Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

WARLEY BASÍLIO BATISTA

PREVISÃO DO DESGASTE DE ELETRODOS DE GRAFITE SUBMETIDOS À ELETRÓLISE

Belo Horizonte
2011

WARLEY BASÍLIO BATISTA

PREVISÃO DO DESGASTE DE ELETRODOS DE GRAFITE SUBMETIDOS À ELETRÓLISE

Monografia apresentada ao Curso de especialização em Estatística do Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Especialista em Estatística.

Área de concentração: Indústria e Mercado

Orientador: Prof^{o}. PhD. Marcelo Azevedo Costa

Belo Horizonte

2011

PREVISÃO DO DESGASTE DE ELETRODOS DE GRAFITE SUBMETIDOS À ELETRÓLISE

Warley Basílio Batista

Marcelo Azevedo Costa

Departamento de Estatística da Universidade Federal de Minas Gerais

Novembro, 2011

Resumo

A eletrólise é um processo que separa os elementos químicos de um composto através do uso da eletricidade. De maneira sumária, procede-se primeiro à decomposição (ionização ou dissociação) do composto em íons e, posteriormente, com a passagem de uma corrente contínua através destes íons, são obtidos os elementos químicos. O entendimento da eletrólise só é possível se conhecermos o comportamento de todas as substâncias envolvidas no processo, pois cada substância se comporta de determinada maneira quando em solução e, em especial quando uma corrente elétrica atravessa essa solução. O propósito geral da **regressão múltipla** é quantificar a relação entre várias variáveis independentes ou preditoras e uma variável dependente ou critério e assim ser possível a previsão de valores futuros. O presente trabalho demonstrou a decomposição da água em seus constituintes e o tratamento estatístico pela técnica computacional da regressão múltipla na previsão de valores futuros do tempo de desgaste de eletrodos submetidos à eletrólise. As equações obtidas descrevem o comportamento do conjunto de dados e a previsão de valores futuros do tempo de desgaste de eletrodos de grafite submetidos à eletrólise.

Palavras-Chave: eletrólise, modelos de previsão, eletrodos de grafite, eletroquímica, regressão multilinear.

PREVISÃO DO DESGASTE DE ELETRODOS DE GRAFITE SUBMETIDOS À ELETRÓLISE

Warley Basílio Batista

Marcelo Azevedo Costa

Departamento de Estatística da Universidade Federal de Minas Gerais

Novembro, 2011

Abstract

Electrolysis is a process that separates the chemical elements of a compound through the use of electricity. In summary manner, proceed to the first decomposition (ionization or dissociation) of the compound into ions and then with the passage of a current through these ions, the chemical elements are obtained. The understanding of electrolysis is only possible if we know the behavior of all substances involved in the process, because each substance behaves a certain way when in solution and in particular when an electric current through this solution. The general purpose of multiple regression is to quantify the relationship between several independent or predictor variables and a dependent variable or criterion and thus be possible to forecast future values. This study demonstrated the decomposition into its constituents and the statistical technique of generalized linear models in forecasting future values of time wear of electrodes subjected to electrolysis. The equations obtained, describes the data set behavior and the prediction of future time wear values of grafita electrodes subjected to electrolysis.

***Keywords*:** electrolysis, prediction models, grafite electrodes, electrochemistry, multlinear regression.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

DMF - Dimetil Formamina

DMSO - Dimetil Sulfóxido

EDTA - Ácido Etilenodiaminotetracético

SSY - Sum Square Y (Soma dos Quadrados de Y)

SSE - Sum Square Error (Soma dos Quadrados dos Erros)

SXY - Somatório XY

SSX - Sum Square X (Soma dos Quadrados de X)

MSE - Mean Square Error (Erro Médio Quadrado)

LISTA DE SÍMBOLOS OU NOTAÇÕES

p.e. - por exemplo

e.d. - exemplo dado

p. - página

SUMÁRIO

1. Introdução

Brett e Brett (1994), afirmam que o escopo da eletroquímica envolve fenômenos químicos associados com separação de cargas. Com freqüência essa separação de cargas direciona a transferência de carga, a qual pode ocorrer homogeneamente em solução, ou heterogeneamente em superfícies de eletrodos. Na realidade, para assumir eletro neutralidade, duas ou mais semi-reações de transferência de carga tomam lugar, em direções opostas. Exceto no caso de reações redox homogêneas, estas são separadas em espaço, usualmente ocorrendo em diferentes eletrodos imersos em solução numa célula.

A eletrólise é um processo eletroquímico, caracterizado pela ocorrência de **reações de oxirredução** em uma solução condutora quando se estabelece uma diferença de potencial elétrico entre dois (ou mais) eletrodos mergulhados nessa solução. Vale lembrar que a denominação **solução eletrolítica**, empregada para designar qualquer solução aquosa condutora de eletricidade, deriva justamente desse processo.

A necessidade de controlar a corrosão quase sempre se reduz a considerações de segurança e economia. Máquinas, equipamentos, e produtos funcionais podem falhar devido à corrosão de tal maneira a resultar em ferimento pessoal. Por causa da escolha dos materiais, reforço de procedimentos de manufatura e controle de produtos para minimizar dano pessoal, tudo envolve considerações econômicas, implantação de medidas seguras e não somente envolvem interesses humanitários, mas, também econômicos. Com todas as decisões econômicas, a base para ação é um ajuste entre os benefícios gerados por certo nível de controle da corrosão versus os custos que irão resultar se aquele nível de controle não for mantido (STANSBURY; BUCHANAN, 2000, p.2).

Estatística, em um sentido limitado, é um ramo da ciência que lida com a tirada de conclusões sobre populações baseada em amostras. Em um sentido mais amplo, estatística encerra coleção, organização, e sumarização de dados; apresentação de dados em forma tabular e gráfica; desenvolvimento de modelos para o propósito de entendimento de fenômenos aleatórios e não-aleatórios; uso de modelos para previsão; aproximação matemática para tomada de decisão e estimação de riscos; e assim por diante. Análise de Regressão (e correlação) é uma área da estatística que lida com métodos para investigar a existência de associações e, se presente, a natureza da associação, entre várias quantidades observáveis. Análise de regressão é um método de investigar a presença de associações se dados apropriados estão disponíveis. A descoberta de associações e a habilidade de expressar

tais associações em uma forma matemática precisa *pode*m possibilitar aquele a prever o valor não observável de uma variável baseado no valor observado de uma ou mais variáveis associadas ou relacionadas. Elas podem também ajudar determinar como uma pode controlar os valores de outra variável por manipular os valores de uma variável relacionada. Isto pode ser confirmado somente por investigações experimentais controladas (GRAYBILL; IYER, 1994; p.1).

Objetiva-se o trabalho, no entendimento da dinâmica de corrosão eletroquímica de eletrodos de grafite quando submetidos à eletrólise. O tratamento estatístico abordado permite a previsão de valores futuros e modelagem do comportamento da série de dados.

Justifica-se no sentido de levantar as informações necessárias para atuar de forma orientada na prevenção, previsão e antecipação do processo de corrosão eletroquímico.

O restante do trabalho está estruturado da seguinte forma, na próxima seção um resumo dos principais conceitos relacionados ao trabalho. A seção 3 descrição dos dados. A seção 4 análise dos resultados da aplicação da metodologia nos dados. Finalmente, a seção 5 alguns comentários finais.

2. Metodologia

Capítulo 1: Considerações Acerca dos Métodos Eletroquímicos de Análise e o Processo de Corrosão Eletroquímica

2.1 Métodos Eletroquímicos de Análise

Segundo Bard e Faulkner (2001) em 20 anos, desde que apareceu sua primeira edição, os campos da eletroquímica e química eletroanalítica se desenvolveu substancialmente. Um entendimento apurado do fenômeno, o providencial desenvolvimento de ferramentas experimentais já conhecidas em 1980, e a introdução de novos métodos foram todos importantes para essa evolução.

Ainda, de acordo com os mesmo autores, eletroquímica é a área da química preocupada com a inter-relação da elétrica e efeitos químicos. A maior parte desse campo lida com o estudo das mudanças químicas causadas pela passagem de uma corrente elétrica e a produção de energia elétrica por reações químicas. O campo da eletroquímica norteia um extenso cursor de diferentes fenômenos (p.e.; eletroforese e corrosão), dispositivos (displays eletrocrômicos, sensores eletroanalíticos, baterias e células combustível), e tecnologias (a eletrodeposição de metais e a produção em larga-escala de alumínio e cloro).

De acordo com Bagotsky (2006), estudos experimentais em eletroquímica lidam com o misto de propriedades dos eletrólitos (condutividade, etc.); potenciais de eletrodo no equilíbrio e não-equilíbrio; a estrutura, propriedades, e condições da interface entre diferentes fases (eletrólitos e condutores eletrônicos, outros eletrólitos, ou isolantes); e a natureza, cinética e mecanismo das reações eletroquímicas. Técnicas eletroquímicas bem como as não eletroquímicas são usadas quando estudando estes aspectos. Técnicas eletroquímicas são comumente usadas, também, em análises químicas, na determinação de propriedades de várias substâncias e para outros propósitos.

Por sua vez, Joseph Wang (2001) situa as técnicas eletroanalíticas no limiar entre eletricidade e química, nomeando-as como medidas de quantidades elétricas, tais como corrente, potencial ou carga e sua relação com os parâmetros químicos. Tal uso das medidas elétricas para fins analíticos tem encontrado vasta faixa de aplicações, incluindo monitoramento ambiental, controle de qualidade industrial, e análises biomédicas. Avanços nas décadas de 1980 e 1990 incluindo o desenvolvimento de ultramicroeletrodos, (...)

lideraram a um substancial crescimento na popularidade da eletroanálise, e à sua expansão em novas fases e empreendimentos.

2.2 Visão Geral da Corrosão Eletroquímica

Definição e exemplos de corrosão

2.2.1 A Necessidade de Controlar a Corrosão

A necessidade de controlar a corrosão quase sempre se reduz a considerações de segurança e economia. Máquinas, equipamentos, e produtos funcionais podem falhar devido à corrosão de tal maneira a resultar em ferimento pessoal. Por causa da escolha dos materiais, reforço de procedimentos de manufatura e controle de produtos para minimizar dano pessoal, tudo envolve considerações econômicas, implantação de medidas seguras e não somente envolvem interesses humanitários, mas, também econômicos. Com todas as decisões econômicas, a base para ação é um ajuste entre os benefícios gerados por certo nível de controle da corrosão versus os custos que irão resultar se aquele nível de controle não for mantido (STANSBURY; BUCHANAN, 2000, p.2).

2.2.2 Processo de Corrosão Eletroquímico e Variáveis

Antes de examinar em detalhes as teorias dos processos de corrosão aquosos e as bases para fazer os cálculos quantitativos das taxas de corrosão, será útil desenvolver qualitativamente o principal fenômeno envolvido. As seções a seguir revisam vários tipos gerais de combinações metal/ambiente corrosivo, as reações químicas envolvidas, mecanismos idealizados para a transferência de íons metálicos ao ambiente, e os processos eletroquímicos ocorrendo na interface entre o metal e o ambiente aquoso.

2.2.2.1 Corrosão Uniforme com o pH como a Variável Principal

Para metais, M, que são termodinamicamente instáveis em água, as mais simples reações de corrosão são:

$$M + \quad \rightarrow \quad + \frac{\cdots}{2} \quad - \qquad \text{no } pH < \qquad \text{(eq. 1.1.1)}$$

$$M + \quad \rightarrow \quad + \quad + \frac{\cdots}{2} \qquad \text{no } pH \geq \qquad \text{(eq. 1.1.2)}$$

Portanto, o metal passa do estado metálico para íons de valência *m* em solução com evolução do hidrogênio. A reação é considerada ser diretamente com íons hidrogênio em solução ácida e progressivamente com moléculas de água assim que o pH aumenta para condições neutras e alcalinas. Dois processos são envolvidos na reação, com cada um

envolvendo uma mudança na carga: M para M^{m+}e mH^{+} para $\frac{m}{2}H_2$ (em solução ácida). As mudanças na carga são acompanhadas por transferência de elétrons de M para H^{+}. Pela razão de a fase metálica ser um condutor elétrico, sustenta a transferência de elétrons, permitindo aos dois processos ocorrerem em sítios separados na superfície do metal. Em casos limitantes, esses processos ocorrem dentro de poucos diâmetros atômicos na superfície com os sítios constantemente mudando com o tempo, desse modo, produzindo corrosão uniforme. De outro modo, a corrosão é não uniforme. Corrosão uniforme sustentada pelo pH é representada esquematicamente na Figura 1. Neste exemplo, oxigênio é excluído por um gás de nitrogênio purgado e encoberto.

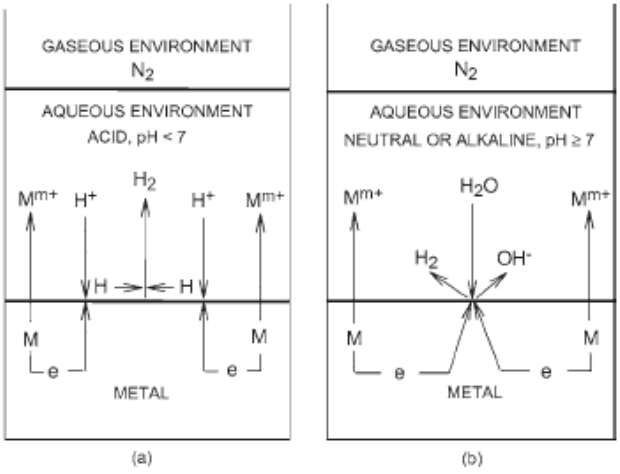


Figura 1: Corrosão Uniforme Sustentada por *pH* controlado (oxigênio excluído, desaerado). (a) Ácido, *pH* < 7. (b) Neutro ou Alcalino, *pH*≥ 7

2.2.2.2 Corrosão Uniforme com pH e Oxigênio Dissolvidos como Variáveis

Quando oxigênio dissolvido está presente na solução, usualmente pelo contato com ar (ambiente aerado), as reações se aplicam *em adição* àquelas já consideradas.

$$M + \frac{\cdots}{4} - + \quad \rightarrow \quad + \frac{\cdots}{2} - \qquad \text{no } pH < \qquad \text{(eq. 1.1.3)}$$

$$M + \frac{\cdots}{4} \quad + \frac{\cdots}{2} \quad \rightarrow \quad + \qquad \text{no } pH \geq \qquad \text{(eq. 1.1.4)}$$

Corrosão uniforme sustentada por oxigênio dissolvido e pH é representada esquematicamente na Figura 2. Visto que os elétrons são agora consumidos por duas reações, a taxa de corrosão do metal aumenta. No caso do aço, oxigênio dissolvido é mais importante no sustento da corrosão do que a presença de íons hidrogênio quando o pH é maior que aproximadamente 4.

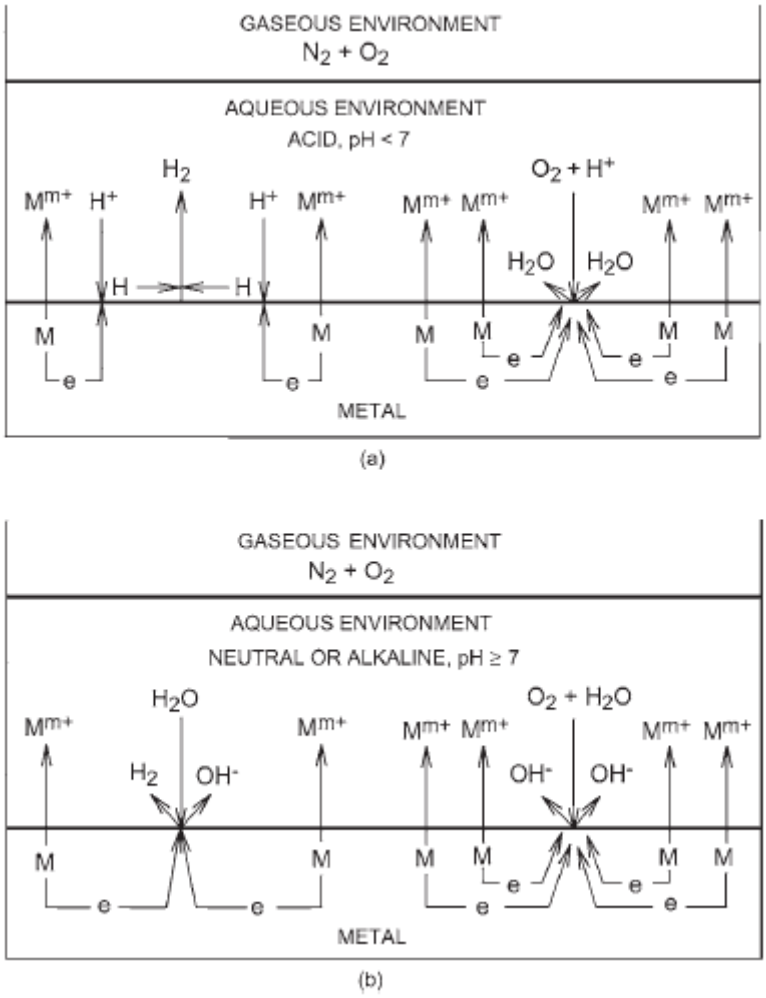


Figura 2: Corrosão Uniforme sustentada pelo *pH* e oxigênio dissolvido (aerado). (a) Ácidos, $pH < 7$. (b) Neutro ou Alcalino, $pH \geq 7$

2.2.3 Corrosão Uniforme com Formação de Produto de Corrosão

Um exemplo da formação de produto de corrosão é a ferrugem do ferro como ilustrado na Figura 3. Quando o *pH* é maior do que aproximadamente 4, e sob condições aeradas, uma camada de Fe_3O_4 negro, e possivelmente $Fe(OH)_2$, se forma em contato com o substrato de ferro. Na presença do oxigênio dissolvido, uma camada externa de Fe_2O_3 vermelho ou FeOOH se forma. A aderência e porosidade dessas camadas mudam com o tempo e podem ser influenciadas por outras espécies químicas no ambiente, tais como íons cloreto e sulfato. Em qualquer caso, a formação da camada de produto de corrosão influencia a taxa de corrosão por introduzir uma barreira pela qual íons e oxigênio devem difundir-se para sustentar o processo de corrosão.

