AN INTRODUCTION TO STATISTICAL LEARNING

Respostas

Author
Bianca Portela
Brasil
2023

Contents

1 Statistical Learning **3**

2 Linear Regression **7**

1 Statistical Learning

1. Para cada uma das partes (a) a (d), indique se, geralmente, esperaríamos que o desempenho de um método flexível de aprendizado estatístico seja melhor ou pior do que um método inflexível. Justifique sua resposta.

(a) O tamanho da amostra n é extremamente grande, e o número de preditores p é pequeno.

Melhor. Esperamos que o desempenho de um método de aprendizado estatístico flexível seja melhor do que um método inflexível quando o tamanho da amostra n é extremamente grande e o número de preditores p é pequeno. Isso ocorre porque um modelo flexível pode aproveitar melhor os dados abundantes e encontrar padrões sutis nos poucos preditores disponíveis.

(b) O número de preditores p é extremamente grande, e o número de observações n é pequeno.

Pior. Modelo menos flexível performa melhor. Por n ser pequeno, a escolha de um modelo mais flexível fará com que o modelo se ajuste bem demais aos dados e eles sobreajustará os resultados da estimação.

(c) A relação entre os preditores e a resposta é altamente não linear.

Melhor. Modelo mais flexível é melhor. Um modelo menos flexível performará de maneira ruim nesse caso. Permitir maior flexibilidade ao modelo faz com que ele se ajuste melhor aos dados, já que ele possui mais graus de liberdade.

(d) A variância dos termos de erro, ou seja, $\sigma^2 = \mathbf{Var}(\epsilon)$, é extremamente alta.

Pior. Quando a variância dos termos de erro, ou seja, $\sigma^2 = \text{Var}(\epsilon)$, é extremamente alta, esperamos que o desempenho de um método de aprendizado estatístico flexível seja pior do que um método inflexível. Um modelo flexível pode se ajustar demais aos erros aleatórios, resultando em overfitting, enquanto um modelo mais inflexível pode ser mais robusto a essas flutuações.

2. Explique se cada cenário é um problema de classificação ou regressão e indique se estamos mais interessados em inferência ou previsão. Por fim, forneça n e p.

Previsão: queremos apenas uma previsão e $\hat{f}$ é uma caixa preta. **Inferência:** queremos entender a associação entre Y e $X_1, \ldots, X_p$. Desejamos estimar f, mas nosso objetivo final não é necessariamente fazer predições em Y. Não podemos tratar $\hat{f}$ como uma caixa preta, pois precisamos saber sua forma exata.

(a) Coletamos um conjunto de dados das 500 principais empresas nos EUA. Para cada empresa, registramos o lucro, o número de funcionários, a indústria e o salário do CEO. Estamos interessados em compreender quais fatores afetam o salário do CEO.

$n = 500$ principais empresas dos EUA

$p = 3$ (lucro, número de funcionários, indústria) Problema de regressão, inferência.

(b) Estamos considerando lançar um novo produto e desejamos saber se ele

será um sucesso ou um fracasso. Coletamos dados de 20 produtos similares lançados anteriormente. Para cada produto, registramos se foi um sucesso ou fracasso, o preço cobrado pelo produto, o orçamento de marketing, o preço da concorrência e outras dez variáveis.

$n = 20$ produtos

$p = 13$ (preço cobrado pelo produto, orçamento de marketing, preço da concorrência e outras dez variáveis) Problema de classificação, previsão.

(c) Estamos interessados em prever a variação percentual na taxa de câmbio USDEuro em relação às mudanças semanais nos mercados de ações mundiais. Portanto, coletamos dados semanais para todo o ano de 2012. Para cada semana, registramos a variação percentual no USDEuro, a variação percentual no mercado dos EUA, a variação percentual no mercado britânico e a variação percentual no mercado alemão.

