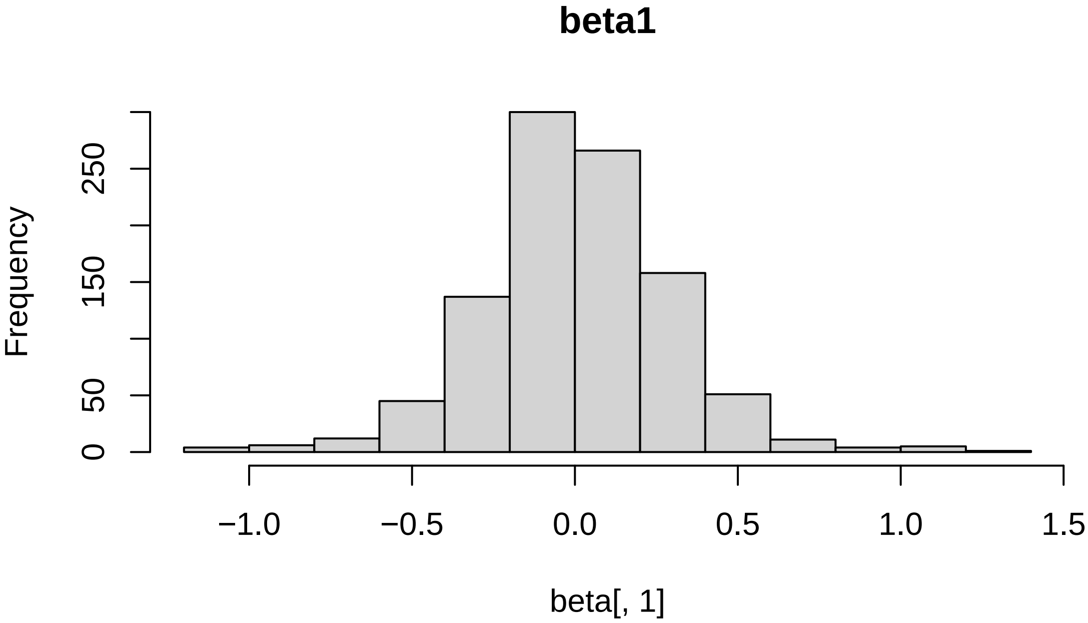
Solution

```r
rm(list = ls())
set.seed(2024)

# Generate inverse gamma function
rinvgamma <- function(n, shape, scale) {
    1 / rgamma(n, shape = shape, rate = 1 / scale)
}
nu <- 2 #transform alpha = nu/2
s2 <- 5 #transform beta = s2/2

tau_star <- 2

#draw simulations
M <- 1000
sigma2 <- numeric(length = M)
beta <- matrix(NA, nrow = M, ncol = 5)
for (i in 1:M) {
    #draw sigma
    sigma2[i] <- rinvgamma(n = 1, shape = nu, scale = s2)
    #draw beta conditional on sigma
    beta[i, 1] <- rnorm(1, 0, sqrt(sigma2[i] / tau_star))
    beta[i, 2] <- rnorm(1, 0, sqrt(sigma2[i] / tau_star))
    beta[i, 3] <- rnorm(1, 0, sqrt(sigma2[i] / tau_star))
    beta[i, 4] <- rnorm(1, 0, sqrt(sigma2[i] / tau_star))
    beta[i, 5] <- rnorm(1, 0, sqrt(sigma2[i] / tau_star))
}

#plot histograms of draws
hist(sigma2, main = "sigma^2")
```

Exercise 5.3: Solution

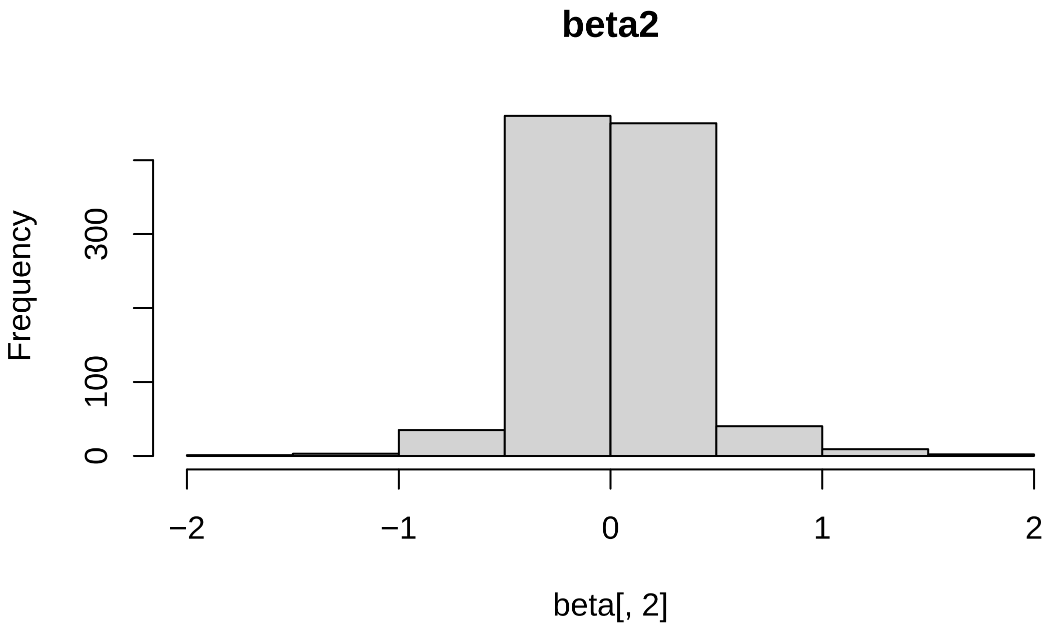


```
hist(beta[ ,1], main = "beta1")
```

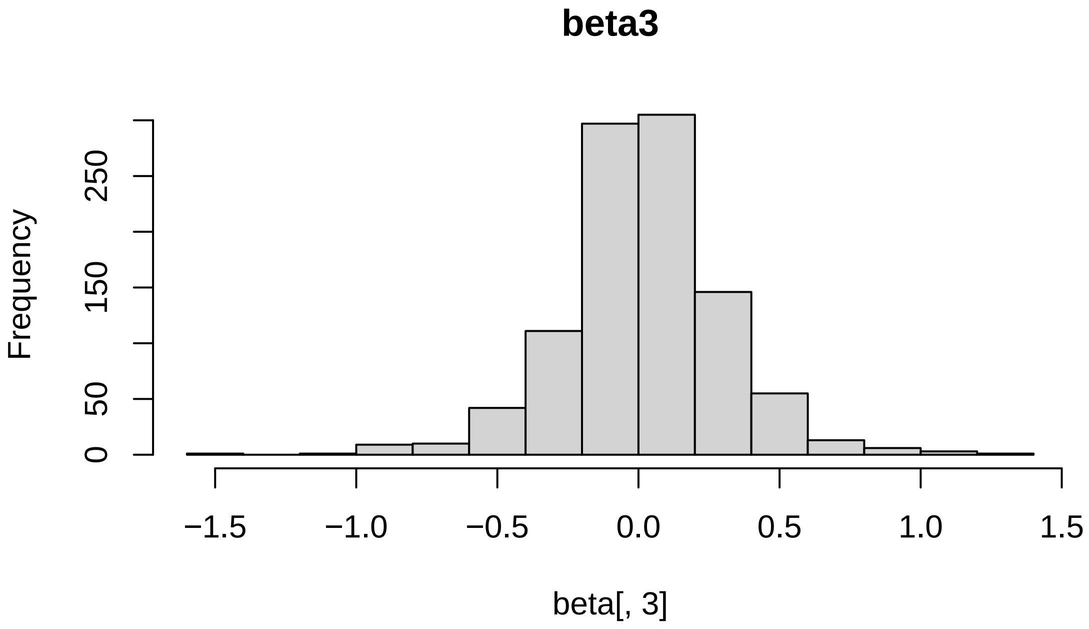


```
hist(beta[ ,2], main = "beta2")
```

Exercise 5.3: Solution

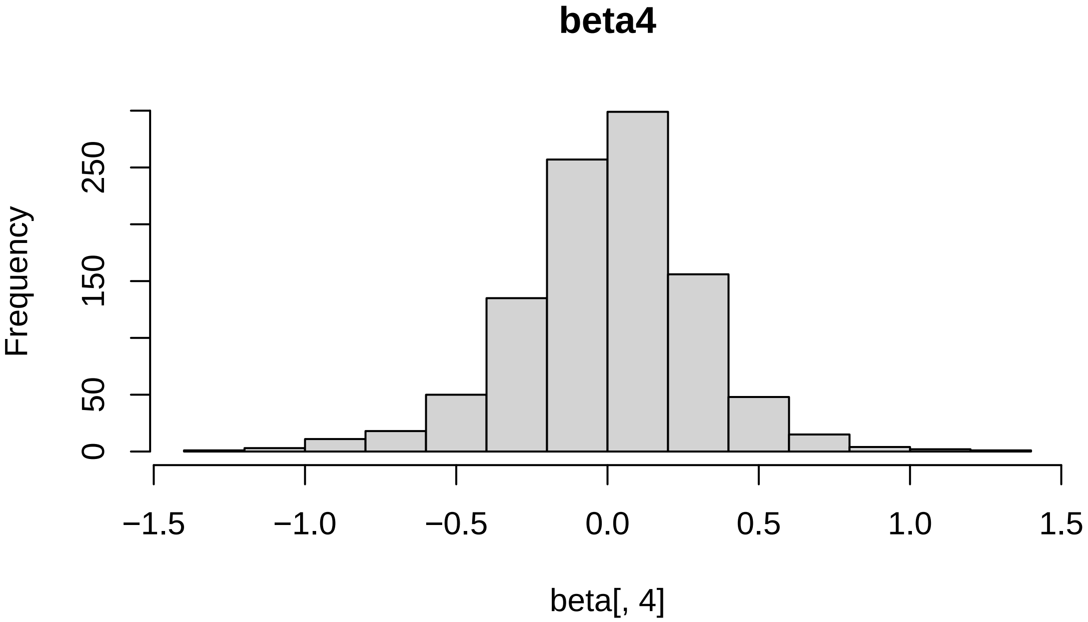


```
hist(beta[ ,3], main = "beta3")
```

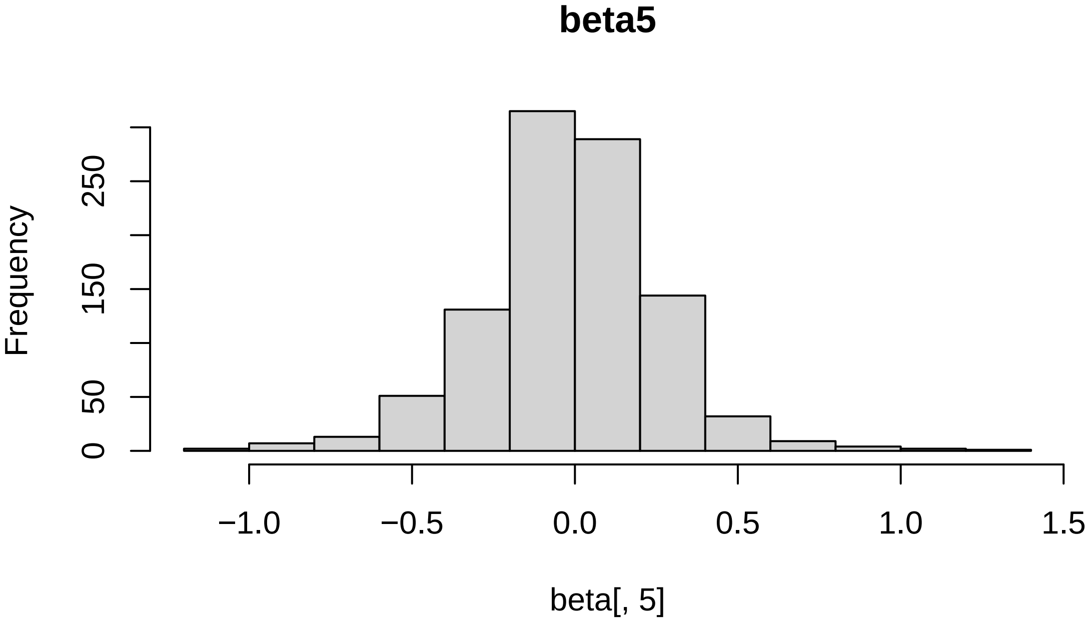


```
hist(beta[ ,4], main = "beta4")
```

Exercise 5.3: Solution



```r
hist(beta[ ,5], main = "beta5")
```



```r
#compute means
mean_beta1 <- (1/M)*sum(beta[,1])
mean_beta2 <- (1/M)*sum(beta[,2])
mean_beta3 <- (1/M)*sum(beta[,3])
mean_beta4 <- (1/M)*sum(beta[,4])
mean_beta5 <- (1/M)*sum(beta[,5])

mean_beta1
```

```
## [1] 0.004857497
```

Exercise 5.3: Solution

```
mean_beta2
```

```
## [1] 0.01632097
```

```
mean_beta3
```

```
## [1] 0.01921502
```

```
mean_beta4
```

```
## [1] 0.000445152
```

```
mean_beta5
```

```
## [1] -0.00933961
```

```
#compute variances
var_beta1 <- (1/M)*sum((beta[,1]-mean_beta1)^2)
var_beta2 <- (1/M)*sum((beta[,2]-mean_beta2)^2)
var_beta3 <- (1/M)*sum((beta[,3]-mean_beta3)^2)
var_beta4 <- (1/M)*sum((beta[,4]-mean_beta4)^2)
var_beta5 <- (1/M)*sum((beta[,5]-mean_beta5)^2)

var_beta1
```

```
## [1] 0.08537771
```

```
var_beta2
```

```
## [1] 0.09899042
```

```
var_beta3
```

```
## [1] 0.08186849
```

```
var_beta4
```

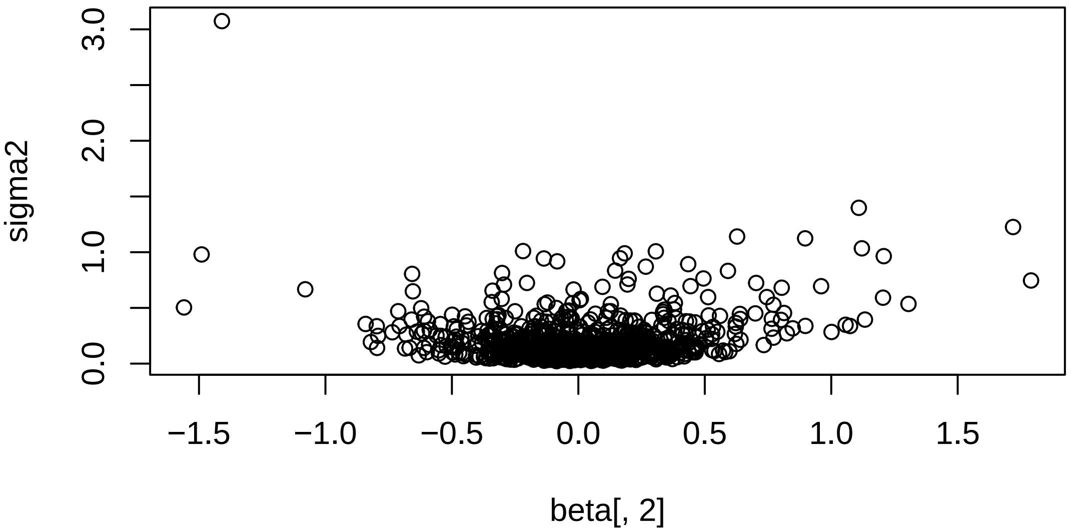
```
## [1] 0.08935099
```

```
var_beta5
```

```
## [1] 0.07569401
```

```
#plot joint prior beta2 and sigma2
plot(beta[,2], sigma2)
```

Exercise 5.3: Solution

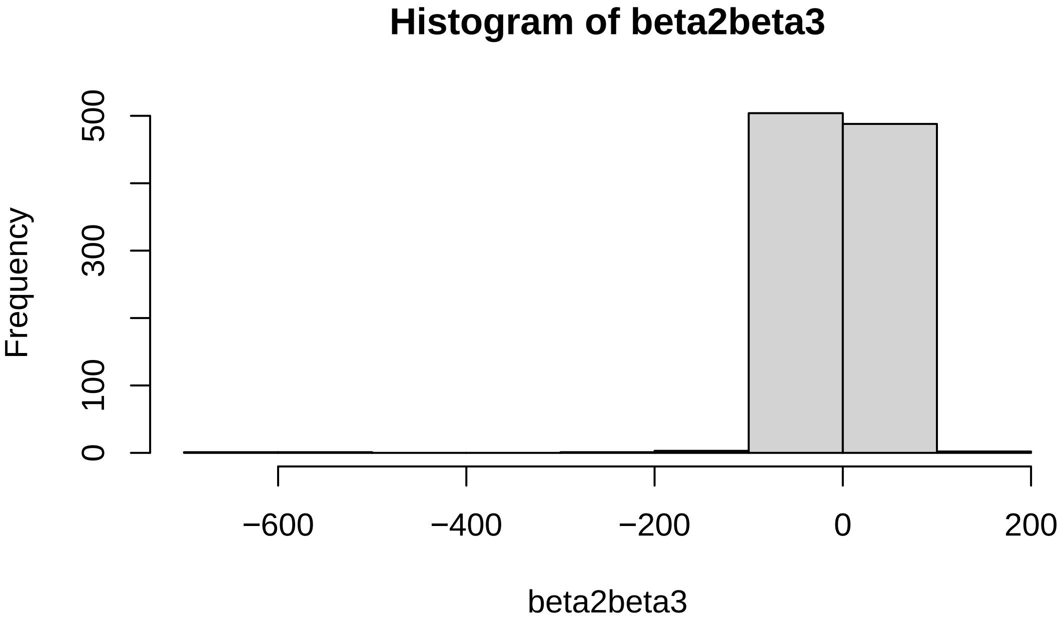


2. Show the prior distribution of β_2/β_3 by plotting a histogram of your draws $\{\beta_{2,(m)}/\beta_{3,(m)}\}_{m=1}^{M}$. Also, use your draws to compute the prior mean and variance of β_2/β_3:

$$\mathbb{E}[\beta_2/\beta_3] \approx \frac{1}{M}\sum_{m=1}^{M} \beta_{2,(m)}/\beta_{3,(m)} \equiv \mu \ , \quad \mathbb{V}[\beta_2/\beta_3] \approx \frac{1}{M}\sum_{m=1}^{M} (\beta_{2,(m)}/\beta_{3,(m)} - \mu)^2 \ .$$

Solution

```
#plot of draws of beta2/beta3
beta2beta3 <- beta[,2]/beta[,3]
hist(beta2beta3)
```



```
#prior mean of beta2/beta3
mean_beta2beta3 <- (1/M)*sum(beta2beta3)
mean_beta2beta3
```

Exercise 5.3: Solution

```
## [1] -1.925752
```

```
#prior variance of beta2/beta3
var_beta2beta3 <- (1/M)*sum((beta2beta3-mean_beta2beta3)^2)
var_beta2beta3
```

```
## [1] 969.2392
```

3. Simulate a dataset of $n = 100$ observations, proceeding as follows:
   (a) let $x_{1i} = 1 \; \forall \; i$ and draw $x_{2i}, x_{3i}, x_{4i}, x_{5i} \sim N(0, 1)$ for $i = 1 : n$,
   (b) draw $u_i \sim N(0, \sigma^2)$ for $i = 1 : n$, with $\sigma^2 = 10$,
   (c) let $\beta_0 = [3, 8, -4, 0, 0]'$ (the true value of β),
   (d) for $i = 1 : n$, compute $y_i = x_i'\beta_0 + u_i$, with $x_i = [x_{1i}, ..., x_{5i}]'$.

Solution

```
#sample size
n <- 100
#explanatory variables
x2 <- rnorm(n)
x3 <- rnorm(n)
x4 <- rnorm(n)
x5 <- rnorm(n)
X <- cbind(1, x2, x3, x4, x5)
#disturbance term
U <- rnorm(n, 0, sqrt(10))
#true parameters
beta0 <- c(3, 8, -4, 0, 0)
#generate y
Y <- X %*% beta0 + U
```

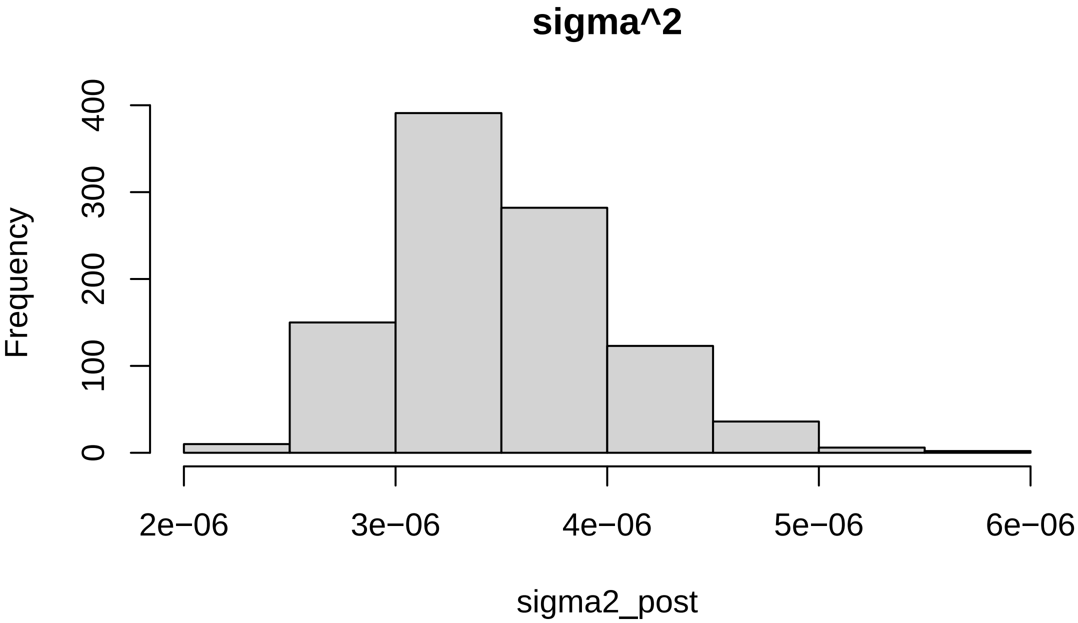
4. Use your dataset to compute the posterior of (β, σ^2). Specifically, compute $\bar{\beta}$, $\bar{V}$, $\bar{\nu}$ and $\bar{s}^2$ using your dataset, and repeat exercises (a) and (b): draw from the posterior $p(\beta, \sigma^2|Y) = p(\beta_1, \beta_2, ..., \beta_5, \sigma^2|Y)$,and use your draws to plot marginal posteriors, to compute means and variances of each parameter, to plot the joint posterior of β_2 and σ^2, and to show the posterior distribution of β_2/β_3.

Solution

```
# Install MASS package if not already installed
if (!requireNamespace("MASS", quietly = TRUE)) {
  install.packages("MASS")
}

# Load MASS package
library(MASS)

#compute posterior means and variances
beta_bar <- solve( t(X) %*% X + tau_star * diag(ncol(X)) ) %*% t(X) %*% Y
```

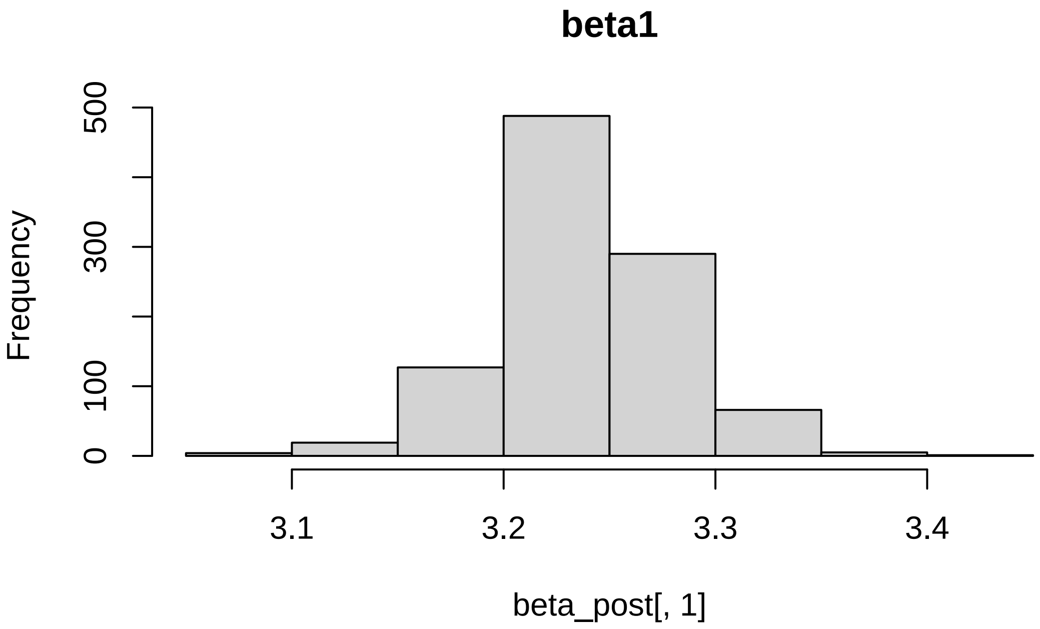
Exercise 5.3: Solution

```r
XtX <- t(X) %*% X
XtY <- t(X) %*% Y
tau_starI <- tau_star * diag(ncol(X))
XtX_tau_starI <- XtX + tau_starI
V_bar <- solve(XtX_tau_starI)
nu_bar <- (4 + n)/2
s2_bar <- (10 + t(Y)%*%Y -t(beta_bar) %*% V_bar %*% beta_bar)/2

#draw simulations
sigma2_post <- numeric(length = M)
beta_post <- matrix(NA, nrow = M, ncol = 5)
for (i in 1:M) {
    #draw sigma posterior
    sigma2_post[i] <- rinvgamma(n = 1, shape = nu_bar, scale = s2_bar)
    #draw beta conditional on sigma
    beta_post[i, ] <- mvrnorm(1, beta_bar, sigma2[i]*V_bar)
}

#plot histograms of draws
hist(sigma2_post, main = "sigma^2")
```

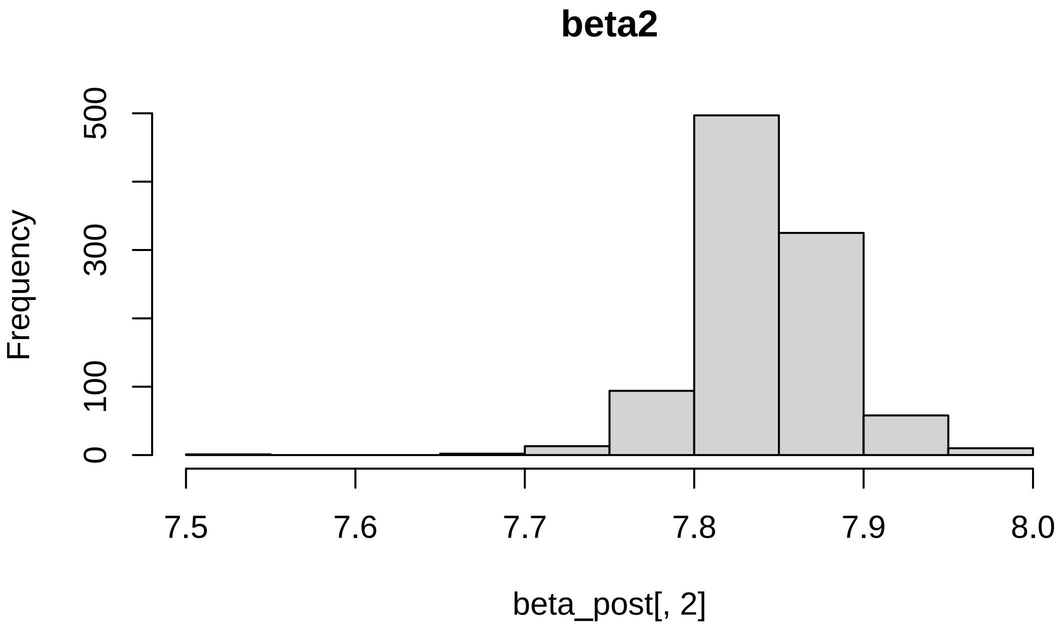


```r
hist(beta_post[ ,1], main = "beta1")
```

Exercise 5.3: Solution

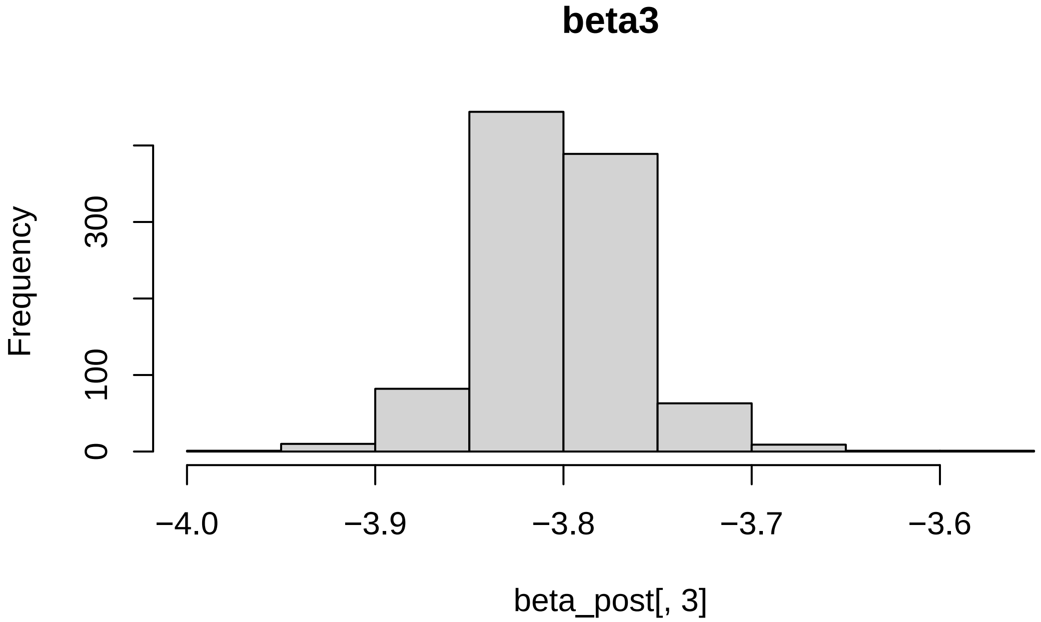


```
hist(beta_post[ ,2], main = "beta2")
```

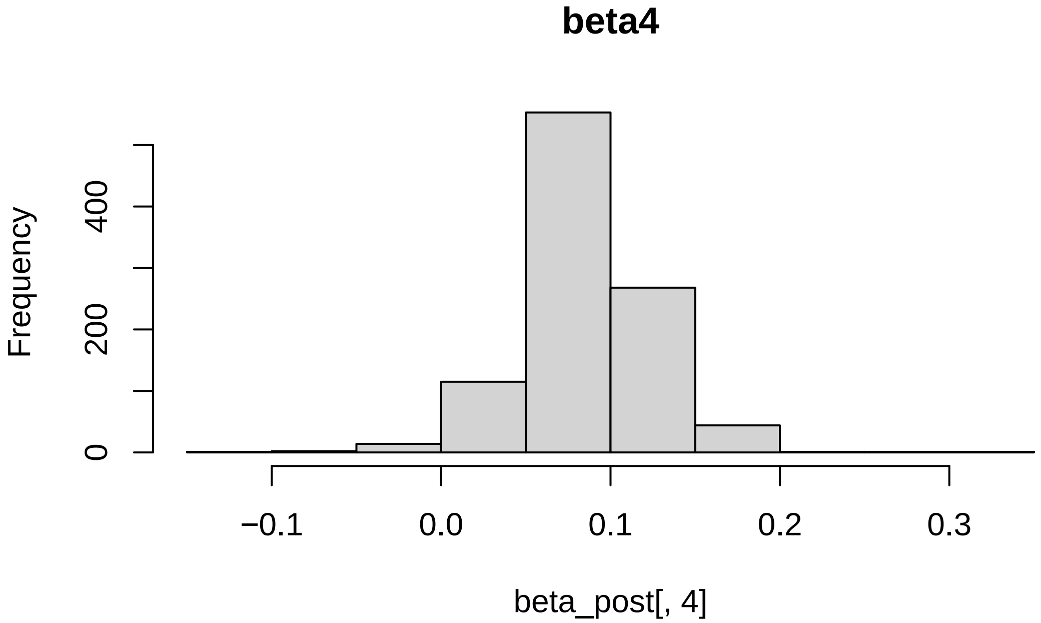


```
hist(beta_post[ ,3], main = "beta3")
```

Exercise 5.3: Solution

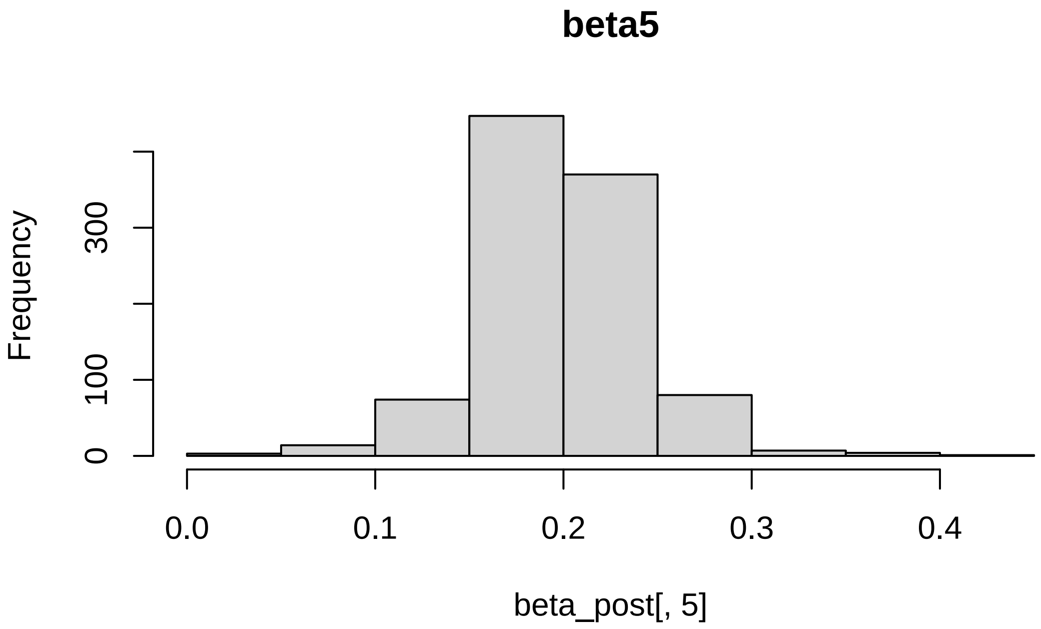


```r
hist(beta_post[ ,4], main = "beta4")
```



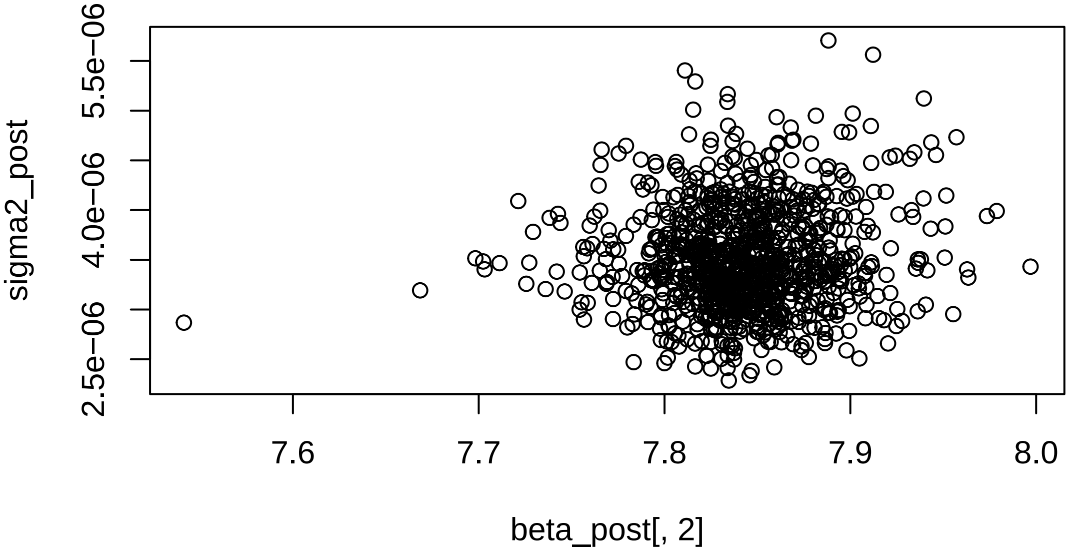
```r
hist(beta_post[ ,5], main = "beta5")
```

Exercise 5.3: Solution

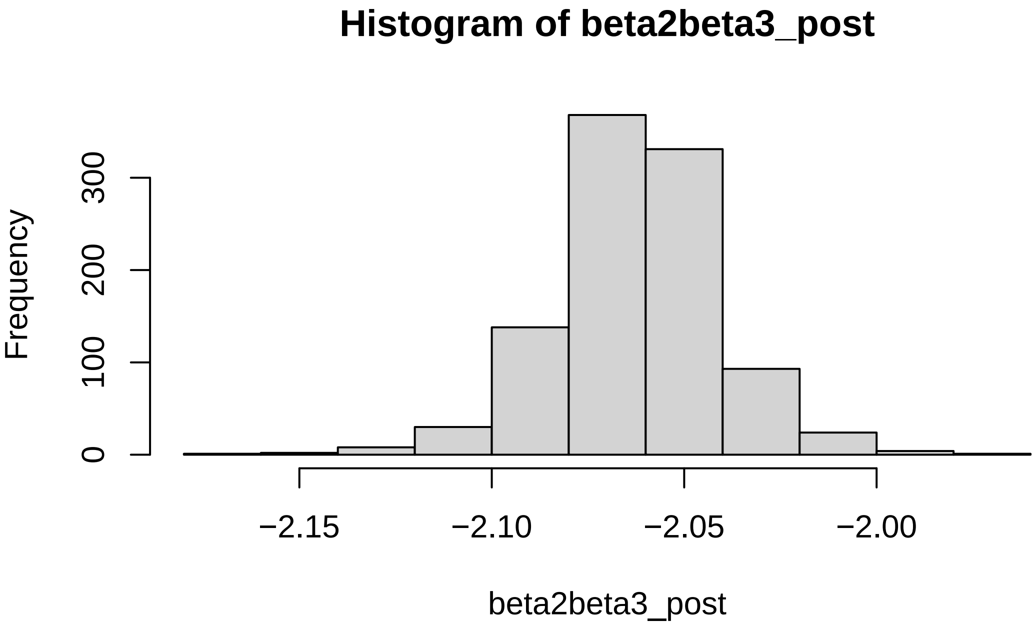


```
#compute means
mean_beta1_post <- (1/M)*sum(beta_post[,1])
mean_beta2_post <- (1/M)*sum(beta_post[,2])
mean_beta3_post <- (1/M)*sum(beta_post[,3])
mean_beta4_post <- (1/M)*sum(beta_post[,4])
mean_beta5_post <- (1/M)*sum(beta_post[,5])

#compute variances
var_beta1_post <- (1/M)*sum((beta_post[,1]-mean_beta1_post)^2)
var_beta2_post <- (1/M)*sum((beta_post[,2]-mean_beta2_post)^2)
var_beta3_post <- (1/M)*sum((beta_post[,3]-mean_beta3_post)^2)
var_beta4_post <- (1/M)*sum((beta_post[,4]-mean_beta4_post)^2)
var_beta5_post <- (1/M)*sum((beta_post[,5]-mean_beta5_post)^2)

#plot joint posterior beta2 and sigma2
plot(beta_post[,2], sigma2_post)
```

Exercise 5.3: Solution



```
#plot of draws of beta2 vs. draws of beta3 posterior
beta2beta3_post <- beta_post[,2]/beta_post[,3]
hist(beta2beta3_post)
```



```
#posterior mean of beta2/beta3
mean_beta2beta3_post <- (1/M)*sum(beta_post[,2]/beta_post[,3])
mean_beta2beta3_post
```

```
## [1] -2.06261
```

```
#posterior variance of beta2/beta3
var_beta2beta3_post <-
(1/M)*sum((beta_post[,2]/beta_post[,3]-mean_beta2beta3_post)^2)
var_beta2beta3_post
```

Exercise 5.3: Solution

```
## [1] 0.0004820653
```

Exercise 5.4: Solution

The linear regression model

$$y_i = x_i'\beta + u_i \quad , \quad u_i|x_i \sim N(0, \sigma^2) \tag{1}$$

and the prior $\beta|\,\sigma^2 \sim N\left(0, \sigma^2 \frac{1}{\lambda} I\right)$ lead to the posterior $\beta|Y, \sigma^2 \sim N(\bar{\beta}, \sigma^2 \bar{V})$ with

$$\bar{V} = [X'X + \lambda I]^{-1} \; , \quad \bar{\beta} = [X'X + \lambda I]^{-1} X'Y \; .$$

(a) The lecture notes claim that

$$\bar{\beta} = \arg\min_{\beta} \, (Y - X\beta)'(Y - X\beta) + \lambda \sum_{j=1}^{k} \beta_j^2 \; , \tag{2}$$

i.e. the posterior mean under the above prior equals the Ridge estimator, which minimizes the sum of squared residuals plus the penalty-term $\lambda \sum_{j=1}^{k} \beta_j^2$. Let's show that. Specifically, derive the likelihood of the linear regression in Eq. (1) and multiply it by the prior to get the posterior up to proportionality:

$$p(\beta|Y) = \frac{p(Y|\beta)p(\beta)}{p(Y)} \propto p(Y|\beta)p(\beta) \; .$$

Because this posterior is a Normal distribution, you know that $\bar{\beta}$ is not only the posterior mean, but also the posterior mode, i.e.

$$\bar{\beta} = \arg\max_{\beta} \; p(\beta|Y) = \arg\max_{\beta} \; p(Y|\beta)p(\beta) \; .$$

Use this result to show the claim from Eq. (2).
Hint: note that the pdf of a multivariate normal distribution $X \sim N(\mu, \Sigma)$ is

$$f(x) = (2\pi)^{-k/2} |\Sigma|^{-1/2} exp \left\{ -\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu) \right\} \; ,$$

where k is the dimension (length) of the vector X. Also, note that $\sum_{j=1}^{k} \beta_j^2 = \beta'\beta$.

Exercise 5.4: Solution

Solution: The posterior mode – here denoted by $\bar{\beta}$ – is the

$$\begin{aligned}
&\arg\max_{\beta}\ p(\beta|Y) \\
=&\arg\max_{\beta}\ p(Y|\beta)p(\beta) \\
=&\arg\max_{\beta}\ \log\ p(Y|\beta)+\log\ p(\beta) \\
=&\arg\min_{\beta}\ -\log\ p(Y|\beta)-\log\ p(\beta)\ .
\end{aligned}$$

This result will be used in all of the exercises in this problem.

Under the Ridge prior $\beta|\ \sigma^2 \sim N\left(0, \sigma^2\frac{1}{\lambda}I\right)$ we have

$$\begin{aligned}
p(\beta) &= (2\pi)^{-k/2}\left|\sigma^2\frac{1}{\lambda}I\right|^{-1/2} exp\left\{-\frac{1}{2}(\beta-0)'\left[\sigma^2\frac{1}{\lambda}I\right]^{-1}(\beta-0)\right\} \\
&= (2\pi)^{-k/2}\left|\sigma^2\frac{1}{\lambda}I\right|^{-1/2} exp\left\{\frac{1}{2\sigma^2}\lambda\beta'\beta\right\}\ ,
\end{aligned}$$

and hence

$$\log\ p(\beta) = c_1 - \frac{1}{2\sigma^2}\lambda\beta'\beta\ ,$$

where $c_1 = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log\ |\sigma^2\frac{1}{\lambda}I|$ does not depend on β and hence can be treated as a constant.

Under the linear regression model, we have the likelihood

$$p(Y|\beta) = (2\pi\sigma^2)^{-n/2}exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\}\ ,$$

and hence

$$\log\ p(Y|\beta) = c_2 - \frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\ .$$

Exercise 5.4: Solution

Putting the two pieces together, we have that the posterior mode, $\bar{\beta}$ is the

$$\begin{aligned}
&\arg\min_{\beta} \; -\log\; p(Y|\beta) - \log\; p(\beta) \\
=&\arg\min_{\beta} \; -c_2 + \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) - c_1 + \frac{1}{2\sigma^2}\lambda\beta'\beta \\
=&\arg\min_{\beta} \; \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) + \frac{1}{2\sigma^2}\lambda\beta'\beta \\
=&\arg\min_{\beta} \; (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta \; .
\end{aligned}$$

(b) Now use analogous steps as above to show that the posterior mode under the prior

$$p(\beta) = \prod_{j=1}^{k} \frac{1}{2}\lambda \, exp\left\{-\lambda|\beta_j|\right\}$$

solves

$$\min_{\beta} \; (Y - X\beta)'(Y - X\beta) + \tilde{\lambda}\sum_{j=1}^{k}|\beta_j| \; ,$$

for $\tilde{\lambda} = 2\lambda\sigma^2$ (a one-to-one transformation of λ). We call this posterior mode the Lasso estimator. Note that it cannot be obtained analytically.

