UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 8: REGRESSÃO POLINOMIAL E SELEÇÃO DE MODELOS

Salvador

2023

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 8: REGRESSÃO POLINOMIAL E SELEÇÃO DE MODELOS

Atividades de laboratório apresentadas ao Instituto de Matemática e Estatística da Universidade Federal da Bahia como parte das exigências da disciplina Análise de Regressão ministrada pela professora Dra. Edleide de Brito.

Salvador

2023

SUMÁRIO

1 INTRODUÇÃO **1**

1.1 Atividade 1 - Regressão polinomial 1

1.2 Atividade 2 - Seleção de modelos 1

2 REGRESSÃO POLINOMIAL **3**

2.1 Análise descritiva 3

2.2 Modelo polinomial com a renda 5

2.2.1 Colinearidade 7

2.3 Modelo polinomial com a renda centrada na média 7

2.3.1 Colinearidade 8

2.4 Análise dos resíduos 9

3 SELEÇÃO DE MODELOS **11**

3.1 Análise descritiva 12

3.2 Modelo de regressão linear múltiplo 14

3.3 Seleção de covariáveis 15

3.3.1 Eliminação *backward* baseada no teste F 15

3.3.2 Seleção *stepwise* baseada no AIC 15

3.3.3 Seleção *stepwise* baseada no BIC 15

3.3.4 Todas as regressões possíveis 16

3.4 Análise dos resíduos 16

4 CONCLUSÃO **18**

4.1 Atividade 1 - Regressão polinomial 18

4.2 Atividade 2 - Seleção de modelos 18

1 Introdução

1.1 Atividade 1 - Regressão polinomial

A primeira atividade tem como objetivo investigar o efeito da renda anual do marido no tempo decorrido entre o casamento e o nascimento do primeiro filho. Para realizar essa análise, técnicas de regressão polinomial serão utilizadas, buscando identificar o modelo que melhor se ajusta aos dados.

Na primeira etapa da análise, será realizada uma análise descritiva dos dados. Posteriormente, serão ajustados modelos de regressão polinomial para explorar a relação entre a renda e o tempo. A qualidade do ajuste de cada modelo será avaliada por meio dos coeficientes de determinação ajustados e da comparação dos valores-p das variáveis de renda.

Além disso, será verificado se há multicolinearidade entre as variáveis de renda, a fim de evitar problemas de estimativas imprecisas e interpretações incorretas dos coeficientes. Serão calculados os valores do fator de inflação da variância (VIF) e coeficientes de determinação para cada variável independente, a fim de identificar a presença de multicolinearidade.

Por fim, será realizada uma análise dos resíduos para verificar se o modelo escolhido atende aos pressupostos da regressão linear. Eventuais pontos discrepantes ou desvios dos pressupostos serão levados em consideração para uma análise mais precisa e confiável dos resultados.

Assim, espera-se obter uma compreensão mais aprofundada sobre o efeito da renda anual do marido no tempo entre o casamento e o nascimento do primeiro filho.

1.2 Atividade 2 - Seleção de modelos

Já a segunda atividade teve como objetivo utilizar diversos métodos de seleção de modelos a fim de analisar o logaritmo do antígeno específico (*laep*) da próstata dado outras variáveis que estão relacionadas com o câncer de próstata.

Primeiramente, será realizado uma análise descritiva a fim de observar a distribuição das variáveis e a relação que essas variáveis têm com a variável resposta *laep*, além da relação que elas têm ente si. Após isso, será estimado um modelo completo com todas as variáveis explicativas contínuas e será avaliado quais variáveis foram significativas, se o modelo foi significativo, ao nível de 5%, e o quão esse modelo explica a variabilidade da variável resposta.

A parte mais crucial será a parte dos métodos de seleção de modelos, onde utilizando os métodos *backwise* baseado no teste F, *stepwise* tanto baseado no AIC quanto no BIC, serão selecionados os melhores conjuntos de variáveis explicativas (os melhores modelos) para predizer o logaritmo do antígeno específico. Ademais, também será observado todas as possibilidades de modelos com essas variáveis e serão selecionados os cinco melhores modelos baseados no R^2 ajustado. Por fim, será realizado uma análise de resíduos, para verificar se os resíduos do modelo que foi considerado o mais adequado por esses métodos no geral não violam as suposições de homoscedasticidade, normalidade e independência.