$n =$ dados semanais durante o ano de 2012 da taxa de câmbio USDEuro (52 semanas)

$p = 3$ (variação percentual dos EUA, variação percentual no mercado britânico, variação percentual no mercado alemão)

3. Agora revisitamos a decomposição viés-variância.

(a) Forneça um esboço das curvas típicas de (viés)^2, variância, erro de treinamento, erro de teste e erro Bayesiano (ou erro irreduzível), em um único gráfico, à medida que avançamos de métodos de aprendizado estatístico menos flexíveis para abordagens mais flexíveis. O eixo-x deve representar o nível de flexibilidade no método, e o eixo-y deve representar os valores para cada curva. Deve haver cinco curvas. Certifique-se de rotular cada uma delas.

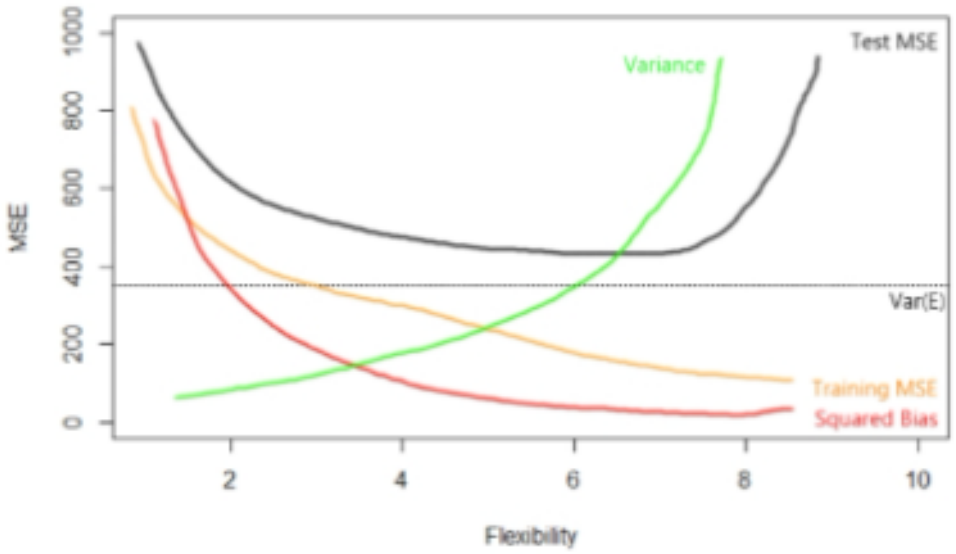

Figure 1: Legenda da imagem.

(b) Explique por que cada uma das cinco curvas tem a forma exibida na parte (a).

As curvas representadas no gráfico mostram o efeito da flexibilidade do modelo nos valores de viés, variância, erro de treinamento, erro de teste e erro Bayesiano.

A curva verde representa a variância, que aumenta à medida que a flexibilidade do modelo aumenta. Isso ocorre porque modelos mais flexíveis tendem a se ajustar mais aos dados de treinamento, resultando em maior variabilidade nas previsões.

A curva vermelha representa o viés, que diminui à medida que aumenta a flexibilidade do modelo. O viés representa o erro introduzido ao simplificar um problema da vida real com um modelo mais simples. Modelos mais flexíveis, com mais graus de liberdade, têm vieses menores, pois se ajustam melhor aos dados de treinamento.

A curva preta representa o erro de teste (MSE), que aumenta à medida que a variância e a flexibilidade do modelo aumentam. Isso ocorre porque modelos excessivamente flexíveis começam a se ajustar ao ruído dos dados de treinamento e não generalizam bem para novos dados.

A curva amarela representa o erro de treinamento, que diminui com o aumento da flexibilidade, apesar do aumento da variância. Isso acontece porque modelos muito flexíveis tendem a se sobreajustar aos dados de treinamento, resultando em um desempenho aparentemente melhor nesse conjunto.

O erro Bayesiano (ou erro irreduzível) é representado por uma curva constante, pois representa o limite mínimo teórico de erro que nenhum modelo pode ultrapassar, independentemente da flexibilidade.