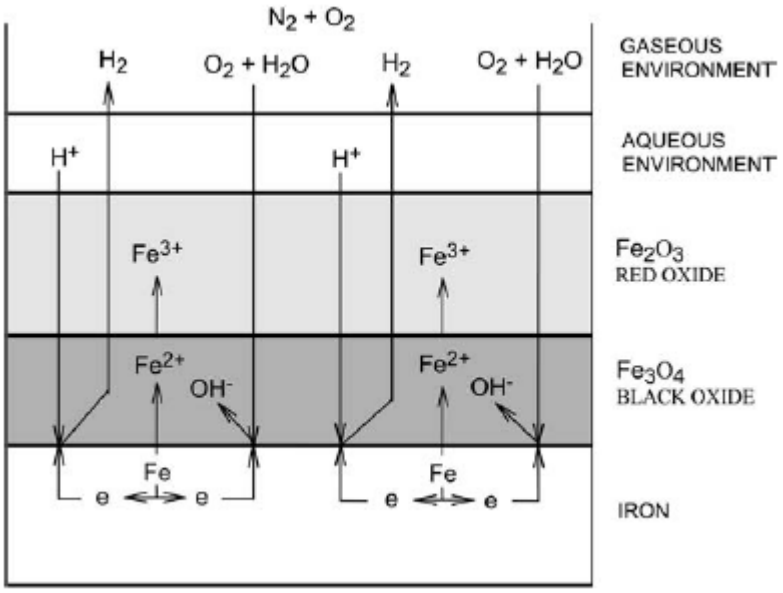


Figura 3: Corrosão Uniforme com Depósito de Produto Sólido de Corrosão. Detalhes da Formação de Espécies de Óxidos não são Consideradas neste Ponto.

2.3 Sistema, Instrumentação e Componentes da Célula Eletroquímica

A instrumentação básica requerida para experimentos potencial-controlado é relativamente barata e prontamente disponível comercialmente.

2.3.1 Solventes e Eletrólitos Facilitadores

Medidas eletroquímicas são comumente levadas a cabo em um meio que consiste de solvente contendo um eletrólito facilitador. A escolha do solvente é ditada primariamente pela solubilidade do analito e sua atividade redox, e pelas propriedades do solvente tais como condutividade elétrica, atividade eletroquímica, e reatividade química. O solvente não deve reagir com o analito (ou produtos) e não deve sofrer reações eletroquímicas acima de uma extensa faixa de potencial.

Enquanto água tem sido usada como um solvente mais do que quaisquer outros meios, solventes não aquosos [p.e., acetonitrila, propileno carbonato, dimetilformamina (DMF), dimetil sulfóxido (DMSO), ou metanol] tem também freqüentemente sido usados. Solventes misturados podem também ser considerados para certas aplicações. Água bi-destilada é adequada para a maioria dos trabalhos em meios aquosos. Água tri-destilada é freqüentemente requerida quando análise traço (stripping) são de interesse. Solventes orgânicos com freqüência requerem secagem ou procedimentos de purificação.

Eletrólitos facilitadores são requeridos em experimentos potencial-controlado para reduzir a resistência da solução, eliminar efeitos de eletromigração, e manter uma constante iônica forte (p.e.., “desatolar” o efeito de variáveis quantidades de acontecimento

naturalmente eletrolítico). Os eletrólitos facilitadores inertes podem ser um sal inorgânico, um ácido mineral, ou um tampão. Enquanto cloreto ou nitrato de potássio, cloreto de amônio, hidróxido de sódio, ou ácido clorídrico são amplamente usados quando usando água como solvente; sais de tetraalquilamonio são empregados em meios orgânicos. Sistemas tampão (tais como acetato, fosfato, ou citrato) são usados quando um controle de *pH* é essencial. A composição do eletrólito pode afetar a seletividade de medidas voltamétricas. Por exemplo, a tendência da maioria dos eletrólitos complexarem com íons de metal podem beneficiar a análise de misturas de metais. Além disso, agentes mascarantes [ácido etilenodiaminotetracético (EDTA)] pode ser adicionado para "remover" interferências indesejáveis. O eletrólito facilitador deve ser preparado a partir de reagentes altamente purificados, e não deve ser facilmente oxidado ou reduzido. A faixa de concentração usual é 0.1 – 1.0M, em outras palavras, em grande excesso de concentração de todas as espécies eletroativas. Significativamente níveis mais baixos podem ser empregados em conexão com eletrodos de ultramicro trabalho.

2.3.2 Instrumentação

Avanços rápidos em microeletrônica, e em particular a introdução de amplificadores operacionais, tem levado a maiores mudanças em instrumentação eletroanalítica. Circuitos integrados minúsculos e baratos podem agora realizar muitas funções que previamente necessitavam de instrumentos muito grandes. Tais instrumentos consistem de dois circuitos: um circuito polarizante que aplica o potencial à célula e um circuito de medição que monitora a corrente da célula. (Figura 4).

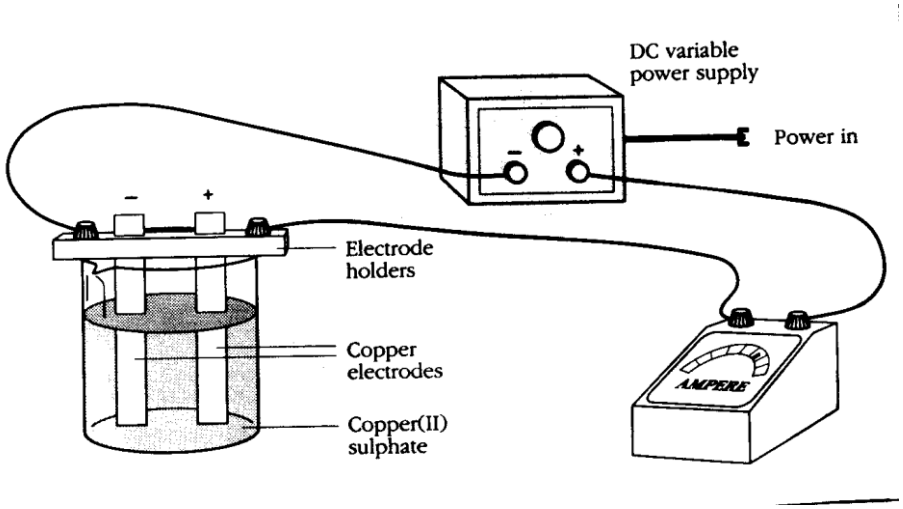


Figura 4: Esquema de Instrumentação Empregada em Células Eletrolíticas para Polarização e Monitoramento da Corrente

2.3.3 Eletrodos de Trabalho

A performance do procedimento voltamétrico é fortemente influenciada pelo material do eletrodo de trabalho. O eletrodo de trabalho deve prover altos sinais de ruído característicos, bem como uma resposta reprodutível. Assim, sua seleção depende primariamente de dois fatores: o comportamento redox do analito alvo e a corrente de fundo sobre a região de potencial requerida para a medida. Outras considerações incluem a janela de potencial, condutividade elétrica, reprodutibilidade de superfície, propriedades mecânicas, custo, disponibilidade, e toxicidade. Uma faixa de materiais tem encontrado aplicação como eletrodos de trabalho para eletroanálises. Os mais populares são aquele envolvendo mercúrio, carbono ou metais nobres (particularmente platina e ouro). A geometria destes eletrodos deve também ser considerada.

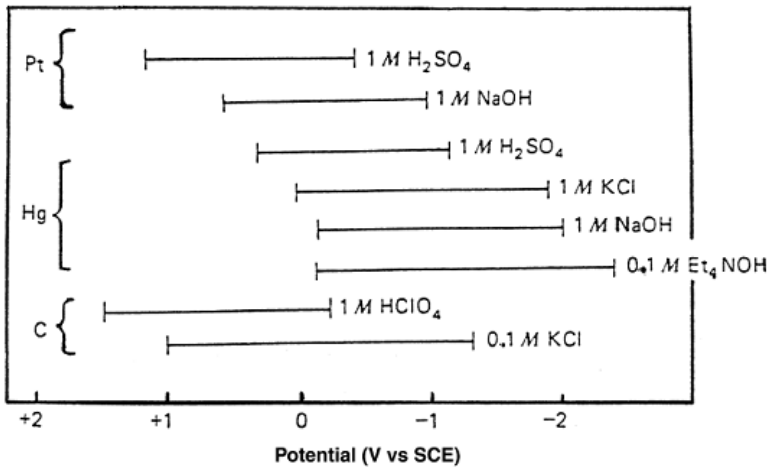


Figura 5: Janela de potencial acessível de eletrodos de platina, mercúrio e carbono em vários eletrólitos suportadores

2.3.4 Eletrodos Sólidos

A limitada faixa de potencial anódico de eletrodos de mercúrio tem impedido sua utilidade para monitoramento de compostos oxidantes. Adequadamente, eletrodos sólidos com janelas de potencial anódico estendido têm atraído interesse analítico considerável. Dos muitos diferentes materiais sólidos que podem ser usados como eletrodos de trabalho, os mais freqüentemente usados são carbono, platina e ouro. Prata, níquel e cobre podem também ser usados para aplicações específicas.

Um importante fator no uso de eletrodos sólidos é a dependência da resposta no estado de superfície do eletrodo. Adequadamente, o uso de tais eletrodos requer preciso pré-tratamento do eletrodo e polimento para obter resultados reprodutíveis. A natureza destes passos de pré-tratamento depende dos materiais envolvidos. Polimento mecânico (para um acabamento liso) e potencial circulante são comumente usados para eletrodos de metal, enquanto vários produtos químicos, eletroquímicos ou procedimentos térmicos de superfície são acrescentados para ativação de eletrodos de base-carbono. Diferente de eletrodos de mercúrio, eletrodos sólidos apresentam uma superfície heterogênea com respeito à atividade eletroquímica. Tal heterogeneidade superficial leva a desvios do comportamento esperado para superfícies homogêneas.

Capítulo 2: Inferência Estatística, Métodos de Previsão e Modelos de Regressão

2.4 Ingredientes Básicos para Inferência Estatística

Um dos propósitos da ciência é relacionar, descrever e prever eventos no mundo em que vivemos. Estas atividades são também importantes em negócios e nossas ocupações diárias. Em quase todo aspecto do empenho humano, é útil ser capaz de prever eventos futuros baseados na informação presente e passada. (GRAYBILL; IYER, 1994)

Para descrever situações de interesse, devemos defini-las precisamente e decidir o que queremos determinar ou prever.

2.4.1 Conceitos

Modelo — Um modelo de população é uma descrição das quantidades (número de atributos) de interesse associado com cada item na população.

Parâmetro Parâmetros são resumo de números que caracterizam vários aspectos da população e são usualmente as quantidades de interesse para o investigador.

Amostra Uma amostra é um conjunto de itens selecionados da população, e observações são feitas neste conjunto de itens. Com base nesta amostra uma decisão é feita sobre os valores dos parâmetros de interesse, e esta decisão é usualmente acompanhada por uma medida da incerteza na resposta.

2.4.1.1 Modelo

Uma população uni-variada para ser estudada consiste de *N* números, onde *N* é geralmente muito grande.

Para estudar uma população de *N* números precisa-se ter algum modo de organizar estes *N* números. Uma útil aproximação é organizá-los em um **histograma de probabilidade** e encontrar uma função matemática que aproxima este histograma. Tal função matemática é usualmente chamada **função densidade de probabilidade**. Estatísticos teóricos usam uma variedade de funções densidade de probabilidade para estudar diferentes tipos de populações, a mais importante entre elas sendo a função densidade de probabilidade **Gaussiana**.

2.4.2 Populações Gaussianas

Uma população Gaussiana teórica Y é completamente especificada pela sua média, denotada por μ , e seu desvio padrão, denotado por σ . A distribuição desta população é descrita pela função densidade de probabilidade

$$f\ Y = \frac{1}{\sqrt{2\pi\sigma}} \quad \text{para} \quad -\infty < \ < \infty \qquad \text{eq. 1.2.1}$$

A área sob a curva, definida pela função em (1.2.1) entre os valores *a* e *b* $a <$, dá a proporção dos valores da população que são maiores que *a*, mas, menores que ou iguais a *b* (Figura 6).

As funções densidade de probabilidade correspondendo a duas diferentes populações Gaussianas teóricas são mostradas na Figura 7. Estas duas populações têm diferentes valores para μ , mas elas têm o mesmo desvio padrão σ . Está claro a partir desta figura que mudando o valor de μ muda-se somente a localização da distribuição da população e não sua forma (curvas (a) e (b) na figura 7 tem a mesma forma, mas, diferentes locais). Por outro lado, mudando o valor de σ muda-se a forma da distribuição da população, mas não a localização,

como exemplificado pelas curvas (c) e (d) na Figura 8. Essas duas curvas têm a mesma localização (mesmo valor de μ), mas, diferentes formas (diferentes valores de σ).

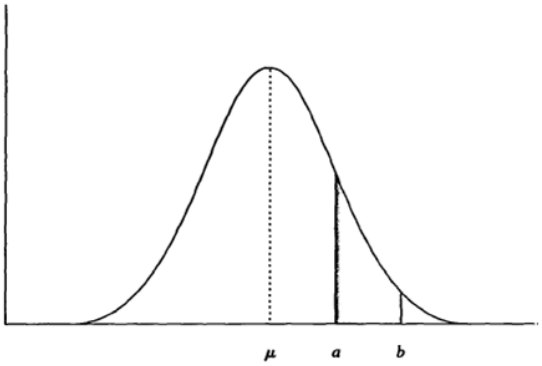


Figura 6: Área sob a curva definida em 1.2.1

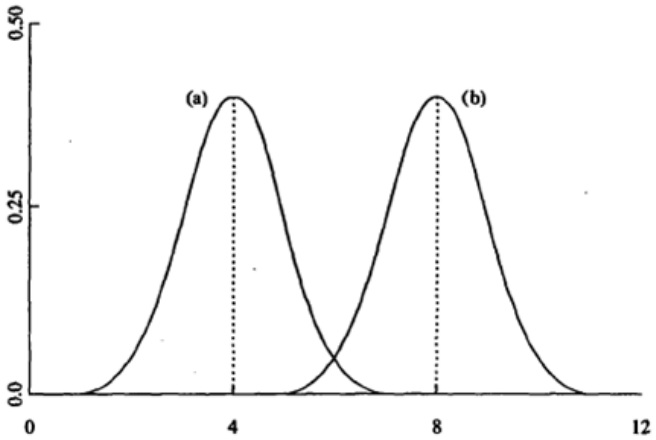


Figura 7: Função densidade de probabilidade de duas diferentes populações Gaussianas teóricas

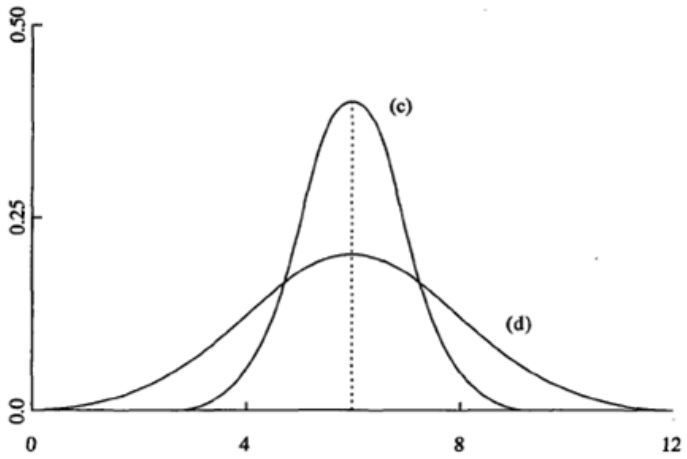


Figura 8: Duas populações com médias iguais e diferentes desvios padrão

2.4.2.1 Média

A média (também chamada *termo médio*) de uma população Y é denotada por μ e é definida por

$$\mu_{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad \text{eq. 1.2.2}$$

2.4.2.2 Desvio padrão

O desvio padrão de uma população Y é denotado por σ_{Y} e é definido por

$$\sigma_{Y} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (Y_i - \mu_{Y})^2} \qquad \text{eq. 1.2.3}$$

2.4.2.3 Variância

A variância de uma população Y é o quadrado do seu desvio padrão. É denotado por σ_{Y}^2 e é definido por

$$\sigma_{Y}^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \mu_{Y})^2 \qquad \text{eq. 1.2.4}$$

As definições de média e desvio padrão para populações infinitas requerem conceitos de cálculo e algum conhecimento sobre convergência de séries infinitas que, não serão abordadas.

