MQAM

1.1 - What is Simple Linear Regression

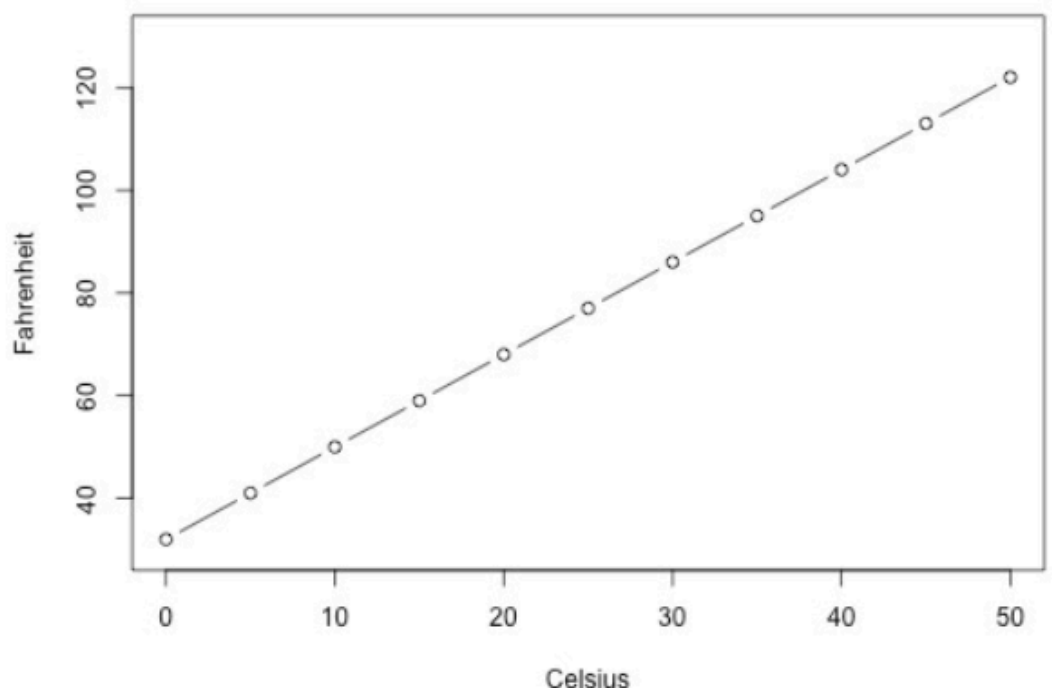
A **Regressão Linear Simples** é um método estatístico usado para resumir e estudar a relação entre duas variáveis quantitativas (numéricas). Uma variável é chamada de **preditora** (`x`) e a outra de **resposta** (`y`).

Tipos de Relações

Relação Determinística (ou Funcional)

Neste tipo de relação, o valor de uma variável pode ser determinado **exatamente** a partir do valor da outra, sem nenhum erro. A relação é perfeita e geralmente descrita por uma fórmula matemática ou lei física.

- **Característica Principal:** Previsão perfeita.
- **Exemplo:** A conversão de temperatura de Celsius para Fahrenheit. Se você sabe a temperatura em Celsius, sabe *exatamente* qual será em Fahrenheit.
- **Gráfico:** Todos os pontos de dados caem perfeitamente sobre uma linha.

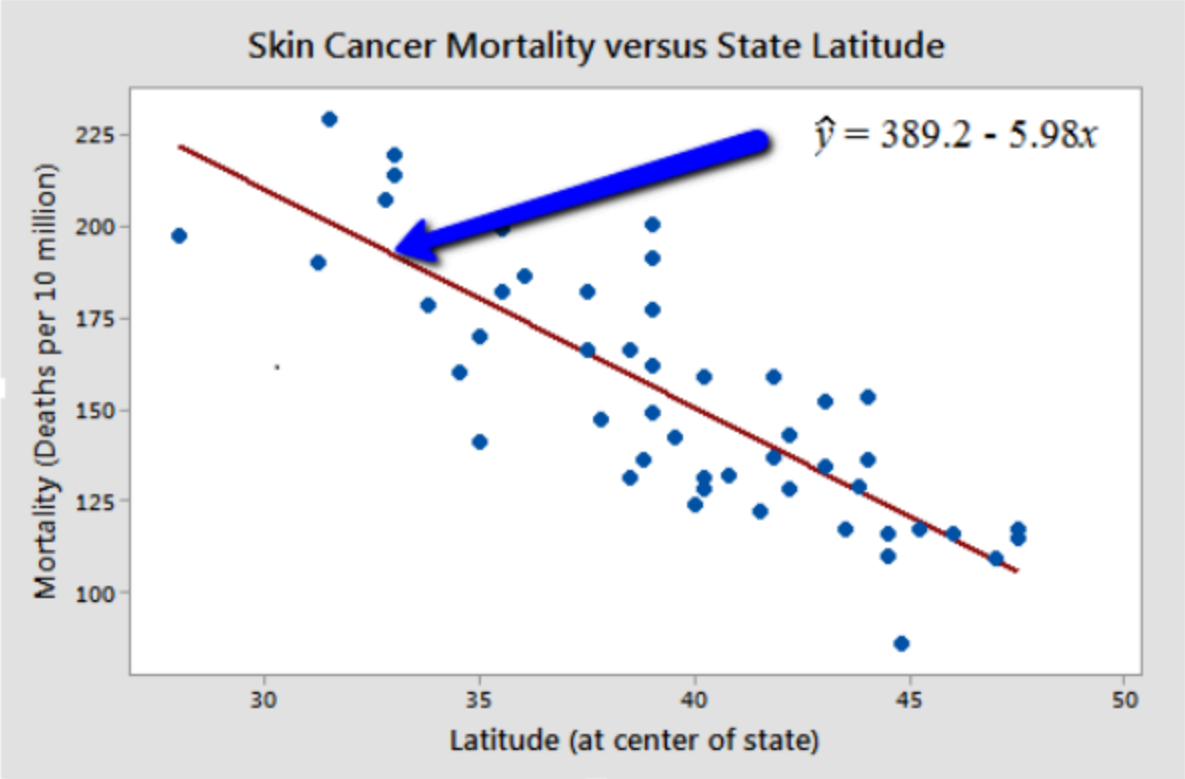


Relação Estatística

Esta é a relação que a regressão linear estuda. Nela, as variáveis estão relacionadas, mas a relação **não é perfeita**. Existe uma **tendência** geral, mas também uma **dispersão** (ou

"scatter") dos dados em torno dessa tendência.

- **Característica Principal:** Previsão imperfeita, mas com uma tendência identificável.
- **Exemplo:** A taxa de mortalidade por câncer de pele e a latitude de um estado. Geralmente, estados com maior latitude têm menor mortalidade, mas não é possível prever a taxa de mortalidade *exatamente* apenas com a latitude.
- **Gráfico:** Os pontos formam uma "nuvem" que segue uma tendência geral, mas não estão perfeitamente alinhados. Em resumo, a regressão linear é a ferramenta que nos ajuda a encontrar e descrever a linha de tendência que melhor se ajusta a uma relação estatística imperfeita.



1.2 - What is The Best Fitting Line

O Critério dos Mínimos Quadrados (Least Squares Criterion)

Para encontrar a **única linha** que melhor descreve a tendência nos dados, usamos um método matemático chamado "critério dos mínimos quadrados".

- **Erro de Previsão (Resíduo):** Para cada ponto de dado, existe uma distância vertical entre o valor real (y) e o valor que a linha prevê ($\hat{y}$). Essa distância é o erro de previsão. * **A Lógica:** O método eleva ao quadrado cada um desses erros (para torná-los todos positivos e penalizar mais os erros grandes) e depois os soma.

- **A "Melhor" Linha:** A **linha de regressão de mínimos quadrados** é aquela que torna a **soma dos erros quadrados (SSE)** a menor possível. De todas as infinitas retas que poderiam ser desenhadas, esta é a que passa "mais perto" de todos os pontos de dados simultaneamente, de acordo com esse critério.

As Fórmulas e a Interpretação dos Coeficientes

A linha de regressão é definida pela equação $\hat{y} = b_0 + b_1 x$, onde b_0 e b_1 são calculados para satisfazer o critério dos mínimos quadrados.

A Inclinação (Slope), b_1

- **O que é:** A inclinação b_1 nos diz o quanto esperamos que a variável resposta (y) **mude**, em média, para **cada aumento de uma unidade** na variável preditora (x).
- **Exemplo (Altura e Peso):** Se $b_1 = 6.14$, significa que para cada polegada (inch) a mais na altura, esperamos que o peso aumente, em média, **6.14 libras (pounds)**.
- **Sinal da Inclinação:**
  - $b_1 > 0$ **(Positiva):** A linha sobe. À medida que x aumenta, y tende a aumentar.
  - $b_1 < 0$ **(Negativa):** A linha desce. À medida que x aumenta, y tende a diminuir.

O Intercepto (Intercept), b_0

- **O que é:** O intercepto b_0 é o valor previsto da variável resposta (y) quando a variável preditora (x) é **igual a zero**.
- **Cuidado com a Interpretação:** Muitas vezes, o intercepto **não tem um significado prático**. No exemplo de altura e peso, o intercepto seria o peso previsto para uma pessoa com altura 0, o que não faz sentido. Isso acontece porque $x=0$ está muito fora da faixa de dados observados (um fenômeno chamado **extrapolação**). O intercepto é, na maioria das vezes, apenas um ponto de partida matemático para a linha.

1.3 - The Simple Linear Regression Model

População vs. Amostra

A Linha de Regressão da População

- **Conceito:** Esta é a linha **verdadeira e desconhecida** que descreve a relação média entre as variáveis x e y para **toda a população** de interesse.
- **Equação:** $\mu_y = \beta_0 + \beta_1 x$
- **Parâmetros:** β_0 (beta-zero) e β_1 (beta-um) são os parâmetros **verdadeiros** do intercepto e da inclinação da população. Nós nunca os conhecemos exatamente, a menos que tenhamos dados de toda a população.

