

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

—*—

TIỂU LUẬN CUỐI KÌ

MÔ HÌNH HÓA THỐNG KÊ

Giảng viên hướng dẫn: **TS. Nguyễn Thị Mộng Ngọc**

Nhóm thực hiện: **Nhóm 4**

Học viên: **Phan Thị Thùy An**

MSHV: 20C29002

Đinh Thị Nữ

MSHV: 20C29013

Lý Phi Long

MSHV: 20C29028

Đặng Khánh Thi

MSHV: 20C29038

TP. Hồ Chí Minh – Tháng 04, 2021

Mục lục

1	Dữ liệu tự chọn	5
1.1	Dữ liệu 1: Mô hình hồi quy đa biến	6
1.2	Dữ liệu 2: Hồi quy thành phần chính	14
2	Dữ liệu có sẵn	19
2.1	Dữ liệu 1	20
2.2	Dữ liệu 2	25
2.3	Dữ liệu 3	31
2.4	Dữ liệu 4	41

BẢNG PHÂN CÔNG CÔNG VIỆC

Thành viên	Công việc	Mã số học viên
1. Phan Thị Thùy An Nhóm trưởng	<ul style="list-style-type: none">– Code và trình bày phần A: data 2– Code và trình bày phần B: data 2– Thảo luận ghi nhận xét 6 bài	20C29002
2. Đinh Thị Nữ	<ul style="list-style-type: none">– Code và trình bày phần A: data 2– Code và trình bày phần B: data 1– Thảo luận ghi nhận xét 6 bài	20C29013
3. Lý Phi Long	<ul style="list-style-type: none">– Code và trình bày phần A: data 1– Code và trình bày phần B: data 4– Thảo luận ghi nhận xét 6 bài	20C29028
4. Đặng Khánh Thi	<ul style="list-style-type: none">– Code và trình bày phần A: data 1– Code và trình bày phần B: data 3– Thảo luận ghi nhận xét 6 bài	20C29038

Chương 1

Dữ liệu tự chọn

- Tên "đề tài", nguồn gốc của dữ liệu, giới thiệu các biến.
- Mô hình chọn được; phân tích kết quả
- Đưa ra những phương pháp/phân tích khác có thể giúp cho kết quả tốt hơn.
- Kết luận.

1.1 Dữ liệu 1: Mô hình hồi quy đa biến

Giới thiệu bộ dữ liệu

Bộ dữ liệu được tìm thấy trên trang Kaggle - một cộng đồng trực tuyến về khoa học dữ liệu và học máy. Đó là bộ dữ liệu **Chi phí Y tế Cá nhân**¹ (*Medical Cost Personal Datasets*). Đây là một bộ dữ liệu được lấy ra từ cuốn *Machine Learning with R* của Brett Lantz, một cuốn sách giới thiệu về học máy bằng R.

Bộ dữ liệu ghi lại các thông tin về thông tin của người đăng kí bảo hiểm và chi phí mà bảo hiểm y tế phải chi trả cho cá nhân đó. Bộ dữ liệu có 1338 quan trắc, gồm 7 biến sau:

1. **age**: tuổi
2. **sex**: giới tính
3. **bmi**: chỉ số đo cân nặng, sử dụng tỷ lệ giữa cân nặng và chiều cao (kg/m), chỉ số BMI lý tưởng là từ 18.5 đến 24.9.
4. **children**: số lượng trẻ em được bao gồm trong bảo hiểm y tế của người đăng kí.
5. **smoker**: 1 nếu người đó có hút thuốc, ngược lại là 0.
6. **region**: vùng miền ở US, bao gồm Đông Bắc (*northeast*), Đông Nam (*southeast*), Tây Nam (*southwest*), Tây Bắc (*northwest*).
7. **charges**: Chi phí y tế của cá nhân được chi trả bởi bảo hiểm y tế.

Nhận thấy biến **region** có bốn giá trị, để thuận tiện cho việc hồi quy mô hình đa biến, chúng ta cần phải tách **region** thành ba biến giả lần lượt là **region_ne** - vùng Đông Bắc, **region_se** - vùng Đông Nam và **region_sw** - vùng Tây Nam, nếu không nằm trong 3 vùng này thì nó là vùng Tây Bắc. Vậy bộ dữ liệu hiện tại có tất cả 9 biến.