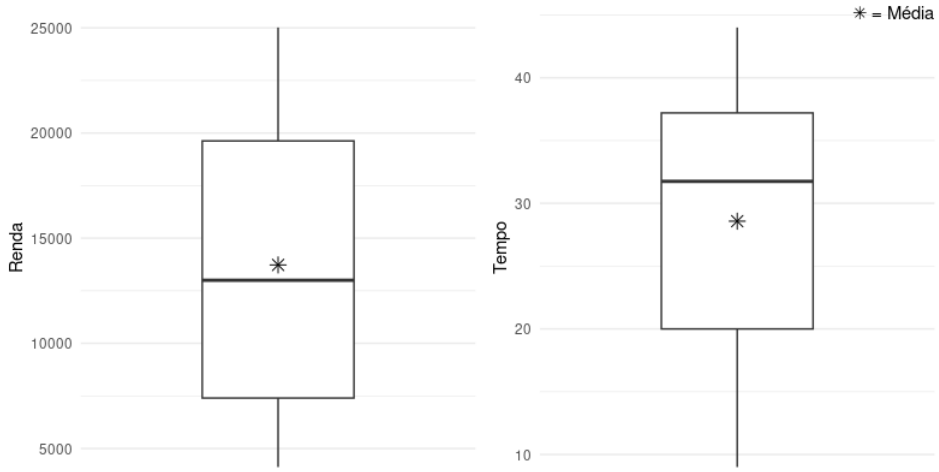
~~*lvolca*~~

2 Efeito da renda anual do marido no tempo entre o casamento e o nascimento do primeiro filho

2.1 Análise descritiva

Observando os *boxplots* da Figura 2.1, é possível notar que existe uma simetria na distribuição da variável renda, pois a mediana e a média são próximas. Além disso, a variabilidade entre o primeiro quartil e a mediana é similar à variabilidade entre a mediana e o terceiro quartil. Enquanto que distribuição do tempo parece ter uma leve assimetria à esquerda, pois a mediana é maior que a média, além de apresentar uma maior variabilidade entre o primeiro quartil e a mediana do que entre a mediana e o terceiro quartil.

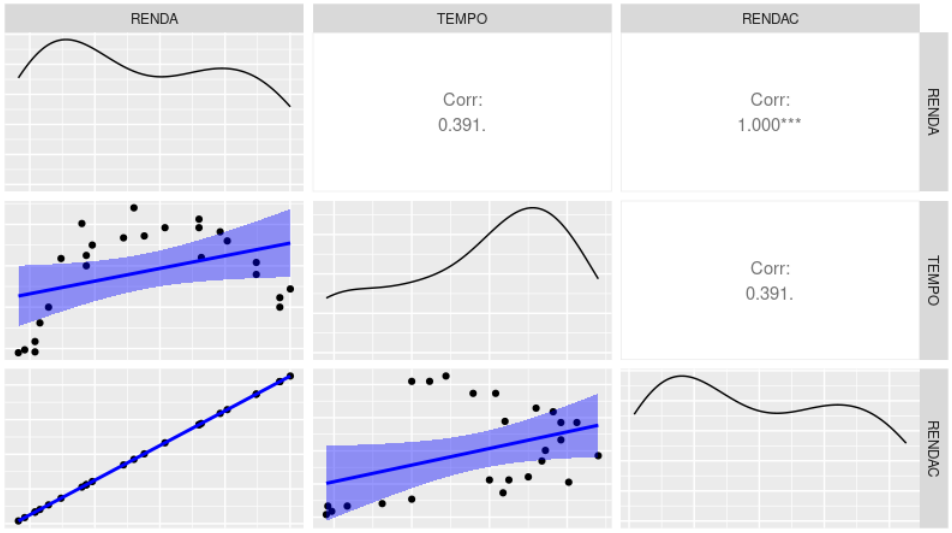
Figura 2.1: *Boxplots* da renda e do tempo até o primeiro filho de casais com pelo menos um filho



A correlação apresentada na Figura 2.2 indica que a renda apresenta uma correlação positiva moderada com a variável reposta, o tempo. O gráfico de dispersão da renda e

o tempo aponta que a relação que existe entre essas variáveis é não linear, bem como a relação entre a variável renda centrada na média, *RENDAC*, também apresenta uma relação não linear com a variável resposta.

Figura 2.2: Gráfico de dispersão, densidade e correlação das características dos casais com pelo menos um filho



2.2 Modelo polinomial com a renda

As Tabelas 2.1, 2.2 e 2.3 apresentam sumarizações das regressões lineares em que as variáveis que denotam a renda são utilizadas para predizer o tempo até o primeiro filho.

Na Tabela 2.1, o coeficiente de intercepto aponta que, quando a renda é zero, o tempo até o primeiro filho é esperado ser de aproximadamente 20,18 meses (1 ano e 8 meses). O coeficiente para a variável de renda, que é significativo ao nível de 10%, indica que um aumento unitário na renda está associado a um aumento de 0,0006 meses (aproximadamente 23 minutos) no tempo até o primeiro filho.

Tabela 2.1: Sumarização da regressão linear em que a renda é utilizada para predizer o tempo até o primeiro filho

Coeficientes	Coeficiente	EP	Est. T	Valor-p	R^2_{ajust}
Intercepto	20,18	4,61	4,38	<0,001	0,12
Renda	0,0006	0,0003	2,04	0,053	

Na Tabela 2.2, o coeficiente de intercepto, neste caso, não tem interpretabilidade, pois indicou um tempo negativo quando a renda é zero. Os valores-p para ambas as variáveis de renda são significativamente menores que 0,001, o que sugere que tanto a renda quanto a renda ao quadrado têm uma relação significativa com o tempo até o primeiro filho ao nível de 10%.

O coeficiente de determinação ajustado para este modelo de regressão linear simples foi substancialmente superior ao coeficiente de determinação do modelo anterior.