A Linha de Regressão da Amostra

- **Conceito:** Como quase nunca temos acesso à população inteira, pegamos uma **amostra** e usamos o critério dos mínimos quadrados para encontrar a melhor linha de ajuste para *aquela amostra específica*.
- **Equação:** $\hat{y} = b_0 + b_1x$
- **Estimativas (Estatísticas):** b_0 e b_1 são as **estimativas** que calculamos a partir dos nossos dados da amostra. Elas são nossa melhor aproximação para os verdadeiros parâmetros da população, β_0 e β_1.
- **Ponto Chave:** Se pegássemos uma amostra diferente, obteríamos valores ligeiramente diferentes para b_0 e b_1, mas ambas seriam estimativas da mesma linha de população verdadeira.

As 4 Suposições do Modelo de Regressão (LINE)

Para que nossas inferências sobre a população (usando os dados da amostra) sejam válidas, o modelo de regressão linear simples assume quatro condições sobre os **termos de erro (ε)**, que são as distâncias de cada ponto até a linha de regressão da **população**. As suposições podem ser lembradas pelo acrônimo **LINE**:

- **L - Linearidade (Linearity):** A relação entre a média da variável resposta y e a variável preditora x é, de fato, **linear**.
- **I - Independência (Independence):** Os erros são **independentes** uns dos outros. O erro de uma observação não fornece nenhuma informação sobre o erro de outra.
- **N - Normalidade (Normality):** Para cada valor de x, os erros são distribuídos segundo uma **curva normal** (distribuição normal) com média zero.
- **E - Igualdade de Variâncias (Equal Variances):** A **dispersão** (ou variância) dos erros é a **mesma** para todos os valores de x. Isso também é chamado de **homocedasticidade**.

What is The Common Variance

O que é a Variância do Erro?

A variância do erro (σ^2) é uma medida de quão **dispersos** os pontos de dados (y) estão em torno da **verdadeira linha de regressão da população**.

- **σ^2 pequeno:** Significa que os pontos de dados estão muito **próximos** da linha de regressão. A relação entre x e y é forte e as previsões do modelo serão mais **precisas e confiáveis**.
- **σ^2 grande:** Significa que os pontos de dados estão muito **espalhados** ao redor da linha. A relação é mais fraca e as previsões do modelo serão **menos precisas**. O exemplo dos

termômetros ilustra isso: a marca B é melhor porque seus dados têm menos dispersão (menor σ2), o que leva a previsões mais consistentes.

Como Estimamos a Variância do Erro?

Como σ^2 é um parâmetro da população (e, portanto, desconhecido), nós o estimamos usando os dados da nossa amostra. A estatística que usamos para estimar σ^2 é chamada de **Erro Quadrático Médio (Mean Square Error - MSE)**.

- **Fórmula:** $MSE = \Sigma(y_i - \hat{y}_i)^2 / (n - 2)$
- **Interpretação da Fórmula:**
  - **Numerador:** É a Soma dos Quadrados dos Erros (SSE), que é a soma de todas as distâncias verticais ao quadrado entre os pontos de dados observados (y_i) e os valores previstos pela linha de regressão da *amostra* ($\hat{y}_i$).
  - **Denominador ($n - 2$):** Dividimos por $n - 2$ em vez de n (como em uma média normal) porque "perdemos" dois graus de liberdade ao usar os dados para estimar os dois parâmetros da linha: o intercepto (β_0) e a inclinação (β_1).
- **Como encontrar no software:** Em uma saída de análise estatística (como a do Minitab), o **MSE** é o valor encontrado na linha "Error" (ou Residual) e na coluna "MS" (Mean Square) da tabela de Análise de Variância (ANOVA).
  A raiz quadrada do MSE é o **desvio padrão do erro (S)**, que nos dá uma ideia do tamanho típico de um erro de previsão, nas mesmas unidades da variável resposta.

1.5 - The Coefficient of Determination, R^2

O que é o R²?

O **R²** é uma medida que nos diz **qual proporção da variação total** na variável resposta (y) é "explicada" pela variável preditora (x) através do modelo de regressão. Em termos simples, ele quantifica o quão bem a linha de regressão se ajusta aos dados.
O valor do R² está sempre entre 0 e 1 (ou 0% e 100%).

A Lógica por Trás do R²: Somas de Quadrados

Para entender o R², precisamos dividir a variação total dos dados em duas partes:

1. **SSTO (Soma Total dos Quadrados - Total Sum of Squares):** Mede a **variação total** dos pontos de dados (y) em torno da média ($\bar{y}$). É a nossa "bagunça" total antes de criarmos o modelo.
2. **SSE (Soma dos Quadrados dos Erros - Error Sum of Squares):** Mede a variação que **sobra**, ou seja, a variação "não explicada" pelos dados em torno da linha de regressão. É a variabilidade aleatória que o modelo não conseguiu capturar.

3. **SSR (Soma dos Quadrados da Regressão - Regression Sum of Squares):** Mede a variação que é **explicada** pelo nosso modelo de regressão.
   A relação fundamental é: **SSTO = SSR + SSE** (Variação Total = Variação Explicada + Variação Não Explicada)
   O R² é simplesmente a razão da variação explicada pela variação total:
   `R² = SSR / SSTO`

Como Interpretar o R²

A interpretação é geralmente feita como uma porcentagem.

- **Exemplo:** Se **R² = 0.68** (ou 68%), isso significa que **68% da variabilidade** na variável resposta (`y`) pode ser explicada ou contabilizada pela relação linear com a variável preditora (`x`). Os outros 32% são devidos a outros fatores ou variabilidade aleatória.
- **R² = 1 (100%):** Um ajuste perfeito. Todos os pontos de dados caem exatamente sobre a linha de regressão.
- **R² = 0 (0%):** O modelo não explica nada. A variável preditora (`x`) não tem poder algum para explicar a variação em `y` (a linha de regressão seria horizontal).

Cuidado Importante: Correlação não é Causalidade!

Um R² alto indica que o modelo se ajusta bem e que há uma forte associação entre as variáveis. No entanto, isso **NÃO significa que `x` CAUSA a mudança em `y`**. A palavra "explicada" em estatística não implica uma relação de causa e efeito.

O que é o Coeficiente de Correlação (r)?

O coeficiente de correlação `r` é uma medida que quantifica a **força** e a **direção** da **relação linear** entre duas variáveis quantitativas. Seu valor está sempre no intervalo entre -1 e +1.

Relação com o R²

O `r` está diretamente ligado ao R² (Coeficiente de Determinação):

- **Cálculo:** `r = ± √R²`
- **Sinal:** O sinal de `r` (positivo ou negativo) é **sempre o mesmo** que o sinal da inclinação da linha de regressão (b_1).
  - Se a linha de regressão sobe (inclinação positiva), `r` é positivo.
  - Se a linha de regressão desce (inclinação negativa), `r` é negativo.

Como Interpretar o Valor de `r`

A interpretação de `r` se baseia em sua proximidade com -1, 0 ou +1.

- `r = 1`: Indica uma relação linear **positiva perfeita**. Todos os pontos de dados caem exatamente em uma linha reta que sobe.
- `r = −1`: Indica uma relação linear **negativa perfeita**. Todos os pontos de dados caem exatamente em uma linha reta que desce.
- `r = 0`: Indica que **não há relação linear** entre as variáveis.
  Valores intermediários indicam o seguinte:
- **Quanto mais próximo `r` estiver de 1 ou -1**, mais **forte** é a relação linear.
- **Quanto mais próximo `r` estiver de 0**, mais **fraca** é a relação linear.
  Exemplo: Uma correlação de `r = −0.825` (como no caso da mortalidade por câncer e latitude) nos diz duas coisas:

1. A relação é **negativa** (à medida que a latitude aumenta, a mortalidade tende a diminuir).
2. A relação é bastante **forte**, pois -0.825 está próximo de -1.

Características Importantes de `r`

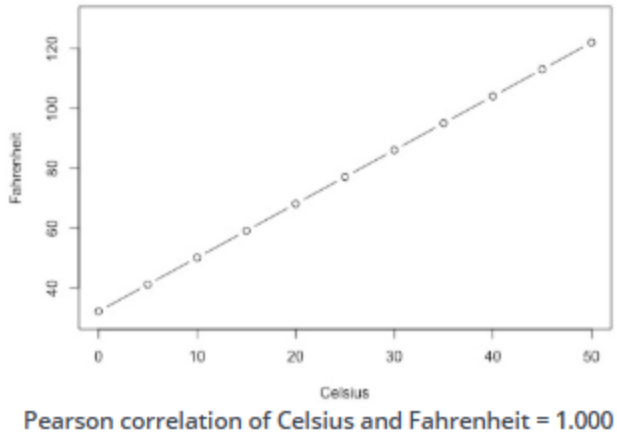
- **Não tem unidade de medida (Unitless):** Como não tem unidades, o `r` pode ser usado para comparar a força da relação linear entre diferentes pares de variáveis, mesmo que sejam medidas em escalas diferentes (ex: quilos vs. metros).
- **Simétrico:** A correlação entre a variável `x` e `y` é a mesma que a correlação entre `y` e `x`

1.7 - Some Examples

Veremos quatro exemplos práticos para ilustrar como interpretar o coeficiente de correlação (`r`) e o coeficiente de determinação (R^2).

Exemplo 1: Temperatura em Celsius e Fahrenheit

- **Relação:** Linear positiva perfeita (determinística).
- **Valores:** `r = 1.0` e `R² = 100%`.