Kiểm tra sự trùng lặp dữ liệu trong bộ dữ liệu, ta có kết quả từ phạm mềm R ở hình 1.1.1, thấy rằng chỉ tồn tại một dữ liệu bị trùng, ta tiến hành loại bỏ dữ liệu này. Vậy bộ dữ liệu hiện tại có 1337 quan trắc.

¹<https://www.kaggle.com/mirichoi0218/insurance>

```
> dup = duplicated(insurance)
> dup_data = insurance[dup,]; dup_data
  age sex  bmi children smoker  region  charges
582 19 male 30.59      0     no northwest 1639.563
```

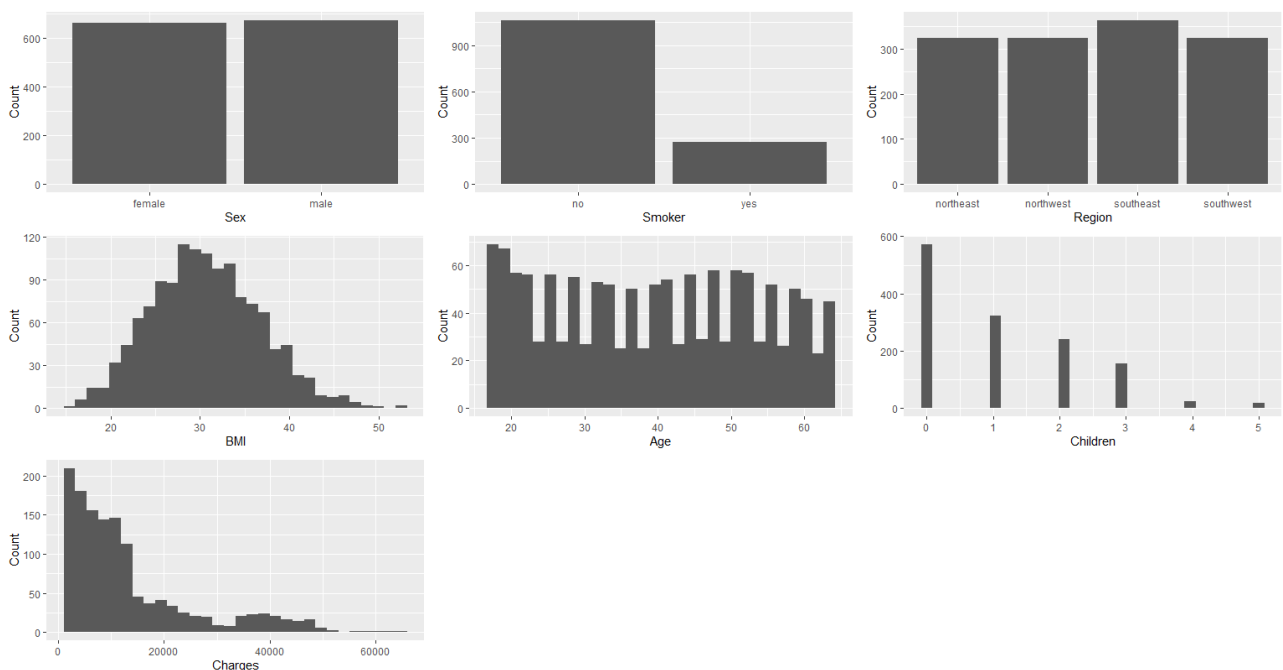
Hình 1.1.1: Dữ liệu bị trùng lặp

Một vài quan trắc đầu tiên trong bộ dữ liệu được thể hiện trong hình 1.1.2 và số chiều của nó: 1337 dòng (quan trắc) và 9 cột (biến).

```
> head(insurance)
  age  sex  bmi children smoker  charges region_ne region_se region_sw
1  19 female 27.900      0      1 16884.924      0      0      1
2  18  male 33.770      1      0 1725.552      0      1      0
3  28  male 33.000      3      0 4449.462      0      1      0
4  33  male 22.705      0      0 21984.471      0      0      0
5  32  male 28.880      0      0 3866.855      0      0      0
6  31 female 25.740      0      0 3756.622      0      1      0
> dim(insurance)
[1] 1337  9
```

Hình 1.1.2: Một vài quan trắc đầu tiên và số chiều của bộ dữ liệu

Phân bố của 7 biến ban đầu ở hình 1.1.3 và trung bình tổng của từng biến theo biến phụ thuộc `charges` ở hình 1.1.4.



