**Chương 2: Triển khai mô phỏng phân tích tương quan & hồi quy**

**2.1 Mô tả dữ liệu và nguồn thu thập**

Trong nghiên cứu này, chúng tôi thực hiện mô phỏng dữ liệu nhằm minh họa quá trình phân tích tương quan và hồi quy tuyến tính đơn.  
Hai biến số được xem xét bao gồm:

* **Biến X (Height)**: Chiều cao tính bằng cm.
* **Biến Y (Weight)**: Cân nặng tính bằng kg.

Dữ liệu được sinh ngẫu nhiên dựa trên giả định mối quan hệ tuyến tính:

Y=0.45X−50+ε*Y*=0.45*X*−50+*ε*

trong đó:

* ε∼N(0,σ2)*ε*∼N(0,*σ*2), với σ=4*σ*=4 kg.
* Số lượng quan sát: n=30*n*=30.

**Bảng 2.1 – Trích mẫu dữ liệu quan sát**

| **STT** | **Height (cm)** | **Weight (kg)** |
| --- | --- | --- |
| 1 | 174.82 | 30.20 |
| 2 | 168.00 | 25.34 |
| ... | ... | ... |
| 30 | 165.90 | 24.80 |

(Nguồn dữ liệu: Mô phỏng bằng Python, xem chi tiết ở Phụ lục A)

**2.2 Tiền xử lý dữ liệu**

Dữ liệu được kiểm tra qua các bước:

1. **Giá trị thiếu (Missing values):** Không phát hiện.
2. **Ngoại lai (Outliers):** Xác định bằng phương pháp IQR. Không có giá trị nào vượt ngưỡng Q1−1.5×IQR*Q*1​−1.5×*IQR* hoặc Q3+1.5×IQR*Q*3​+1.5×*IQR*.
3. **Đồng nhất đơn vị đo:** Cả hai biến đều ở đơn vị gốc, không cần chuẩn hóa.

**2.3 Phân tích tương quan**

**2.3.1 Công thức**

Hệ số tương quan Pearson được tính bằng:

rxy=∑i=1n(xi−xˉ)(yi−yˉ)(n−1)sxsy*rxy*​=(*n*−1)*sx*​*sy*​∑*i*=1*n*​(*xi*​−*x*ˉ)(*yi*​−*y*ˉ​)​

trong đó:

* xˉ*x*ˉ, yˉ*y*ˉ​: Trung bình mẫu của X*X* và Y*Y*.
* sx*sx*​, sy*sy*​: Độ lệch chuẩn mẫu.
* n*n*: Số quan sát.

**2.3.2 Kết quả tính toán**

* xˉ=168.683*x*ˉ=168.683, sx=6.300*sx*​=6.300
* yˉ=25.423*y*ˉ​=25.423, sy=4.899*sy*​=4.899
* Hiệp phương sai: sxy=20.180*sxy*​=20.180
* Hệ số tương quan Pearson: r=0.6539*r*=0.6539

**2.3.3 Kiểm định ý nghĩa**

Giả thuyết:

* H0:ρ=0*H*0​:*ρ*=0 (không có tương quan)
* H1:ρ≠0*H*1​:*ρ*=0 (có tương quan)

Thống kê kiểm định:

t=rn−21−r2=4.5733,df=28*t*=1−*r*2​*rn*−2​​=4.5733,*df*=28

p-value = 8.90×10−58.90×10−5 < 0.05  
→ Bác bỏ H0*H*0​, chấp nhận H1*H*1​: có mối tương quan tuyến tính dương có ý nghĩa thống kê.

**2.4 Xây dựng mô hình hồi quy tuyến tính đơn**

**2.4.1 Phương pháp**

Phương pháp bình phương tối thiểu (OLS) tìm b0,b1*b*0​,*b*1​ để tối thiểu hóa:

SSE=∑i=1n(yi−y^i)2*SSE*=*i*=1∑*n*​(*yi*​−*y*^​*i*​)2

**2.4.2 Công thức**

b1=sxysx2,b0=yˉ−b1xˉ*b*1​=*sx*2​*sxy*​​,*b*0​=*y*ˉ​−*b*1​*x*ˉ

**2.4.3 Kết quả mô hình**

* b1=0.5094*b*1​=0.5094
* b0=−60.3382*b*0​=−60.3382

Phương trình hồi quy:

Y^=−60.3382+0.5094X*Y*^=−60.3382+0.5094*X*

**Bảng 2.2 – Kết quả ước lượng tham số**

| **Tham số** | **Ước lượng** | **Sai số chuẩn** | **t-stat** | **p-value** |
| --- | --- | --- | --- | --- |
| b0*b*0​ | -60.3382 | 14.302 | -4.218 | 0.0002 |
| b1*b*1​ | 0.5094 | 0.1114 | 4.573 | 0.00009 |

**2.5 Độ phù hợp mô hình**

Hệ số xác định:

R2=SSRSST=r2=0.4276*R*2=*SSTSSR*​=*r*2=0.4276

Ý nghĩa: 42.76% biến thiên của cân nặng được giải thích bởi chiều cao, phần còn lại do yếu tố khác.

**2.6 Minh họa kết quả**

**Hình 2.1 – Biểu đồ phân tán và đường hồi quy**

(Chèn hình ch02\_regression.png)

**2.7 Thảo luận**

1. **Kết quả:** Mối tương quan dương vừa phải, phù hợp giả định ban đầu.
2. **Giới hạn:** R2*R*2 chưa cao, mô hình chưa tính đến nhiều yếu tố quan trọng khác.
3. **Hướng mở rộng:** Dùng hồi quy đa biến, thêm các biến như độ tuổi, giới tính, chế độ ăn uống.

**2.8 Kết luận chương**

Chương này đã triển khai mô phỏng dữ liệu, phân tích tương quan và xây dựng mô hình hồi quy tuyến tính đơn.  
Kết quả chỉ ra mối quan hệ tuyến tính có ý nghĩa thống kê giữa chiều cao và cân nặng, nhưng độ giải thích còn hạn chế.  
Nghiên cứu tiếp theo sẽ mở rộng mô hình và kiểm tra các giả định hồi quy để tăng độ chính xác.