EST507 – Modelos de Regressão

PIPGEs ICMC/USP UFSCar

2017

Ementa

- Conceitos básicos e notações.
- Modelos lineares.
- Método dos mínimos quadrados.
- Testes de hipóteses e intervalos de confiança.
- Família exponencial de distribuições.
- Componentes dos modelos lineares generalizados.
- Método de máxima verossimilhança.
- Estimação e inferência em modelos lineares generalizados.
- Verificação da adequação de modelos.
- Modelos para respostas binárias.
- Modelos para dados de contagens.

Bibliografia

- Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W. Applied Linear Statistical Models. 5th ed., McGraw-Hill Irwin: Boston, 2005.
- McCullagh, P., Nelder, J. A. Generalized Linear Models. 2nd ed., Chapman & Hall: London, 1989.
- Paula, G.A. Modelos de Regressão com Apoio Computacional. IME/USP: São Paulo-SP, 2013 (`https://www.ime.usp.br/~giapaula/texto_2013.pdf`).
- Searle, S. R. Linear Models. Wiley: New York, 1997.

Avaliação

Sendo MF a média final, P_1 a nota da primeira prova e P_2 a nota da segunda prova, tem-se

$$MF = 0,5 \times (P_1 + P_2).$$

O conceito na disciplina utilizará o seguinte critério:

- $MF \geqq 8,5$: conceito A.
- $7,0 \leqq MF < 8,5$: conceito B.
- $5,0 \leqq MF < 7,0$: conceito C.
- $MF < 5,0$: reprovação.

Avaliação

Sendo MF a média final, P_1 a nota da primeira prova e P_2 a nota da segunda prova, tem-se

$$MF = 0,5 \times (P_1 + P_2).$$

O conceito na disciplina utilizará o seguinte critério:

- $MF \geqq 8,5$: conceito A.
- $7,0 \leqq MF < 8,5$: conceito B.
- $5,0 \leqq MF < 7,0$: conceito C.
- $MF < 5,0$: reprovação.

Datas das provas: 27/5/2017 e 8/7/2017 (dois sábados).

Avaliação

Sendo MF a média final, P_1 a nota da primeira prova e P_2 a nota da segunda prova, tem-se

$$MF = 0,5 \times (P_1 + P_2).$$

O conceito na disciplina utilizará o seguinte critério:

- $MF \geqq 8,5$: conceito A.
- $7,0 \leqq MF < 8,5$: conceito B.
- $5,0 \leqq MF < 7,0$: conceito C.
- $MF < 5,0$: reprovação.

Datas das provas: 27/5/2017 e 8/7/2017 (dois sábados).

Análise de regressão

Análise de regressão é um conjunto de técnicas estatísticas utilizadas para estudar a relação entre alguma característica de interesse de uma variável e uma ou mais outras variáveis.

A variável de interesse, aquela que queremos explicar em função de outras variáveis, é denominada variável resposta (variável dependente, regressando, . . .).

Análise de regressão

Análise de regressão é um conjunto de técnicas estatísticas utilizadas para estudar a relação entre alguma característica de interesse de uma variável e uma ou mais outras variáveis.

A variável de interesse, aquela que queremos explicar em função de outras variáveis, é denominada variável resposta (variável dependente, regressando, . . .).

As variáveis que utilizamos para explicar a variável resposta são denominadas variáveis explicativas (variáveis independentes, regressores, variáveis preditoras, preditores, variáveis explanatórias, covariáveis, . . .).

Análise de regressão

Análise de regressão é um conjunto de técnicas estatísticas utilizadas para estudar a relação entre alguma característica de interesse de uma variável e uma ou mais outras variáveis.

A variável de interesse, aquela que queremos explicar em função de outras variáveis, é denominada variável resposta (variável dependente, regressando, ...).

As variáveis que utilizamos para explicar a variável resposta são denominadas variáveis explicativas (variáveis independentes, regressores, variáveis preditoras, preditores, variáveis explanatórias, covariáveis, ...).

