PROJETO 3 ANÁLISE DE REGRESSÃO

Neste projeto você deverá trabalhar em **DUPLA** sem mesclar alunos de turmas diferentes e sem repetir membros de grupos formados em projetos ou miniprojetos anteriores.

Este Projeto 3 está composto por três etapas, as quais estão claramente definidas a seguir:

ESCOLHA DAS VARIÁVEIS (1ª. ETAPA)

Você e sua equipe receberão uma variável do professor que deverá fazer parte da sua base de dados. Vocês podem utilizar essa variável como variável resposta.

Essa variável recebida do professor junto de mais outras duas variáveis devem ser devem ser retiradas do site GapMinder[1]. 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.

Para a **1ª. Etapa** do trabalho, faça:

- Escolha outras duas variáveis que complementem a variável atribuída a você para modelar um efeito que pareça fazer sentido. **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.**

  Faça a seleção de pelos menos duas variáveis explicativas e uma variável resposta utilizando os gráficos do Gap Minder como recurso descritivo. Lembre-se que uma dessas variáveis já é uma que recebeu do professor. O grupo não pode escolher outras variáveis já designadas pelo professor aos outros grupos.

  IMPORTANTE: Elabore um problema 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 apenas o ano de 2007 ou apenas o ano de 2010. 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.

O tipo mais simples de análise de regressão, chamado de **regressão linear simples**, envolve uma variável explicativa (comumente chamada de *X*) e uma variável resposta (comumente chamada de *Y*). Aqui, vale ressaltar que o termo **regressão linear** significa **regressão linear nos parâmetros**. Assim, a equação a seguir representa um modelo de regressão linear simples:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \hat{y}_i + \varepsilon_i, \qquad (1)$$

em que

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

x_i - valor da variável explicativa associada ao i-*ésimo* elemento da amostra;

β_0 - parâmetro que denota o intercepto da equação;

β_1 - parâmetro que denota o coeficiente angular da equação;

$\hat{y}_i$ - é chamada regressão de Y em x, que interpretamos como o valor médio de Y dado um determinado valor x;

ε_i - erro estocástico (aleatório);

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

n – tamanho da amostra.

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

a) Obtenha os estimadores de β_0 e β_1 a partir do **Método dos Mínimos Quadrados**, cujo objetivo é encontrar a reta que passa mais próxima ao mesmo tempo de todos os pontos. Neste caso, encontre os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam a soma dos erros ao quadrado dada por:

$$\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

b) Descreva as suposições feitas sobre os erros em termos de distribuição, valor esperado e variância e responda como pode ser checada, na prática, a adequação dessas suposições.

c) Como ficam os testes de hipóteses na regressão simples e o que a rejeição ou não da particular hipótese nula H_0 significa no caso?

d) É possível fazer uma regressão com mais do que uma variável explicativa, ou seja, fazer uma regressão múltipla? Se sim, o que muda ou não muda quando comparada com a regressão simples em termos de: modelo conforme descrito na equação (1), suposições do modelo (item b) e teste de hipóteses (item c)?

ANÁLISE DE REGRESSÃO (3ª. ETAPA)

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

- 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 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.
- Escreva um texto que resuma o seu objetivo em termos das variáveis escolhidas, o modelo ajustado e interpretação das estimativas que foram significantes no modelo.
- Avalie se modelo de regressão múltipla obtido acima é igualmente bom quando os países são separados em subgrupos (com critérios consistentes a definir pelo grupo). Ou seja, avalie se modelo global é igualmente válida para um suposto modelo local.
- Elabore uma conclusão sobre seu estudo em função dos resultados inferenciais observados.

CRONOGRAMA

Datas	Fases	Formato
21/11	Entrega da 1ª. etapa	Github na pasta Projeto 3 de todos alunos **até às 23h59** ○ Arquivo .docx ou .pdf com os gráficos do GapMinder que auxiliaram na escolha das variáveis e análise dos mesmos. ○ Arquivo contendo base de dados.
24/11	Entrega da 2ª. etapa	Github na pasta Projeto 3 de todos alunos **até às 23h59** ○ Arquivo .docx ou .pdf contendo 2ª. etapa. O item a pode ser feito a mão e anexado foto do exercício.
06/12	Entrega da 3ª. etapa	Github na pasta Projeto 3 de todos alunos **até às 23h59** ○ Arquivo .ipynb com análises desenvolvidas com estrutura de RELATÓRIO (use Markdown). **IMPORTANTE:** **NÃO usar exemplo do Python disponibilizado no github como layout da 3ª estapa do seu Projeto 3!** **Por exemplo, espero ver gráfico 3D muito mais sofisticado do que o utilizado no exemplo!!**

LISTA DE VARIÁVEIS QUE DEVE SER CONSIDERADA COMO RESPOSTA

Segue lista de variáveis dentre as quais sua dupla deverá escolher uma para ser utilizada como variável resposta.

- Fertilidade (Children per women)
- Expectativa de Vida (Life expectancy)
- Mortalidade infantil (Child mortality)
- Índice de percepção de corrupção (Corruption Perception Index - CPI)
- Taxas de emprego ou de desemprego (employment or unemployment rates)
- Escore de democracia (Democracy score)

Observação: Caso o grupo tenha uma outra proposta para ser feita utilizando alguma outra variável do Gapminder como resposta, converse com o professor até dia 21 antes da entrega da 1ª. etapa.

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 de beta 0 e beta 1 somente com equações, sem esclarecer de forma clara qual o contexto.. 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. Explica o que muda na equação (1) para a regressão múltipla sem discutir os efeitos no cálculo dos betas	Foi mostrado de forma clara como os estimadores de beta 0 e beta 1 foram encontrados A notação matemática usada é coerente e todos os passos intermediários foram detalhados. As suposições e a natureza das variáveis foram deixados claros. Existe texto que explica o por quê dos passos usados na demonstração de beta 0 e beta 1 As suposições sobre distribuição, valor esperado e variância dos erros são colocadas de forma clara e na notação adequada É apresentado em detalhes

					como se poderia checar as suposições sobre os erros, preferencialmente com um exemplo ou relacionando com técnicas de análise (por exemplo recorrendo aos conhecimentos do Projeto 2) Explicou adequadamente como se pode fazer um teste de hipótese na regressão simples e o que a rejeição da hipótese nula significa. Menciona quais tipos de distribuições e estatística (cálculos) de teste seriam úteis para decidir o teste de hipótese. Explicou o que muda na equação (1) no caso de regressão múltipla em termos da equação e do cálculo dos betas
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.	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

			Ainda que haja a tentativa de uma análise de diagnóstico (verificação das suposições), apresenta resultados incompletos.		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