Tabela 2.2: Sumarização da regressão linear em que a renda e a renda ao quadrado são utilizadas para predizer o tempo até o primeiro filho

Coeficientes	Coeficiente	EP	Est. T	Valor-p	R^2_{ajust}
Intercepto	-19,87	3,90	-5,09	<0,001	
Renda	0,0078	0,001	12,29	<0,001	0,87
$Renda^2$	-2,528e-07	2,195e-08	-11,52	<0,001	

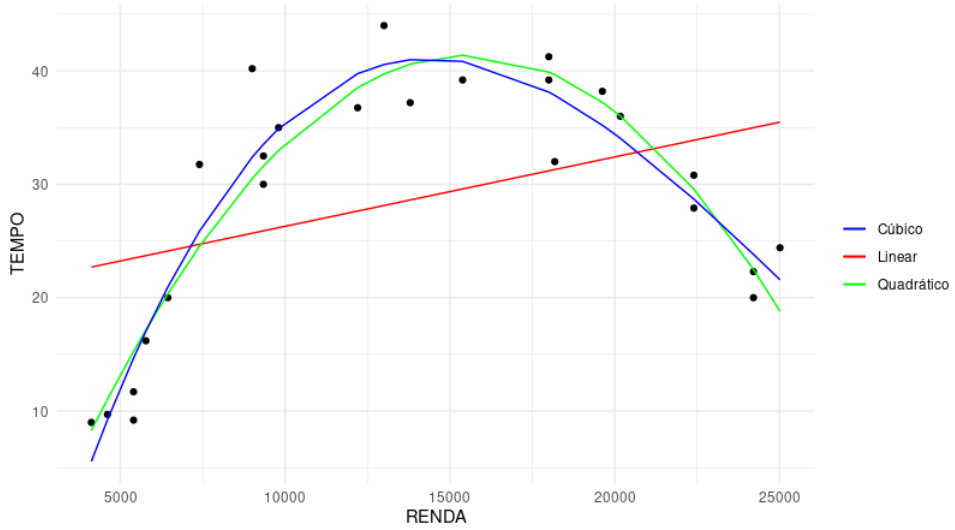
Na tabela 2.3, o coeficiente de intercepto novamente não tem interpretabilidade prática, pois, quando a renda é igual a 0, o tempo é negativo. Todos os coeficientes das variáveis de renda, renda ao quadrado e renda ao cubo são estatisticamente significativos (valores-p menores que 0,10).

Tabela 2.3: Sumarização da regressão linear com a variáveis explicativas renda, renda ao quadrado e renda ao cubo

Coeficientes	Coeficiente	EP	Est. T	Valor-p	R^2_{ajust}
Intercepto	-35,29	8,147	-4,33	<0,001	0,89
Renda	0,01	0,002	5,70	<0,001	
Renda^2	-5,932e-07	1,622e-07	-3,66	0,001	
Renda^3	7,807e-12	3,692e-12	2,12	0,047	

O coeficiente de determinação ajustado mostrou que a inclusão da variável de renda ao cubo melhorou a capacidade do modelo em explicar os dados em comparação com os modelos anteriores. Mas, não houve um aumento consideravelmente grande em relação ao modelo quadrático. Isso também pode ser visto na Figura 2.3, a qualidade do ajuste do modelo quadrático é similar à qualidade do ajuste do modelo cúbico.

Figura 2.3: Ajuste de diferentes modelo polinomiais de diferentes ordens



2.2.1 Colinearidade

Para analisar a multicolinearidade no modelo cúbico, podemos observar o fator de inflação da variância (VIF) na Tabela 2.4. É possível notar que, devido aos coeficientes de determinação serem bem próximos de 1, o valor de VIF para cada variável está bem alto, o que indica uma presença de forte colinearidade no modelo polinomial cúbico.

Tabela 2.4: Coeficiente de determinação e valores VIF

Variável	R_j^2	VIF
Renda	0,9974	397,94
Renda^2	0,9995	1941,08
Renda^3	0,9984	630,94

Para tentar resolver o problema de multicolinearidade presente no modelo polinomial, foi proposto refazer todos esses modelos utilizando a renda centrada na média (Rendac) como variável explicativa.

2.3 Modelo polinomial com a renda centrada na média

Na Tabela 2.5, podemos observar que o modelo de regressão linear simples com a variável explicativa sendo a renda centrada na média é significativo ao nível de 10%. Assim como para o modelo quadrático que está sumarizado na Tabela 2.6, os coeficientes da renda centrada na média e renda centrada na média ao quadrado são significativos ao nível de 10%.

Tabela 2.5: Sumarização da regressão linear em que a renda centrada na média é utilizada para predizer o tempo até o primeiro filho

Coeficientes	Coeficiente	EP	Est. T	Valor-p	R^2_{ajust}
Intercepto	28,58	2,06	13,85	<0,001	0,12
Rendac	0,0006	0,0003	2,04	0,053	

Tabela 2.6: Sumarização da regressão linear em que a renda centrada na média e a renda centrada na média ao quadrado são utilizadas para predizer o tempo até o primeiro filho

Coeficientes	Coeficiente	EP	Est. T	Valor-p	R^2_{ajust}
Intercepto	40,54	1,308	31,00	<0,001	
Rendac	9,287e-04	1,189e-04	7,81	<0,001	0,87
Rendac^2	-2,528e-07	2,195e-08	-11,52	<0,001	

Na Tabela 2.7, onde está sumarizado o modelo cúbico com a variável renda centrada na média, é possível notar que o coeficiente da renda centrada na média não é significativa

ao nível de 10%, entretanto os coeficientes da renda centrada na média ao quadrado e ao cubo são. Todavia, não é aconselhável retirar o coeficiente da renda centrada na média do modelo, pois este componente está presente nos coeficientes de maior grau, já que se ao adicionar uma unidade ao polinômio de maior grau, neste caso o de grau 3, obtém-se todos os polinômios de menor grau a este (os polinômios de grau 1 e grau 2).