Como declarado antecipadamente, em problemas práticos os N desconhecidos números que constituem uma população poderiam conceitualmente ser usados para formar um **histograma de probabilidade** (um histograma para o qual a área total se iguala a 1). Se este histograma é bem aproximado pela função matemática definida na eq. 1.2.1 quando o valor da média e o desvio padrão destes N números da população são substituídos por μ_{Y} e σ_{Y}, então *podemos proceder como se a população sob estudo seja Gaussiana.*

A população teórica Gaussiana é uma abstração matemática e tal população não pode existir em investigações reais, mas é útil como uma aproximação para populações finitas em muitos problemas aplicados. É também usada em teoria estatística para derivar pontos estimados, intervalos de confiança, e testes para μ e σ. Aproximações similares são comuns em outras situações.

Pela razão de nunca sabermos os valores da população em um problema real, é impossível estar certo que a população sob estudo é na verdade Gaussiana. Contudo, mesmo quando a população não é exatamente Gaussiana, os procedimentos de inferência estatística são freqüentemente precisos o bastante para tomar decisões contanto que a população seja *aproximadamente* Gaussiana. Em alguns casos procedimentos estatísticos estão disponíveis para detecção de sérias violações dos pressupostos Gaussianos.

2.4.3 Parâmetros (Resumo de Números)

(...). Ainda se tivéssemos a população inteira disponível (o que é raramente o caso em um problema real), existiriam números demais para um investigador usar para tomar decisões sem resumi-los em um menor grupo. Assim uma criteriosa sumarização de características da população é extremamente importante.

2.4.3.1 Parâmetros para Populações Univariadas

Considere a população uni-variada Y e assuma para o momento que todos os números na população estão disponíveis. Para entender as características importantes da população de números, é conveniente resumi-los por um menor conjunto de números ou talvez pelo uso de técnicas gráficas adequadas. Histogramas de probabilidade e curvas de freqüência cumulativa são duas das comumente usadas descrições gráficas das populações.

Geralmente se gostaria de usar um ou mais (dito m) resumo de números, θ θ , para descrever várias características da população inteira. Estes m números, chamados **parâmetros** (populacionais), poderiam ser computados a partir da população inteira se fossem conhecidos.

Embora os m parâmetros θ θ não possam, em geral, dizer nos tudo sobre a população inteira, eles com freqüência resumem adequadamente certas características importantes da população que são relevantes ao estudo em mãos. A **média** e o **desvio padrão** são dois particularmente úteis e importantes parâmetros da população, especialmente no caso de populações Gaussianas.

2.4.3.1.1 Média

Muitas vezes gostaríamos de usar um *parâmetro único* para *representar* a população inteira de números. Neste texto usamos a **média** da população (também chamada **termo médio**), definida na eq. 1.2.2, como um só número que melhor representa a população inteira de números. O símbolo μ (μ é a letra Grega mu) representa a média populacional, e o subscrito Y indica qual população está sob estudo. A média μ é com freqüência também usada para prever o valor Y de qualquer item populacional escolhido aleatoriamente. Na maioria de problemas reais a média de uma população não é conhecida porque nem todos os elementos Y da população são conhecidos. Todavia, é um número que gostaríamos de ter disponível para usar na tomada de decisões sobre uma população. Então selecionamos uma amostra da população e estimamos μ .

Em muitas situações, números singulares tal como a média (ou a mediana) são freqüentemente usados para representar o valor de cada item em uma população.

2.4.3.1.2 Desvio padrão

Embora a média μ é com freqüência usada como o melhor número único para representar a população inteira de números Y, nenhum só número pode adequadamente descrever ou representar uma população inteira. Conseqüentemente um adicional resumo de números é usado para nos dizer quão bem μ representa a população inteira. Este número é σ, o desvio padrão da população, o qual foi definido na eq. 1.2.3. Intuitivamente, quanto menor σ é, mais útil μ é como um valor representativo para a população inteira. Da mesma forma, quanto maior σ é, menos útil μ é como um representativo para toda a população. Note que se $\sigma =$, então todos os números na população Y são o mesmo e iguais a μ, e a média é uma perfeita representação da população inteira. Mas se a proporção substancial de valores de Y são muito menores do que a média e outros são muito maiores, então σ serão grandes e μ pode não representar adequadamente cada valor populacional.

2.4.3.2 Parâmetros para Populações Multivariadas

Suponha que uma população de tamanho *N* é uma população *k*-variada $k >$. Então existem *k* quantidades associadas com cada item da população, resultando em *Nk* números no todo. Cada uma das *k* quantidades associadas com os itens da população originam uma população uni-variada. Deixando as *k* quantidades para o item *I* denotadas por $X_{I1},...,X_{IK}$. Então os números $X_{I1},...,X_{N1}$ formam uma população uni-variada com média μ e desvio padrão σ, os números $X_{I2},...,X_{N2}$ formam outra população uni-variada com média μ e desvio padrão σ, etc. Assim μ μ, e σ σ são parâmetros associados com a população *k*-variada. Os *Nk* números podem ser esquematicamente representados como na Tabela 1.

Tabela 1: Representação Esquemática de uma População *k*-variada de Tamanho *N*

	K medidas em cada item			
Itens	1	2	…	*k*
1	X_{11}	X_{12}	…	X_{1k}
2	X_{21}	X_{22}	…	X_{2k}
⋮	⋮	⋮	⋮	⋮
I	X_{I1}	X_{I2}	…	X_{Ik}
⋮	⋮	⋮	⋮	⋮
N	X_{N1}	X_{N2}	…	X_{Nk}
Média	μ	μ	…	μ
Desvio padrão	σ	σ	…	σ

Fonte: Adaptado de (GRAYBILL; IYER, 1994)

2.4.3.2.1 Coeficiente de Correlação

Quando lidando com populações multivariadas existe freqüentemente a necessidade de resumir associações entre várias quantidades medidas no mesmo item. Um sumário de medida que é algumas vezes usado para este propósito é o **coeficiente de correlação** entre duas variáveis (também chamado **correlação de Pearson** ou **correlação produto do momento**). Esta medida de associação é denotada por $\rho_{\cdot,\cdot\cdot}$ (ρ é a letra Grega rho) quando resumindo o relacionamento entre as variáveis Y e X. É definido por

$$\rho_{\cdot,\cdot\cdot} = \frac{\sum \quad \_ \quad \_\_}{\sqrt{[\sum \quad \_\_ \sum \quad \_}} \qquad \text{eq. 1.2.5}$$

Para uma população bivariada Y, X de N itens. Note que a eq. 1.2.5 implica que $\rho_{\cdot,\cdot\cdot} = \_\_,\_\_$; p.e., o coeficiente de correlação entre Y e X é o mesmo que o coeficiente de correlação entre X e Y. Pode ser mostrado que $\rho_{\cdot,\cdot\cdot}$ é um número entre -1 e +1. É igual a +1 quando os valores da população Y_I, X_I todos sobrepõem uma linha reta que tem uma inclinação positiva, e é igual a -1 quando todos os valores populacionais sobrepõem uma linha reta com inclinação negativa. Figuras 9-11 são *gráficos de dispersão* de populações bivariadas Y, X consistindo de 500 itens, cada uma com diferente valor de $\rho_{\cdot,\cdot\cdot}$.

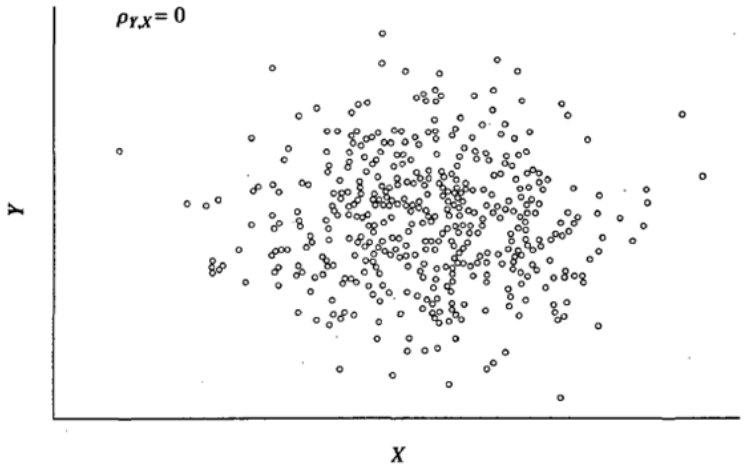


Figura 9: População com Coeficiente de Correlação $\rho_{\cdot,\cdot\cdot} =$

Observe que um valor positivo do coeficiente de correlação entre Y e X indica que, geralmente falando, maiores valores de Y são associados com maiores valores de X, e menores valores de Y são associados com menores valores de X. Da mesma forma, quando o coeficiente de correlação é negativo, encontramos que maiores valores de Y são associados com menores valores de X. Perceba que uma carência geral de associação *linear* é indicada quando a magnitude do coeficiente de correlação é próxima à zero. *Enquanto coeficientes de correlação podem ter proveitosas interpretações para alguns problemas* (*particularmente quando* Y *e* X *aproximadamente relacionados linearmente*)*, eles podem não fornecer proveitosa interpretação e, de fato, ser um resumo enganoso quantitativo em outros problemas.*

Em resumo, quando lidando com uma população *k*-variada, dito $X_1,...,X_k$, o resumo básico de quantidades ou parâmetros que são freqüentemente usados são as médias, μ μ, e desvios padrão, σ σ, de populações k uni-variadas, junto com os $\binom{\cdot}{\cdot}$ coeficientes de correlação, $\rho_{\cdot\cdot,\cdot\cdot}$, $\rho_{\cdot\cdot,\cdot\cdot}$ $\rho_{\cdot\cdot}$, entre os k $k-$ pares de variáveis X_1, X_2, X_1, X_3, ..., X_{k-} .

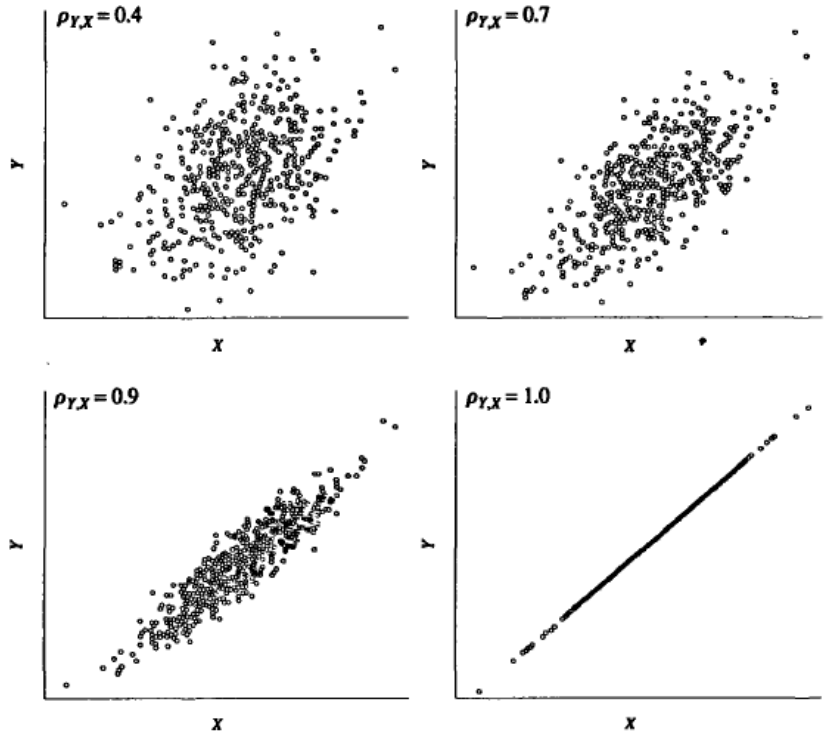


Figura 10: Populações com crescentes Coeficientes de correlação

2.5 Amostra e Inferências

Como declarado anteriormente, um investigador deve primeiro definir a população de estudo. Sempre que possível a população alvo por si deveria ser a população de estudo. Se isto não é possível, estão a população de estudo deveria parecer-se a população alvo tão estreitamente quanto possível. Depois que a população de estudo é definida, é descrita com um modelo, e parâmetros que são necessários para se tomar decisões são identificados. O próximo passo é determinar os valores desses parâmetros. Visto que a população nunca é completamente conhecida em uma situação real, um investigador pode nunca saber os valores de parâmetro exatamente. Um procedimento comumente usado é selecionar um subconjunto dos itens, referidos como **amostra**, a partir da população de estudo e usar as medidas associadas com esses itens amostrais para *inferir* os valores dos parâmetros populacionais de interesse. Se a amostra é selecionada da população usando um dos vários procedimentos de amostragem aleatória, então é possível atribuir uma *medida de incerteza* para as conclusões derivadas usando tal amostra. O processo de fazer inferências sobre os valores dos parâmetros populacionais baseado em amostras aleatórias é chamado *inferência estatística*.

2.5.1 Amostra Aleatória Simples

O tamanho *n* da amostra a ser selecionado é determinado pelo investigador baseado em uma cuidadosa consideração de custos e objetivos do estudo. Amostras podem ser selecionadas da população de estudo usando qualquer um dos vários procedimentos de amostragem aleatória que são discutidos em livros texto em métodos de amostragem. Um profundo entendimento das vantagens e desvantagens dos vários métodos de amostragem é necessário antes de um dos procedimentos ser selecionado. O mais simples de todos os procedimentos de amostragem é um chamado *amostragem aleatória simples*, e uma amostra obtida usando este procedimento é chamada **amostra aleatória simples**. A definição de uma amostra aleatória simples segue:

DEFINIÇÃO **Amostra Aleatória Simples**

Se uma população tem *N* elementos, existem $H = \binom{\ }{\ }$ $\frac{\cdots \quad +}{1x2x\cdots}$ amostras distintas de tamanho *n* que podem ser obtidas. Se cada uma dessas *H* amostras têm uma igual chance de serem selecionadas, então a amostra atualmente obtida é chamada *amostra aleatória simples de tamanho n*.

2.5.2 Ponto Estimador

Supondo θ ser um parâmetro populacional desconhecido (θ poderia ser μ , σ , σ ρ , etc.). Um **ponto estimado** de θ é um *número* calculado a partir de amostra de dados, para ser usado por um investigador como o valor de θ (porque θ é desconhecido e daí indisponível) na tomada de decisões. Procedimentos de estimativa para calcular pontos estimadores são úteis quando seus valores estão próximos aos valores atuais de parâmetros populacionais desconhecidos.

2.5.2.1 Estimadores da Média e Desvio Padrão de uma População Uni-variada

Se $y_1, y_2, ..., y_n$ é uma amostra aleatória simples de uma população Y cuja média é μ e cujo desvio padrão é σ , o estimador comumente usado de μ e σ , denotado por $\hat{\mu}$ e $\hat{\sigma}$, respectivamente são

$$\hat{\mu} = = \frac{1}{n}\sum_{i=} \qquad \text{eq. 1.2.6}$$

e

$$\hat{\sigma} = \sqrt{\frac{SSY}{n-}} \qquad \text{eq. 1.2.7}$$

Onde SSY é chamado soma dos quadrados ajustado para Y, e é definido por

$$SSY = \sum_{i=}$$

O estimado de σ é $\hat{\sigma}^2$, o quadrado do desvio padrão estimado $\hat{\sigma}$. Note que

$$SSY = \quad - \quad \hat{}^2$$

2.5.2.2 Estimador do Coeficiente de Correlação ρ em uma População Bivariada

Se $y_1, x_1, \ldots, y_n, x_n$ é uma amostra aleatória simples de uma população bivariada Y, X, o estimador do coeficiente de correlação entre Y e X que é muito utilizado é dado por

$$\rho_{\_,\ldots} = \frac{\sum \quad -}{\sqrt{\sum \quad \sum}}$$

2.5.3 Estimador Não Viciado

Para avaliar se um *procedimento* para estimar um parâmetro populacional é um bom procedimento, *devemos investigar o estimador dado pelo procedimento para cada amostra que pudesse ser obtida*. Considere o procedimento exatamente dado de usar a média amostral para estimar a média populacional. Se uma população Y tem N itens e se o tamanho da amostra é n, então existem $H = \binom{\ }{\ }$ amostras possíveis de tamanho n, qualquer uma das quais pudesse ser selecionada e, sob amostragem aleatória simples, cada uma destas amostras possíveis tem a mesma probabilidade de ser a verdadeira amostra escolhida. Conceitualmente, para uma dessas H amostras possíveis de tamanho n, poderíamos computar a média amostral. Isto resultaria em H médias amostrais $\bar{y}_1, \ldots, \bar{y}_H$, algumas das quais estariam próximas à média populacional e outras não estariam; algumas seriam maiores do que a média populacional e outras seriam menores, mas qualquer uma dessas H médias amostrais possíveis são iguais à média populacional. Expressamos este fato por dizer que a *média amostral é um estimador não-viciado da média populacional*. Muitos investigadores e estatísticos consideram procedimentos de estimação não-viciados desejáveis.

Mais geralmente, supondo que queremos estimar um parâmetro θ de uma população. Selecionamos uma amostra aleatória de tamanho *n* da população e computa-se um estimador θ de θ usando os valores amostrais de acordo com algum procedimento ou fórmula. Em princípio, para cada amostra de tamanho *n* possível (recordar que existem *H* destas), um estimador pode ser obtido usando o mesmo procedimento. Isto resultaria em *H* estimadores θ θ. Se a média destes *H* estimadores, θ θ, de θ são iguais a θ, então o procedimento usado para computar estes estimadores é dito ser um *procedimento de estimação não-viciado*. É Claro, em qualquer problema dado calculamos somente um destes *H* valores de θ porque temos somente uma amostra de tamanho *n* disponível. Se calcularmos este único valor de θ usando um procedimento de estimação não-viciado, então se diz que θ é um **estimador não-viciado** de θ.

Para resumir, um ponto estimado de um parâmetro θ é um número θ, obtido dos dados amostrais, que é com freqüência usado como o valor de θ para tomar várias decisões.