Hình 1.1.3: Phân bố của 7 biến ban đầu

```
> aggregate(charges ~ sex, insurance, mean)
  sex  charges
1 female 12569.58
2  male 13975.00
```

(a) sex

```
> aggregate(charges ~ smoker, insurance, mean)
 smoker  charges
1      0  8440.66
2      1 32050.23
```

(b) smoker

```
> aggregate(charges ~ region_se + region_ne + region_sw, insurance, mean)
 region_se region_ne region_sw  charges
1         0         0         0 12450.84
2         1         0         0 14735.41
3         0         1         0 13406.38
4         0         0         1 12346.94
```

(c) region

```
> aggregate(charges ~ bmi_ranges, insurance, mean)
 bmi_ranges  charges
1  [15,20]  8756.355
2  (20,25]  9529.920
3  (25,30] 10086.132
4  (30,35] 10929.365
5  (35,40] 12366.023
6  (40,45] 14538.651
7  (45,50] 15013.603
8  (50,55] 16988.006
```

(d) bmi

```
> aggregate(charges ~ age_ranges, insurance, mean)
 age_ranges  charges
1  [18,28]  9115.278
2  (28,38] 11598.555
3  (38,48] 14334.586
4  (48,58] 15887.954
```

(e) age

```
> aggregate(charges ~ children, insurance, mean)
 children  charges
1         0 12384.695
2         1 12731.172
3         2 15073.564
4         3 15355.318
5         4 13850.656
6         5  8786.035
```

(f) children

Hình 1.1.4: Trung bình tổng của từng biến theo biến phụ thuộc charges

Một số điều thú vị thấy được ở hai hình 1.1.3 và 1.1.4, ta xét lần lượt từng biến:

- **sex**: dù là nam hay nữ thì phân bố giữa hai giới này đều xấp xỉ nhau, đồng thời chi phí trung bình mà bảo hiểm y tế chi trả cũng xấp xỉ nhau.
- **smoker**: số lượng người hút thuốc ít hơn số lượng người không hút thuốc có trong bộ dữ liệu, nhưng chi phí trung bình mà bảo hiểm y tế chi trả cho nhóm này thì hoàn toàn cao hơn rất nhiều, điều này khá hiển nhiên.
- **region**: phân bố của các vùng và chi phí trung bình mà bảo hiểm y tế chi trả ở từng vùng cũng đều xấp xỉ nhau.
- **bmi**: chỉ số BMI có phân bố dạng chuẩn, và chi phí trung bình mà bảo hiểm y tế chi trả cũng tăng dần đều theo chỉ số này, điều này cũng hợp lý vì khi chỉ số BMI càng cao thì khả năng bị béo phì cũng tăng.
- **age**: tuổi tác có phân bố ngẫu nhiên, và chi phí trung bình mà bảo hiểm y tế chi trả cũng tăng dần theo tuổi, điều này cũng khá hiển nhiên.
- **children**: phân bố của trẻ em được hưởng theo bảo hiểm bị lệch hẳn về bên trái, nên chi phí trung bình mà bảo hiểm y tế chi trả cho 4-5 trẻ em có thể bị sai lệch do mất cân bằng dữ liệu.

Nhận xét tổng quan, ta thấy rằng chi phí bảo hiểm y tế chi trả **charges** có khả năng phụ thuộc vào các đặc tính như người hút thuốc **smoker**, chỉ số **bmi**, tuổi tác **age** và số trẻ em phụ thuộc **children**. Các đặc tính còn lại như vùng miền **region** và giới tính **sex** có thể sẽ không ảnh hưởng nhiều đến **charges**.