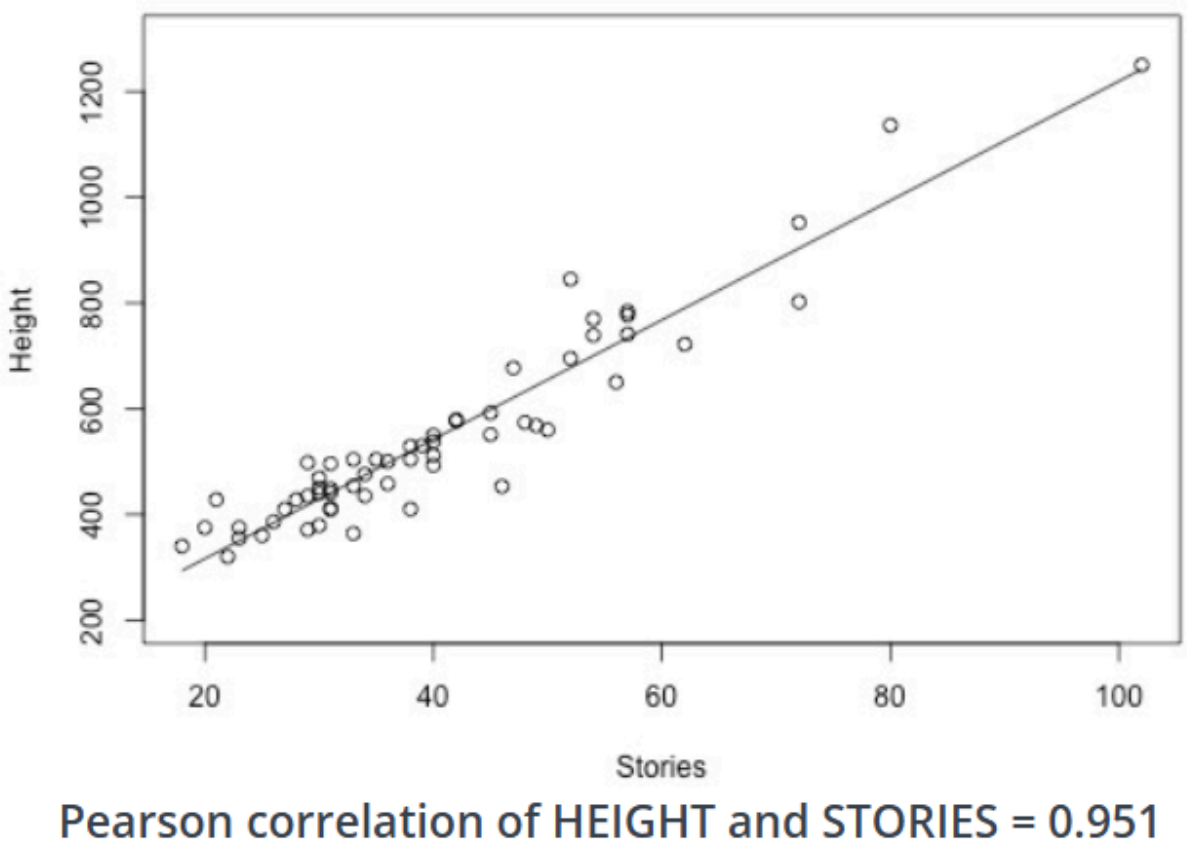


Pearson correlation of Celsius and Fahrenheit = 1.000

- **Interpretação:** O valor `r = 1` confirma uma relação linear positiva perfeita. O `R² = 100%` indica que 100% da variação na temperatura em Fahrenheit é perfeitamente explicada pela temperatura em Celsius.

Exemplo 2: Andares e Altura de Prédios

- **Relação:** Linear positiva forte.
- **Valores:** `r = 0.951` e `R² = 90.4%`.

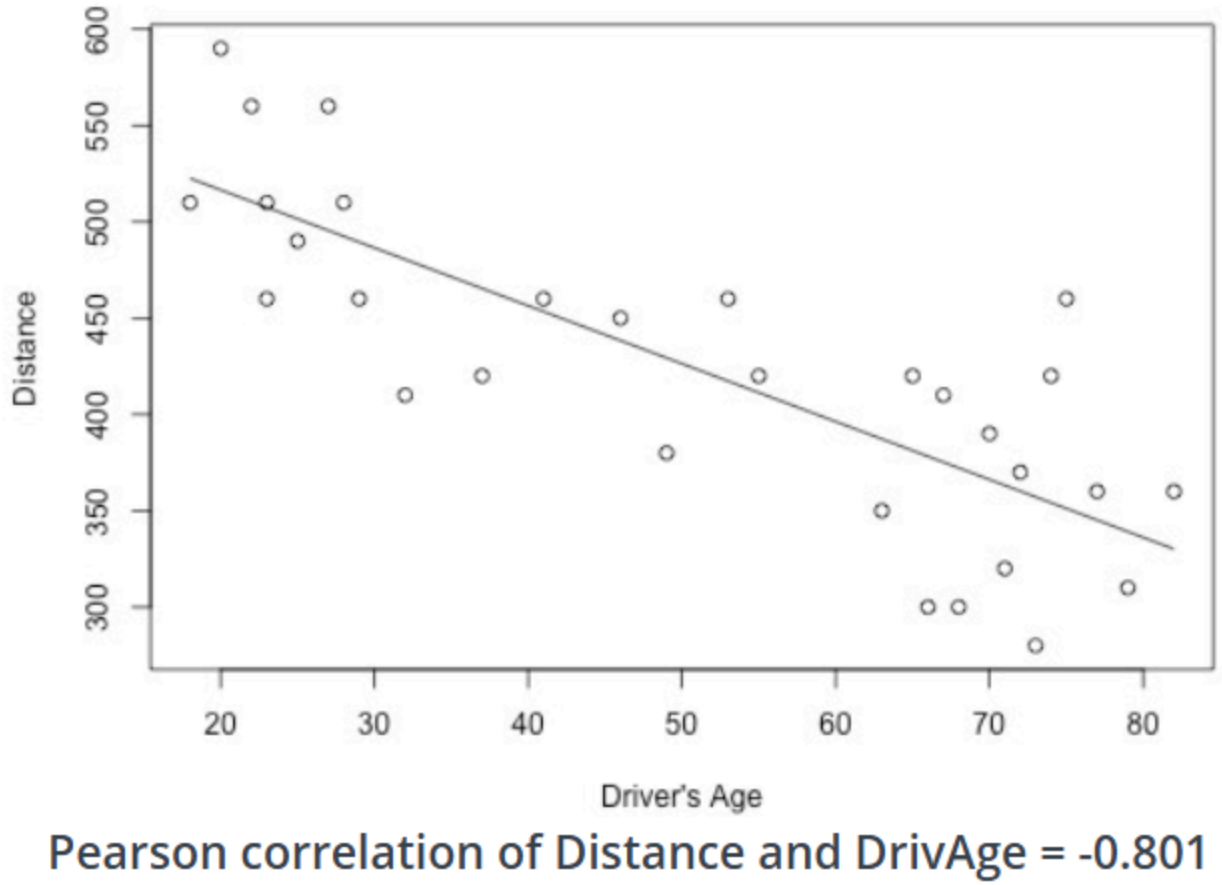


- **Interpretação:** O `r` positivo e próximo de 1 indica uma relação linear muito forte: à medida que o número de andares aumenta, a altura do prédio também aumenta. O `R²` mostra que 90.4% da variação na altura dos prédios é explicada pelo número de andares.

Exemplo 3: Idade e Distância de Visão dos Motoristas

- **Relação:** Linear negativa moderadamente forte.

- **Valores:** `r = −0.801` e `R² = 64.2%`.

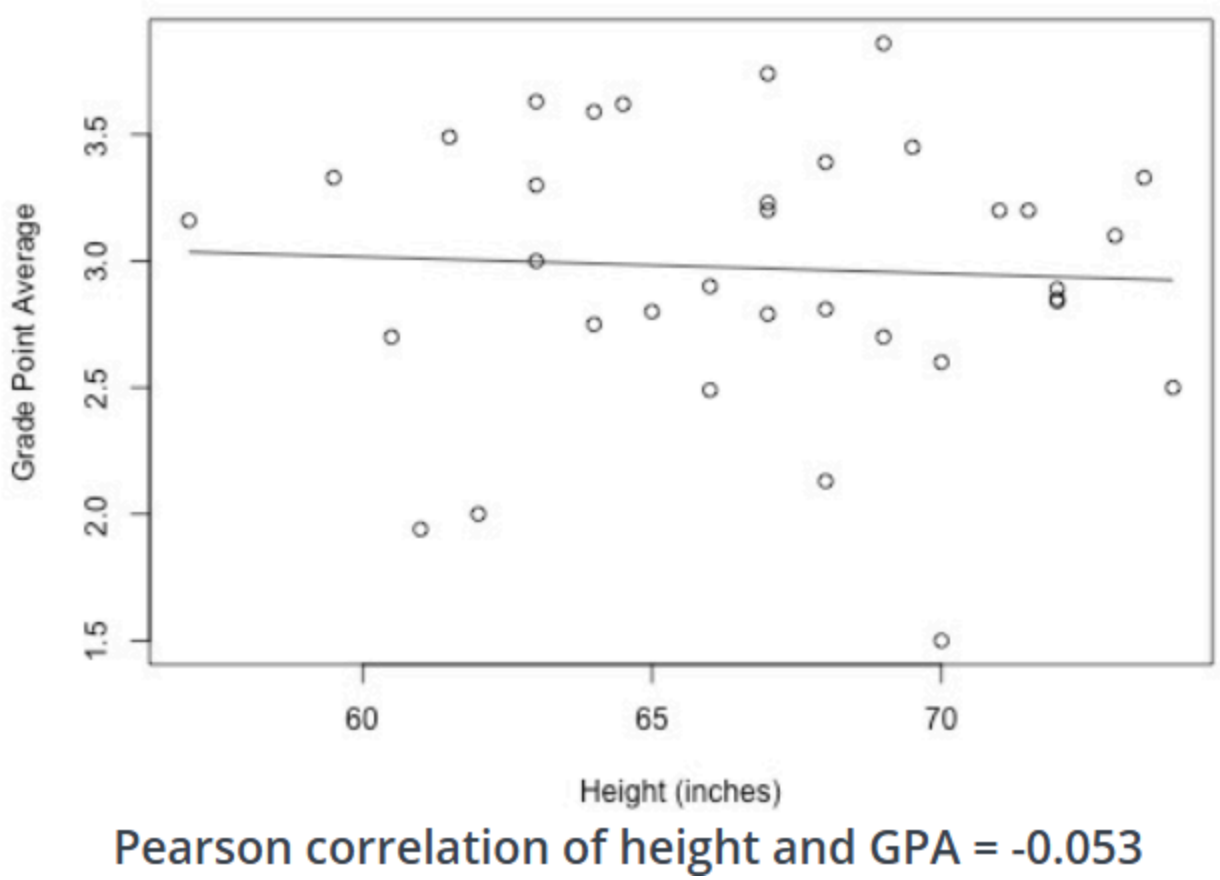


- **Interpretação:** O `r` negativo indica que, à medida que a idade do motorista aumenta, a distância que ele consegue enxergar tende a diminuir. O valor, por ser razoavelmente próximo de -1, sugere que a relação é moderadamente forte. O `R²` indica que 64.2% da variação na distância de visão é explicada pela idade do motorista.