Tabela 2.7: Sumarização da regressão linear com a variáveis explicativas renda centrada na média, renda centrada na média ao quadrado e renda centrada na média ao cubo

Coeficientes	Coeficiente	EP	Est. T	Valor-p	R^2_{ajust}
Intercepto	40,97	1,232	33,25	<0,001	0,89
Rendac	3,539e-04	2,934e-04	1,21	0,241	
$Rendac^2$	-2,717e-07	2,225e-08	-12,21	0,001	
$Rendac^3$	7,807e-12	3,692e-12	2,12	0,047	

2.3.1 Colinearidade

Considerando a variável renda centrada na média, a Tabela 2.8 apresenta o coeficiente de determinação e o fator de inflação da variância (VIF) para cada uma das variáveis independentes nos modelos. Observa-se que os valores de VIF para este modelo cúbico com a renda centrada na média são substancialmente menores do que os valores de VIF para o modelo cúbico com a renda. Portanto, é possível dizer que a colinearidade foi bastante reduzida para este novo modelo.

Ainda assim, o VIF's para as variáveis renda e $renda^3$ foram altos. O VIF para o modelo quadrático com a renda centrada na média como variável explicativa é igual a aproximadamente 1,06, o que indica praticamente ausência de colinearidade no modelo quadrático.

Tabela 2.8: Coeficiente de determinação e valores VIF

Variável	R^2_j	VIF
Rendac	0,87	7,45
$Rendac^2$	0,20	1,26
$Rendac^3$	0,88	8,12

Em resumo, com base nos resultados apresentados, é possível concluir que o modelo de regressão polinomial de terceira ordem, tanto com renda quanto com a renda centrada na média, apresenta um ajuste melhor aos dados em comparação com os modelos anteriores.

Realizando o teste exato da razão de máxima verossimilhança para testar o modelo cúbico contra o quadrático foi obtido um valor-p de 0,047. Ou seja, rejeita-se hipótese nula de equivalência entre o modelo quadrático com o modelo cúbico, ao nível de 10%.

No entanto, é importante considerar a presença de multicolinearidade no modelo cúbico. O modelo quadrático não apresenta multicolinearidade, além disso apresentou um ajuste apenas um pouco inferior ao modelo cúbico e é menos complexo, ou seja, é um modelo mais parcimonioso.

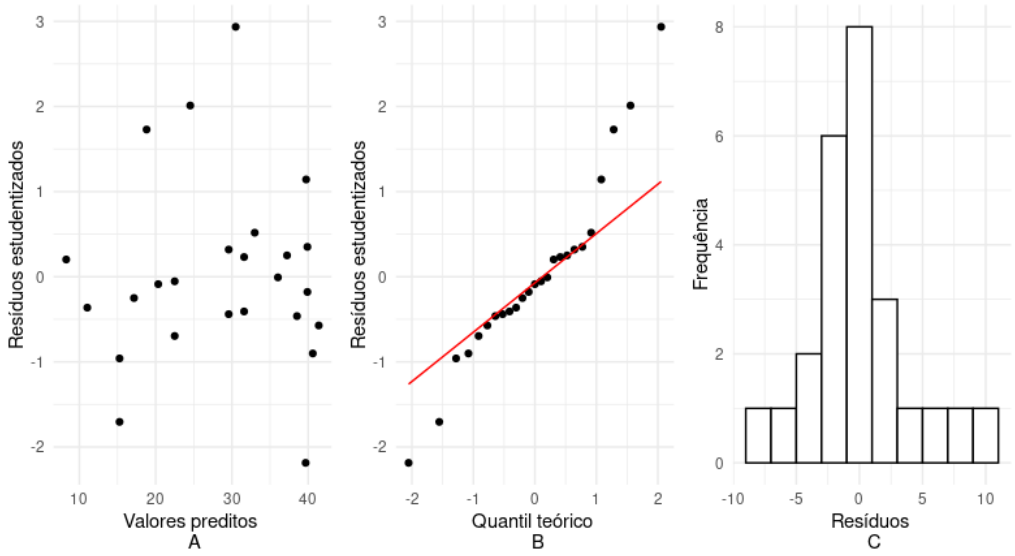
Então, o modelo quadrático com a variável renda centrada na média é o mais adequado para predizer o tempo até o primeiro filho dos casais. Mas, uma análise dos resíduos é imprescindível para verificar se esse modelo atende a todos os pressupostos e se há pontos não usuais.

2.4 Análise dos resíduos

A análise de resíduos foi feita para o modelo quadrático que considera a renda centrada na média como a variável explicativa, pois foi o modelo mais adequado segundo às análises realizadas no tópico anterior.

Observando o gráfico dos resíduos versus valores preditos, Figura 2.4 (A), é possível notar que a variabilidade dos resíduos é aparentemente constante, embora haja alguns valores discrepantes com valores de resíduos estudentizados acima de 2. O teste de *Goldfeld-Quandt* apresentou um valor-p de aproximadamente 0,19, ou seja, a homocedasticidade dos resíduos não é rejeitada ao nível de 10%.

Figura 2.4: Gráfico dos resíduos versus valores ajustados, resíduos estudentizados versus quantil teórico e histograma dos resíduos da regressão linear em que a renda é utilizada para predizer o tempo até o primeiro filho



Alguns pontos discrepantes afetam a normalidade dos resíduos na Figura 2.4. Esses

pontos discrepantes tornam a cauda da distribuição dos resíduos mais pesada. Mas, o teste *Shapiro-Wilk* apresentou um valor-p de 0,26, ou seja, a normalidade dos resíduos não é rejeitada ao nível de 10%.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente 0,95, apontando que não há autocorrelação serial nos resíduos a um nível de 10%. Logo, todos os pressupostos do modelo são atendidos.

