PROJETO 5 ANÁLISE DE REGRESSÃO

Neste projeto você deverá trabalhar em até **DUPLA**.

Este Projeto 5 está composto por três etapas, as quais estão claramente definidas a seguir:

ESCOLHA DAS VARIÁVEIS (1ª. ETAPA)

Você e seu par deverão escolher três variáveis do site GapMinder[1], sendo uma fazendo papel de resposta e duas como explicativas (causas). É importante justificar a maneira que as duas variáveis influenciam na variável resposta.

Esse site traz a base de dados do Banco Mundial[2] que descreve diversas características dos diversos países do mundo (educação, trabalho, saúde, meio ambiente, entre vários outros temas) ao longo do tempo.

Assim, para a **1ª. Etapa** do trabalho, faça:

- Escolha as três variáveis e justifiquem como as duas escolhidas a variável escolhida como resposta. **Por exemplo, estudar como o gasto total com saúde por pessoa (em $) e o percentual da população com acesso ao saneamento podem explicar a expectativa de vida (em anos) de um país. É importante ressaltar que a equipe pode sim escolher as variáveis citadas, mas não todas obviamente!**

  Faça a seleção de pelos menos duas variáveis explicativas e uma variável resposta utilizando os gráficos do Gap Minder como recurso descritivo.

  IMPORTANTE: Elabore um questionamento/hipótese que envolva as variáveis escolhidas. **No caso do exemplo acima poderia ser: Qualidade de vida melhora sobrevida?**

- Construa a base de dados com as suas variáveis utilizando um ano mais recente e com informações para muitos países. Leve essa base de dados para o Python.

[1] Site do Gap Minder: http://goo.gl/zNwLAZ

[2] http://www.gapminder.org/data/

PARTE TEÓRICA (2ª. ETAPA)

De maneira bastante simplificada, a técnica estatística chamada de regressão nada mais é do que uma ferramenta que costuma ser bastante empregada quando se objetiva modelar o efeito que algumas variáveis exercem nas outras (no geral, uma variável em função de outras). Basicamente, este estudo consiste na construção e análise de uma relação matemática entre as tais variáveis de interesse.

Na terminologia de regressão, a variável que se deseja estudar (efeito) é chamada de variável dependente ou resposta. Já as variáveis que são usadas para explicar a variável dependente são chamadas de regressores, de variáveis independentes ou de variáveis explicativas (causas). Dessa forma, a análise de regressão consiste em estudar como alterações nas variáveis explicativas influenciam o comportamento médio da variável resposta.

A análise de regressão com mais do que uma variável explicativa, chamado de **regressão linear múltipla**, envolve pelo menos duas variáveis explicativas (comumente chamadas de X_1, X_2,..., X_p) e uma variável resposta (comumente chamada de Y). Aqui, vale ressaltar que o termo **regressão linear** significa **regressão linear nos parâmetros**. O ajuste de vocês irá levar em consideração duas variáveis explicativas e, nesse caso, a equação a seguir representa um modelo de regressão linear:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i = \\ &= \hat{y}_i + \varepsilon_i \end{aligned}, \qquad (1)$$

em que

y_i - valor da variável resposta associada ao i-*ésimo* elemento da amostra;

x_{1i} - valor da variável explicativa X_1 associada ao i-*ésimo* elemento da amostra;

x_{2i} - valor da variável explicativa X_2 associada ao i-*ésimo* elemento da amostra;

β_0, β_1 e β_2 - parâmetros que denotam o ajuste linear;

$\hat{y}_i$ - é interpretado como o valor médio ou esperado de Y dado um determinado valor de X_1 e X_2; e

ε_i - erro estocástico (aleatório);

i = 1, 2, ..., n; e

n – tamanho da amostra.

Para a **2ª. Etapa** do trabalho, faça:

a) Consulte em livros como se calculam os estimadores de β_0, β_1 e β_2 a partir dos dados. Não é necessário demostrar as expressões.

b) Como ficam os testes de hipóteses na regressão múltipla e o que a rejeição ou não da particular hipótese nula H_0 significa nesse caso?

c) Qual será a interpretação das estimativas dos coeficientes que serão estimados no seu problema. Aqui, faça a interpretação em termos do problema ainda que a estimativa não tenha sido calculada.

d) Quais as suposições feitas sobre os erros em termos de: distribuição, valor esperado e variância e, ainda responda, como a adequação dessas suposições pode ser checada na prática?

ANÁLISE DE REGRESSÃO (3ª. ETAPA)

Para a **3ª. Etapa** do trabalho, faça:

- Escreva um texto que resuma o seu objetivo em termos das variáveis escolhidas.
- Considerando a sua base de dados construída na 1ª etapa, faça uma análise descritiva aos dados de acordo com o problema definido pelo grupo.
- Ajuste um modelo de regressão múltipla aos dados de acordo com o problema definido pelo grupo e de acordo com a análise descritiva (necessidade de transformação na variável, por exemplo, para deixa-la linear). Avalie, via teste de hipóteses, se há variáveis relevantes ao modelo.
- Verifique a adequação das suposições do modelo e a qualidade do ajuste.
- Interprete os parâmetros utilizando agora as estimativas.
- Elabore uma conclusão sobre seu estudo em função dos resultados inferenciais observados.