Solution: The log of the Laplace prior,

$$p(\beta) = \prod_{j=1}^{k} \frac{1}{2}\lambda \exp\left\{-\lambda|\beta_j|\right\} = \left(\frac{\lambda}{2}\right)^k \exp\left\{-\lambda\sum_{j=1}^{k}|\beta_j|\right\} \; ,$$

is given by

$$\log p(\beta) = c_1 - \lambda\sum_{j=1}^{k}|\beta_j| \; ,$$

Exercise 5.4: Solution

where $c_1 = k(\log(\lambda) - \log(2))$. Hence, for the posterior mode $\bar{\beta}$ we get

$$\begin{aligned}
&\arg\min_{\beta} -\log p(Y|\beta) - \log p(\beta) \\
&= \arg\min_{\beta} -c_2 + \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) - c_1 + \lambda \sum_{j=1}^{k} |\beta_j| \\
&= \arg\min_{\beta} \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) + \lambda \sum_{j=1}^{k} |\beta_j| \\
&= \arg\min_{\beta} (Y - X\beta)'(Y - X\beta) + \tilde{\lambda} \sum_{j=1}^{k} |\beta_j| \ ,
\end{aligned}$$

where $\tilde{\lambda} = 2\sigma^2\lambda$, and $c_2 = -\frac{n}{2}\log(2\pi\sigma^2)$, as before.

(c) One can use the two Ridge- and Laplace-priors above not only to obtain the Ridge/Lasso estimator of β in the linear regression model, but also in many other models, e.g. the Probit model. Again, use analogous steps as above to show that the posterior mode under the Probit model and the Ridge-prior solves

$$\max_{\beta} \ \log \ p(Y|X, \beta) - \tilde{\lambda} \sum_{j=1}^{k} \beta_j^2 \ ,$$

where $\tilde{\lambda} = \lambda/2$ is a simple one-to-one transformation of λ, and log $p(Y|X, \beta)$ is the log-likelihood under the Probit model. Note that you do not have to derive this likelihood! Also, show that the posterior mode under the Probit model and the Laplace-prior solves

$$\max_{\beta} \ \log \ p(Y|X, \beta) - \lambda \sum_{j=1}^{k} |\beta_j| \ .$$

Solution: For the Ridge-prior, we have

$$\begin{aligned}
\bar{\beta} &= \arg\max_{\beta} \log p(Y|\beta) + \log p(\beta) \\
&= \arg\max_{\beta} \log p(Y|\beta) + c_1 - \frac{1}{2\sigma^2}\lambda\beta'\beta \\
&= \arg\max_{\beta} \log p(Y|\beta) - \tilde{\lambda}\beta'\beta \ ,
\end{aligned}$$

where $\tilde{\lambda} = \frac{1}{2\sigma^2}\lambda$ and $c_1 = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2\frac{1}{\lambda}I|$, and we have made use of the fact that in the Probit model σ^2 is a scaling parameter that cannot be identified

Exercise 5.4: Solution

independently of β and is therefore set equal to 1.

For the Laplace-prior, we have

$$\begin{aligned}\bar{\beta} &= \arg\max_{\beta} \log p(Y|\beta) + \log p(\beta) \\ &= \arg\max_{\beta} \log p(Y|\beta) + c_1 - \lambda \sum_{j=1}^{k} |\beta_j| \\ &= \arg\max_{\beta} \log p(Y|\beta) - \lambda \sum_{j=1}^{k} |\beta_j| \ .\end{aligned}$$

Note that in both of these expressions the penalty enters with a minus-sign, whereas in the previous two exercises it entered with a plus-sign. This is because we are maximizing the likelihood, whereas the optimization problems in the previous two exercises were about minimizing the sum of squared residuals.

Exercise 5.5: Solution

Consider the linear regression model

$$y_i = x_i'\beta + u_i \ ,$$

with $\mathbb{E}[x_i u_i] = 0$ and $\mathbb{V}[u_i|x_i] = \sigma^2$. Suppose you observe n i.i.d. observations: $\{(y_i, x_i)\}_{i=1:n}$.

(a) (5 points) Derive the OLS estimator $\hat{\beta}_{OLS}$. What do you know about the finite sample distribution of $\hat{\beta}_{OLS}|X$? Is $\hat{\beta}_{OLS}$ unbiased? Is it consistent?

Solution:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_i (y_i - x_i'\beta)^2 = \arg\min_{\beta} (Y - X\beta)'(Y - X\beta) = (X'X)^{-1}X'Y \ ,$$

obtained by solving the FOC $-2X'(Y - X\beta) = 0$, provided that $X'X$ is of full rank. **[1p]**

Regarding the finite sample distribution of $\hat{\beta}_{OLS}|X$, we know the following. Under $Y = X\beta + U$, we have $\hat{\beta}_{OLS} = \beta + (X'X)^{-1}X'U$. Hence, $\mathbb{E}[\hat{\beta}_{OLS}|X] = \beta + (X'X)^{-1}X'\mathbb{E}[U|X]$, which may or may not be equal to β, i.e. $\hat{\beta}_{OLS}$ may or may not be unbiased. Also, owing to the homoskedasticity assumption $\mathbb{V}[u_i|x_i] = \sigma^2$, we have

$$\mathbb{V}[\hat{\beta}_{OLS}|X] = (X'X)^{-1}\mathbb{V}[X'U|X](X'X)^{-1} = (X'X)^{-1}\sigma^2$$

as $\mathbb{V}[X'U|X] = \mathbb{E}[X'UU'X|X] = X'\mathbb{E}[U'U|X]X = X'\sigma^2 I X = X'X\sigma^2$. **[2p]**

Yes, $\hat{\beta}_{OLS}$ is consistent:

$$\hat{\beta}_{OLS} = \beta + (X'X)^{-1}X'U = \beta + \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \frac{1}{n}\sum_i x_i u_i \xrightarrow{p} \beta$$

by Slutsky's theorem, whereby $\frac{1}{n}\sum_i x_i x_i' \xrightarrow{p} \mathbb{E}[x_i x_i']$ by WLLN, $\left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \xrightarrow{p} \mathbb{E}[x_i x_i']^{-1}$ by CMT and $\frac{1}{n}\sum_i x_i u_i \xrightarrow{p} \mathbb{E}[x_i u_i] = 0$ by WLLN. **[2p]**

(b) (4 points) Assuming $u_i|x_i \sim N(0, \sigma^2)$, derive the ML estimator $\hat{\beta}_{MLE}$. What are the reasons why one might want to make this assumption on the distribution of $u_i|x_i$, in general (for the linear regression model and beyond)?

Exercise 5.5: Solution

Solution: The likelihood is

$$\begin{aligned}\mathcal{L}(\beta|X,Y) = p(Y|\beta) &= \prod_i p(y_i|\beta) \\ &= \prod_i (2\pi\sigma^2) exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\} \\ &= (2\pi\sigma^2)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - x_i'\beta)^2\right\} .\end{aligned}$$

(σ^2 and X (and x_i) are dropped from the conditioning set for notational simplicity.) **[2p]** Hence, we get

$$\begin{aligned}\hat{\beta}_{MLE} = \arg\max_\beta \mathcal{L}(\beta|X,Y) &= \arg\max_\beta (2\pi\sigma^2)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right\} \\ &= \arg\min_\beta (Y - X\beta)'(Y - X\beta) \\ &= (X'X)^{-1}X'Y \\ &= \hat{\beta}_{OLS} \ . \textbf{[1p]}\end{aligned}$$

This assumption is made to enable likelihood-based testing and because some models can only be estimated using ML, not (O)LS. **[1p]**

(c) (8 points) Assuming in addition $\beta|\sigma^2 \sim N(0, \tau\sigma^2 I)$, derive the (conditional) posterior distribution $p(\beta|Y, \sigma^2)$. Define the Bayes estimator $\hat{\beta}_B$ to be the posterior mean. What is the (finite sample) distribution of $\hat{\beta}_B|X$ under the Bayesian paradigm? What is its distribution under the frequentist paradigm? Is $\hat{\beta}_B$ unbiased? Is it consistent? How could you choose τ and what does this choice signify?

Solution: First, we drive the posterior

$$\begin{aligned}p(\beta|Y) &\propto p(Y|\beta)p(\beta) \\ &\propto exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right\} exp\left\{-\frac{1}{2\tau\sigma^2}\beta'\beta\right\} \\ &\propto exp\left\{-\frac{1}{2\sigma^2}\left(\beta'\left(X'X + \frac{1}{\tau}I\right)\beta - 2\beta'X'Y\right)\right\} ,\end{aligned}$$

which let's us conclude that $\beta|Y \sim N(\mu, \sigma^2 V)$ with

$$V = \left[X'X + \frac{1}{\tau}I\right]^{-1} , \quad \mu = VX'Y \ . \textbf{[2p]}$$

Exercise 5.5: Solution

We define

$$\hat{\beta}_B = \mathbb{E}[\beta|Y] = \mu = \left[X'X + \frac{1}{\tau}I\right]^{-1} X'Y \ .$$

Under the Bayesian paradigm, $\hat{\beta}_B$ is a constant, it has no distribution, because it only depends on data (which is regarded as constant), not on the parameter β (which is regarded as random). **[1p]**

Under the frequentist paradigm, using similar arguments as above for $\hat{\beta}_{OLS}$, we have

$$\mathbb{E}[\hat{\beta}_B|X] = \left[X'X + \frac{1}{\tau}I\right]^{-1} [X'X\beta + X'\mathbb{E}[U|X]] \ ,$$

and

$$\mathbb{V}[\hat{\beta}_B|X] = \left[X'X + \frac{1}{\tau}I\right]^{-1} \mathbb{V}[X'U|X] \left[X'X + \frac{1}{\tau}I\right]^{-1} \ .$$

Note that even if $\mathbb{E}[U|X] = 0$, $\hat{\beta}_B$ is biased. However, for any $\tau < \infty$, it has a lower variance than $\hat{\beta}_{OLS} = \hat{\beta}_{MLE}$. **[2p]** Also, $\hat{\beta}_B$ is consistent as

$$\hat{\beta}_B = \beta - \left[\frac{1}{n}\sum_i x_i x_i' + \frac{1}{n}\frac{1}{\tau}I\right]^{-1} \left[\frac{1}{n}\frac{1}{\tau}I\right]\beta,$$

and this second-term (the bias) converges in probability to zero. **[1p]**

One can derive the MDD as a function of τ, $p(Y|\tau)$, and choose τ to maximize this quantity. This optimizes the bias-variance trade-off. The bias decreases in τ, but the variance increases in τ, whereby for $\tau \to \infty$, we get $\hat{\beta}_B = \hat{\beta}_{MLE}$ and for $\tau = 0$, we get $\hat{\beta}_B = 0$ with variance zero. **[2p]**

(d) (3 points) How would you estimate σ^2 under OLS vs. MLE vs. the Bayesian paradigm? You do not need to derive the estimators.

Solution: Under OLS, we can use the analogy principle to get

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (y_i - x_i'\hat{\beta}_{OLS})^2 \ .\textbf{[1p]}$$

Under MLE, we get the same result by maximizing the likelihood w.r.t. σ^2. **[1p]** Finally, under Bayesian inference, we specify a prior $p(\sigma^2)$ and given $p(Y|\sigma^2)$, obtained as part of the inference on $\beta|\sigma^2$, we can derive the marginal posterior $p(\sigma^2|Y) \propto p(Y|\sigma^2)p(\sigma^2)$. **[1p]**

6 Extremum Estimation

6.1 Asymptotics in Linear Regression Model via Extremum Estimation Theory

Remark: excercise builds on 4.2.

Suppose that you cannot find an analytical expression for $\hat{\beta}$. Use the simplified version of the extremum estimation theory discussed in class to analyze whether $\hat{\beta}$ is consistent and to find its asymptotic distribution. For simplicity, assume you know σ^2 (i.e. you estimate only β, conditioning on σ^2).

Exercise 6.1: Solution

Suppose that you cannot find an analytical expression for $\hat{\beta}$. Use the simplified version of the extremum estimation theory discussed in class to analyze whether $\hat{\beta}$ is consistent and to find its asymptotic distribution. For simplicity, assume you know σ^2 (i.e. you estimate only β, conditioning on σ^2).

Solution:

The extremum estimation theory shows asymptotic properties (consistency and asymptotic Normality) for an estimator defined as

$$\hat{\theta} = \arg\min_{\theta} \; Q_n(\theta|Z^n) \; ,$$

where Z^n denotes all the data. In our case, we have $\hat{\beta}_{ML} = \arg\min_{\beta} \; Q_n(\beta|Y^n, X^n)$ for

$$Q_n(\beta|Y^n, X^n) = -\frac{1}{n}\ell(\beta|X, Y) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \; . \quad \textbf{[1p]}$$

Replacing $\frac{1}{n}\sum\limits_{i=1}^{n}$ by $\mathbb{E}$, we can see that $Q_n(\beta|Y^n, X^n)$ converges in probability to $Q(\beta) = \mathbb{E}[-(y_i - x_i'\beta)^2]$.[1] Moreover, $Q(\beta)$ is clearly a continuous function. Finally, $Q(\beta)$ is uniquely minimised by β_0, because $\mathbb{E}[y_i|x_i] = x_i'\beta_0$ and we know that $\mathbb{E}[(y_i - f(x_i))^2]$ is uniquely

[1] The more precise calculation establishes uniform convergence in probability, i.e.

$$\mathbb{P}\left[\sup_{\beta}|Q_n(\beta|Y^n, X^n) - Q(\beta)| < \varepsilon\right] \to 1 \; , \quad \text{or} \;\; \sup_{\beta}|Q_n(\beta|Y^n, X^n) - Q(\beta)| \xrightarrow{p} 0 \; .$$

To show this, first note that

$$Q(\beta) = \mathbb{E}[(y_i - x_i'\beta)^2] = \mathbb{E}[y_i^2] - 2\mathbb{E}[y_i x_i'\beta] + \mathbb{E}[(x_i'\beta)^2] \; .$$

Using this, we can see that

$$\begin{aligned}&\sup_{\beta}\left|\frac{1}{n}\sum_i y_i^2 - \frac{2}{n}\sum_i y_i x_i'\beta + \frac{1}{n}\sum_i (x_i'\beta)^2 - \mathbb{E}[y_i^2] + 2\mathbb{E}[y_i x_i'\beta] - \mathbb{E}[(x_i'\beta)^2]\right| \\ &= \sup_{\beta}\left|\frac{1}{n}\sum_i y_i^2 - \mathbb{E}[y_i^2]\right| - \sup_{\beta}\left|\frac{2}{n}\sum_i y_i x_i'\beta - 2\mathbb{E}[y_i x_i'\beta]\right| + \sup_{\beta}\left|\frac{1}{n}\sum_i (x_i'\beta)^2 - \mathbb{E}[(x_i'\beta)^2]\right| \xrightarrow{p} 0\end{aligned}$$

by WLLN.

Exercise 6.1: Solution

minimised by $f(x_i) = \mathbb{E}[y_i|x_i]$.[2] This proves consistency. **[2p]**

To calculate the asymptotic distribution, we make use of the score,

$$Q_n^{(1)}(\beta, Y^n, X^n) = \frac{\partial Q_n}{\partial \beta} = \frac{2}{n}\sum_{i=1}^n x_i'(y_i - x_i'\beta) \quad ,$$

and the Hessian,

$$Q_n^{(2)}(\beta, Y^n, X^n) = \frac{\partial^2 Q_n}{\partial \beta \partial \beta'} = \frac{\partial Q_n^{(1)}}{\partial \beta'} = -\frac{2}{n}\sum_i x_i x_i' \; .$$

By CLT, the score function evaluated at β_0 and scaled by $\sqrt{n}$ converges in distribution to

$$\begin{aligned}\sqrt{n}Q_n^{(1)}(\beta_0, Y^n, X^n) &= \frac{2}{\sqrt{n}}\sum_i x_i' u_i \\ &= 2\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n x_i'(y_i - x_i'\beta) - \mathbb{E}[x_i'(y_i - x_i'\beta)]\right) \\ &\xrightarrow{d} N(0, M) \; ,\end{aligned}$$

where $M = 4\mathbb{E}[(x_i'u_i)(x_i'u_i)'] = 4\sigma^2\mathbb{E}[x_i x_i'] = 4\sigma^2 Q$, whereas – by WLLN – the Hessian converges in probability to

$$Q_n^{(2)}(\beta, Y^n, X^n) \xrightarrow{p} H = -2\mathbb{E}[x_i x_i'] = -2Q \; .$$

[1p each: score & Hessian] The extremum estimation theory tells us then that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, H^{-1}MH^{-1}) = N(0, \sigma^2 Q^{-1}) \; . \quad \textbf{[1p]}$$

[2] Here's an alternative way to show that, without making use of the stated result:

$$\begin{aligned}Q(\tilde{\beta}) = \mathbb{E}[(y_i - x_i'\tilde{\beta})^2] &= \mathbb{E}[(y_i - x_i'\beta_0 - x_i'(\tilde{\beta} - \beta_0))^2] \\ &= \mathbb{E}_X[\mathbb{E}[(y_i - x_i'\beta_0 - x_i'(\tilde{\beta} - \beta_0))^2|X]] \\ &= \mathbb{E}_X[\mathbb{E}[(u_i - x_i'(\tilde{\beta} - \beta_0))^2|X]] \\ &= \mathbb{E}_X[\mathbb{E}[u_i^2|X] + \mathbb{E}[(x_i'(\tilde{\beta} - \beta_0))^2|x_i]] \\ &= \sigma^2 + \mathbb{E}_X[\mathbb{E}[(\tilde{\beta} - \beta_0)'x_i x_i'(\tilde{\beta} - \beta_0)|X]] \\ &= \sigma^2 + (\tilde{\beta} - \beta_0)'\mathbb{E}[x_i x_i'](\tilde{\beta} - \beta_0) \\ &\geq \sigma^2 = Q(\beta_0) \; ,\end{aligned}$$

where we used the fact that $(x_i'(\tilde{\beta} - \beta_0))^2$ is a scalar and hence can be written as $(x_i'(\tilde{\beta} - \beta_0))'(x_i'(\tilde{\beta} - \beta_0))$ as well as the fact that $\mathbb{E}[x_i x_i']$ is positive-definite, which implies that $(\tilde{\beta} - \beta_0)'\mathbb{E}[x_i x_i'](\tilde{\beta} - \beta_0) \geq 0$ and is equal to zero only at $\tilde{\beta} = \beta_0$.

7 Further Topics in (Cross-Sectional) Econometrics

7.1 Finite Sample-Distribution in Linear Regression Model

Remark: excercise builds on 4.2.

Suppose that you cannot find the finite sample distribution of $\hat{\beta}$. Describe two approaches to approximate it.

Exercise 7.1: Solution

Suppose that you cannot find the finite sample distribution of $\hat{\beta}$. Describe two approaches to approximate it.

Solution:

Approach 1: Bootstrapping. **[1p]** Take M times n random draws with replacement from the original dataset (X, Y). This gives you M different samples of size n. For each of these M newly-generated samples, find $\hat{\beta}^{(m)}$. If your original sample is truly random and representative of the true population, then the sequence of the estimated $\left\{\hat{\beta}^{(m)}\right\}_{m=1}^{M}$ approximates the finite sample distribution of $\hat{\beta}$. **[1p]**

Approach 2: Approximate the finite sample distribution using the asymptotic distribution of $\hat{\beta}$. **[1p]** From $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$, we can argue that $\hat{\beta} \overset{approx.}{\sim} N\left(\beta, \frac{1}{n}\sigma^2 Q^{-1}\right)$. Replacing σ^2 and Q^{-1} with consistent estimators $\hat{\sigma}^2$ and $\hat{Q}^{-1} = \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1}$, we get

$$\hat{\beta} \overset{approx.}{\sim} N\left(\beta, \frac{1}{n}\hat{\sigma}^2 \hat{Q}^{-1}\right) .$$

This approach works well for n large. **[1p]**

8 Numerical Methods

8.1 Applied Probit Model I

Remark: dataset `dat_SalesCustomers.csv` required.

The dataset `dat_SalesCustomers.csv` contains data on sales of shopping malls in Instanbul. It includes the following variables: *invoice_no* (identifier of transaction/invoice), *customer_id* (identifier of customer), *category* (type of goods sold), *price* (in TRY, Turkish Lira), *invoice_date*, *shopping_mall*, *gender*, *age* and *payment_method* (cash- vs. credit-card- vs. debit-card-payment).

You are interested in shedding light on the determinants of cash- vs card-payment. For this purpose, you set up a probit model:

$$y_i^* = x_i'\beta + u_i \quad , \quad u_i|x_i \sim N(0,1) \ , \tag{8.1}$$

whereby we observe $y_i = \mathbf{1}\{y_i^* \geqslant 0\}$, a dummy variable for cash payment. Recall that the Maximum Likelihood (ML) estimator for β solves

$$\hat{\beta} = \arg\min_{\beta} Q_n(\beta; Z_n) \quad \text{for} \quad Q_n(\beta; Z_n) = -\frac{1}{n}\ell(\beta; Z_n) \ , \tag{8.2}$$

where

$$\ell(\beta; Z_n) = \sum_{i=1}^{n} y_i \log(\Phi(x_i'\beta)) + (1 - y_i) \log(\Phi(-x_i'\beta))$$

is the log-likelihood and $Z_n = \{y_i, x_i\}_{i=1}^n$ comprises all of the data you have available (outcome-variables and covariates for the n observations in your sample).

(a) Are there missing values in your data? Delete all observations with a missing value in the variables *category*, *price*, *gender*, *age* or *payment_method.* How many observations do you have left?

(b) Based on the variable *payment_method*, generate a dummy variable for cash payment and call it *paid_in_cash.* Also, based on *gender*, create a dummy for males, *male.* What fraction of transactions were carried out in cash? What fraction of the overall sales (in TRY) were carried out in cash?

To decrease computational costs, consider only the first $n = 1000$ observations for the following questions.

(c) Based on the variable *category*, create a dummy for each of the following four categories: i) clothes and shoes, ii) cosmetics, iii) food, iv) technology. In this way, we divide the categories into five groups, whereby the fifth is made up by the rest, i.e. goods that do not belong to either of the four categories. How are the transactions split across these five categories? How are the sales split across these five categories?

(d) Taking *paid_in_cash* as your outcome variable y_i and *price*, *male*, *age* and the four category-dummies as your covariates x_i, use a numerical optimization-command from the software of your choice to solve the optimization problem in Eq. (8.2) and obtain $\hat{\beta}$ for your sample.[1] If manual optimization does not work, you can use a pre-programmed command to estimate the probit model.
Hint: instead of computing first $\Phi(x)$ using a software-command for the cdf of a $N(0,1)$ RV (`pnorm(x)` in R) and then taking logs, it's better to directly use a software-command for the log of the cdf of a $N(0,1)$ RV (`pnorm(x,log.p=TRUE)` in R). This way, you avoid having to compute the log of a number very close to zero, which can result in `-Inf`.[2]
Hint: to ensure convergence, you might want to supply a the gradient of your objective function.

(e) Based on your estimate, compute the effect of age doubling on the expected probability of using cash for a 30 year-old male who bought clothes/shoes for 500 TRY, i.e. for an observation with $x_i = x_i^* \equiv [1, 500, 1, 30, 1, 0, 0, 0]$. Put differently, this is the difference in expected probabilities of cash payment between a 60 year-old and a 30 year-old male who bought clothes/shoes for 500 TRY. We will call this quantity $\gamma_1(\hat{\beta})$. Also, compute the same effect without conditioning on the category of goods sold in two steps: (i) compute the effect for each of the five categories and (ii) take a weighted average of them, with weights given by the proportions of these goods-categories in overall sales (see your answer to (c)). We will call this quantity $\gamma_2(\hat{\beta})$.

(f) Suppose that your probit model in Eq. (8.1) is correctly specified. Is your estimator $\hat{\beta}$ consistent? Use the simplified version of the extremum estimation theory we discussed in class to answer this question.

[1] As part of your derivations for exercise (f), you have to find the score and the Hessian of the objective function in Eq. (8.2),

$$s_n(\beta) \equiv Q_n^{(1)}(\beta; Z_n) \equiv \frac{\partial Q_n(\beta; Z_n)}{\partial \beta} \quad \text{and} \quad H_n(\beta) \equiv Q_n^{(2)}(\beta; Z_n) \equiv \frac{\partial^2 Q_n(\beta; Z_n)}{\partial \beta \partial \beta'} = \frac{\partial s_n(\beta)}{\partial \beta'} .$$

You can also use them to construct your own numerical optimization algorithm to find $\hat{\beta}$.

[2] The alternative is to do manual adjustments, coding -Inf as a very large negative number, but this can be imprecise.

(g) Use bootstrapping to find a numerical approximation of the finite sample distribution of $\hat{\beta}$ as well as the two marginal effects $\gamma_1(\hat{\beta})$ and $\gamma_2(\hat{\beta})$: draw $M = 100$ different samples of n observations with replacement from your dataset and compute (numerically) $\hat{\beta}$, $\gamma_1(\hat{\beta})$ and $\gamma_2(\hat{\beta})$ for each of them. Plot the finite sample distributions you obtained (regarding $\hat{\beta}$, you can limit yourself to the coefficient on *age*).

(h) Another approach to approximate the finite sample distribution of $\hat{\beta}$ and functions of it like the marginal effects is to use their asymptotic distribution. Use the simplified version of the extremum estimation theory we discussed in class to show that the asymptotic distribution of $\hat{\beta}$ is given by

$$\sqrt{n}(\hat{\beta} - \beta_0) \overset{d}{\to} N\left(0, H^{-1}\right) \quad \text{with} \quad H = \mathbb{E}\left[\frac{\phi(x_i'\beta_0)^2}{\Phi(x_i'\beta_0)\Phi(-x_i'\beta_0)} x_i x_i'\right] . \tag{8.3}$$

Then, use the asymptotic distribution in Eq. (8.3) to approximate the finite sample distribution of $\hat{\beta}$ in your sample. Plot the approximate finite sample distribution of the estimated coefficient on *age* using a histogram. How does it compare to the one obtained via bootstrapping?
Hint: The numerator and the denominator in the fraction that appears in H are often both very close to zero. Rather than computing it as-is, first compute the log of it and then take the exponential, i.e. compute

$$\frac{\phi(x_i'\beta_0)^2}{\Phi(x_i'\beta_0)\Phi(-x_i'\beta_0)} \quad as \quad exp\left\{2log\phi(x_i'\beta_0) - log\Phi(x_i'\beta_0) - log\Phi(-x_i'\beta_0)\right\} .$$

To compute log $\phi(x)$ and log $\Phi(x)$, as before in exercise (b), it's better practice to use the log-pdf/cdf software-commands than to compute first the pdf/cdf and then take logs manually (i.e. in R, use `dnorm(x,log=TRUE)` and `pnorm(x,log.p=TRUE)`).

(i) Use the asymptotic distribution of $\hat{\beta}$ from Eq. (8.3) and the Delta method to find the asymptotic distribution of $\gamma_1(\hat{\beta})$. Then, use it to approximate the finite sample distribution of $\gamma_1(\hat{\beta})$ in your sample. How does this approximate finite sample distribution compare to the one obtained via bootstrapping?

(j) Now let's test whether the true partial effect $\gamma_1(\hat{\beta})$ is significantly different from 0 at the $\alpha = 0.05$ level:

$$\mathscr{H}_0 : \gamma_1(\beta) = 0 \quad \text{vs.} \quad \mathscr{H}_1 : \gamma_1(\beta) \neq 0 .$$

(In other words, we are testing whether the expected probabilities of cash payment for a 30 year-old and a 60 year-old male buying clothes for 500 TRY are different.) One approach to do so uses the finite sample distribution of $\gamma_1(\hat{\beta})$ approximated via its asymptotic distribution, which you found in the exercise before:

$$\gamma_1(\hat{\beta}) \overset{approx.}{\sim} N\left(\gamma_1(\beta), \frac{1}{n}\hat{V}\right) ,$$

for some $\hat{V}$ you had to find. Use this expression to construct a t-test. What do you conclude? Also, use the above expression to construct a 95% confidence interval for $\gamma_1(\beta)$.[3] (If you couldn't find $\hat{V}$, just state the test statistic and critical value for a general $\hat{V}$.)

8.2 Frequentist and Bayesian Inference for General Likelihood Functions

Suppose you have a complicated model which yields the likelihood function $p(Z_n|\theta)$, where θ denotes the parameter you would like to estimate and Z_n denotes an $n-$dimensional object containing all your data. You are able to evaluate $p(Z_n|\theta)$, but not to manipulate it analytically.

(a) How can you find the MLE $\hat{\theta}$? How would you analyze its asymptotic properties? How can you conduct inference on some function of θ, $g(\theta)$? Can you do that for arbitrary functions g?

(b) Suppose your prior is $\theta \sim N(0, \lambda^{-1}I)$. How would you find the posterior $p(\theta|Y)$ and conduct inference on $g(\theta)$? Can you do that for arbitrary functions g?

8.3 Frequentist and Bayesian Inference: Numerical Computations

Consider the linear regression model

$$y_i = x_i'\beta + u_i \ ,$$

with $u_i|x_i \sim N(0, \sigma_i^2)$. Suppose you observe n i.i.d. observations: $\{(y_i, x_i)\}_{i=1:n}$.

(a) Derive the likelihood and define the ML estimator for $\theta = (\beta, \{\sigma_i^2\}_{i=1}^n)$. Taking first-order conditions, derive the conditional estimators $\hat{\beta}|\{\sigma_i^2\}_{i=1}^n$ and $\hat{\sigma}_i^2|\ \beta$. How could you obtain the joint estimator based on these conditional estimators? How could you obtain the joint estimator without relying on these conditional estimators?

(b) Suppose that you successfully obtained a numerical value for the joint ML estimator. How can you analyze the asymptotic properties of your estimator? Why do you care about these asymptotic properties?

(c) Suppose you took priors for β and $\{\sigma_i^2\}_{i=1}^n$ and derived analytically the conditional posteriors $p(\beta|Y, \{\sigma_i^2\}_{i=1}^n)$ and $p(\sigma_i^2|Y, \beta)$. How can you obtain the joint posterior $p(\beta, \{\sigma_i^2\}_{i=1}^n|Y)$ based on these conditionals?

(d) Let β_3 be the third-element in β. Once you found the posterior $p(\beta, \{\sigma_i^2\}_{i=1}^n|Y)$, how can you find the posterior of β_3/σ_4^2?

[3] Note that in general, we would use the Wald-test. Here we can use the t-test because we are testing a single thing, i.e. our testing function $g(\beta) = \gamma_1(\beta) = 0$ is a scalar. Our t-test will give the same result as the Wald test, because the asymptotic distribution of the Wald-test-statistic is derived in the same way as that of our t-test statistic here (i.e. it also uses the Delta method), except that it squares things in the end to go from a Normal to a Chi-Squared distribution.

Exercise 7.1: Solution

Suppose that you cannot find the finite sample distribution of $\hat{\beta}$. Describe two approaches to approximate it.

Solution:

Approach 1: Bootstrapping. [**1p**] Take M times n random draws with replacement from the original dataset (X, Y). This gives you M different samples of size n. For each of these M newly-generated samples, find $\hat{\beta}^{(m)}$. If your original sample is truly random and representative of the true population, then the sequence of the estimated $\left\{\hat{\beta}^{(m)}\right\}_{m=1}^{M}$ approximates the finite sample distribution of $\hat{\beta}$. [**1p**]

Approach 2: Approximate the finite sample distribution using the asymptotic distribution of $\hat{\beta}$. [**1p**] From $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$, we can argue that $\hat{\beta} \overset{approx.}{\sim} N\left(\beta, \frac{1}{n}\sigma^2 Q^{-1}\right)$. Replacing σ^2 and Q^{-1} with consistent estimators $\hat{\sigma}^2$ and $\hat{Q}^{-1} = \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1}$, we get

$$\hat{\beta} \overset{approx.}{\sim} N\left(\beta, \frac{1}{n}\hat{\sigma}^2 \hat{Q}^{-1}\right) .$$

This approach works well for n large. [**1p**]

Exercise 8.2: Solution

Suppose you have a complicated model which yields the likelihood function $p(Z_n|\theta)$, where θ denotes the parameter you would like to estimate and Z_n denotes an $n-$dimensional object containing all your data. You are able to evaluate $p(Z_n|\theta)$, but not to manipulate it analytically.

(a) (4 points) How can you find the MLE $\hat{\theta}$? How would you analyze its asymptotic properties? How can you conduct inference on some function of θ, $g(\theta)$? Can you do that for arbitrary functions g?

Solution: We can obtain $\hat{\theta}$ by maximizing $p(Z_n|\theta)$ numerically w.r.t. θ. We can analyze its asymptotic properties using extremum estimation theory. Finally, to conduct inference on $g(\theta)$, given $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$, we'd need to find the (asymptotic) distribution of $\sqrt{n}(g(\hat{\theta}) - g(\theta_0))$. We can do that for rather simple functions, for which we can either find the distribution manually or invoke a theorem like the Delta method.

(b) (3 points) Suppose your prior is $\theta \sim N(0, \lambda^{-1} I)$. How would you find the posterior $p(\theta|Y)$ and conduct inference on $g(\theta)$? Can you do that for arbitrary functions g?

Solution: We can use a numerical posterior sampling algorithm, for which we only need to evaluate $p(Z_n|\theta)$ and $p(\theta)$. This gives draws $\{\theta^i\}_{i=1}^M$, i.e. a numerical approximation of the posterior $p(\theta|Z_n)$, based on which we can get $\{g(\theta^i)\}_{i=1}^M$, i.e. a numerical approximation of the posterior $p(g(\theta)|Z_n)$. This works for arbitrary functions (as long as we can compute $g(\theta)$).

Exercise 8.3: Solution

Consider the linear regression model

$$y_i = x_i'\beta + u_i \ ,$$

with $u_i|x_i \sim N(0, \sigma_i^2)$. Suppose you observe n i.i.d. observations: $\{(y_i, x_i)\}_{i=1:n}$.

(a) (6 points) Derive the likelihood and define the ML estimator for $\theta = (\beta, \{\sigma_i^2\}_{i=1}^n)$. Taking first-order conditions, derive the conditional estimators $\hat{\beta}|\{\sigma_i^2\}_{i=1}^n$ and $\hat{\sigma}_i^2|\ \beta$. How could you obtain the joint estimator based on these conditional estimators? How could you obtain the joint estimator without relying on these conditional estimators?

Solution: Using the fact that observations are i.i.d., the likelihood is

$$\begin{aligned} p(Y|\beta, \{\sigma_i^2\}_{i=1}^n) &= \prod_{i=1}^n p(y_i|x_i, \beta, \sigma_i^2) \\ &= \prod_{i=1}^n (2\pi\sigma_i^2)^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma_i^2}(y_i - x_i'\beta)^2\right\} \\ &= (2\pi)^{-\frac{1}{2}} \left[\prod_{i=1}^n (\sigma_i^2)^{-\frac{1}{2}}\right] exp\left\{-\frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - x_i'\beta}{\sigma_i}\right)^2\right\} \\ &= (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta)\right\} \ , \end{aligned}$$

where $\Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$. To ease notation, we can write Σ for $\{\sigma_i^2\}_{i=1}^n$. Note that the exercise can also be solved without defining Σ. **[1.5p]** The ML estimator is defined as

$$(\hat{\beta}, \hat{\Sigma}) = \arg\max_{\beta,\Sigma} p(Y|\beta, \Sigma) \ .\textbf{[0.5p]}$$

Taking FOCs, we get the conditional estimators

$$\hat{\beta}|\Sigma = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y = \left(\sum_{i=1}^n x_i x_i'/\sigma_i^2\right)^{-1} \left(\sum_{i=1}^n x_i y_i/\sigma_i^2\right) \ , \textbf{[1p]}$$

and

$$\hat{\sigma}_i^2|\beta = (y_i - x_i'\beta)^2 \ .\textbf{[1p]}$$

Note that we are essentially running a regression $\tilde{y}_i = \tilde{x}_i\beta + \tilde{u}_i$ with the adjusted variables $\tilde{y}_i = y_i/\sigma_i$ and $\tilde{x}_i = x_i/\sigma_i$, and with $\tilde{u}_i = u_i/\sigma_i \sim N(0, 1)$.

The joint estimator $(\hat{\beta}, \hat{\Sigma})$ can be obtained by iterating on these two conditional estimators using the procedure by Meng & Rubin (1993). Given an initial value for, say, Σ (e.g. the identity matrix), one would compute $\hat{\beta}$, based on which one would compute

Exercise 8.3: Solution

$\hat{\Sigma}$, etc., until convergence. **[1p]**

Alternatively, without relying on these conditional estimators, we can maximize the likelihood above numerically w.r.t. β, Σ, e.g. using a Newton-type algorithm. **[1p]**

(b) (4 points) Suppose that you successfully obtained a numerical value for the joint ML estimator. How can you analyze the asymptotic properties of your estimator? Why do you care about these asymptotic properties?

Solution: Asymptotic properties of the ML estimator – even if not available analytically but only numerically – can be analyzed using the extremum estimation theory. By verifying a set of conditions we can establish consistency and asymptotic Normality.**[2p]** We care about the asymptotic properties because we want our estimator to be consistent and because we may want to use its asymptotic distribution to approximate its finite sample distribution. The latter is relevant if we want to conduct inference, i.e. test hypotheses, construct confidence intervals, and generally obtain a measure of uncertainty surrounding our estimates of θ and functions of θ. **[2p]**

(c) (4 points) Suppose you took priors for β and $\{\sigma_i^2\}_{i=1}^n$ and derived analytically the conditional posteriors $p(\beta|Y, \{\sigma_i^2\}_{i=1}^n)$ and $p(\sigma_i^2|Y, \beta)$. How can you obtain the joint posterior $p(\beta, \{\sigma_i^2\}_{i=1}^n|Y)$ based on these conditionals?

Solution: We can obtain the joint posterior by iteratively drawing from the conditionals using Gibbs sampling. **[2p]** This is analogous to the procedure by Meng & Rubin (1993) outlined above. Given an initial value for, say, Σ (e.g. the identity matrix), we would draw β from its conditional posterior, based on which we would draw Σ from the conditional posterior for Σ, etc. Doing so, we obtain a set of draws $\{\beta^m, \Sigma^m\}_{m=1}^M$, which in the limit as $M \to \infty$ approximates the joint posterior. We stop the procedure when the moments of our draws have stabilized and we only consider the last, say, 1000 draws, as the first iterations are influenced by our arbitrary initialization. **[2p]**

(d) (2 points) Let β_3 be the third-element in β. Once you found the posterior $p(\beta, \{\sigma_i^2\}_{i=1}^n|Y)$, how can you find the posterior of β_3/σ_4^2?

Solution: Given a set of draws $\{\beta^m, \Sigma^m\}_{m=1}^M$ that approximate the joint posterior $p(\beta, \{\sigma_i^2\}_{i=1}^n|Y)$, we can compute $\beta_3^m/\sigma_4^{2,m}$ for each draw m. This set of draws, $\{\beta_3^m/\sigma_4^{2,m}\}_{m=1}^M$, approximates the posterior $p(\beta_3/\sigma_4^2|Y)$. **[2p]**

9 Univariate Time Series Analysis

9.1 Time Series Basics I

(a) (3 points) Define weak stationarity (WS).

(b) (3 points) Why do we need WS in our work with time series data?

(c) (2 points) Are daily returns of the S&P 500 stock index WS? (see plot below)

(d) (2 points) Are weekly hours worked in the US economy WS? (see plot below)

(e) (2 points) Is female labor force participation in the US economy WS? (see plot below)

9.2 Time Series Basics II

This excercise is inspired by McConnell, Margaret and Gabriel Perez-Quiros (2000): "Output Fluctuations in the United States: What has changed since the early 1980's?" *American Economic Review*, 90(5), 1464-76.

Download some aggregate time series in quarterly frequency from FRED: Real GDP (GDPC1), Real Personal Consumption Expenditure (PCECC96), Real Gross Private Domestic Investment (GPDIC1).

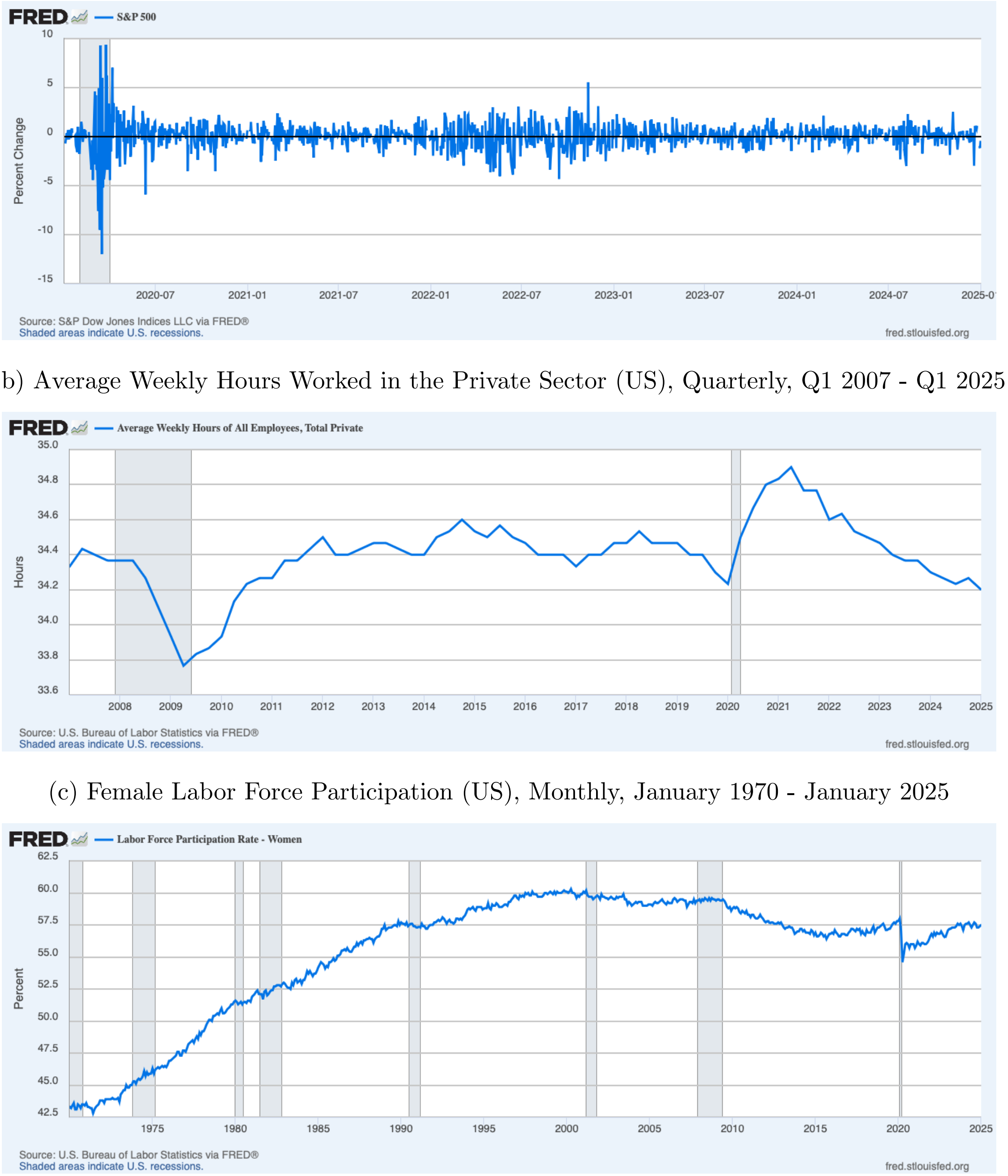
(a) Take logs of GDP, Consumption, and Investment. Plot the three series (you can generate the plots directly in FRED). What are the most striking features?

(b) For each series (in logs), estimate a model of the form

$$y_t = \beta_1 + \beta_2 t + u_t$$

using OLS, based on the samples 1965:Q1 to 2006:Q4, 2007:Q1 to 2019:Q4, and 2007:Q1 to 2022:Q2.

(c) According to your estimates, what are the annualized average growth rates (in percent) of GDP, consumption, and investment? Are these series growing, approximately, at the same



(a) S&P 500 Returns, Daily, 1 January 2020 - 1 January 2025

Source: S&P Dow Jones Indices LLC via FRED®
Shaded areas indicate U.S. recessions.

(b) Average Weekly Hours Worked in the Private Sector (US), Quarterly, Q1 2007 - Q1 2025

Source: U.S. Bureau of Labor Statistics via FRED®
Shaded areas indicate U.S. recessions.

(c) Female Labor Force Participation (US), Monthly, January 1970 - January 2025

Source: U.S. Bureau of Labor Statistics via FRED®
Shaded areas indicate U.S. recessions.

rate?

(d) For each subsample, compute sample autocorrelation functions for the deviations of output, consumption, and investment (the $\hat{u}_t$'s) from their estimated deterministic trend.

(e) Now compute quarter-on-quarter growth rates of these three series as $\ln y_t - \ln y_{t-1}$, plot the growth rates. What are the most striking features?

(f) Compute the sample means of the growth rates for each of the above subsamples. Compare the growth-rate results to (ii). Also compute the sample standard deviations for the growth rates.

(g) Repeat the analysis in (vi) for the subsamples “before 1984”, “between 1984 and 2006”. Did the means and the volatility of the series change?

9.3 Basic Calculations under Time Series Models: MA(3)

Consider the MA(3) process

$$y_t = (1 - 2.4L + 0.8L^2 - 0.4L^3)u_t \ ,$$

where L denotes the lag operator and

$$\mathbb{E}[u_t u_\tau] = \begin{cases} 1 & \text{if } t = \tau \\ 0 & \text{otherwise} \end{cases} .$$

(a) Define Weak and Strict Stationarity. Is the process y_t weakly stationary? Is it strictly stationary? You may impose additional conditions on the u_t's.