Phân tích và chọn mô hình

Xét mô hình đầy đủ sau:

$$\begin{aligned} \text{charges} = & \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{sex} + \beta_3 \times \text{bmi} + \beta_4 \times \text{children} + \beta_5 \times \text{smoker} \\ & + \beta_6 \times \text{region_ne} + \beta_7 \times \text{region_se} + \beta_8 \times \text{region_sw} + \epsilon \end{aligned} \quad (1.1.1)$$

Mô hình hồi quy đầy đủ có các thông số ở hình 1.1.5, đúng như dự đoán, biến vùng miền **region** và giới tính **sex** không có ý nghĩa thống kê, và các biến còn lại có ý nghĩa thống kê khá cao.

```

> mod_full = lm(charges ~ ., insurance)
> summary(mod_full)

Call:
lm(formula = charges ~ ., data = insurance)

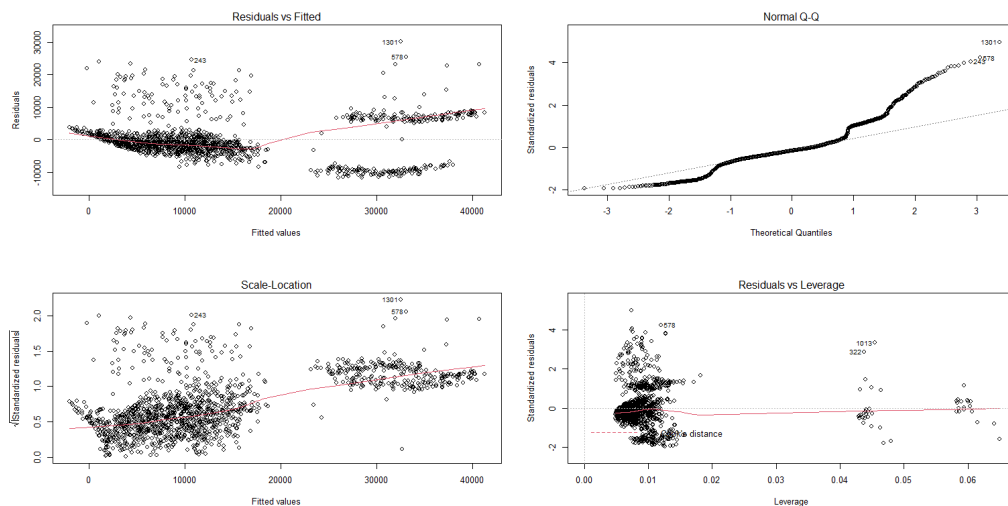
Residuals:
    Min       1Q   Median       3Q      Max
-11687  -2906   -938    1492   30042

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12301.48    995.05  -12.363  < 2e-16 ***
age           257.11     11.93   21.556  < 2e-16 ***
sexmale      -126.44    333.09   -0.380  0.704296
bmi           336.96     28.62   11.772  < 2e-16 ***
children1     389.07    421.63    0.923  0.356295
children2    1633.74    466.97    3.499  0.000483 ***
children3     962.44    548.39    1.755  0.079489 .
children4    2945.16   1239.67    2.376  0.017655 *
children5    1114.26   1456.58    0.765  0.444417
smoker       23835.29   414.33   57.527  < 2e-16 ***
region_ne     376.51    477.11    0.789  0.430170
region_se    -656.82    480.71   -1.366  0.172064
region_sw    -576.41    479.00   -1.203  0.229051
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6061 on 1324 degrees of freedom
Multiple R-squared:  0.7518,    Adjusted R-squared:  0.7495
F-statistic: 334.1 on 12 and 1324 DF,  p-value: < 2.2e-16

```

Hình 1.1.5: Mô hình đầy đủ



Hình 1.1.6: Các biểu đồ của mô hình lựa chọn

Tiến hành sử dụng phương pháp tính hệ số VIF để kiểm tra hiện tượng đa cộng tuyến có trong mô hình này, kết quả từ phần mềm R ở hình 1.1.7 cho thấy các hệ số VIF đều dưới 5, chứng tỏ không tồn tại hiện tượng này trong mô hình.

```
> vif(mod_full)
          GVIF Df GVIF^(1/(2*Df))
age      1.020542 1      1.010219
sex      1.009380 1      1.004679
bmi      1.108950 1      1.053067
children 1.024871 5      1.002460
smoker   1.018045 1      1.008982
region_ne 1.521188 1      1.233364
region_se 1.666349 1      1.290871
region_sw 1.536442 1      1.239533
```