2.5.4 Funções

Sendo *D* um conjunto de números. Uma função $f \cdot$ no conjunto *D* é uma *regra* que descreve como os números no conjunto *D* são *mudados*, *transformados*, ou *mapeados* para produzir outros números. Se *x* representa um número em *D*, então o resultado da aplicação da regra, p.e., a função $f \cdot$, para *x* resultar em um número, dito *z*, e este é simbolicamente denotado por escrever $f\ x =$ ou $z =$. O conjunto *D* é chamado *domínio* da função $f \cdot$.

Quando uma função $f \cdot$ é especificada, seu domínio *D* deve ser especificado também. Qualquer letra pode ser usada para representar uma função; alguns exemplos são $f \cdot$ $\cdot$ e $\mu \cdot$. Algumas vezes uma letra com um subscrito é usada para representar funções; alguns exemplos são $f_1 \cdot$ $\cdot$ $\cdot$ e $g_t \cdot$

2.5.4.1 Função de Muitas Variáveis

As funções já discutidas são funções de uma (independente) variável, dita *x*. Uma função de três variáveis independentes, ditas x_1, x_2, x_3, poderiam ser denotadas por $Y =$ $,x_3$, $z =$ x_3 , ou $z =$ x_3 , etc.

2.5.4.2 Funções Lineares

Uma função $f\ x$ de uma única variável x é dita ser **linear em** x se ela pode ser escrita como

$f\ x = \quad +$ eq. 1.2.8

onde o valor de a e b não dependem do valor de x.

Uma função $f\ x_1,...,x_k$ ou k variáveis $x_1,...,x_k$ é dita ser **simultaneamente linear em** $x_1,...,x_k$ se ela pode ser escrita como

$f\ x_1,...x_k = \quad + \quad +\cdots$ eq. 1.2.9

onde os valores $a_0,a_1,...,a_k$ não dependem dos valores de $x_1,...,x_k$.

2.5.5 Populações Gaussianas Multivariadas

Considerando uma população k-variada

$X_{11},X_{12},...,X_{1k}\ ,...,\ X_{N1},X_{N2},...,X_{Nk}$ escrita $X_1,X_2,...X_k$ para abreviar.

É dito que esta k-variada população é uma ***k-variada população Gaussiana*** *(também chamada um k-variada população normal) se a coleção de números*

$Z_1 = \quad + \quad +\cdots \quad ,$

para $I = \quad$ *é uma uni-variada população Gaussiana para cada escolha possível dos valores das constantes* $a_1,a_2,...,a_k$.

Na realidade uma população pode ser Gaussiana somente se N é infinito. Então usam-se populações Gaussianas apenas como aproximações. É mais fácil determinar se uma k-variada população é aproximadamente Gaussiana quando $k =$ do que quando $k >$. Assim para determinar se uma k-variada $k >$ população é aproximadamente Gaussiana, examina-se Z para ver se ela é aproximadamente uma população Gaussiana uni-variada para cada escolha das constantes $a_1,a_2,a_3,...,a_k$. Se for encontrado que Z é aproximadamente Gaussiana para cada escolha das constantes a_i, então pode se concluir que a k-variada população $X_1,X_2,...,X_k$ é aproximadamente Gaussiana.

Em particular, isto implica que se $X_1,X_2,...,X_k$ é uma população Gaussiana k-variada, então cada uma das k populações uni-variadas X_1 , X_2 ,..., X_k devem ser uma população Gaussiana. Para o caso, por escolher $a_j =$ para todo $j \neq$ e tomando $a_i =$, pode

se concluir que a população X_i é uma população Gaussiana. Contudo, ainda se cada uma das *k* populações uni-variadas X_1 , X_2 ,..., X_k é uma população Gaussiana, não pode se concluir que $X_1, X_2, ..., X_k$ é uma *k*-variada população Gaussiana a menos que também se verifique que a coleção de números $a_1X_{I1} +$ [illegible] $+\cdots$ [illegible] para $I =$ [illegible] é uma população Gaussiana uni-variada para cada conjunto possível de valores das constantes $a_1, a_2, ..., a_k$.

2.6 Regressão e Previsão

2.6.1 Previsão

Razões para previsão – um resumo

Existem pelo menos três razões porque previsão é útil.

1. *Os valores verdadeiros de Y são muito caros para se obter*, mas os valores da variável preditora *X* (ou X_1,..., X_k no caso de múltiplas variáveis preditoras) são relativamente econômicos de se obter, então nós podemos usar os valores econômicos para prever os valores caros de *Y*. Isto é especialmente útil em casos quando um item tem que ser destruído para medir o valor da variável resposta *Y* (...).
2. *Os valores resposta são impossíveis de medir já que eles são usualmente valores futuros e assim não estão disponíveis agora.* Contudo, para propósitos de tomada de decisão investigadores querem saber os valores *antes* de eles ficarem disponíveis. Naturalmente se *Y* for conhecido, não estaremos interessados em predizer-lhe. Mas, em muitos exemplos onde previsão é necessária e usada, o valor verdadeiro da variável resposta não é conhecido porque é um valor *futuro* que queremos saber *agora*. Se *X* está disponível agora e *Y* não está, então podemos usar o valor da função de previsão, $Py\ x$, para prever o valor de *Y agora*.
3. *Previsão não é de interesse imediato, mas a função de previsão é a importante quantidade*.

2.6.1.1 O que é necessário para previsão?

1 As variáveis preditoras, ditas $X_1, ..., X_k$, e os valores observados dessas variáveis.

2 Uma função de previsão ou fórmula, denotada por $P_Y\ x_1,...,x_k$, para prever a variável resposta Y usando as variáveis preditoras $X_1,...,X_k$.

Função de Previsão

A *melhor* (previsão) função para prever Y usando $X_1,...,X_k$ pode ser obtida usando análise de regressão.

2.6.2 Análise de Regressão

Análise de regressão é um método comumente usado para obter uma função de previsão para prever os valores de uma variável resposta Y usando variáveis preditoras $X_1,...,X_k$.

2.6.2.1 Subpopulações

Para cada valor distinto de X na população existe uma *sub-população* de valores Y. *A sub-população correspondendo a* $X =$ *é o conjunto de todos os valore Y daqueles itens na população com* $X =$.

2.6.2.2 Função de Regressão

Para qualquer valor X dado, dito $X =$, o *significado* (p.e., média) dos valores Y nesta sub-população é denotada por μ , e o desvio padrão destes valores Y é denotado por σ .

DEFINIÇÃO

A função μ é chamada **função de regressão** de Y em X e é o *significado* da sub-população de valores Y para cada valor distinto de X.

Em particular, a média de uma sub-população de valores Y, todos dos quais tem $X =$, é igual a μ .

Note que embora os valores atuais Y dos itens na sub-população com $X =$ são em geral, nem todos o mesmo, o valor predito para qualquer desses itens será o mesmo e igual a μ porque eles todos tem o mesmo valor X, nomeado x.

Contudo, se σ é pequeno, a maioria dos valores Y nesta sub-população estarão perto de μ , e a probabilidade é alta de que o valor Y a ser previsto estará próximo ao valor previsto μ .

As médias e desvios padrão de sub-populações são de interesse em uma variedade de situações.

2.6.2.3 Sub-populações, Previsão, e Regressão – Um Resumo de Conceitos

São resumidos os conceitos e idéias já discutidos para o caso de uma única variável preditora (explicativa).

1 A população bi-variável Y, X é particionada em sub-populações – uma sub-população de valores Y para cada valor distinto de X.

2 A sub-população de valores Y correspondendo a qualquer valor dado da variável preditora X, dito $X =$, tem média μ e desvio padrão σ .

3 μ *é chamada função de regressão de* Y *em* X, e é o melhor valor singular a representar (prever) qualquer valor Y na sub-população cujo valor X é x.

4 Se $Y\ x$ denota o valor Y de um item que seja aleatoriamente escolhido da sub-população com $X =$, então o melhor valor previsto de $Y\ x$ é a média, μ , da sub-população de todos os itens cujo valor X se iguala a x.

5 σ é o desvio padrão da sub-população de valores Y cujo valor X é x, e é usado para determinar quão bem μ representa toda a coleção de valores Y na sub-população cujos valores X se igualam a x.

6 Na maioria, se não em todas as aplicações, μ e σ são desconhecidos e devem ser estimados a partir de amostra de dados.

7 Nos livros teóricos em estatística, a distribuição da sub-população $Y\ x$ é chamada *distribuição condicional* de Y dado $X =$.

2.6.2.4 Subpopulações e Regressão no Caso de Várias Variáveis Preditoras

Estendem-se os conceitos de sub-populações e funções de regressão no caso onde o número de variáveis preditoras é maior que uma. Quando existem *k* variáveis preditoras (explicativas), dita $X_1,...,X_k$, *cada combinação distinta de valores de* $X_1,...,X_k$ *na população determina uma sub-população de valores Y*. A sub-população de valores Y determinado por $x_1,...,x_k$ é a coleção de valores Y na população para a qual $X_1 =$ $=$. A média dos valores Y pertencentes a esta população é denotada por μ $_k$, e o desvio padrão é denotado por σ x_k . O melhor valor a usar para prever $Y\ x_1,...,x_k$ é μ $_k$, a

média da sub-população de valores Y com $X_1 =$ $=$. O desvio padrão σ x_k desta população é uma medida de quão bem μ c_k representa cada valor Y nesta sub-população (p.e., quão boa a função de previsão é).

Quando o número de variáveis preditoras é maior que 1 (um) , podemos considerar sub-populações determinadas por qualquer subconjunto dessas variáveis.

2.6.2.5 Métodos de Amostragem em Regressão

Na prática, inferências sobre parâmetros populacionais são baseados na informação provida por amostras. É conseqüentemente muito importante assegurar que recursos disponíveis são usados eficientemente e toda aquela informação relevante é reunida. Idealmente a coleção de dados envolve amostragem aleatória de muito bem definidas populações.

2.6.2.5.1 Amostragem Aleatória Simples

Amostras de dados são obtidas pela seleção de uma simples amostra aleatória de n itens da população inteira de N itens e registrando os valores para a variável resposta Y e as variáveis preditoras $X_1,...,X_k$, para cada item na amostra.

2.6.2.5.2 Amostragem Aleatória com Valores *X* Pré-selecionados

Valores específicos das variáveis preditoras $X_1,...,X_k$ são pré-selecionados pelo investigador, e cada conjunto de valores destes pré-selecionados determina uma sub-população de valores Y. Uma amostra aleatória simples de um ou mais valores Y é selecionado de cada uma dessas sub-populações. O número de observações a ser amostrada de cada sub-população é também pré-determinada pelo investigador.

2.6.3 Regressão Linear e Não-Linear

Raramente sabemos a verdadeira função de regressão em um problema aplicado, mas, podemos freqüentemente pressupor uma classe de funções tal que uma das funções nesta classe servirá como uma aproximação para a verdadeira função de regressão e é precisa bastante para o problema em mãos. As mais simples classes de funções que são úteis em muitos problemas são funções de linha reta, funções quadráticas, etc. Isto significa que podemos escrever uma equação para a função de regressão sob estudo, mas envolverá algumas constantes desconhecidas (chamadas parâmetros). Por exemplo, se um investigador sabe que a função de regressão sob estudo é uma linha reta (para todo propósito prático), mas não sabe a inclinação ou o intercepto desta linha reta, então poderia anotar a função de

regressão como $\mu_{\quad} = {}_{\cdot} + {}_{\cdot}$ onde β e β são parâmetros desconhecidos para serem determinados ou estimados. Neste caso a função de regressão é uma *função linear de parâmetros desconhecidos*. Em geral, regressão linear significa a função de regressão ser simultaneamente linear nos parâmetros desconhecidos $\beta_{\cdot}$, e regressão não-linear significa a função de regressão não ser simultaneamente linear nos parâmetros desconhecidos $\beta_{\cdot}$.

2.7 Métodos Sistemáticos de Estimativa

Embora se encontre um estimador da linha de regressão populacional usando uma linha reta que era *visualmente* julgada prover um bom ajuste aos dados amostrais, isto foi feito para ilustração somente. É desejável obter um estimador da linha de regressão populacional em um mais objetivo e científico semelhante procedimento. Vários semelhantes métodos estão disponíveis na literatura, e um método que é muito usado e tem uma longa história é baseado no tão falado **método dos quadrados mínimos**. Outro método que está se tornando crescentemente popular é o **método dos desvios mínimos absolutos**. Usa-se primeiramente o método dos quadrados mínimos para estimar parâmetros desconhecidos porque é matematicamente mais simples do que a maioria das aproximações alternativas e porque estimadores obtidos usando o método dos mínimos quadrados são melhores estimadores quando certas suposições sobre a população e a amostra são satisfeitas.

Na regressão de linha reta, o interesse está usualmente na estimativa de $\mu_{\quad} \beta_{\cdot} \beta_{\cdot} \sigma_{\cdot}$, e $Y\ x$, onde $Y\ x$ é o valor Y de um item escolhido de modo aleatório de uma sub-população na qual o valor X é x. O interesse pode também estar em estimar $\rho_{\cdot,\cdot} \mu_{\cdot} \mu_{\cdot,\cdot} \sigma_{\cdot}$, e σ. A estimativa de σ para todo x é, para todos os propósitos práticos, impossível a menos que feitas algumas suposições simplificadoras considerando a população Y, X.

2.7.1 Ponto Estimador

O objetivo primário em um estudo de regressão é usar dados amostrais para obter ponto e intervalo de confiança estimados para as quantidades desconhecidas $\beta_{\cdot} \beta \sigma \mu$, $Y\ x$ e também para as várias funções significativas destas quantidades. Estes estimadores se tornam socorro aos investigadores em ganhar discernimento em questões muito complicadas acerca da população sob estudo.

Recordar que um ponto estimado de um parâmetro desconhecido é um número, calculado de um dado amostral observado, que pode ser usado em lugar de um valor desconhecido do parâmetro de interesse para tomar decisões práticas. Usando estes estimadores pode-se obter o melhor estimador de outras quantidades de interesse.

2.7.2 Método dos Quadrados Mínimos

Supondo $y_1, x_1, ..., y_n, x_n$ ser uma amostra de tamanho n de uma população bivariada Y, X, selecionada usando ou amostragem simples aleatória ou amostragem com valores X pré-selecionados, e a função de regressão da população é

μ = + eq. 1.2.10

A quantidade e_i definida por

e_i = − = − + eq. 1.2.11

é o erro de predição quando se usa μ = + para predizer y_i para i = . A relação entre o valor observado y_i, o valor da função de regressão em x_i, β + e o erro de predição e_i, é dado por

y_i = + + eq. 1.2.12

Isto é aludido como **modelo de regressão amostral**.

Visto que β e β são parâmetros desconhecidos, deseja-se usar dados amostrais para obter estimadores deles. Os estimadores resultantes são denotados por $\hat{\beta}$ e $\hat{\beta}$, respectivamente. Estimador correspondente da função de regressão é denotado por

$\hat{\mu}$ = + eq. 1.2.13

O erro de predição quando usando $\hat{\mu}$ = + para predizer y_i é denotado por $\hat{e}_i$ e é dado por

$\hat{e}_i$ = − − eq. 1.2.14

As quantidades $\hat{e}_i$ para i = são chamados resíduos.

O *estimador quadrado mínimo* $\hat{\beta}$ e $\hat{\beta}$ são escolhidos de tal modo que a quantidade SSE X, chamada de soma dos quadrados dos erros de predição quando X é usado para predizer Y, são definidos por

$$SSE\ X = \sum_{i=} \qquad \sum \qquad \text{eq. 1.2.15}$$

Tem o menor valor possível entre todas as escolhas possíveis que poderiam ser feitas para $\hat{\beta}$ e $\hat{\beta}$. Quando não há possibilidade de confusão, simplesmente se escreve *SSE* ao invés de *SSE X* e se referir a ele como a **soma dos erros quadrados** ou **soma dos quadrados dos erros.**

O princípio dos quadrados mínimos declara que o melhor estimador da função de regressão populacional $\mu \quad = \quad + \quad$ é obtida pela escolha de $\hat{\beta}$ e $\hat{\beta}$ de tal modo que a soma dos quadrados dos erros de predição,

$$SSE = \sum_{i=} \hat{} \qquad \qquad \text{eq. 1.2.16}$$

alcançam os *mínimos* valores possíveis.

2.7.3 Pontos Estimados para Funções Lineares de β e β

Enquanto β e β são parâmetros importantes na reta de regressão, investigadores estão muito freqüentemente interessados em fazer inferências sobre certas combinações lineares de β e β . Supondo que θ denota a combinação linear $a_0\beta +$ de β e β , onde a_0 e a_1 são números conhecidos. O melhor ponto estimador de θ é igual a θ onde

$$\theta = \quad \hat{} \quad + \quad \hat{} \qquad \text{eq. 1.2.17}$$

Observe que μ é uma quantidade da forma $a_0\beta +$ com $a_0 =$ e $a_1 =$; β é também um caso especial com $a_0 =$ e $a_1 =$; β é um caso especial com $a_0 =$ e $a_0 =$; $\mu \quad - \quad = \quad + \quad - \quad + \quad = \quad - \quad$ é um caso especial com $a_0 =$ e $a_1 = \quad -$.