As observações 05, 18 e 19 foram apontadas por diferentes métricas como observações não usuais. Então, faz-se necessário remover essas observações e verificar a mudança no ajuste do modelo de regressão.

3 Cirurgia de câncer de próstata

O conjunto de dados "Prostate cancer surgery"possui 97 linhas e 9 colunas. O conjunto de dados foi obtido a partir de um estudo realizado em 97 homens com câncer de próstata que estavam programados para receber uma prostatectomia radical.

O conjunto de dados possui as seguintes colunas:

- *lvolca*: log(volume do câncer). Esta coluna representa o logaritmo do volume do câncer de próstata.
- *lpeso*: log(peso da próstata). Esta coluna representa o logaritmo do peso da próstata.
- *Idade*. Esta coluna representa a idade dos pacientes.
- *lhpb*: log(quantidade de hiperplasia prostática benigna). Esta coluna representa o logaritmo da quantidade de hiperplasia prostática benigna.
- *svi*: invasão vesicular seminal. Esta coluna indica se houve invasão da vesícula seminal ou não.
- *lpc*: log(penetração capsular). Esta coluna representa o logaritmo da penetração capsular.
- *gleason*: escore de Gleason. Esta coluna representa o escore de Gleason, que é uma pontuação utilizada para avaliar a agressividade do câncer de próstata.
- *pg45*: porcentagem de escores de Gleason 4 ou 5. Esta coluna indica a porcentagem de escores de Gleason iguais a 4 ou 5.
- *laep*: log(antígeno específico da próstata). Esta coluna representa o logaritmo do antígeno específico da próstata.

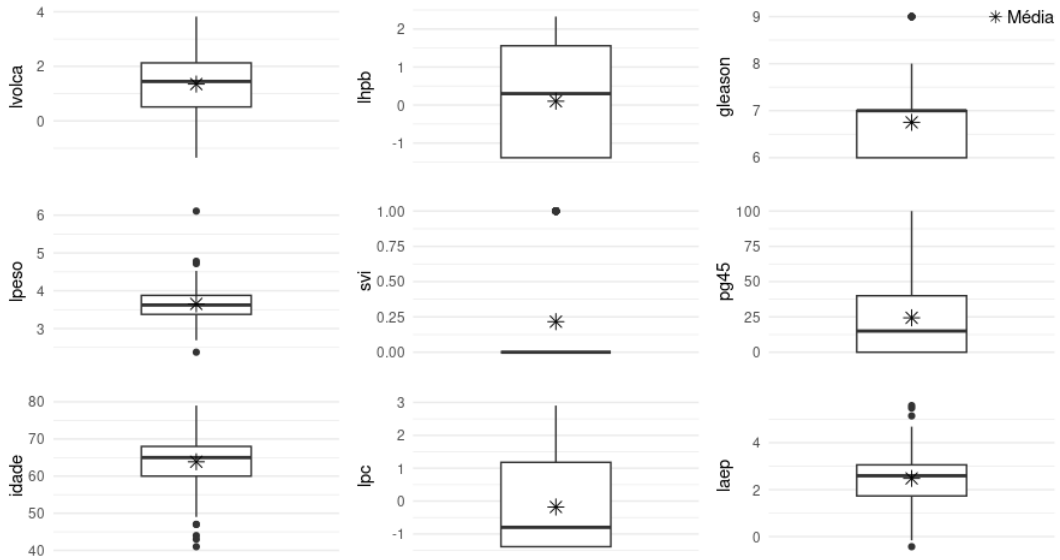
Esses dados foram coletados para realizar análises e investigações sobre o câncer de próstata, visando compreender os fatores relacionados a essa doença e sua progressão, além de explorar a eficácia da prostatectomia radical como tratamento.

3.1 Análise descritiva

As variáveis *lvolca*, *lpeso*, *idade* e *laep* apresentam simetria, com médias próximas as medianas. Mas, há pontos discrepantes nessas variáveis. As variáveis *lhpb*, *lpc* e *pg45* possuem médias próximas as medianas, entretanto nessas variáveis é possível notar a presença de mais variabilidade entre um dos quartis do que no outro, indicando falta de simetria.

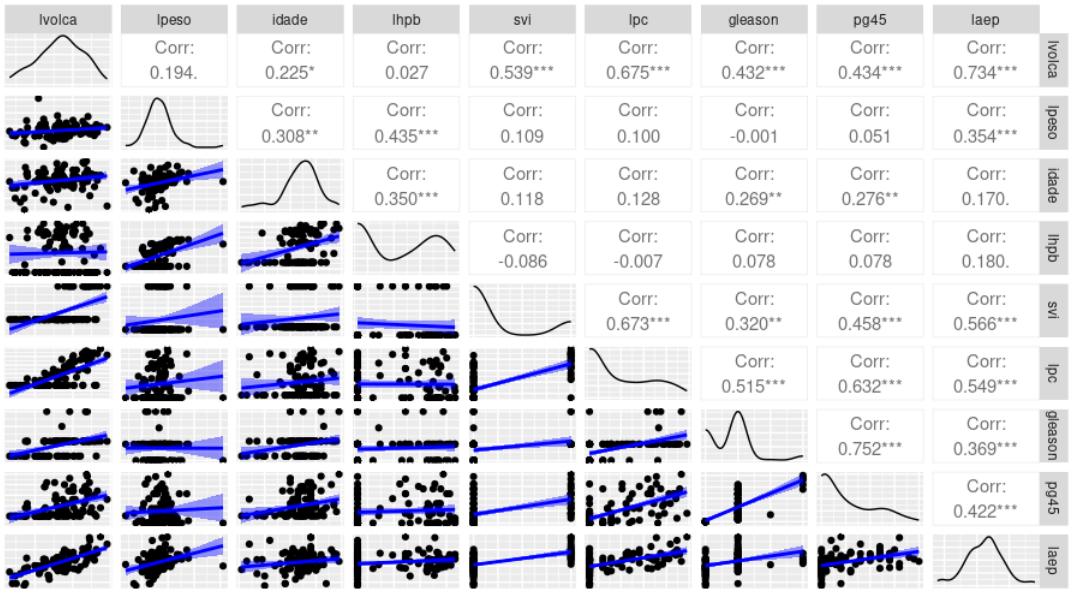
Já a variável *gleason* também possui média próxima a mediana, mas o seu limite inferior coincide com o primeiro quartil e a mediana coincide com o terceiro quartil, o que significa que há pouquíssima variabilidade entre o limite inferior e o primeiro quartil e entre a mediana e o terceiro quartil. Na variável *svi*, que representa a invasão na vesícula seminal, e assume os valores 0 (não apresenta) e 1 (apresenta), é possível notar que a maior parte dos homens no estudo não apresentou invasão na vesícula seminal, uma vez que a sua média é inferior a 0,5.