(b) Calculate the autocovariance function of y_t.

(c) Calculate $\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_t\right]$.

(d) Suppose $\{u_t\}$ is i.i.d. with $\mathbb{E}[u_t] = 0$ and $\mathbb{V}[u_t] = \sigma^2$, and define $x_t = u_t u_{t-4}$. Is the process x_t strictly stationary? Is x_t ergodic? Is x_t a White Noise process?

(e) Calculate $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x_t\right]$ and $\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right]$.

9.4 Asymptotics under Independent vs. Dependent Processes

Consider the following processes:

$$\begin{aligned} &x_t \overset{i.i.d.}{\sim} N(0, 4) \ , \\ &y_t = \phi y_{t-1} + u_t \ , \quad u_t \overset{i.i.d.}{\sim} N(0, 4) \ . \end{aligned}$$

Let $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$.

(a) (2 points) What is the probability limit of $\bar{x}$?

(b) (5 points) Under $\phi = 0.8$, what is the probability limit of $\bar{y}$? How does your answer change for $\phi = 0.9$? And for $\phi = 1$?

(c) (3 points) What is the asymptotic distribution of $\bar{x}$, i.e. what is the limit distribution of $\sqrt{T}(x_t - \mathbb{E}[x_t])$?

(d) (6 points) Under $\phi = 0.8$, what is the asymptotic distribution of $\bar{y}$, i.e. what is the limit distribution of $\sqrt{T}(y_t - \mathbb{E}[y_t])$? How does your answer change for $\phi = 0.9$? And for $\phi = 1$?

9.5 Frequentist Inference in AR(1): Basic Calculations & Misspecified Model I

Consider the AR(1) model

$$y_t = \phi y_{t-1} + u_t \ , \quad u_t \overset{i.i.d.}{\sim} N(0,1) \ \ ,$$

where $|\phi| < 1$.

(a) What is the difference between a conditional and unconditional likelihood function? Derive them for the present AR(1) model.

(b) Derive the conditional ML estimator of ϕ, denoted by $\hat{\phi}$, using the conditional likelihood function.

(c) Show that $\hat{\phi}$ is a consistent estimator of ϕ.

(d) Derive the limit distribution of $\hat{\phi}$.

(e) Suppose data is generated from a stationary AR(2) model

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t \ \ , \quad \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2) \ ,$$

but the econometrician estimates the above AR(1) model. What are the probability limits of $\hat{\phi}$ and $\hat{\sigma}_u^2$ in terms of $(\rho_1, \rho_2, \sigma_\epsilon^2)$?
Hint: note that this requires you to find the variance and first-order covariance of y_t under the AR(2) process.

9.6 Frequentist Inference in AR(1): Basic Calculations & Misspecified Model II

Suppose you model Swiss CPI inflation, y_t, as an AR(1) process:

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t \ , \quad u_t \sim WN(0, \sigma^2) \ ,$$

where $|\phi_1| < 1$.

(a) (3 points) What does $u_t \sim WN(0, \sigma^2)$ mean? What does it tell you about the properties of u_t?

(b) (2 points) What does the parameter ϕ_1 signify? No derivations are needed.

(c) (2 points) What does the parameter ϕ_0 signify? Suppose that Swiss inflation averages to around 0.3 percent over the sample that you consider. Assuming that $\phi_1 = 0.8$, where do you expect ϕ_0 to lie?

(d) (3 points) Let us fix $\phi_0 = 0$. Derive the OLS estimator for ϕ_1, $\hat{\phi}_1$.

(e) (4 points) Assuming that your model is correct – i.e. y_t is actually determined as

$$y_t = \phi_1 y_{t-1} + u_t \ , \quad \text{with} \quad u_t \sim WN(0, \sigma^2)$$

–, what is the probability limit of $\hat{\phi}_1$? State clearly any assumptions you need to make to find this limit.

(f) (6 points) Instead, suppose now that you misspecified your model and inflation actually follows an MA(1) process:

$$y_t = (1 + \theta L)\varepsilon_t \quad \text{for} \quad \varepsilon_t \sim WN(0, \sigma^2) \ .$$

Derive the probability limit of $\hat{\phi}_1$ in terms of θ. Again, state clearly any assumptions you need to make to find this limit.
Hint: you will need to compute moments of the MA(1) process. Note that $y_t = (1 + \theta L)\varepsilon_t$ holds for every period t, and so $y_{t-1} = (1 + \theta L)\varepsilon_{t-1}$.

9.7 Frequentist Inference in AR(1): Near-Unit-Roots

Consider the AR(1) model

$$y_t = \phi y_{t-1} + u_t \ , \quad u_t \overset{i.i.d.}{\sim} N(0, 1) \quad , \quad y_0 = 0 \quad .$$

(a) Take $\phi = 0.8$, and generate $m = 1, \ldots, M = 1,000$ samples of length $T = 80$. For each sample m compute the OLS/ML estimator of ϕ, denoted by $\hat{\phi}^m$. This gives you a set of $M = 1000$ estimators, $\{\hat{\phi}^m\}_{m=1}^M$.

(b) The asymptotic distribution suggests that

$$\sqrt{T}(\hat{\phi} - \phi) \overset{approx.}{\sim} N(0, \hat{\gamma}_0^{-1}) \ , \quad \gamma_0 = \mathbb{V}[y_t] = \mathbb{V}[y_{t-1}] \ ,$$

and hence, for the t-statistic, it should hold that

$$\frac{\sqrt{T}(\hat{\phi} - \phi)}{1/\sqrt{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}} \overset{approx.}{\sim} N(0, 1) \ .$$

Generate the actual finite sample distribution of this t-statistic, i.e. compute the t-statistic for each of your estimators $\{\hat{\phi}^m\}_{m=1}^M$, and plot a histogram. Is it close to a standard Normal distribution? Overlay the pdf of a $N(0, 1)$ on your histogram to see.

(c) Repeat the previous two exercises for the parameter values $\phi \in \Phi = \{0.2, 0.5, 0.9, 0.98, 0.999\}$. Discuss your results in view of the asymptotic calculations for $|\phi| < 1$ vs. $\phi = 1$. What do the simulations say about the accuracy of the asymptotic distribution as an approximation of the finite-sample distribution?

9.8 Frequentist Inference in AR(1): Assumptions & Model-Selection

Consider the AR(1) model

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t \ , \quad u_t \overset{iid}{\sim} N(0, \sigma^2) \ .$$

(a) Derive the conditional likelihood function $p(Y_{1:T}|y_0, \phi_0, \phi_1, \sigma^2)$ and the MLE $(\hat{\phi}_0, \hat{\phi}_1, \hat{\sigma}^2)$.

(b) What assumptions do you need to establish consistency and find the asymptotic distribution of the MLE? Derive the asymptotic distribution of $\hat{\phi}_1|\sigma^2$ (i.e. assuming σ^2 is a constant you know). Is it a good approximation of its finite sample distribution?

(c) Suppose you are unsure whether the true model contains one or two lags of y_t, i.e. whether y_t follows an AR(1) or AR(2). Describe two possible ways to choose among the two models (under the frequentist paradigm).

9.9 Frequentist Inference in Time Series Regression

Consider the time series regresssion

$$y_t = x_t'\beta + u_t \ , \quad \text{or} \quad Y = X\beta + U \ ,$$

where x_t and β are k-dimensional vectors, Y and U are n-dimensional vectors, and X is $n \times k$. Let $\mathbb{E}[u_t x_t] = 0$ and $\mathbb{V}[u_t] = \sigma_u^2$.

(a) Would it be reasonable to assume that u_t is i.i.d.? Why (not)?

(b) Derive the OLS estimator for β, $\hat{\beta}_{OLS}$.

(c) Is $\hat{\beta}_{OLS}$ consistent? What properties do the processes y_t, x_t and u_t need to satisfy so that you can analyze consistency? Define these properties.

(d) Does the question whether or not u_t is i.i.d. affect the asymptotic variance of $\hat{\beta}_{OLS}$? Explain.

9.10 Frequentist Inference in Time Series Regression with Autocorrelated Errors

Remark: excercise continued in 9.11.

Consider the time series regresssion

$$y_t = x_t'\beta + u_t \ , \quad \text{or} \quad Y = X\beta + U \ , \tag{9.1}$$

where x_t and β are k-dimensional vectors, Y and U are n-dimensional vectors, and X is $n \times k$. In addition, suppose that u_t follows an AR(1) process:

$$u_t = \rho u_{t-1} + \varepsilon_t \ , \quad \text{with} \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_e^2) \quad \text{and} \quad |\rho| < 1 \ . \tag{9.2}$$

(a) Define weak stationarity. Is ε_t weakly stationary? Is u_t weakly stationary?

(b) Compute the variance and first-order autocovariance of u_t.

(c) Derive an equation for y_t in terms of y_{t-1}, x_t and x_{t-1} and the i.i.d. error term ε_t (rather than the autocorrelated error term u_t).
Hint: solve the time series regression equation (9.1) for u_t and u_{t-1}, plug these expressions into the AR(1) equation (9.2), and solve for y_t.

(d) Let $\theta = (\beta', \rho, \sigma_e^2)'$. Derive the likelihood $p(Y|X, y_0, \theta)$ associated with the model you derived in exercise (c).
Hint: recall how the likelihood of an AR(1) model $z_t = \phi z_{t-1} + \varepsilon_t$ is derived as the product of conditionals: $p(Z|z_0, \phi, \sigma_e^2) = \prod_{t=1}^{T} p(z_t|z_{t-1}, \phi, \sigma_e^2)$. Here, the analogous applies for y_t. Also, note that by conditioning on X we treat $x_t \ \forall \ t$ as given.

9.11 Bayesian Inference in Time Series Regression with Autocorrelated Errors

Remark: excercise builds on 9.10.

Suppose you know σ_e^2 (you may set $\sigma_e^2 = 1$ if you want), and your prior for $(\beta', \rho)'$ is

$$p(\beta, \rho) = p(\beta)p(\rho) \ , \quad \text{with} \quad \beta \sim N(0, \sigma_e^2 \lambda^{-1} I) \quad \text{and} \quad \rho \sim N(0, \sigma_e^2 \tau^{-1}) \ .$$

To simplify notation, write your above likelihood $p(Y|X, y_0, \theta)$ as $p(Y|\beta, \rho)$.

(e) Suppose no further analytical calculations are possible; you have the likelihood $p(Y|\beta, \rho)$ and prior $p(\beta, \rho)$, but nothing more. How could you obtain the joint posterior $p(\beta, \rho|Y)$[1] numerically? Describe the procedure you would use.

(f) Under the prior $\beta \sim N(0, \sigma_e^2 \lambda^{-1} I)$, derive the conditional posterior of $\beta|\rho$. Given this conditional posterior, what do you expect the conditional posterior of $\rho|\beta$ to be? You do not need to do any derivations for $\rho|\beta$.
Hint: to obtain $p(\beta|Y, \rho)$, write the model as $\tilde{y}_t = \tilde{x}_t \beta + \varepsilon_t$, for some $\tilde{y}_t$ and $\tilde{x}_t$.

(g) Given the conditional posteriors $p(\beta|Y, \rho)$ and $p(\rho|Y, \beta)$, how could you obtain the joint posterior $p(\beta, \rho|Y)$? Describe the procedure you would use.

[1] To be precise, this posterior is also conditional on y_0 and X, just like the likelihood. Same for the conditional posteriors mentioned subsequently.

Exercise 9.1: Solution

(a) Define weak stationarity (WS).

Solution: A stochastic process y_t is WS if its first and second moments are independent of time, i.e. if its mean $\mathbb{E}[y_t] \equiv \mu_t = \mu \ \forall \ t$ and its autocovariances $Cov(y_t, y_{t-h}) \equiv \gamma_{t,h} = \gamma_h \ \forall \ t$.

(b) Why do we need WS in our work with time series data?

Solution: On a rather intuitive and somewhat abstract level, we need WS (and ergodicitiy) as a replacement for the i.i.d. assumption from cross-sectional inference. Time series data is clearly not i.i.d. over time, but with WS we ensure that at least the first and second moments are constant over time, which means that the dependence of data is stable over time.

More concretely, we need WS (together with ergodicitiy) to apply the LLN and CLT for time series data. (In fact, we need strict stationarity for them, which subsumes weak stationarity).

(c) Are daily returns of the S&P 500 stock index WS? (see plot below)

Solution: No. You can see that their variance is not constant over time. For instance, they exhibit a lot of volatility at the beginning of the sample around 2020, then their variance decreases in 2021, and it increases again in 2022.

(d) Are weekly hours worked in the US economy WS? (see plot below)

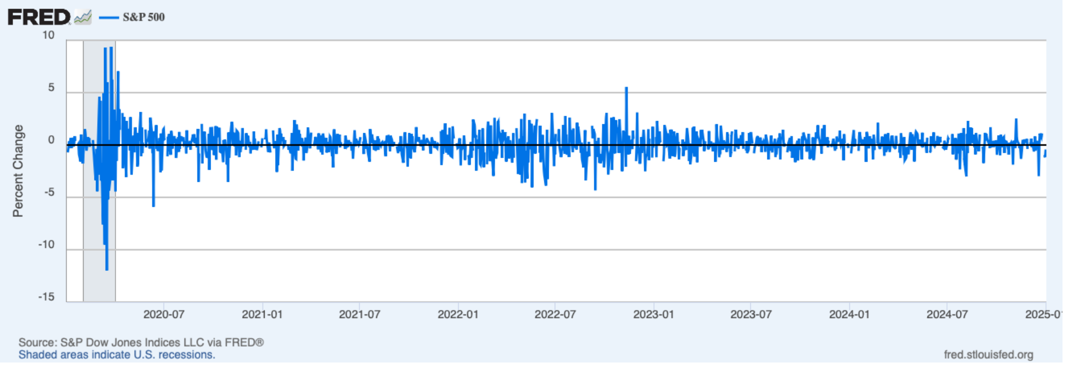
Solution: Yes, they likely are, at least judging based on this plot; both the mean and the variance seem to be stable over time (and there is no graphical reason to believe that the autocovariances are changing over time either).

(e) Is female labor force participation in the US economy WS? (see plot below)

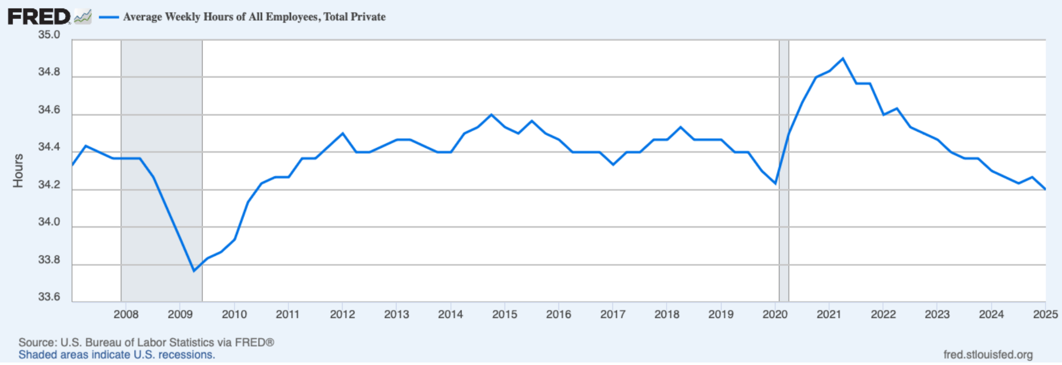
Solution: No. There is a clear upward time trend from the beginning of the sample up until the early 2000's, which means that the mean is changing over time.

Exercise 9.1: Solution

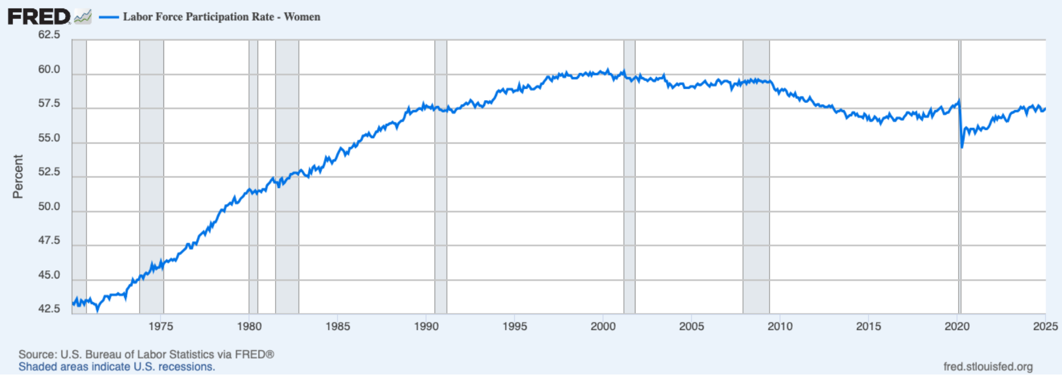
(a) S&P 500 Returns, Daily, 1 January 2020 - 1 January 2025



(b) Average Weekly Hours Worked in the Private Sector (US), Quarterly, Q1 2007 - Q1 2025



(c) Female Labor Force Participation (US), Monthly, January 1970 - January 2025



Exercise 9.2: Solution

This excercise is inspired by McConnell, Margaret and Gabriel Perez-Quiros (2000): "Output Fluctuations in the United States: What has changed since the early 1980's?" *American Economic Review*, 90(5), 1464-76.

Download some aggregate time series in quarterly frequency from FRED: Real GDP (GDPC1), Real Personal Consumption Expenditure (PCECC96), Real Gross Private Domestic Investment (GPDIC1).

1. Take logs of GDP, Consumption, and Investment. Plot the three series (you can generate the plots directly in FRED). What are the most striking features?

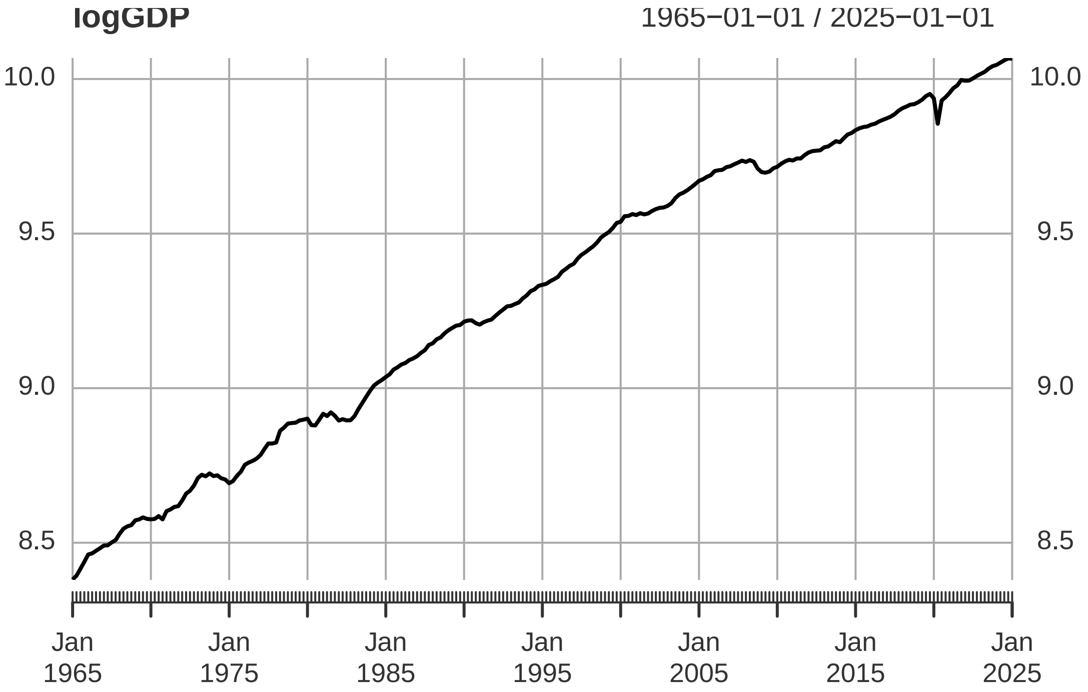
Solution: We use the following variables:

- Real Gross Domestic Product (GDPC1): https://fred.stlouisfed.org/series/GDPC1
- Real Personal Consumption Expenditures (PCECC96): https://fred.stlouisfed.org/series/PCECC96
- Real Gross Private Domestic Investment (GPDIC1): https://fred.stlouisfed.org/series/GPDIC1

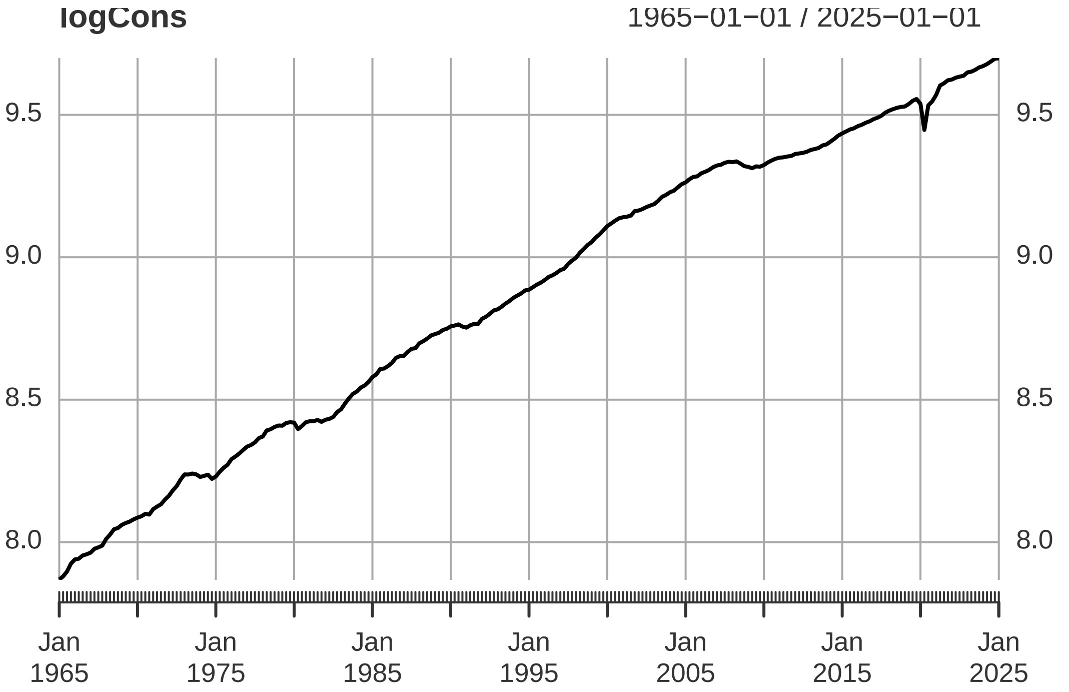
This code below reads the data and plots the series for the time period 1965:1-2022:2.

```
# Download series:

GDP <- getSymbols("GDPC1", src = "FRED", auto.assign = FALSE)
Cons <- getSymbols("PCECC96", src = "FRED", auto.assign = FALSE)
NFI <- getSymbols("GPDIC1", src = "FRED", auto.assign = FALSE)


# Define desired date range:

date.start <- "1965-01-01"
date.end <- "2025-03-01"
mytimerange <- function(x) {
    x[paste(date.start, date.end, sep = "/")]
}


# Take logs and take date range defined above:

logGDP <- log(mytimerange(GDP))
logCons <- log(mytimerange(Cons))
logNFI <- log(mytimerange(NFI))


# Plot series:

plot(logGDP)
```

Exercise 9.2: Solution

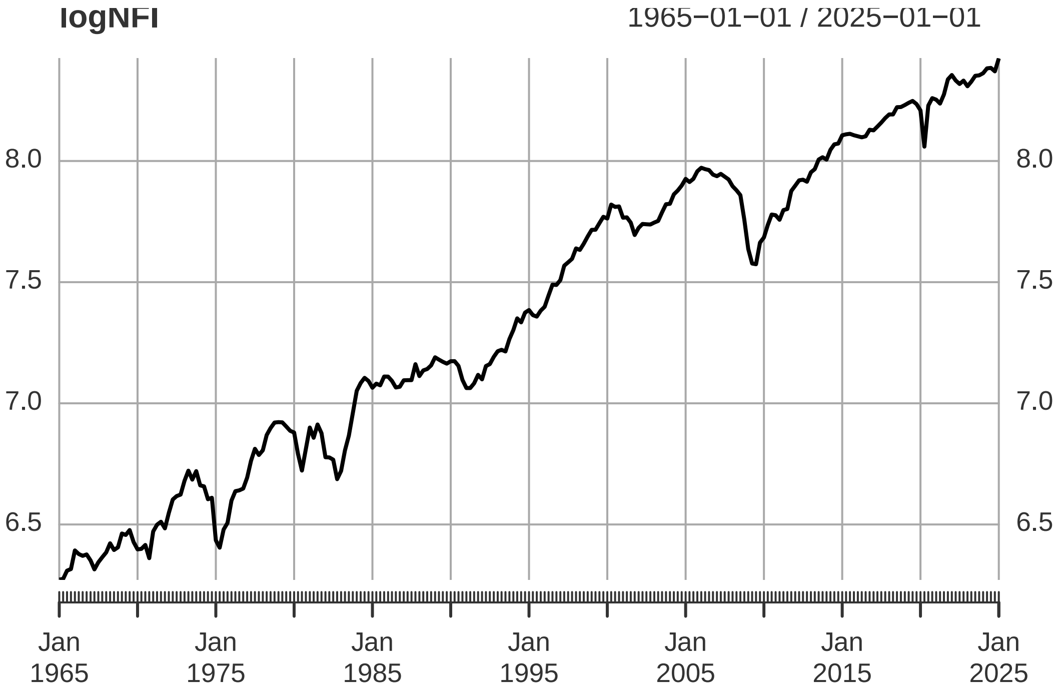


```
plot(logCons)
```



```
plot(logNFI)
```

Exercise 9.2: Solution



2. For each series (in logs), estimate a model of the form

$$y_t = \beta_1 + \beta_2 t + u_t$$

using OLS, based on the samples 1965:Q1 to 2006:Q4, 2007:Q1 to 2019:Q4, and 2007:Q1 to 2022:Q2.

Solution: The estimated coefficients for the sample 1965:1 - 2006:4 are printed below. To be concise, I will just print the results for this sample, but you can easily obtain the results for other samples by changing "date.start" and "date.end" in the code below. Note how including the Covid-19 episode changes the estimates.

```
# Define desired date range:

date.start <- "1965-01-01"
date.end <- "2006-12-01"
mytimerange <- function(x){ x[paste(date.start,date.end,sep="/")] }


# Take logs and take date range defined above:

logGDP <- log(mytimerange(GDP))
logCons <- log(mytimerange(Cons))
logNFI <- log(mytimerange(NFI))


# Estimate models:

GDPest <- lm(logGDP~seq(1,length(logGDP),by=1))
names(GDPest$coefficients)=c("beta1_GDP","beta2_GDP")
```

Exercise 9.2: Solution

```
Consest <- lm(logCons~seq(1,length(logCons),by=1))
names(Consest$coefficients)=c("beta1_Cons","beta2_Cons")

NFIest <- lm(logNFI~seq(1,length(logNFI),by=1))
names(NFIest$coefficients)=c("beta1_NFI","beta2_NFI")


# Print coefficients:

GDPest$coefficients
```

```
##   beta1_GDP   beta2_GDP
## 8.410386598 0.007770652
```

```
Consest$coefficients
```

```
##  beta1_Cons  beta2_Cons
## 7.903089677 0.008351974
```

```
NFIest$coefficients
```

```
##  beta1_NFI  beta2_NFI
## 6.22048140 0.01002159
```

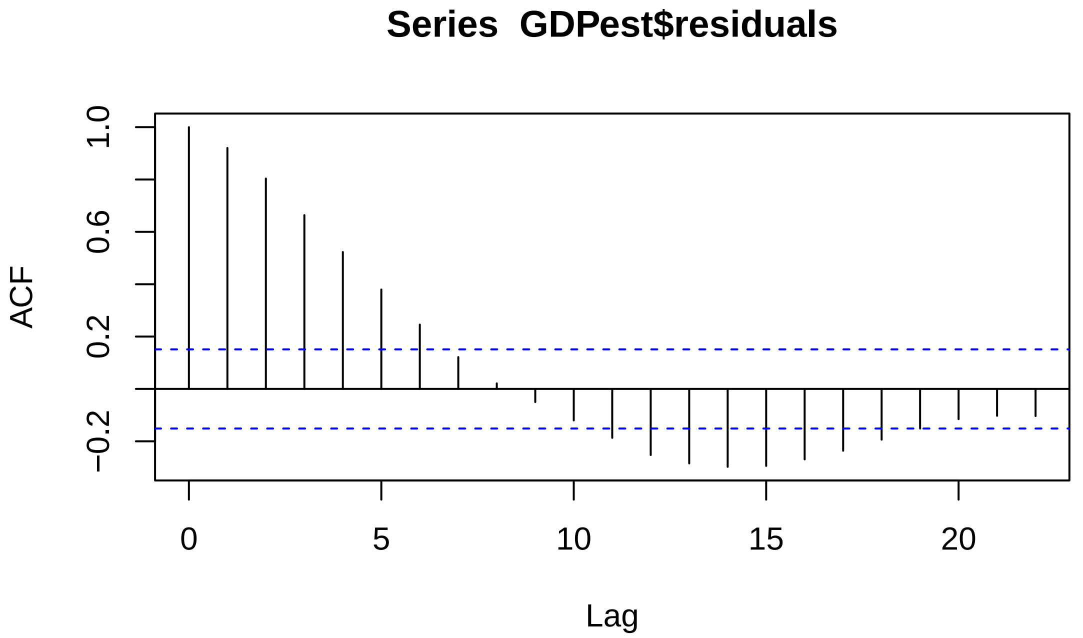
3. According to your estimates, what are the annualized average growth rates (in percent) of GDP, consumption, and investment? Are these series growing, approximately, at the same rate?

Solution: The annualized growth rate is $\log(y_t) - log(y_{t-4})$. According to our model, it corresponds to $400 * \beta_2$. The estimated growth rates are $g_{GDP} = 3.108261$, $g_{Cons} = 3.3407897$ and $g_{NFI} = 4.008636$. While GDP and consumption grow at very similar rates, the growth rate of investment is higher

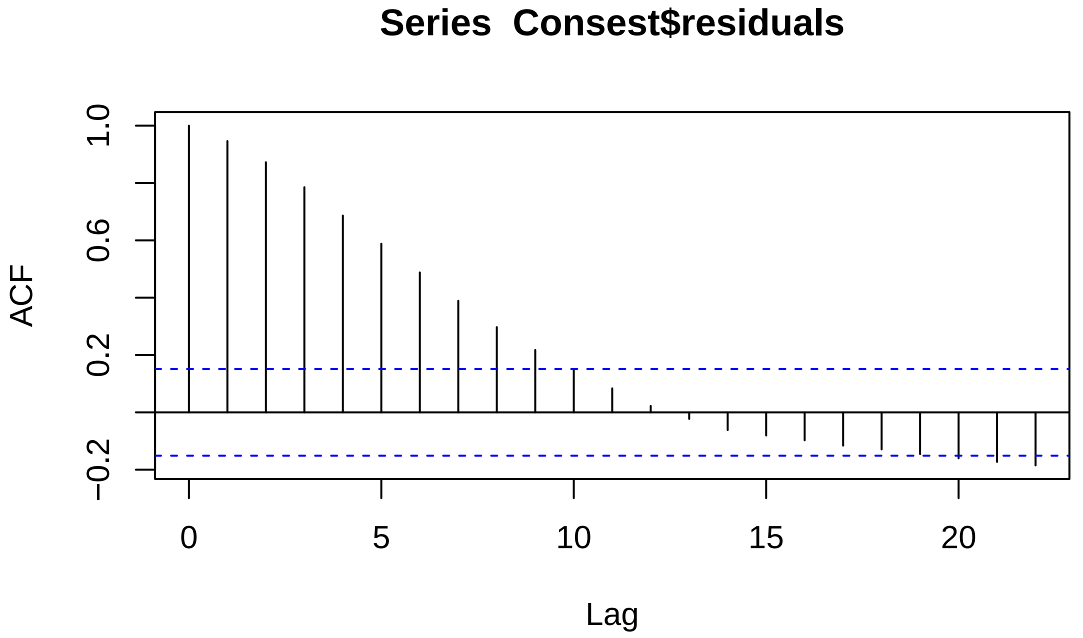
4. For each subsample, compute sample autocorrelation functions for the deviations of output, consumption, and investment (the $\hat{u}_t$'s) from their estimated deterministic trend.

```
acf(GDPest$residuals)
```

Exercise 9.2: Solution

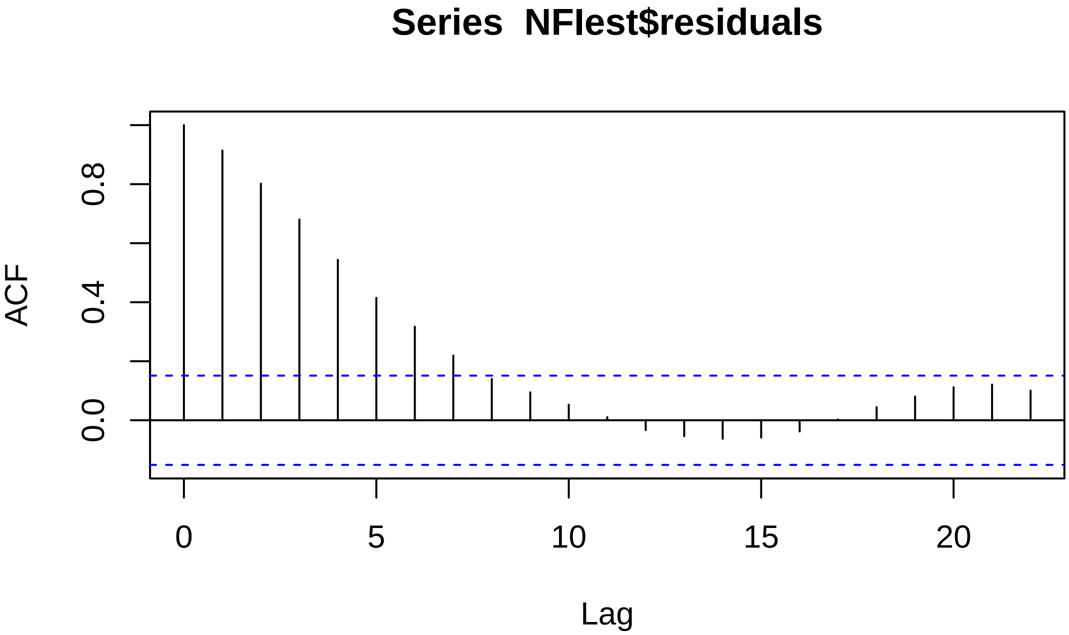


```
acf(Consest$residuals)
```



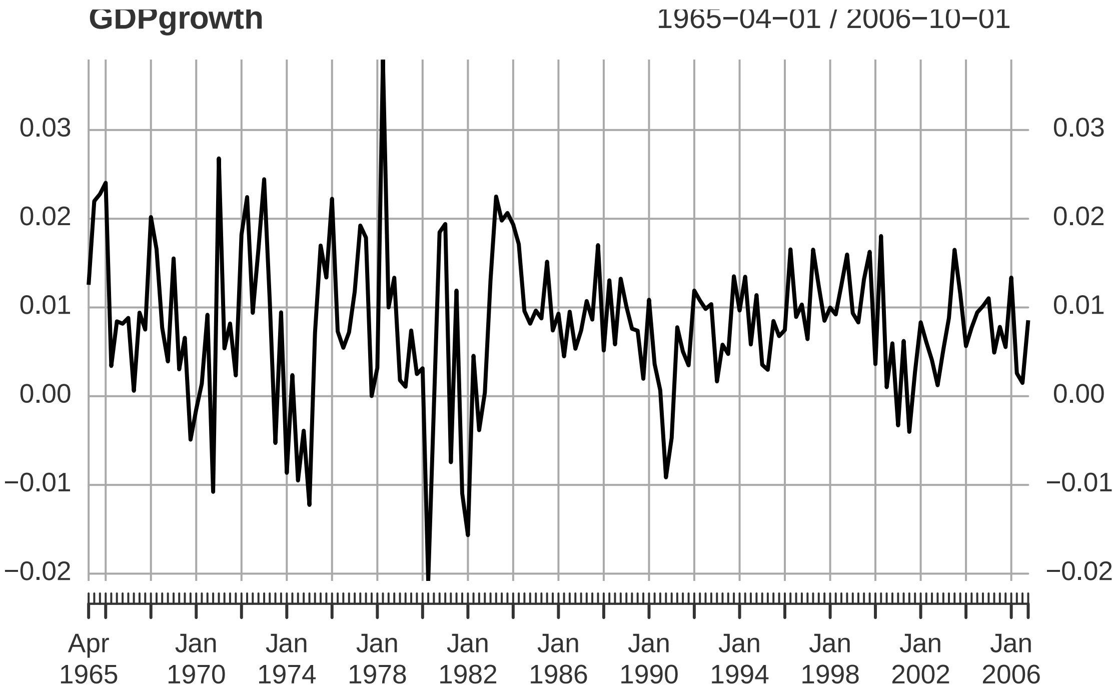
```
acf(NFIest$residuals)
```

Exercise 9.2: Solution



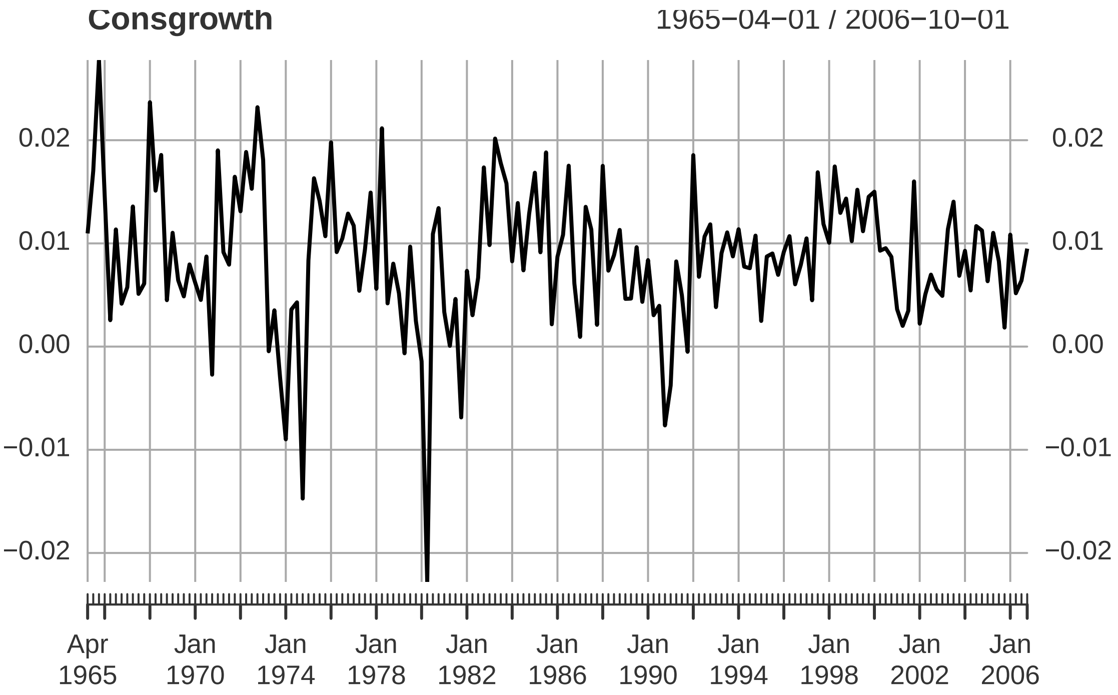
5. Now compute quarter-on-quarter growth rates of these three series as $\ln y_t - \ln y_{t-1}$, plot the growth rates. What are the most striking features?

```
GDPgrowth <- (logGDP - lag(logGDP))[-1]
Consgrowth <- (logCons - lag(logCons))[-1]
NFIgrowth <- (logNFI - lag(logNFI))[-1]

plot(GDPgrowth)
```

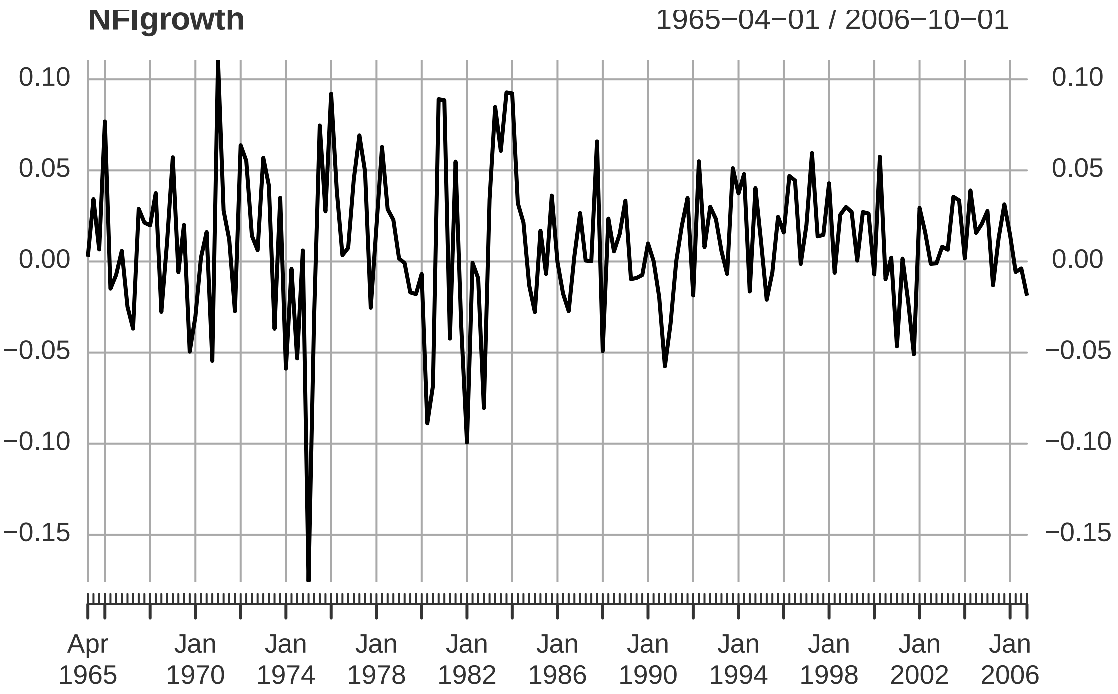


Exercise 9.2: Solution

```
plot(Consgrowth)
```



```
plot(NFIgrowth)
```



6. Compute the sample means of the growth rates for each of the above subsamples. Compare the growth-rate results to (ii). Also compute the sample standard deviations for the growth rates.

```
mean(GDPgrowth)
```

```
## [1] 0.007988902
```

Exercise 9.2: Solution

```
mean(Consgrowth)
```

```
## [1] 0.008666419
```

```
mean(NFIgrowth)
```

```
## [1] 0.0100079
```

```
sd(GDPgrowth)
```

```
## [1] 0.008264141
```

```
sd(Consgrowth)
```

```
## [1] 0.0067471
```

```
sd(NFIgrowth)
```

```
## [1] 0.03910535
```

We get pretty much the same growth rates as the ones obtained using the fitted deterministic trend model.

7. Repeat the analysis in (vi) for the subsamples “before 1984", “between 1984 and 2006". Did the means and the volatility of the series change?

Solution: Again, to be concise, I will not print results here in this document. By changing “date.start” and “date.end” you can obtain the results I refer to below.

Before 1970, the growth rates of GDP and consumption were higher and the standard deviation of growth rates was lower.

Analyzing the interval 1970 - 1982, one sees that the growth rates of GDP and in particular those of consumption decreased considerably, and that their standard deviations were higher.

Finally, analyzing only the sample after 1982 one obtains similar results as indicated in the above analysis using the whole sample. This period is referred to as “the Great Moderation” because of the decrease in GDP and consumption volatility post-1982.

Importantly, the above results pertaining to standard deviations of growth rates are contingent on not including the recent Covid-recession. Including it substantially increases the standard deviation in the post-1982 period (and whole sample since 1960) so that the period 1970 - 1982 actually gives a lower standard deviation.

Exercise 9.3: Solution

Consider the MA(3) process

$$y_t = (1 - 2.4L + 0.8L^2 - 0.4L^3)u_t \ ,$$

where L denotes the lag operator and

$$\mathbb{E}[u_t u_\tau] = \begin{cases} 1 & \text{if } t = \tau \\ 0 & \text{otherwise} \end{cases} \ .$$

(a) Define Weak and Strict Stationarity. Is the process y_t weakly stationary? Is it strictly stationary? You may impose additional conditions on the u_t's.

Solution: Weak stationarity (WS) means that the mean and autocovariances do not depend on time t. Strict stationarity (SS) means that joint CDFs for sets of observations depend only on displacement between the included observations, not time (i.e. not on their time subscripts).

Under the given assumptions, y_t is WS, since its mean and autocovariance do not depend on time (the mean is zero, and for the ACF see below). However, y_t is not necessarily SS. We know u_t is a white noise process but it might have higher order dependencies which would render the process of y_t not SS. One assumption that guarantees SS is assuming u_t is Gaussian. A Gaussian White Noise process is SS since higher order moments of a Gaussian distribution are determined by its first two moments. Then, an MA(q) process with an SS disturbance process also becomes SS.

(b) Calculate the autocovariance function of y_t.