2.7.3.1 Notação

É costumeiro usar a notação

$$SXY = \sum_{i=} \qquad \text{eq. 1.2.18}$$

$$SSX = \sum_{i=} \qquad \text{eq. 1.2.19}$$

e

$$SSY = \sum_{i=} \qquad \text{eq. 1.2.20}$$

Então que a fórmula para $\hat{\beta}$ pode ser convenientemente escrita como

$$\hat{\beta}_{-} = \frac{SXY}{SSX} \qquad \text{eq. 1.2.21}$$

Observação

As seguintes alternativas (mas equivalente) expressões para *SXY*, *SSX* e *SSY* são algumas vezes úteis.

$$SXY = \sum_{i=} \qquad \sum_{=} \qquad \text{eq. 1.2.22}$$

$$SSX = \sum_{i=} \qquad \sum_{=} \qquad \text{eq. 1.2.23}$$

e

$$SSY = \sum_{i=} \qquad \sum_{=} \qquad \text{eq. 1.2.24}$$

2.7.4 Ponto Estimador de σ

Recorde que σ é o desvio padrão comum das populações de valores de *Y* determinados pelos valores distintos de *X*. O estimador $\hat{\sigma}$ de σ pode ser calculado usando a fórmula

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-}} \qquad \text{eq. 1.2.25}$$

onde *SSE* é dado por qualquer uma das seguintes expressões equivalentes:

$$SSE = \sum_{i=} \qquad \sum_{=} \qquad \sum \qquad \text{eq. 1.2.26}$$

A quantidade $\frac{SSE}{n-}$, que está sob o símbolo de raiz quadrada na eq. 1.2.26, é chamada **erro médio quadrado** para predizer *Y* usando *X* e é denotado por $MSE\ X$, ou *MSE* para simplificar.

Assim

$$MSE = \frac{SSE}{n-} \qquad \text{eq. 1.2.27}$$

Com esta notação o estimador $\hat{\sigma}$ de σ é dado por

$\hat{\sigma} =$ eq. 1.2.28

Uma fórmula conveniente para calcular *SSE* usando uma calculadora de mão é

$SSE =$ $-$ eq. 1.2.29

Comentário

Na função de regressão amostral na eq. 1.2.10, pode-se substituir $\bar{y} -$ por $\hat{\beta}$ e escrever

μ $=$ $+$ $-$ eq. 1.2.30

Assim μ $=$, que demonstra que o gráfico de função de regressão amostral passa através do ponto $\bar{x}, \bar{y}$, o qual é o "centro" dos dados.

2.8 Coeficiente de Determinação e Coeficiente de Correlação

Para uma população bi-variada Y, X , a melhor predição do valor *Y* de um item escolhido aleatoriamente, dado que seu valor *X* é *x* é μ , o valor da função de regressão de *Y* em *X* avaliado a $X =$. Se o valor *X* do item em questão não é usado, então a melhor predição do valor *Y* do item é μ . Claramente, investigadores têm uma escolha. Eles podem usar μ para prever o valor *Y* do item selecionado, ou eles podem usar μ para prever seu valor *Y*. Naturalmente, para usar μ para prever o valor *Y* do item, se deve saber seu valor *X*, e pode haver alguns custos envolvidos na determinação do valor *X*. Também, não há garantia que usando μ em vez de μ para prever *Y* irá melhorar a previsão suficientemente para justificar o custo associado com a medida ou observação *X*.

Assim a decisão de escolher entre μ e μ para previsão de *Y* usualmente depende, pelo menos em parte, do (a) custo da observação do valor *X*, e (b) o melhoramento na previsão que é feita possível pelo uso do valor *X*. Neste contexto o investigador pode estar interessado em saber as respostas para as seguintes perguntas:

1. Quão bom é μ como um preditor do valor *Y* de um item que é para ser escolhido aleatoriamente da população?

2. Quão bom é μ como um preditor do valor *Y* de um item que é para ser escolhido aleatoriamente de uma população (note que para usar μ para prever o valor de *Y* de um item, se deve saber seu valor *X*)?
3. Quanto melhor é μ do que μ para previsão do valor *Y* de um item aleatoriamente escolhido?
4. É μ um preditor adequado de *Y*?
5. É μ um preditor adequado de *Y*?

Estas questões são respondidas usando desvios padrão populacionais dos erros de previsão como *resumo de medidas* do quão bom preditores são.

1. A quantidade σ é a medida de quão bom μ é como um preditor do valor de *Y*.
2. A quantidade $\sigma =$ é a medida do quão bom μ é como um preditor do valor *Y* porque, se for sabido que o valor *X* daquele item escolhidor for *x*, então a atenção é restrita a sub-população de todos os itens com $X =$; μ é a média e σ é o desvio padrão para estas sub-populações. Recordar que para usar μ , se deve conhecer o valor de *X* para o item que o valor *Y* está sendo predito.
3. σ σ, ou σ σ, ou σ $-$, ou σ $-$ (ou alguma outra função significativa de σ e σ) é uma medida do quanto melhor μ é do que μ para prever o valor de *Y*.
4. μ adequado ou não para previsão do valor de *Y* depende do problema em particular. Um investigador pode considerar μ ser um adequado preditor do valor de *Y* se a maioria dos valores de *Y*, dito pelo menos uma proporção *p* da população, encontra-se próximo a μ , dita dentro de uma distância de *d* unidades de μ (*p* e *d* são especificados pelo investigador).

Nota Tenha em mente que μ e μ podem *ambos* ser adequados na previsão de *Y* ou *nenhum* pode ser adequado. Também note que estão sendo discutidas funções de predição populacional. Na prática, estimadores são usados destas funções de predição baseados em dados amostrais. As funções de predição amostral não desempenham tão bem como funções de predição da população, mas se os estimadores são baseados em amostras suficientemente grandes, então pode-se esperar que as funções de predição amostral se desempenhem quase tão bem como as funções de predição populacionais.

2.8.1 Coeficiente de Determinação

Uma medida comumente usada que resume o desempenho de $\mu_{\cdot}$ como um preditor de *Y*, *relativo* a $\mu_{\cdot}$, é o *coeficiente de determinação* de *Y* com *X*, denotado por $\eta_{\cdot\cdot}$.

Este é definido por (η é a letra grega eta)

$$\eta_{\cdot\cdot} = \frac{\bar{}}{\sigma_{\cdot}} \qquad \text{eq. 1.2.31}$$

Assim $\eta_{\cdot}$ é a *redução proporcional na variância dos erros de previsão quando usando* $\mu_{\cdot}$ *em vez de* μ *para prever Y.*

Uma medida alternativa da performance relativa de $\mu_{\cdot}$ relativa a $\mu_{\cdot}$ é $\delta_{\cdot}$, a *redução proporcional no desvio padrão dos erros de previsão*, e é definido por

$$\delta_{\cdot} = \frac{\bar{}}{\sigma_{\cdot}}$$

Estas duas medidas são relacionadas pela equação

$$\delta_{\cdot} = -\sqrt{} \bar{} \cdot$$

Assim que qualquer quantidade pode ser obtida a partir do conhecimento da outra. Pela razão de $\eta_{\cdot\cdot}$ ser a medida que é tradicionalmente usada por estatísticos e praticantes, apenas esta medida será considerada embora se acredite que $\delta_{\cdot\cdot}$ é também uma medida significativa.

2.8.2 Relação de $\eta_{\cdot\cdot}$ ao Coeficiente de Correlação $\rho_{\cdot\cdot}$ (Pearson)

Recorde que o coeficiente de Pearson de correlação (também chamado *coeficiente de simples correlação* ou *coeficiente de correlação de produto dos momentos*), definido e denotado por $\rho_{\cdot}$, é uma medida da *associação linear* entre *Y* e *X*. Pode ser mostrado que *quando a função de regressão de* ***Y*** *em* ***X*** *é da forma*

$$\mu_{\cdot} = \cdot + \cdot$$

Esta é a razão para que o símbolo $\rho_{\cdot\cdot}$ seja frequentemente usado para denotar o coeficiente de determinação de *Y* com *X*, mas deve-se estar ciente que se a função de regressão de *Y* em *X* não é da forma $\mu_{\cdot} = \cdot + \cdot$ (p.e., não é linear em *x*), então $\eta_{\cdot\cdot}$ não

é igual a $\rho_{\cdot,\ldots}$. Para evitar qualquer possibilidade de confusão, deve-se usar o símbolo $\eta_{\cdot,\ldots}$ para denotar o coeficiente de determinação de Y com X.

2.8.3 Ponto de Estimativa para $\rho_{\cdot,\ldots}$

As quantidades $\rho_{\cdot}$ e $\rho_{\cdot,\ldots}$ são parâmetros populacionais, e a discussão sobre eles tem centrado em torno de seus usos e significado na população. Pontos estimados válidos de $\rho_{\cdot,\ldots}$ e $\rho_{\cdot,\ldots}$ podem ser calculados de dados amostrais de acordo com as fórmulas na eq. 1.2.32 e eq. 1.2.33, respectivamente, (...). Em particular, os dados devem ser obtidos por amostragem aleatória simples.

$$\rho_{\cdot,\ldots} = \frac{SSY}{SSY} = \frac{SSR_{\cdot,\ldots}}{SSY} = \frac{SYY^{2}}{SSX \quad SSY} \qquad \text{eq. 1.2.32}$$

e

$$\hat{\rho}_{\cdot} = \frac{SYY}{\sqrt{SSX \quad SSY}} \qquad \text{eq. 1.2.33}$$

Se dados são obtidos por amostragem com valores pré-selecionados de X, então nenhum estimador válido de $\rho_{\cdot,\ldots}$ ou $\rho_{\cdot}$ está disponível a partir de dados amostrais.

2.9 O Propósito da Regressão Múltipla

O propósito geral da **regressão múltipla** (o termo foi primeiro usado por Pearson, 1908) é quantificar a relação entre várias variáveis independentes ou preditoras e uma variável dependente ou critério.

2.9.1 Computação para Resolver a Equação de Regressão Múltipla

Uma superfície unidimensional em uma bi-dimensional ou espaço de duas variáveis é uma linha definida pela equação $Y = \quad + \quad$. De acordo com esta equação, a variável Y pode ser expressa em termos de ou como função de uma constante (b_0) e uma inclinação (b_1) multiplica a variável X. A constante é também referida como o intercepto, e a inclinação como o coeficiente de regressão. No da **regressão múltipla**, quando existem múltiplas variáveis preditoras, a superfície de regressão usualmente não pode ser visualizada em um espaço bi-dimensional, mas os cálculos são uma simples extensão dos cálculos no caso de um único preditor. Em geral então, procedimentos de **regressão múltipla** irão estimar uma equação linear da forma:

$$Y = \quad + \quad + \quad + \quad +$$

onde k é o número de preditores. Note que nesta equação, os coeficientes de regressão (ou coeficientes $b_1 ... b_k$) representam as contribuições *independentes* de cada **variável dependente** para a previsão da **variável dependente**. Outro modo de expressar este fato é dizer que, por exemplo, variável X_1 está correlacionada com variável Y, depois do controle para todas as outras **variáveis independentes**. Este tipo de correlação é também referida como uma *correlação parcial* (este termo foi usado primeiro por Yule, 1907).

A superfície de regressão (uma linha na regressão simples, um plano ou superfície de mais alta dimensão na **regressão múltipla**) expressa a melhor predição da **variável dependente** (Y), dadas as **variáveis independentes** (X,s). Contudo, natural é raramente (se nunca) perfeitamente previsível, e usualmente existe substancial variação dos pontos observados dos próximos pontos correspondentes na superfície de regressão predita (seus valores preditos) é chamado de valor *residual*. Visto que o objetivo do procedimento de regressão linear é ajustar uma superfície, a qual é uma função linear das variáveis X, tão próximas quanto possível à variável observada Y, os valores residuais para os pontos observados podem ser usados para elaborar um critério para o "melhor ajuste". Especificamente, em problemas de regressão a superfície é calculada para a qual a soma dos desvios quadrados dos pontos observados de uma superfície são minimizados. Assim, o procedimento geral é algumas vezes também referido com *estimativa dos mínimos quadrados*.

Os cálculos atuais envolvidos na resolução de problemas de regressão podem ser expressos compactamente e convenientemente usando notação matricial. Suponha que existem n valores de Y observados e n valores observados associados para cada um das k diferentes X variáveis. Então, Y_i, X_{ik} e e_i podem representar a i-ésima observação da variável Y, a i-ésima observação de cada uma das X variáveis, e o i-ésimo valor residual desconhecido, respectivamente. Recolhendo estes termos em matrizes tem-se

$$Y = \begin{bmatrix} \\ \\ \\ \\ \\ \\ \end{bmatrix} \quad \begin{bmatrix} & & \cdots & \cdots & \cdots & & \\ & & & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix} \quad \begin{bmatrix} \\ \\ \\ \\ \\ \\ \end{bmatrix}$$

O modelo de **regressão múltipla** em notação matricial então pode ser expresso como

$Y =$ $\quad +$

onde b é um vetor de 1 coluna (para o intercepto) + k coeficientes de regressão desconhecidos. Recordando que o objetivo da **regressão múltipla** é minimizar a soma dos

quadrados residuais, coeficientes de regressão que satisfazem este critério são encontrados pela resolução do conjunto de equações normais

$X'Xb =$

Quando as variáveis X são linearmente independentes (p.e., elas são não redundante, sustentando uma matriz $X'X$ que é condensada) existe uma única solução para as equações normais. Pré-multiplicando ambos os lados da fórmula matricial para as equações normais pelo inverso de $X'X$ dá

$X'X^{-} X'Xb =$ $X'Y$

ou

$b =$ $X'Y$

Este último resultado é muito satisfatório em vista de sua simplicidade e sua generalidade. Com ressalvas à sua simplicidade, ele expressa a solução para a equação de regressão em termos de exatamente 2 matrizes (X e Y) e 3 operações matriciais básicas, (1) transposição matricial, que envolve intercambiar os elementos nas linhas e colunas de uma matriz, (2) multiplicação matricial, que envolve descoberta da soma dos produtos do elementos para cada combinação de linha e coluna de duas similares (p.e., multiplicáveis) matrizes, e (3) matriz de inversão, a qual envolve descoberta da matriz equivalente de um recíproco numérico, que é, a matriz que satisfaz

$A^{-}AA =$

para uma matriz A

Com considerações à generalidade do modelo de **regressão múltipla**, suas limitações notáveis são que (1) pode ser usada para analisar somente uma única **variável dependente**, (2) não pode prover uma solução para os coeficientes de regressão quando as variáveis X não são linearmente independentes e o inverso $X'X$ consequentemente não existe. Estas restrições, contudo, podem ser superadas, e fazendo então o modelo de **regressão múltipla** é transformado no modelo linear generalizado.

2.10 Regressão Linear Múltipla

2.10.1 Notação e Definições

A palavra *múlptipla* em regressão linear múltipla significa que há mais de uma variável preditora. Recorde que a palavra *linear* significa que a função de regressão, denotada por μ c_k , é *linear nos parâmetros* β β β . A palavra *regressão* significa que o

estudo é tratado com a predição de variáveis resposta Y usando a relação entre Y e k variáveis preditoras $X_1, X_2, ..., X_k$.

μ = + +⋯ eq. 1.2.34

2.10.2 Variáveis Básicas Observáveis e Variáveis Derivadas

É útil fazer distinção entre *variáveis básicas observáveis* e *variáveis derivadas.* Se Z representa a altura de um indivíduo, então Z é uma variável básica observável e $\sqrt{Z}$ ou $\log Z$ são variáveis derivadas. Assim variáveis derivadas são funções conhecidas de variáveis observáveis. As quantidades x_i na função de regressão linear múltipla em (4.1.1) podem ser valores de variáveis básicas observáveis ou variáveis derivadas. Para o caso, se as variáveis Z_1 e Z_2 são variáveis básicas observáveis e a função de regressão é

μ = + + +

Então pode-se definir as variáveis X_1, X_2, X_3 por

X_1 = X_2 = X_3 =

e seus valores x_1, x_2, x_3 por

x_1 = x_2 = x_3 =

então que a função de regressão pode ser escrita como

μ = + + +

Isto demonstra a versatilidade da função de regressão na eq. 1.2.34. Mais exemplos são dados na eq. 1.2.35.

μ = + + + eq. 1.2.35a

μ = + + + + eq. 1.2.35b

μ = + + + + eq. 1.2.35c

Cada uma dessas funções de regressão podem ser escritas na forma dada na eq. 1.2.34 pela definição de variáveis $X_1, X_2, ...,$ etc. devidamente. Para o caso, na eq. 1.2.35a definindo X_1 = = e X_3 = então que a função de regressão pode ser reescrita como

μ = + + +

Aqui cada X_i é uma função *conhecida* de Z_i e não envolve nenhum parâmetro desconhecido. Na eq. 1.2.35b usa-se $X_1 =$ $=$ $+$, e $X_3 =$, que permite reescrever a função de regressão como

μ $=$ $+$ $+$ $+$

Similarmente, na eq. 1.2.35c usa-se $X_1 =$ $=$ $=$ $+$ $+$, e $X_4 =$ e reescrever a função de regressão como

μ $=$ $+$ $+$ $+$ $+$

2.10.3 O Modelo de Regressão Populacional

Para o *I-ésimo* item na população, o verdadeiro valor da variável resposta é Y_I, e o valor predito usando a função de regressão na eq. 1.2.34, a qual é a melhor função de predição, é

μ $=$ $+$ $+$ $+\cdots$

O erro de predição para o *I*-ésimo elemento da população é denotado por E_I; é a diferença entre o valor verdadeiro Y_I e o valor predito μ $,\cdots$ e é dado por

$$E_I = - ,X_{I,k}$$

ou

$$E_I = - + +\cdots$$

Esta equação é geralmente escrita como

$Y_I = + +\cdots +$ eq. 1.2.36

para $I =$ e é chamada **modelo de regressão populacional**. É usado o símbolo $Y\ x_1,...,x_k$ para denotar valor Y de um item escolhido aleatoriamente com $X_1 = = .$

2.10.4 Amostra

Para fazer inferências estatísticas sobre os parâmetros desconhecidos na função de regressão populacional na eq. 1.2.34, deve-se usar dados amostrais para computar pontos estimados e intervalos estimados destes parâmetros desconhecidos, então seleciona-se uma amostra de tamanho *n* a partir da população.