Figura 3.1: *Boxplots* das variáveis dos dados de homens com câncer de próstata



É possível observar, através da Figura 3.2, que a maioria das variáveis apresentam uma correlação positiva moderada com a variável resposta (*laep*). Os gráfico de dispersão, apontam que há relações de linearidade entre as variáveis explicativas e a variável resposta, sobretudo a relação com as variáveis *lvolca*, *lpc* e *pg45*.

Figura 3.2: Gráficos de dispersão, histogramas e correlações das variáveis dos dados de homens com câncer de próstata



É possível notar, ainda, que as variáveis explicativas são correlacionadas entre si. A variável *lvolca*, *svi*, *lpc* e *gleason* são as que mais estão correlacionadas.

3.2 Modelo de regressão linear múltiplo

A Tabela 3.1 apresenta a sumarização da regressão linear em que o logaritmo do antígeno específico da próstata é utilizado para predizer os dados dos homens com câncer de próstata. É possível notar que apenas duas covariáveis são significativas ao nível de 5%.

Tabela 3.1: Sumarização da regressão linear em que os dados dos homens com câncer de próstata são utilizados para predizer o logaritmo do antígeno específico da próstata

	Coeficiente	EP	Est. T	Valor-p
Intercepto	0,83	0,87	0,96	0,340
lvolca	0,63	0,09	7,04	**<0,001**
lpeso	0,50	0,18	2,83	**0,006**
idade	-0,02	0,01	-1,58	0,118
lhpb	0,08	0,06	1,31	0,192
lpc	0,01	0,09	0,16	0,872
pg45	0,01	0,004	1,68	0,096

A Tabela 3.2 apresenta a Tabela ANOVA do modelo, o valor-p do teste F obtido foi menor que o nível 0,05. Portanto, rejeita-se a hipótese nula de igualdade de todos os coeficientes angulares do modelo a zero, a um nível de significância de 5%. Ou seja, rejeita-se a hipótese de que o modelo não tem nenhuma validade para explicar a variável resposta.

Tabela 3.2: Tabela ANOVA

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	Estatística F	Valor-p
Regressão	6	78,81	13,13	23,87	**<0,001**
Resíduos	90	49,10	0,55		
Total	96	127,91			

O coeficiente de determinação indica que 62% da variabilidade do logaritmo do antígeno específico da próstata é explicada por esse modelo de regressão e o coeficiente de determinação ajustado foi de 0,59.

Algumas covariáveis nesse modelo podem ser consideradas desnecessárias. Então, faz-se necessário uma seleção de covariáveis, a fim de obter um modelo mais parcimonioso e que explique melhor a o logaritmo do antígeno específico da próstata.

3.3 Seleção de covariáveis

3.3.1 Eliminação *backward* baseada no teste F

Realizando a seleção de covariáveis através da eliminação *backward* baseada no teste F, pemaneceram no modelo final as covariáveis *lvolca,lpeso* e *pg45*. Sendo que apenas a *lvolca* e *lpeso* são significativas ao nível de 5%.

Tabela 3.3: Modelo de regressão selecionado através da eliminação *backward* baseada no teste F

	Coeficiente	EP	Est. T	Valor-p
Intercepto	-0,40	0,56	-0,71	0,48
lvolca	0,62	0,07	8,53	**<0,001**
lpeso	0,52	0,16	3,36	**0,001**
pg45	0,006	0,003	1,89	0,06

O coeficiente de determinação ajustado desse modelo foi 0,59, o AIC 222,95 e o BIC 235,82.

3.3.2 Seleção *stepwise* baseada no AIC

A seleção de covariáveis através do método *stepwise* baseada no AIC, resultou no mesmo conjunto de variáveis independentes selecionadas no modelo anterior, Figura 3.4.

Tabela 3.4: Modelo de regressão selecionado através do método *stepwise* baseada no AIC

	Coeficiente	EP	Est. T	Valor-p
Intercepto	-0,40	0,56	-0,71	0,48
lvolca	0,62	0,07	8,53	**<0,001**
lpeso	0,52	0,16	3,36	**0,001**
pg45	0,006	0,003	1,89	0,06

3.3.3 Seleção *stepwise* baseada no BIC

Já a seleção *stepwise* baseada no BIC foi mais criteriosa e eliminou a covariável *pd45*, que não foi significativa no modelo anterior ao nível de 5%.