Solution: Under covariance stationarity, we can multiply both sides with y_{t+h} and take expectations to get $E[y_t y_{t+h}]$ which equals $\gamma(h)$ since y_t is a zero-mean process.

$$\mathbb{E}[y_t y_{t+h}] = E\big[[u_t - 2.4u_{t-1} + 0.8u_{t-2} - 0.4u_{t-3}][u_{t+h} - 2.4u_{t+h-1} + 0.8u_{t+h-2} - 0.4u_{t+h-3}]\big]$$

Using $\mathbb{E}[u_t u_\tau] = 1$ for $t = \tau$ and $= 0$ for $t \neq \tau$, we can get

Exercise 9.3: Solution

$$\begin{aligned}
\gamma(0) &= \mathbb{V}[y_t] = 1 + \theta_1^2 + \theta_2^2 + \theta_3^2 \\
\gamma(1) &= \theta_1 + \theta_2\theta_1 + \theta_3\theta_2 \\
\gamma(2) &= \theta_2 + \theta_1\theta_3 \\
\gamma(3) &= \theta_3 \\
\gamma(h) &= 0 \;\; \forall h > 3 \\
\gamma(h) &= \gamma(-h) \; ,
\end{aligned}$$

where in our application $(\theta_1, \theta_2, \theta_3) = (-2.4, 0.8, -0.4)$.

(c) Calculate $\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_t\right]$.

Solution:

$$\begin{aligned}
\mathbb{V}[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_t] &= \frac{1}{T}\mathbb{E}[\left[\sum_{t=1}^{T} y_t\right]^2] \\
&= \frac{1}{T}\mathbb{E}[(y_1 + y_2 + ... + y_T) * (y_1 + y_2 + ... + y_T)] \\
&= \frac{1}{T}2\left[\frac{T}{2}\gamma(0) + (T-1)\gamma(1) + ... + \gamma(T-1)\right] \\
&= \frac{1}{T}2\left[\frac{T}{2}\gamma(0) + (T-1)\gamma(1) + (T-2)\gamma(2) + (T-3)\gamma(3)\right] \\
&= 1 + \frac{4.64}{T}
\end{aligned}$$

(d) Suppose $\{u_t\}$ is i.i.d. with $\mathbb{E}[u_t] = 0$ and $\mathbb{V}[u_t] = \sigma^2$, and define $x_t = u_t u_{t-4}$. Is the process x_t strictly stationary? Is x_t ergodic? Is x_t a White Noise process?

Solution: Since u_t is SS and ergodic and $f(x, y) = xy$ is a measurable function, $\{x_t\}$ is also strictly stationary.

Yes, x_t is a WN process: we have

$$\begin{aligned}
\mathbb{E}[x_t] &= \mathbb{E}[u_t u_{t-4}] \\
&= \mathbb{E}[u_t]\mathbb{E}[u_{t-4}] \\
&= 0 \; ,
\end{aligned}$$

Exercise 9.3: Solution

$$\begin{aligned}\mathbb{V}[x_t] &= \mathbb{E}[x_t^2] - E[x_t]^2 \\ &= \mathbb{E}[x_t^2] \\ &= \mathbb{E}[u_t^2 u_{t-4}^2] \\ &= \mathbb{E}[u_t^2]\mathbb{E}[u_{t-4}^2] \\ &= \sigma^4 \ ,\end{aligned}$$

and

$$\begin{aligned}\gamma(h) &= \mathbb{E}[x_t x_{t-h}] \\ &= \mathbb{E}[u_t u_{t-4} u_{t-h} u_{t-h-4}] \\ &= \mathbb{E}[u_t]\mathbb{E}[u_{t-4} u_{t-h} u_{t-h-4}] \\ &= 0 \ ,\end{aligned}$$

for all $h > 0$.

(e) Calculate $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x_t\right]$, and $\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right]$.

Solution:

$$\begin{aligned}\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} x_t] &= \sum_{t=1}^{T}\frac{1}{T}\mathbb{E}[x_t] \\ &= \mathbb{E}[x_t] = 0\end{aligned}$$

$$\begin{aligned}\mathbb{V}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right) &= \mathbb{E}\left(\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right]^2\right) - \left[\mathbb{E}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right)\right]^2 \\ &= \mathbb{E}\left(\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right]^2\right) \\ &= \frac{1}{T}\left[\mathbb{E}\left(\sum_{t=1}^{T} x_t^2\right) + \mathbb{E}\left(\sum_{t=h+1}^{T} x_t x_{t-h}\right)\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}x_t^2 = \sigma^4 \ .\end{aligned}$$

Exercise 9.4: Solution

Consider the following processes:

$$x_t \overset{i.i.d.}{\sim} N(0,4) \ ,$$
$$y_t = \phi y_{t-1} + u_t \ , \quad u_t \overset{i.i.d.}{\sim} N(0,4) \ .$$

Let $\bar{x} = \frac{1}{T}\sum_{t=1}^T x_t$ and $\bar{y} = \frac{1}{T}\sum_{t=1}^T y_t$.

(a) What is the probability limit of $\bar{x}$?

Solution: x_t is an i.i.d. series, which means that it is SS and ergodic. Hence, we can apply the LLN for SS and ergodic time series to get

$$\bar{x} = \frac{1}{T}\sum_{t=1}^T x_t \xrightarrow{p} \mathbb{E}[x_t] = 0 \ .$$

(b) Under $\phi = 0.8$, what is the probability limit of $\bar{y}$? How does your answer change for $\phi = 0.9$? And for $\phi = 1$?

Solution: Under $\phi = 0.8$, we have $|\phi| < 1$, and so y_t is WS. In fact, since u_t is Normal, y_t is SS, and it is ergodic. Therefore, we can apply the LLN for SS and ergodic time series to get

$$\bar{y} = \frac{1}{T}\sum_{t=1}^T y_t \xrightarrow{p} \mathbb{E}[y_t] = \frac{c}{1-\phi} = \frac{0}{1-0.8} = 0 \ .$$

This result does not change under $\phi = 0.9$, as the mean remains zero. However, for $\phi = 1$, y_t is a unit root process (a random walk), which is not stationary. This prevents us from applying the LLN for SS and ergodic time series to see to what $\bar{y}$ converges.

(c) What is the asymptotic distribution of $\bar{x}$, i.e. what is the limit distribution of $\sqrt{T}(\bar{x} - \mathbb{E}[x_t])$?

Solution: Since x_t is SS and ergodic, we can use the CLT for SS and ergodic time series to get[1]

$$\sqrt{T}(\bar{x} - \mathbb{E}[x_t]) \xrightarrow{d} N\left(0, \bar{\sigma}_x^2\right) \ ,$$

where $\bar{\sigma}_x^2 = \mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T x_t\right]$.

(d) Under $\phi = 0.8$, what is the asymptotic distribution of $\bar{y}$, i.e. what is the limit distribution of $\sqrt{T}(\bar{y} - \mathbb{E}[y_t])$? How does your answer change for $\phi = 0.9$? And for

[1] Another requirement, which is satisfied here, is that its variance is finite. Same for y_t in the next exercise.

Exercise 9.4: Solution

$\phi = 1$?

Solution: Again, for both $\phi = 0.8$ and $\phi = 0.9$, the process y_t is SS and ergodic. Hence, once again, we can invoke the CLT for SS and ergodic time series to get

$$\sqrt{T}(\bar{y} - \mathbb{E}[y_t]) \xrightarrow{d} N(0, \bar{\sigma}_y^2) \ ,$$

where $\bar{\sigma}_y^2 = \mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_t\right]$.

However, for $\phi = 1$, the process y_t is a unit root, which is not stationarity and which therefore prevents us from applying the CL for SS and ergodic time series to find out the desired asymptotic distribution.

Exercise 9.5: Solution

Consider the AR(1) model

$$y_t = \phi y_{t-1} + u_t \ , \quad u_t \overset{i.i.d.}{\sim} N(0,1) \ \ ,$$

where $|\phi| < 1$.

(a) What is the difference between a conditional and unconditional likelihood function? Derive them for the present AR(1) model.

Solution: The conditional likelihood function for the parameter vector θ is $p(Y_{1:T}|\theta, y_0)$. This likelihood function is defined conditional on the initial observation y_0. In contrast, the unconditional likelihood function also takes into account the likelihood of initial observations: $p(Y_{1:T}, y_0|\theta) = p(Y_{1:T}|\theta, y_0)p(y_0|\theta)$.

For an AR(1) model, the conditional likelihood function takes the form:

$$\begin{aligned}
p(Y_{1:T}|\theta, y_0) &= p(y_T|\theta, y_{T-1}, y_{T-2}, ..., y_0)p(y_{T-1}, y_{T-2}, ..., y_1|\theta, y_0) \\
&= p(y_T|\theta, y_{T-1}, y_{T-2}, ..., y_0)p(y_{T-1}|y_{T-2}, ..., y_1, \theta, y_0)p(y_{T-1}, y_{T-2}, ..., y_1|\theta, y_0) \\
&= ... \\
&= \prod_{t=1}^{T} p(y_t|\theta, y_{t-1}, ..., y_1, y_0) \\
&= (2\pi)^{-T/2}(\sigma^2)^{-T/2} \exp\left\{ \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \phi y_{t-1})^2 \right\} .
\end{aligned}$$

We know that for the AR(1) model, under stationarity, the unconditional distribution of any observation y_t – including notably y_0 – is

$$y_t \sim N\left(0, \frac{\sigma^2}{1-\phi^2}\right) .$$

Hence, the unconditional likelihood function is:

$$p(Y_{1:T}|\theta, y_0) = (2\pi)^{-(T+1)/2}(\sigma^2)^{-(T+1)/2}(1-\phi^2)^{1/2} \exp\left\{ \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \phi y_{t-1})^2 \right\} \exp\left\{ \frac{(1-\phi^2)}{2\sigma^2} y_0^2 \right\} .$$

(b) Derive the conditional ML estimator of ϕ, denoted by $\hat{\phi}$, using the conditional likelihood function.

Exercise 9.5: Solution

Solution: The conditional log-likelihood function is

$$\begin{aligned}\log p(Y_{1:T}|\theta, y_0) &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \phi y_{t-1})^2 \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\phi)'(Y - X\phi) \ ,\end{aligned}$$

where Y stacks $\{y_1, ..., y_T\}$ and X stacks $\{x_1, ..., x_T\}$, with $x_t = y_{t-1}$. Maximizing it w.r.t ϕ gives the FOC

$$\sum_{t=1}^{T}(y_t - \phi y_{t-1})y_{t-1} = X'(Y - X\phi) = 0 \ .$$

Solving this for ϕ, we get the conditional ML estimator which coincides with the OLS estimator:

$$\widehat{\phi} = (X'X)^{-1}X'Y = \frac{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}y_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2} \ .$$

(c) Show that $\hat{\phi}$ is a consistent estimator of ϕ.

Solution: We can write

$$\widehat{\phi} = \frac{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}y_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2} = \phi + \frac{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}u_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2} \ .$$

Given that u_t is i.i.d. and Normal, it is strictly stationary and ergodic (by the ergodicity theorem). Because $|\phi| < 1$, we know y_t is weakly stationary. In fact, since we can write y_t as an infinite sum of u_t's (i.e. y_t is a measurable transformation of u_t), we know that y_t is also strictly stationary and ergodic. The same holds for y_{t-1}^2 and $y_{t-1}u_t$. Thus, we can apply the LLN and CLT for SS and ergodic series to these terms. Using the LLN for the denominator, we get

$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{p} \mathbb{E}[y_{t-1}^2] = \gamma_0 = \frac{\sigma^2}{1 - \phi^2} \ .$$

Using the LLN for the numerator, we get

$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}u_t \xrightarrow{p} \mathbb{E}[y_{t-1}u_t] = 0 \ .$$

Exercise 9.5: Solution

Overall, we obtain the result that $\widehat{\phi}$ is consistent:

$$\widehat{\phi} - \phi \xrightarrow{p} 0 \ .$$

(d) Derive the limit distribution of $\hat{\phi}$.

Solution: We can write

$$\sqrt{T}(\widehat{\phi} - \phi) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1} u_t}{\frac{1}{T} \sum_{t=1}^{T} y_{t-1}^2} \ .$$

Using the CLT for ergodic and SS time series, we get

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1} u_t \xrightarrow{d} N\left(0, \mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1} u_t\right]\right) \quad ,$$

where

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1} u_t\right] &= \frac{1}{T}\mathbb{V}\left[\sum_{t=1}^{T} y_{t-1} u_t\right] \\
&= \frac{1}{T} \sum_{t=1}^{T} \mathbb{V}[y_{t-1}u_t] + \frac{1}{T} \sum_{t=1}^{T} \sum_{\tau \neq t} \mathrm{Cov}(y_{t-1}u_t, x_\tau u_\tau) \\
&= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[u_t^2 y_{t-1} y_{t-1}'] + \frac{1}{T} \sum_{t=1}^{T} \sum_{\tau \neq t} \mathbb{E}[y_{t-1} u_t x_\tau u_\tau] \\
&= \mathbb{E}[u_t^2]\mathbb{E}[y_{t-1} y_{t-1}'] \\
&= \sigma^2 \gamma_0 \ ,
\end{aligned}$$

whereby $\mathbb{E}[u_t^2 y_{t-1} y_{t-1}'] = \mathbb{E}[u_t^2]\mathbb{E}[y_{t-1} y_{t-1}']$ and $\mathbb{E}[y_{t-1} u_t x_\tau u_\tau] = \mathbb{E}[y_{t-1} x_\tau]\mathbb{E}[u_t]\mathbb{E}[u_\tau] = 0$ because u_t is independent across t.[1] Combining this result with the probability limit of the denominator in the expression for $\hat{\phi}$, we can deduce that $\hat{\phi}$ is asymptotically

[1] Alternatively, and more simply, applying the CLT for Martingale Difference Sequences (MDS) gives

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1} u_t \Rightarrow N\left(0, \frac{\sigma^4}{1-\phi^2}\right) ,$$

where we used the fact that the second moment of $y_{t-1}u_t$ is

$$\mathbb{E}[y_{t-1}^2 u_t^2] = \mathbb{E}[y_{t-1}^2]\mathbb{E}[u_t^2] = \sigma^2 \frac{\sigma^2}{1-\phi^2} \ .$$

Exercise 9.5: Solution

normal:

$$\sqrt{T}(\widehat{\phi} - \phi) \Rightarrow N(0, v) \ ,$$

where $v = \sigma^2 \gamma_0 \gamma_0^{-2} = \sigma^2 \left(\frac{\sigma^2}{1-\phi^2} \right)^{-1} = 1 - \phi^2$.

(e) Suppose data is generated from a stationary AR(2) model

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t \quad , \quad \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2) \ ,$$

but the econometrician estimates the above AR(1) model. What are the probability limits of $\hat{\phi}$ and $\hat{\sigma}_u^2$ in terms of $(\rho_1, \rho_2, \sigma_\epsilon^2)$?
Hint: note that this requires you to find the variance and first-order covariance of y_t under the AR(2) process.

Solution: Recall that the estimator takes the form

$$\widehat{\phi} = \frac{\frac{1}{T}\sum_{t=1}^T y_{t-1} y_t}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2} \ .$$

As before, by the LLN for strictly stationary and ergodic processes, the denominator converges to the variance of y_t,

$$\frac{1}{T}\sum_{t=1}^T y_{t-1}^2 \xrightarrow{p} \mathbb{E}[y_{t-1}^2] = \gamma_0 \ ,$$

while the numerator converges to the first-order covariance,

$$\frac{1}{T}\sum_{t=1}^T y_t y_{t-1} \xrightarrow{p} \mathbb{E}[y_t y_{t-1}] = \gamma_1 \ .$$

The variance and first-order covariance of a stationary AR(2) are given by the expressions:

$$\gamma_0 = \frac{1-\rho_2}{(1+\rho_2)[(1-\rho_2)^2 - \rho_1^2]} \sigma_\epsilon^2 \ ,$$

$$\gamma_1 = \frac{\rho_1}{1-\rho_2} \gamma_0 \ .$$

(These are obtained by solving the system of equations implied by the Yule-Walker equations.) Summing up, we get

$$\widehat{\phi} \xrightarrow{p} \frac{\rho_1}{1-\rho_2} \ .$$

Exercise 9.5: Solution

Now, let's find the ML estimator of σ_u^2 by taking derivative of the conditional log likelihood function w.r.t. σ^2. This gives the FOC

$$-\frac{T}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{1}{\sigma^4}\sum_{t=1}^{T}(y_t-\phi y_{t-1})^2=0$$

Hence,

$$\begin{aligned}\widehat{\sigma^2}&=\frac{1}{T}\sum_{t=1}^{T}(y_t-\widehat{\phi}y_{t-1})^2\\&=\frac{1}{T}\sum_{t=1}^{T}(y_t^2-2\widehat{\phi}y_ty_{t-1}+\widehat{\phi}^2y_{t-1}^2)\\&=\frac{1}{T}\sum_{t=1}^{T}y_t^2-2\widehat{\phi}\frac{1}{T}\sum_{t=1}^{T}y_ty_{t-1}+\widehat{\phi}^2\frac{1}{T}\sum_{t=1}^{T}y_{t-1}^2\ .\end{aligned}$$

Now, we could directly find the probability limit of this expression by using the probability limits of $\widehat{\phi}$, $\frac{1}{T}\sum_{t=1}^{T}y_t^2$, $\frac{1}{T}\sum_{t=1}^{T}y_ty_{t-1}$ and $\frac{1}{T}\sum_{t=1}^{T}y_{t-1}^2$. However, it turns out that it is simpler to first plug in for $\widehat{\phi}$ and simplify this expression. We get

$$\begin{aligned}\widehat{\sigma^2}&=\frac{1}{T}\sum_{t=1}^{T}y_t^2-2\frac{\frac{1}{T}\sum_{t=1}^{T}y_{t-1}y_t}{\frac{1}{T}\sum_{t=1}^{T}y_{t-1}^2}\frac{1}{T}\sum_{t=1}^{T}y_ty_{t-1}+\left[\frac{\frac{1}{T}\sum_{t=1}^{T}y_{t-1}y_t}{\frac{1}{T}\sum_{t=1}^{T}y_{t-1}^2}\right]^2\frac{1}{T}\sum_{t=1}^{T}y_{t-1}^2\\&=\frac{1}{T}\sum_{t=1}^{T}y_t^2-\frac{\left[\frac{1}{T}\sum_{t=1}^{T}y_{t-1}y_t\right]^2}{\frac{1}{T}\sum_{t=1}^{T}y_{t-1}^2}\end{aligned}$$

By LLN:

$$\frac{1}{T}\sum_{t=1}^{T}y_t^2\xrightarrow{p}\mathbb{E}[y_t^2]=\gamma_0\quad\text{and}\quad\frac{1}{T}\sum_{t=1}^{T}y_{t-1}^2\xrightarrow{p}\gamma_0\ .$$

By LLN and Slutsky's theorem,

$$\left[\frac{1}{T}\sum_{t=1}^{T}y_{t-1}y_t\right]^2\xrightarrow{p}\mathbb{E}[y_{t-1}y_t]^2=\gamma_0^2\Big(\frac{\rho_1}{1-\rho_2}\Big)^2\ .$$

Exercise 9.5: Solution

Putting everything together, we get (by Slutsky's theorem),

$$\begin{aligned}\widehat{\sigma^2} &\xrightarrow{p} \gamma_0 - \left(\frac{\rho_1}{1-\rho_2}\right)^2 \gamma_0 \\ &= \gamma_0 \Big[1 - \frac{\rho_1^2}{(1-\rho_2)^2}\Big] \\ &= \sigma_\epsilon^2 \frac{1-\rho_2}{(1+\rho_2)[(1-\rho_2)^2-\rho_1^2]} \frac{(1-\rho_2)^2-\rho_1^2}{(1-\rho_2)^2} \\ &= \frac{\sigma_\epsilon^2}{1-\rho_2^2} \ .\end{aligned}$$

Exercise 9.6: Solution

Suppose you model Swiss CPI inflation, y_t, as an AR(1) process:

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t \ , \quad u_t \sim WN(0, \sigma^2) \ ,$$

where $|\phi_1| < 1$.

(a) What does $u_t \sim WN(0, \sigma^2)$ mean? What does it tell you about the properties of u_t?

Solution: It tells us that u_t is a white noise process (with mean zero and variance σ^2), which means that it is uncorrelated over time: $\mathbb{E}[u_t u_s] = 0$ for all $t \neq s$.

(b) What does the parameter ϕ_1 signify? No derivations are needed.

Solution: ϕ_1 captures the autocorrelation of y_t, which is a measure of persistence. The larger $|\phi_1|$, the more persistent is the process, i.e. the more it is determined by the past (embodied by y_{t-1}) rather than by news, innovations to the process (embodied by u_t).

(c) What does the parameter ϕ_0 signify? Suppose that Swiss inflation averages to around 0.3 percent over the sample that you consider. Assuming that $\phi_1 = 0.8$, where do you expect ϕ_0 to lie?

Solution: The intercept ϕ_0 is related to the mean of y_t (or, put differently, to the "equilibrium value" of the process y_t); it is given by $\mathbb{E}\left[y_t\right] = \frac{\phi_0}{1-\phi_1}$. Therefore, if in our sample we find $\bar{y} = 0.3$, and assuming that $\phi_1 = 0.8$, then it is reasonable to expect ϕ_0 to lie in the neighborhoood of $0.3 \times (1 - 0.8) = 0.06$.

(d) Let us fix $\phi_0 = 0$. Derive the OLS estimator for ϕ_1, $\hat{\phi}_1$.

Solution: We rewrite the model in matrix notation as $Y = \phi X + U$, where Y stacks y_t for $t = 1 : T$, X stacks y_{t-1}, and U stacks u_t. The OLS estimator $\hat{\phi}_1$ is defined as:

$$\begin{aligned} \hat{\phi}_1 &= \arg\min_{\phi_1} \sum_{t=1}^{T} u_t^2 \\ &= \arg\min_{\phi_1} U'U \\ &= \arg\min_{\phi_1} (Y - \phi X)'(Y - \phi X) \ . \end{aligned}$$

By taking the FOC, we get

$$\hat{\phi}_1 = (X'X)^{-1} X'Y = \left(\sum_{t=1}^{T} y_{t-1}^2 \right)^{-1} \sum_{t=1}^{T} y_{t-1} y_t \ .$$

Exercise 9.6: Solution

(e) Assuming that your model is correct – i.e. y_t is actually determined as

$$y_t = \phi_1 y_{t-1} + u_t \ , \quad \text{with} \quad u_t \sim WN(0, \sigma^2)$$

–, what is the probability limit of $\hat{\phi}_1$? State clearly any assumptions you need to make to find this limit.

Solution: Under the assumption that the model for y_t is indeed correct, we can write the OLS estimator as:

$$\begin{aligned}\hat{\phi}_1 &= \left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1}y_t \\ &= \phi_1 + \left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1}u_t \ .\end{aligned}$$

Next, under the assumption that y_t and u_t are SS and ergodic time series,[1] we can apply the LLN for such series to get

$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{p} \mathbb{E}[y_{t-1}^2]$$

and

$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}u_t \xrightarrow{p} \mathbb{E}[y_{t-1}u_t] = 0 \ .$$

By Slutsky's theorem, then, we can combine the two results above to get that $\hat{\phi}_1 \xrightarrow{p} \phi_1$.

(f) Instead, suppose now that you misspecified your model and inflation actually follows an MA(1) process:

$$y_t = (1 + \theta L)\varepsilon_t \quad \text{for} \quad \varepsilon_t \sim WN(0, \sigma^2) \ .$$

Derive the probability limit of $\hat{\phi}_1$ in terms of θ. Again, state clearly any assumptions you need to make to find this limit.
Hint: you will need to compute moments of the MA(1) process. Note that $y_t = (1+\theta L)\varepsilon_t$ holds for every period t, and so $y_{t-1} = (1 + \theta L)\varepsilon_{t-1}$.

Solution: To find the probability limit of $\hat{\phi}_1$, we insert the true model for y_t, which is

[1] This would, for example, be given if we assume that u_t is Normal. This implies that it is an i.i.d. series, which makes it SS and ergodic, and which in turn makes y_t SS and ergodic, since y_t – under $|\phi_1| < 1$ – can be written as the sum of all past u_ts.
Also, note that this implies then that y_t^2 and $y_{t-1}u_t$ are SS and ergodic as well.

Exercise 9.6: Solution

now given by this MA(1) process:

$$\begin{aligned}\hat{\phi}_1 &= \left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1}y_t \\ &= \left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1}(\varepsilon_t + \theta\varepsilon_{t-1}) \\ &\xrightarrow{p} \mathbb{E}[y_{t-1}^2]^{-1}\left(\mathbb{E}[y_{t-1}\varepsilon_t] + \theta\mathbb{E}[y_{t-1}\varepsilon_{t-1}]\right) \ ,\end{aligned}$$

where the last line follows by the LLN for SS and ergodic time series (and Slutsky's theorem), for which we need to assume that y_t is SS and ergodic.[2] By stationarity, $\mathbb{E}[y_{t-1}^2] = \mathbb{E}[y_t^2]$, and, for an MA(1) process, this quantity (the variance of y_t) equals

$$\mathbb{V}[y_t] = \mathbb{E}[y_t^2] = \mathbb{E}[(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_t + \theta\varepsilon_{t-1})] = \mathbb{E}[\varepsilon_t^2 + \theta^2\varepsilon_{t-1}^2] = (1+\theta^2)\sigma^2 \ .$$

Also, we have

$$\mathbb{E}[y_{t-1}\varepsilon_t] = \mathbb{E}[(\varepsilon_{t-1} + \theta\varepsilon_{t-2})\varepsilon_t] = 0$$

and

$$\mathbb{E}[y_{t-1}\varepsilon_{t-1}] = \mathbb{E}[(\varepsilon_{t-1} + \theta\varepsilon_{t-2})\varepsilon_{t-1}] = \sigma^2 \ .$$

Therefore, overall we have $\hat{\phi}_1 \xrightarrow{p} \dfrac{\theta}{1+\theta^2}$.

[2] Just as above, this is, for example, given as soon as we assume that ε_t is Normal.

Exercise 9.7: Solution

Consider the AR(1) model

$$y_t = \phi y_{t-1} + u_t \;, \quad u_t \overset{i.i.d.}{\sim} N(0,1) \;, \quad y_0 = 0 \;.$$

1. Take $\phi = 0.8$, and generate $m = 1, \ldots, M = 1{,}000$ samples of length $T = 80$. For each sample m compute the OLS/ML estimator of ϕ, denoted by $\hat{\phi}^m$. This gives you a set of $M = 1000$ estimators, $\{\hat{\phi}^m\}_{m=1}^{M}$.

Solution

```
#required packages
library(stats)
library(quantmod)
library(ggplot2)
library(patchwork) # to have several ggplots side by side

rm(list = ls())
set.seed(121)
T <- 80 + 1
M <- 1000
mInnovations <- matrix(rnorm(T * M), T, M)
phi_ <- 0.8

#simulate the distribution of phi_hat and compute t-stat
simulateDistribution <- function(phi, T, M, mInnovations) {
  tStat <- rep(0, M)
  for (i in 1:M) {
    y <- rep(mInnovations[1 , i], T)
    for (t in 2:T) {
      y[t] <- phi * y[t - 1] + mInnovations[t, i]
    }
    phiHat <- solve(t(y[1:(T - 1)]) %*% y[1:(T - 1)],
                    t(y[1:(T - 1)]) %*% y[2:(T)])
    tStat[i] <- (phiHat - phi) / sqrt(1 / T * (1 - phi ^ 2))
  }
  return(data.frame(tStat))
}
```

2. The asymptotic distribution suggests that

$$\sqrt{T}(\hat{\phi} - \phi) \overset{approx.}{\sim} N(0, \hat{\gamma}_0^{-1}) \;, \quad \gamma_0 = \mathbb{V}[y_t] = \mathbb{V}[y_{t-1}] \;,$$

and hence, for the t-statistic, it should hold that

$$\frac{\sqrt{T}(\hat{\phi} - \phi)}{1/\sqrt{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}} \overset{approx.}{\sim} N(0,1) \;.$$

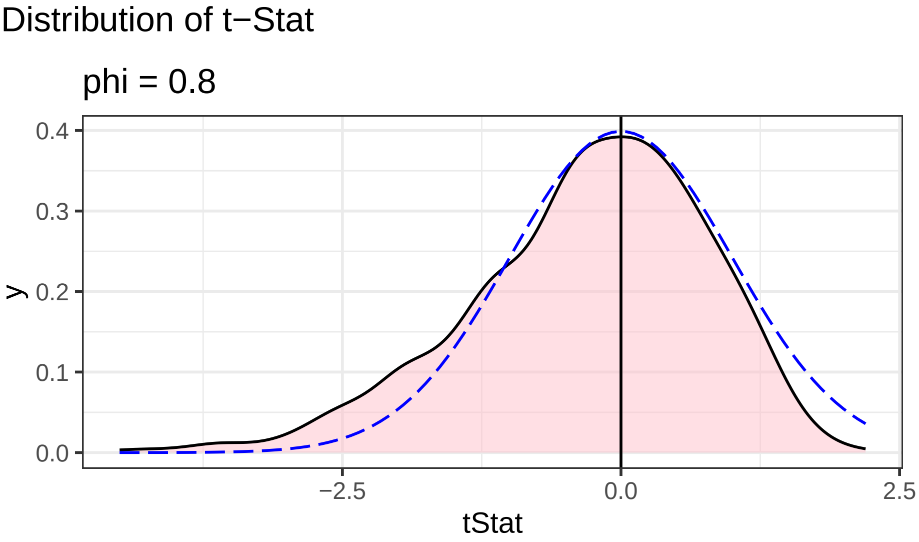
Generate the actual finite sample distribution of this t-statistic, i.e. compute the t-statistic for each of

Exercise 9.7: Solution

your estimators $\{\hat{\phi}^m\}_{m=1}^M$, and plot a histogram. Is it close to a standard Normal distribution? Overlay the pdf of a $N(0,1)$ on your histogram to see.

Solution

```
results <- simulateDistribution(phi_, T, M, mInnovations)

 gg <- ggplot(results) +  geom_density(aes(x = tStat), alpha= 1/2, fill = "pink") +
     geom_vline(xintercept  = 0) +
    theme_bw() + theme(legend.position="none") + #theme(aspect.ratio=5/8) +
    stat_function(fun = dnorm, args=list(mean = 0, sd = 1),
                              colour="blue", linetype = "longdash") +
    ggtitle(paste0("phi = ", phi_))

gg+ plot_annotation(title = "Distribution of t-Stat")
```



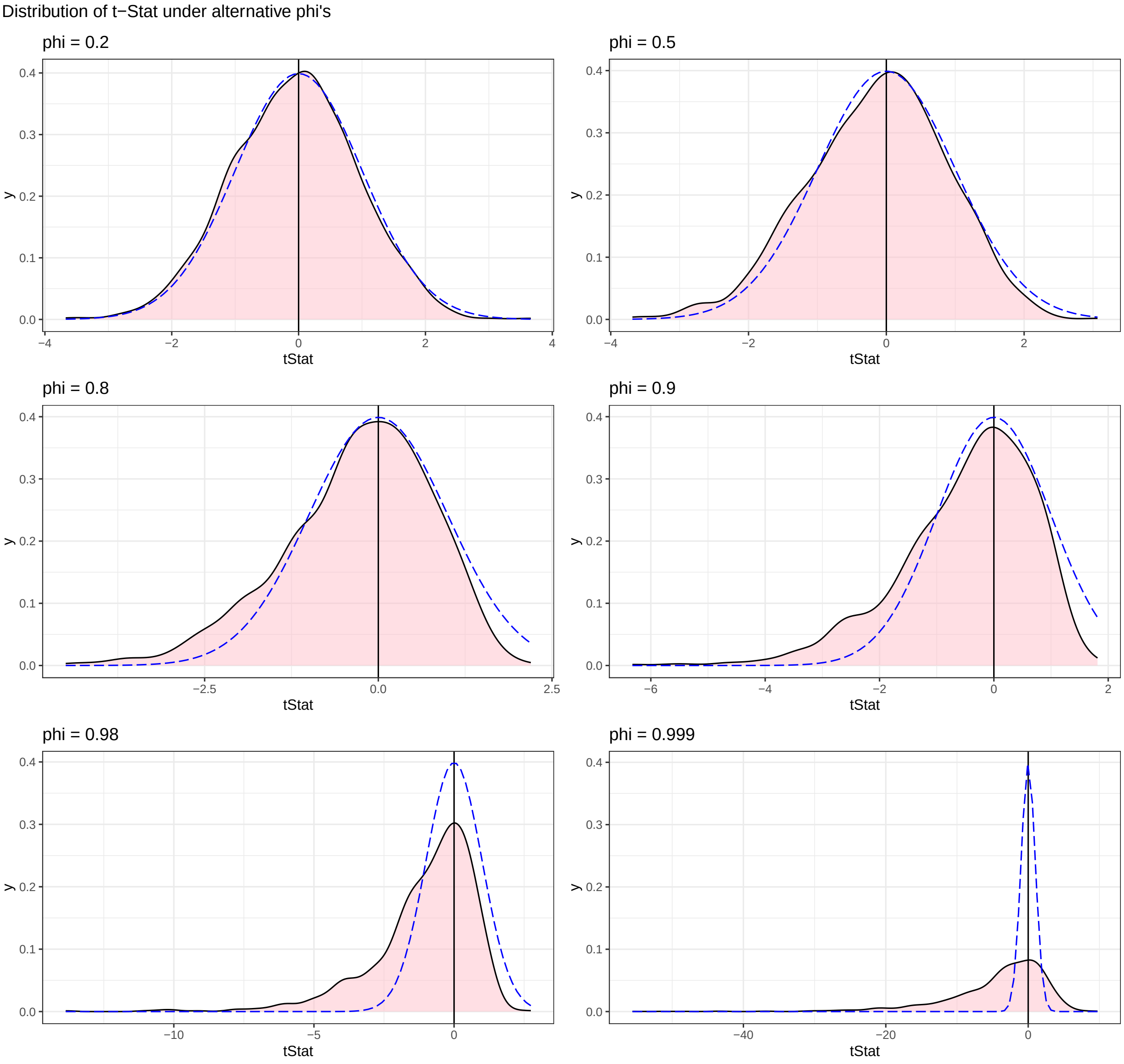
3. Repeat the previous two exercises for the parameter values $\phi \in \Phi = \{0.2, 0.5, 0.9, 0.98, 0.999\}$. Discuss your results in view of the asymptotic calculations for $|\phi| < 1$ vs. $\phi = 1$. What do the simulations say about the accuracy of the asymptotic distribution as an approximation of the finite-sample distribution?

Solution

```
phis <- c(0.2, 0.5, 0.8, 0.9, 0.98, 0.999)
results <- lapply(phis,
                  FUN = function(x) simulateDistribution(x, T, M, mInnovations))
lPlots <- list()

for (i in 1:length(phis)) {
  gg <- ggplot(results[[i]]) +  geom_density(aes(x = tStat), alpha= 1/2, fill = "pink") +
     geom_vline(xintercept  = 0) +
    theme_bw() + theme(legend.position="none") + #theme(aspect.ratio=5/8) +
    stat_function(fun = dnorm, args=list(mean = 0, sd = 1),
                              colour="blue", linetype = "longdash") +
```

Exercise 9.7: Solution

```
    ggtitle(paste0("phi = ", phis[i]))
  #ggsave(paste0("RPlotDistrPhi", phis[i] * 100, ".pdf"), gg)

  lPlots <- append(lPlots,list(gg))

}

ppp <- (lPlots[[1]] | lPlots[[2]] ) /
    (lPlots[[3]] | lPlots[[4]]) /
    (lPlots[[5]] | lPlots[[6]] )
ppp + plot_annotation(title = "Distribution of t-Stat under alternative phi's")
```



We know that for $|\phi| < 1$, the asymptotic distribution of the Student-t statistic above should indeed be $N(0,1)$, whereas for the unit root case $\phi = 1$ we get the Dickey-Fuller distribution. The plots show that in

Exercise 9.7: Solution

finite samples even for $|\phi| < 1$ the statistic's distribution is very poorly approximated by the asymptotic distribution, ever more so the closer ϕ is to 1. Note that this suggests that the OLS/ML estimator is downward biased.

In contrast, for the Bayesian computations, the supposedly true value of the parameter is irrelevant (i.e. whether the process is stationary or not is irrelevant): the posterior that we derive is correct regardless of that value. Intuitively, this is because the Baysesian approach conditions on the data, whereas the frequentist approach analyzes an estimator's properties under the assumption that nature provides it repeatedly with random samples from the population, for which the properties of the process under the supposedly true value of the parameter are important.

Exercise 9.8: Solution

Consider the AR(1) model

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t \ , \quad u_t \overset{iid}{\sim} N(0, \sigma^2) \ .$$

(a) Derive the conditional likelihood function $p(Y_{1:T}|y_0, \phi_0, \phi_1, \sigma^2)$ and the MLE $(\hat{\phi}_0, \hat{\phi}_1, \hat{\sigma}^2)$.

Solution:

$$\begin{aligned} p(Y_{1:T}|y_0, \theta) &= \prod_t p(y_t|Y_{0:t-1}, \theta) \\ &= \prod_t (2\pi\sigma^2)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}(y_t - x_t'\phi)^2\right\} \\ &= (2\pi\sigma^2)^{-T/2} exp\left\{-\frac{1}{2\sigma^2}(Y - X\phi)'(Y - X\phi)\right\} \ , \end{aligned}$$

which (through the FOCs) yields

$$\hat{\phi} = (X'X)^{-1}X'Y \ , \quad \hat{\sigma^2} = \frac{1}{T}(Y - X\hat{\phi})'(Y - X\hat{\phi}) \ .$$

(b) What assumptions do you need to establish consistency and find the asymptotic distribution of the MLE? Derive the asymptotic distribution of $\hat{\phi}_1|\sigma^2$ (i.e. assuming σ^2 is a constant you know). Is it a good approximation of its finite sample distribution?

Solution: We need strict stationarity (SS) and ergodicity.

To find the asymptotic distribution of $\hat{\phi}_1|\sigma^2$, we first find the (joint) asymptotic distribution of $\hat{\phi}|\sigma^2$. We get

$$\sqrt{T}(\hat{\phi} - \phi) = \phi + \left(\frac{1}{T}\sum_t x_t x_t'\right)^{-1} \frac{1}{\sqrt{T}}\sum_t x_t u_t \xrightarrow{d} N(0, V) \ ,$$

by Slutsky's theorem, assembling the convergences of the numerator and denominator, respectively. For the denominator, we get $\frac{1}{T}\sum_t x_t x_t' \xrightarrow{p} \mathbb{E}[x_t x_t']$ by the LLN for SS and ergodic TS, and $\left(\frac{1}{T}\sum_t x_t x_t' \xrightarrow{p} \mathbb{E}[x_t x_t']\right)^{-1} \xrightarrow{p} \mathbb{E}[x_t x_t']^{-1}$ then by the CMT. For the numerator, we get

$$\frac{1}{\sqrt{T}}\sum_t x_t u_t \xrightarrow{d} N(0, S) \ ,$$

Exercise 9.8: Solution

by the CLT for SS and ergodic TS. The mean is zero because $\frac{1}{T}\sum_t x_t u_t \xrightarrow{p} \mathbb{E}[x_t u_t] = 0$ by the mentioned LLN. Also, we have

$$S = \mathbb{V}[\frac{1}{\sqrt{T}}\sum_t x_t u_t] = \frac{1}{T}\sum_t \mathbb{V}[x_t u_t] = \sigma^2 \mathbb{E}[x_t x_t'] \ ,$$

which holds because $Cov(x_t u_t, x_s u_s) = \mathbb{E}[x_t x_s' u_t u_s] = 0 \ \forall \ t \neq s$ by LIE and $\mathbb{E}[u_t x_t] = 0$. Overall, we have $V = \mathbb{E}[x_t x_t']^{-1} S \mathbb{E}[x_t x_t']^{-1} = \mathbb{E}[x_t x_t']^{-1}\sigma^2$.

The sought-for asymptotic distribution is then

$$\sqrt{T}(\hat{\phi}_1 - \phi_1)|\sigma^2 \xrightarrow{d} N(0, \sigma^2 \left(\mathbb{E}[x_t x_t']^{-1}\right)_{22}) \ ,$$

where M_{22} denotes the element (2,2) of matrix M. (Note that one could simply this further based on $x_t = [1 y_{t-1}]'$.)

This is a good approximation of the finite sample distribution of $\hat{\phi}_1|\sigma^2$ only if ϕ_1 is not too close to 1 and if the sample size is large enough.

(c) Suppose you are unsure whether the true model contains one or two lags of y_t, i.e. whether y_t follows an AR(1) or AR(2). Describe two possible ways to choose among the two models (under the frequentist paradigm).

Solution: One way to proceed is to test $\phi_2 = 0$ in the AR(2) model. The other way is to compare the two models using information criteria like the BIC (or an approximation thereof).

Exercise 9.9: Solution

Consider the time series regresssion

$$y_t = x_t'\beta + u_t \;, \quad \text{or} \quad Y = X\beta + U \;,$$

where x_t and β are k-dimensional vectors, Y and U are n-dimensional vectors, and X is $n \times k$. Let $\mathbb{E}[u_t x_t] = 0$ and $\mathbb{V}[u_t] = \sigma_u^2$.

(a) (3 points) Would it be reasonable to assume that u_t is i.i.d.? Why (not)?

Solution: For u_t to be i.i.d., it must not be dependent over time and it must follow the same distribution for every t. As y_t is likely dependent over time, for u_t to be independent, all the dependency in y_t must be captured by x_t, which is unlikely.
[1p] for NO, [2p] for explanation, [-1p] for saying we can assume WN

(b) (3 points) Derive the OLS estimator for β, $\hat{\beta}_{OLS}$.

Solution:

$$\hat{\beta}_{OLS} := \arg\min_{\beta} \sum_{t=1}^{T} (y_t - x_t'\beta)^2 := \arg\min_{\beta} (Y - X\beta)'(Y - X\beta) \;.\textbf{[1p]}$$

The FOC is $-2X'(Y - X\beta) = 0$ **[1p]** and leads to $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ **[1p]**. In alternative notation, we have

$$\frac{\partial}{\partial\beta} \sum_{t=1}^{T} (y_t - x_t'\beta)^2 = -2\sum_{t=1}^{T} x_t(y_t - x_t'\beta) = \sum_{t=1}^{T} x_t y_t - \beta \sum_{t=1}^{T} x_t x_t' = 0 \;,$$

leading to

$$\hat{\beta}_{OLS} = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t y_t \;.$$

(c) (6 points) Is $\hat{\beta}_{OLS}$ consistent? What properties do the processes y_t, x_t and u_t need to satisfy so that you can analyze consistency? Define these properties.

Solution: To have convergence in probability (LLN) we need y_t, x_t and u_t are strictly stationary (i.e. that the joint distribution of a bunch of observations is stable over time; it only depends on the time periods separating the observations displacement, not on time itself) and ergodic (i.e. that observations become independent asymptotically as the displacement goes to ∞). **[3p]**

Exercise 9.9: Solution

Under these assumptions, we have that $\hat{\beta}_{OLS}$ is consistent:

$$\hat{\beta}_{OLS} - \beta = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \left(\sum_{t=1}^{T} x_t u_t \right) \xrightarrow{p} 0 \; [\mathbf{1p}]$$

because

$$\frac{1}{T} \sum_{t=1}^{T} x_t x_t' \xrightarrow{p} \mathbb{E}[x_t x_t'] \;, \qquad \text{and} \quad \frac{1}{T} \sum_{t=1}^{T} x_t u_t \xrightarrow{p} \mathbb{E}[x_t u_t] = 0 \;,$$

and assembling these two pieces by Slutsky's theorem yields the above result. **[2p]**

(d) (4 points) Does the question whether or not u_t is i.i.d. affect the asymptotic variance of $\hat{\beta}_{OLS}$? Explain.

Solution: Yes. **[1p]** If x_t, y_t and u_t are SS and ergodic, $\hat{\beta}_{OLS}$ also converges in distribution:

$$\sqrt{T}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} \mathbb{E}[x_t x_t']^{-1} N \left(0, \mathbb{V} \left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t u_t \right] \right) \mathbb{E}[x_t x_t']^{-1} \;.$$

Thereby,

$$\mathbb{V} \left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t u_t \right] = \frac{1}{T} \mathbb{V} \left[\sum_{t=1}^{T} x_t u_t \right] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{V}\left[x_t u_t \right] + \frac{1}{T} \sum_{t=1}^{T} \sum_{\tau \neq t} cov(x_t u_t, x_\tau u_\tau) \;.$$

If $x_t u_t$ is i.i.d. and hence uncorrelated over time, the second term is zero. Otherwise, it is not. **[2p]** When $x_t u_t$ is not i.i.d., the asymptotic variance is greater compared to the i.i.d.-case because $x_t u_t$ is likely positively correlated over time. **[1p]**

Exercise 9.10: Solution

Consider the time series regresssion

$$y_t = x_t'\beta + u_t \;, \quad \text{or} \quad Y = X\beta + U \;, \tag{1}$$

where x_t and β are k-dimensional vectors, Y and U are n-dimensional vectors, and X is $n \times k$. In addition, suppose that u_t follows an AR(1) process:

$$u_t = \rho u_{t-1} + \varepsilon_t \;, \quad \text{with} \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_e^2) \quad \text{and} \quad |\rho| < 1 \;. \tag{2}$$

(a) (3 points) Define weak stationarity. Is ε_t weakly stationary? Is u_t weakly stationary?