2.10.5 Ponto Estimador

Um dos principais objetivos da análise de regressão múltipla é usar uma amostra de dados para obter ponto e intervalo de confiança estimados para quantidades desconhecidas $\beta_0\ \beta_1\ \ldots\ \beta_k\ \mu_{Y|x_1,\ldots,x_k}$, e σ, e também para funções selecionadas destas quantidades. Consequentemente, a função de regressão populacional é da forma dada na eq. 1.2.34. Uma amostra de tamanho *n* é selecionada usando amostragem aleatória simples ou por amostragem com valores pré-selecionados de $X_1,\ldots,X_k$ a partir da população $Y, X_1,\ldots,X_k$ de *N* itens.

O *método dos mínimos quadrados* é usado para obter os pontos estimados de $\beta_0\ \beta_1\ \ldots\ \beta_k$.

2.10.6 Estimadores Mínimos Quadrados de $\beta_0\ \beta_1\ \ldots\ \beta_k$

Considere a função de regressão populacional

$$\mu_{Y|x_1,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad \text{eq. 1.2.37}$$

a qual se gostaria de usar para prever *Y* com $X_1,\ldots,X_k$ como fatores de previsão. Contudo, visto que β_0 não é conhecido, esta função não pode ser usada. Assim deve-se usar amostra de dados dada na Tabela 2. Para obter estimadores de β_0, usa-se o **princípio dos quadrados mínimos**. Os estimadores resultantes são denotados por $\hat{\beta}_0\ \hat{\beta}_1\ \ldots\ \hat{\beta}_k$, respectivamente. O estimador correspondente da função de regressão é denotado por

$$\hat{\beta}_{Y|x_1,\ldots,x_k} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \qquad \text{eq. 1.2.38}$$

Tabela 2: Representação Esquemática de Valores de Dados Amostrais

Y	X_1	X_2	$\cdots$	X_k
y_1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$
y_2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	$x_{i,1}$	$x_{i,2}$	$\cdots$	$x_{i,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$

Fonte: Adaptado de (GRAYBILL; IYER, 1994)

Considere as *n* observações na Tabela 2. O valor predito *Y* do *i*-ésimo item amostral, com $x_{i,1},\ldots,x_{i,k}$ como os valores das variáveis preditoras, é dado por

$\hat{\mu}$ = $\hat{}$ + $\hat{}$ +⋯ $\hat{}$. O correspondente erro de previsão é denotado por $\hat{e}_i$ e dado por

e_i = − ⌈ ⋯ ⌉ eq. 1.2.39

As quantidades $\hat{e}_i, i =$ são chamadas *resíduos*. Eles são úteis no exame da validade das suposições assumidas para modelos de regressão.

Os estimadores mínimos quadrados $\hat{\beta}$ $\hat{\beta}$ $\hat{\beta}$ de β β β são escolhidos de tal modo que a quantidade $SSE\ X_1,...,X_k$, que é definida por

$SSE\ X_1,...,X_k = \sum_{i=}$ ⋯ $\hat{\sum}$ eq. 1.2.40

alcança seu menor valor possível entre todas as escolhas possíveis que podem ser feitas estimando β β β. O mínimo valor correspondente de $SSE\ X_1,...,X_k$ é chamado **soma dos erros quadrados** para predizer *Y* usando a função de regressão estimada

$\hat{\beta}$ = $\hat{}$ + $\hat{}$ +⋯ $\hat{}$

Note que não se está realmente interessado em predizer os valores *Y* dos itens amostrais porque já são conhecidos seus valores verdadeiros, a saber $y_1,...,y_n$. Mas se a função de regressão estimada

$\hat{\mu}$ = $\hat{}$ + $\hat{}$ +⋯ $\hat{}$

é uma boa preditora dos *valores amostrais conhecidos* y_i correspondendo a *n* itens amostrais, então há razão para esperar que será uma boa função de predição para *todos* os valores de *Y* na população. Assim usam-se os valores y_i e $x_{i,j}, j =$ dos itens na amostra para avaliar a performance da função de regressão estimada.

O princípio dos quadrados mínimos afirma que o melhor estimador da função de regressão populacional μ = + +⋯ , usando dados amostrais, é obtida por escolher $\hat{\beta}$ $\hat{\beta}$ $\hat{\beta}$ na eq. 1.2.38 de tal modo que a soma dos quadrados dos erros de previsão

$$SSE = \sum_{i=} \hat{} \quad \angle_{=} \quad \hat{} \quad \hat{} \quad \text{eq. 1.2.41}$$

é o mínimo possível.

2.10.7 Ponto Estimador para Funções Lineares de β β β

Uma questão de interesse para um investigador pode frequentemente ser formulada em termos da questão envolvendo uma combinação linear destes parâmetros β β β, dado por

$$\theta = \quad + \quad + \cdots$$

onde os componentes a_i no vetor $a^T = \quad ., a_k$ são especificados pelo investigador.

O ponto estimador de $\theta =$ é

$$\theta = \hat{} = \hat{} + \hat{} + \cdots \hat{} \quad \text{eq. 1.2.42}$$

onde os $\hat{\beta}$ são computados pela fórmula eq. 1.2.38.

Observe que μ c_k por si só é uma quantidade da forma $a^T\beta$ com $a_0 = = = .$ Da mesma maneira, cada β é um caso especial de $a^T\beta$ com $a_i =$ e todos os elementos remanescentes de *a* iguais a zero.

2.10.8 Resíduos

Se for usada a função de regressão estimada $\hat{\mu}$ para prever os valores *Y* do item amostral *i*, que tem $x_{i,1},...,x_{i,k}$ como o valor para as variáveis preditoras, o valor previsto *Y* é $\hat{\mu}$ $x_{i,k}$, e o erro de previsão é $\hat{}$ $\hat{}$ $x_{i,k}$. Assim

$$\hat{} \quad \hat{} \quad = \quad - \hat{} - \hat{} \quad - \cdots \hat{} \quad \text{para } i = \quad \text{eq. 1.2.43}$$

As quantidades $\hat{}$ são chamadas **resíduos** e eles podem ser computados a partir de dados amostrais porque $y_i, x_{i,1},...,x_{i,k}$ e $\hat{\beta}$ são todos conhecidos.

2.11 Design da Célula

Optou-se pela utilização de uma célula única pela maior facilidade de operação. Além de se excluir a necessidade de uma ponte salina entre duas semi-células a distância entre os eletrodos é um fator fundamental no ganho de produção.

2.12 Determinação Experimental do Eletrólito

A escolha do eletrólito se baseou em características, sobretudo termodinâmicas mais favoráveis, levando em consideração propriedades físico-químicas, toxicidade, custo, produtos de decomposição e impacto ambiental na escolha final. Cada eletrólito candidato foi submetido ao processo de eletrólise e aquele que apresentou o melhor desempenho foi o escolhido.

2.13 Determinação Experimental do Tempo de Desgaste

O sistema de eletrólise foi montado como mostra a Figura 11. O tempo de desgaste foi determinado pelo uso de cronômetro, iniciando a contagem assim que aplicada a diferença de potencial e finalizada a contagem com o momento de ruptura do eletrodo e registro da menor corrente. Outras variáveis foram mensuradas (Temperatura inicial e final, Corrente máxima e mínima, massa inicial e final) e um banco de dados foi montado.

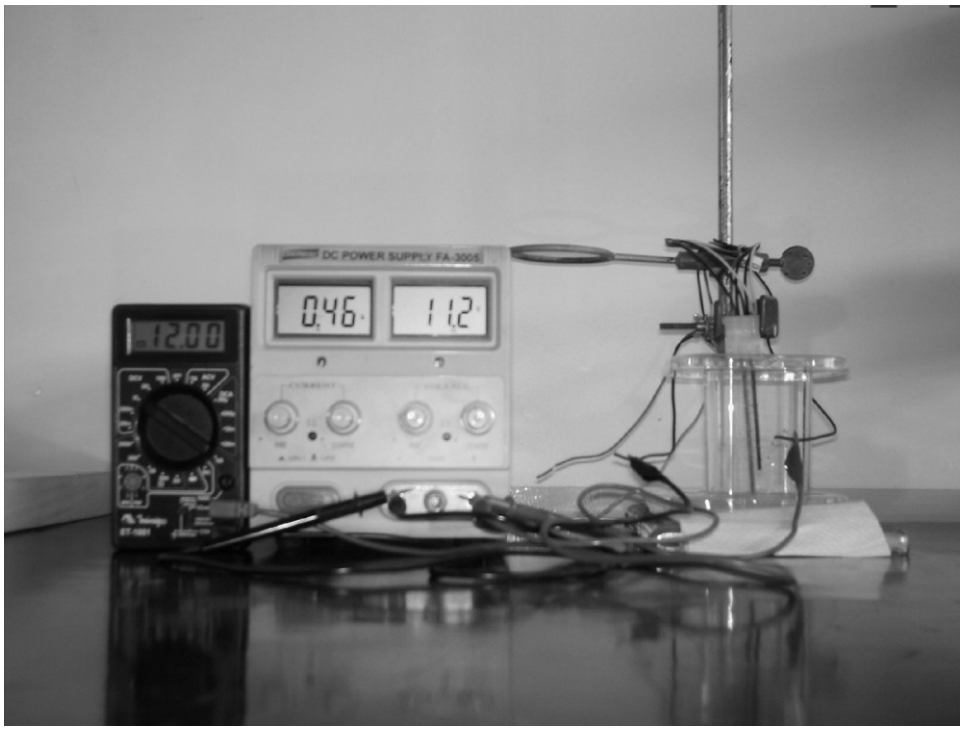


Figura 11: Módulo de Eletrólise

2.13.1 Análise dos Dados Experimentais

Os dado experimentais foram tratados por meio dos softwares Minitab (versão 15) e R (versão 2.13.2) com o emprego de rotinas de análise estatística e algoritmos. Os resultados foram expressos em saídas computacionais e gráficos.

3. Descrição dos Dados

Banco de dados: eletrodos.xls

Período Amostral: 15/02/2011 a 30/06/2011

Variáveis: t: tempo de desgaste (variável resposta); **Ti:** temperatura inicial do experimento;

Tf: temperatura final do experimento; **Cmax:** corrente máxima; **Cmin:** corrente mínima;

mi: massa inicial; **mf:** massa final; **aT:** amplitude temperatura; **aC:** amplitude corrente;

am: amplitude massa

dia	Ti(°C)	Tf(°C)	Tensão(V)	corrente máx (A)	corrente min (A)	tempo (min)	volume (mL)	[ ] mol/L	massa inicial (g)	massa residual (g)
15/fev	23	45	12	0,92	0,44	33,44	100	1,46962	0,552	0,411
16/fev	32	46	12	0,93	0,35	24,42	100	1,46962	0,56	0,438
17/fev	30	44	12	0,91	0,32	26,46	100	1,46962	0,562	0,429
18/fev	26	41	12	0,91	0,38	20,13	100	1,46962	0,567	0,429
21/fev	28	43	12	0,91	0,38	21,16	100	1,46962	0,566	0,429
22/fev	31	43	12	0,99	0,32	17,24	100	1,46962	0,553	0,428
23/fev	31	44	12	0,97	0,41	19,03	100	1,46962	0,563	0,424
24/fev	27	42	12	0,91	0,38	21,53	100	1,46962	0,563	0,42
25/fev	28	40	12	0,9	0,42	16,52	100	1,46962	0,567	0,441
28/fev	25	42	12	1,01	0,38	22,04	100	1,46962	0,568	0,432
01/mar	30	43	12	0,95	0,37	22,01	100	1,46962	0,563	0,431
02/mar	30	44	12	0,91	0,36	22,07	100	1,46962	0,569	0,437
03/mar	22	37	12	0,82	0,42	24,06	100	1,46962	0,558	0,419
04/mar	24	40	12	0,98	0,42	20,44	100	1,46962	0,547	0,418
10/mar	29	39	12	0,91	0,38	19,48	100	1,46962	0,558	0,43

11/mar	24	42	12	1,11	0,42	18,38	100	1,46962	0,563	0,423
14/mar	30	46	12	1,26	0,38	16,08	100	1,46962	0,558	0,416
15/mar	29	45	12	1,03	0,41	21,06	100	1,46962	0,555	0,408
16/mar	29	45	12	1,01	0,4	20,46	100	1,46962	0,558	0,405
17/mar	29	45	12	1,01	0,38	22,24	100	1,46962	0,554	0,394
18/mar	30	46	12	0,99	0,42	22,04	100	1,46962	0,558	0,398
21/mar	29	44	12	0,96	0,38	22,23	100	1,46962	0,552	0,422
22/mar	24	41	12	0,72	0,32	28,05	100	1,46962	0,569	0,437
23/mar	24	44	12	0,93	0,37	23,45	100	1,46962	0,56	0,405
24/mar	29	44	12	0,94	0,41	21,37	100	1,46962	0,561	0,417
25/mar	25	48	12	1,47	0,31	17,26	100	1,46962	0,55	0,402
28/mar	30	51	12	1,56	0,44	15,53	100	1,46962	0,561	0,408
29/mar	29	51	12	1,71	0,49	13,3	100	1,46962	0,554	0,414
30/mar	30	50	12	1,83	0,42	13,54	100	1,46962	0,554	0,4
31/mar	30	49	12	1,81	0,98	12,54	100	1,46962	0,553	0,399
01/abr	21	43	12	1,57	0,4	14,26	100	1,46962	0,542	0,386
04/abr	30	48	12	1,61	0,43	15,19	100	1,46962	0,554	0,4
05/abr	30	49	12	1,63	0,47	15,24	100	1,46962	0,55	0,4
06/abr	28	50	12	1,65	0,38	15,59	100	1,46962	0,532	0,378
07/abr	28	49	12	1,65	0,4	16,45	100	1,46962	0,555	0,393
08/abr	28	44	12	1,24	0,78	17,29	100	1,46962	0,556	0,415
11/abr	30	38	12	0,71	0,56	19,42	100	1,46962	0,558	0,386
12/abr	27	45	12	1,18	0,77	15,53	100	1,46962	0,549	0,416
13/abr	27	44	12	1	0,8	19,05	100	1,46962	0,546	0,456
14/abr	25	37	12	0,88	0,58	13,47	100	1,46962	0,56	0,415
15/abr	30	46	12	1,4	0,67	13,42	100	1,46962	0,55	0,4
18/abr	26	44	12	1,34	0,58	16,39	100	1,46962	0,546	0,396
19/abr	26	41	12	1,15	0,81	13,47	100	1,46962	0,56	0,408
20/abr	31	44	12	1,27	0,78	13,3	100	1,46962	0,561	0,432
25/abr	24	43	12	1,22	0,8	14,17	100	1,46962	0,558	0,429

26/abr	31	43	12	1,19	0,78	12,02	100	1,46962	0,556	0,41
27/abr	27	51	12	1,58	0,78	18,42	100	1,46962	0,555	0,393
28/abr	28	50	12	1,56	0,68	17,07	100	1,46962	0,556	0,407
29/abr	25	49	12	1,38	0,68	19,3	100	1,46962	0,556	0,399
02/mai	31	42	12	1,52	0,38	15,36	100	1,46962	0,548	0,39
03/mai	27	39	12	1,32	0,68	8,35	100	1,46962	0,558	0,406
04/mai	29	51	12	1,67	0,42	15,59	100	1,46962	0,552	0,401
05/mai	30	53	12	1,79	0,2	16,3	100	1,46962	0,555	0,391
06/mai	33	51	12	1,94	0,4	15,31	100	1,46962	0,555	0,384
09/mai	26	45	12	1,39	0,31	14,57	100	1,46962	0,554	0,387
10/mai	31	49	12	2	0,64	15,21	100	1,46962	0,553	0,398
11/mai	25	43	12	1,6	0,8	15,41	100	1,46962	0,555	0,39
12/mai	31	52	12	1,77	0,42	15,34	100	1,46962	0,549	0,401
13/mai	30	51	12	1,7	0,78	12,35	100	1,46962	0,554	0,391
16/mai	25	41	12	1,26	0,14	20,07	100	1,46962	0,552	0,396
17/mai	30	49	12	1,62	0,44	15,44	100	1,46962	0,552	0,395
18/mai	31	51	12	1,63	0,48	15,04	100	1,46962	0,55	0,385
19/mai	22	48	12	1,58	0,49	15,49	100	1,46962	0,55	0,398
20/mai	30	46	12	1,42	0,44	13,12	100	1,46962	0,55	0,404
23/mai	28	50	12	1,7	0,34	15,17	100	1,46962	0,564	0,413
24/mai	32	51	12	1,81	0,48	14,33	100	1,46962	0,55	0,393
25/mai	24	51	12	1,71	0,58	16,02	100	1,46962	0,555	0,399
26/mai	25	40	12	1,27	0,43	13,12	100	1,46962	0,54	0,412
27/mai	28	42	12	1,71	0,4	13,1	100	1,46962	0,55	0,404
30/mai	23	46	12	1,42	0,44	15,44	100	1,46962	0,562	0,421
31/mai	30	45	12	1,27	0,42	16,15	100	1,46962	0,56	0,418
01/jun	24	41	12	1,04	0,38	19,03	100	1,46962	0,564	0,42
02/jun	25	43	12	1,23	0,44	18,29	100	1,46962	0,545	0,391
03/jun	20	41	12	1,16	0,58	20,03	100	1,46962	0,554	0,419
06/jun	23	42	12	1,12	0,38	17,29	100	1,46962	0,563	0,417

