**Nomes**: João Rodrigues, Marcos de Campos, Vinícius Ferreira

**Observação:** a lista foi realizada usando um notebook em Python. Ao final do documento há o código com as definições das funções utilizadas para resolver os problemas da lista. O notebook com todos os códigos e cálculos foi enviado juntamente com esse PDF.

## **Exercício C2**

### **Questão (i)**

Ao regredirmos *lwage* sobre a lista de variáveis exógenas *educ, exper, tenure, married, South, urban, black e IQ*, obtém-se a seguinte regressão:

Tabela

Descrição gerada automaticamente

Substituindo *IQ* por *KWW* como proxy de aptidão individual, o coeficiente de educação passa de 0,0544 para 0,0576, ou seja, há um aumento no retorno da educação sobre o salário de aproximadamente 0,32%. Além disso, o coeficiente *t* de educação com *KWW* como proxy é maior do que o de *IQ*, o que sugere uma estimação mais precisa.

Tabela

Descrição gerada automaticamente

### **Questão (ii)**

Estimando a regressão com *IQ* e *KWW* como proxies para aptidão individual, obtém-se os resultados abaixo. O retorno marginal da educação sobre o salário diminui para menos de aproximadamente 5%.

Tabela

Descrição gerada automaticamente

### **Questão (iii)**

Fazendo o teste de significância conjunta de *IQ* e *KWW*, vê-se que as variáveis são conjuntamente estatisticamente significantes até mesmo ao nível de 1% de significância, com um p-valor baixíssimo de 0,0002. Assim, ambas as variáveis são boas de serem usadas como proxies para aptidão intelectual e devem ser incluídas no modelo.

Uma imagem contendo Gráfico

Descrição gerada automaticamente

## **Exercício C3**

### **Questão (i)**

O modelo de regressão simples de *lscrap* sobre *grant* pode violar a hipótese RLM.4, uma vez que características da empresa que podem influenciar na obtenção ou não de um subsídio – como seu tamanho, aproximado pela coluna de *sales* – foram relegados ao termo de erro e, portanto, é provável que E(u|x) ≠ 0.

### **Questão (ii)**

Os resultados da estimação usando apenas os dados de 1988 podem ser vistos abaixo. Como não se pude supor causalidade, não se pode dizer que uma concessão de subsídio aumente a taxa de refugo da empresa. Contudo, o coeficiente de *grant* indica uma correlação: a concessão de subsídio aumenta em aproximadamente 5,66% a taxa de refugo da empresa – o que vai contra a intuição -, apesar de seu coeficiente não ser estatisticamente significante.

Tabela

Descrição gerada automaticamente com confiança baixa

### **Questão (iii)**

Usando o *lscrap\_87* como variável exógena ao lado de *grant*, o coeficiente de *grant* se torna negativo (o que era esperado, uma vez que se espera que o subsídio governamental diminua a o *turnover* da empresa), além de se tornar estatisticamente significante ao nível de 10% contra uma alternativa bilateral. Isso indica que há características estruturais da empresa que permanecem ao longo do tempo e afeta, de forma significativa a sua taxa de refugo, como pode ser visto pela estatística *t* de *lscrap\_87*.

Dividindo o p-valor da hipótese bilateral por 2, o p-valor da alternativa unilateral é 0,045, ou seja, rejeita-se a hipótese nula de que o coeficiente de *grant* é maior que 0 em favor de H1: ao nível de 5%.

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

### **Questão (iv)**

O coeficiente de *lscrap\_87* é 0,8312 e seu erro padrão é 0,044. Ao nível de significância de 5%, seu intervalo de confiança não atinge 1, ou seja, rejeita-se H0: em favor de H1: .

### **Questão (v)**

Fazendo a regressão usando erros robustos em relação a heteroscedasticidade:

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

O p-valor dividido por 2 do coeficiente de *grant* é 0,0485, ou seja, o coeficiente é ainda mais significantemente diferente de 0 com os erros padrões robustos, indicando uma estimação mais precisa do coeficiente.

O intervalo de confiança de *lscrap\_87* se aproxima mais de 1, mas o coeficiente ainda é estatisticamente diferente de um ao nível de significância de 5%.

## **Exercício C4**

### **Questão (i)**

O coeficiente de *DC* é muito grande e estatisticamente muito significante, o que indica que o estado de *DC* possuirá uma mortalidade infantil muito maior que os outros estados (todos os outros fatores mantidos constantes), como mostra a estimativa abaixo:

Tabela

Descrição gerada automaticamente

### **Questão (ii)**

As estimativas para o ano de 1990 sem o estado de DC na regressão estão abaixo:

Tabela

Descrição gerada automaticamente

Comparando com a regressão da questão (i), todas as estimativas e erros padrões são idênticos, ou seja, a inclusão de uma dummy para apenas uma observação equivale a retirar essa observação da regressão.