Exemplos

Assunto: renda presumida.
Variável resposta: renda de um indivíduo.
Variáveis preditoras: profissão, sexo, idade, volume médio de investimentos, valor médio da fatura do cartão de crédito, valor médio mensal dos depósitos, etc.

Assunto: internação de clientes de planos de saúde.
Variável resposta: dias de internação.
Variáveis preditoras: tipo de internação (eletiva, urgência ou obstétrica), idade do paciente e UF da internação.

Exemplos

Assunto: renda presumida.
Variável resposta: renda de um indivíduo.
Variáveis preditoras: profissão, sexo, idade, volume médio de investimentos, valor médio da fatura do cartão de crédito, valor médio mensal dos depósitos, etc.

Assunto: internação de clientes de planos de saúde.
Variável resposta: dias de internação.
Variáveis preditoras: tipo de internação (eletiva, urgência ou obstétrica), idade do paciente e UF da internação.

Exemplos

Assunto: risco de crédito.
Variável resposta: indicador de ser bom ou mau cliente de cartão de crédito.
Variáveis preditoras: número máximo de dias de atraso no pagamentos das faturas nos últimos seis meses, indicador de saque com o cartão, percentual da fatura pago no mês anterior ao da coleta dos dados, tempo que o cliente tem o cartão, etc.

Exemplos

Uma consultoria desenvolveu para uma indústria um teste de aptidão para funcionários de um determinado setor. Para avaliar se o teste poderia ser útil na seleção de novos funcionários, a indústria aplicou o teste a oito novos funcionários e três meses depois observou um índice de produtividade de cada um desses funcionários. Os dados se encontram abaixo.

Aptidão	22	25	15	19	23	18	17	21
Produtividade	36	41	25	34	41	30	29	35

Baseado nos resultados obtidos, esse teste de aptidão deve ser usado na seleção de novos funcionários?

Exemplos

Uma consultoria desenvolveu para uma indústria um teste de aptidão para funcionários de um determinado setor. Para avaliar se o teste poderia ser útil na seleção de novos funcionários, a indústria aplicou o teste a oito novos funcionários e três meses depois observou um índice de produtividade de cada um desses funcionários. Os dados se encontram abaixo.

Aptidão	22	25	15	19	23	18	17	21
Produtividade	36	41	25	34	41	30	29	35

Baseado nos resultados obtidos, esse teste de aptidão deve ser usado na seleção de novos funcionários?

O resultado do teste de aptidão é um bom preditor do índice de produtividade?

Exemplos

Uma consultoria desenvolveu para uma indústria um teste de aptidão para funcionários de um determinado setor. Para avaliar se o teste poderia ser útil na seleção de novos funcionários, a indústria aplicou o teste a oito novos funcionários e três meses depois observou um índice de produtividade de cada um desses funcionários. Os dados se encontram abaixo.

Aptidão	22	25	15	19	23	18	17	21
Produtividade	36	41	25	34	41	30	29	35

Baseado nos resultados obtidos, esse teste de aptidão deve ser usado na seleção de novos funcionários?
O resultado do teste de aptidão é um bom preditor do índice de produtividade?