As curvas se comportam dessa maneira devido ao trade-off entre viés e variância. À medida que aumentamos a flexibilidade do modelo, o viés diminui mais rapidamente do que a variância aumenta inicialmente, resultando em uma diminuição do erro esperado. No entanto, em algum ponto, aumentar a flexibilidade tem pouco impacto no viés, mas começa a aumentar significativamente a variância. Quando isso acontece, o erro de teste começa a aumentar. Portanto, encontrar um modelo com baixa variância e baixo viés é um desafio, pois ao diminuir um, o outro tende a aumentar.

4. Agora, pense em algumas aplicações da vida real para aprendizado estatístico.

(a) Descreva três aplicações da vida real em que a classificação pode ser útil. Descreva a variável de resposta, bem como os preditores. O objetivo de cada aplicação é inferência ou previsão? Explique sua resposta.

1. Prever se a pessoa desenvolverá diabetes ou não. O Y é se a pessoa tem ou não diabetes. Os Xs podem ser características da pessoa e hábitos de vida. É uma aplicação para previsão.

2. O consumidor irá cancelar o serviço ou não. Previsão. A variável resposta é se o serviço foi ou não cancelado. O X pode ser variáveis de comportamento de compra e características individuais.

3. Classificar imagens (é uma maçã ou não é uma maçã). Previsão. Y é se é ou não é

uma maçã. X pode ser características da maçã (vermelha, redonda, peso, etc.)

(b) Descreva três aplicações da vida real em que a regressão pode ser útil. Descreva a variável de resposta, bem como os preditores. O objetivo de cada aplicação é inferência ou previsão? Explique sua resposta.

1. Prever preço de alguma coisa (ex. casa). Previsão. Y será o preço da casa, X as características da casa.

2. Relação entre educação e salário. Inferência. Y é o salário, X educação e outras características pessoais.

3. Previsão de demanda de produtos: A variável de resposta é a quantidade de vendas, e os preditores podem incluir dados históricos de vendas, preços, promoções e fatores sazonais. O objetivo é a previsão para otimizar os estoques e produção.

(c) Descreva três aplicações da vida real em que a análise de agrupamento (cluster analysis) pode ser útil.

1. Segmentação de clientes. Inferência. Agrupar clientes com base em seus comportamentos de compra, preferências e características demográficas para melhorar estratégias de marketing direcionadas.

2. Análise de mercado. Inferência. Agrupar produtos ou serviços semelhantes com base em características e demandas para entender a concorrência e identificar oportunidades.

3. Detecção de padrões de comportamento em tráfego de rede. Inferência. Agrupar fluxos de dados para identificar atividades incomuns ou maliciosas na rede e melhorar a segurança cibernética.

5. Quais são as vantagens e desvantagens de uma abordagem muito flexível (em comparação com uma abordagem menos flexível) para regressão ou classificação? Em que circunstâncias uma abordagem mais flexível pode ser preferida em relação a uma abordagem menos flexível? Quando uma abordagem menos flexível pode ser preferida?

Vantagens: Modelos mais flexíveis se ajustam melhor aos dados, identificando padrões em relações não lineares e reduzindo o viés.

Desvantagem: Porém, existe o risco de sobreajuste. Isso acontece quando o modelo se esforça demais para encontrar padrões nos dados de treinamento, podendo detectar padrões causados apenas por acaso, em vez de serem representativos da realidade. Isso resulta em um alto erro ao testar o modelo em novos dados. Em casos de sobreajuste, um modelo menos flexível poderia ter obtido melhores resultados de previsão. Além disso, modelos muito flexíveis podem ser difíceis de interpretar.

Quando os dados têm comportamento não linear, uma abordagem mais flexível é preferível. Por outro lado, em situações de sobreajuste ou quando o objetivo é inferência (compreender as relações entre os preditores e a resposta), uma abordagem menos flexível é a mais indicada.

7. Descreva as diferenças entre uma abordagem de aprendizado estatístico paramétrica e não paramétrica. Quais são as vantagens de uma abordagem paramétrica para regressão ou classificação (em oposição a uma abordagem não paramétrica)? Quais são suas desvantagens?

Aprendizado Paramétrico: Nesse tipo de aprendizado, fazemos uma hipótese específica sobre a forma do modelo, onde os parâmetros são estimados a partir dos dados. Isso reduz a estimação da função desconhecida f a um conjunto de parâmetros.