07/jun	28	42	12	1,09	0,4	15,01	100	1,46962	0,561	0,427
08/jun	23	39	12	1,03	0,38	16,08	100	1,46962	0,564	0,428
09/jun	29	42	12	1,1	0,4	13,21	100	1,46962	0,568	0,445
10/jun	31	41	12	1,26	0,4	14,23	100	1,46962	0,552	0,418
13/jun	22	46	12	1,36	0,41	21,17	100	1,46962	0,559	0,407
14/jun	30	49	12	1,48	0,5	17,2	100	1,46962	0,552	0,409
15/jun	29	51	12	1,53	0,43	18,31	100	1,46962	0,561	0,399
16/jun	27	50	12	1,64	0,74	16,18	100	1,46962	0,551	0,384
17/jun	28	52	12	1,61	0,8	16,21	100	1,46962	0,552	0,38
20/jun	29	49	12	1,61	0,85	11,53	100	1,46962	0,545	0,411
21/jun	30	53	12	1,62	0,7	16,4	100	1,46962	0,551	0,387
22/jun	30	52	12	1,63	0,38	16,07	100	1,46962	0,551	0,395
27/jun	24	49	12	1,44	0,43	19,17	100	1,46962	0,554	0,401
28/jun	30	51	12	1,63	0,88	14,23	100	1,46962	0,557	0,385
29/jun	30	53	12	1,63	0,38	17,1	100	1,46962	0,553	0,425
30/jun	30	53	12	1,79	0,42	15,17	100	1,46962	0,555	0,374

4. Análise dos Resultados

Análise Exploratória

Estatísticas Descritivas

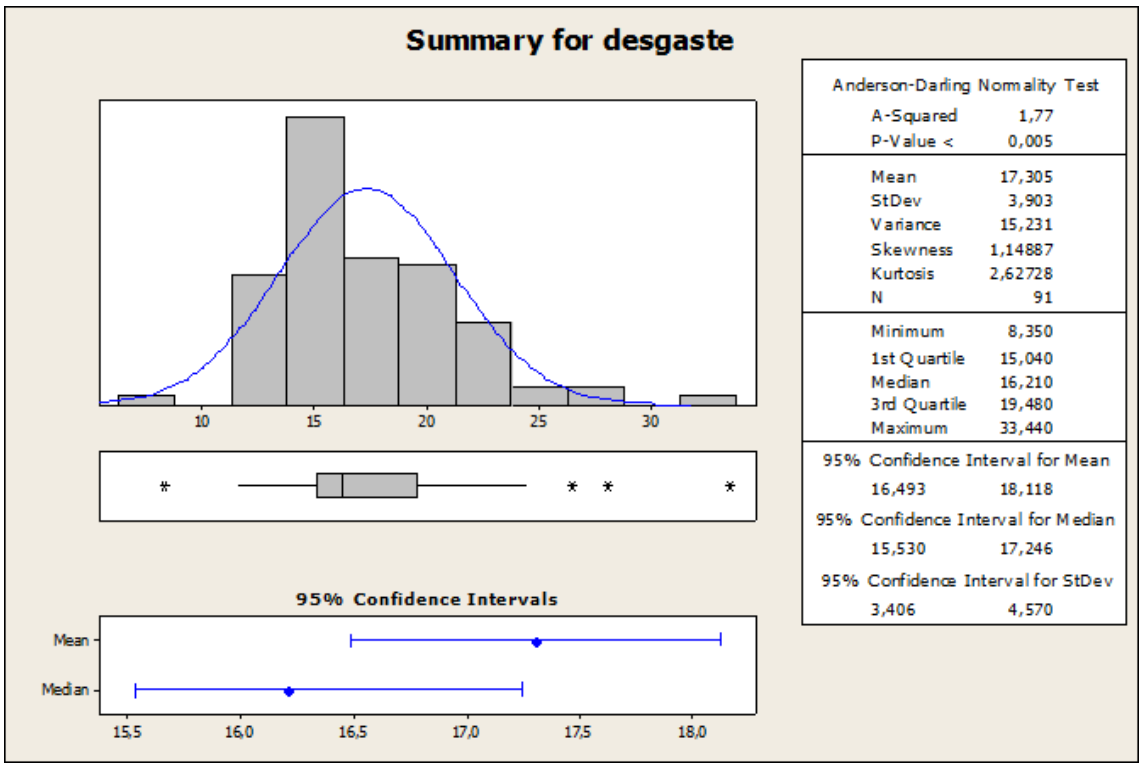


Figura 12: Estatísticas descritivas

Discussão: pela análise das estatísticas descritivas é possível observar que a série tem uma alta variabilidade, evidenciada pelo desvio padrão de cerca de 3,903 unidades de tempo; uma média de 17,305min para o desgaste dos eletrodos com um valor mínimo de 8,350min e máximo de 33,44min. A série não apresenta uma distribuição normal (p-valor < 0,005). Em comparação à distribuição normal apresenta um valor "Skewness" positivo (deslocamento direita/positivo) caracterizando uma concentração de valores maiores do que a média e um valor de curtose positivo (mais "apontada" que a normal) caracterizando uma concentração de valores mais próximos à média. A alta variabilidade da série pode ser identificada pela constatação de que a mediana não faz parte do mesmo intervalo de confiança da média ao nível de significância de 5%.

Comportamento da Série de Dados

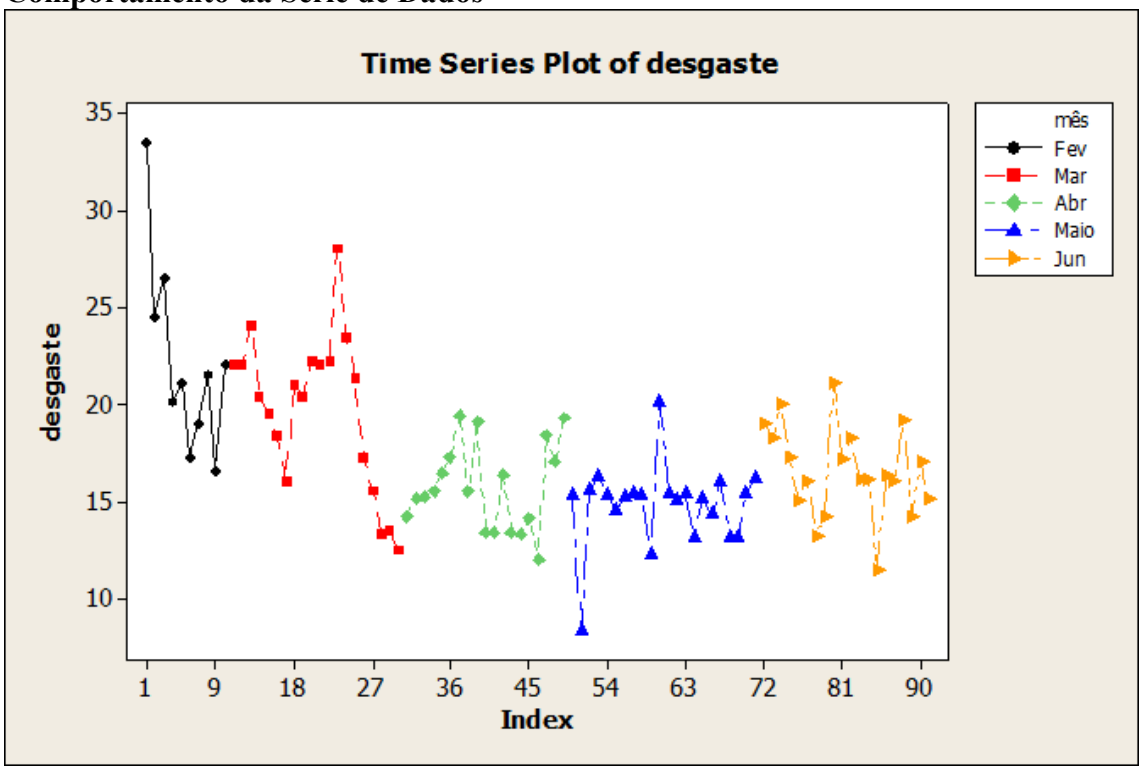


Figura 13: Comportamento da série ao longo do período amostral

Discussão: é possível notar uma variabilidade (descontrole) nos dois meses iniciais (Fevereiro e Março). Com as proposições de melhoramento e otimização do sistema a partir do terceiro mês (abril) se alcançou uma estabilidade até o final do período amostral. É observado maior valor de tempo de desgaste no mês de fevereiro (primeira amostra) e menor valor no mês de maio.

Correlação entre as Variáveis

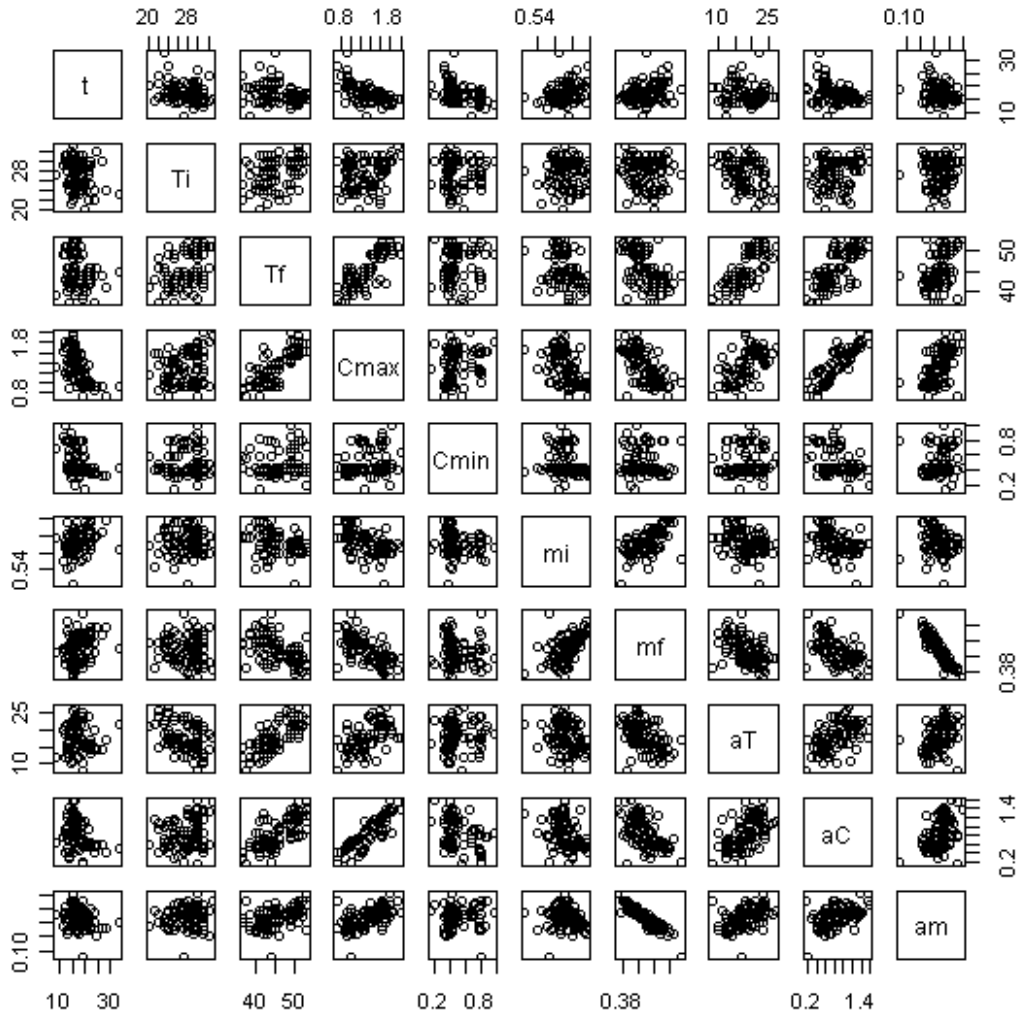


Figura 14: Matriz de Dispersão Gerada no Software R

Discussão: visualmente, parece haver correlação linear entre as variáveis Tf e Cmax (positiva), Tf e aT (positiva), Cmax e aC (positiva), mf e am (negativa) e fraca correlação não-linear entre a variável resposta t e a variável preditora Cmax (negativa).

Análise da Correlação entre as Variáveis Preditoras e Variável Resposta

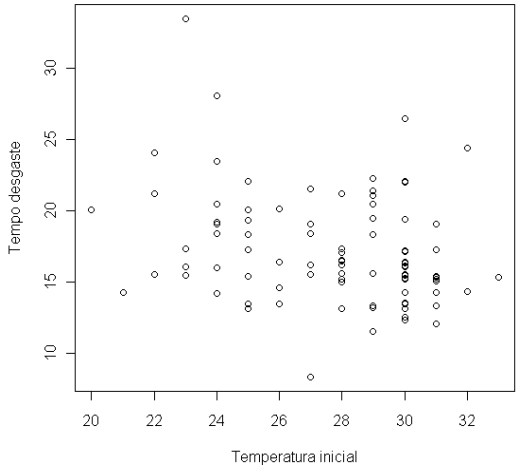


Gráfico 1: Preditora Ti *x* Resposta t

Discussão: visualmente parece não haver correlação entre a variável reposta e a temperatura inicial.

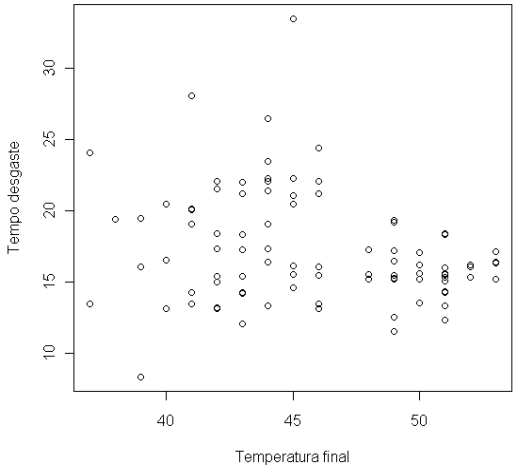


Gráfico 2: Preditora Tf *x* Resposta t

Discussão: visualmente parece não haver correlação entre a variável reposta e a temperatura final.

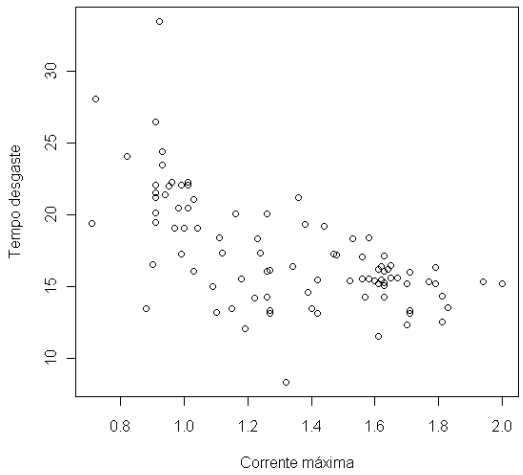


Gráfico 3: Preditora Cmax *x* Resposta t

Discussão: visualmente parece haver uma correlação não linear entre a variável reposta e a corrente máxima.

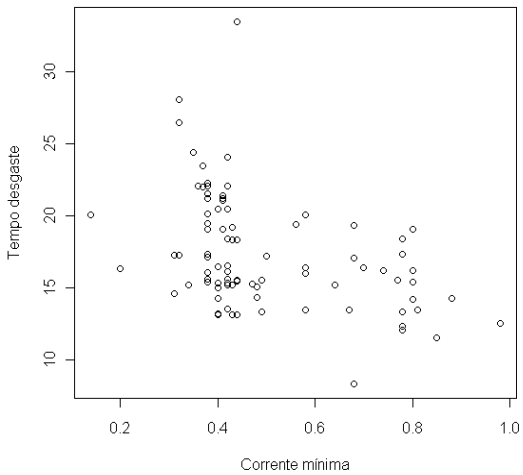


Gráfico 4: Preditora Cmin *x* resposta t

Discussão: visualmente parece haver correlação não linear entre a variável reposta e a corrente mínima.

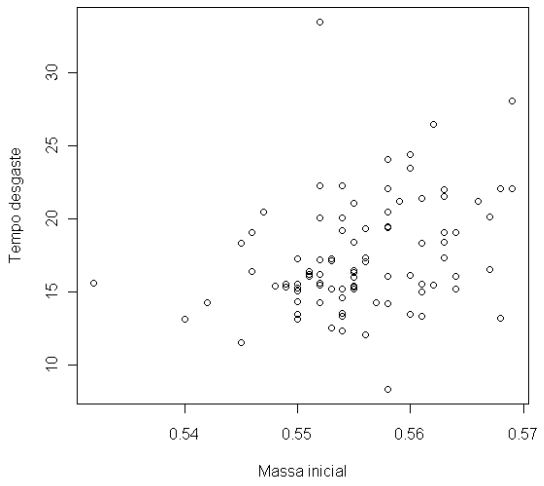


Gráfico 5: Preditora mi *x* Resposta t

Discussão: visualmente parece não haver correlação entre a variável reposta e a massa inicial.