Exemplos

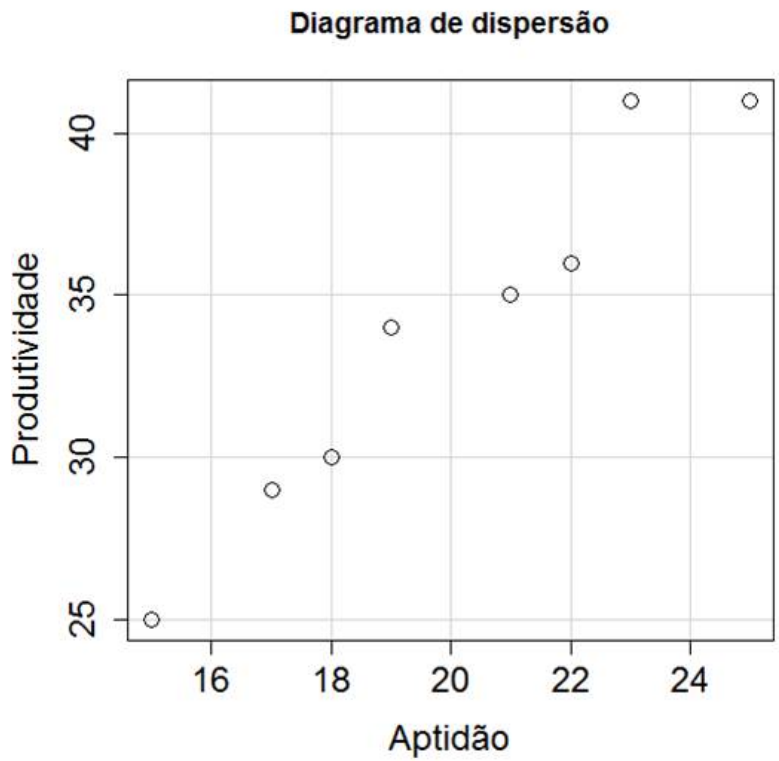

Figura 1: Produtividade e aptidão de funcionários.

Modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \ i = 1, \ldots, n, \tag{1}$$

em que
n é o número de observações (tamanho da amostra),
Y_i é a variável resposta para a unidade observacional i,
β_0 e β_1 são parâmetros desconhecidos (coeficientes da regressão),
x_i é o valor da variável preditora (não aleatória) para a unidade observacional i e
ϵ_i é o erro aleatório (inobservável) para a unidade observacional i.

Suposições usuais:

$\mathrm{E}(\epsilon_i) = 0$,
$\mathrm{Var}(\epsilon_i) = \sigma^2$, desconhecida, $0 < \sigma^2 < \infty$ e
$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, para todo $i \neq j$, $i, j = 1, \ldots, n$.

Modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \ i = 1, \ldots, n, \tag{1}$$

em que
n é o número de observações (tamanho da amostra),
Y_i é a variável resposta para a unidade observacional i,
β_0 e β_1 são parâmetros desconhecidos (coeficientes da regressão),
x_i é o valor da variável preditora (não aleatória) para a unidade observacional i e
ϵ_i é o erro aleatório (inobservável) para a unidade observacional i.
Suposições usuais:
$\mathrm{E}(\epsilon_i) = 0$,
$\mathrm{Var}(\epsilon_i) = \sigma^2$, desconhecida, $0 < \sigma^2 < \infty$ e
$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, para todo $i \neq j$, $i, j = 1, \ldots, n$.

Modelo de regressão linear simples

Na expressão (1), $\mathrm{E}(Y) = \beta_0 + \beta_1 x$ é uma função linear de β_0 e β_1.
O modelo em que

$$Y_i = \beta_0 + \beta_1 \log(x_i + 1) + \epsilon_i, \;\; x_i \geqq 0, \;\; i = 1, \ldots, n, \qquad (2)$$

com as mesmas suposições usuais de antes, também é linear.

Modelo de regressão linear simples

Na expressão (1), $\mathrm{E}(Y) = \beta_0 + \beta_1 x$ é uma função linear de β_0 e β_1.
O modelo em que

$$Y_i = \beta_0 + \beta_1 \log(x_i + 1) + \epsilon_i, \;\; x_i \geqq 0, \;\; i = 1, \ldots, n, \tag{2}$$

com as mesmas suposições usuais de antes, também é linear.
O modelo em que

$$Y_i = \frac{\beta_0}{1 + \exp(\beta_1 - \beta_2 x_i)} + \epsilon_i, \;\; i = 1, \ldots, n, \tag{3}$$

com as mesmas suposições usuais de antes, é não linear.