O coeficiente de determinação ajustado desse modelo foi 0,58, o AIC 224,58 e o BIC 234,88.

Tabela 3.5: Modelo de regressão selecionado através do método *stepwise* baseada no BIC

	Coeficiente	EP	Est. T	Valor-p
Intercepto	-0,30	0,57	-0,53	0,596
lvolca	0,68	0,07	10,23	**<0,001**
lpeso	0,51	0,16	3,25	**0,002**

3.3.4 Todas as regressões possíveis

A Tabela 3.6 apresenta os cinco melhores subconjuntos de variáveis de acordo com o R^2 ajustado para todos os modelos possíveis com essas variáveis. É possível notar que as variáveis lvolca, lpeso e pg45, que foram as variáveis escolhidas pelo métodos *backwise* baseada no teste F e *stepwise* baseada no AIC, aparecem em todos os cinco melhores subconjuntos (até mesmo aparecendo apenas as três sozinhas).

Tabela 3.6: Best Subsets Regression

Preditores	R^2_{ajust}
lvolca lpeso idade lhpb pg45	0,5949299
lvolca lpeso idade pg45	0,5917192
lvolca lpeso idade lhpb lpc pg45	0,5905471
lvolca lpeso pg45	0,5883048
lvolca lpeso lhpb pg45	0,5876523

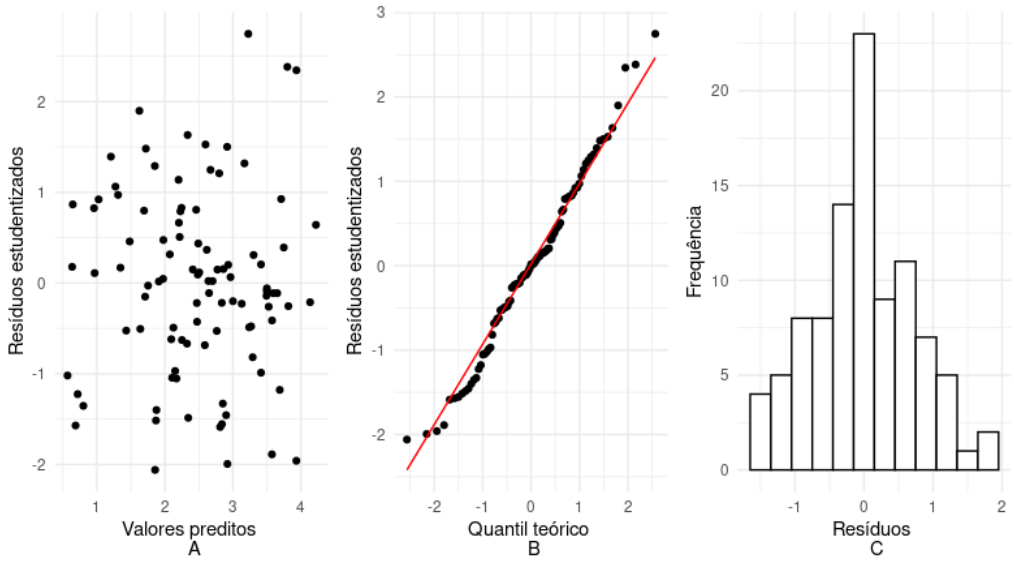
3.4 Análise dos resíduos

A análise de resíduos está sendo feita para o modelo de regressão linear com as variáveis lvolca, lpeso e pg45, pelo fato de terem selecionados por mais de um método de seleção de modelos aqui apresentado.

Então, observando a Figura 3.3 (A), é possível notar que os resíduos não parecem ter a mesma variabilidade ao longo dos valores preditos, ocorrendo de modo aleatório em torno de zero. O teste de *Goldefeld-Quandt* ,que avalia a se há variância constante dos resíduos, indicou um valor-p de 0,52, não rejeitando a hipótese nula de homoscedasticidade dos resíduos ao nível de significância de 5%. Portanto, é plausível dizer que os resíduos são homoscedásticos

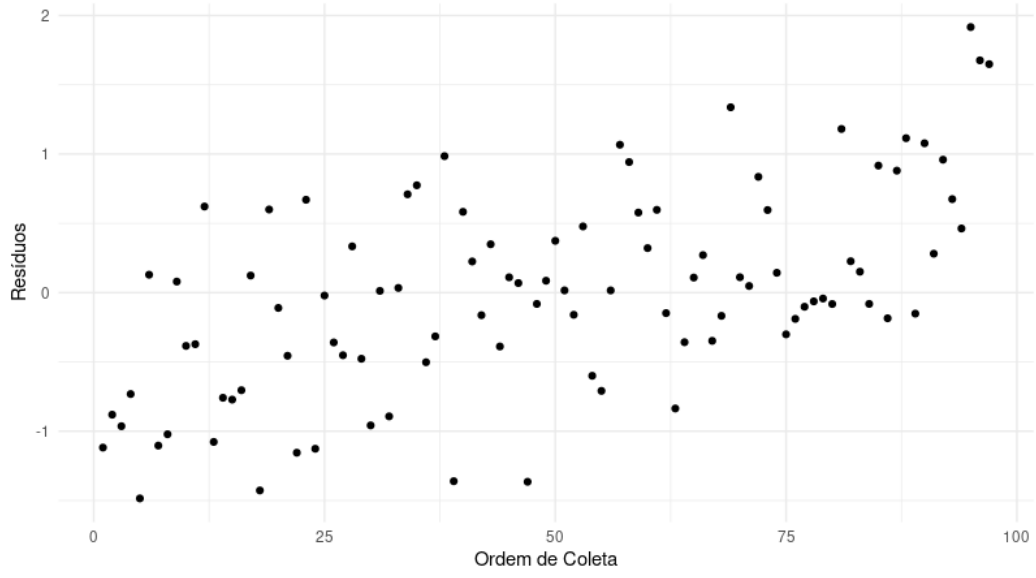
No qq-plot da Figura 3.3 (B), é observável que os quantis dos resíduos studentizados estão bem alinhados com os quantis teóricos da distribuição normal. Realizando o teste de *Shapiro-Wilk*, obteve-se um valor-p igual a 0,54, não rejeitando a hipótese nula de normalidade ao nível de significância de 5%. Logo, é possível dizer que os resíduos seguem uma distribuição normal