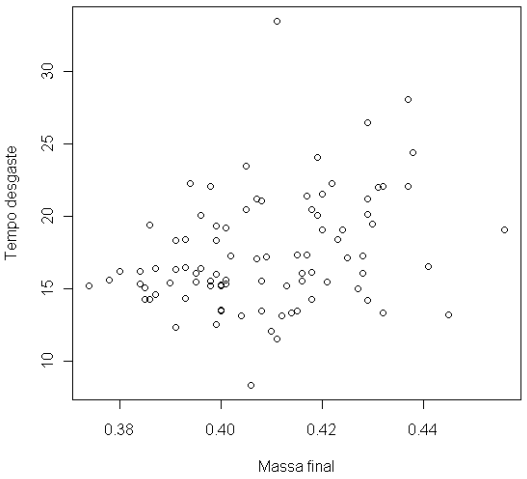


Gráfico 6: Preditora mf *x* Reposta t

Discussão: visualmente parece não haver correlação entre a variável reposta e a massa final.

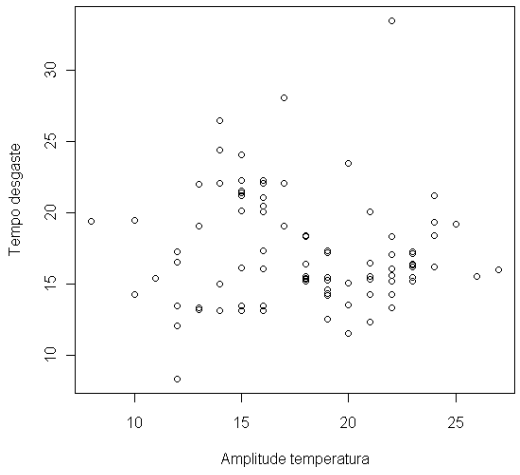


Gráfico 7: **Preditora aT *x* Resposta t**

Discussão: visualmente parece não haver correlação entre a variável reposta e a amplitude de temperatura.

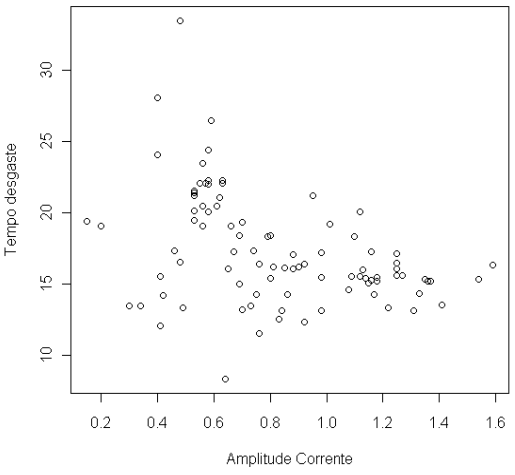
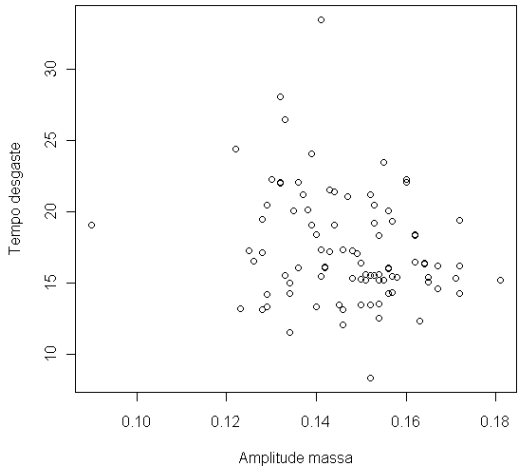


Gráfico 8: Preditora aM *x* Resposta t

Discussão: visualmente parece não haver correlação entre a variável reposta e a amplitude de corrente.



Discussão: visualmente parece não haver uma correlação entre a variável resposta e a variável amplitude massa.

Gráfico 9: Preditora am *x* Resposta t

Matriz de Correlação

Correlations: t; Ti; Tf; Cmax; Cmin; mi; mf; aT; aC; am

	t	Ti	Tf	Cmax	Cmin	mi	mf	aT	aC
Ti	-0,251								
	0,016								
Tf	-0,234	0,408							
	0,026	0,000							
Cmax	-0,617	0,271	0,791						
	0,000	0,009	0,000						
Cmin	-0,401	0,035	0,157	0,242					
	0,000	0,741	0,138	0,021					
mi	0,315	-0,042	-0,311	-0,486	-0,184				
	0,002	0,691	0,003	0,000	0,081				
mf	0,316	-0,114	-0,532	-0,661	-0,157	0,555			
	0,002	0,283	0,000	0,000	0,138	0,000			
aT	-0,064	-0,293	0,754	0,633	0,139	-0,295	-0,476		
	0,544	0,005	0,000	0,000	0,190	0,004	0,000		
aC	-0,403	0,250	0,702	0,865	-0,278	-0,386	-0,573	0,555	
	0,000	0,017	0,000	0,000	0,008	0,000	0,000	0,000	
am	-0,227	0,115	0,484	0,555	0,100	-0,192	-0,923	0,425	0,497
	0,030	0,280	0,000	0,000	0,346	0,068	0,000	0,000	0,000

Cell Contents: Pearson correlation
P-Value

Discussão: A partir dos valores de correlação e p-value demonstrados acima, verifica-se moderada e significativa correlação entre a variável resposta t e as variáveis Cmax, Cmin e aC (negativa); significativa correlação moderada entre as variáveis Ti e Tf (positiva); significativa correlação moderada

entre as variáveis mf e Tf (negativa), am e Tf (positiva) e, significativa correlação forte entre as variáveis Cmax e Tf (positiva), aT e Tf (positiva), aC e Tf (positiva); significativa correlação moderada entre as variáveis mi e Cmax (negativa), mf e Cmax (negativa), aT e Cmax (postitiva), am e Cmax (positiva) e, significativa correlação forte entre as variáveis aC e Cmax (positiva); significativa correlação moderada entre as variáveis mf e mi (positiva); moderada correlação significativa entre as variáveis aT e mf (negativa), aC e mf (negativa) e, significativa correlação forte entre as variáveis am e mf (negativa); ainda moderada correlação significativa entre as variáveis aC e aT (positiva) am e aT (positiva) am e aC (positiva).

Ajuste do modelo

Método stepwise

Stepwise Regression: t versus Ti; Tf; Cmax; Cmin; mi; mf; aT; aC; am

Alpha-to-Enter: 0,15 Alpha-to-Remove: 0,15

Response is t on 9 predictors, with N = 91

Step	1	2	3	4
Constant	27,09	23,22	25,20	14,48
Cmax	-7,4	-11,5	-10,7	-13,3
T-Value	-7,39	-10,50	-10,29	-10,65
P-Value	0,000	0,000	0,000	0,000
aT		0,520	0,514	0,351
T-Value		5,95	6,32	3,85
P-Value		0,000	0,000	0,000
Cmin			-6,0	-5,7
T-Value			-3,82	-3,85
P-Value			0,000	0,000
Tf				0,37
T-Value				3,37
P-Value				0,001
S	3,09	2,62	2,44	2,31
R-Sq	38,04	55,80	62,15	66,56
R-Sq(adj)	37,34	54,79	60,85	65,01
PRESS	889,916	667,646	582,853	533,562
R-Sq(pred)	35,08	51,29	57,48	61,08

Discussão: com 4 iterações ao nível de significância de 15% o ajuste apresentou as variáveis Cmax, aT, Cmin, e Tf como as possíveis variáveis preditoras do modelo de regressão. Todas as estatísticas p-value são significativas e o modelo com estas possíveis variáveis apresentam um poder de explicação de 65,01%.

Método best subset

Não foi possível, pois as variáveis preditoras são altamente correlacionadas. Solução: retirar do ajuste as variáveis manipuladas **aT**, **aC** e **am** que possuem a mesma informação das demais variáveis por serem delas derivadas, por isso, a alta correlação.

Ajuste do Modelo com Variáveis Transformadas

Best Subsets Regression: t versus Ti; Tf; ...

Response is t

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	Ti	Tf	aTx	Cmin	amx	Cmin^2	1/Cmax
1	43,6	42,9	97,8	2,9479							X
1	38,0	37,3	115,9	3,0893			X				
2	59,8	58,9	46,6	2,5016		X					X
2	55,2	54,2	61,7	2,6413		X	X				
3	67,9	66,8	22,2	2,2492	X	X					X
3	64,3	63,1	33,8	2,3701		X		X			X
4	72,4	71,1	9,4	2,0974	X	X		X			X
4	71,9	70,6	11,1	2,1170	X	X				X	X
5	73,6	72,0	7,6	2,0651	X	X	X	X		X	
5	73,1	71,5	9,1	2,0834	X	X	X			X	X
6	74,4	72,6	6,9	2,0440	X	X	X	X	X	X	
6	74,1	72,3	7,7	2,0542	X	X	X	X		X	X
7	74,8	72,7	7,5	2,0397	X	X	X	X	X	X	X
7	74,6	72,5	8,2	2,0480	X	X	X	X	X	X	X
8	75,0	72,6	8,8	2,0434	X	X	X	X	X	X	X
8	74,9	72,5	9,1	2,0470	X	X	X	X	X	X	X
9	75,3	72,5	10,0	2,0458	X	X	X	X	X	X	X

***Em destaque, possíveis modelos que se ajustam adequadamente.**

Modelo de Regressão Ajustado

The regression equation is

t = - 21,7 - 0,418 Ti + 0,723 Tf + 21,8 1/Cmax

Predictor	Coef	SE Coef	T	P
Constant	-21,691	5,027	-4,32	0,000
Ti	-0,41794	0,08940	-4,68	0,000
Tf	0,72308	0,09137	7,91	0,000
1/Cmax	21,797	1,716	12,70	0,000

S = 2,24923 R-Sq = 67,9% R-Sq(adj) = 66,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	930,63	310,21	61,32	0,000
Residual Error	87	440,14	5,06		
Total	90	1370,77			

Source	DF	Seq SS
Ti	1	86,25
Tf	1	28,49
1/Cmax	1	815,89

Discussão: verificando a menor estatística Cp de Mallows para o maior R-Sq ajustado (poder de explicação do modelo), quatro modelos mostraram-se interessantes. Cada modelo foi ajustado para o modelo de regressão e com a análise de resíduos se optou por um dos modelos. Todas as estatísticas p do modelo ajustado significativas.

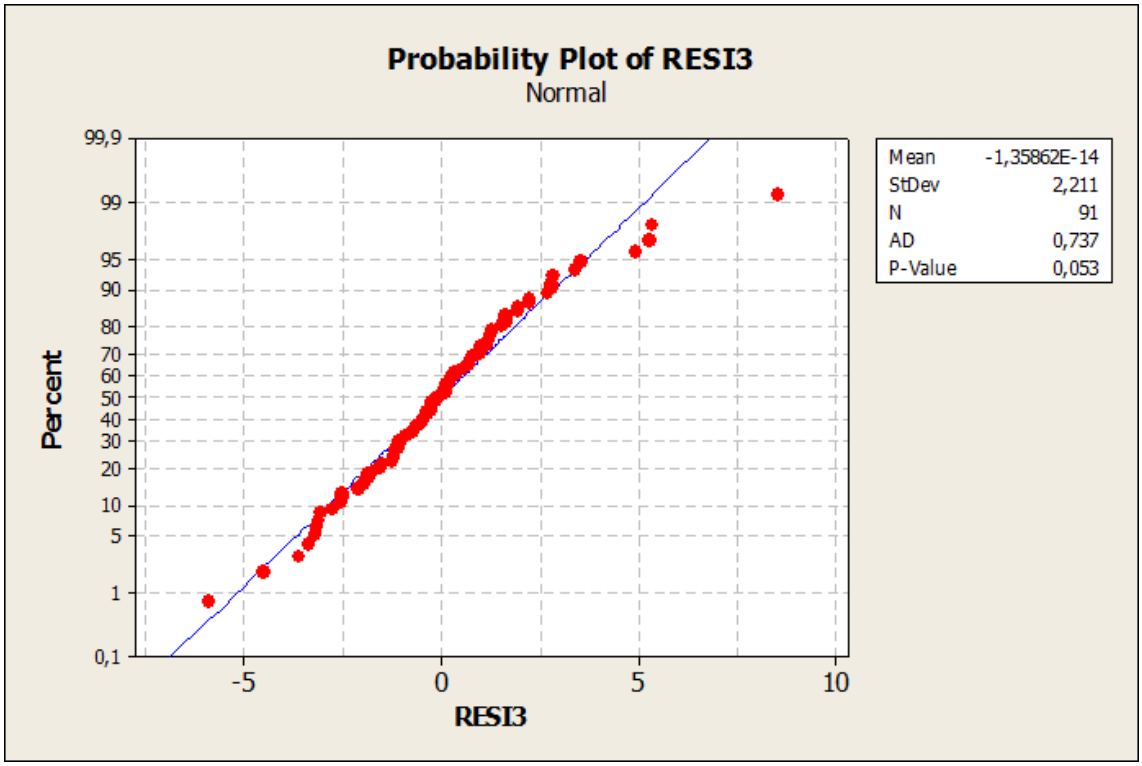


Figura 15: Distribuição Normal dos Resíduos do Modelo Ajustado

Discussão: Os resíduos do modelo ajustado seguem uma distribuição normal ao nível de significância de 5%. O ajuste adequado do modelo gera resíduos com distribuição normal.

Por fim, a equação de regressão do modelo é explicativa e preditiva:

```
t = - 21,7 - 0,418 Ti + 0,723 Tf + 21,8 1/Cmax
```

Discussão: a equação do modelo ajustado tem a capacidade de explicar o comportamento da série de dados e prever valores futuros. O tempo de desgaste tem uma relação inversa com a Ti (temperatura inicial) com um fator $\beta_{\_} = -$; o tempo de desgaste gasta uma relação proporcional com a Tf (temperatura final) e o inverso da Cmax (corrente máxima) com respectivos fatores $\beta_{\_} = +$ e $\beta_{\_} = +$.

5. Considerações Finais

Durante a eletrólise, o desgaste de eletrodos, sobretudo os de grafite, se configura o fator limitante na viabilidade do método na obtenção de hidrogênio, um vetor energético de destaque. Devido à grande complexidade da relação entre as variáveis envolvidas no processo, métodos mais simples de inferência estatística são limitados. As técnicas computacionais envolvendo os modelos lineares generalizados demonstraram-se adequadas para descrever o comportamento da série de dados e prever valores futuros do desgaste de eletrodos.

Com o entendimento da dinâmica de desgaste dos eletrodos é possível atuar no processo, inserindo melhoramentos e propostas no sentido de aumentar o tempo de desgaste de eletrodos de grafite quando submetidos à eletrólise, ou ainda melhor, propor a substituição por materiais de engenharia mais modernos, se possível mais baratos e mais versáteis.

A variabilidade inicial da série, devido ao “descontrole” da eletrólise, pode ser diminuída com mais amostragens no período “sobre controle” do experimento. As melhorias propostas contribuíram com a estabilidade do sistema.

Para trabalhos futuros é possível avaliar outras variáveis do processo; considerar os balanços mássico e energético; trabalhar com diferentes fabricantes de eletrodos de grafite, gerando mais dinâmicas entre as variáveis; considerar múltiplos eletrodos sofrendo corrosão; verificar o desgaste utilizando água potável e deionizada e ainda considerar outras materiais na fabricação de eletrodos.

6. Referências

ATKINS, P.; PAULA, J. de. **Atkin's Physical Chemistry**. Oxford. 7th edition.

BAGOTSKY, V. S. **Fundamentals of Electrochemistry**. 2nd ed. Wiley-Interscience. 2006.

BARD, A. J.; FAULKNER, L. R. **Electrochemical Methods: Fundamentals and Applications**. 2nd ed. John Wiley & Sons, INC. 2001.

BARRANTE, J. R. **Applied Mathematics for Physical Chemistry**. Prentice-Hall. 1998. 2nd edition.

BRETT, C. M. A.; BRETT, A. M. O. **Electrochemistry: Principles, Methods, and Applications.** 1994. Oxford University Press.

CASTELLAN, G. W. **Fundamentos de Físico-Química: Sistema SI**. Rio de Janeiro: LTC, 1996. 527p.

DAVID, R. L. **Handbook of CHEMISTRY and PHYSICS**.CRC PRESS. 2003-2004. 84th edition.

Encyclopedia of Physical Science and Technology Analytical Chemistry. 3rd edition.

GRAYBILL, F. A.; IYER, H. K. **Regression Analysis: Concepts and Applications**. Duxbury Pr; 1st edition; 1994. 650p.

HALE, A. J. **The Manufacture of Chemicals by Electrolysis.** Constable & Company LTD, London. 1919.

LEWICKI, P.; HILL, T. **Statistics: Methods and Applications.** Statsof, Inc. 2006. 832p.

MITTAL, H. V. **R Graphs Cookbook.** Packt Publishing: Birmingham-Mumbai. 2011.

MONTGOMERY, D. C; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros.** 4ª Ed. LTC. 2009. 514p.

MYERS, R. T.; OLDHAM, K. B.; TOCCI, S. **HOLT CHEMISTRY**. Holt, Rineheart an Wisnton, 2006.

STANSBURY, E. E.; BUCHANAN R.A. **Fundamentals of Electrochemical Corrosion**. ASM international. 2000.

TEED, P. L. **The Chemistry and Manufature of Hydrogen**. New York, Longmans, Green and Co: London, Arnold. 1919.

WANG, J. **Analytical Electrochemistry**. 2nd ed. Wiley-VCH. 2000.

WEST, D. M.; HOLLER, J. F.; SKOOG, D. A**. Fundamentos de Química Analítica.** Philadelphia: Thomsom, 2005, 1124p. 8ª edição.