CRONOGRAMA

Datas	Fases	Formato
18/05	Entrega da 1ª. etapa	Github na pasta Projeto 5 decada aluno **até às 15h30** ○ Arquivo .docx ou .pdf com os gráficos do GapMinder que auxiliaram na escolha das variáveis e análise dos mesmos. ○ Arquivo contendo base de dados.
23/05	Entrega da 2ª. etapa	Github na pasta Projeto 5 de todos alunos **até às 23h59** ○ Arquivo .docx ou .pdf contendo 2ª. etapa.
30/05	Entrega da 3ª. etapa	Github na pasta Projeto 5 de todos alunos **até às 23h59** ○ Arquivo .ipynb com análises desenvolvidas com estrutura de RELATÓRIO (use Markdown). **IMPORTANTE:** **Espero ver gráfico 3D sofisticado com vários ângulos que ajude na interpretação do problema!!**

Engenharia
Ciência dos Dados

Insper

Fases	Insatisfatório (I)	Em desenvolvimento (D)	Essencial (C)	Proficiente (B)	Avançado (A)
Entrega 1ª etapa: **Seleção de variáveis**	Não fez a entrega		Selecionou as variáveis e entregou IPython mas a entrega foi incompleta (por exemplo não leu todas as variáveis) ou não fez a entrega adequadamente no prazo		Entregou na data adequada a seleção de **3** variáveis alternativas e justificou a escolha com plots de dispersão do Gapminder Entregou na data um IPython Notebook ou **arquivo com base de dados em outra extensão** que demonstra terem conseguido ler a variável atribuída ao grupo juntamente com as demais variáveis escolhidas
Entrega 2ª etapa: **Entender modelos de regressão**	Não entregou	Entrega com atraso considerável ou com parte significativa dos itens faltantes o incorretos	Entrega os itens da rubrica B mas com parte pequena dos itens faltantes ou não perfeitamente corretos.	Apresentou o cálculo dos betas. Menciona parcialmente as suposições sobre os erros mas sem notação adequada. Apresenta sem detalhes como verificar as suposições sobre os erros. Mencionou o que significa rejeitar a hipótese nula sem maiores detalhes.	Fez rubrica B de forma exemplar.

Entrega 3ª etapa: **Objetivo de aprendizado:** **Aplicar e analisar modelos de regressão**	Não fez a entrega	Modelo e diagnóstico muito pobres! Bastante incompletos! .	Apenas executa as regressões usando a função do *statsmodels* mas não deixa completamente claras as intenções, o significado e as conclusões das análises a serem feitas. Explora apenas parcialmente as possíveis combinações de variáveis explicativas e resposta. Uso insuficiente de gráficos para esclarecer as relações entre variáveis, retas de regressão e suposições sobre o erro. Ainda que haja tentativa de uma verificação das do diagnóstico (se suposições foram adequadas), apresenta resultados incompletos. Ainda que haja a tentativa de uma análise de diagnóstico (verificação das suposições), apresenta resultados incompletos.	Descreve muito bem o modelo de regressão quanto ao significado das estimativas significantes e interpretações ao problema. Usa o R2 ou outros para explicar qualidade do ajuste. Ainda que haja a tentativa de uma análise de diagnóstico (verificação das suposições), apresenta resultados incompletos.	Verificou a adequação das suposições feitas sobre o erro de forma quantitativa e conclui se os dados a satisfazem de forma válida ou não, deixando clara a conclusão e justificando com plots ou análise dos parâmetros gerados pelo OLS (veja). Tentou uma combinação variada e representativa de regressores e variáveis explicativas na regressão linear simples e múltipla. Enriqueceu o relatório enunciando quantas combinações teriam sido possíveis e qual estratégia usou para selecionar quais testou. Verifica se os resultados da regressão múltipla são melhores que os da regressão simples com apenas uma das explicativas. Sumarizou bem as tentativas (com uma tabela, por exemplo) e apontou o melhor modelo encontrado No contexto de pelo menos um exemplo, Demonstrou entender o que significam coeficiente de determinação, coeficiente de determinação ajustado (se aplicável), valores p (p-values) dos coeficientes e resultados dos testes de hipótese a respeito dos betas para regressão. Apresentou plots de dispersão e da reta de regressão, pelo menos para o caso de modelos com bom R

					quadrado. Apontou claramente qual o melhor modelo de regressão obtido com base em argumentos quantitativos. Procurou formular uma hipótese plausível sobre por que as variáveis do melhor modelo obtido se relacionam da forma que se relacionaram.
Relatórios (todas as fases)	Não entregou ou realizou uma entrega muito incompleta	Apresentou somente cálculos, comandos do IPython, tabelas e gráficos sem texto	Acrescentou texto que apresenta os passos da análise sem boa conexão com as intenções para as mesmas	Descreveu adequadamente a motivação das análises e discutiu seus resultados.	Realizou as ações da rubrica B, acrescidos de objetivos claros e conclusões claras para as análises. Os objetivos dos textos são colocados de forma clara É dada uma motivação clara sobre porque se faz os cálculos e análises (sem ficarem plots e tabelas jogados) Sempre que cabível e adequado os resultados das análises e cálculos são explicados