## **Exercício C8**

### **Questão (i)**

A média de *stotal* é 0,047 e seu erro padrão é 0,853, como mostra a imagem abaixo. Um fato interessante é que a mediana da variável é 0 e, portanto, menor que sua média:

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

### **Questão (ii)**

A regressão de *jc* sobre *stotal* faz com que o coeficiente da variável exógena seja 0,0112 e seu erro padrão seja 0,011, resultando em um valor tanto estatisticamente quanto economicamente não significante (próximo de 0).

A regressão de *univ* sobre *stotal* faz com que o coeficiente da variável exógena seja 1,16 e seu erro padrão seja 0,029, resultando em um valor estatisticamente e economicamente bastante significante.

Assim, *univ* é bastante correlacionada com a proxy de aptidão individual, o que, a partir da Solução Plugada do Problema de Variáveis Omitidas, pode tornar os estimadores viesados, mas o viés pode ser menor do que aquele que ocorreria caso se omitisse a variável proxy. Isso ocorre em razão de pessoas com maior aptidão – aproximada por *stotal* - terem maior chance de irem a uma faculdade completa de 4 anos.

Tabela

Descrição gerada automaticamente

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

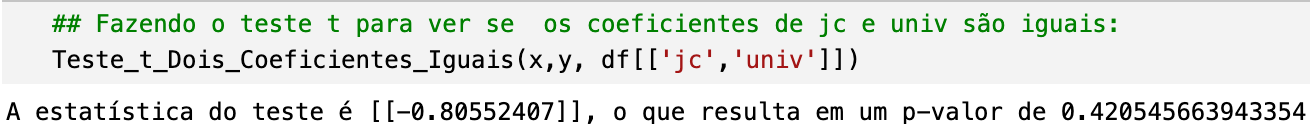
### **Questão (iii)**

Adicionando *stotal* à equação 4.17 do livro, obtém-se as estimativas abaixo:

Tabela

Descrição gerada automaticamente

O teste *t* bilateral de H0: contra H1: resulta em um p-valor bilateral de 0,42, de modo que o p-valor unicaudal do teste é 0,42/2 = 0,21. Assim, não se rejeita a 5% de significância a hipótese de que os retornos são iguais, como é possível ver no código abaixo, que fornece o p-valor bilateral:



### **Questão (iv)**

Adicionando *stotal2* à equação da questão (iii), obtém-se as estimativas abaixo:

Tabela

Descrição gerada automaticamente

A transformação quadrática de *stotal* produz uma estimativa nada estatisticamente significante, de modo que ela pode ser omitida e prefere-se o modelo mais simples da questão (iii), visando diminuir a variância dos estimadores dos coeficientes significantes.

### **Questão (v)**

Adicionando os termos de interação *stotal\*jc* e *stotal\*univ* à equação e fazendo um teste F, vê-se que os novos termos são estatisticamente não conjuntamente significantes ao nível de 5%, uma vez que o p-valor do teste bilateral é 0,141.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

### **Questão (vi)**

O modelo “mais correto” para esse caso é o da questão (iii), uma vez que é o mais simples e não perde em qualidade com relação aos modelos mais complexos usados na questão (iv) e (v) (cujas adições são estatisticamente não significantes ao nível de 5%). A adição de variáveis não significantes não torna os estimadores viesados, mas aumenta a variância das estimações, especialmente caso haja um alto grau de multicolinearidade entre as variáveis exógenas, como é o caso entre *stotal2* e *univ*, por exemplo.