Modelo de regressão linear simples

Na expressão (1), $\mathrm{E}(Y) = \beta_0 + \beta_1 x$ é uma função linear de β_0 e β_1.
O modelo em que

$$Y_i = \beta_0 + \beta_1 \log(x_i + 1) + \epsilon_i, \;\; x_i \geqq 0, \;\; i = 1, \ldots, n, \tag{2}$$

com as mesmas suposições usuais de antes, também é linear.
O modelo em que

$$Y_i = \frac{\beta_0}{1 + \exp(\beta_1 - \beta_2 x_i)} + \epsilon_i, \;\; i = 1, \ldots, n, \tag{3}$$

com as mesmas suposições usuais de antes, é não linear.

O que é regressão?

Em muitas situações, a característica de interesse da variável resposta Y é sua esperança $\mathrm{E}(Y)$, que é relacionada com variáveis explicativas $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$.

Este é possivelmente o significado mais usual do termo regressão.

O que é regressão?

Em muitas situações, a característica de interesse da variável resposta Y é sua esperança $\mathrm{E}(Y)$, que é relacionada com variáveis explicativas $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$.

Este é possivelmente o significado mais usual do termo regressão.

Pode haver interesse em outras características da variável resposta, como sua variância e alguns de seus quantis.

O que é regressão?

Em muitas situações, a característica de interesse da variável resposta Y é sua esperança $\mathrm{E}(Y)$, que é relacionada com variáveis explicativas $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$.
Este é possivelmente o significado mais usual do termo regressão.
Pode haver interesse em outras características da variável resposta, como sua variância e alguns de seus quantis.
Se $\boldsymbol{\theta}$ é o vetor de parâmetros da distribuição de Y, pode haver interesse em relacionar elementos de $\boldsymbol{\theta}$ com variáveis explicativas.

O que é regressão?

Em muitas situações, a característica de interesse da variável resposta Y é sua esperança $\mathrm{E}(Y)$, que é relacionada com variáveis explicativas $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$.
Este é possivelmente o significado mais usual do termo regressão.
Pode haver interesse em outras características da variável resposta, como sua variância e alguns de seus quantis.
Se $\boldsymbol{\theta}$ é o vetor de parâmetros da distribuição de Y, pode haver interesse em relacionar elementos de $\boldsymbol{\theta}$ com variáveis explicativas.
O modelo

$$Y_i = r(x_i) + \epsilon_i, \;\; i = 1, \ldots, n, \tag{4}$$

com as mesmas suposições usuais de antes, em que $r(\cdot)$ é um suavizador, é um modelo não paramétrico.

O que é regressão?

Em muitas situações, a característica de interesse da variável resposta Y é sua esperança $\mathrm{E}(Y)$, que é relacionada com variáveis explicativas $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$.
Este é possivelmente o significado mais usual do termo regressão.
Pode haver interesse em outras características da variável resposta, como sua variância e alguns de seus quantis.
Se $\boldsymbol{\theta}$ é o vetor de parâmetros da distribuição de Y, pode haver interesse em relacionar elementos de $\boldsymbol{\theta}$ com variáveis explicativas.
O modelo

$$Y_i = r(x_i) + \epsilon_i, \;\; i = 1, \ldots, n, \tag{4}$$

com as mesmas suposições usuais de antes, em que $r(\cdot)$ é um suavizador, é um modelo não paramétrico.

O que é regressão?

Se Y representa tempo de vida ($Y \geqq 0$), o interesse pode ser a função risco de Y, dada por $h(y) = f(y)/S(y)$, em que $S(y) = 1 - F(y)$ é a função sobrevivência de Y.

Um modelo bastante utilizado tem expressão $h(y) = h_0(y)exp(\boldsymbol{x}^\top \beta)$, em que $h_0(\cdot)$ representa a função risco basal. As covariáveis $\boldsymbol{x}$ modificam o risco de base.

O que é regressão?