Hình 1.1.7: Hiện tượng đa cộng tuyến trong mô hình đầy đủ

Do mô hình đầy đủ đã có nhiều biến có ý nghĩa thống kê, nên nhóm em sẽ sử dụng phương pháp Stepwise lùi để chọn mô hình phù hợp nhanh hơn, và tiêu chuẩn BIC cho mô hình đơn giản trên. Các bước và kết quả chọn mô hình, các thông số của nó được thể hiện ở hình 1.1.8

```
> mod_BIC$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
charges ~ age + sex + bmi + children + smoker + region_ne + region_se + region_sw
Final Model:
charges ~ age + bmi + smoker

Step  Df  Deviance Resid. Df  Resid. Dev  AIC
1      5  636297214    1324  48637825279 23370.06
2 - children 5  636297214    1324  48637825279 23370.06
3 - sex      1  3935431    1330  49274122492 23351.44
4 - region_ne 1  14387939    1331  49292445862 23337.54
5 - region_sw 1  126087384    1332  49418533246 23333.76
6 - region_se 1  92479000    1333  49511012246 23329.06

> summary(mod_BIC)
Call:
lm(formula = charges ~ age + bmi + smoker, data = insurance)

Residuals:
    Min       1Q   Median       3Q      Max
-12415.2  -2974.4   -981.3   1490.3  28972.6

Coefficients:
(Intercept) -11671.69      259.43      322.64    23822.18
            938.14      11.95      27.50      413.06
            -12.44     21.71     11.73     57.67
            <2e-16 ***   <2e-16 ***   <2e-16 ***   <2e-16 ***
age          bmi        smoker

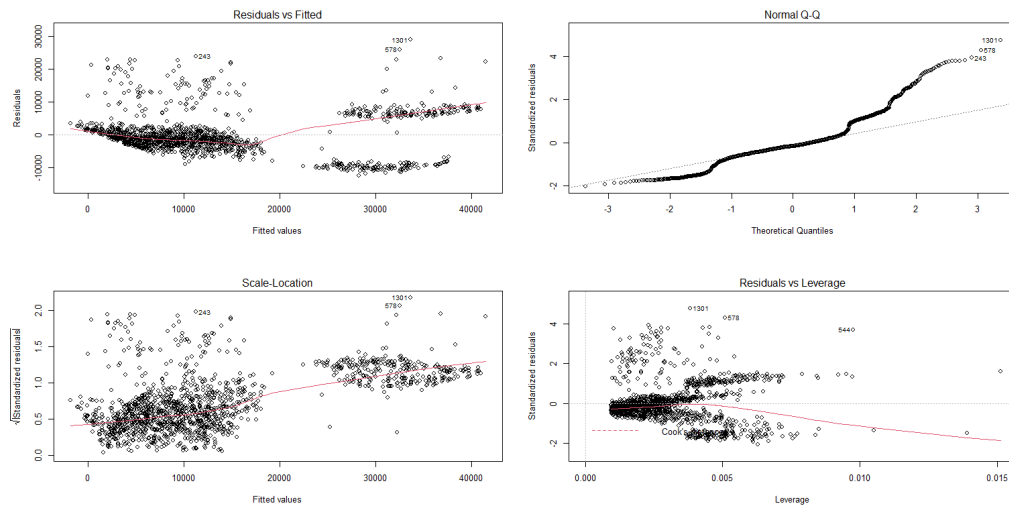
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6094 on 1333 degrees of freedom
Multiple R-squared:  0.7473,    Adjusted R-squared:  0.7467
F-statistic: 1314 on 3 and 1333 DF,  p-value: < 2.2e-16
```