## **Exercício C10**

### **Questão (i)**

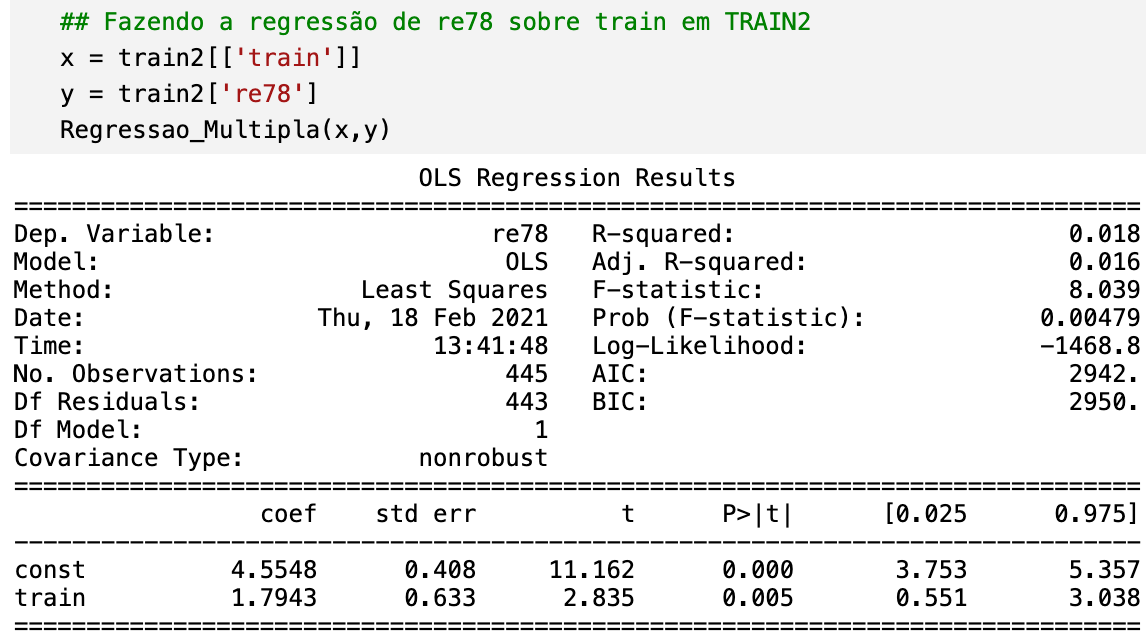
Em JTRAIN2 há cerca de 445 homens na amostra, dentre os quais 41,57% receberam treinamento. Em JTRAIN3, há 2675, dentre os quais apenas 6,92% receberam treinamento (em termos totais, a quantidade de pessoas que receberam treinamento foi igual em ambas as amostras). Essa diferença percentual pode ser explicada pelo fato de que as pessoas em JTRAIN2 foram parte de um experimento direcionado, enquanto a amostra de JTRAIN3 diz respeito a uma subpopulação da população ocupada em 1978.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

### **Questão (ii)**

Segundo a estimativa do modelo simples de regressão de *re78* (ganhos reais em 1978) sobre *train*, há um efeito positivo (e marginalmente estatisticamente significante ao nível de 5%) do treinamento de cerca de $ 1.794,3 na renda anual, como mostram os resultados abaixo:



### **Questão (iii)**

Controlando para outras variáveis exógenas, o coeficiente de *train* não muda muito, sendo ligeiramente menor que o estimado pela regressão simples. Contudo, o coeficiente estimado não é mais estatisticamente significante ao nível de 5%, mas é ao nível de 10% de significância.

Essa falta de mudança pode ser explicada em razão dos dados serem experimentais: caso o grupo do treinamento e o grupo controle sejam divididos aleatoriamente, o recebimento ou não de treinamento não é correlacionado com as características individuais das observações da amostra.

### **Questão (iv)**

Usando a base de dados JTRAIN3, as estimativas da regressão simples de *re78* sobre *train* dão resultados completamente opostos (mas muito significantes) àquelas obtidas a partir dos dados de JTRAIN2. Agora, o treinamento teria sido responsável por diminuir a renda anual de 1978 em cerca de $ 15.204,8:

Tabela

Descrição gerada automaticamente

Controlando para outros fatores, o sinal do estimador de *train* se torna positivo, mas ele é estatisticamente não significante, com um p-valor bastante alto de 0,8:

Tabela

Descrição gerada automaticamente

A mudança em relação ao que se observou usando os dados de JTRAIN2 é justamente o caráter não-experimental das observações de JTRAIN3, o que pode ser visto pelas estimativas da segunda regressão.

*Train* estaria, portanto, correlacionada com outras variáveis em JTRAIN3, como mostram as estimativas de uma regressão auxiliar de *train* sobre as demais variáveis exógenas do modelo 2 da questão (iv): segundo elas, pessoas que receberam treinamento já possuíam menores rendas nos anos anteriores (o que pode ter motivado a escolha pelo treinamento), são de minorias étnicas e possuíam um menor nível de educação e idade.