Solution: A stochastic process y_t is WS when:

- $\mathbb{E}[y_t] = \mu \qquad \forall t$
- $\gamma_{h,t} = \mathbb{E}[y_t y_{t-h}] = \gamma_h \qquad \forall t$

i.e. first and second moments do not depend on time. **[1p]**

ε_t is WS: for all t, we have $\mathbb{E}[\varepsilon_t] = 0$, $\gamma_{0,t} = \sigma_\varepsilon^2$ and $\gamma_{h,t} = 0$ for $h \neq 0$. **[1p]** u_t is also WS because we know that an AR(1) process is WS as long as $|\rho| < 1$. **[1p]**

(b) (4 points) Compute the variance and first-order autocovariance of u_t.

Solution: Because u_t is WS, we know $\mathbb{E}[u_t] = \mathbb{E}[u_{t-1}]$ and $\mathbb{V}[u_t] = \mathbb{V}[u_{t-1}]$. Hence, we get

$$\mathbb{E}[u_t] = \rho\mathbb{E}[u_t] + 0 \quad \Rightarrow \quad \mathbb{E}[u_t] = 0$$

(to be used below) and

$$\mathbb{V}[u_t] = \rho^2\mathbb{V}[u_t] + \sigma_\varepsilon^2 \quad \Rightarrow \quad \mathbb{V}[u_t] = \frac{\sigma_\varepsilon^2}{1-\rho^2} \;.\textbf{[2p]}$$

We can get the first-order autocovariance using the Yule-Walker equation. To get it, we multiply the AR(1)-equation on both sides by u_{t-1} and take expectations:

$$\begin{aligned} &\mathbb{E}[u_t u_{t-1}] = \rho\mathbb{E}[u_{t-1}^2] + \mathbb{E}[\varepsilon_t u_{t-1}] \\ \Leftrightarrow \quad &\gamma_1 = \rho\gamma_0 + 0 \;, \end{aligned}$$

because u_{t-1} depends only on $\varepsilon_{t-1}, \varepsilon_{t-2}, ...$ and hence is uncorrelated with ε_t. We also used $\mathbb{E}[u_t] = 0$, which implies that the second term in $\gamma_1 = \text{Cov}(u_t, u_{t-1}) =$

Exercise 9.10: Solution

$\mathbb{E}[u_t u_{t-1}] - \mathbb{E}[u_t]\mathbb{E}[u_{t-1}]$ is zero. Overall, we get

$$\gamma_1 = \rho\gamma_0 = \frac{\rho\sigma_\varepsilon^2}{1-\rho^2}\ .\textbf{[2p]}$$

The alternative, more cumbersome approach is to compute the two objects using the MA(∞)-representation of u_t. Using again the fact that $|\rho| < 1$ (i.e. u_t being WS), we have

$$u_t = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}$$

and therefore

$$\mathbb{E}[u_t] = \sum_{j=0}^{\infty} \rho^j \mathbb{E}[\varepsilon_{t-j}] = 0\ ,$$

$$\mathbb{V}[u_t] = \mathbb{E}[u_t^2] = \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}\right)^2\right] = \sum_{j=0}^{\infty} \rho^{2j}\sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2}\ ,$$

and, more generally, for all h,

$$\begin{aligned}\gamma_h = \mathbb{E}[u_t u_{t-h}] &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \rho^j u_{t-j}\right)\left(\sum_{k=0}^{\infty} \rho^k u_{t-h-k}\right)\right] \\ &= \sum_{k=0}^{\infty} \rho^{h+k}\rho^k\sigma_\varepsilon^2 \\ &= \sigma_\varepsilon^2 \rho^h \sum_{k=0}^{\infty} \rho^{2k} \\ &= \frac{\rho^h}{1-\rho^2}\sigma_\varepsilon^2\ ,\end{aligned}$$

which gives the above expression for γ_1.

(c) (2 points) Derive an equation for y_t in terms of y_{t-1}, x_t and x_{t-1} and the i.i.d. error term ε_t (rather than the autocorrelated error term u_t).
Hint: solve the time series regression equation (1) for u_t and u_{t-1}, plug these expressions into the AR(1) equation (2), and solve for y_t.

Solution: We have $u_t = y_t - x_t'\beta$ and therefore

$$(y_t - x_t'\beta) = \rho(y_{t-1} - x_{t-1}'\beta) + \varepsilon_t\ ,$$

Exercise 9.10: Solution

or, rearranged,

$$y_t = \rho y_{t-1} + x_t'\beta - \rho(x_{t-1}'\beta) + \varepsilon_t \ .$$

(d) (4 points) Let $\theta = (\beta', \rho, \sigma_e^2)'$. Derive the likelihood $p(Y|X, y_0, \theta)$ associated with the model you derived in exercise (c).
Hint: recall how the likelihood of an AR(1) model $z_t = \phi z_{t-1} + \varepsilon_t$ is derived as the product of conditionals: $p(Z|z_0, \phi, \sigma_e^2) = \prod_{t=1}^{T} p(z_t|z_{t-1}, \phi, \sigma_e^2)$. Here, the analogous applies for y_t. Also, note that by conditioning on X we treat $x_t \ \forall \ t$ as given.

Solution: Analogously to the derivation mentioned in the hint, we have

$$\begin{aligned}
p(Y|X, y_0, \theta) &= p(y_t|X, \theta, y_{t-1}, ..., y_0)p(y_{t-1}, ..., y_1|X, \theta, y_0) \\
&= ... \\
&= \prod_{t=1}^{\infty} p(y_t|X, \theta, y_{t-1}, ..., y_0) \\
&= \prod_{t=1}^{\infty} (2\pi\sigma_e^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_e^2}(y_t - \rho y_{t-1} + x_t'\beta - \rho(x_{t-1}'\beta)^2\right\} \\
&= (2\pi\sigma_e^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{t=1}^{T}(y_t - \rho y_{t-1} + x_t'\beta - \rho(x_{t-1}'\beta)^2\right\} \\
&= (2\pi\sigma_e^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma_e^2}(Y - \rho X_Y - X\beta + \rho X_X\beta)'(Y - \rho X_Y - X\beta + \rho X_X\beta)\right\} \ ,
\end{aligned}$$

where:

- Y is $(T \times 1)$ and stacks $(y_T, y_{T-1}, ..., y_1)$,
- X_Y is $(T \times 1)$ and stacks $(y_{T-1}, y_{T-2}, ..., y_0)$,
- X is $(T \times k)$ and stacks $(x_T, x_{T-1}, ..., x_1)$ along rows, and
- X_X is $(T \times k)$ and stacks $(x_{T-1}, x_{T-2}, ..., x_0)$ along rows.

Exercise 9.11: Solution

Suppose you know σ_e^2 (you may set $\sigma_e^2 = 1$ if you want), and your prior for $(\beta', \rho)'$ is

$$p(\beta, \rho) = p(\beta)p(\rho) \; , \quad \text{with} \quad \beta \sim N(0, \sigma_e^2 \lambda^{-1} I) \quad \text{and} \quad \rho \sim N(0, \sigma_e^2 \tau^{-1}) \; .$$

To simplify notation, write your above likelihood $p(Y|X, y_0, \theta)$ as $p(Y|\beta, \rho)$.

(a) (4 points) Suppose no further analytical calculations are possible; you have the likelihood $p(Y|\beta, \rho)$ and prior $p(\beta, \rho)$, but nothing more. How could you obtain the joint posterior $p(\beta, \rho|Y)$[1] numerically? Describe the procedure you would use.

Solution: We could use the Random-Walk-Metropolis-Hastings (RWMH) algorithm. **[1p]** To apply it, we only need to be able to evaluate numerically the prior $p(\beta, \rho)$ and the likelihood $p(Y|\beta, \rho)$. The idea is to start from some initial value $\theta^0 = (\beta^0, \rho^0)$ chosen by the researcher and iteratively draw a new value for θ using the preceding value. More concretely, given a previous draw $\theta^{i-1} = (\beta^{i-1}, \rho^{i-1})$, we draw a candidate-draw v from some proposal density and as our θ^i we then either take this draw v (we "accept" the candidate v) or we just stay with θ^{i-1} (we "reject" the candidate v). We continue with the iterations until the set of draws we obtained have "stabilized", i.e. their distribution (moments) does not change. By setting the probability of acceptance to the ratio of posteriors under v and θ^{i-1},

$$\alpha(v|\beta^{i-1}, \rho^{i-1}) = \min\left\{1, \frac{p(v|Y)}{p(\theta^{i-1}|Y)}\right\} = \min\left\{1, \frac{p(Y|v)p(v)}{p(Y|\theta^{i-1})p(\theta^{i-1})}\right\} \; ,$$

the algorithm is more likely to accept draws which have a lot of probability mass under the posterior. As a result, the set of values (draws) we are left with in the end approximates well the desired posterior. **[3p]**

(b) (6 points) Under the prior $\beta \sim N(0, \sigma_e^2 \lambda^{-1} I)$, derive the conditional posterior of $\beta|\rho$. Given this conditional posterior, what do you expect the conditional posterior of $\rho|\beta$ to be? You do not need to do any derivations for $\rho|\beta$.
Hint: to obtain $p(\beta|Y, \rho)$, write the model as $\tilde{y}_t = \tilde{x}_t \beta + \varepsilon_t$, for some $\tilde{y}_t$ and $\tilde{x}_t$.

Solution: Above, we derived the likelihood:

$$p(Y|\rho, \beta) = (2\pi\sigma_e^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma_e^2}(Y - \rho X_Y - X\beta + \rho X_X \beta)'(Y - \rho X_Y - X\beta + \rho X_X \beta)\right\} \; .$$

As indicated in the hint, conditional on ρ, we can write our model as $\tilde{Y} = \tilde{Y}\beta + U$,

[1] To be precise, this posterior is also conditional on y_0 and X, just like the likelihood. Same for the conditional posteriors mentioned subsequently.

Exercise 9.11: Solution

whereby we know (i.e. can treat as known data) $\tilde{Y} = Y - \rho X_Y$ and $\tilde{X} = X - \rho X_X$. Correspondingly, we write the above likelihood as

$$p(Y|\rho, \beta) = (2\pi\sigma_e^2)^{-T/2} \exp\left\{ -\frac{1}{2\sigma_e^2}(\tilde{Y} - \tilde{X}\beta)'(\tilde{Y} - \tilde{X}\beta) \right\} .\textbf{[1p]}$$

Under the prior $\beta \sim N(0, \sigma_e^2 \lambda^{-1} I)$, we then get the posterior

$$\begin{aligned} p(\beta|Y, \rho) &\propto \exp\left\{ -\frac{1}{2\sigma_e^2}(\tilde{Y} - \tilde{X}\beta)'(\tilde{Y} - \tilde{X}\beta) \right\} \exp\left\{ -\frac{1}{2\sigma_e^2}\beta'\lambda I \beta \right\} \\ &\propto \exp\left\{ -\frac{1}{2\sigma_e^2}[-\beta\tilde{X}'\tilde{Y} - \tilde{Y}'\tilde{X}\beta + \beta'\tilde{X}'\tilde{X}\beta + \beta'\lambda I\beta] \right\} \\ &\propto \exp\left\{ -\frac{1}{2\sigma_e^2}[\beta'[\tilde{X}'\tilde{X} + \lambda I]\beta - 2\beta\tilde{Y}'\tilde{X}] \right\} , \textbf{[2p]} \end{aligned}$$

which lets us conclude that

$$\beta|Y, \rho \sim N(\bar{\beta}, \sigma_e^2 \bar{V}_\beta), \quad \bar{\beta} = [\tilde{X}'\tilde{X} + \lambda I]^{-1}\tilde{X}'\tilde{Y}, \quad \bar{V}_\beta = [\tilde{X}'\tilde{X} + \lambda I]^{-1} .\textbf{[1p]}$$

Analogously, conditional on β, we can write our model as $Y^* = X^*\rho + U$ with $Y^* = Y - X\beta$ and $X^* = X_Y - X_X\beta$. This yields the posterior

$$\rho|Y, \beta = N(\bar{\rho}, \sigma_e^2 \bar{V}_\rho), \quad \bar{\rho} = [X^{*'}X^* + \tau]^{-1}X^{*'}Y^*, \quad \bar{V}_\rho = [X^{*'}X^* + \tau]^{-1} .$$

(Compared to the above posterior for β, we just replace $\tilde{Y}$ with Y^*, $\tilde{X}$ with X^* and λ with τ. We can also drop the identity matrix because, while β is a vector, ρ is a scalar.) **[2p]**

(c) (4 points) Given the conditional posteriors $p(\beta|Y, \rho)$ and $p(\rho|Y, \beta)$, how could you obtain the joint posterior $p(\beta, \rho|Y)$? Describe the procedure you would use.

Solution: We can use the Gibbs sampling algorithm: **[1p]** we take some initial value $\theta^0 = (\beta^0, \rho^0)$, and we then iteratively draw a new value for θ using the preceding value. More concretely, given a previous draw $\theta^{i-1} = (\beta^{i-1}, \rho^{i-1})$, we draw β^i from the conditional density $p(\beta|Y, \rho^{i-1})$ (i.e. using the most recent draw of ρ we have, ρ^{i-1}) and we draw ρ^i from $p(\rho|Y, \beta^i)$ (i.e. conditioning on the most recent draw of β we have, β^i). We continue with this procedure until the set of draws we obtained have "stabilized", i.e. their distribution (moments) does not change and so we can be reasonably sure that they have converged to the desired posterior. **[3p]**

10 Multivariate & Nonlinear Time Series Analysis

10.1 Bivariate VAR(1): Basic Calculations and Identification-Discussion

Consider the following bivariate VAR(1):

$$y_t = \Phi_1 y_{t-1} + u_t \ , \quad u_t = \Phi_\varepsilon \varepsilon_t \ , \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0, I).$$

where $y_t = (w_t, h_t)'$ is composed of log wages w_t and log hours h_t. The vector of structural shocks $\varepsilon_t = (\varepsilon_{a,t}, \varepsilon_{b,t})'$ is composed of a labor demand shock (technology shok) $\varepsilon_{a,t}$ and a labor supply shock (preference shock) $\varepsilon_{b,t}$.

(a) What condition does Φ_1 have to satisfy so that y_t is stationary?

(b) Suppose y_t is stationary, derive the autocovariances of order zero and one, denoted by $\Gamma_{yy}(0)$ and $\Gamma_{yy}(1)$.

(c) Derive the impulse response function of y_{t+h}, $h = 0, 1, \ldots$ with respect to the vector of structural shocks ε_t. How do log wages react to a labor supply shock that occurred 3 periods before?

(d) Describe the identification problem in the context of this VAR.

(e) Supose we are willing to assume that, contemporaneously, hours worked are only affected by preferences, not technology. What restrictions does this assumption impose on Φ_ε? Is this enough to uniquely identify Φ_ε?

(f) Alternatively, suppose that we assume that a labor supply shock $\varepsilon_{b,t}$ moves wages and hours in opposite directions upon impact, whereas a demand shock $\varepsilon_{a,t}$ moves wages and hours in the same direction. What restrictions does this assumption impose on Φ_ε? Is this enough to uniquely identify Φ_ε?

We can think of wages and hours being determined by an interplay of labor supply and labor demand. Let $h_t = \varphi^D(w_t, y_{t-1}; \varepsilon_{a,t}, \varepsilon_{b,t})$ be the demand and $h_t = \varphi^S(w_t, y_{t-1}; \varepsilon_{a,t}, \varepsilon_{b,t})$ the supply function. They show the relationship between hours and wages (i.e. quantity and price in the labor market), whereby these functions (think of labor/supply curves) depend on current technology and preference shocks as well as past hours and wages.[1]

(h) Suppose we assume that the labor demand is only affected by the technology shock, not the preference shock, whereas labor supply is affected by both shocks:

$$h_t = \varphi^D(w_t, y_{t-1}; \varepsilon_{a,t}) \quad \text{and} \quad h_t = \varphi^S(w_t, y_{t-1}; \varepsilon_{a,t}, \varepsilon_{b,t}) \; .$$

Is this enough to uniquely identify Φ_ε?
Hint: remember that demand must equal supply at all times.

(i) Does your answer change if, on top of the above assumption, we assume that

$$\frac{\partial w_t}{\partial \varepsilon_{b,t}} = (\alpha - 1)\frac{\partial h_t}{\partial \varepsilon_{b,t}}$$

for some given α; conditional on α, is it possible to uniquely identify the elements of Φ_ε? If yes, show how you can solve for Φ_ε based on α and the reduced-form VAR parameters.

10.2 Bivariate VAR(1): Identification under Sign Restrictions

Consider the following bivariate VAR(1)

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} \Sigma_{11}^{tr} & 0 \\ \Sigma_{12}^{tr} & \Sigma_{22}^{tr} \end{bmatrix} \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \; ,$$

written more compactly as

$$y_t = \Phi y_{t-1} + \Sigma_{tr}\Omega(\varphi)\varepsilon_t \; , \quad \varepsilon_t \sim N(0, I) \; ,$$

where I is the identity matrix.

(a) In the context of the above equation, what is a reduced-form VAR and what is a structural VAR?

(b) Show that $\Omega\Omega' = I$ and $\Omega'\Omega = I$ for any $\varphi \in [-\pi, \pi]$.
Hint: we have $\cos(\varphi)^2 + \sin(\varphi)^2 = 1 \; \forall \; \varphi \in \mathbb{R}$.

[1] In a structural economic model for the labor market, we would typically assume exogenous preference and technology processes (not i.i.d., but persistent, e.g. AR(1)s), which then, combined with the households' (workers') and firms' utility and profit maximization problems, lead to such demand and supply functions. Owing to the persistence preference and technology processes, the demand and supply functions will depend on past preference and technology shocks, which are summarized by past hours and wages in y_{t-1}.

(c) Is φ identifiable from the data? Are Φ and Σ_{tr} identifiable? Explain carefully.

(d) Suppose the parameters $(\Phi, \Sigma_{tr}, \varphi)$ are given. What is the definition of an impulse response function (IRF) and how would you compute an IRF for a one-standard deviation shock $\epsilon_{1,t}$.

(e) Suppose we want to impose that both y_{1t} and y_{2t} respond positively upon impact to a shock ϵ_{1t}. What restrictions does this assumption impose on φ?
Hint: based on the fact that $\Sigma_{tr}\Sigma_{tr}' = \Sigma$, where Σ is a variance-covariance matrix, you can deduce the signs of Σ_{11}^{tr} and Σ_{22}^{tr}, which allows you to give a more detailed response. Also, see Fig. 10.1 below.

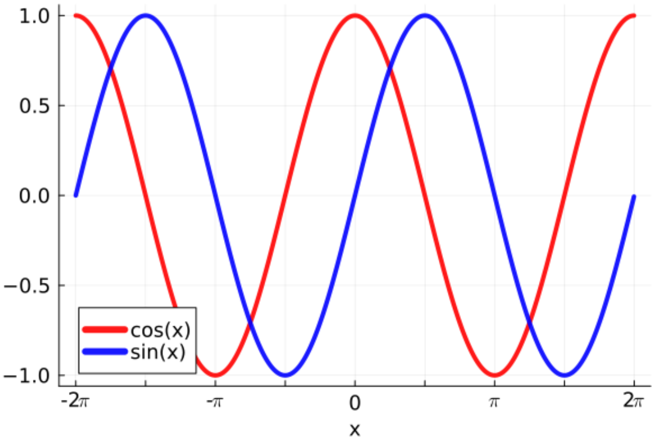


Figure 10.1: Sine and Cosine Functions

10.3 Bivariate VAR(1): Frequentist & Bayesian Estimation & IRF-Computation

Let y_t be an $n \times 1$ vector and consider the following VAR(1):

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} N(0, \Sigma) \; .$$

(a) Derive the conditional likelihood $p(Y_{1:T}|y_0, \Phi_0, \Phi_1, \Sigma)$ and the ML estimator(s) for (Φ_0, Φ_1, Σ),

$$(\hat{\Phi}_0, \hat{\Phi}_1, \hat{\Sigma}) = \arg \max_{(\Phi_0, \Phi_1, \Sigma)} p(Y_{1:T}|y_0, \Phi_0, \Phi_1, \Sigma) \; .$$

(b) Download quarterly observations on US real GDP growth and inflation from FRED[2] for the period 1985:I to 2019:IV. Estimate a VAR(1) with intercept by OLS. Conditional on the OLS estimate, compute the eigenvalues of Φ_1, and the unconditional mean and variance of y_t. Compare the unconditional mean and variance implied by the estimated VAR to sample mean and variance of y_t. Would you expect them to be approximately the same? Are they approximately the same?

[2] This is the FRED database of the Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org. Use the series “GDPC1” for the former and the series “GDPDEF” or “CPIAUCSL” for the latter.

(c) Derive the MNIW-posterior under the Normal likelihood above and the improper prior

$$p(\Phi, \Sigma) \propto |\Sigma|^{-(n+1)/2} .$$

(d) Write a program that generates draws $m = 1, \dots, M$ from the posterior $p(\Phi, \Sigma|Y)$ you derived using direct sampling from the MNIW distribution. (If you were unable to derive the posterior in the previous exercise, take the posterior under the improper prior $p(\Phi, \Sigma) \propto c$ shown in the script.) For each parameter in Φ and Σ, compute the posterior mean and a 90% credible interval. Tabulate your posterior mean estimates along with the OLS estimates.

(e) For each posterior draw m, compute the lower-triangular Cholesky factor Σ_{tr} and compute Impulse Response Functions (IRFs) based on $\Phi_\epsilon = \Sigma_{tr}$. Concretely, for each draw m, you should get an IRF_{ij}^h for the four pairs (i, j) and for the horizons $h = 1 : 12$. Then, for each (i, j) pair, generate a plot with h on the x-axis and, for each h, the mean as well as the 90% Bayesian credible set of IRF_{ij}^h on the y-axis. (To obtain the 90% credible set, you may assume that the distribution of IRF_{ij}^h is symmetric).

Show the result in a single figure with 2 × 2 subplots. Comment on your plots. Which identification assumption is embodied in taking $\Phi_\epsilon = \Sigma_{tr}$? How does that assumption manifest itself in your plots?

(f) How do the IRFs change if you re-estimate the VAR with $p = 4$ lags?

10.4 Simple State Space Model: Filtering, Smoothing & ML-EM-Estimation

The following exercises are inspired by Aruoba, Diebold, Nalewaik, Schorfheide & Song (2016): "Improving GDP measurement: A measurement-error perspective," *Journal of Econometrics,* 191, 384-397.

A country's GDP can be measured in two ways: i) by measuring the *expenditures* of households, firms and governments on products and services produced in that country during a specific period, or ii) by measuring the *income* obtained by households, firms and governments from products and services produced in that country during a specific period.[3] Even though in theory, these two measures should coincide, in practice they do not exactly. In the case of the US, the former is somewhat confusingly referred to as "GDP", while the latter is referred to as "GDI" (see for example the website of the Bureau of Economic Analysis (BEA): https://www.bea.gov/data/economic-accounts/national).

(a) Download the two quarterly series "GDPC1" (real gross domestic product) and "A261RX1Q020SBEA" (real gross domestic income) from FRED[4] for the period 1979:I to 2023:IV, and compute their respective growth rates relative to the same quarter in the previous year (giving you two series

[3] The latter includes not only labor income, but also income from the supply of other productive factors.

[4] This is the FRED database of the Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org.

starting from 1980:I). We will call the former $GDP_{E,t}$ and the latter $GDP_{I,t}$. Plot the two series as well as their difference.

(b) Estimate an AR(1) with intercept for each of the two series. Comment on the estimated means, variances and autocorrelations of the two series.

(c) Consider the first model from the mentioned paper, intended to extract some underlying true GDP growth series, GDP_t, based on data on $GDP_{E,t}$ and $GDP_{I,t}$:

$$\begin{bmatrix} GDP_{E,t} \\ GDP_{I,t} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} GDP_t + \begin{bmatrix} \epsilon_{E,t} \\ \epsilon_{I,t} \end{bmatrix} ,$$
$$GDP_t = \phi_0 + \phi_1 GDP_{t-1} + \epsilon_t ,$$

where $\epsilon_{E,t} \sim N(0, \sigma_E^2)$, $\epsilon_{I,t} \sim N(0, \sigma_I^2)$ and $\epsilon_t \sim N(0, \sigma^2)$ are mutually independent.

(a) What is the measurement equation and what is the transition equation in this state space model? Conceptually, how does the approach of extracting a hidden series for GDP_t based on this model compare to taking a simple average of $GDP_{E,t}$ and $GDP_{I,t}$?

(b) Let $s_t = GDP_t$ and $y_t = (GDP_{E,t}, GDP_{I,t})'$. Write a program that implements the Kalman filter. Concretely, given a value for $\theta = (\phi_0, \phi_1, \sigma^2, \sigma_E^2, \sigma_I^2)$, your program should generate the means and variances of the sequences of predicted states, $\{s_{t|t-1}\}_{t=1}^T$, predicted observations $\{y_{t|t-1}\}_{t=1}^T$ and updated states $\{s_{t|t}\}_{t=1}^T$. As a by-product, it should also return the (conditional) likelihood $p(Y_{1:T}|\theta, y_0)$ evaluated at that particular value for θ. You can initialize your Kalman filter by assuming that

$$s_0 = GDP_0 = (GDP_{E,0} + GDP_{I,0})/2 = (1/2, 1/2)y_0$$

is fixed (i.e. its mean is equal to that expresssion and its variance is zero).

Verify that your code works by computing the mean of updated states (i.e. the series of updated estimates of GDP_t) at $\theta = (\phi_0, \phi_1, 1, 1, 1)$ and $\phi = (\phi_0, \phi_1)'$ taken as the estimate from one of your two AR(1) models in (b). Is it close to $GDP_{E,t}$ and $GDP_{I,t}$?

(c) Write a program that implements the Kalman smoother, i.e. given a value for $\theta = (\phi_0, \phi_1, \sigma^2, \sigma_E^2, \sigma_I^2)$, your program should generate the means and variances of smoothed states, $\{s_{t|T}\}_{t=1}^T$.

Again, verify that your code works by computing and plotting the mean of smoothed states (i.e. the series of smoothed estimates of GDP_t) using the same value for θ as in the previous exercise. Is it close to $GDP_{E,t}$ and $GDP_{I,t}$?

(d) Fixing $(\sigma^2, \sigma_E^2, \sigma_I^2) = (1, 1, 1)$, compute the Maximum Likelihood (ML) estimate for ϕ by numerically maximizing the likelihood evaluated by the Kalman filter. How do your estimated mean, variance and autocorrelation of GDP_t compare to the ones obtained for $GDP_{E,t}$ and $GDP_{I,t}$ in (b)?[5]

(e) Fixing $(\sigma^2, \sigma_E^2, \sigma_I^2) = (1, 1, 1)$, compute the Maximum Likelihood (ML) estimate for ϕ using the Expectation Maximization (EM) algorithm. Concretely, initialize ϕ at some value $\hat{\phi}^0$ and iterate for $m = 1, 2, ..., M$ between

- computing $\{\hat{s}_t^m\}_{t=1}^T$, the means of smoothed states $\{s_{t|T}\}_{t=1}^T$ given ϕ, and
- computing $\hat{\phi}^m$, the ML estimate of ϕ taking $\{\hat{s}_t^m\}_{t=1}^T$ as your data for $\{GDP_t\}_{t=1}^T$.

Your final estimate for ϕ is $\hat{\phi} = \hat{\phi}^M$.

You can take $M = 100$ or do these iterations until the values for $\hat{\phi}$ stabilize.[6] How does your estimate $\hat{\phi}$ compare to the one obtained using numerical maximization of the likelihood in (d)?

(f) Taking one of your estimates $\hat{\phi}$ from (d) or (e), compute and plot the means of the predicted, updated and smoothed states at $\theta = (\hat{\phi}_0, \hat{\phi}_1, 1, 1, 1)$. How do these estimates of GDP_t compare to one another? How do they compare to a simple average of $GDP_{E,t}$ and $GDP_{I,t}$?

[5] If needed, you can also focus on the estimation of ϕ_1 and fix ϕ_0 to some sensible value.

[6] Again, if needed, you can also focus on the estimation of ϕ_1 and fix ϕ_0 to some sensible value.

Exercise 10.1: Solution

Consider the following bivariate VAR(1):

$$y_t = \Phi_1 y_{t-1} + u_t \ , \quad u_t = \Phi_\varepsilon \varepsilon_t \ , \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0, I).$$

where $y_t = (w_t, h_t)'$ is composed of log wages w_t and log hours h_t. The vector of structural shocks $\varepsilon_t = (\varepsilon_{a,t}, \varepsilon_{b,t})'$ is composed of a labor demand shock (technology shock) $\varepsilon_{a,t}$ and a labor supply shock (preference shock) $\varepsilon_{b,t}$.

(a) What condition does Φ_1 have to satisfy so that y_t is stationary?

Solution: All Eigenvalues of Φ_1 have to be less than 1 in absolute value.

(b) Suppose y_t is stationary, derive the autocovariances of order zero and one, denoted by $\Gamma_{yy}(0)$ and $\Gamma_{yy}(1)$.

Solution: Clearly, $\mathbb{E}[y_t] = 0$. Therefore, $\Gamma(h) = \mathbb{E}[y_t y_{t-h}']$. There are two ways to obtain $\Gamma(h)$.

Inserting for y_{t-1} in the RHS of the VAR(1) equation, we get

$$\begin{aligned} y_t &= \Phi_1(\Phi_1 y_{t-2} + u_{t-1}) + u_t \\ &= \Phi_1^2 y_{t-2} + u_t + \Phi_1 u_{t-1} \ . \end{aligned}$$

Repeating this same process infinitely many times, we get

$$y_t = \sum_{l=0}^{\infty} \Phi_1^l u_{t-l} \ ,$$

since $\lim_{l\to\infty} \Phi_1^l y_{t-l} = 0$ under stationarity. Using this equation, we get

$$\begin{aligned} \Gamma(h) &= \mathbb{E}[y_t y_{t-h}'] \\ &= \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \Phi_1^l u_{t-l} \right) \left(\sum_{k=0}^{\infty} \Phi_1^l u_{t-h-k} \right)' \right] \\ &= \sum_{l=0}^{\infty} \sum_{k=0}^{\infty} \Phi_1^l \mathbb{E}[u_{t-l} u_{t-h-k}'] \Phi_1^{l\prime} \\ &= \sum_{k=0}^{\infty} \Phi_1^{k+h} \Sigma \Phi_1^{k\prime} \ , \end{aligned}$$

since u_t is a WN processs and, therefore, $\mathbb{E}[u_{t-l} u_{t-h-k}'] = 0$ unless $l = h + k$. We can evaluate this expression for $h = 0$ and $h = 1$. Practically, we have to cut off this infinite

Exercise 10.1: Solution

sum after some finite number of terms, but this has little bearing on our results since, by stationarity, Φ_1^k converges to zero as $k \to \infty$.

Using the second approach, we have

$$\begin{aligned}\Gamma(0) = \mathbb{E}[y_t y_t'] &= \mathbb{E}[(\Phi_1 y_{t-1} + \Phi_\varepsilon \varepsilon_t)(\Phi_1 y_{t-1} + \Phi_\varepsilon \varepsilon_t)'] \\ &= \Phi_1 \Gamma(0) \Phi_1' + \Phi_\varepsilon \Phi_\varepsilon' \ .\end{aligned}$$

Using $vec(ABC) = (A \otimes C')vec(B)$, we have

$$vec(\Gamma(0)) = (I - \Phi_1 \otimes \Phi_1)^{-1} vec(\Phi_\varepsilon \Phi_\varepsilon') \ .$$

The covariance matrices can be obtained using the Yule-Walker equations:

$$\Gamma(\tau) = \mathbb{E}[y_t y_{t-\tau}'] = \mathbb{E}[\Phi_1 y_{t-1} y_{t-\tau}'] + \mathbb{E}[\Phi_\varepsilon \varepsilon_t y_{t-\tau}'] = \Phi_1 \Gamma(\tau - 1) \ \forall \ \tau > 0 \ .$$

(c) Derive the impulse response function of y_{t+h}, $h = 0, 1, \ldots$ with respect to the vector of structural shocks ε_t. How do log wages react to a labor supply shock that occurred 3 periods before?

Solution: Using our derivations above, we have

$$y_{t+h} = \sum_{l=0}^{\infty} \Phi_1^l u_{t+h-l} = \sum_{l=0}^{\infty} \Phi_1^l \Phi_\varepsilon \varepsilon_{t+h-l} \ .$$

Therefore,

$$\frac{\partial y_{t+h}}{\partial \varepsilon_t} = \Phi_1^h \Phi_\varepsilon \ .$$

Since we are interested in the effect of the shock after 3 periods, we have $h = 3$. Also, since w_t is the first element of y_{t+h} and $\varepsilon_{b,t}$ is the second element in ε_t, we seek

$$\frac{\partial w_t}{\partial \varepsilon_{b,t}} = [\Phi_1^3 \Phi_\varepsilon]_{12} \ ,$$

i.e. the element in the first row and second column of the matrix $\Phi_1^3 \Phi_\varepsilon$

(d) Describe the identification problem in the context of this VAR.

Solution: It holds that $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$. When we estimate the reduced-form VAR, $y_t =$

Exercise 10.1: Solution

$\Phi_1 y_{t-1} + u_t$, we only obtain estimates of Φ_1 and $\mathbb{V}[u_t] = \Sigma$. This is not enough to pin down Φ_ε as Σ has $n(n+1)/2$ distinct elements, while Φ_ε has n^2. In our case, we have we 3 elements in Σ and 4 elements in Φ_ε. Therefore, we need (at least) one restriction so that we can identify all 4 elements in Φ_ε.

To see the identification problem more clearly, w.l.o.g. we can write

$$\Phi_\varepsilon = \Sigma_{tr}\Omega \ ,$$

where Σ_{tr} is the lower-triangular Cholesky factor of Σ and Ω is s.t. $\Omega\Omega' = I$. Since Σ is identified by the data, so is Σ_{tr}. However, any Ω that satisfies $\Omega\Omega' = I$ is consistent with the data (and there is a whole continuum of such matrices)!

(e) Supose we are willing to assume that, contemporaneously, hours worked are only affected by preferences, not technology. What restrictions does this assumption impose on Φ_ε? Is this enough to uniquely identify Φ_ε?

Solution: Yes. If h_t is contemporaneously only affected by $\varepsilon_{a,t}$, but not by $\varepsilon_{b,t}$, then this implies that $[\Phi_\varepsilon]_{22} = 0$. Therefore, the matrix Φ_ε has 3 non-zero elements. Since $n = 2$, we have $n(n+1)/2 = 3$ elements in Σ, which is exactly what we need to point-identify the remaining 3 elements in Φ_ε.

Coming back to our decomposition of Φ_ε into Σ_{tr} and Ω (which is a bit of an overkill when there is point-identification), it turns out that our assumption gives us a unique Ω that is consistent with the imposed restriction.

(f) Alternatively, suppose that we assume that a labor supply shock $\varepsilon_{b,t}$ moves wages and hours in opposite directions upon impact, whereas a demand shock $\varepsilon_{a,t}$ moves wages and hours in the same direction. What restrictions does this assumption impose on Φ_ε? Is this enough to uniquely identify Φ_ε?

Solution: Then we only know that Φ_ε has signs $\begin{bmatrix} + & + \\ + & - \end{bmatrix}$ (for example).

In this case we don't have point-identification but set-identification, as there is not a unique Φ_ε, i.e. unique values for each of its parameters, but a whole set of parameter-values in Φ_ε that satisfy the imposed restriction.

Now the decomposition of Φ_ε into Σ_{tr} and Ω is useful. As before (as always), we can perfectly tell Σ_{tr} from the data. In contrast to the identification assumption in the previous exercise, our assumptions here restrict the permissible set of Ωs, but do not perfectly tell us Ω: Ω now needs to satisfy $\Omega\Omega' = I$ as well as the imposed

Exercise 10.1: Solution

sign-restrictions on Φ_ε (which translate into sign-restrictions on Ω since $\Phi_\varepsilon = \Sigma_{tr}\Omega$).

We can think of wages and hours being determined by an interplay of labor supply and labor demand. Let $h_t = \varphi^D(w_t, y_{t-1}; \varepsilon_{a,t}, \varepsilon_{b,t})$ be the demand and $h_t = \varphi^S(w_t, y_{t-1}; \varepsilon_{a,t}, \varepsilon_{b,t})$ the supply function. They show the relationship between hours and wages (i.e. quantity and price in the labor market), whereby these functions (think of labor/supply curves) depend on current technology and preference shocks as well as past hours and wages.[1]

(h) Suppose we assume that the labor demand is only affected by the technology shock, not the preference shock, whereas labor supply is affected by both shocks:

$$h_t = \varphi^D(w_t, y_{t-1}; \varepsilon_{a,t}) \quad \text{and} \quad h_t = \varphi^S(w_t, y_{t-1}; \varepsilon_{a,t}, \varepsilon_{b,t}) \; .$$

Is this enough to uniquely identify Φ_ε?
Hint: remember that demand must equal supply at all times.

Solution: We don't get point-identification as Φ_ε still has 4 unrestricted elements, since w_t and h_t both depend on both shocks. For h_t, that's obvious when you look at the labour supply function. For w_t, you can see this by noting that if you set supply and demand equal and solve for w_t, the resulting function that determines w_t will also depend on both $\varepsilon_{a,t}$ and $\varepsilon_{b,t}$.

(i) Does your answer change if, on top of the above assumption, we assume that

$$\frac{\partial w_t}{\partial \varepsilon_{b,t}} = (\alpha - 1)\frac{\partial h_t}{\partial \varepsilon_{b,t}}$$

for some given α; conditional on α, is it possible to uniquely identify the elements of Φ_ε? If yes, show how you can solve for Φ_ε based on α and the reduced-form VAR parameters.

Solution: Under this additional assumption, we have that

$$\frac{\partial w_t}{\partial \varepsilon_{b,t}} = [\Phi_\varepsilon]_{12} = (1-\alpha)\frac{\partial h_t}{\partial \varepsilon_{b,t}} = (1-\alpha)\,[\Phi_\varepsilon]_{22} \;\; ,$$

where $[\Phi_\varepsilon]_{ij}$ is the (i,j)th element in Φ_ε. This is just a reparameterization of Φ_ε; it still has 4 parameters, which cannot be identified using only the 3 distinct ele-

[1] In a structural economic model for the labor market, we would typically assume exogenous preference and technology processes (not i.i.d., but persistent, e.g. AR(1)s), which then, combined with the households' (workers') and firms' utility and profit maximization problems, lead to such demand and supply functions. Owing to the persistence preference and technology processes, the demand and supply functions will depend on past preference and technology shocks, which are summarized by past hours and wages in y_{t-1}.

Exercise 10.1: Solution

ments/parameters in Σ. However, conditional on α, $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$ forms a system of 3 equations in 3 unknowns, which means that given Σ and α, we can solve for Φ_ε.

Exercise 10.2: Solution

Consider the following bivariate VAR(1)

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} \Sigma_{11}^{tr} & 0 \\ \Sigma_{12}^{tr} & \Sigma_{22}^{tr} \end{bmatrix} \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} ,$$

written more compactly as

$$y_t = \Phi y_{t-1} + \Sigma_{tr}\Omega(\varphi)\varepsilon_t \ , \quad \varepsilon_t \sim N(0, I) \ , \tag{1}$$

where I is the identity matrix.

(a) (3 Points) In the context of the above equation, what is a reduced-form VAR and what is a structural VAR?

Solution: The structural VAR is the specification with uncorrelated shocks $\varepsilon_t \sim N(0, I)$. It can be written as in the equation above,

$$y_t = \Phi y_{t-1} + \Sigma_{tr}\Omega(\varphi)\varepsilon_t \ ,$$

or

$$Ay_t = By_{t-1} + \varepsilon_t \ ,$$

with $A = (\Sigma_{tr}\Omega(\varphi))^{-1}$ and $B = (\Sigma_{tr}\Omega(\varphi))^{-1}\Phi$. **[1.5p]** Instead, the reduced-form VAR contains one-step ahead forecasting errors $u_t \sim N(0, \Sigma)$, which are (typically) not uncorrelated:

$$y_t = \Phi y_{t-1} + u_t \ .\textbf{[1.5p]}$$

The notation is chosen such that Σ_{tr} is the lower-triangular Cholesky factor of Σ and Ω is an orthogonal matrix: $\Omega\Omega' = I$:

$$\Sigma = \mathbb{V}[u_t] = \mathbb{V}[\Sigma_{tr}\Omega(\varphi)\varepsilon_t] = \Sigma_{tr}\Omega I \Omega' \Sigma_{tr}' = \Sigma_{tr}\Sigma_{tr}' \ .$$

(b) (1 Point) Show that $\Omega\Omega' = I$ and $\Omega'\Omega = I$ for any $\varphi \in [-\pi, \pi]$.
Hint: we have $\cos(\varphi)^2 + \sin(\varphi)^2 = 1 \ \forall \ \varphi \in \mathbb{R}$.

Exercise 10.2: Solution

Solution: We have

$$\begin{aligned}\Omega\Omega' &= \begin{bmatrix}\cos(\varphi) & -\sin(\varphi)\\ \sin(\varphi) & \cos(\varphi)\end{bmatrix}\begin{bmatrix}\cos(\varphi) & \sin(\varphi)\\ -\sin(\varphi) & \cos(\varphi)\end{bmatrix}\\ &= \begin{bmatrix}\cos(\varphi)^2+\sin(\varphi)^2 & 0\\ 0 & \cos(\varphi)^2+\sin(\varphi)^2\end{bmatrix},\end{aligned}$$

and $\cos(\varphi)^2 + \sin(\varphi)^2 = 1 \; \forall \; \varphi \in \mathbb{R}$. **[1p]**

(c) (4 Points) Is φ identifiable from the data? Are Φ and Σ_{tr} identifiable? Explain carefully.

Solution: Φ and Σ (and hence Σ_{tr}) are identifiable from the data. They appear in the likelihood function for the reduced-form VAR. **[2p]** In contrast, when writing down the likelihood function for the structural VAR, $\Omega(\varphi)$ does not appear, because $\Sigma_{tr}\Omega\Omega'\Sigma_{tr}' = \Sigma_{tr}\Sigma_{tr}' = \Sigma$ for any φ. Therefore, φ is not identified. **[2p]**

The intuition is as follows. Decomposing the errors u_t into shocks ε_t requires four parameters: we can write $u_t = \Phi_\varepsilon \varepsilon_t$, where Φ_ε is a 2×2 matrix and hence has 4 parameters. W.l.o.g. we can write $\Phi_\varepsilon = \Sigma_{tr}\Omega(\varphi)$, whereby three parameters are in Σ_{tr} and one is in Ω. The variance-covariance matrix Σ of reduced-form errors u_t, which is identified, contains 3 free parameters (as it is symmetric), which are used to determine Σ_{tr}, so that nothing is left to pin down φ.

Essentially, the data tells us the variances of u_{1t} and u_{2t} and their covariance, whereas we would like to know to what extent these three quantities are the result of four forces: shock 1 (ε_{1t}) affecting variable 1 (u_{1t}), shock 2 affecting variable 1, shock 1 affecting variable 2, and shock 2 affecting variable 2.

(d) (4 Points) Suppose the parameters $(\Phi, \Sigma_{tr}, \varphi)$ are given. What is the definition of an impulse response function (IRF) and how would you compute an IRF for a one-standard deviation shock $\varepsilon_{1,t}$.

Solution: The IRF shows the reaction of y_t to a shock in ε_t over time (horizons h):

$$IRF_{ij}^h \equiv \frac{\partial y_{t+h}}{\partial \varepsilon_t} \text{ .}\mathbf{[1p]}$$

Exercise 10.2: Solution

Writing y_{t+h} as a function of past errors and shocks gives

$$\begin{aligned} y_{t+h} &= \Phi^{h+1} y_{t-1} + \sum_{j=0}^{h} \Phi^j u_{t+h-j} \\ &= \Phi^{h+1} y_{t-1} + \sum_{j=0}^{h} \Phi^j \Sigma_{tr} \Omega(\varphi) \varepsilon_{t+h-j} \ .\textbf{[1p]} \end{aligned}$$

Therefore,

$$IRF_{ij}^h \equiv \frac{\partial y_{t+h}}{\partial \varepsilon_t} = \Phi^h \Sigma_{tr} \Omega(\varphi) \ .\textbf{[2p]}$$

(e) (5 Points) Suppose we want to impose that both y_{1t} and y_{2t} respond positively upon impact to a shock ε_{1t}. What restrictions does this assumption impose on φ?
Hint: based on the fact that $\Sigma_{tr}\Sigma_{tr}' = \Sigma$, where Σ is a variance-covariance matrix, you can deduce the signs of Σ_{11}^{tr} and Σ_{22}^{tr}, which allows you to give a more detailed response. Also, see Fig. 1 below.

Solution: The responses at impact are given by

$$\begin{aligned} \frac{\partial y_t}{\partial \varepsilon_t} &= \Sigma_{tr} \Omega(\varphi) \\ &= \begin{bmatrix} \Sigma_{11}^{tr} & 0 \\ \Sigma_{12}^{tr} & \Sigma_{22}^{tr} \end{bmatrix} \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \ . \end{aligned}$$

The responses of $y_t = [y_{1t}, y_{2t}]'$ to ε_{1t} are contained in the first column of this matrix. The assumptions require

$$\begin{aligned} \frac{\partial y_{1,t}}{\partial \varepsilon_{1,t}} &= \Sigma_{11}^{tr} cos(\varphi) > 0 \ , \\ \frac{\partial y_{2,t}}{\partial \varepsilon_{1,t}} &= \Sigma_{12}^{tr} cos(\varphi) + \Sigma_{22}^{tr} sin(\varphi) > 0 \ .\textbf{[2p]} \end{aligned}$$

As Σ_{tr} is the Cholesky factor of Σ, we have $\Sigma_{11}^{tr}, \Sigma_{22}^{tr} > 0$. Therefore, the assumption that y_{1t} responds positively requires $cos(\varphi) > 0$, which implies $\varphi \in (-\pi/2, \pi/2)$. **[1p]**

Given that $cos(\varphi) > 0$, the assumption that y_{2t} responds positively requires

$$\frac{\sin(\varphi)}{\cos(\varphi)} > -\frac{\Sigma_{12}^{tr}}{\Sigma_{22}^{tr}} \ .$$

Exercise 10.2: Solution

If $\Sigma_{21}^{tr} = 0$, we can refine the set of φ consistent with our assumptions to $\varphi \in (0, \pi/2)$. As $\Sigma_{21}^{tr} \rightarrow \infty$, we get no further refinement: $\varphi \in (-\pi/2, \pi/2)$. As $\Sigma_{21}^{tr} \rightarrow -\infty$, we get point identification: $\varphi \rightarrow \pi/2$. The intuition is as follows. $\Sigma_{21}^{tr}\Sigma_{11}^{tr}$ is the (contemporaneous) covariance of reduced form errors u_{1t} and u_{2t}. As $\Sigma_{21}^{tr} \rightarrow \infty$, the two become perfectly positively correlated. Then, imposing that there is a shock that moves both in the same direction does not contain any additional information useful for disentangling u_{1t} and u_{2t} into uncorrelated shocks ε_{1t} and ε_{2t}. As opposed to that, as $\Sigma_{21}^{tr} \rightarrow -\infty$, u_{1t} and u_{2t} become perfectly negatively correlated and our assumption that there is a shock that lets both move in the same direction, despite this negative correlation, allows us to perfectly tell apart the two shocks ε_{1t} and ε_{2t} which underlie the movements in u_{1t} and u_{2t}. **[2p]**

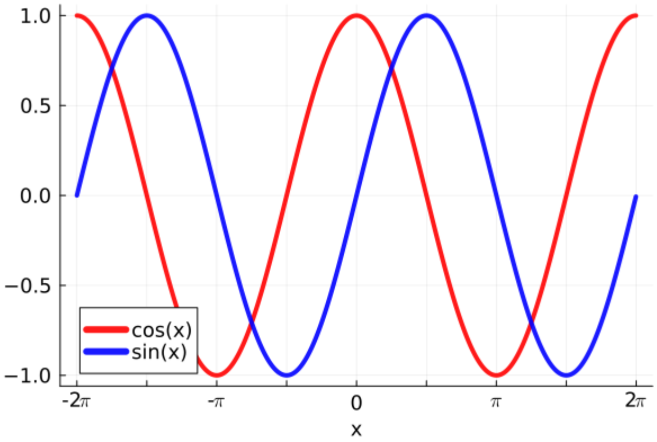


Figure 1: Sine and Cosine Functions

Exercise 10.3: Solution

Let y_t be an $n \times 1$ vector and consider the following VAR(1):

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} N(0, \Sigma) \; .$$

1. Derive the conditional likelihood $p(Y_{1:T}|y_0, \Phi_0, \Phi_1, \Sigma)$ and the ML estimator(s) for (Φ_0, Φ_1, Σ),

$$(\hat{\Phi}_0, \hat{\Phi}_1, \hat{\Sigma}) = \arg \max_{(\Phi_0, \Phi_1, \Sigma)} p(Y_{1:T}|y_0, \Phi_0, \Phi_1, \Sigma) \; .$$

Solution

We can write the VAR(p) in linear regression form as follows. Let $k = np + 1$ and define

$$\underset{(T\times n)}{Y} = [y_1, ..., y_T]' \; , \quad \underset{(T\times n)}{U} = [u_1, ..., u_T]' \; ,$$

as well as

$$\underset{(k\times 1)}{x_t} = [y'_{t-1}, ..., y'_{t-p}, 1]' \; , \quad \underset{(T\times k)}{X} = [x_1, ..., x_T]' \; , \quad \text{and} \quad \underset{(k\times n)}{\Phi} = [\Phi_1, ..., \Phi_p, \Phi_0]' \; .$$

Then

$$y'_t = x'_t \Phi + u'_t \; , \quad \text{and} \quad Y = X\Phi + U \; .$$

Under normality of u_t, we have the conditional density

$$\begin{aligned} p(y_t|Y_{t-p:t-1}, \Phi, \Sigma) &= (2\pi)^{-n/2}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(y'_t - x'_t\Phi)\Sigma^{-1}(y'_t - x'_t\Phi)'\right\} \\ &= (2\pi)^{-n/2}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(y'_t - x'_t\Phi)'(y'_t - x'_t\Phi)\right]\right\} \; , \end{aligned}$$

where the second line uses the fact that $a'Ba = tr[Baa']$. We get the conditional likelihood

$$\begin{aligned} p(Y_{1:T}|Y_{-p+1:0}, \Phi, \Sigma) &= \prod_{t=1}^{T} p(y_t|Y_{t-p:t-1}, \Phi, \Sigma) \\ &= (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\sum_{t=1}^{T}\frac{1}{2}tr\left[\Sigma^{-1}(y'_t - x'_t\Phi)'(y'_t - x'_t\Phi)\right]\right\} \\ &= (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\sum_{t=1}^{T}\Sigma^{-1}(y'_t - x'_t\Phi)'(y'_t - x'_t\Phi)\right]\right\} \\ &= (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(Y - X\Phi)'(Y - X\Phi)\right]\right\} \; , \end{aligned}$$

using the facts that $tr[A + B] = tr[A] + tr[B]$ and $\sum_{t=1}^{T}(y'_t - x'_t\Phi)'(y'_t - x'_t\Phi) = (Y - X\Phi)'(Y - X\Phi)$. Maximizing this expression by taking derivatives gives the ML estimators

$$\hat{\Phi} = (X'X)^{-1}X'Y \; , \quad \hat{\Sigma} = \frac{1}{T}(Y - X\hat{\Phi})'(Y - X\hat{\Phi}) \; .$$

Exercise 10.3: Solution

The derivations use the rules $tr[A] = tr[A']$, $tr[A+B] = tr[A] + tr[B]$ as well as

$$\frac{\partial tr[AXB]}{\partial X} = A'B' \ , \quad \frac{\partial tr[X'BXC]}{\partial X} = BXC{+}B'XC' \ , \quad \frac{\partial tr[AX^{-1}B]}{\partial X} = -(X^{-1}BAX^{-1})' \ , \quad \frac{\partial ln|X|}{\partial X} = (X')^{-1} \ .$$

2. Download quarterly observations on US real GDP growth and inflation from FRED[1] for the period 1985:I to 2019:IV. Estimate a VAR(1) with intercept by OLS. Conditional on the OLS estimate, compute the eigenvalues of Φ_1, and the unconditional mean and variance of y_t. Compare the unconditional mean and variance implied by the estimated VAR to sample mean and variance of y_t. Would you expect them to be approximately the same? Are they approximately the same?

   Solution I use the annualized rate both for inflation and GDP growth. Inflation is computed using the GDP deflator. The following results look different if CPI inflation is used (you'd get Eigenvalues of 0.37 and 0.07).

```
gdp <- diff(log(getSymbols("GDPC1", src = "FRED", auto.assign = FALSE)))  * 400
# cpi <- getSymbols("CPIAUCSL", src = "FRED", auto.assign = FALSE)
# # use last observations in the quarter to compute inflation rate
# cpi <- apply.quarterly(cpi, last)
# # adjust the time label
# month(index(cpi)) <- lubridate::month(index(cpi)) - 2
# # compute annualized inflation
# infl <- diff(log(cpi)) * 4
infl <- diff(log(getSymbols("GDPDEF", src= "FRED", auto.assign = FALSE))) *400
data <- merge(gdp, infl)
colnames(data) <- c("gdp", "infl")
data <- data["1985-01-01/2019-10-01"]
data <- unclass(data)
```

```
computeVAR1Mean <- function(modelVAR1) {
  means <- solve(diag(ncol(modelVAR1$F1Hat)) - modelVAR1$F1Hat,
                 modelVAR1$F0Hat)
  colnames(means) <- "mean"
  return(means)
}

computeVAR1Variance <- function(modelVAR1) {
  n <- ncol(modelVAR1$F1Hat)
  vecGamma0 <- solve(diag(n ^ 2) - kronecker(modelVAR1$F1Hat, modelVAR1$F1Hat),
                     c(modelVAR1$Sigma))
  Gamma0 <- matrix(vecGamma0, n, n)
  colnames(Gamma0) <- rownames(Gamma0) <- colnames(modelVAR1$F1Hat)
  return(Gamma0)
}
```

[1] This is the FRED database of the Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org. Use the series "GDPC1" for the former and the series "GDPDEF" or "CPIAUCSL" for the latter.

Exercise 10.3: Solution

```
bigT <- nrow(data) - 1
Y <- data[2:(bigT + 1), ]
X <- cbind(1, data[1:(bigT), ])
bigK <- ncol(X)
bigD <- ncol(Y)
modelVAR1 <- list()
modelVAR1$FHat <- solve(t(X) %*% X, t(X) %*% Y)
modelVAR1$F0Hat <- t(modelVAR1$FHat[1, , drop = FALSE])
modelVAR1$F1Hat <- t(modelVAR1$FHat[-1, ])
uHat <- Y -  X %*% modelVAR1$FHat
modelVAR1$S <- t(uHat) %*% uHat
modelVAR1$Sigma <- modelVAR1$S / (bigT - bigK)
```

Estimation results:

	gdp	infl
intercept	1.9401	0.5706
gdp	0.3713	0.0806
infl	-0.1353	0.6214

Eigenvalues are far from one:

```
## [1] 0.57 0.43
```

Mean implied by the model and sample mean are very close:

	model	sample
gdp	2.6408	2.6433
infl	2.0693	2.1146

Variance implied by the model:

	gdp	infl
gdp	5.1694345	0.1748524
infl	0.1748524	0.9200014

Sample variance:

	gdp	infl
gdp	5.0688538	0.1885877
infl	0.1885877	0.9388674

Again, the variance implied by the model is very close to the sample variance.

Exercise 10.3: Solution

3. Derive the MNIW-posterior under the Normal likelihood above and the improper prior

$$p(\Phi, \Sigma) \propto |\Sigma|^{-(n+1)/2} \,.$$

Solution

The (conditional) likelihood for a reduced-form VAR is

$$p(Y|\Phi, \Sigma) = (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(Y - X\Phi)'(Y - X\Phi)\right]\right\} \,.$$

First, we derive the conditional posterior of $\Phi|\Sigma$. The prior pdf of $\Phi|\Sigma$ is $p(\Phi|\Sigma) \propto c$, i.e. it is proportional to a constant. This must be the case because $p(\Phi|\Sigma) = p(\Phi, \Sigma)/p(\Sigma)$, whereby we know that the joint distribution $p(\Phi, \Sigma)$ is not a function of Φ, and the marginal $p(\Sigma) = \int p(\Phi, \Sigma)d\Phi$ can never be a function of Φ. An alternative way to see that is to note that, by Bayes' rule, conditional distributions are proportional to joint distributions. Hence, $p(\Phi|\Sigma) \propto p(\Phi, \Sigma) = |\Sigma|^{-(n+1)/2} \propto c$.

In turn, we get

$$\begin{aligned}
p(\Phi|Y, \Sigma) &= \frac{p(Y|\Phi, \Sigma)p(\Phi|\Sigma)}{p(Y|\Sigma)} \propto p(Y|\Phi, \Sigma)p(\Phi|\Sigma) \\
&\propto p(Y|\Phi, \Sigma)p(\Phi|\Sigma) \\
&\propto exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(Y - X\Phi)'(Y - X\Phi)\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(\Phi'X'X\Phi - 2\Phi'X'Y)\right]\right\} \,.
\end{aligned}$$

Defining $\bar{P} = X'X$ and $\bar{P}\bar{\mu} = X'Y$ – which yields $\bar{\mu} = \bar{P}^{-1}\left[X'Y\right] = (X'X)^{-1}X'Y = \hat{\Phi}$ –, this expression looks analogous to that for the pdf of a Matrix-Normal distribution $MN(\bar{\mu}, \bar{P}^{-1}, \Sigma)$, from which we conclude that $\Phi|Y, \Sigma \sim MN(\bar{\mu}, \bar{P}^{-1}, \Sigma)$.

Second, we invert Bayes' formula above to find

$$\begin{aligned}
p(Y|\Sigma) &= \frac{p(Y|\Phi, \Sigma)p(\Phi|\Sigma)}{p(\Phi|Y, \Sigma)} \\
&= (2\pi)^{-nT/2}(2\pi)^{nk/2}|\Sigma|^{-T/2}|\bar{P}|^{-n/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}M\right]\right\} \,,
\end{aligned}$$

where $M = Y'Y - \bar{\mu}'\bar{P}\bar{\mu}$. This expression is obtained by first inserting the full expressions for the pdfs $p(Y|\Phi, \Sigma)$ and $p(\Phi|Y, \Sigma)$ on the RHS and then cancelling all terms that involve Φ, as $p(Y|\Sigma)$ is not a function of Φ.

Third, we use $p(Y|\Sigma)$ to derive the marginal posterior of Σ. Note that the prior pdf of Σ is $p(\Sigma) \propto |\Sigma|^{-(n+1)/2}$. This is because

$$p(\Sigma) = \int p(\Phi, \Sigma)d\Phi = \int c|\Sigma|^{-(n+1)/2}d\Phi = |\Sigma|^{-(n+1)/2}\int c d\Phi \propto |\Sigma|^{-(n+1)/2} \,.$$

Exercise 10.3: Solution

For the posterior, we then get

$$\begin{aligned} p(\Sigma|Y) &\propto p(Y|\Sigma)p(\Sigma) \\ &\propto |\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}M\right]\right\}|\Sigma|^{-(n+1)/2} \\ &= |\Sigma|^{-\frac{T+n+1}{2}} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}M\right]\right\} , \end{aligned}$$

from which we can deduce that $\Sigma|Y \sim IW(\bar{S}, \bar{\nu})$ with $\bar{\nu} = T$ and $\bar{S} = M$.

Overall, we get the MNIW-posterior

$$\Phi|Y, \Sigma \sim MN(\bar{\mu}, \bar{P}^{-1}, \Sigma) , \quad \Sigma|Y \sim IW(\bar{S}, \bar{\nu}) ,$$

Note that

$$M = Y'Y - Y'X(X'X)^{-1}(X'X)(X'X)^{-1}X'Y = Y'\left[I - X'(X'X)^{-1}X'\right]Y = Y'M_XY = (M_XY)'(M_XY) = \hat{U}'\hat{U} ,$$

where $M_X = I - P_X$ and $P_X = X'(X'X)^{-1}X'$ are projection matrices with the property that $M_X = M_X'$ and $M_XM_X = M_X$ (and analogously for P_X), and $M_XY = Y - P_XY = Y - X\hat{\Phi} = Y - \hat{Y} = \hat{U}$.

Hence, much more intuitively, we have the posterior

$$\Phi|Y, \Sigma \sim MN(\hat{\Phi}, (X'X)^{-1}, \Sigma) , \quad \Sigma|Y \sim IW(\hat{U}'\hat{U}, T) .$$

As usual, the Matrix-Normal distribution $\Phi|Y, \Sigma \sim MN(\hat{\Phi}, (X'X)^{-1}, \Sigma)$ is equivalent to the multivariate (vector-) Normal

$$vec(\Phi)|\Sigma, Y \sim N\left(vec(\hat{\Phi}), \Sigma \otimes (X'X)^{-1}\right) .$$

4. Write a program that generates draws $m = 1, \ldots, M$ from the posterior $p(\Phi, \Sigma|Y)$ you derived using direct sampling from the MNIW distribution. (If you were unable to derive the posterior in the previous exercise, take the posterior under the improper prior $p(\Phi, \Sigma) \propto c$ shown in the script.) For each parameter in Φ and Σ, compute the posterior mean and a 90% credible interval. Tabulate your posterior mean estimates along with the OLS estimates.

Solution

```
bigN <- 10000
SigmaDraws <- array(0, dim = c(bigD, bigD, bigN))
PhiDraws <- array(0, dim = c(bigK, bigD, bigN))
XXinv <- solve(t(X) %*% X)
for (n in 1:bigN) {
  SigmaDraws[ , , n] <- MCMCpack::riwish(v = bigT - bigK, S = modelVAR1$S)
  PhiDraws[ , , n] <- matrix(MASS::mvrnorm(1, c(modelVAR1$FHat),
                                           kronecker(SigmaDraws[ , , n], XXinv)),
                             bigK, bigD)
}
```

Poserior means of the parameter estimates are very close to the OLS estimates which is unsurprising

Exercise 10.3: Solution

given that the true posterior mean is the OLS estimator.

```
postMean <- apply(PhiDraws, 1:2, mean)
lb <- apply(PhiDraws, 1:2, function(x) quantile(x, 0.05))
ub <- apply(PhiDraws, 1:2, function(x) quantile(x, 0.95))
tbl <- cbind(c(postMean), c(lb), c(ub), c(modelVAR1$FHat))
colnames(tbl) <- c("mean", "LB", "UB", "OLS")
rownames(tbl) <- paste0("Phi_", apply(expand.grid(1:3, 1:2), 1,
                                      function(x) paste0(x[1], ",", x[2])))
knitr::kable(tbl, digits = 4)
```

	mean	LB	UB	OLS
Phi_1,1	1.9383	1.1576	2.7134	1.9401
Phi_2,1	0.3725	0.2401	0.5054	0.3713
Phi_3,1	-0.1360	-0.4439	0.1719	-0.1353
Phi_1,2	0.5675	0.3023	0.8347	0.5706
Phi_2,2	0.0807	0.0356	0.1262	0.0806
Phi_3,2	0.6225	0.5173	0.7278	0.6214

Similarly, for the variance:

```
postMean <- apply(SigmaDraws, 1:2, mean)
lb <- apply(SigmaDraws, 1:2, function(x) quantile(x, 0.05))
ub <- apply(SigmaDraws, 1:2, function(x) quantile(x, 0.95))
tbl <- cbind(c(postMean), c(lb), c(ub), c(modelVAR1$Sigma))
colnames(tbl) <- c("mean", "LB", "UB", "OLS")
rownames(tbl) <- paste0("Sigma_", apply(expand.grid(1:2, 1:2), 1,
                                      function(x) paste0(x[1], ",", x[2])))
knitr::kable(tbl, digits = 7)
```

	mean	LB	UB	OLS
Sigma_1,1	4.5625699	3.7214462	5.5593632	4.4573603
Sigma_2,1	0.0602507	-0.1594769	0.2840465	0.0590917
Sigma_1,2	0.0602507	-0.1594769	0.2840465	0.0590917
Sigma_2,2	0.5259174	0.4306395	0.6404686	0.5136647

5. For each posterior draw m, compute the lower-triangular Cholesky factor Σ_{tr} and compute Impulse Response Functions (IRFs) based on $\Phi_\epsilon = \Sigma_{tr}$. Concretely, for each draw m, you should get an IRF_{ij}^h for the four pairs (i,j) and for the horizons $h = 1:12$. Then, for each (i,j) pair, generate a plot with h on the x-axis and, for each h, the mean as well as the 90% Bayesian credible set of IRF_{ij}^h on the y-axis. (To obtain the 90% credible set, you may assume that the distribution of IRF_{ij}^h is symmetric).

   Show the result in a single figure with 2×2 subplots. Comment on your plots. Which identification assumption is embodied in taking $\Phi_\epsilon = \Sigma_{tr}$? How does that assumption manifest itself in your plots?

Exercise 10.3: Solution

Solution

Using the identification scheme given by the Cholesky factorization gives the following relationship between reduced form errors ε_t and structural innovations e_t:

$$\varepsilon_t = \Sigma_{tr} e_t \ .$$

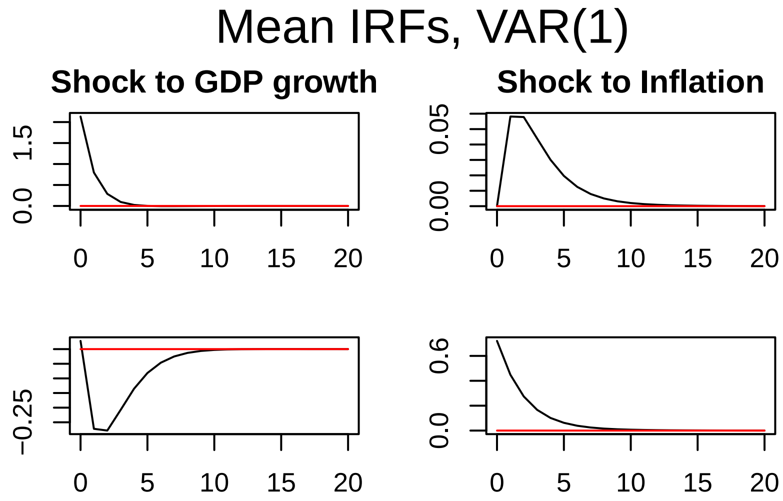
The IRF at horizon j is then given by

$$\frac{\partial y_{t+j}}{\partial e_t} = \Phi_1^j \Sigma_{tr} \ .$$

Below I plot the mean responses across all draws.

```r
H <- 20

IRFs <- array(0, dim = c(bigD, bigD, H+1, bigN))

for (nn in 1:bigN){
  Sigma_tr <- t(chol(SigmaDraws[,,nn])) # transpose to get lower- not upper-triangular part

  Phi1 <- PhiDraws[-1,,nn]

  IRFs[,,1,nn] <- Sigma_tr

  for (h in 1:H){
    IRFs[,,h+1,nn] <- Phi1 %*% IRFs[,,h,nn]
  }
}

meanIRFs <- apply(IRFs,1:3,mean)

par(mfrow = c(2,2),oma = c(2,2,2,2),mar = c(2,2,2,2))
plot(0:H,meanIRFs[1,1,],type='l',xlab="h",ylab="GDP growth",main="Shock to GDP growth")
lines(0:H,rep(0,H+1),type='l',col='red')
plot(0:H,meanIRFs[1,2,],type='l',xlab="h",ylab="",main="Shock to Inflation")
lines(0:H,rep(0,H+1),type='l',col='red')
plot(0:H,meanIRFs[2,1,],type='l',xlab="h",ylab="Inflation")
lines(0:H,rep(0,H+1),type='l',col='red')
plot(0:H,meanIRFs[2,2,],type='l',xlab="h",ylab="")
lines(0:H,rep(0,H+1),type='l',col='red')
mtext("Mean IRFs, VAR(1)", outer = TRUE, cex = 1.5)
```

Exercise 10.3: Solution



```
par(mfrow = c(1, 1))
```

6. How do the IRFs change if you re-estimate the VAR with $p = 4$ lags?

   Solution

   In order to obtain the IRFs for a VAR(p) with $p > 1$, it's easiest to write the model in companion form as a VAR(1):

$$X_t = F_c + F_1 X_{t-1} + F_\varepsilon v_t \ ,$$

where $X_t = [y_t', y_{t-1}', ..., y_{t-p+1}']'$, $v_t = [e_t', 0, ..., 0]'$ are $np \times 1$ vectors, and

$$F_c = \begin{bmatrix} \Phi_c \\ 0 \\ ... \\ 0 \end{bmatrix} \ , \quad F_1 = \begin{bmatrix} \Phi_1 & \Phi_2 & ... & \Phi_{p-1} & \Phi_p \\ I & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & I & 0 \end{bmatrix} \ , \quad F_\varepsilon = \begin{bmatrix} \Phi_\varepsilon & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{bmatrix} \ ,$$

where $\Phi_\varepsilon = \Sigma_{tr}$ in our case. Analogously to before, the IRF of X_t is given by

$$\frac{\partial X_{t+j}}{\partial v_t} = F_1^j F_\varepsilon \ .$$

The IRF of y_t can then be obtained using the $n \times np$ selection matrix $M = [I, 0, ..., 0]$:

$$\frac{\partial y_{t+j}}{\partial e_t} = \frac{\partial y_{t+j}}{\partial X_{t+j}} \frac{\partial X_{t+j}}{\partial v_t} \frac{\partial v_t}{\partial e_t} = M F_1^j F_\varepsilon M' \ .$$

```
    fObtainXandY <- function(mData,p){
      n <- ncol(mData)
      bigT <- nrow(mData)-p

      k <- n*p + 1

      X <- matrix(NA,bigT,k)
      for (pp in 1:p){
```

Exercise 10.3: Solution

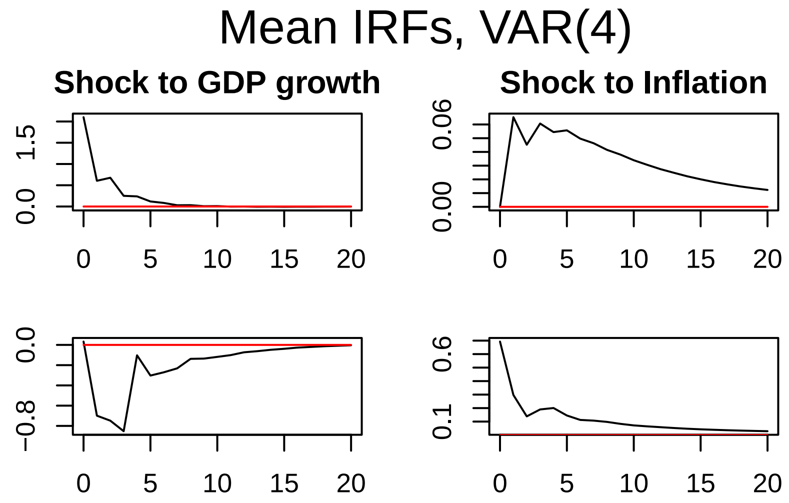
```
    X[,(pp-1)*n +(1:n)] <- mData[(p+1-pp):(bigT+p-pp),]
  }
  X[,k] <- rep(1,bigT)
  return(list(X=X,Y=mData[(p+1):(bigT+p),]))
}

p <- 4


VARobjects <- fObtainXandY(data,p)
Y <- VARobjects$Y
X <- VARobjects$X
bigT <- nrow(Y)
bigK <- ncol(X)
bigD <- ncol(Y)
modelVAR <- list()
modelVAR$FHat <- solve(t(X) %*% X, t(X) %*% Y)
uHat <- Y -  X %*% modelVAR$FHat
modelVAR$S <- t(uHat) %*% uHat

SigmaDraws <- array(0, dim = c(bigD, bigD, bigN))
PhiDraws <- array(0, dim = c(bigK, bigD, bigN))
XXinv <- solve(t(X) %*% X)
for (n in 1:bigN) {
  SigmaDraws[ , , n] <- MCMCpack::riwish(v = bigT - bigK, S = modelVAR$S)
  PhiDraws[ , , n] <- matrix(MASS::mvrnorm(1, c(modelVAR$FHat),
                                           kronecker(SigmaDraws[ , , n], XXinv)),
                             bigK, bigD)
}

fCompanionForm <- function(mPhi,mSig){

  k <- nrow(mPhi)
  n <- ncol(mPhi)
  p <- (k-1)/n

  vPhiC <- mPhi[k,]
  mFc <- matrix(0,p*n,1)
  mFc[1:n] <- vPhiC

  Sigma_tr <- t(chol(mSig))
  mFeps <- matrix(0,p*n,p*n)
  mFeps[1:n,1:n] <- Sigma_tr
```

Exercise 10.3: Solution

```
  mF <- matrix(0,p*n,p*n)
  mF[(n+1):(n*p),1:(n*p-n)] <- diag(rep(1,n*p-n))

  for (pp in 1:p){
    mF[1:n,(pp-1)*n+(1:n)] <- mPhi[(pp-1)*n+(1:n),]
  }

  mM <- matrix(0,n,p*n)
  mM[1:n,1:n] <- diag(rep(1,n))

  return(list(Fc=mFc,F=mF,Feps=mFeps,M=mM))
}


IRFs <- array(0, dim = c(bigD, bigD, H+1, bigN))
IRFsForX <- array(0, dim = c(bigD*p, bigD*p, H+1, bigN))

for (nn in 1:bigN){

  companionFormObjects <- fCompanionForm(PhiDraws[,,nn],SigmaDraws[,,nn])

  mF1 <- companionFormObjects$F
  mM <- companionFormObjects$M
  mFeps <- companionFormObjects$Feps



  IRFsForX[,,1,nn] <- mFeps
  IRFs[,,1,nn] <- mM %*% IRFsForX[,,1,nn] %*% t(mM)

  for (h in 1:H){
    IRFsForX[,,h+1,nn] <- mF1 %*% IRFsForX[,,h,nn]
    IRFs[,,h+1,nn] <- mM %*% IRFsForX[,,h+1,nn] %*% t(mM)
  }
}

meanIRFs <- apply(IRFs,1:3,mean)

par(mfrow = c(2,2),oma = c(2,2,2,2),mar = c(2,2,2,2))
plot(0:H,meanIRFs[1,1,],type='l',xlab="h",ylab="GDP growth",main="Shock to GDP growth")
lines(0:H,rep(0,H+1),type='l',col='red')
plot(0:H,meanIRFs[1,2,],type='l',xlab="h",ylab="",main="Shock to Inflation")
lines(0:H,rep(0,H+1),type='l',col='red')
plot(0:H,meanIRFs[2,1,],type='l',xlab="h",ylab="Inflation")
```

Exercise 10.3: Solution

```
lines(0:H,rep(0,H+1),type='l',col='red')
plot(0:H,meanIRFs[2,2,],type='l',xlab="h",ylab="")
lines(0:H,rep(0,H+1),type='l',col='red')
mtext("Mean IRFs, VAR(4)", outer = TRUE, cex = 1.5)
```



```
par(mfrow = c(1, 1))
```

Exercise 10.4: Solution

The following exercises are inspired by Aruoba, Diebold, Nalewaik, Schorfheide & Song (2016): "Improving GDP measurement: A measurement-error perspective," *Journal of Econometrics,* 191, 384-397.

A country's GDP can be measured in two ways: i) by measuring the *expenditures* of households, firms and governments on products and services produced in that country during a specific period, or ii) by measuring the *income* obtained by households, firms and governments from products and services produced in that country during a specific period.[1] Even though in theory, these two measures should coincide, in practice they do not exactly. In the case of the US, the former is somewhat confusingly referred to as "GDP", while the latter is referred to as "GDI" (see for example the website of the Bureau of Economic Analysis (BEA): https://www.bea.gov/data/economic-accounts/national).

1. Download the two quarterly series "GDPC1" (real gross domestic product) and "A261RX1Q020SBEA" (real gross domestic income) from FRED[2] for the period 1979:I to 2023:IV, and compute their respective growth rates relative to the same quarter in the previous year (giving you two series starting from 1980:I). We will call the former $GDP_{E,t}$ and the latter $GDP_{I,t}$. Plot the two series as well as their difference.

Solution

```
GDPE <- getSymbols("GDPC1", src = "FRED", auto.assign = FALSE)
GDPI <- getSymbols("A261RX1Q020SBEA", src = "FRED", auto.assign = FALSE)

fMyTimeRange <- function(x,sStartDate,sEndDate){ x[paste(sStartDate,sEndDate,sep="/")] }

sStartDate = "1979-01-01"
sEndDate = "2023-12-01"
GDPE <- fMyTimeRange(GDPE$GDPC1,sStartDate,sEndDate)
GDPI <- fMyTimeRange(GDPI$A261RX1Q020SBEA,sStartDate,sEndDate)

T <- length(GDPE)

vGDPEgr <- diff(log(GDPE),4)[5:T] * 100
vGDPIgr <- diff(log(GDPI),4)[5:T] * 100

sStartDate <- "1980-01-01"

mData <- data.frame("date"=seq(as.Date(sStartDate),to=as.Date(sEndDate),by="quarter"),
            "GDPE"=vGDPEgr,
            "GDPI"=vGDPIgr)
colnames(mData) <- c("date","GDPE","GDPI")

dat <- melt(mData, id="date")

lMyPlotOptions <- list(theme_bw(),
               theme(aspect.ratio=5/8),
```

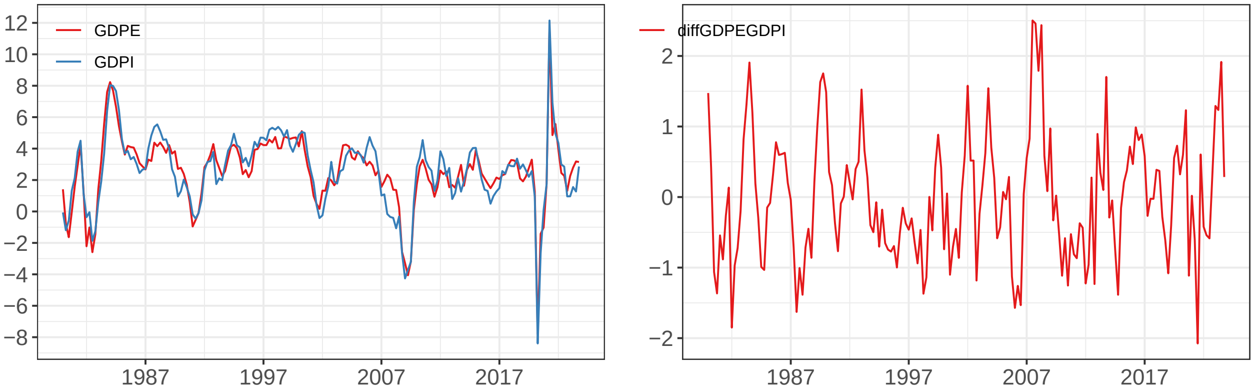
[1] The latter includes not only labor income, but also income from the supply of other productive factors.

[2] This is the FRED database of the Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org.

Exercise 10.4: Solution

```
               theme(legend.background = element_rect(fill = "transparent"), legend.justification=c(1,1
        scale_colour_brewer(palette="Set1"),
        labs(x="", y=""),
        scale_x_date(date_labels = "%Y",date_breaks="10 years",limits = as.Date(c(sStartDate, NA)))
        )

plot1 <- ggplot(dat, aes(x=date, y=value, group=variable, colour=variable)) +
        geom_line(aes(colour=variable)) +
        scale_y_continuous(breaks=seq(-20,20,2)) +
        lMyPlotOptions



mData2 <- data.frame(mData$date,mData$GDPE-mData$GDPI)
colnames(mData2) <- c("date","diffGDPEGDPI")
dat <- melt(mData2, id="date")

plot2 <- ggplot(dat, aes(x=date, y=value)) +
        geom_line(aes(colour=variable)) +
        scale_y_continuous(breaks=seq(-20,20,1)) +
        lMyPlotOptions
```

```
ppp <- (plot1 | plot2 )

ppp
```



2. Estimate an AR(1) with intercept for each of the two series. Comment on the estimated means, variances and autocorrelations of the two series.

Solution

```
fComputeAR1 <- function(vSeries){

    ar1 <- lm(vSeries ~ lag(vSeries,1))

    print(summary(ar1))

```

Exercise 10.4: Solution

```
    vPhis <- coefficients(ar1)
    phi0 <- vPhis[1]
    phi1 <- vPhis[2]

    sigma2 <- mean(ar1$residuals^2)

    print(paste("Mean = ",phi0/(1-phi1) ))

    print(paste("Variance = ",sigma2/(1-phi1^2) ))

    print(paste("Autocorrelation Coefficient = ",phi1 ))

    return(ar1)

}

ar1_GDPE <- fComputeAR1(vGDPEgr)
```

```
##
## Call:
## lm(formula = vSeries ~ lag(vSeries, 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4195 -0.4488  0.0876  0.5421  9.5721
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.67810    0.17194   3.944 0.000116 ***
## lag(vSeries, 1)  0.74098    0.05098  14.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.465 on 173 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5498, Adjusted R-squared:  0.5472
## F-statistic: 211.2 on 1 and 173 DF,  p-value: < 2.2e-16
##
## [1] "Mean =  2.61797105918274"
## [1] "Variance =  4.70630673621827"
## [1] "Autocorrelation Coefficient =  0.740981264850859"
```

```
ar1_GDPI <- fComputeAR1(vGDPIgr)
```

```
##
## Call:
```

Exercise 10.4: Solution

```
## lm(formula = vSeries ~ lag(vSeries, 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7387 -0.5849  0.0349  0.6656 10.2058
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.60543    0.16783   3.608 0.000404 ***
## lag(vSeries, 1)  0.77360    0.04774  16.204  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.494 on 173 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6028, Adjusted R-squared:  0.6005
## F-statistic: 262.6 on 1 and 173 DF,  p-value: < 2.2e-16
##
## [1] "Mean =  2.67413090789856"
## [1] "Variance =  5.49477794629578"
## [1] "Autocorrelation Coefficient =  0.773597557646432"
```

The means and autocorrelations are reasonably close one to another. However, $GDP_{I,t}$ has a noticeably higher variance than $GDP_{E,t}$.

3. Consider the first model from the mentioned paper, intended to extract some underlying true GDP growth series, GDP_t, based on data on $GDP_{E,t}$ and $GDP_{I,t}$:

$$\begin{bmatrix} GDP_{E,t} \\ GDP_{I,t} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} GDP_t + \begin{bmatrix} \epsilon_{E,t} \\ \epsilon_{I,t} \end{bmatrix} \ ,$$
$$GDP_t = \phi_0 + \phi_1 GDP_{t-1} + \epsilon_t \ ,$$

where $\epsilon_{E,t} \sim N(0, \sigma_E^2)$, $\epsilon_{I,t} \sim N(0, \sigma_I^2)$ and $\epsilon_t \sim N(0, \sigma^2)$ are mutually independent.

3a. What is the measurement equation and what is the transition equation in this state space model? Conceptually, how does the approach of extracting a hidden series for GDP_t based on this model compare to taking a simple average of $GDP_{E,t}$ and $GDP_{I,t}$?

Solution

The first equation is the measurement equation, as it relates the observed/measured $GDP_{E,t}$ and $GDP_{I,t}$ to the unobserved/hidden GDP_t series. The second equation is the transition equation. It shows how the hidden state evolves over time.

Running a filtering or smoothing algorithm for this state space model means will return the series GDP_t that best describes the dynamics of two observed series $GDP_{E,t}$ and $GDP_{I,t}$, so that the leftover parts, the error terms $\epsilon_{E,t}$ and $\epsilon_{I,t}$, are not autocorrelated, but any autocorrelation of both $GDP_{E,t}$ and $GDP_{I,t}$ is due to the autocorrelation of GDP_t. Intuitively, one can imagine this as taking a weighted average of the observed

Exercise 10.4: Solution

series, where the weights are dynamic and optimized so as to best describe the observed series by the stated measurement equation.

Note that filtering and smoothing is done conditional on the parameters ϕ_0, ϕ_1, σ_E^2, σ_I^2 and σ^2. By estimating these parameters, we have more leeway in finding the underlying series that best accounts for the dynamics of the observed series, and, even more importantly, we attach a very precise meaning to "finding the hidden states that best describe the dynamics of observables": we treat the hidden states as parameters as well and we find the values of parameters (including hidden states) that maximize the likelihood of observing our data or we find the posterior distribution of parameters (including hidden states) by updating our prior with the likelihood.

3b. Let $s_t = GDP_t$ and $y_t = (GDP_{E,t}, GDP_{I,t})'$. Write a program that implements the Kalman filter. Concretely, given a value for $\theta = (\phi_0, \phi_1, \sigma^2, \sigma_E^2, \sigma_I^2)$, your program should generate the means and variances of the sequences of predicted states, $\{s_{t|t-1}\}_{t=1}^T$, predicted observations $\{y_{t|t-1}\}_{t=1}^T$ and updated states $\{s_{t|t}\}_{t=1}^T$. As a by-product, it should also return the (conditional) likelihood $p(Y_{1:T}|\theta, y_0)$ evaluated at that particular value for θ. You can initialize your Kalman filter by assuming that

$$s_0 = GDP_0 = (GDP_{E,0} + GDP_{I,0})/2 = (1/2, 1/2)y_0$$

is fixed (i.e. its mean is equal to that expression and its variance is zero).

Verify that your code works by computing the mean of updated states (i.e. the series of updated estimates of GDP_t) at $\theta = (\phi_0, \phi_1, 1, 1, 1)$ and $\phi = (\phi_0, \phi_1)'$ taken as the estimate from one of your two AR(1) models in (b). Is it close to $GDP_{E,t}$ and $GDP_{I,t}$?

Solution

```
# Generic Kalman Filtering function:

fKalmanFilter <- function(mA,mB,mH,mC,mD,mQ,mY,mPupd0=NA,vXupd0=NA){

    # Y_t = mA + mB X_t + N(0,mH)
    # X_t = mC + mD X_{t-1} + N(0,mQ)

    # Note: all arguments must be matrices! i.e. code scalars as matrix(scalar,1,1)

    nY <- nrow(mA)
    nX <- nrow(mC)
    p <- nX/nY
    T <- nrow(mY)

    mXpr <- matrix(0,T,nX) # predicted states
    mXupd <- matrix(0,T+1,nX)     # updated states

    aPpr <- array(0,dim=c(T,nX,nX)) # variance of predicted states
    aPupd <- array(0,dim=c(T+1,nX,nX)) # variance of updated states

    mYpr <- matrix(0,T,nY) # predicted data
```

Exercise 10.4: Solution

```
vLL <- matrix(0,T,1) # ll increments


# Initialize state

if (is.null(dim(mPupd0))){
    aPupd[1,,] <- matrix(solve(diag(1,nX^2)-kronecker(mD,mD)) %*% as.vector(mQ),nrow=nX,ncol=nX)
}else{
    aPupd[1,,] <- mPupd0
}

if (is.null(dim(vXupd0))){
    mXupd[1,] <- mC
}else{
    mXupd[1,] <- vXupd0
}


for (tt in 1:T){

    ttt <- tt + 1

    #compute the one-period-ahead prediction (mean) and covariance matrix of the latent state vecto

    mXpr[tt,] <- t( mC + mD %*% mXupd[ttt-1,] )
    aPpr[tt,,] <- mD %*% aPupd[ttt-1,,] %*% t(mD) + mQ

    #compute the one-period-ahead prediction (mean) and covaraince matrix of the observed variables

    mYpr[tt,] <- mA + mB %*% mXpr[tt,]
    mF <- mB %*% aPpr[tt,,] %*% t(mB) + mH

    # compute updated states, i.e. mean and covariance matrix:

    vFerr <- matrix( as.double(mY[tt,]) - mYpr[tt,] , nY,1) #forecast error

    mHelper <- aPpr[tt,,] %*% t(mB) %*% solve(mF)

    mXupd[ttt,] <- mXpr[tt,] + mHelper %*% vFerr
    aPupd[ttt,,] <- aPpr[tt,,] - mHelper %*% mB %*% aPpr[tt,,]

    #Likelihood increment:
    vLL[tt] <- -0.5*log(abs(det(mF)))-0.5*t(vFerr)%*%solve(mF)%*%vFerr
```

Exercise 10.4: Solution

```
    }

    return(list('mXpr'=mXpr,'aPpr'=aPpr,'mXupd'=mXupd,'aPupd'=aPupd,'mYpr'=mYpr,'vLL'=vLL))

}


fMyKalmanFiltering <- function(vTheta){

    phi0 <- vTheta[1]
    phi1 <- vTheta[2]
    sig2 <- vTheta[3]
    sig2E <- vTheta[4]
    sig2I <- vTheta[5]

    s0 <- matrix( (mData[1,2] + mData[1,3]) / 2 , 1,1)
    s0_var <- matrix(0,1,1)

    mA <- matrix(0,nrow=2,ncol=1)
    mB <- matrix(1,nrow=2,ncol=1)

    mH <- diag(c(sig2E,sig2I))

    mC <- matrix(phi0,1,1)
    mD <- matrix(phi1,1,1)

    mQ <- matrix(sig2,1,1)


    lOutput <- fKalmanFilter(mA,mB,mH,mC,mD,mQ,mY=mData[,-1],mPupd0=s0_var,vXupd0=s0)

    return(lOutput)

}


vMyThetaTest <- c(ar1_GDPE$coefficients[1],ar1_GDPE$coefficients[2],1,1,1)

lOutput <- fMyKalmanFiltering(vMyThetaTest)

mean(lOutput$mXupd)
```

```
## [1] 2.580788
```

Exercise 10.4: Solution

```
mean(vGDPEgr)
```

```
## [1] 2.582899
```

```
mean(vGDPIgr)
```

```
## [1] 2.601775
```

Yes, the mean is reasonably close to those of the observed series.

3c. Write a program that implements the Kalman smoother, i.e. given a value for $\theta = (\phi_0, \phi_1, \sigma^2, \sigma_E^2, \sigma_I^2)$, your program should generate the means and variances of smoothed states, $\{s_{t|T}\}_{t=1}^T$.

Again, verify that your code works by computing and plotting the mean of smoothed states (i.e. the series of smoothed estimates of GDP_t) using the same value for θ as in the previous exercise. Is it close to $GDP_{E,t}$ and $GDP_{I,t}$?