Figura 3.3: Gráfico dos resíduos versus valores ajustados, resíduos estudentizados versus quantil teórico e histograma dos resíduos da regressão linear em que os dados dos homens com câncer de próstata são utilizados para predizer o logaritmo do antígeno específico da próstata



Realizando o teste de *Durbin-Watson* para avaliar a independência dos resíduos, obteve-se um valor-p de 0,001, portanto, rejeitando a hipótese de não correlação serial dos resíduos ao nível de significância de 5%. Entretanto, isso,possivelmente,se deve ao fato de que as observações foram ordenadas crescentemente de acordo com a variável resposta, por isso é possível observar uma tendência positiva no gráfico de resíduos versus ordem de coletas na Figura 3.4. Portanto, não é possível dizer que os resíduos não são independentes

Figura 3.4: Gráfico dos resíduos versus ordem de coleta da regressão linear em que os dados dos homens com câncer de próstata são utilizados para predizer o logaritmo do antígeno específico da próstata



4 Conclusão

4.1 Atividade 1 - Regressão polinomial

Inicialmente, uma análise descritiva dos dados mostrou uma correlação positiva moderada entre a renda e o tempo. Em seguida, foram realizadas regressões lineares e polinomiais para modelar essa relação. Com base na análise realizada, pode-se concluir que o efeito da renda anual do marido no tempo entre o casamento e o nascimento do primeiro filho é não linear.

Os resultados das regressões indicaram que um modelo polinomial de terceira ordem apresentou um ajuste melhor aos dados em comparação com os modelos lineares. No entanto, foi identificada a presença de multicolinearidade no modelo cúbico que utilizou a variável renda como a variável resposta. Portanto, foi proposto realizar novamente os modelos polinomiais até a ordem 3 considerando a renda centrada na média como variável explicativa. Com essa variável, o modelo cúbico obteve menores valores de VIF do que o anterior, indicando que é menos afetado por colinearidade. Considerando essa questão, o modelo polinomial quadrático com a variável renda centrada na média foi considerado o mais adequado para prever o tempo até o primeiro filho. Uma análise dos resíduos mostrou que os pressupostos do modelo foram atendidos.

Portanto, com base nessas conclusões, pode-se dizer que a renda anual do marido tem um efeito significativo e não linear no tempo entre o casamento e o nascimento do primeiro filho, sendo que um aumento na renda está associado a um aumento no tempo até o primeiro filho, mas esse efeito diminui à medida que a renda aumenta.

4.2 Atividade 2 - Seleção de modelos

Na análise descritiva, verificou-se as variáveis que apresentaram distribuição simétrica, tais quais *lvolca*,*lpeso*,*idade* e *laep*. Assim como, as variáveis que estavam mais correlacionadas linearmente com a variável resposta *laep* (o logaritmo do antígeno específico da próstata), que foram as variáveis *lvolca*,*lpc* e *pg45*. Também foi observado que há variáveis

explicativas bem correlacionadas entre si, como é o caso entre variáveis *pg45* e *gleason*, por exemplo.

Foi estimado o modelo de regressão linear múltiplo completo (com todas as variáveis explicativas contínuas) e, apenas as variáveis *lvolca* e *lpeso* foram significativas ao nível de significância de 5%. O modelo completo foi capaz de explicar 62% da variabilidade da variável resposta.

Partindo para os métodos de seleção de modelo, o método *backwise* baseado no teste F escolheu as mesmas variáveis (o mesmo modelo) que o método de *stepwise* baseado no AIC, que foram as variáveis *lvolca*, *lpeso* e *pg45*. No modelo com essas três variáveis, apenas as variáveis *lvolca* e *pg45* foram significativas ao nível e 5%. Enquanto que o método *stepwise* baseado no BIC, escolheu apenas as variáveis *lvolca* e *lpeso* para o modelo ideal. Por fim, realizando todas as regressões possíveis e escolhendo os melhores modelos baseado no R^2 ajustado, foi possível notar que todos os cinco melhores modelos tiveram as três variáveis citadas anteriormente incluídas no modelo.

Portanto, o modelo com as variáveis *lvolca*, *lpeso* e *pg45* foi considerado o mais adequado. Realizando o teste de resíduos para esse modelo, verificou que os resíduos não violaram as suposições de homocedasticidade e normalidade, entretanto o teste de *Durbin-Watson* indicou que os resíduos não eram independentes. Entretanto, isso se deve ao fato de que as observações foram ordenadas de acordo com a variável resposta *laep*. Desse modo, não foi possível dizer que os resíduos violaram a independência.