Tabela

Descrição gerada automaticamente

### **Questão (v)**

Fazendo *avgre = (re74+re75)/2*, obtém-se:

* O mínimo de *avgre* em JTRAIN2 e JTRAIN3 são equivalentes entre si e iguais a 0;
* O máximo em JTRAIN2 é $ 24.376,45, enquanto em JTRAIN3 é $146.901,00
* A média em JTRAIN2 é $ 1.739,70, enquanto em JTRAIN3 é $18.040,44
* O desvio padrão em JTRAIN2 é $ 3.900,09, enquanto em JTRAIN3 é $ 13.293,44

Assim, vê-se que a amostra em JTRAIN3 possui rendas muito maiores e espalhadas do que aquela de JTRAIN2, corroborando o que foi discutido no item (iv) - JTRAIN2 possui observações de indivíduos com rendas historicamente menores. Assim, as bases de dados não representam as mesmas populações, como fica ainda mais evidente em seus boxplots:

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

### **Questão (vi)**

Selecionando apenas os homens de JTRAIN2 com *avgre* menor que 10 e fazendo a regressão solicitada na questão, vê-se que a estimativa de *train* é 1,583 com uma estatística *t* de 2,503 e um p-valor de 0,013.

Fazendo o mesmo para os dados de JTRAIN3, o coeficiente de *train* é 1,8445 com uma estatística *t* de 2,065 e um p-valor de 0,039.

Assim, para a subpopulação de homens de baixa renda, as amostras de JTRAIN2 (dados experimentais) e JTRAIN3 (dados reais) são muito parecidas, uma vez que é justamente essa população que mais procura treinamento (de fato, 179 das 185 pessoas que realizaram o treinamento tem *avgre* menor que 10), o que vai de acordo com as hipóteses da questão (iv).

### **Questão (vii)**

Usando apenas os homens que estavam desempregados em 1974 e 1975 (*unemp\_74 = unemp\_75 = 1*), a estimativas de *train* usando os dados de JTRAIN2 é similar à original - 1,84 agora vs. 1,79 com o conjunto de dados completo - e é ainda mais significante. Já usando a base JTRAIN3, a estimativa de *train* salta para 3,8 – antes, era -15, 2 - com uma estatística *t* de 4,3.

Os sumários das regressões podem ser vistos abaixo (JTRAIN2 e JTRAIN3, respectivamente):

Tabela

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

### **Questão (viii)**

É importante termos populações representativas porque, como mostrado no item (v), JTRAIN3 possui observações de homens com renda muito alta e que, por isso, não escolheram participar do treinamento, o que faz com que a estimativa de treinamento fique negativa – uma vez que apenas homens de menor renda escolheram participar.

Quando controlado pela renda média de anos anteriores ou até mesmo pelo histórico de emprego, as duas bases de dados entregam estimações muito parecidas e que indicam o efeito benéfico do treinamento para aqueles que participaram.

## **Exercício C11**

### **Questão (i)**

Fazendo a regressão de *mrdrte* sobre *exec* e *unem*, vê-se que a quantidade de execuções nos últimos 3 anos é tanto economicamente como estatisticamente não significante, ou seja, o modelo mostra que não há evidência de um efeito dissuasor da pena de morte sobre a taxa de homicídios por 100.000 pessoas.

Tabela

Descrição gerada automaticamente

### **Questão (ii)**

De 1990 a 1993, o Texas registrou 34 execuções, mais que o triplo do segundo estado mais próximo (11, Virginia):

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

Adicionando uma *dummy* para o Texas (equivalente a retirá-lo da amostra, como visto no exercício C2), obtém-se os resultados abaixo, indicando que o coeficiente da *dummy* não é estatisticamente significante. Assim, Texas não é um *outlier* importante para a nossa análise e, portanto, deve ser mantido no modelo.

Tabela

Descrição gerada automaticamente

### **Questão (iii)**

Adicionando a taxa de homicídios por 100.000 habitantes de três anos atrás nas variáveis exógenas, se torna negativo e significante ao nível de 5%, registrando um p-valor de 0,023. *Mrtde\_90* também é estatisticamente muito significativo, com uma estatística *t* de 65,319, o que indica que há características estruturais e temporalmente rígidas de cada estado que afetam a taxa de homicídios.

Tabela

Descrição gerada automaticamente

### **Questão (iv)**

Adicionando uma *dummy* para Texas no modelo da questão (iii), o coeficiente de *exec* se torna estatisticamente não significante - uma vez que boa parte das execuções acontecem no Texas. O coeficiente da *dummy* é estatisticamente não significante, o que faz com que seja melhor deixá-lo na análise, ou seja, a observação do estado do Texas continua não sendo um *outlier* significante.

Tabela

Descrição gerada automaticamente

## **Anexo: definições das funções utilizadas**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente