Questão 1
Implemente em Python o modelo de regressão linear múltipla com todas as variáveis explicativas do conjunto de dados Prestigio. Use blue collar como a categoria de referência para a variável "type", centralize as variáveis quantitativas, subtraindo as de sua média amostral e não considere interação entre variáveis. Exclua os casos (toda a linha) nos quais a informação da variável "type" está faltando. Realize o processo de seleção de variáveis explicativas pelo método backward (manualmente, eliminando cada variável, com base no p-valor do teste de sua significância no modelo), considerando um nível de significância de 5%. Cada passo do processo de seleção resultou na eliminação das seguintes variáveis, com os respectivos p-valores:

Sugestões de código em Python:

```
# com os dados no dataframe 'df'

# use este comando para excluir os dados faltantes
df = df.dropna()

# use este comando para criar uma variável dummy para cada categoria de 'type';
# na regressão, lembre-se que você deverá usar apenas 2 dessas variáveis criadas
# e aquela que ficar de fora, representará a categoria de referência
df = pd.concat([df, pd.get_dummies(df['type'].astype('category'), prefix = 'd')], axis =
    1)
```

Você partirá deste modelo:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	46.1442	1.887	24.453	0.000	42.396	49.892
income_c	0.0010	0.000	3.976	0.000	0.001	0.002
education_c	3.6624	0.646	5.671	0.000	2.380	4.945
df.d_prof	5.9052	3.938	1.500	0.137	−1.915	13.726
df.d_wc	−2.9171	2.665	−1.094	0.277	−8.211	2.377
women_c	0.0064	0.030	0.212	0.832	−0.054	0.067

Alternativas:

(a) women (p-valor=0,832) e intercepto (p=0,879)

(b) women (p-valor=0,832) e wc (p-valor=0,279)

(c) women (p-valor=0,832), wc (p-valor=0,277) e prof (p-valor=0,137)

(d) women (p-valor=0,832), wc (p-valor=0,279) e intercepto (p=0,552)

(e) intercepto (p-valor=0,879), women (p-valor=0,832), wc (p-valor=0,277) e prof (p-valor=0,137)

Solução: **Alternativa b.**

Questão 2
Estudando possíveis interações entre as variáveis explicativas e excluindo do modelo as variáveis não significativas, chegou-se ao seguinte modelo final, considerando variáveis centradas:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	46.9029	1.170	40.074	0.000	44.579	49.227
income_c	0.0027	0.000	6.483	0.000	0.002	0.003
education_c	3.0282	0.428	7.083	0.000	2.179	3.877
df.d_prof	8.7224	2.530	3.447	0.001	3.698	13.747
income_c:df.d_prof	−0.0020	0.000	−4.416	0.000	−0.003	−0.001

NÃO se pode afirmar que:

Alternativas:

(a) cada ano adicional de educação, supondo renda e tipo de profissão mantidos os mesmos, espera-se um aumento de 3 unidades no escore de prestígio da ocupação

(b) para o tipo de ocupação profissional, de gerenciamento e técnica, a renda se associa de maneira negativa com o prestígio, ou seja, quanto maior a renda, menor o prestígio nessa categoria

(c) ocupações com maior renda, mais anos de educação e do tipo de categoria profissional, gerenciamento ou técnico estão associadas a um maior prestígio

(d) a estimativa do parâmetro de regressão associado à variável profissão tem uma variabilidade grande

(e) as ocupações associadas a blue e white collar não diferem quanto ao escore de prestígio da ocupação

Solução: **Alternativa b.**

Questão 3
Use o modelo da Questão 2 (com as 98 ocupações com dados completos) para predizer o prestígio de uma ocupação do tipo profissional com tempo médio de educação igual a $15,22$ e renda média de 9593.

Use a função *predict()* em Python;
res.predict(ynovo), se os resultados estiverem armazenados em "res" e os novos dados em "ynovo".

Alternativas:

(a) 71

(b) 108

(c) nenhuma das alternativas

(d) 76

(e) 67

Solução: **Alternativa a.**

Questão 4
As análises de resíduos NÃO servem para:

Alternativas:

(a) para verificar se há pontos que influenciam demais nos valores preditos

(b) verificar se a suposição de distribuição normal dos erros está satisfeita

(c) verificar a suposição de homocedasticidade (variância constante dos erros)

(d) para a seleção de variáveis explicativas no modelo

(e) para verificar se há valores extremos em relação às variáveis explicativas

Solução: **Alternativa d.**

Note que[1] nos modelos de regressão linear (simples ou múltiplo), as suposições do modelo ajustado precisam ser validadas para que os resultados sejam confiáveis. As análises de resíduos são um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos (o resíduo é dado pela diferença entre a variável resposta observada e a variável resposta estimada $\varepsilon = Y - \widehat{Y}$).

Se o modelo for apropriado, os resíduos devem refletir as propriedades impostas pelo termo de erro do modelo (independência, homocedasticidade, normalidade), além de satisfazer a suposição de linearidade do modelo e a não existência de pontos influentes; no caso regressão múltipla, é preciso ainda diagnosticar colinearidade e multicolinearidade entre as variáveis explicativas para que a relação existente entre elas não interfira nos resultados, causando inferências errôneas ou pouco confiáveis.

Questão 5
Nos modelos de regressão linear (simples ou múltiplo):

Alternativas:

(a) não é possível usar uma transformação para as variáveis explicativas, pois ele passaria a ser não linear

(b) os métodos de estimação de mínimos quadrados ordinários e de máxima verossimilhança geram os mesmos estimadores dos parâmetros de regressão

(c) não pode ser ajustado sem intercepto

(d) não se pode prever valores da variável resposta com as estimativas obtidas dos parâmetros de regressão se a observação não pertencer à amostra usada para treino

(e) não se pode comparar a importância das variáveis explicativas para predição da resposta

Solução: **Alternativa b.**

Note que:

- o termo constante na análise de regressão linear, também chamado de intercepto, é o valor no qual a linha ajustada cruza o eixo y; ao não incluir a constante, a linha de regressão será forçada a percorrer a origem (o que significa que todos os preditores e a variável resposta devem ser iguais a zero nesse ponto); se a linha ajustada não passar naturalmente pela origem, os coeficientes de regressão e as predições serão viesados caso não seja inclusa a constante[2];

[1] Mais detalhes podem ser consultados em: `http://www.portalaction.com.br/analise-de-regressao/analise-dos-residuos`.

[2] Mais em https://blog.minitab.com/pt/analise-de-regressao-como-interpretar-a-constante-intercepto-y.

- pode-se fazer transformação na variável resposta e/ou nas preditoras para contornar problemas com variância não constante, não normalidade dos erros ou não linearidade do modelo, por exemplo;
- supondo o modelo de regressão linear simples[3] dado por $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, o objetivo é estimar os parâmetros β_0 e β_1 de modo que os desvios entre os valores observados e estimados sejam mínimos (quer-se minimizar $\varepsilon_i = Y_i - \widehat{Y}_i$, isto é, $\varepsilon_i = Y_i - [\widehat{\beta}_0 + \widehat{\beta}_1 X_i]$).
  - No método de mínimos quadrados ordinários não é necessário conhecer a forma da distribuição dos erros e o método consiste em minimizar a soma dos quadrados dos desvios:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i]^2$$

cujas estimativas dos parâmetros são obtidas diferenciando-se parcialmente a expressão em relação aos parâmetros e igualando-se a zero. Assim, tem-se:

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) X_i = 0.$$

Segue que:

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum\limits_{i=1}^{n} X_i = \sum\limits_{i=1}^{n} Y_i \\ \widehat{\beta}_0 \sum\limits_{i=1}^{n} X_i + \widehat{\beta}_1 \sum\limits_{i=1}^{n} X_i^2 = \sum\limits_{i=1}^{n} X_i Y_i \end{cases}$$

e então:

$$\widehat{\beta}_0 = \frac{\sum\limits_{i=1}^{n} Y_i}{n} - \frac{\widehat{\beta}_1 \sum\limits_{i=1}^{n} X_i}{n} = \bar{Y} - \widehat{\beta}_1 \bar{X} \quad \text{e} \quad \widehat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum\limits_{i=1}^{n} X_i^2 - n\bar{X}^2}.$$

  - No método de máxima verossimilhança, sob as suposições $Y_i \sim N(\beta_0 + \beta_1 X_i; \sigma^2)$ (Y_i independentes) e $\varepsilon_i \sim N(0; \sigma_\varepsilon^2)$, o método consiste em obter as estimativas dos parâmetros que maximizam o logaritmo natural da função de verossimilhança:

$$L(\beta_0, \beta_1, \sigma^2 | (X_i, Y_i)) = \prod_{i=1}^{n} f((X_i, Y_i); \beta_0, \beta_1, \sigma^2),$$

isto é, que maximizam:

$$l(\beta_0, \beta_1, \sigma^2 | (X_i, Y_i)) = \ln L(\beta_0, \beta_1, \sigma^2 | (X_i, Y_i)) = \sum_{i=1}^{n} f((X_i, Y_i); \beta_0, \beta_1, \sigma^2),$$

[3] As propriedades também se verificam para o modelo de regressão múltiplo.

em que $f((X_i, Y_i); \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}.e^{-\frac{1}{2}\frac{(Y_i-\beta_0-\beta_1 X_i)^2}{\sigma^2}}$; ou seja,

$$l(\beta_0, \beta_1, \sigma^2|(X_i, Y_i)) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i-\beta_0-\beta_1 X_i)^2}{\sigma^2}.$$

Assim, as estimativas dos parâmetros são obtidas diferenciando-se parcialmente a expressão em relação aos parâmetros e igualando-se a zero. Tem-se:

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 = 0.$$

Segue que:

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum\limits_{i=1}^{n} X_i = \sum\limits_{i=1}^{n} Y_i \\ \widehat{\beta}_0 \sum\limits_{i=1}^{n} X_i + \widehat{\beta}_1 \sum\limits_{i=1}^{n} X_i^2 = \sum\limits_{i=1}^{n} X_i Y_i \\ \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 \end{cases}$$

e então:

$$\widehat{\beta}_0 = \frac{\sum\limits_{i=1}^{n} Y_i}{n} - \frac{\widehat{\beta}_1 \sum\limits_{i=1}^{n} X_i}{n} = \bar{Y} - \widehat{\beta}_1\bar{X}, \quad \widehat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum\limits_{i=1}^{n} X_i^2 - n\bar{X}^2} \quad \text{e} \quad \widehat{\sigma}^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2.$$