Exemplo 4: Altura e GPA (Média de Notas)

- **Relação:** Ausência de relação linear.

- **Valores:** `r = −0.053` e `R² = 0.3%`.



- **Interpretação:** O valor de `r` é extremamente próximo de 0, o que sugere que não há praticamente nenhuma relação linear entre a altura de um estudante e sua média de notas. O `R²` confirma isso, mostrando que apenas 0.3% da variação na média de notas pode ser explicada pela altura, um valor insignificante.

1.8 - R^2 Cautions

Alguns **cuidados e erros comuns** na interpretação do coeficiente de determinação (R²) e do coeficiente de correlação (`r`).

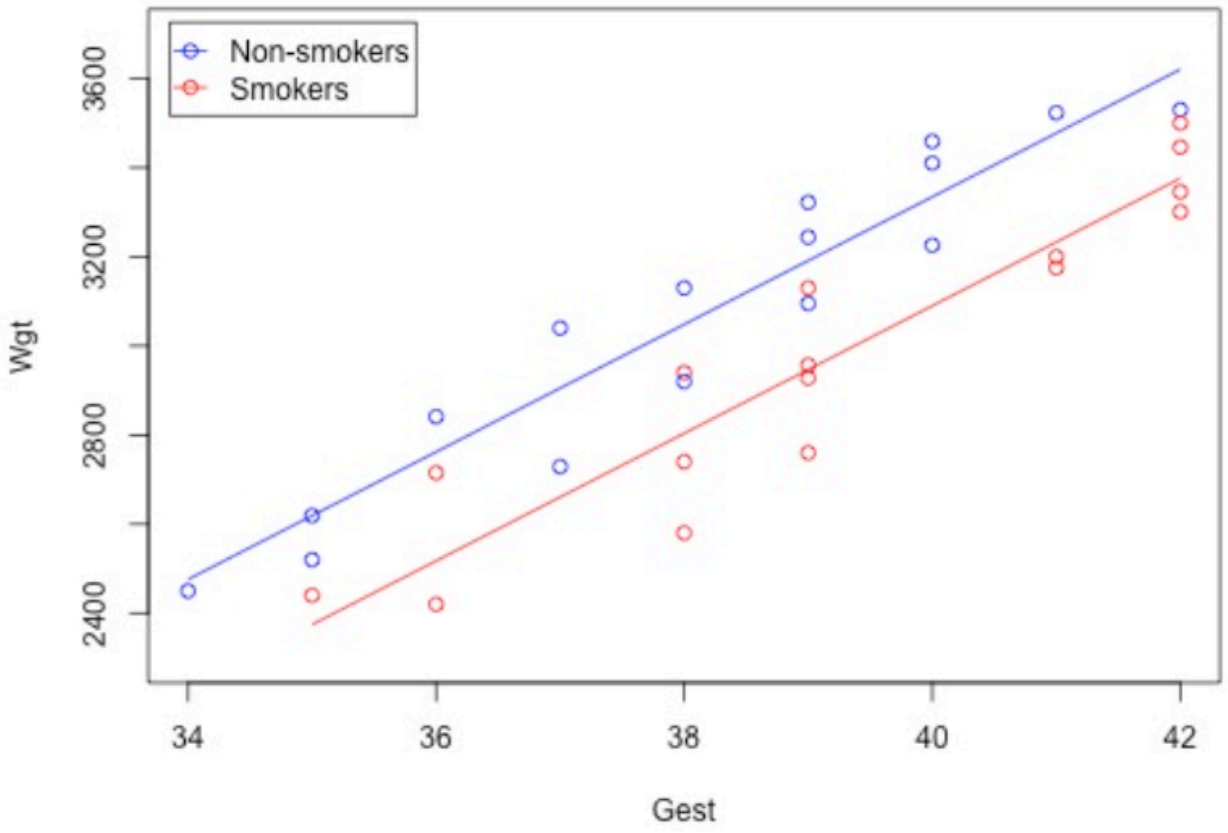
1. `r` **e R² Medem Apenas Relações Lineares**: É possível existir uma relação curva (curvilínea) perfeita entre duas variáveis, mas o `r` e o R² serem iguais a zero. Essas métricas falham em detectar relações que não são lineares.
2. **R² Alto Não Garante um Bom Ajuste do Modelo**: Um valor de R² elevado não significa, por si só, que a linha de regressão é o melhor modelo para os dados. Pode haver um padrão curvo nos dados que seria melhor descrito por outro tipo de função. É essencial sempre analisar o gráfico de dispersão.
3. **Sensibilidade a Pontos de Dados Atípicos (Outliers)**: Os valores de `r` e R² podem ser drasticamente afetados por um único ou poucos pontos de dados influentes. A remoção de

um único ponto pode mudar o sinal da correlação e alterar drasticamente o valor de R^2, como mostrado no exemplo dos terremotos.

4. **Correlação Não Implica Causalidade**: Uma forte associação entre duas variáveis (um `r` ou R^2 alto) não prova que uma variável **causa** a outra. Fatores ocultos (variáveis de confusão) podem estar influenciando ambas. Apenas experimentos controlados, e não estudos observacionais, podem estabelecer uma relação de causa e efeito.
5. **Cuidado com Correlações Ecológicas**: Correlações baseadas em dados agregados (médias ou taxas de grupos, como países ou regiões) tendem a **superestimar** a força da associação em comparação com correlações baseadas em dados individuais.
6. **Significância Estatística Não Implica Significância Prática**: Com conjuntos de dados muito grandes, é possível obter um resultado "estatisticamente significante" (um valor-p baixo) mesmo quando a relação é muito fraca (um R^2 muito baixo) para ter qualquer importância ou utilidade prática.
7. **R^2 Alto Não Garante Previsões Úteis**: Mesmo com um R^2 alto, se a variabilidade geral dos dados for muito grande, os intervalos de confiança e de predição gerados pelo modelo podem ser tão amplos que se tornam inúteis para fins práticos de previsão.

8.1 - Example on Birth Weight and Smoking

O exemplo analisa a relação entre o **peso de bebês ao nascer**, o **tempo de gestação** e o **hábito de fumar da mãe**. Utiliza-se um modelo de regressão múltipla com uma variável quantitativa (gestação) e uma variável qualitativa binária (fumar: sim/não).



Pontos Principais:

1. **Modelo de Regressão:** O modelo gera duas retas paralelas. Uma para mães não fumantes e outra para mães fumantes. A distância vertical entre as retas representa a diferença média no peso dos bebês devido ao fumo, mantendo o tempo de gestação constante.
2. **Equação de Regressão:** `Peso = −2390 + 143.10 * Gestação − 244.5 * Fumo`
   Logo, se a mãe fuma (Fumo = 1) -> `Peso = −2635 + 143 * Gest`
   Se não, (Fumo = 0) -> `Peso = −2390 + 143 * Gest`
3. **Interpretação:**
   - Para um mesmo tempo de gestação, o modelo estima que bebês de **mães fumantes pesam, em média, 244.5 gramas a menos** que bebês de mães não fumantes.
   - Tanto o tempo de gestação quanto o hábito de fumar são preditores estatisticamente significativos para o peso do bebê (valores-p < 0.001).
4. **Conclusão:** Após ajustar pelo tempo de gestação, há uma diferença significativa no peso médio de bebês nascidos de mães fumantes e não fumantes. O fumo durante a gravidez está associado a um menor peso ao nascer.

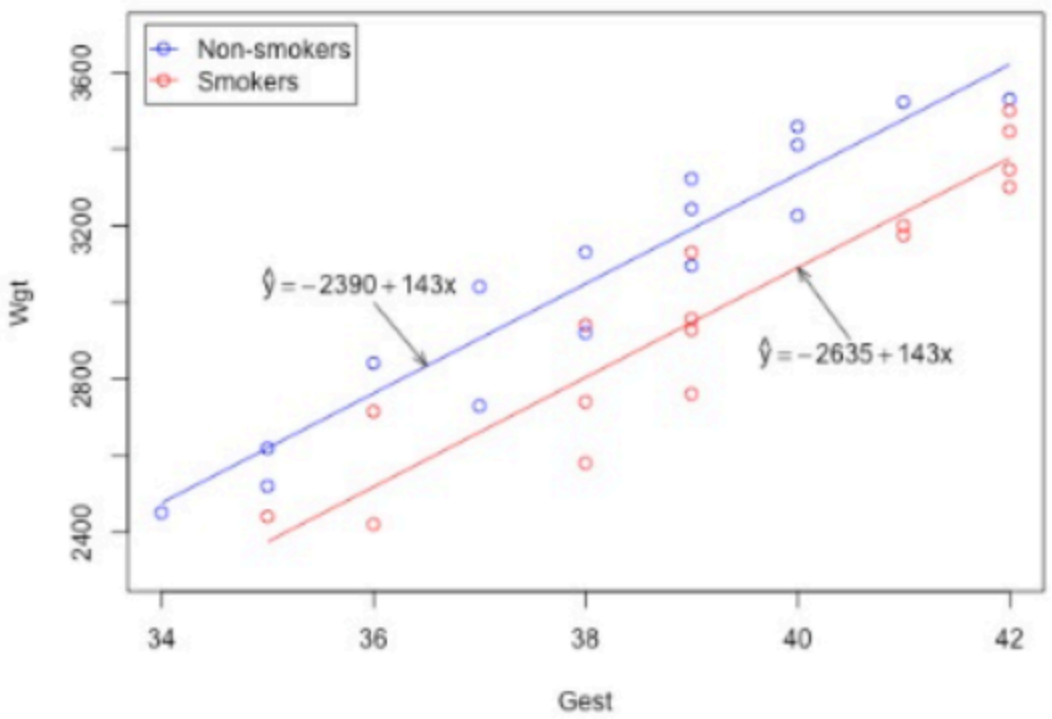
8.2 - The Basics

O conteúdo explica como usar uma **regressão linear múltipla** para analisar o efeito de uma variável categórica (Fumante/Não Fumante) e uma contínua (Tempo de Gestação) no peso de bebês.

Pontos Chave:

1. **Variável Binária:** A variável qualitativa "Fumar" é codificada como uma variável binária (ou "indicadora"): `1` se a mãe fumou, `0` se não.
2. **Modelo com Retas Paralelas:** A inclusão dessa variável no modelo de regressão gera duas equações de reta que são paralelas:
   - **Não Fumantes (código 0):** Média do Peso = (β0) + β1× Gestação
   - **Fumantes (código 1):** Média do Peso = (β0+β2) + β1× Gestação
3. **Interpretação dos Coeficientes:**

   - **β1**: Representa a inclinação, ou seja, o aumento médio no peso do bebê para cada semana adicional de gestação. Este efeito é o **mesmo** para ambos os grupos.
   - **β2**: É o ponto mais importante. Representa a **diferença média no peso** entre bebês de mães fumantes e não fumantes, para qualquer tempo de gestação fixo.



4. **Conclusão da Análise:**
   - O teste de hipótese para β2 verifica se essa diferença é estatisticamente significante.
   - No exemplo, o coeficiente para "Fumo" (`Smoke`) foi **-244.5**, significando que, mantendo a gestação constante, bebês de mães fumantes pesam, em média, 244.5 gramas a menos.

- O p-value baixo (< 0.001) confirma que essa diferença é estatisticamente significante.

8.3 - Two Separate Advantages

O texto demonstra as **vantagens de usar um único modelo de regressão múltipla** (com a variável binária "Fumo") em vez de criar **dois modelos separados**, um para fumantes e outro para não fumantes.

Vantagens do Modelo Único ("Pooled"):

1. **Maior Precisão:** Ao usar todos os 32 pontos de dados de uma só vez, o erro padrão dos coeficientes (como o da "Gestação") é menor. Isso resulta em **intervalos de confiança mais estreitos** e estimativas mais precisas em comparação com os modelos separados, que usam amostras menores (apenas 16 pontos cada).
2. **Eficiência para Responder à Pergunta de Pesquisa:** O modelo único responde diretamente à pergunta principal — "existe diferença entre os grupos?" — através de um único teste de hipótese para o coeficiente da variável binária (no caso, `Smoke`). Se esse coeficiente for estatisticamente significante, conclui-se que há uma diferença.
   Em resumo, **agrupar os dados em um único modelo é mais poderoso e eficiente** porque utiliza toda a informação disponível para obter estimativas mais precisas e permite testar a diferença entre os grupos de forma direta.

8.4 - Coding Qualitative Variables

O conteúdo foca em diferentes maneiras de **codificar variáveis qualitativas** para uso em modelos de regressão e como a escolha do método de codificação afeta a interpretação dos coeficientes do modelo.

1. O Problema de Usar Variáveis Demais (Multicolinearidade)

- **Regra geral:** Para uma variável qualitativa com `c` **categorias** (ou níveis), você deve criar `c − 1` **variáveis indicadoras**.
- **Exemplo:** Para a variável "Fumo", que tem 2 categorias (fumante, não fumante), usamos apenas `2 − 1 = 1` variável indicadora.
- **Por que isso é importante?** Se você criar uma variável indicadora para cada categoria (ex: uma para "fumante" e outra para "não fumante"), elas serão perfeitamente correlacionadas. Isso causa um problema chamado **multicolinearidade perfeita**, e o software estatístico não conseguirá estimar o modelo, sendo forçado a remover uma das variáveis.

2. Comparação de Esquemas de Codificação

O texto compara dois esquemas de codificação para a variável "Fumo":

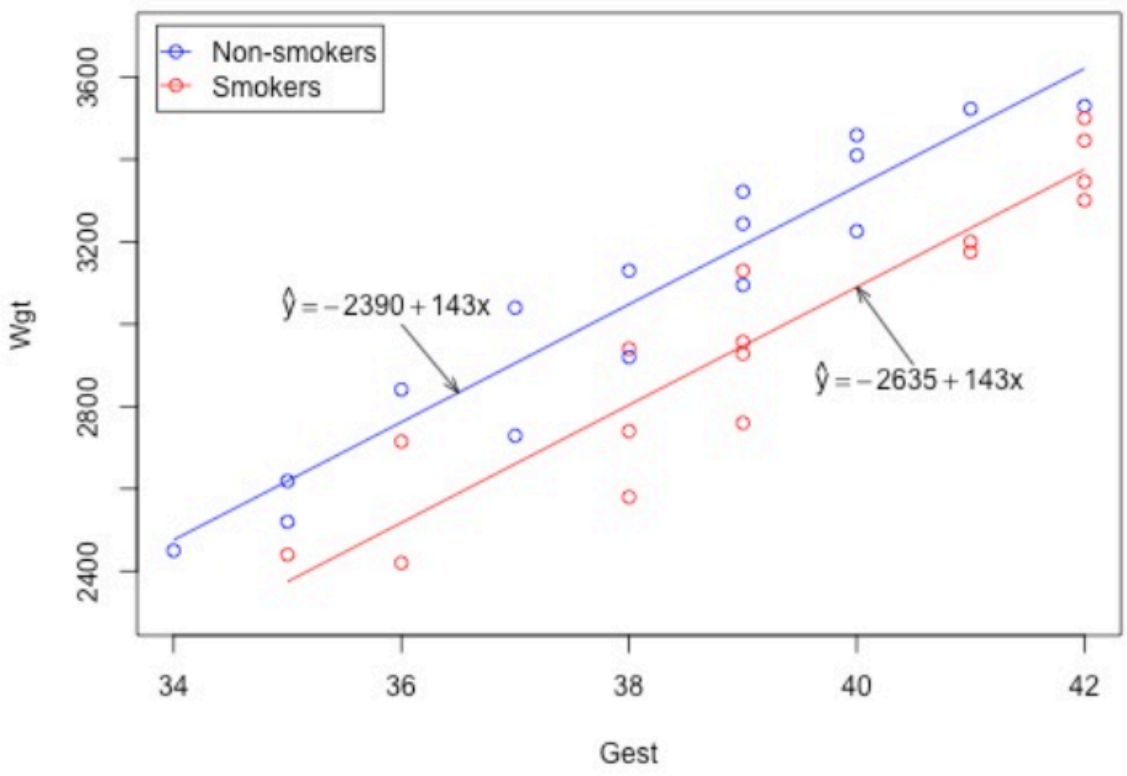
A) Codificação Padrão (0, 1)

- Não Fumante = 0 (*grupo de referência*)
- Fumante = 1
- **Interpretação dos Coeficientes:**
  - **β0 (Intercepto):** É o intercepto **específico do grupo de referência**. No exemplo, é o peso estimado de um bebê de uma mãe não fumante com 0 semanas de gestação.
  - **β1 (Gestação):** É o efeito da gestação no peso (a inclinação), que é o mesmo para ambos os grupos.
  - **β2 (Fumo):** Representa a **diferença** entre o intercepto do grupo "1" (fumantes) e o grupo de referência "0" (não fumantes). No exemplo, o valor de **-245g** é a diferença direta no peso.

B) Codificação Alternativa (-1, 1)

- Não Fumante = -1
- Fumante = 1
- **Interpretação dos Coeficientes:**
  - **β0 (Intercepto):** Representa o **intercepto médio geral**, ou seja, a média dos interceptos dos dois grupos. Não representa mais um grupo específico.
  - **β1 (Gestação):** Sua interpretação **não muda**. Continua sendo a inclinação da reta.
  - **β2 (Fumo):** Representa o **"desvio" (offset)** de cada grupo em relação ao intercepto médio. A diferença total entre os grupos é o **dobro** deste coeficiente (2 x β2). No exemplo, β2 é -122, então a diferença total é 2 x 122 = 244g, que é essencialmente o

mesmo resultado (-245g) da codificação (0, 1).

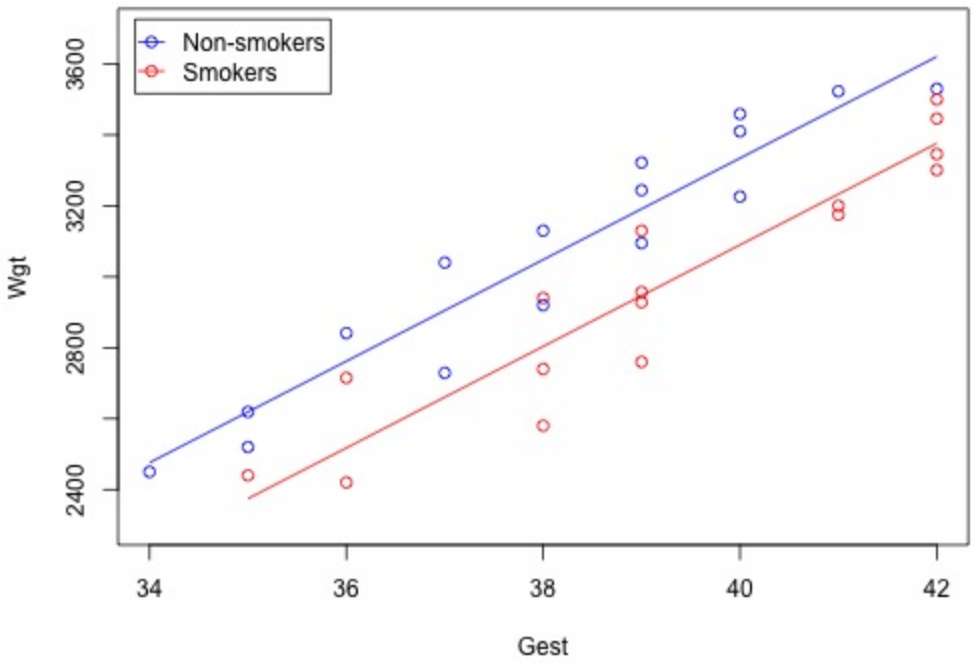


Conclusão Principal

Independentemente do esquema de codificação escolhido (0,1 ou -1,1), o resultado final e as conclusões científicas são os mesmos. As retas de regressão ajustadas para cada grupo serão idênticas. O que muda é apenas o significado matemático e a interpretação dos coeficientes β0 e β2. Portanto, é crucial saber qual codificação está sendo usada para interpretar os resultados de uma regressão corretamente.

8.5 - Additive Effects

1. **Modelo Aditivo:** Um modelo é considerado aditivo quando o efeito de uma variável preditora no resultado **não depende** do valor de outra variável preditora. Elas não interagem entre si; seus efeitos simplesmente "se somam".
2. **Aplicação no Exemplo:**
   - O efeito do **tempo de gestação** no peso do bebê (um aumento de 143g por semana) é o **mesmo**, independentemente de a mãe ser fumante ou não.
   - O efeito do **fumo** no peso do bebê (uma redução de 245g) é o **mesmo**, não importa qual seja o tempo de gestação.
3. **Representação Gráfica:** A característica visual de um modelo aditivo com uma variável categórica (como o fumo) é que as **retas de regressão para cada grupo são paralelas**. O paralelismo das retas é a evidência gráfica da falta de interação entre os preditores.

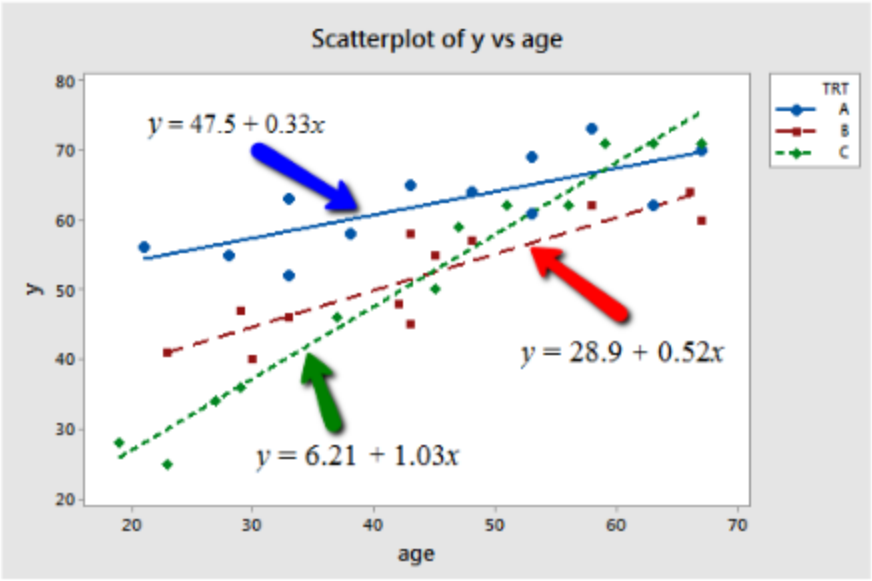


Em resumo, o modelo `Peso = −2390 + 143 Gest − 245 Fumo` é um **modelo aditivo** porque o impacto da gestação e o impacto do fumo são independentes e simplesmente se somam para prever o peso final.

8.6 - Interaction Effects

O que é um Efeito de Interação?

- Acontece quando o efeito de uma variável preditora (ex: Idade) sobre a variável resposta (ex: Eficácia do tratamento) **depende do nível ou valor de outra variável preditora** (ex: Qual tratamento foi usado - A, B ou C).
- Em outras palavras, os preditores não atuam de forma independente; o efeito de um modifica o efeito do outro.



- **Visualmente**, um modelo com interação resulta em retas de regressão que **não são paralelas**. Cada grupo tem sua própria inclinação.

O Exemplo: Tratamentos para Depressão

- **Objetivo:** Analisar a eficácia de três tratamentos diferentes para depressão, levando em conta a idade do paciente.
- **Variáveis:**
  - **Resposta (y):** Medida da eficácia do tratamento.
  - **Preditor Quantitativo (x1):** Idade do paciente.
  - **Preditor Qualitativo:** Tratamento (A, B ou C), codificado com duas variáveis indicadoras (x2 e x3), usando o Tratamento A como grupo de referência.

Construindo o Modelo com Interação

O modelo de regressão é formulado para incluir não apenas os "efeitos principais" (idade e tratamento), mas também os "termos de interação":
$y = \beta_0 + \beta_1 \cdot idade + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot (idade \cdot x_2) + \beta_5 \cdot (idade \cdot x_3)$

- $idade \cdot x_2$ **e** $idade \cdot x_3$ são os **termos de interação**. Eles permitem que a inclinação da reta (o efeito da idade) mude para cada tratamento.
  Essa única equação geral gera três equações de reta distintas, uma para cada tratamento, com **interceptos e inclinações diferentes**:
- **Tratamento A (Referência):** A inclinação é β1.
- **Tratamento B:** A inclinação é β1+β4.
- **Tratamento C:** A inclinação é β1+β5.

Conclusões

- **Pergunta de Pesquisa:** O efeito da idade na eficácia do tratamento depende de qual tratamento o paciente recebeu?
- **Teste de Hipótese:** Para responder a isso, é feito um **teste F parcial** para verificar se os coeficientes dos termos de interação (β4 e β5) são, em conjunto, diferentes de zero.
- **Resultado:** A análise da variância (ANOVA) mostra um **p-value muito baixo (< 0.001)** para os termos de interação.
- **Conclusão Final:** Rejeita-se a hipótese nula. Há forte evidência estatística de que existe uma **interação significativa**. Portanto, o efeito da idade na eficácia do tratamento realmente **depende** do tratamento administrado. Por exemplo, para pacientes do Tratamento C, a eficácia aumenta em 1.03 unidades para cada ano a mais de idade, enquanto para o Tratamento A, o aumento é de apenas 0.33 unidades.

8.7 - Leaving an Important Interaction Out of a Model

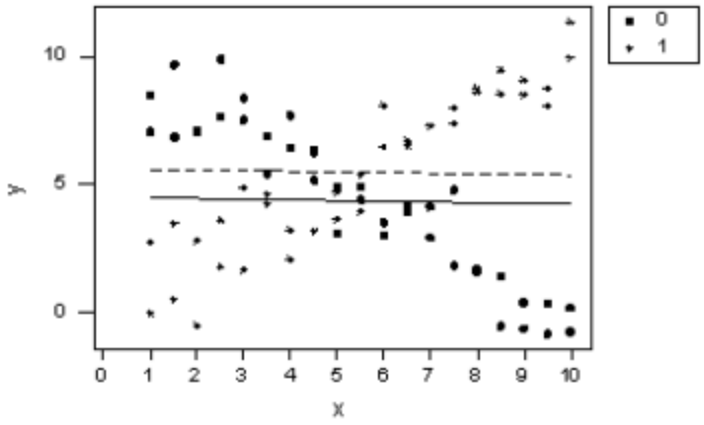
Aqui vemos o **grande perigo de omitir um termo de interação necessário** de um modelo de regressão. A principal lição é que essa omissão pode levar a conclusões completamente erradas.

O Cenário: Uma Interação Óbvia

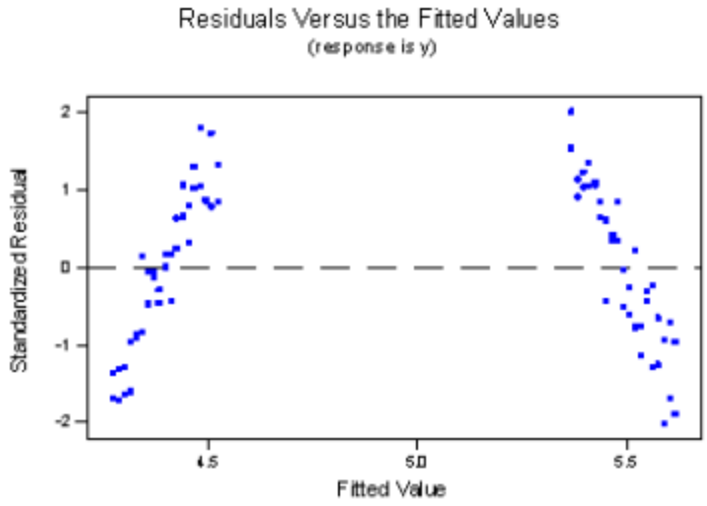
- O texto apresenta um conjunto de dados onde a relação entre `x` e `y` é visivelmente diferente para dois grupos (grupo 0 e grupo 1).
- **Inspeção Visual:** No gráfico, fica claro que para o **grupo 0**, `y` diminui à medida que `x` aumenta. Para o **grupo 1**, `y` aumenta à medida que `x` aumenta.
- Isso é um sinal clássico de que existe uma **interação forte** entre a variável `x` e a variável `grupo`.

O Modelo Errado (Sem Interação)

- Primeiro, os dados são analisados de propósito com um **modelo aditivo (errado)**, que força as retas de regressão a serem paralelas: `ŷ = 4.55 − 0.028x + 1.10 grupo`
- **Resultados Enganosos:** Ao olhar os p-values deste modelo, tanto o coeficiente de `x` (p = 0.831) quanto o de `grupo` (p = 0.125) **não são estatisticamente significantes**.
- **Conclusão Errada:** Um pesquisador que usasse este modelo concluiria, incorretamente, que nem `x` nem o `grupo` têm efeito sobre `y`. Os gráficos de resíduos deste modelo também mostrariam um padrão claro, indicando um mau ajuste.

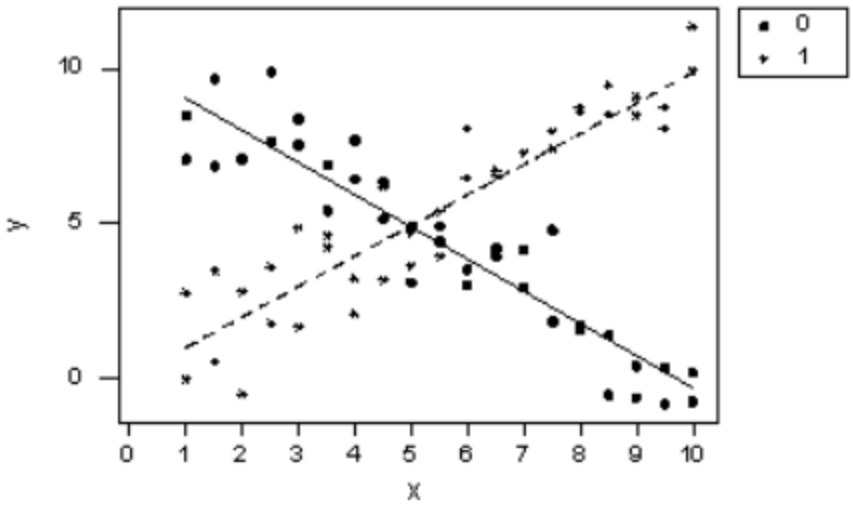


ell. By leaving the interaction term out of the model, v
The residuals versus fits plot:

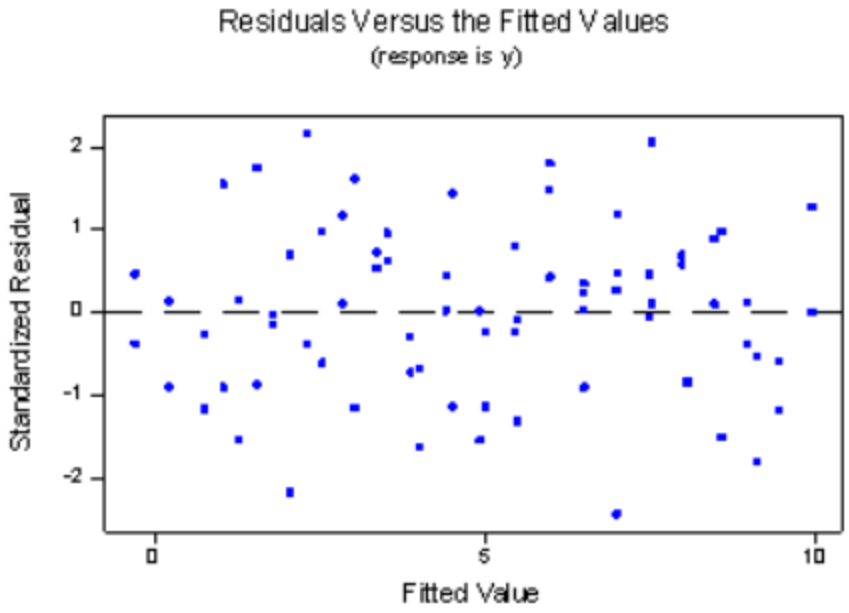


O Modelo Certo (Com Interação)

- Em seguida, os dados são reanalisados com o **modelo correto**, que inclui um termo de interação: `ŷ = 10.1 − 1.04x − 10.1 grupo + 2.03 groupx`
- **Resultados Corretos:** Este modelo se ajusta muito melhor aos dados. O p-value para o termo de interação (`groupx`) é **altamente significante (p < 0.001)**.
- **Conclusão Correta:** A análise agora revela a verdade: existe uma interação. O efeito de `x` sobre `y` definitivamente **depende** do grupo. O modelo produz duas retas com inclinações opostas, descrevendo perfeitamente o que foi visto no gráfico inicial.



ed regression model now allows the slopes of the two lines
e trend in the data. The residuals versus fits plot is about a



Moral da História

A principal mensagem é que **ignorar uma interação importante pode mascarar efeitos que são reais e significantes**. Omitir o termo de interação no exemplo levou o modelo a "anular" os efeitos opostos dos dois grupos, resultando em uma conclusão falsa de que não havia efeito algum.

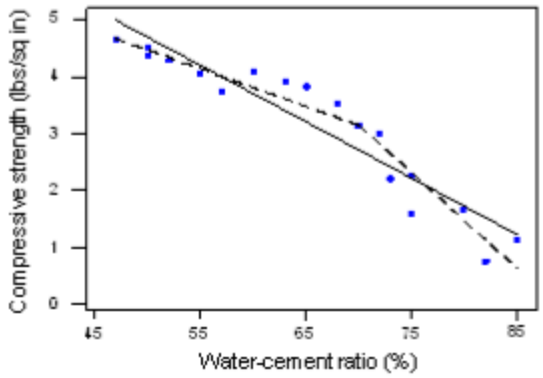
Isso reforça a regra de que **"nunca se deve interpretar os efeitos principais isoladamente na presença de uma interação significativa"**, pois eles não representam a história completa e podem ser enganosos.

8.8 - Piecewise Linear Regression Models

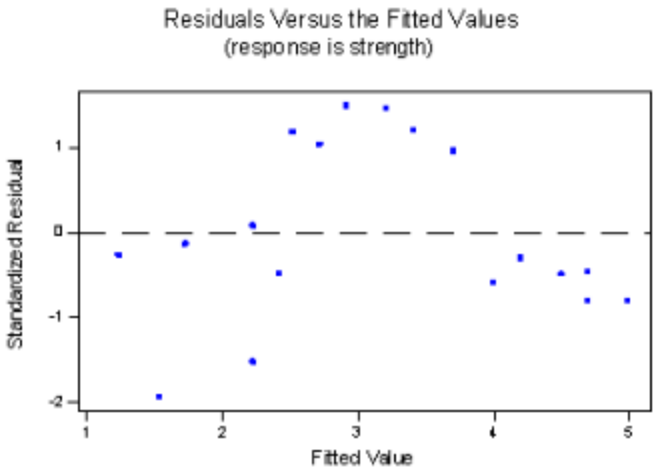
Aqui nos é apresentado uma técnica para modelar dados que seguem diferentes tendências lineares em diferentes intervalos da variável preditora (x).

O Problema: Uma Única Reta Não é Suficiente

- **Exemplo:** A análise da resistência do concreto (`strength`) em função da proporção de água/cimento (`ratio`).
- **Inspeção Visual:** Um gráfico de dispersão dos dados mostra que a relação entre as variáveis não é perfeitamente linear. Parece haver uma "quebra" ou mudança na inclinação da tendência.
- **Diagnóstico:** Ao ajustar uma regressão linear simples, o gráfico de resíduos mostra um padrão curvo claro, confirmando que o modelo linear simples não é adequado e que é necessário um modelo mais complexo.



id line —appears to fit the data fairly well in some overall sen



A Solução: Regressão por Partes (Piecewise)

- **Conceito:** Em vez de uma única reta, o modelo ajusta duas (ou mais) retas diferentes que se conectam em um ponto específico.
- **"Nó" (Knot):** É o ponto no eixo x onde as retas se conectam. No exemplo, o **nó** está no `ratio = 70`.

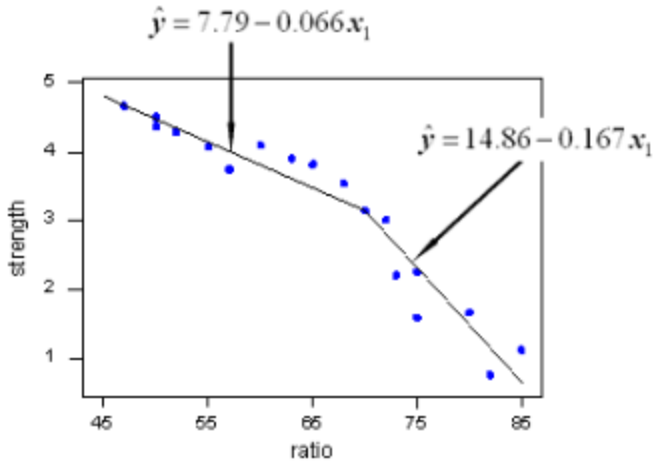
Como o Modelo é Construído

O modelo usa de forma inteligente uma **variável dummy** e um **termo de interação** para criar as duas retas em uma única equação:
$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot (x_1 - 70) \cdot x_2$

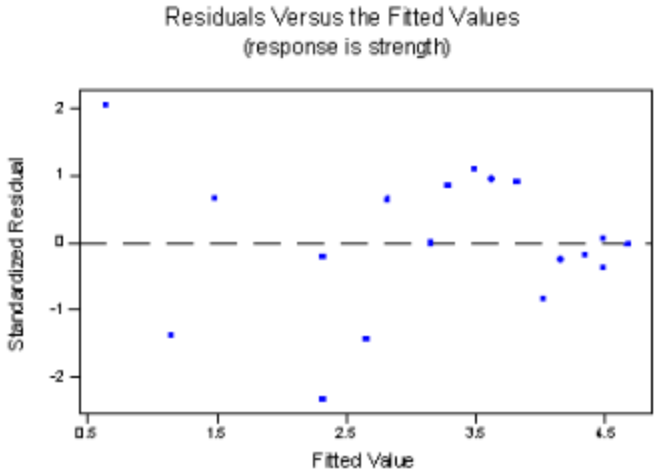
- x_1: É a variável original (ratio).
- x_2: É uma variável dummy que "ativa" a mudança na inclinação.
  - $x_2 = 0$ se $ratio \leq 70$
  - $x_2 = 1$ se $ratio > 70$
- $(x_1 - 70) \cdot x_2$: É o termo de interação. Ele só tem efeito quando $ratio$ é maior que 70. O coeficiente β_2 representa a **mudança na inclinação** que ocorre após o nó.

Resultados e Conclusão

- A equação estimada (`strength = 7.79 − 0.0663 ratio − 0.101 x2*`) gera duas funções de reta conectadas:
  1. Para $ratio \leq 70$: $\hat{y} = 7.79 - 0.0663\ ratio$
  2. Para $ratio > 70$: $\hat{y} = 14.86 - 0.167\ ratio$



s a significant improvement in the fit of the model:



- **Melhora Significativa:** Este modelo se ajusta muito melhor aos dados, o que é confirmado visualmente no gráfico (as duas retas seguem os pontos de perto) e pelo novo gráfico de resíduos, que não apresenta mais nenhum padrão óbvio.
  Em resumo, a regressão por partes é uma ferramenta poderosa para modelar relações que são lineares, mas que mudam de inclinação em determinados pontos.

8.9 - Further Examples

Aqui vemos três exemplos distintos que ilustram o uso de variáveis categóricas e interações em modelos de regressão múltipla.

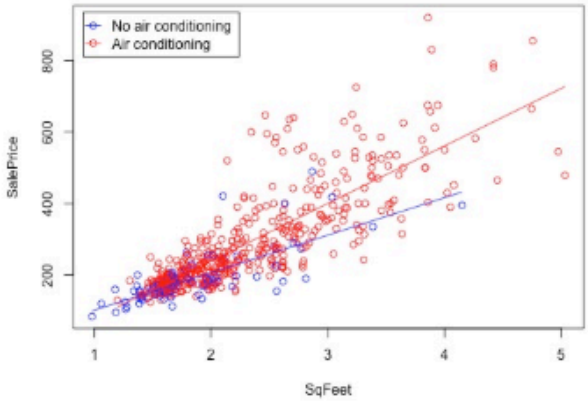
Exemplo 1: Dados de Massa Muscular (Modelo Aditivo Simples)

- **Objetivo:** Descrever a massa muscular em função da idade e do gênero (masculino/feminino).

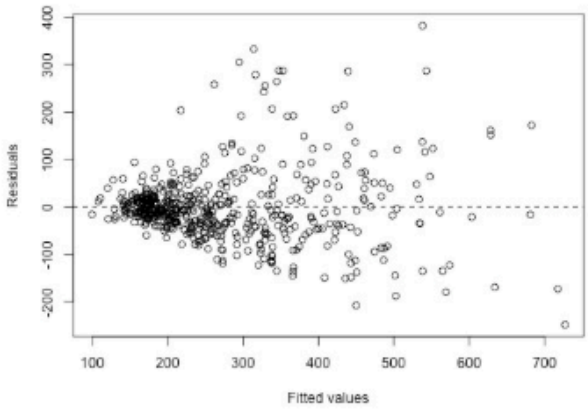
- **Modelo:** Um modelo aditivo simples é proposto. $\text{Massa Muscular} = \beta_0 + \beta_1 \cdot \text{Idade} + \beta_2 \cdot \text{Gênero}$
- **Interpretação:**
  - β_1: Representa a mudança na massa muscular para cada ano adicional de idade.
  - β_2: Representa a **diferença média** na massa muscular entre homens e mulheres, mantendo a idade constante.
  - Este modelo assume que o efeito da idade na massa muscular é o mesmo para ambos os gêneros (retas paralelas).

Exemplo 2: Dados de Imóveis (Modelo com Interação)

- **Objetivo:** Analisar o preço de venda de casas (`SalePrice`) com base no seu tamanho (`SqFeet`) e na presença ou ausência de ar-condicionado (`Air`).
- **Modelo:** Um modelo com **interação** é utilizado, pois suspeita-se que o efeito do tamanho do imóvel no preço dependa da presença de ar-condicionado. $\text{Preço} = \beta_0 + \beta_1 \cdot \text{Tamanho} + \beta_2 \cdot \text{Ar} + \beta_3 \cdot (\text{Tamanho} \cdot \text{Ar})$
- **Interpretação:**
  - O termo de interação ($\text{Tamanho} \cdot \text{Ar}$) permite que as retas de regressão **não sejam paralelas**.
  - A análise mostra que a diferença de preço entre casas com e sem ar-condicionado **aumenta** à medida que o tamanho da casa aumenta. Não é uma diferença constante.
- **Observação Importante:** O texto aponta um problema neste modelo: a **heterocedasticidade**, ou seja, a variabilidade dos erros aumenta à medida que o preço previsto aumenta (o gráfico de resíduos mostra um formato de cone). Isso indica que o modelo precisa de ajustes.



There is an increasing variance problem apparent in the above plot, which is even more obvious in the following residual plot:



Exemplo 3: Dados de Risco de Infecção Hospitalar (Variável Categórica com Múltiplos Níveis)

- **Objetivo:** Analisar o risco de infecção hospitalar (Y) com base no tempo de internação (X_1), frequência de raios-X (X_2) e a **região do hospital** (Nordeste, Centro-Norte, Sul, Oeste).
- **Modelo:** Para incluir a variável "região" com 4 categorias, são criadas **3 variáveis indicadoras (dummy)**, usando a região Nordeste como **grupo de referência**.
- **Interpretação dos Coeficientes:**
  - Cada coeficiente das variáveis indicadoras (β_3, β_4, β_5) representa a **diferença média no risco de infecção** entre aquela região (ex: Sul) e a região de referência (Nordeste), mantendo as outras variáveis constantes.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.134259   0.877347  -2.433  0.01668 *
Stay         0.505394   0.081455   6.205 1.11e-08 ***
Xray         0.017587   0.005649   3.113  0.00238 **
i2           0.171284   0.281475   0.609  0.54416
i3           0.095461   0.288852   0.330  0.74169
i4           1.057835   0.378077   2.798  0.00612 **
                         ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   Residual standard error: 1.036 on 105 degrees of freedom
 Multiple R-squared:  0.4198,    Adjusted R-squared:  0.3922
   F-statistic: 15.19 on 5 and 105 DF,  p-value: 3.243e-11
```

- **Teste de Hipótese (Teste F Parcial):**
  - Para verificar se existe uma diferença *geral* entre as regiões (ou seja, se a variável "região" como um todo é significante), um **teste F parcial** é utilizado.
  - O teste compara um modelo completo (com as variáveis de região) contra um modelo reduzido (sem elas).
  - **Conclusão:** O teste F mostra que existe uma diferença regional significativa. Análises posteriores indicam que a região Oeste tem um risco de infecção significativamente maior que a região Nordeste, enquanto as outras regiões não apresentam diferenças significantes em relação à referência.

8.10 - Summary

Basicamente nesse capítulo aprendemos os seguintes três tópicos:

1. Modelos com Variáveis Qualitativas (Categóricas)

Aprendemos a incluir preditores que não são numéricos (como "Fumo" ou "Região") em um modelo de regressão. A técnica principal é o uso de **variáveis indicadoras (dummy)**, geralmente codificadas como 0 ou 1.

- **Regra Chave:** Para uma variável com **c** categorias, usamos **c - 1** variáveis indicadoras.
- **Interpretação:** O grupo codificado com todos os zeros se torna o **grupo de referência**, e os coeficientes das outras variáveis indicadoras representam a **diferença** em relação a esse grupo.

2. Modelos com Efeitos de Interação

Este foi um grande passo além dos modelos aditivos (de retas paralelas). Uma interação ocorre quando o efeito de um preditor na resposta **depende do nível de outro preditor**.

- **O que significa:** A relação não é "tamanho único". Por exemplo, o efeito da idade na eficácia de um tratamento pode ser diferente para cada tipo de tratamento.

- **Como modelar:** Adicionamos um **termo de multiplicação** (ex: `idade * tratamento`) à equação.
- **Resultado Visual:** As retas de regressão para cada grupo **não são paralelas**, cada uma tendo sua própria inclinação.

3. Modelos de Regressão por Partes (Piecewise)

Esta é uma aplicação especial e inteligente dos termos de interação. É usada quando uma relação linear parece "quebrar" ou mudar de inclinação em um ponto específico.

- **Quando usar:** Quando os dados mostram duas ou mais tendências lineares distintas em diferentes intervalos de `x`.
- **Componente Chave:** O **"nó" (knot)**, que é o ponto `x` onde a inclinação muda.
- **Como funciona:** O modelo usa um termo de interação para "ativar" uma mudança na inclinação somente após o ponto do nó, criando duas ou mais retas conectadas que se ajustam melhor aos dados do que uma única reta.
  Essencialmente, a lição mostrou como construir modelos de regressão muito mais flexíveis e realistas, capazes de capturar relações complexas que uma regressão linear simples não conseguiria.