Se Y representa tempo de vida ($Y \geqq 0$), o interesse pode ser a função risco de Y, dada por $h(y) = f(y)/S(y)$, em que $S(y) = 1 - F(y)$ é a função sobrevivência de Y.
Um modelo bastante utilizado tem expressão $h(y) = h_0(y)exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, em que $h_0(\cdot)$ representa a função risco basal. As covariáveis $\boldsymbol{x}$ modificam o risco de base.
Modelos semiparamétricos.

O que é regressão?

Se Y representa tempo de vida ($Y \geqq 0$), o interesse pode ser a função risco de Y, dada por $h(y) = f(y)/S(y)$, em que $S(y) = 1 - F(y)$ é a função sobrevivência de Y.

Um modelo bastante utilizado tem expressão $h(y) = h_0(y)exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, em que $h_0(\cdot)$ representa a função risco basal. As covariáveis $\boldsymbol{x}$ modificam o risco de base.

Modelos semiparamétricos.

Modelos aditivos generalizados (GAM).

O que é regressão?

Se Y representa tempo de vida ($Y \geqq 0$), o interesse pode ser a função risco de Y, dada por $h(y) = f(y)/S(y)$, em que $S(y) = 1 - F(y)$ é a função sobrevivência de Y.
Um modelo bastante utilizado tem expressão $h(y) = h_0(y)exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, em que $h_0(\cdot)$ representa a função risco basal. As covariáveis $\boldsymbol{x}$ modificam o risco de base.
Modelos semiparamétricos.
Modelos aditivos generalizados (GAM).
Modelos aditivos generalizados para localização, escala e forma (GAMLSS).

O que é regressão?

Se Y representa tempo de vida ($Y \geqq 0$), o interesse pode ser a função risco de Y, dada por $h(y) = f(y)/S(y)$, em que $S(y) = 1 - F(y)$ é a função sobrevivência de Y.
Um modelo bastante utilizado tem expressão $h(y) = h_0(y)exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, em que $h_0(\cdot)$ representa a função risco basal. As covariáveis $\boldsymbol{x}$ modificam o risco de base.
Modelos semiparamétricos.
Modelos aditivos generalizados (GAM).
Modelos aditivos generalizados para localização, escala e forma (GAMLSS).
Modelos de regressão bayesianos.
E muito mais . . . .

O que é regressão?

Se Y representa tempo de vida ($Y \geqq 0$), o interesse pode ser a função risco de Y, dada por $h(y) = f(y)/S(y)$, em que $S(y) = 1 - F(y)$ é a função sobrevivência de Y.
Um modelo bastante utilizado tem expressão $h(y) = h_0(y)exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, em que $h_0(\cdot)$ representa a função risco basal. As covariáveis $\boldsymbol{x}$ modificam o risco de base.
Modelos semiparamétricos.
Modelos aditivos generalizados (GAM).
Modelos aditivos generalizados para localização, escala e forma (GAMLSS).
Modelos de regressão bayesianos.
E muito mais ....
Na área de Aprendizagem de Máquina (*Machine Learning*), análise de regressão é um método de aprendizado supervisionado (*supervised learning*).

O que é regressão?

Se Y representa tempo de vida ($Y \geqq 0$), o interesse pode ser a função risco de Y, dada por $h(y) = f(y)/S(y)$, em que $S(y) = 1 - F(y)$ é a função sobrevivência de Y.
Um modelo bastante utilizado tem expressão $h(y) = h_0(y)exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, em que $h_0(\cdot)$ representa a função risco basal. As covariáveis $\boldsymbol{x}$ modificam o risco de base.
Modelos semiparamétricos.
Modelos aditivos generalizados (GAM).
Modelos aditivos generalizados para localização, escala e forma (GAMLSS).
Modelos de regressão bayesianos.
E muito mais ....
Na área de Aprendizagem de Máquina (*Machine Learning*), análise de regressão é um método de aprendizado supervisionado (*supervised learning*).