Solution

```
# Generic Kalman Smoother:

fKalmanSmoother <- function(mXpr,aPpr,mXupd,aPupd,mD){

    T <- nrow(mXpr)
    nX <- ncol(mXpr)

    mXsm <- matrix(NA,T,nX)
    aPsm <- array(NA,dim=c(T,nX,nX))

    mXsm[T,] <- mXupd[T+1,]
    aPsm[T,,] <- aPupd[T+1,,]

    for (tt in c((T-1):1)){
        ttt = tt+1
        mJ <- aPupd[ttt,,] %*% t(mD) %*% solve(aPpr[tt+1,,])
        mXsm[tt,] <- mXupd[ttt,] + mJ %*% (mXsm[tt+1,]-mXpr[tt+1,])
        aPsm[tt,,] <- aPupd[ttt,,] + mJ %*% (aPsm[tt+1,,]-aPpr[tt+1,,]) %*% t(mJ)
    }

    return(list('mXsm'=mXsm,'aPsm'=aPsm))

}


fMyKalmanSmoothing <- function(vTheta){

    phi1 <- vTheta[2]

    mD <- matrix(phi1,1,1)
```

Exercise 10.4: Solution

```
    lOutputKF <- fMyKalmanFiltering(vTheta)

    lOutputKS <- fKalmanSmoother(mXpr=lOutputKF$mXpr,aPpr=lOutputKF$aPpr,mXupd=lOutputKF$mXupd,aPupd=lOu

    return(lOutputKS)

}

lOutput <- fMyKalmanSmoothing(vMyThetaTest)

mean(lOutput$mXsm)
```

```
## [1] 2.593504
```

```
mean(vGDPEgr)
```

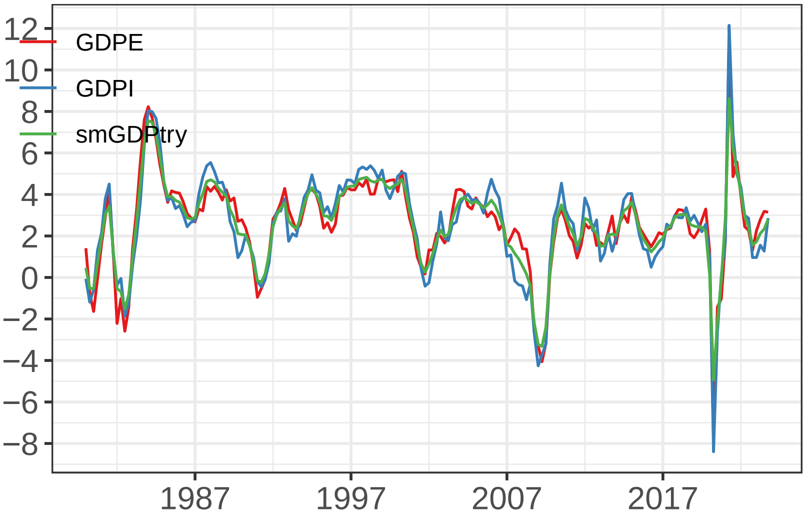
```
## [1] 2.582899
```

```
mean(vGDPIgr)
```

```
## [1] 2.601775
```

```
mData <- cbind(mData,lOutput$mXsm)
colnames(mData)[ncol(mData)] <- "smGDPtry"

dat <- melt(mData, id="date")
plot3 <- ggplot(dat, aes(x=date, y=value, group=variable, colour=variable)) +
        geom_line(aes(colour=variable)) + theme_bw() + theme(aspect.ratio=5/8) +
        theme(legend.background = element_rect(fill = "transparent"), legend.justification=c(1,1), lege
        scale_colour_brewer(palette="Set1") + labs(x="", y="") +
        scale_y_continuous(breaks=seq(-8,14,2), limits=c(min(mData[,-1]),max(mData[,-1]))) +
        scale_x_date(date_labels = "%Y",date_breaks="10 years",limits = as.Date(c(sStartDate, NA)))

plot3
```

Exercise 10.4: Solution



```
mData <- mData[,-ncol(mData)] #delete added "trial" series again
```

The mean of the smoothed series is also close to the means of the observed series. Also, the smoothed series looks as expected; it closely tracks the evolutions of the observed series, but it is smoother than the latter.

3d. Fixing $(\sigma^2, \sigma^2_E, \sigma^2_I) = (1, 1, 1)$, compute the Maximum Likelihood (ML) estimate for ϕ by numerically maximizing the likelihood evaluated by the Kalman filter. How do your estimated mean, variance and autocorrelation of GDP_t compare to the ones obtained for $GDP_{E,t}$ and $GDP_{I,t}$ in (b)?[3]

Solution

```
fMyLL <- function(vPhis){

    if (abs(vPhis[2]) > 0.99){

        return(-Inf)

    }else{

        vTheta <- c(vPhis,1,1,1)

        lOutputKF <- fMyKalmanFiltering(vTheta)

        return(sum(lOutputKF$vLL))

    }

}

fMyMinusLL <- function(vPhis){

    return(-fMyLL(vPhis))
```

[3] If needed, you can also focus on the estimation of ϕ_1 and fix ϕ_0 to some sensible value.

Exercise 10.4: Solution

```
}

vPhiInit <- vMyThetaTest[1:2]

estPhis <- optim(vPhiInit, fMyMinusLL, method = "Nelder-Mead")

vPhisMLE_num <- estPhis$par

phi0 <- vPhisMLE_num[1]
phi1 <- vPhisMLE_num[2]

print(paste("MLE phi0 = ",phi0))
```

```
## [1] "MLE phi0 =  0.449302089156763"
```

```
print(paste("MLE phi1 = ",phi1))
```

```
## [1] "MLE phi1 =  0.830699834460211"
```

```
print(paste("LL at MLE = ",-estPhis$value))
```

```
## [1] "LL at MLE =  -256.177936071001"
```

```
sig2 <- 1

print(paste("Mean = ",phi0/(1-phi1) ))
```

```
## [1] "Mean =  2.65387861685917"
```

```
print(paste("Variance = ",sig2/(1-phi1^2) ))
```

```
## [1] "Variance =  3.22645397982172"
```

```
print(paste("Autocorrelation Coefficient = ",phi1 ))
```

```
## [1] "Autocorrelation Coefficient =  0.830699834460211"
```

The mean of GDP_t is close to the ones implied by the AR(1) models estimated on each of the observed series. The variance is noticeably smaller, as expected, given that part of the variance of the observed series is picked up by the error terms $\varepsilon_{E,t}$ and $\varepsilon_{I,t}$. The autocorrelation coefficient is also a bit higher than that of $GDP_{E,t}$ and $GDP_{I,t}$. This is also expected and due to the presence of the error terms, because if we add iid noise to an autocorrelated series, the autocorrelation of the resulting series will be smaller. (Intuitively, think of an OLS estimation with measurement error in the covariates, where the coefficient becomes biased towards zero).

3e. Fixing $(\sigma^2, \sigma_E^2, \sigma_I^2) = (1, 1, 1)$, compute the Maximum Likelihood (ML) estimate for ϕ using the Expectation Maximization (EM) algorithm. Concretely, initialize ϕ at some value $\hat{\phi}^0$ and iterate for $m = 1, 2, ..., M$ between

- computing $\{\hat{s}_t^m\}_{t=1}^T$, the means of smoothed states $\{s_{t|T}\}_{t=1}^T$ given ϕ, and
- computing $\hat{\phi}^m$, the ML estimate of ϕ taking $\{\hat{s}_t^m\}_{t=1}^T$ as your data for $\{GDP_t\}_{t=1}^T$.

Exercise 10.4: Solution

Your final estimate for ϕ is $\hat{\phi} = \hat{\phi}^M$.

You can take $M = 100$ or do these iterations until the values for $\hat{\phi}$ stabilize.[4] How does your estimate $\hat{\phi}$ compare to the one obtained using numerical maximization of the likelihood in (d)?

Solution

```
fMyEM <- function(vPhis){

    myTol <- 0.001

    vPhisOld <- vPhis

    while (TRUE){

        vTheta <- c(vPhisOld,1,1,1)

        lOutput <- fMyKalmanSmoothing(vTheta)

        vStates <- lOutput$mXsm

        vStates <- xts(x=vStates,order.by=mData$date)

        ar1 <- lm(vStates ~ lag(vStates,1))

        vPhisNew <- coefficients(ar1)

        vDiff <- abs(vPhisNew-vPhisOld)
        maxDiff <- max(vDiff)

        if (maxDiff < myTol){
            return(vPhisNew)
        }else{
            vPhisOld <- vPhisNew
        }

    }

}


vPhisMLE_EM <- fMyEM(vPhiInit)

phi0 <- vPhisMLE_EM[1]
phi1 <- vPhisMLE_EM[2]
```

[4] Again, if needed, you can also focus on the estimation of ϕ_1 and fix ϕ_0 to some sensible value.

Exercise 10.4: Solution

```
print(paste("MLE phi0 = ",phi0))
```

```
## [1] "MLE phi0 =  0.329872463229866"
```

```
print(paste("MLE phi1 = ",phi1))
```

```
## [1] "MLE phi1 =  0.87805404048036"
```

```
print(paste("LL at MLE = ",fMyLL(vPhisMLE_EM)))
```

```
## [1] "LL at MLE =  -256.936683983894"
```

```
sig2 <- 1

print(paste("Mean = ",phi0/(1-phi1) ))
```

```
## [1] "Mean =  2.70507087343667"
```

```
print(paste("Variance = ",sig2/(1-phi1^2) ))
```

```
## [1] "Variance =  4.36640986915244"
```

```
print(paste("Autocorrelation Coefficient = ",phi1 ))
```

```
## [1] "Autocorrelation Coefficient =  0.87805404048036"
```

It is quite a bit different than the estimates obtained by brute-force numerical likelihood maximization, but roughly in the same ballpark.

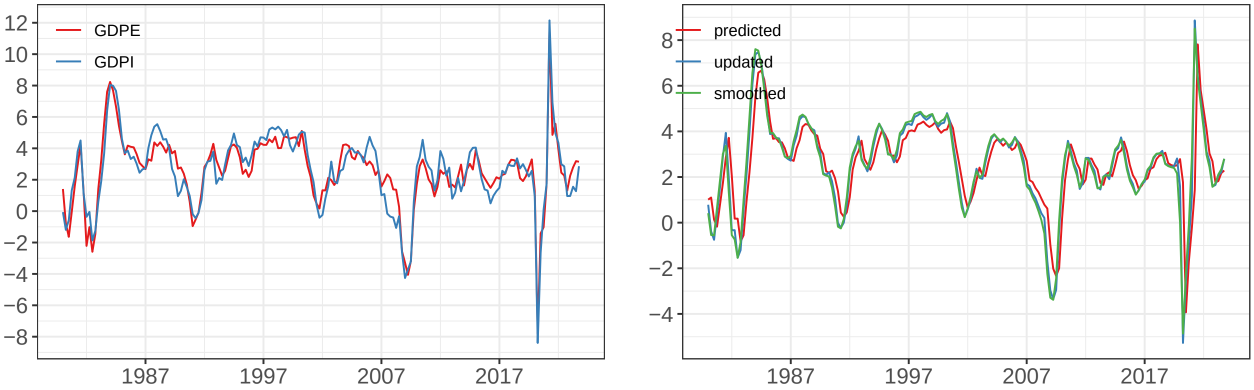
3f. Taking one of your estimates $\hat{\phi}$ from (d) or (e), compute and plot the means of the predicted, updated and smoothed states at $\theta = (\hat{\phi}_0, \hat{\phi}_1, 1, 1, 1)$. How do these estimates of GDP_t compare to one another? How do they compare to a simple average of $GDP_{E,t}$ and $GDP_{I,t}$?

Solution

```
vTheta <- c(vPhisMLE_num,1,1,1)

lOutputKF <- fMyKalmanFiltering(vTheta)

lOutputKS <- fMyKalmanSmoothing(vTheta)


mData <- data.frame(mData$date,lOutputKF$mXpr,lOutputKF$mXupd[2:length(lOutputKF$mXupd)],lOutputKS$mXsm

colnames(mData) <- c("date","predicted","updated","smoothed")

dat <- melt(mData, id="date")

plot4 <- ggplot(dat, aes(x=date, y=value, group=variable, colour=variable)) +
    geom_line(aes(colour=variable)) +
    scale_y_continuous(breaks=seq(-20,20,2)) +
```

Exercise 10.4: Solution

```
    lMyPlotOptions



ppp <- (plot1 | plot4 )

ppp
```



The updated and the smoothed states are virtually indistinguishable. They are both a bit more volatile than the predicted states. All three series are smoother than the observed series but track the latter closely.

11 Further Topics in Time-Series Econometrics

12 Panel Data Analysis

12.1 Panel Data Analysis: Basics (IPP, RE and FE-W)

Suppose you are interested in the determinants of kids' performance in high school and in particular its relation to the student-teacher-ratio and poverty rate. To investigate, you use a dataset of $n = 3306$ school districts in Brazil, followed over $T = 7$ years. You set up the following regression:

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it} \ ,$$

where y_{it} is the average math score in district i at time t, α_i denotes unobserved district-level heterogeneity, and x_{it} contains overall enrolment (number of students enrolled in high schools in district i), the enrolment per teacher, school expenditures and the fraction of citizens in district i below some defined poverty line.

(a) (5 Points) What is (are) the incidental parameter(s) and the incidental parameters problem (IPP) in this setting?

(b) (3 Points) Would the IPP change if you added time-fixed effects (time-dummies) to the above regression? Concretely, consider the regression

$$y_{it} = \alpha_i + \sum_{s=1}^{T} \delta_s \ \mathbf{1}\left\{t = s\right\} + x_{it}'\beta + u_{it} \ ,$$

where $\mathbf{1}\left\{t = s\right\}$ is an indicator variable equal to one if $t = s$ and zero otherwise.

(c) (6 Points) Derive the Random Effects (RE) estimator, $\hat{\beta}_{RE}$. How does it deal with the IPP? Is its core assumption reasonable in this particular example?

(d) (4 Points) Derive the Fixed Effects - Within (FE-W) estimator, $\hat{\beta}_{FE-W}$. How does it deal with the IPP?

(e) (3 Points) What is the difference between strictly and sequentially exogenous regressors?

(f) (6 Points) Derive the probability limit of $\hat{\beta}_{FE-W}$. Is it consistent under sequential exogeneity? In this concrete example, is $\hat{\beta}_{FE-W}$ consistent? Why (not)?

(g) (6 Points) Suppose that β is estimated using the FE-W estimator $\hat{\beta}_{FE-W}$, and suppose that all conditions for consistency of $\hat{\beta}_{FE-W}$ are satisfied. Also, let $u_{it} \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Propose an estimator for σ^2. Is it consistent?

(h) (4 Points) Propose another approach to estimate β consistently.

12.2 Dynamic Panel Data Regression

Consider the panel data regression

$$y_{it} = \alpha_i + \beta_1 \tilde{x}_{it} + \beta_2 y_{i,t-1} + u_{it} = \alpha_i + x_{it}'\beta + u_{it} \ ,$$

where $\mathbb{E}[\tilde{x}_{it} u_{is}] = 0 \ \forall \ t, s$, $\mathbb{E}[\tilde{x}_{it}\alpha_i] = 0 \ \forall \ t$, and $\mathbb{E}[u_{it}^2|\tilde{x}_{it}, y_{i,t-1}] = 1$. Suppose you have $T = 3$ time periods and $n = 10'000$ cross-sectional units.

(a) Can you estimate $\beta = (\beta_1, \beta_2)$ consistently by applying (pooled) OLS to the above regression model? Why (not)? And what if you assumed Normality of u_{it} and estimated the model by ML?

(b) How would you proceed under the Random Effects paradigm? Under what conditions can you consistently estimate β in this model?

(c) Consider the Fixed Effects First Difference transformation

$$\Delta y_{it} = \beta_1 \Delta\tilde{x}_{it} + \beta_2 \Delta y_{i,t-1} + \Delta u_{it} \ .$$

Find $\hat{\beta}_{FD}$, the (pooled) OLS estimator of this transformed model. Is it consistent?

(d) Consider using $y_{i,t-2}$ as an IV for $\Delta y_{i,t-1}$. Find the GMM-IV estimator $\hat{\beta}_{IV,GMM}$. Can you use its asymptotic variance to conduct reliable inference on β?

(e) Suppose you have a consistent estimator $\hat{\beta}$. Provide an estimator for α_i. Is it consistent? Can you think of an alternative estimator with a lower variance?

12.3 Applied Panel Data Analysis: Griliches & Mairesse I

This problem is based on Griliches and Mairesse (1995, NBER Working Paper 5067, "Production Functions: The Search for Identification"). You can download the data from Moodle either in ASCII format GMdata.RAW or in Stata format GMdata.dta. There are nine variables: *index* (firm ID), *sic3* (3 digit SIC), *yr* (year $\in \{73, 78, 83, 88\}$), *ldsal* (log of deflated sales), *lemp* (log of employment), *ldnpt* (log of deflated capital), *ldrnd* (log of deflated R&D), *ldinv* (log of deflated investment). Consider the model

$$ldsal_{it} = \alpha_i + \beta_1 lemp_{it} + \beta_2 ldnpt_{it} + u_{it}.$$

(a) Compute cross-sectional summary statistics (by year) for the following variables: *ldsal*, *lemp*, *ldnpt*. For each year and each variable report: mean, median, standard deviation, minimum, maximum, 5th percentile, 95th percentile. Generate box plots, one for each variable, placing the years next to each other. Do you see a time trend?

(b) Now, let's create a balanced panel and eliminate firms for which you don't have observations for all four years. How many firms do you loose?

(c) Compute the Random Effects (RE) estimator of $\beta = (\beta_1, \beta_2)'$, i.e., estimate a pooled OLS regression of $ldsal_{it}$ on $lemp_{it}$ and $ldnpt_{it}$ along with an intercept, putting α_i into the error term. State the assumptions needed for consistency of the RE estimator. Are they likely to hold?

(d) Compute the Fixed Effects (FE) Within (FE-W) estimator of β. State the assumptions needed for its consistency. Are they likely to hold?

(e) Compute the FE First Difference (FE-FD) estimator of β. State the assumptions needed for its consistency. Are they likely to hold?

(f) Derive the asymptotic distribution of the FE-W estimator.

(g) Compute the standard errors (i.e. estimates of the standard deviations) of your FE-W estimates of β_1 and β_2. You can base your calculations on the asymptotic variance you derived in the previous exercise, or you can use a command from a software package as long as you can make sure it is based on an appropriate formula.[1]

(h) Compute the standard errors for your FE-W estimates also based on clustered bootstrapping. This is analogous to classical bootstrapping, but, to get a valid panel dataset, you only draw cross-sectional units (firms) with replacement, and for the drawn firms you take all the time periods. Set the number of bootstrap samples $B = 1000$ and take a sample size of n (your actual sample size).
Hint: To facilitate your coding process, first consider a single bootstrap sample, then verify your code works for $B = 10$ or $B = 100$, and only once you solved the whole problem set, take $B = 1000$, as it might take a long time to execute.

(i) Now, instead of creating a balanced panel as in (b), use the full data set (an unbalanced panel)

[1] Pay attention not to use any simpler formula that assumes homoskedasticity and/or no serial correlation. In addition, built-in commands in most statistical packages may have finite-sample degree-of-freedom-adjustment terms (in the style of dividing by $(n-1)$ instead of n in the sample variance). While these do not matter asymptotically, in finite sample you may see differences. If you would like to code by hand such finite-sample adjustment terms, see Chapter 17.12 of the Hansen textbook.

to re-compute the FE-W estimator in (d).

(j) Compute the standard error estimates for your FE-W estimates in (i) based on clustered bootstrap.

12.4 Applied Panel Data Analysis: Griliches & Mairesse II

This is a continuation of the last problem set, based on Griliches and Mairesse (1995, NBER Working Paper 5067, "Production Functions: The Search for Identification"). You can download the data from Moodle either in ASCII format GMdata.RAW or in Stata format GMdata.dta. There are nine variables: *index* (firm ID), *sic3* (3 digit SIC), *yr* (year $\in \{73, 78, 83, 88\}$), *ldsal* (log of deflated sales), *lemp* (log of employment), *ldnpt* (log of deflated capital), *ldrnd* (log of deflated R&D), *ldinv* (log of deflated investment). Compared to the last exercise, we add time fixed effects:

$$ldsal_{it} = \alpha_i + \beta_1 lemp_{it} + \beta_2 ldnpt_{it} + \beta_3 ldrst_{it} + \sum_{\tau=1}^{T} f_t \, \mathbf{1}\{t = \tau\} + u_{it} \; , \tag{12.1}$$

where $\mathbf{1}\{t = \tau\}$ for $\tau = 1 : T$ are time-dummies.

(a) Compare the modeling of unit- and time-fixed effects in Eq. (12.1). How does the addition of time-fixed effects relate to the incidental parameters problem (IPP)?

(b) Construct a balanced panel dataset, and take first differences of Eq. (12.1). What happens to α_i and the year dummies?

(c) Suppose for a moment that you added a dummy for the computer industry to the model: $\delta d357_{it}$, where $d357_{it}$ is a dummy variable equal to one if firm i belongs to industry (3-digit SIC code) 357. Does this have an effect on the IPP? How does implied specification for first differences change? Do your conclusions change if we consider time-varying effects of the industry-dummy, $\delta_t d357_{it}$?

(d) Using the balanced panel, re-compute the FE-FD estimator, the FE-W estimator, the RE estimator and the respective standard error estimates.

(e) Conduct the Hausman test, comparing the RE- and FE-W-estimators. What do you conclude? *Hint: note that you need the estimated covariances of the RE- and FE-W-estimators to set up this hypothesis-test.*

(f) Is there support for the null hypothesis of constant returns to scale, $\mathscr{H}_0 : \beta_1 + \beta_2 = 1$? *Hint: note that you need the estimated covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ to set up this hypothesis-test.*

Exercise 12.1: Solution

Suppose you are interested in the determinants of kids' performance in high school and in particular its relation to the student-teacher-ratio and poverty rate. To investigate, you use a dataset of $n = 3306$ school districts in Brazil, followed over $T = 7$ years. You set up the following regression:

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it} \ ,$$

where y_{it} is the average math score in district i at time t, α_i denotes unobserved district-level heterogeneity, and x_{it} contains overall enrolment (number of students enrolled in high schools in district i), the enrolment per teacher, school expenditures and the fraction of citizens in district i below some defined poverty line.

(a) (5 Points) What is (are) the incidental parameter(s) and the incidental parameters problem (IPP) in this setting?

Solution: Because we have a panel data analysis with a large n and small T **[1p]**, the α_is are the indicental parameters. There are n of them; one for each i. **[1p]**

The IPP refers to the fact that we cannot estimate α_i consistently unless $T \to \infty$ **[2p]** (and even β we cannot estimate consistently without further assumptions). Under T fixed and $n \to \infty$ – the asymptotics relevant to the present analysis –, the number of parameters explodes.

The IPP is illustrated best when we attempt at estimating the above equation "directly" by OLS. The model can be written as

$$y_{it} = \sum_{j=1}^{n} \alpha_j \ \mathbf{1}\{i = j\} + x_{it}'\beta + u_{it} \ ,$$

where $\mathbf{1}\{i = j\}$ is a dummy for observation j. Therefore, we might be tempted to estimate β and $\{\alpha_j\}_{j=1}^n$ by regressing y_{it} on x_{it} and n unit-dummies. However, this does not yield consistent estimators since our usual theorems for analyzing consistency do not work for the case where the number of parameters increases with n. **[1p]**

(b) (3 Points) Would the IPP change if you added time-fixed effects (time-dummies) to the above regression? Concretely, consider the regression

$$y_{it} = \alpha_i + \sum_{s=1}^{T} \delta_s \ \mathbf{1}\{t = s\} + x_{it}'\beta + u_{it} \ ,$$

where $\mathbf{1}\{t = s\}$ is an indicator variable equal to one if $t = s$ and zero otherwise.

Solution: No, the inclusion of time-fixed effect dummies neither ameliorates nor at-

Exercise 12.1: Solution

tenuates the IPP. **[2p]** Still, we have that the number of parameters goes to infinity as $n \to \infty$. The parameters in front of the time-dummies, $\{\delta_s\}_{s=1}^T$, do not represent a problem; as $n \to \infty$ and for T fixed, we could, in principle, estimate them consistently (in absence of the incidental parameters, $\{\alpha_j\}_{j=1}^n$). **[1p]**

(c) (6 Points) Derive the Random Effects (RE) estimator, $\hat{\beta}_{RE}$. How does it deal with the IPP? Is its core assumption reasonable in this particular example?

Solution: The key assumption underlying the RE estimator is that $\mathbb{E}[\alpha_i|x_{it}] = 0$ for all t, i.e. that the entity-specific time-invariant heterogeneity α_i is uncorrelated with the regressors x_{it} at every point in time t. **[1.5p]**

The RE approach deals with the IPP by putting α_i into the error term. This can is justified by the above assumption. **[1.5p]** In this case, the model becomes:

$$y_{it} = x_{it}'\beta + v_{it} \quad \text{or} \quad y_i = X_i\beta + v_i \ ,$$

where $v_{it} = \alpha_i + u_{it}$. We can estimate this regression by (pooled) OLS: we minimize the sum of squared residuals

$$\min_\beta \sum_{i=1}^n \sum_{t=1}^T u_{it}^2 = \min_\beta \sum_{i=1}^n (y_i - X_i\beta)'(y_i - X_i\beta) \ ,$$

which leads to the RE(-POLS) estimator

$$\hat{\beta}_{RE-POLS} = \left(\sum_{i=1}^n X_i'X_i\right)^{-1} \sum_{i=1}^n X_i'y_i \ . \quad [\mathbf{1.5p}]$$

The RE estimator is consistent as long as $\mathbb{E}[v_{it}|x_{it}] = 0$, which in turn implies that $\mathbb{E}[u_{it}|x_{it}, \alpha_i] = 0$ for all t and that $\mathbb{E}[\alpha_i|x_{it}] = 0$ for all t. The latter assumption is likely to be violated, as α_i likely contains determinants of school performance that are correlated with, e.g., poverty, such as crime. **[1.5p]**

(d) (4 Points) Derive the Fixed Effects - Within (FE-W) estimator, $\hat{\beta}_{FE-W}$. How does it deal with the IPP?

Solution: The FE-Within estimator deals with the IPP by eliminating the parameters α_i by subtracting, for each unit i and time period t, the average of the model equation over time from the original model equation (i.e. by deriving the implied model for time-demeaned variables based on the original model equation). **[2p]**

Exercise 12.1: Solution

In this case, the model becomes:

$$\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{u}_{it} \quad \text{or} \quad \ddot{y}_i = \ddot{X}_i\beta + \ddot{u}_i \ ,$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and analogously for $\ddot{x}_{it}$ and $\ddot{u}_{it}$. Applying (pooled) OLS to this equation leads to the FE-W estimator:

$$\hat{\beta}_{FE-W} = \left(\sum_{i=1}^{n} \ddot{X}_i'\ddot{X}_i\right)^{-1} \sum_{i=1}^{n} \ddot{X}_i'\ddot{y}_i \ . \quad \textbf{[2p]}$$

(e) (3 Points) What is the difference between strictly and sequentially exogenous regressors?

Solution: Consider an individual regressor x_{it}^k in x_{it}. Strict exogeneity requires $\mathbb{E}[u_{it}|x_{i,1:T}^k] = 0$ for each regressor k (and each unit i and each time period t). In other words, u_{it} needs to be uncorrelated with past, current as well as future values of all regressors **[1.5p]**. In contrast, sequential exogeneity requires u_{it} only to be uncorrelated with past and current regressors: $\mathbb{E}[u_{it}|x_{i,1:t}^k] = 0$ (again for each regressor k). **[1.5p]**

(f) (6 Points) Derive the probability limit of $\hat{\beta}_{FE-W}$. Is it consistent under sequential exogeneity? In this concrete example, is $\hat{\beta}_{FE-W}$ consistent? Why (not)?

Solution: We can write the FE-W estimator as

$$\hat{\beta}_{FE-W} = \left(\frac{1}{n}\sum_{i=1}^{n} \ddot{X}_i'\ddot{X}_i\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \ddot{X}_i'\ddot{y}_i \ .$$

Inserting the true model for $\ddot{y}_i = \ddot{X}_i\beta + \ddot{u}_i$ yields

$$\hat{\beta}_{FE-W} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} \ddot{X}_i'\ddot{X}_i\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \ddot{X}_i'\ddot{u}_i \ .$$

Provided that we have n i.i.d. observations of y_i and X_i, we can apply the WLLN to get

$$\hat{\beta}_{FE-W} - \beta \xrightarrow{p} \mathbb{E}\left[\ddot{X}_i'\ddot{X}_i\right]^{-1} \mathbb{E}\left[\ddot{X}_i'\ddot{u}_i\right] \ . \quad \textbf{[2p]}$$

Exercise 12.1: Solution

The latter term can be written out as follows:

$$\mathbb{E}\left[\ddot{X}_i'\ddot{y}_i\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ddot{x}_{it}\ddot{u}_{it}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\ddot{x}_{it}\ddot{u}_{it}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\left(x_{it} - \frac{1}{T}\sum_{t=1} x_{it}\right)\left(u_{it} - \frac{1}{T}\sum_{t=1} u_{it}\right)\right] .$$

This term is zero under strict exogeneity, but not (necessarily) under sequential exogeneity. Hence, $\hat{\beta}_{FE-W}$ is consistent under strict exogeneity, but not (necessarily) under sequential exogeneity. **[2p]**

Note that strict exogeneity is violated as soon as we have some feedback loop by which some lagged outcome $y_{i,t-l}$ – which is by definition a function of the corresponding, lagged error term $u_{i,t-l}$ – affects the present covariates x_{it}. In our concrete example, we cannot exclude a feedback loop from math scores on enrolment and student-teacher ratio (as students/parents move to districts with better schools), or on school expenditures and number of teachers per student (well performing schools might get additional funding, or on the contrary poorly performing schools might get funding from the Government). These compromising feedback loops could be investigated with additional data or information about the institutional setting. **[2p]**

(g) (6 Points) Suppose that β is estimated using the FE-W estimator $\hat{\beta}_{FE-W}$, and suppose that all conditions for consistency of $\hat{\beta}_{FE-W}$ are satisfied. Also, let $u_{it} \overset{i.i.d.}{\sim} N(0,\sigma^2)$. Propose an estimator for σ^2. Is it consistent?

Solution: The regression with demeaned variables gives rise to error terms $\ddot{u}_{it}$ with

$$\mathbb{V}[\ddot{u}_{it}] = \mathbb{V}\left[u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it}\right] = \mathbb{V}\left[\left(1 - \frac{1}{T}\right)u_{it} - \frac{1}{T}\sum_{s=1,s\neq t}^{T} u_{is}\right] = \frac{T-1}{T}\sigma^2 . \quad \textbf{[2p]}$$

Therefore, we can construct the estimator

$$\hat{\sigma}^2 = \frac{T}{T-1}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\hat{\ddot{u}}_{it}^2 ,$$

where $\hat{\ddot{u}}_{it} = \ddot{y}_{it} - \ddot{x}_{it}\hat{\beta}$. **[2p]** Provided that $\hat{\beta}$ is consistent, by the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\ddot{u}}_{it}^2 \xrightarrow{p} \mathbb{E}[\ddot{u}_{it}^2] ,$$

Exercise 12.1: Solution

and so $\hat{\sigma}^2$ is consistent:

$$\hat{\sigma}^2 \to \frac{1}{T-1} \sum_{t=1}^{T} \mathbb{E}[\ddot{u}_{it}^2] = \frac{T}{T-1} \mathbb{V}[\ddot{u}_{it}] = \sigma^2 \ .$$

Provided that strict exogeneity holds, we might take $\hat{\beta}_{FE-W}$. **[2p]**

(h) (4 Points) Propose another approach to estimate β consistently.

Solution: We could use a Fixed Effects - Instrumental Variable (FE-IV) approach. **[1p]**

Taking first differences of our original model equation yields

$$\Delta y_{it} = \Delta x_{it}' \beta + \Delta u_{it} \quad \text{or} \quad y_i^{\Delta} = X_i^{\Delta} \beta + u_i^{\Delta} \ ,$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$, and analogously for Δx_{it} and Δu_{it}. Estimating this model with pooled OLS would yield the FE-FD estimator, which also requires strict exogeneity for consistency. Under sequential exogeneity, the FE-FD estimator is (in general) inconsistent since $u_{i,t-1}$ might be correlated with x_{it}, in which case Δu_{it} is correlated with Δx_{it}. In particular, this situation arises if x_{it} is a function of $y_{i,t-1}$. **[1p]**

As a remedy, we can take $y_{i,t-2}$ as an IV for Δx_{it}. This IV is relevant since $y_{i,t-2}$ is correlated with $x_{i,t-1}$ and therefore Δx_{it}. The IV is exogenous because, provided that u_{it} is uncorrelated over time, $y_{i,t-2}$ is only correlated with $u_{i,t-2}$, but not $u_{i,t-1}$ nor u_{it} and therefore not Δu_{it}. **[2p]**

(Could also describe another FE-IV approach; e.g. using also further lags – like $y_{i,t-3}$ – as IVs for Δx_{it}, or using $\Delta y_{i,t-1}$ as IV for x_{it} in the original equation of interest (whereby α_i is put into the error term).)

(Could also state that we could use the FE-FD estimator as-is (without the IV part), since it is consistent under milder assumptions than the FE-W estimator; it only requires x_{it} to be uncorrelated with $u_{i,t-1}$ (i.e. $y_{i,t-1}$), while it is allowed to be correlated with higher lags of u_{it} (y_{it}), e.g. $y_{i,t-2}$. In the present case, this means that current enrolment or school expenditures can be a function of math scores in the more distant but not immediate past.)

Exercise 12.2: Solution

Consider the panel data regression

$$y_{it} = \alpha_i + \beta_1 \tilde{x}_{it} + \beta_2 y_{i,t-1} + u_{it} = \alpha_i + x_{it}'\beta + u_{it} \ ,$$

where $\mathbb{E}[\tilde{x}_{it} u_{is}] = 0 \ \forall \, t, s$, $\mathbb{E}[\tilde{x}_{it}\alpha_i] = 0 \ \forall \, t$, and $\mathbb{E}[u_{it}^2 | \tilde{x}_{it}, y_{i,t-1}] = 1$. Suppose you have $T = 3$ time periods and $n = 10'000$ cross-sectional units.

(a) (3 points) Can you estimate $\beta = (\beta_1, \beta_2)$ consistently by applying (pooled) OLS to the above regression model? Why (not)? And what if you assumed Normality of u_{it} and estimated the model by ML?

Solution: No, because we have an incidental parameters problem; $\{\alpha_i\}_{i=1}^n$ appear and as $n \to \infty$, we have infinitely many parameters. This holds regardless of the distribution of u_{it} (i.e. regardless of whether we use OLS or MLE).

(b) (4 points) How would you proceed under the Random Effects paradigm? Under what conditions can you consistently estimate β in this model?

Solution: We would put α_i into the error term –

$$y_{it} = x_{it}'\beta + v_{it} \ , \quad v_{it} = \alpha_i + u_{it}$$

– and estimate this model using OLS.

In this particular model, this does not result in a consistent estimator because one of the regressors, $y_{i,t-1}$, and α_i are obviously correlated.

(c) (4 points) Consider the Fixed Effects First Difference transformation

$$\Delta y_{it} = \beta_1 \Delta \tilde{x}_{it} + \beta_2 \Delta y_{i,t-1} + \Delta u_{it} \ .$$

Find $\hat{\beta}_{FD}$, the (pooled) OLS estimator of this transformed model. Is it consistent?

Solution: Using the standard OLS-derivations, we get

$$\begin{aligned} \hat{\beta}_{FD} &= \left(\sum_i \sum_t \Delta x_{it} \Delta x_{it}' \right)^{-1} \sum_i \sum_t \Delta x_{it} \Delta y_{it} \\ &= \beta + \left(\sum_i \sum_t \Delta x_{it} \Delta x_{it}' \right)^{-1} \sum_i \sum_t \Delta x_{it} \Delta u_{it} \ . \end{aligned}$$

Exercise 12.2: Solution

By WLLN and CLT, $\left(\sum_i \sum_t \Delta x_{it} \Delta x_{it}'\right)^{-1} \xrightarrow{p} \left(\mathbb{E}\left[\sum_t \Delta x_{it} \Delta x_{it}'\right]\right)^{-1}$. Also, by WLLN,

$$\sum_i \sum_t \Delta x_{it} \Delta u_{it} \xrightarrow{p} \mathbb{E}\left[\sum_t \Delta x_{it} \Delta u_{it}\right] \neq 0 \ ,$$

because x_{it} and $u_{i,t-1}$ are correlated because x_{it} includes $y_{i,t-1}$. Hence, $\hat{\beta}_{FD}$ is not consistent.

(d) (4 points) Consider using $y_{i,t-2}$ as an IV for $\Delta y_{i,t-1}$. Find the GMM-IV estimator $\hat{\beta}_{IV,GMM}$. Can you use its asymptotic variance to conduct reliable inference on β?

Solution: Let $z_{it} = (\Delta \tilde{x}_{it}, y_{i,t-2})$. We obtain the GMM-IV estimator $\hat{\beta}_{IV,GMM}$ by solving for β in the sample analogue of $\mathbb{E}[z_{it}\Delta u_{it}] = \mathbb{E}[z_{it}(\Delta y_{it} - \Delta x_{it}'\beta)] = 0$, i.e.

$$\frac{1}{n}\sum_i z_{it}(\Delta y_{it} - \Delta x_{it}'\beta) = \frac{1}{n} Z'(Y^{\Delta} - X^{\Delta}\beta) = 0 \ .$$

This yields

$$\hat{\beta}_{IV,GMM} = (Z'X^{\Delta})^{-1} Z'Y^{\Delta} \ .$$

No, we cannot use its asymptotic distribution to conduct reliable inference on β because its finite sample distribution will be very different from its asymptotic one due to weak IV issues.

(e) (3 points) Suppose you have a consistent estimator $\hat{\beta}$. Provide an estimator for α_i. Is it consistent? Can you think of an alternative estimator with a lower variance?

Solution: One possibility is to take $\hat{\alpha}_i = \frac{1}{T}\sum_t (y_{it} - x_{it}'\hat{\beta})$. It is not consistent, unless we are willing to consider large T asymptotics, i.e. $T \to \infty$. And yes, an alternative estimator with a lower variance is the (conditional) posterior mean under the prior(s) $\alpha_i \sim N(\mu, \sigma_{\alpha}^2)$, as it pulls $\hat{\alpha}_i$ toward the common cross-sectional mean of all $\{\hat{\alpha}_i\}_{i=1}^n$.

Exercise 12.3: Solution

This problem is based on Griliches and Mairesse (1995, NBER Working Paper 5067, "Production Functions: The Search for Identification"). You can download the data from Moodle either in ASCII format GMdata.RAW or in Stata format GMdata.dta. There are nine variables: *index* (firm ID), *sic3* (3 digit SIC), *yr* (year $\in \{73, 78, 83, 88\}$), *ldsal* (log of deflated sales), *lemp* (log of employment), *ldnpt* (log of deflated capital), *ldrnd* (log of deflated R&D), *ldinv* (log of deflated investment). Consider the model

$$ldsal_{it} = \alpha_i + \beta_1 lemp_{it} + \beta_2 ldnpt_{it} + u_{it}. \tag{1}$$

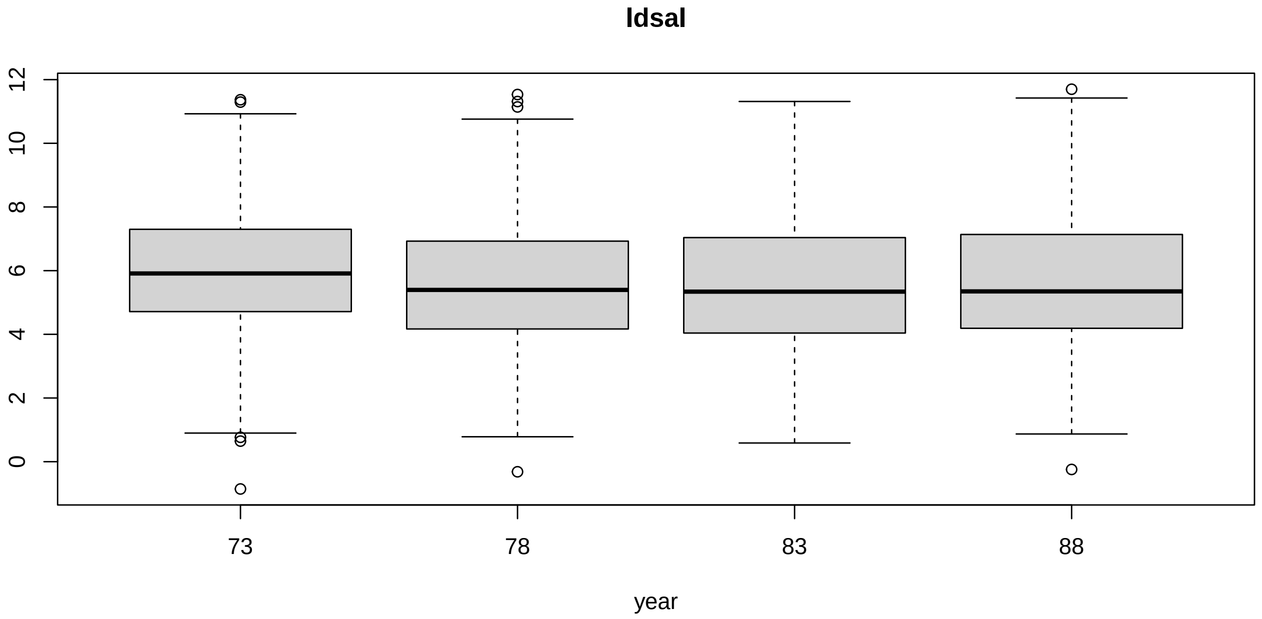
(a) Compute cross-sectional summary statistics (by year) for the following variables: *ldsal*, *lemp*, *ldnpt*. For each year and each variable report: mean, median, standard deviation, minimum, maximum, 5th percentile, 95th percentile. Generate box plots, one for each variable, placing the years next to each other. Do you see a time trend?

```
mData <- read_dta('../ProblemsDataPapers/GMdata.dta')
mData <- as.matrix(mData)
mData <- as.data.frame(mData)

describe(mData)
```

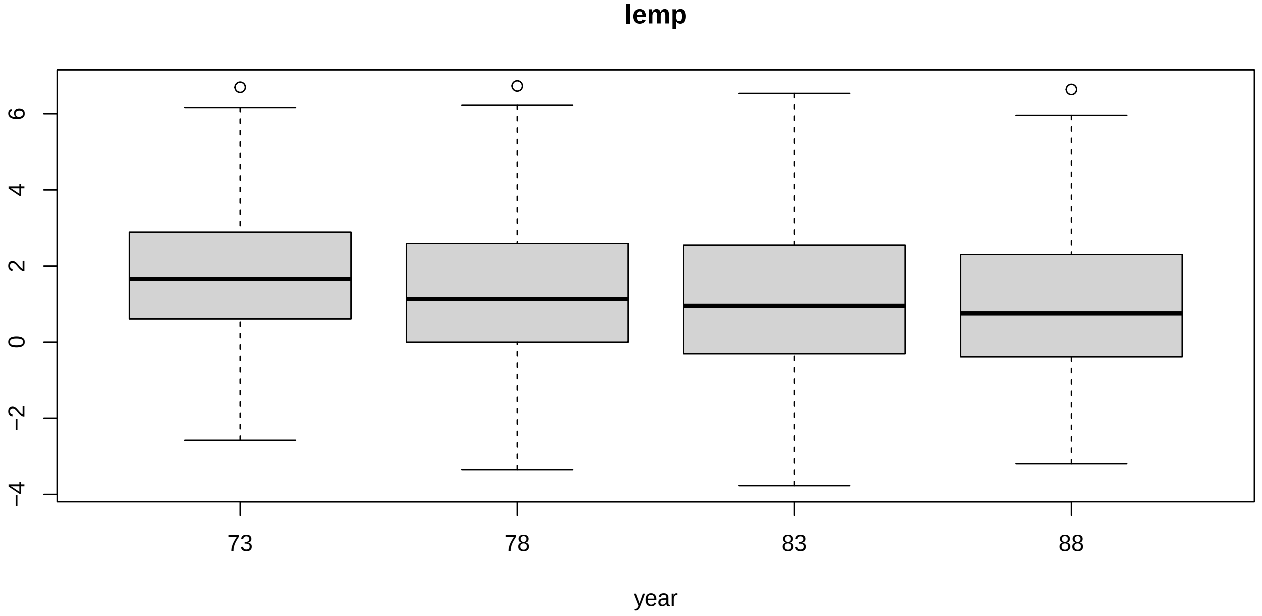
```
##       vars    n   mean     sd median trimmed    mad    min     max   range
## index    1 2971 696.20 404.78 696.00  695.56 521.88   1.00 1400.00 1399.00
## sic3     2 2971 331.46  51.95 356.00  338.71  34.10 200.00  399.00  199.00
## yr       3 2971  80.49   5.35  78.00   80.49   7.41  73.00   88.00   15.00
## ldsal    4 2971   5.67   1.96   5.53    5.64   2.06  -0.86   11.70   12.56
## lemp     5 2971   1.26   1.78   1.11    1.24   1.92  -3.77    6.73   10.50
## ldnpt    6 2971   4.47   2.22   4.21    4.41   2.22  -1.39   11.11   12.50
## ldrst    7 2971   3.40   2.03   3.18    3.34   2.06  -4.29    9.97   14.25
## ldrnd    8 2971   1.79   2.05   1.63    1.74   2.07  -5.31    8.43   13.75
## ldinv    9 2971   2.67   2.17   2.51    2.63   2.25  -3.84    8.99   12.83
##        skew kurtosis   se
## index  0.01    -1.22 7.43
## sic3  -1.03    -0.02 0.95
## yr     0.04    -1.25 0.10
## ldsal  0.17    -0.33 0.04
## lemp   0.14    -0.53 0.03
## ldnpt  0.26    -0.29 0.04
## ldrst  0.25    -0.18 0.04
## ldrnd  0.22    -0.19 0.04
## ldinv  0.18    -0.21 0.04
```

```
boxplot(ldsal ~yr,data=mData,xlab="year",ylab="",main="ldsal")
```

Exercise 12.3: Solution

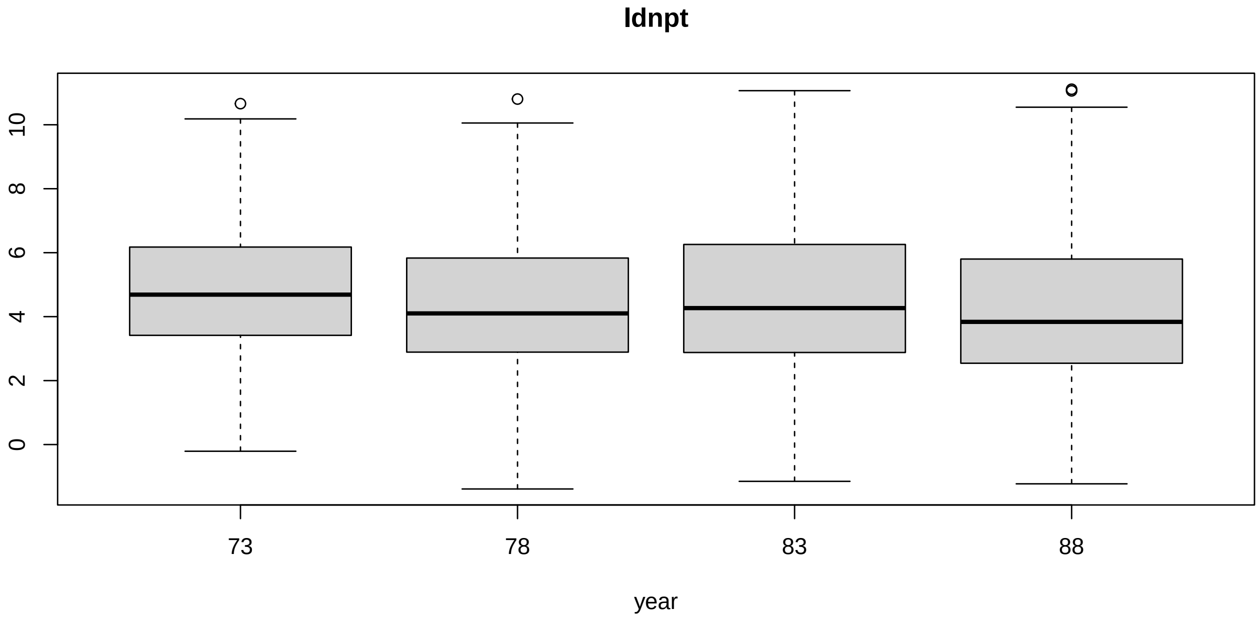


```
boxplot(lemp ~yr,data=mData,xlab="year",ylab="",main="lemp")
```



```
boxplot(ldnpt ~yr,data=mData,xlab="year",ylab="",main="ldnpt")
```

Exercise 12.3: Solution



No clear time trend visible.