Vantagens: - Modelos paramétricos são mais simples e fáceis de interpretar, pois têm uma estrutura definida com parâmetros específicos.

Desvantagem: - Não sabemos a verdadeira forma da função f. Se o modelo escolhido estiver distante da verdadeira f, a estimação pode ser imprecisa. Tentar resolver esse problema com modelos mais flexíveis pode levar ao overfitting, pois eles requerem a estimação de mais parâmetros, resultando em modelos mais complexos e menos generalizáveis.

Aprendizado Não Paramétrico: Nesse tipo de aprendizado, não fazemos hipóteses explícitas sobre a forma funcional da função f.

Vantagens: - Modelos não paramétricos podem se ajustar a uma ampla variedade de funções f, permitindo a captura de padrões complexos e não lineares. Isso resulta em previsões mais precisas e maior flexibilidade para lidar com dados complexos.

Desvantagem: - A abordagem não paramétrica não reduz o problema de estimação da função f a um número limitado de parâmetros. Essa abordagem requer um grande número de observações para estimar f de maneira precisa. Portanto, modelos não paramétricos podem exigir mais dados do que os modelos paramétricos para obter resultados confiáveis.

2 Linear Regression

1. Descreva as hipóteses nulas às quais os valores de p fornecidos na Tabela 3.4 correspondem. Explique quais conclusões podem ser tiradas com base nesses valores de p. Sua explicação deve ser expressa em termos de vendas, TV, rádio e jornal, em vez dos coeficientes do modelo linear.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

Figure 2: Tabela 3.4

A tabela 3.4 mostra os coeficientes estimados de regressão múltipla quando o orçamento de TV, rádio e jornais são usados para prever vendas utilizando dados de Avertising.

No caso de uma regressão múltipla temos o modelo como

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Onde: - Y é a variável dependente. - $X_1, X_2, \ldots, X_p$ são as variáveis independentes. - $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes de regressão correspondentes. - ε é o termo de erro.

A hipótese nula (H_0) em um contexto de regressão múltipla afirmaria que $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. Em outras palavras, cada coeficiente de regressão individual é zero, o que implicaria que a variável correspondente não tem impacto no valor da variável dependente.

O teste de hipótese associado é geralmente realizado calculando o valor-p (*p-value*). Se o valor-p resultante for menor do que um certo nível de significância (por exemplo, 0.05), você pode rejeitar a hipótese nula e concluir que pelo menos uma das variáveis independentes tem um efeito significativo na variável dependente.

Assim temos que tanto TV quanto radio tem efeito significativo na variável dependente (Advertising), enquanto newspaper não tem impacto no valor da variável dependente. Assim, aumentar o orçamento gasto em TV e radio aumenta vendas, enquanto newspaper não causa efeito nehum.

2. Explique detalhadamente as diferenças entre os métodos classificadores KNN (*K-Nearest Neighbors*) e os métodos de regressão KNN.

O modelo KNN é um método não paramétrico que oferece mais flexibilidade em comparação com as técnicas paramétricas, adequando-se bem tanto a problemas de classificação quanto de regressão.

No método de regressão KNN, dado um valor K e um ponto de previsão x_0, o processo envolve as seguintes etapas:

1. Identificação dos K vizinhos mais próximos de x_0, representados por N. 2. Estimativa da resposta $f(x_0)$ utilizando a média das respostas das observações de treinamento em N.

A escolha ideal de K depende do trade-off entre viés e variância. Pequenos valores de K proporcionam um ajuste mais flexível, resultando em baixo viés mas alta variância. Isso ocorre porque a previsão em uma determinada região depende principalmente de uma única observação. Por outro lado, valores maiores de K geram um ajuste mais suave e menos variável. Nesse caso, a previsão em uma região é a média de várias observações, diminuindo o impacto de uma única observação atípica. No entanto, a suavização excessiva pode mascarar a estrutura subjacente em $f(X)$, causando um viés.