(b) Now, let's create a balanced panel and eliminate firms for which you don't have observations for all four years. How many firms do you lose?

```
#Select firms existing all four periods:

mNumberYears <- summaryBy(yr ~ index, FUN=length, data=mData)
vIndicesBalancedPanel <- mNumberYears[mNumberYears$yr.length==4,1]

mDataBP <- mData[mData$index %in% vIndicesBalancedPanel,]
```

We lose 2115 observations, or 1186 firms, which leaves us with 214 firms.

(c) Compute the Random Effects (RE) estimator of $\beta = (\beta_1, \beta_2)'$, i.e., estimate a pooled OLS regression of $ldsal_{it}$ on $lemp_{it}$ and $ldnpt_{it}$ along with an intercept, putting α_i into the error term. State the assumptions needed for consistency of the RE estimator. Are they likely to hold?

```
mX <- cbind(1, as.matrix(mDataBP[,c("lemp","ldnpt")]) )
mY <- as.matrix(mDataBP[,c("ldsal")])

beta_POLS <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)

beta_POLS
```

```
##            [,1]
##       2.7234565
## lemp  0.4289314
## ldnpt 0.5334541
```

Exercise 12.3: Solution

Our model of interest is:

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}$$

Where α_i are entity-specific intercepts. To estimate RE-POLS we push the entity specific intercepts in the error term and replace them with a common intercept. The model therefore becomes:

$$y_{it} = \beta_0 + x_{it}'\beta + v_{it}$$

Where $v_{it} = u_{it} + \alpha_i - \beta_0$. By stacking observations over time and adding a unit vector to the regressors matrix: $\tilde{X}_i = (1, X_i)'$, the RE-POLS estimator is then written as:

$$\hat{\hat{\beta}}_{RE-POLS} = \left(\frac{1}{N}\sum_{i=1}^{n}\tilde{X}_i'\tilde{X}_i\right)^{-1}\frac{1}{N}\sum_{i=1}^{n}\tilde{X}_i'\tilde{y}_i$$

As $N \to \infty$, with an i.i.d. sample, we have that:

$$(\hat{\hat{\beta}}_{RE-POLS} - \beta) \xrightarrow{p} \mathbb{E}[\tilde{X}_i'\tilde{X}_i]^{-1}\mathbb{E}[\tilde{X}_i'v_i]$$

Therefore, the RE-POLS estimator is consistent as long as $\mathbb{E}[\tilde{X}_i'v_i]$ (or $\mathbb{E}[\tilde{x}_{it}v_{it}]$). By recalling that $v_{it} = u_{it} + \alpha_i - \beta_0$ and substituting it into the condition above, we obtain that the two critical assumptions needed for consistency are:

$$\mathbb{E}[\tilde{x}_{it}u_{it}] \ \forall \ t \quad \text{and} \quad \mathbb{E}[\tilde{x}_{it}\alpha_i] = 0$$

In words, the first assumption means that the regressors and the error term should not be correlated at every time horizon. The second assumption implies that the unobserved heterogeneity across individuals should not be correlated with the regressors. Otherwise, have omitted variable bias.

If we assume that there are economies of scale, larger firms (i.e. with more employees) are typically more productive than smaller firms. Therefore, the unobserved firm-specific productivity level correlates both with $ldsal_{it}$ on $lemp_{it}$. This makes the second assumption unlikely to hold.

(d) Compute the Fixed Effects (FE) Within (FE-W) estimator of β. State the assumptions needed for its consistency. Are they likely to hold?

```r
mX1 <- summaryBy(lemp ~ index, FUN=function(x){x-mean(x)}, data=mDataBP)
mX1 <- matrix(as.matrix(mX1[,-1]),nrow(mDataBP),1)
mX2 <- summaryBy(ldnpt ~ index, FUN=function(x){x-mean(x)}, data=mDataBP)
mX2 <- matrix(as.matrix(mX2[,-1]),nrow(mDataBP),1)
mX <- cbind(mX1,mX2)
mY <- summaryBy(ldsal ~ index, FUN=function(x){x-mean(x)}, data=mDataBP)
mY <- matrix(as.matrix(mY[,-1]),nrow(mDataBP),1)

beta_Within <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)
```

Exercise 12.3: Solution

```
beta_Within
```

```
##            [,1]
## [1,] 0.7650351
## [2,] 0.4081587
```

The FE-W estimator assumes that α_i is correlated with x_{it}. Therefore, it is included in the regression:

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}$$

By FWL, regressing y_{it} on x_{it} and on N dummies (one for each entity i) is equivalent to regressing the entity-demeaned y_{it} on the entity-demeaned x_{it}:

$$\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{u}_{it}$$

Where:

$$\ddot{y}_{it} \equiv y_{it} - \bar{y}_i = y_{it} - \sum_{t=1}^{T} y_{it}, \quad \ddot{x}_{it} \equiv x_{it} - \bar{x}_i = x_{it} - \sum_{t=1}^{T} x_{it}, \quad \ddot{u}_{it} \equiv u_{it} - \bar{u}_i = u_{it} - \sum_{t=1}^{T} u_{it}$$

Thus:

$$\hat{\beta}_{FE-W} = \left(\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{x}_{it}' \right)^{-1} \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{y}_{it}$$

And, under the usual assumptions, for $N \to +\infty$:

$$\hat{\beta}_{FE-W} \xrightarrow{p} \beta + \mathbb{E}[\ddot{x}_{it}\ddot{x}_{it}']^{-1}\mathbb{E}[\ddot{x}_{it}\ddot{u}_{it}]$$

The FE-W estimator is consistent also in the case that $\mathbb{E}[x_{it}\alpha_i] \neq 0$, since the individual intercept is included in the model. However, it requires the assumption that:

$$\mathbb{E}[\ddot{x}_{it}\ddot{u}_{it}] = 0, \quad \forall \ i, t$$

In words, the regressors at time t for individual i need to be uncorrelated with all residuals at all time horizons and for all individuals.

If we assume a dynamic economic structure, where firms' sales are a function of past sales (possibly proxying persistent rent positions or customer relations in specific market segments), then the assumption needed for consistency will not hold.

(e) Compute the FE First Difference (FE-FD) estimator of β. State the assumptions needed for its consistency. Are they likely to hold?

```
mDX1 <- summaryBy(lemp ~ index, FUN=function(x){diff(x)},
                  data=mDataBP)
mDX1 <- matrix(as.matrix(mDX1[,-1]),nrow(mDataBP)-1,1)
```

Exercise 12.3: Solution

```
mDX2 <- summaryBy(ldnpt ~ index, FUN=function(x){diff(x)},
                  data=mDataBP)
mDX2 <- matrix(as.matrix(mDX2[,-1]),nrow(mDataBP)-1,1)
mDX <- cbind(mDX1,mDX2)
mDY <- summaryBy(ldsal ~ index, FUN=function(x){diff(x)},
                 data=mDataBP)
mDY <- matrix(as.matrix(mDY[,-1]),nrow(mDataBP)-1,1)

beta_FD <- solve(t(mDX)%*% mDX) %*% (t(mDX) %*% mDY)

beta_FD
```

```
##           [,1]
## [1,] 0.8513505
## [2,] 0.1980678
```

The FE-FD estimator is defined as:

$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta u_{it}$$

With:

$$\Delta y_{it} \equiv y_{it} - y_{it-1}, \quad \Delta x_{it} \equiv x_{it} - x_{it-1}, \quad \Delta u_{it} \equiv u_{it} - u_{it-1}$$

Therefore:

$$\hat{\beta}_{FE-FD} = \left(\frac{1}{N} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)^{-1} \frac{1}{N} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=2}^{T} \Delta x_{it} \Delta y_{it}$$

And, again under the usual i.i.d. assumption and for $N \to +\infty$:

$$\hat{\beta}_{FE-FD} \xrightarrow{p} \beta + \mathbb{E}\left[\sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right]^{-1} \mathbb{E}\left[\sum_{t=2}^{T} \Delta x_{it} \Delta u_{it} \right]$$

The FD estimator is thus consistent if:

$$\mathbb{E}\left[\sum_{t=2}^{T} \Delta x_{it} \Delta u_{it} \right] = \sum_{t=2}^{T} \mathbb{E}[(x_{it} - x_{it-1})(u_{it} - u_{it-1})] = 0$$

This implies that $\mathbb{E}[x_{it}u_{it-1}]$ (and hence $\mathbb{E}[x_{it}u_{it-1}]$) must be zero. However, covariances between x_{it} and lags of u_{it} (and y_{it}) larger than 2 can be nonzero.

This is a looser assumption compared to the one needed for the FE-W estimator to be consistent. However, it still requires no simultaneity between the regressors and the dependent variable. If we assume a simultaneity between sales and employment (i.e. firms' hiring decisions are a function of previous years' sales), then

Exercise 12.3: Solution

FE-FD is not consistent.

(f) Derive the asymptotic distribution of the FE-W estimator.

Under the (strict exogeneity) assumption that $\mathbb{E}[\ddot{x}_{it}\ddot{u}_{it}] = 0$ for all leads and lags, we have that, as $N \to +\infty$, the finite-sample variance of the FE-W estimator:

$$\mathbb{E}[(\hat{\beta}_{FE-W} - \beta)(\hat{\beta}_{FE-W} - \beta)'] =$$

$$= \left(\left(\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{x}_{it}' \right)^{-1} \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{u}_{it} \right) \left(\left(\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{x}_{it}' \right)^{-1} \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{u}_{it} \right)' =$$

$$= \left(\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{x}_{it}' \right)^{-1} \left(\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{u}_{it} \right) \left(\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{u}_{it} \right)' \left(\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_{it}\ddot{x}_{it}' \right)^{-1}$$

Converges to the asymptotic variance:

$$\mathbb{V}_W = \mathbb{E}\left[\sum_{t=1}^{T} \ddot{x}_{it}\ddot{x}_{it}' \right]^{-1} \mathbb{E}\left[\left(\sum_{t=1}^{T} \ddot{x}_{it}\ddot{u}_{it} \right) \left(\sum_{s=1}^{T} \ddot{x}_{is}\ddot{u}_{is} \right)' \right] \mathbb{E}\left[\sum_{t=1}^{T} \ddot{x}_{it}\ddot{x}_{it}' \right]^{-1}$$

Therefore,

$$(\hat{\beta}_{FE-W} - \beta) \xrightarrow{d} N(0, \mathbb{V}_W)$$

With asymptotic variance-covariance matrix $\mathbb{V}_W$ defined above.

(g) Compute the standard errors (i.e. estimates of the standard deviations) of your FE-W estimates of β_1 and β_2. You can base your calculations on the asymptotic variance you derived in the previous exercise, or you can use a command from a software package as long as you can make sure it is based on an appropriate formula.[1]

```
vErrors <- mY - mX %*% beta_Within

N <- length(unique(mDataBP$index))
T <- length(unique(mDataBP$yr))

term1 <- matrix(0,2,2)
term2 <- matrix(0,2,2)
vIndices <- unique(mDataBP$index)
for (ii in 1:N){
```

[1] Pay attention not to use any simpler formula that assumes homoskedasticity and/or no serial correlation. In addition, built-in commands in most statistical packages may have finite-sample degree-of-freedom-adjustment terms (in the style of dividing by $(n-1)$ instead of n in the sample variance). While these do not matter asymptotically, in finite sample you may see differences. If you would like to code by hand such finite-sample adjustment terms, see Chapter 17.12 of the Hansen textbook.

Exercise 12.3: Solution

```
    vDataThisIndex <- mDataBP$index == vIndices[ii]
    mXii <- mX[vDataThisIndex,]
    mUii <- matrix(vErrors[vDataThisIndex],T,1)

    # term1here <- 0
    # for (tt in 1:T){
    #   term1here <- term1here + matrix(mXii[tt,],2,1) %*% matrix(mXii[tt,],1,2)
    # }
    term1 <- term1 + t(mXii) %*% mXii

    term2helperhere <- 0
    for (tt in 1:T){
        term2helperhere <- term2helperhere + matrix(mXii[tt,] * mUii[tt],2,1)
    }
    term2 <- term2 + term2helperhere %*% t(term2helperhere)

}

beta_Within_var <- solve(term1) %*% term2 %*% t(solve(term1))

beta_Within_var
```

```
##                 [,1]          [,2]
## [1,]  0.0022712783 -0.0006266563
## [2,] -0.0006266563  0.0020733553
```

```
cbind(beta_Within, sqrt(diag(beta_Within_var)))
```

```
##            [,1]       [,2]
## [1,] 0.7650351 0.04765793
## [2,] 0.4081587 0.04553411
```

The analytical standard errors of the FE-W estimator computed according to the formula above are 0.0476579, 0.0455341 for $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

(h) Compute the standard errors for your FE-W estimates also based on clustered bootstrapping. This is analogous to classical bootstrapping, but, to get a valid panel dataset, you only draw cross-sectional units (firms) with replacement, and for the drawn firms you take all the time periods. Set the number of bootstrap samples $B = 1000$ and take a sample size of n (your actual sample size).

Hint: To facilitate your coding process, first consider a single bootstrap sample, then verify your code works for $B = 10$ or $B = 100$, and only once you solved the whole problem set, take $B = 1000$, as it might take a long time to execute.

```
B <- 1000
mBetas <- matrix(NA,2,B)

```

Exercise 12.3: Solution

```r
for(bb in 1:B){

    vSampleIndicesHere <- sample(unique(mDataBP$index), N, replace = TRUE)

    mDataBP_thisbatch <- matrix(NA,1,ncol(mDataBP))
    colnames(mDataBP_thisbatch) <- colnames(mDataBP)
    for (ii in 1:N){
        mDataBP_thisbatch <- rbind(mDataBP_thisbatch,mDataBP[mDataBP$index==vSampleIndicesHere[ii],])
    }
    mDataBP_thisbatch <- mDataBP_thisbatch[-1,]

    vNewIndices <- rep(1:N,each=4)
    mDataBP_thisbatch <- cbind(vNewIndices,mDataBP_thisbatch)
    colnames(mDataBP_thisbatch)[1] <- "newindex"

    mX1 <- summaryBy(lemp ~ newindex, FUN=function(x){x-mean(x)}, data=mDataBP_thisbatch)
    mX1 <- matrix(as.matrix(mX1[,-1]),nrow(mDataBP_thisbatch),1)
    mX2 <- summaryBy(ldnpt ~ newindex, FUN=function(x){x-mean(x)}, data=mDataBP_thisbatch)
    mX2 <- matrix(as.matrix(mX2[,-1]),nrow(mDataBP_thisbatch),1)
    mX <- cbind(mX1,mX2)
    mY <- summaryBy(ldsal ~ newindex, FUN=function(x){x-mean(x)}, data=mDataBP_thisbatch)
    mY <- matrix(as.matrix(mY[,-1]),nrow(mDataBP_thisbatch),1)

    beta_Within_thisbatch <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)

    mBetas[,bb] <- beta_Within_thisbatch

}

beta_Within_var_bootstrap <- cov(t(mBetas))

cbind(beta_Within, sqrt(diag(beta_Within_var_bootstrap)))
```

```
##           [,1]       [,2]
## [1,] 0.7650351 0.08019540
## [2,] 0.4081587 0.05946622
```

Here we are resampling entire clusters (i.e. all the time-observations of a single entity, as opposed to individual observations) in our dataset. Specifically, we are drawing with replacement N individuals from the dataset $\{X_{it}, y_{it}\}_{i=1}^{N}$. For each individual $\{X_i, y_i\}$ we include all time periods $t = 1, ..., T$.

The draws with replacement allow us to construct B new samples, each of size $(N \times T)$. We use these new B samples (in our case $B = 1000$) to estimate a new set of parameters $\{\hat{\beta}_b\}_{b=1}^{B}$. These new parameters are then used to estimate the variance-covariance matrix of the FE-W estimator according to the formula:

Exercise 12.3: Solution

$$\hat{V}_{boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}^{(b)} - \bar{\beta})(\hat{\beta}^{(b)} - \bar{\beta})'$$

The clustered bootstrap at the entity level has the advantage to account for conditional heteroskedasticity and autocorrelation withing clusters, i.e. it allows to relax the assumptions that $\mathbb{E}[\ddot{u}_{it}\ddot{u}_{is}] = \sigma^2_{\ddot{u}}$ for $t = s$ and for all t (homoskedasticity), and that $\mathbb{E}[\ddot{u}_{it}\ddot{u}_{is}] = 0$ for $t \neq s$ (no autocorrelation).

We see that the bootstrapped standard errors (0.0801954, 0.0594662 for $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively) are quite a bit higher compared to the analytical ones.

While this difference does not change the significance level of the test of the null hypothesis that either coefficient is equal to zero, it might still change the outcomes of some hypothesis tests, e.g. testing $\beta_1 + \beta_2 = 1$ (constant returns to scale).

(i) Now, instead of creating a balanced panel as in (b), use the full data set (an unbalanced panel) to re-compute the FE-W estimator in (d).

```
# Use only cross-sectional units with at least two years
# of observations:
mObsPerIndex <- summaryBy(yr ~ index, FUN=function(x){length(x)}, data=mData)
vIndicesToUse <- mObsPerIndex[mObsPerIndex[,2] > 1 , 1]

mDataMinTwoYears <- mData[mData$index %in% vIndicesToUse,]

# Here need to create mX and mY manually (not via summaryBy)
# bcs panel is unbalanced:
mX <- matrix(NA,nrow(mDataMinTwoYears),2)
mY <- matrix(NA,nrow(mDataMinTwoYears),1)

vIndices <- unique(mDataMinTwoYears$index)
N <- length(vIndices)
for (ii in 1:N){
    vIndicesHere <- mDataMinTwoYears$index==vIndices[ii]

    mXhere <- mDataMinTwoYears[vIndicesHere,c("lemp","ldnpt")]
    mXhere_demeaned <- mXhere - matrix(apply(mXhere,2,mean),nrow(mXhere),2,byrow=TRUE)
    mX[vIndicesHere,] <- as.matrix(mXhere_demeaned)

    vYhere <- mDataMinTwoYears[vIndicesHere,"ldsal"]
    mY[vIndicesHere] <- as.matrix(vYhere - mean(vYhere))

}

beta_Within <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)

beta_Within
```

Exercise 12.3: Solution

```
##           [,1]
## [1,] 0.7525983
## [2,] 0.3112203
```

We see that the coefficient estimates from the unbalanced panel are relatively close to the ones from the balanced sample. However, they are more cumbersome to compute because the presence of missing observations in the $(N \times T)$ sample implies that the matrix X_{it} has an irregular structure. Therefore, computing the FE-W estimator is not so straightforward, as it requires manual demeaning entity-by-entity because the length of the time dimension $t = 1, ..., T$ varies across individuals.

On top of this, sample-selection issues might occur. For instance, missing values can pertain to firms that discontinued their activity over time due to factors that are correlated with the variables present in the dataset. In this case, including the unbalanced part of the panel might make the sample non-random.

(j) Compute the standard error estimates for your FE-W estimates in (i) based on clustered bootstrap.

```
B <- 1000
mBetas <- matrix(NA,2,B)

for(bb in 1:B){

    vSampleIndicesHere <- sample(unique(mDataMinTwoYears$index), N, replace = TRUE)

    mDataMinTwoYears_thisbatch <- matrix(NA,1,ncol(mDataMinTwoYears))
    colnames(mDataMinTwoYears_thisbatch) <- colnames(mDataMinTwoYears)
    vNewIndices <- NA
    for (ii in 1:N){
        mDataThisii <- mDataMinTwoYears[mDataMinTwoYears$index==vSampleIndicesHere[ii],]
        mDataMinTwoYears_thisbatch <- rbind(mDataMinTwoYears_thisbatch,mDataThisii)
        vNewIndices <- c(vNewIndices,rep(ii,nrow(mDataThisii)))
    }
    mDataMinTwoYears_thisbatch <- mDataMinTwoYears_thisbatch[-1,]
    vNewIndices <- vNewIndices[-1]

    mDataMinTwoYears_thisbatch <- cbind(vNewIndices,mDataMinTwoYears_thisbatch)
    colnames(mDataMinTwoYears_thisbatch)[1] <- "newindex"

# here need to create mX and mY manually
# (not via summaryBy) bcs panel is unbalanced
    mX <- matrix(NA,nrow(mDataMinTwoYears_thisbatch),2)
    mY <- matrix(NA,nrow(mDataMinTwoYears_thisbatch),1)

    vIndices <- unique(mDataMinTwoYears_thisbatch$newindex)
    N <- length(vIndices)
    for (ii in 1:N){
        vIndicesHere <- mDataMinTwoYears_thisbatch$newindex==vIndices[ii]
```

Exercise 12.3: Solution

```
        mXhere <- mDataMinTwoYears_thisbatch[vIndicesHere,c("lemp","ldnpt")]
        mXhere_demeaned <- mXhere - matrix(apply(mXhere,2,mean),nrow(mXhere),2,byrow=TRUE)
        mX[vIndicesHere,] <- as.matrix(mXhere_demeaned)

        vYhere <- mDataMinTwoYears_thisbatch[vIndicesHere,"ldsal"]
        mY[vIndicesHere] <- as.matrix(vYhere - mean(vYhere))

    }

    beta_Within_thisbatch <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)

    mBetas[,bb] <- beta_Within_thisbatch

}

beta_Within_var_bootstrap <- cov(t(mBetas))

cbind( beta_Within , sqrt(diag(beta_Within_var_bootstrap)) )
```

```
##           [,1]       [,2]
## [1,] 0.7525983 0.05906368
## [2,] 0.3112203 0.03633472
```

Also in this case, bootstrapping standard errors in an unbalanced panel requires a manual manipulation of the X_{it}, y_{it} sample to account for missing observations when demeaning entities.

Exercise 12.4: Solution

This is a continuation of the last problem set, based on Griliches and Mairesse (1995, NBER Working Paper 5067, “Production Functions: The Search for Identification”). You can download the data from Moodle either in ASCII format GMdata.RAW or in Stata format GMdata.dta. There are nine variables: *index* (firm ID), *sic3* (3 digit SIC), *yr* (year $\in \{73, 78, 83, 88\}$), *ldsal* (log of deflated sales), *lemp* (log of employment), *ldnpt* (log of deflated capital), *ldrnd* (log of deflated R&D), *ldinv* (log of deflated investment). Compared to the last exercise, we add time fixed effects:

$$ldsal_{it} = \alpha_i + \beta_1 lemp_{it} + \beta_2 ldnpt_{it} + \beta_3 ldrst_{it} + \sum_{\tau=1}^{T} f_t \, \mathbf{1}\{t=\tau\} + u_{it} \ , \tag{1}$$

where $\mathbf{1}\{t=\tau\}$ for $\tau = 1 : T$ are time-dummies.

1. Compare the modeling of unit- and time-fixed effects in (1). How does the addition of time-fixed effects relate to the incidental parameters problem (IPP)?

Solution

The model (1) can be written more compactly as:

$$y_{it} = \alpha_i + x_{it}'\beta + \sum_{\tau=1}^{T} f_t \, \mathbf{1}\{t=\tau\} + u_{it}$$

And in matrix notation as:

$$y = D\alpha + X\beta + Tf + u$$

By stacking the observations for each individual over time. Thus the vector y has dimension $NT \times 1$ and the matrix X has dimension $NT \times k$. The matrix D stacks T dummies for each individual and has dimension $NT \times N$. The matrix T stacks N dummies for each time-period and has dimension $NT \times T$.

The IPP arises because the FE-W estimator for $\hat{\beta}$:

$$\hat{\beta}_W = \left(\frac{1}{N}\frac{1}{T}X'M_{DT}X\right)^{-1}\left(\frac{1}{N}\frac{1}{T}X'M_{DT}y\right) \xrightarrow{p} \beta$$

(Where, clearly, the residual matrix M_{DT} is defined as $I - Z(Z'Z)^{-1}Z'$, the matrix being $Z \equiv [D, T]$). $\hat{\beta}_W$ Consistently estimates the true β for $N \to +\infty$ and for fixed T, while the estimator for the entity specific intercepts:

$$\hat{\alpha}_i = \bar{y}_i - \bar{x}_i'\hat{\beta}_W - \frac{1}{T}\sum_{t=1}^{T}\hat{f}_t = \frac{1}{T}\sum_{t=1}^{T} y_{it} - \left(\frac{1}{T}\sum_{t=1}^{T} x_{it}\right)'\hat{\beta}_W - \sum_{t=1}^{T}\hat{f}_t \nrightarrow \alpha_i$$

Does not consistently estimate the true entity specific intercept α_i. This happens because, for fixed T and $N \to +\infty$, we have that the number of parameters to estimate α_i goes to infinity as well. Hence, α_i are said to be *nuisance parameters*, as they cannot be consistently estimated and do not represent the main object of inference.

Exercise 12.4: Solution

On the other hand, for the same reason, by including time fixed effects we do not incur in the IPP because the estimator $\hat{f}_t$ converges in probability to f_t for $N \to +\infty$ and fixed T. In fact,

$$\hat{f}_t = \bar{y}_t - \bar{x}_t' \hat{\beta}_W - \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_i = \frac{1}{N}\sum_{i=1}^{N} y_{it} - \left(\frac{1}{N}\sum_{i=1}^{N} x_{it}\right)' \hat{\beta}_W - \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_i \xrightarrow{p} f_t$$

2. Construct a balanced panel dataset, and take first differences of (1). What happens to α_i and the year dummies?

Solution

```
# Load data:
mData <- read_dta('../ProblemsDataPapers/GMdata.dta')
mData <- as.matrix(mData)
mData <- as.data.frame(mData)

#Select firms existing all four periods:
mNumberYears <- summaryBy(yr ~ index, FUN=length, data=mData)
vIndicesBalancedPanel <- mNumberYears[mNumberYears$yr.length==4,1]

mDataBP <- mData[mData$index %in% vIndicesBalancedPanel,]
```

We start from model (1):

$$y_{it} = \alpha_i + x_{it}'\beta + \sum_{\tau=1}^{T} f_t \, \mathbf{1}\{t = \tau\} + u_{it}$$

When taking first differences we have:

$$\begin{aligned} y_{it} - y_{i,t-1} &= \alpha_i + x_{it}'\beta + \sum_{\tau=2}^{T} f_t \, \mathbf{1}\{t = \tau\} + u_{it} - \alpha_i + x_{i,t-1}'\beta + \sum_{\tau=1}^{T} f_t \, \mathbf{1}\{t-1 = \tau\} + u_{i,t-1} = \\ &= (\alpha_i - \alpha_i) + (x_{it} - x_{i,t-1})'\beta + \sum_{\tau=2}^{T} f_t (\, \mathbf{1}\{t = \tau\} - \, \mathbf{1}\{t-1 = \tau\}) + (u_{it} - u_{i,t-1}) = \\ &= \Delta x_{it}'\beta + \sum_{\tau=2}^{T} f_t \, \mathbf{1}\{\Delta t = \tau\} + \Delta u_{it} \end{aligned}$$

The entity-specific intercepts α_i wash out because they are time-invariant, while the year dummies become fist-difference dummies, capturing the effect of changing from one year to the other. Notice that these start from the second year of the sample, as we lost one year due to the differencing.

3. Suppose for a moment that you added a dummy for the computer industry to the model: $\delta d357_{it}$, where $d357_{it}$ is a dummy variable equal to one if firm i belongs to industry (3-digit SIC code) 357. Does this have an effect on the IPP? How does implied specification for first differences change? Do your conclusions change if we consider time-varying effects of the industry-dummy, $\delta_t d357_{it}$?

Exercise 12.4: Solution

Solution

By adding a dummy for the computer industry, the model (1) becomes:

$$y_{it} = \alpha_i + x_{it}'\beta + \sum_{\tau=1}^{T} f_t\, \mathbf{1}\left\{t = \tau\right\} + \delta d357_{it} + u_{it}$$

The dummy $d357_{it}$ is likely to be time-invariant, as it's hard for a firm to enter the computer industry or to exit it to enter another one. Or at least to have very little time-variation. Any time-invariant dummy that is introduced in a FE-W model will cause perfect multicollinearity, as it will be a linear combination of the individual intercepts α_i.

For the same reason, the implied FD specification would not change, because, by differencing all variables, time-invariant covariates wash out.

If we want to estimate the time-varying model:

$$y_{it} = \alpha_i + x_{it}'\beta + \sum_{\tau=1}^{T} f_t\, \mathbf{1}\left\{t = \tau\right\} + \delta_t d357_{it} + u_{it}$$

We can create a set of time-industry interaction terms $\delta_t d357_{it} = \sum_{\tau=1}^{T} d_\tau d357_{it}\, \mathbf{1}\left\{t = \tau\right\}$ that vary over time for the same firm. Therefore, they are not perfectly collinear with the individual fixed effects α_i.

Moreover, the implied FD specification now becomes:

$$\Delta y_{it} = \Delta x_{it}'\beta + \sum_{\tau=2}^{T} f_t\, \mathbf{1}\left\{\Delta t = \tau\right\} + \sum_{\tau=2}^{T} d_\tau d357_{it}\, \mathbf{1}\left\{\Delta t = \tau\right\} + \Delta u_{it}$$

In this case taking first differences in the model with time-dummy interactions does not eliminate the $d357_{it}$ variable.

4. Using the balanced panel, re-compute the FE-FD estimator, the FE-W estimator, the RE estimator and the respective standard error estimates.

Solution

```r
# Create year dummies
vYears <- unique(mData$yr)
T <- length(vYears)
N <- length(unique(mData$index))

for (tt in 1:T){
  sYearDummyHere <- paste('yr_',vYears[tt],sep='')
  mDataBP[,sYearDummyHere] <- as.numeric( mDataBP$yr == vYears[tt] )
}



# -- FD Estimator --
```

Exercise 12.4: Solution

```
fDiffByUnit <- function(mData,indUnit,indCol){
mHelp <- summaryBy(list(c(colnames(mData)[indCol]),
                        c(colnames(mData)[indUnit])),
                   FUN=function(x){diff(x)}, data=mData)
matrix(t(as.matrix(mHelp[,-1])),nrow(mData),1)
}

mX <- cbind( mDataBP[,c("index","yr","lemp","ldnpt","ldrst")],
             mDataBP[,startsWith(colnames(mDataBP),'yr')]   )
mX <- mX[,-6] # get rid of year variable as regressor

for (cc in 3:ncol(mX)){
  mX[,cc] <- fDiffByUnit(mX,1,cc)
}
mX <- as.matrix(mX[,-c(1,2)])
mX <- mX[,-4] #get rid of one year dummy to avoid singularity

mY <- mDataBP[,c("index","yr","ldsal")]
mY <- fDiffByUnit(mY,1,3)

beta_FD <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)

# Compute FD standard errors
vErrors <- mY - mX %*% beta_FD
N <- length(unique(mDataBP$index))
T <- length(unique(mDataBP$yr))
k <- ncol(mX)  # number of coefficients
term1 <- matrix(0, k, k)
term2 <- matrix(0, k, k)
vIndices <- unique(mDataBP$index)
for (ii in 1:N){
    vDataThisIndex <- mDataBP$index == vIndices[ii]
    mXii <- mX[vDataThisIndex,]
    mUii <- matrix(vErrors[vDataThisIndex], T, 1)

    term1 <- term1 + t(mXii) %*% mXii

    term2helperhere <- matrix(0, k, 1)
    for (tt in 1:T){
        term2helperhere <- term2helperhere +
          matrix(mXii[tt,] * mUii[tt], k, 1)
    }
    term2 <- term2 + term2helperhere %*% t(term2helperhere)
}
```

Exercise 12.4: Solution

```
beta_FD_var <- solve(term1) %*% term2 %*% t(solve(term1))
cbind(beta_FD, sqrt(diag(beta_FD_var)))
```

```
##             [,1]       [,2]
## lemp  0.73816184 0.04733453
## ldnpt 0.09623454 0.04878127
## ldrst 0.28574805 0.06992956
## yr_78 0.09078790 0.01759902
## yr_83 0.07877562 0.03667714
## yr_88 0.32386321 0.04077437
```

```
# -- Within Estimator --

fDemeanColumnByUnit <- function(mData,indUnit,indCol){
mHelp <- summaryBy(list(c(colnames(mData)[indCol]),
                        c(colnames(mData)[indUnit])),
                   FUN=function(x){x-mean(x)}, data=mData)
matrix(t(as.matrix(mHelp[,-1])),nrow(mData),1)
}

mX <- cbind( mDataBP[,c("index","yr","lemp","ldnpt","ldrst")],
             mDataBP[,startsWith(colnames(mDataBP),'yr')]  )
mX <- mX[,-6] # get rid of year variable as regressor

for (cc in 3:ncol(mX)){
  mX[,cc] <- fDemeanColumnByUnit(mX,1,cc)
}
mX <- as.matrix(mX[,-c(1,2)])
mX <- mX[,-4] #get rid of one year dummy to avoid singularity

mY <- mDataBP[,c("index","yr","ldsal")]
mY <- fDemeanColumnByUnit(mY,1,3)

beta_Within <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)

# Compute Within standard errors
vErrors <- mY - mX %*% beta_Within
N <- length(unique(mDataBP$index))
T <- length(unique(mDataBP$yr))
k <- ncol(mX)  # number of coefficients
term1 <- matrix(0, k, k)
term2 <- matrix(0, k, k)
vIndices <- unique(mDataBP$index)
for (ii in 1:N){
    vDataThisIndex <- mDataBP$index == vIndices[ii]
    mXii <- mX[vDataThisIndex,]
```

Exercise 12.4: Solution

```
    mUii <- matrix(vErrors[vDataThisIndex], T, 1)

    term1 <- term1 + t(mXii) %*% mXii

    term2helperhere <- matrix(0, k, 1)
    for (tt in 1:T){
        term2helperhere <- term2helperhere +
          matrix(mXii[tt,] * mUii[tt], k, 1)
    }
    term2 <- term2 + term2helperhere %*% t(term2helperhere)
}
beta_Within_var <- solve(term1) %*% term2 %*% t(solve(term1))
cbind(beta_Within, sqrt(diag(beta_Within_var)))
```

```
##             [,1]       [,2]
## lemp  0.77081830 0.07622195
## ldnpt 0.03627170 0.10115318
## ldrst 0.40645724 0.11136221
## yr_78 0.07246191 0.02353267
## yr_83 0.05882430 0.05847224
## yr_88 0.26810584 0.06223677
```

```
# -- RE-GLS --

# First Step: use POLS estimator to obtain residuals
# and then estimate of \Omega (\sigma_\alpha, \sigma_u)

#get POLS:
mX <- cbind( mDataBP[,c("lemp","ldnpt","ldrst")],
             mDataBP[,startsWith(colnames(mDataBP),'yr')]   )
mX <- mX[,-4]  # get rid of year variable as regressor
mX <- mX[,-4] # get rid of one year dummy to avoid singularity

mX <- as.matrix(mX)
mY <- as.matrix(mDataBP[,c("ldsal")])

beta_POLS <- solve(t(mX)%*% mX) %*% (t(mX) %*% mY)


#compute residuals, and cast from long (T*Nx1) to short (NxT) format:
mV <- mY - mX %*% beta_POLS

mV <- data.frame(mDataBP$index,mDataBP$yr,mV)
colnames(mV) <- c("index","yr","v")

mV <- dcast(mV, index  ~ yr)
```

Exercise 12.4: Solution

```
## Using v as value column: use value.var to override.
```

```r
mV <- as.matrix(mV[,-1])


#estimate \sigma_\alpha and \sigma_u:
T <- ncol(mV)
N <- nrow(mV)
k <- length(beta_POLS)
sig2_v_hat <- 1/(N*T-k) * sum(mV^2) #sum of each element squared

sig2_alpha_hat <- 0
for(ii in 1:N){
  mvi2 <- mV[ii,] %*% t(mV[ii,])
  sig2_alpha_hat <- sig2_alpha_hat +
    #sum of all cross-products = sum of
    #(lower-)triangular part of matrix vi*vi'
    sum(mvi2[lower.tri(mvi2)])
}
sig2_alpha_hat <- 1/(N*T*(T-1)/2 -k) * sig2_alpha_hat

sig2_u_hat <- sig2_v_hat - sig2_alpha_hat

#assemble mOmega:
mOmega_hat <- matrix(0, N*T, N*T)
for(ii in 1:N){
  # Create the T×T Omega matrix for this unit
  unit_omega <- matrix(sig2_alpha_hat, T, T) + diag(sig2_u_hat, T, T)

  # Place this T×T matrix in the correct block of the large matrix
  row_start <- (ii-1)*T + 1
  row_end <- ii*T
  col_start <- (ii-1)*T + 1
  col_end <- ii*T

  mOmega_hat[row_start:row_end, col_start:col_end] <- unit_omega
}

mOmega_hat_inv <- solve(mOmega_hat)

# Second Step: compute feasible GLS-R.E. estimator:
mXall <- cbind( mDataBP[,c("index","yr","lemp","ldnpt","ldrst")],
                mDataBP[,startsWith(colnames(mDataBP),'yr')]   )
mXall <- mXall[,-6]  # get rid of year variable as regressor
```

Exercise 12.4: Solution

```r
mXall <- mXall[,-6] # get rid of one year dummy to avoid singularity
mYall <- mDataBP[,c("index","yr","ldsal")]
denominator <- matrix(0, ncol(mXall)-2, ncol(mXall)-2)
numerator <- matrix(0, ncol(mXall)-2, 1)
vIndex <- unique(mDataBP$index)
for (ii in 1:N){
  ind <- vIndex[ii]
  mXi <- as.matrix(mXall[mXall$index==ind,-c(1,2)])

  # Extract corresponding block of Omega_hat_inv
  row_start <- (ii-1)*T + 1
  row_end <- ii*T
  Omega_block_inv <- mOmega_hat_inv[row_start:row_end,
                                    row_start:row_end]

  mYi <- matrix(mYall[mYall$index==ind,3], T, 1)

  # Compute using the unit-specific Omega block
  denominator <- denominator + t(mXi) %*% Omega_block_inv %*% mXi
  numerator <- numerator + t(mXi) %*% Omega_block_inv %*% mYi
}

beta_GLS <- solve(denominator) %*% numerator

# Compute RE-GLS standard errors

term <- t(mX) %*% mOmega_hat_inv %*% mX

beta_GLS_var <- solve(term)
cbind(beta_GLS, sqrt(diag(beta_GLS_var)))
```

```
##               [,1]       [,2]
## lemp  -0.19338513 0.04651181
## ldnpt  0.88603337 0.03484115
## ldrst  0.39802140 0.04304327
## yr_78  0.10801324 0.04779859
## yr_83 -0.18697793 0.05027236
## yr_88  0.08475667 0.05147674
```

5. Conduct the Hausman test, comparing the RE- and FE-W-estimators. What do you conclude?
   Hint: note that you need the estimated covariances of the RE- and FE-W-estimators to set up this hypothesis-test.

Solution

To test the null hypothesis $\mathcal{H}_0$ that $(\hat{\beta}_W - \hat{\beta}_{RE}) = 0$ against the alternative that $(\hat{\beta}_W - \hat{\beta}_{RE}) \neq 0$, we set up the Hausman test statistic:

Exercise 12.4: Solution

$$T_H = (\hat{\beta}_W - \hat{\beta}_{RE})'(A\mathbb{V}[\hat{\beta}_W] - A\mathbb{V}[\hat{\beta}_{RE}])^{-1}(\hat{\beta}_W - \hat{\beta}_{RE}) \sim \chi^2_k$$

Where $A\mathbb{V}[\hat{\beta}_W]$ and $A\mathbb{V}[\hat{\beta}_{RE}]$ are the asymptotic variance-covariance matrices of the Within and Random Effects estimators, respectively.

The test statistic is asymptotically distributed under the null hypothesis as a chi-squared with k degrees of freedom. In our case, $k = 3$.

```r
# Use package plm:
mPDataBP <- pdata.frame(mDataBP, index = c("index", "yr"),
                        drop.index = F, row.names = T)

#Perform regressions:
regWithin <- summary( plm(ldsal ~  lemp + ldnpt + ldrst + yr_78
                          + yr_83 + yr_88 , mPDataBP, model="within") )

regRE <- summary( plm(ldsal ~ lemp + ldnpt + ldrst + yr_78
                      + yr_83 + yr_88 -1, mPDataBP, model="random") )

betaWithin <- as.matrix( regWithin$coefficients[,1] )
betaRE <- as.matrix( regRE$coefficients[,1] )
hausman <- t(betaWithin-betaRE) %*%
  solve( as.matrix(regWithin$vcov - regRE$vcov)) %*%
  (betaWithin-betaRE)

hausman
```

```
##          [,1]
## [1,] 392.3956
```

```r
qchisq(0.9,df=6)
```

```
## [1] 10.64464
```

```r
qchisq(0.95,df=6)
```

```
## [1] 12.59159
```

```r
qchisq(0.99,df=6)
```

```
## [1] 16.81189
```

We reject the H_0-hypothesis that the more efficient random effect estimator is consistent at the 1% significance level.

6. Is there support for the null hypothesis of constant returns to scale, $\mathcal{H}_0 : \beta_1 + \beta_2 = 1$?
   Hint: note that you need the estimated covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ to set up this hypothesis-test.

Solution

Exercise 12.4: Solution

To test $\mathscr{H}_0 : \beta_1 + \beta_2 = 1$ against the alternative $\mathscr{H}_1 : \beta_1 + \beta_2 \neq 1$, we set up the two-tailed Wald test statistic:

$$T_W = g(\hat{\beta}_W)'(G(\hat{\beta}_W)A\mathbb{V}[\widehat{\beta_W}]G(\hat{\beta}_W)')^{-1}g(\hat{\beta}_W) \xrightarrow{d} \chi^2_1$$

Where:

$$g(\hat{\beta}_W) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \end{bmatrix} - 1, \quad G(\hat{\beta}_W) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Given the result from the Hausman test, the test here is conducted using the Within estimate.

```r
# Construct Wald test statistic:
mW <- matrix(c(1,1,0,0,0,0), 1, 6)

waldstat <- t(mW %*% betaWithin -1) %*%
  solve( mW %*% regWithin$vcov %*% t(mW) ) %*%
  (mW %*% betaWithin -1)

waldstat
```

```
##          [,1]
## [1,] 12.13317
```

```r
qchisq(0.9,df=1)
```

```
## [1] 2.705543
```

```r
qchisq(0.95,df=1)
```

```
## [1] 3.841459
```

```r
qchisq(0.99,df=1)
```

```
## [1] 6.634897
```

We reject the hypothesis of constant returns to scale at the 1% significance level.

13 Causal Inference

14 Non-Parametric Estimation & Statistical Learning Methods

15 Further Topics