Já no contexto da classificação com KNN, o método é um pouco diferente. Aqui, as etapas são as seguintes:

1. Identificação dos K vizinhos mais próximos de x_0, representados por N. 2. Estimativa da probabilidade condicional da classe j usando a fração de pontos em N que pertencem à classe j:

$$Pr(Y = j | x = x_0) = \frac{1}{K} \sum_{i \in N} I(u_i = j)$$

O classificador KNN atribui a observação de teste x_0 à classe com a maior probabilidade

calculada acima.

De maneira similar à regressão KNN, a escolha apropriada de K é crucial. Um valor baixo de K pode levar a um ajuste excessivamente sensível ao ruído, enquanto um valor alto de K pode suavizar demais as fronteiras de decisão, levando a um viés elevado.

Em resumo, embora ambos os métodos, regressão KNN e classificação KNN, compartilhem a abordagem geral de encontrar os K vizinhos mais próximos, a maneira como eles fazem as estimativas e as considerações sobre o valor de K são distintas devido à natureza dos problemas de regressão e classificação.

3. Suponha que temos um conjunto de dados com cinco preditores: X1 = GPA, X2 = IQ, X3 = Nível (1 para Faculdade e 0 para Ensino Médio), X4 = Interação entre GPA e IQ e X5 = Interação entre GPA e Nível. A variável de resposta é o salário inicial após a graduação (em milhares de dólares). Suponha que usamos mínimos quadrados para ajustar o modelo e obtemos os seguintes coeficientes estimados: $\hat{\beta}_0 = 50$**,** $\hat{\beta}_1 = 20$**,** $\hat{\beta}_2 = 0.07$**,** $\hat{\beta}_3 = 35$**,** $\hat{\beta}_4 = 0.01$ **e** $\hat{\beta}_5 = -10$**.**

(a) Qual das seguintes afirmações está correta e por quê?

i. Para um valor fixo de IQ e GPA, os graduados do ensino médio ganham, em média, mais do que os graduados universitários.

ii. Para um valor fixo de IQ e GPA, os graduados universitários ganham, em média, mais do que os graduados do ensino médio.

iii. Para um valor fixo de QI e GPA, os graduados do ensino médio ganham, em média, mais do que os graduados universitários, desde que o GPA seja suficientemente alto.

iv. Para um valor fixo de QI e GPA, os graduados universitários ganham, em média, mais do que os graduados do ensino médio, desde que o GPA seja suficientemente alto.

A afirmação ii está correta. A afirmação ii diz que os graduados universitários ganham mais do que os do ensino médio para um valor fixo de IQ e GPA. Isso corresponde ao valor positivo do coeficiente estimado de "Nível" ($\hat{\beta}_3$) **(b) Preveja o salário de um graduado universitário com QI de 110 e GPA de 4.0.**

$$salario = 50 + (20 * 4) + (0.07 * 110) + (35 * 1) + 0.01(4 * 110)) - 10 * (4 * 1)$$

$= 137.1$

(c) Verdadeiro ou falso: Como o coeficiente para o termo de interação GPA/QI é muito pequeno, há pouca evidência de um efeito de interação. Justifique sua resposta.

Falso. O coeficiente pequeno para o termo de interação GPA/QI não implica necessariamente que não há um efeito de interação. Para afirmar com certeza que não existe efeito de interação entre os termos GPA/QI é necessário avaliar a significância estatística, que leva em consideração o erro padrão e o valor-p associado ao coeficiente de interação.

4. Eu coletei um conjunto de dados (n = 100 observações) contendo um único preditor e uma resposta quantitativa. Em seguida, ajustei um modelo de regressão linear aos dados, bem como uma regressão cúbica separada, ou seja, $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$**.**

(a) Suponha que a verdadeira relação entre X e Y seja linear, ou seja, $Y = \beta_0 + \beta_1 X + \epsilon$**. Considere a soma dos resíduos de treinamento (RSS) para a regressão linear e também o RSS de treinamento para a regressão cúbica. Esperamos que um seja menor que o outro, que sejam iguais, ou não há informações suficientes para dizer? Justifique sua resposta.**

RSS é o resíduo da soma dos quadrados. É definido como

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

em que $e_i = y_i - \hat{y}_i$. A OLS escolhe $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizem o RSS. Se a verdadeira relação é linear, não faz sentido adicionar um termo cúbico à regressão. Isso apenas fará com que o modelo tenha problemas com overfitting. Ainda assim, o RSS de treino da regressão cúbica terá um valor baixo, pois o modelo está se ajustando demais aos dados, podendo ser inclusive menor do que o da RSS da regressão linear.

(b) Responda (a) utilizando o RSS de teste em vez do RSS de treinamento. A verdadeira função da regressão é uma função linear. O modelo cúbico teve um bom desempenho nos dados de treino pois está se ajustando aos dados, mas tem pouca capacidade de generalização. Assim, quando é inserido novos dados ao modelo através do dataset de teste, o RSS aumenta - evidenciando o problema de sobreajuste. O modelo de regressão linear é menos flexível,mas nesse caso tem melhor ajuste nos dados, apresentando um RSS menor.

(c) Suponha que a verdadeira relação entre X e Y não seja linear, mas não sabemos quão distante ela está da linearidade. Considere o RSS de treinamento para a regressão linear e também o RSS de treinamento para a regressão cúbica. Esperamos que um seja menor que o outro, que sejam iguais, ou não há informações suficientes para dizer? Justifique sua resposta. Caso a verdadeira relação entre X e Y não seja linear, a adição de um polinômio apresentará um aumento de perfomance no RSS de treinamento. Assim o RSS da regressão cúbica será menor que o da regressão linear.

(d) Responda (c) utilizando o RSS de teste em vez do RSS de treinamento.

Não há informações suficientes para responder completamente a isso.

Se a verdadeira relação for altamente não linear e houver baixo ruído (ou erro irreduzível) em nossos dados de treinamento, podemos esperar que o modelo cúbico mais flexível forneça um melhor RSS de teste.

No entanto, se a relação for apenas ligeiramente não linear ou o ruído em nossos dados de treinamento for alto, um modelo linear poderá oferecer melhores resultados.

5. Considere os valores ajustados que resultam da realização de uma regressão linear o intercepto. Nesse cenário, o i-ésimo valor ajustado assume a forma

$$\hat{y}_i = x_i\hat{\beta},$$

onde

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_i'^2}.$$

Mostre que podemos escrever

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}.$$

O que é $a_{i'}$?

3.4

(1) $\hat{y}_i = x_i \hat{\beta}$

(2) $\hat{\beta} = \left(\sum x_i y_i \right) \Big/ \sum_{i'=1}^{n} x_{i'}^2$

(3) $\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$

A. Substituindo (2) → (1)

$$\hat{y}_i = x_i \left(\frac{\sum x_i y_i}{\sum x_i^2} \right)$$

x_i é uma constante, inserindo ela dentro do somatório temos

$$\hat{y}_i = \frac{\sum x_i x_{i'} y_i}{\sum x_{i''}^2}$$

reorganizando os termos temos que

$$\hat{y}_i = a_i y_i$$

em que $a_i = \dfrac{\sum x_i x_{i'}}{\sum x_{i''}^2}$

6. Usando a equação (3.4), argumente que no caso da regressão linear simples, a linha dos mínimos quadrados sempre passa pelo ponto $(\bar{x}, \bar{y})$.

A equação (3.4) é

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

e

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

3.6 ① $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

② $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

a reta de regressão é

③ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

prove que quando $x = \bar{x} \Rightarrow \hat{y} = \bar{y}$

substituindo ② ⇒ ③

$\hat{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x$

quando $x = \bar{x}$ temos

$\hat{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}$

$\hat{y} = \bar{y}$

(Quando $x = \bar{x}$, o valor previsto $\hat{y} = \bar{y}$. Ou seja, a linha de regressão passa pelo ponto médio $(\bar{x}, \bar{y})$

7. É afirmado no texto que no caso da regressão linear simples de Y em relação a X, a estatística R^2 (3.17) é igual ao quadrado da correlação entre X e Y (3.18). Prove que isso é verdade. Para simplificar, você pode assumir que $\bar{x} = \bar{y} = 0$.

References