(a) Các bước và kết quả chọn mô hình

(b) Mô hình lựa chọn

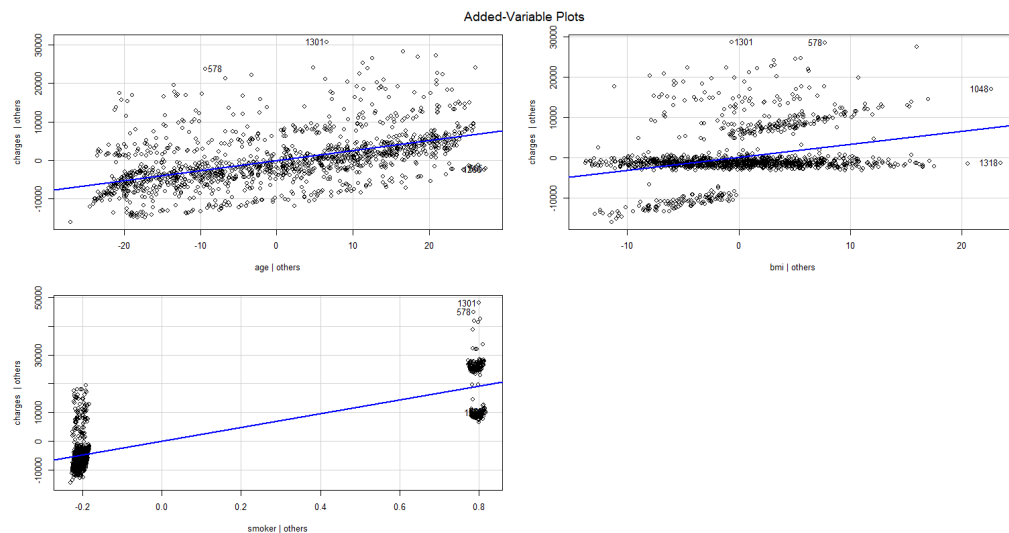
Hình 1.1.8: Mô hình lựa chọn với phương pháp Stepwise và tiêu chuẩn BIC



Hình 1.1.9: Các biểu đồ của mô hình lựa chọn

```
> vif(mod_BIC)
age      bmi      smoker
1.012793 1.012145 1.000699
```

Hình 1.1.10: Hiện tượng đa cộng tuyến trong mô hình lựa chọn



Hình 1.1.11: Mối quan hệ tuyến tính của từng biến đối với biến phụ thuộc charges

Nhận xét và kết luận

Ý nghĩa mô hình?

Mô hình giải thích đc bn?

Giải thích các plot của mô hình?

Có thể mở rộng bài toán bằng cách xem xét các giá trị ngoại lai và hiểu được sự tồn tại của các giá trị đó. Sau đó, chọn phương pháp tối ưu nhằm lựa chọn mô hình phù hợp hơn.

1.2 Dữ liệu 2: Hồi quy thành phần chính

Giới thiệu bộ dữ liệu

Hiện nay, Xe đạp cho thuê được giới thiệu ở nhiều thành phố để nâng cao sự thoải mái khi di chuyển. Điều cần quan tâm khi cho thuê xe đạp là xe đạp phải luôn sẵn sàng và tiếp cận được người dùng vào đúng thời điểm, giúp giảm bớt thời gian chờ. Do đó, việc đảm bảo một nguồn cung cấp xe đạp cho thuê ổn định cho thành phố trở thành mối quan tâm lớn. Phần quan trọng là cần dự đoán được số lượng xe đạp cần thiết tại mỗi giờ, để có được nguồn cung cấp xe đạp cho thuê ổn định.

Nhóm em sử dụng bộ dữ liệu Nhu cầu thuê xe đạp ở Seoul² (**Seoul Bike Sharing Demand Dataset**).

Bộ dữ liệu ghi lại các thông tin về thời tiết, số lượng xe đạp được thuê mỗi giờ theo từng ngày, từ 01/12/2017 đến 31/11/2018. Bộ dữ liệu có 8760 quan trắc, gồm 14 biến sau:

1. **Date** - Ngày ghi lại số lượng xe đạp cho thuê
2. **Rented Bike count** - Số lượng xe đạp được thuê được ghi lại theo mỗi giờ
3. **Hour** - Giờ trong ngày
4. **Temperature** - Nhiệt độ ($^{\circ}C$)
5. **Humidity** - Độ ẩm (%)
6. **Windspeed** - Tốc độ gió (m/s)
7. **Visibility** - Tầm nhìn xa ($10m$)
8. **Dew point temperature** - Nhiệt độ điểm sương ($^{\circ}C$)
9. **Solar radiation** - Bức xạ mặt trời (Mj/m^2)
10. **Rainfall** - Lượng mưa (mm)

²<https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand>

11. **Snowfall** - Lượng tuyết rơi (*cm*)

12. **Seasons** - Mùa (Winter, Spring, Summer, Autumn)

13. **Holiday** - Ngày lễ (Holiday/No holiday)

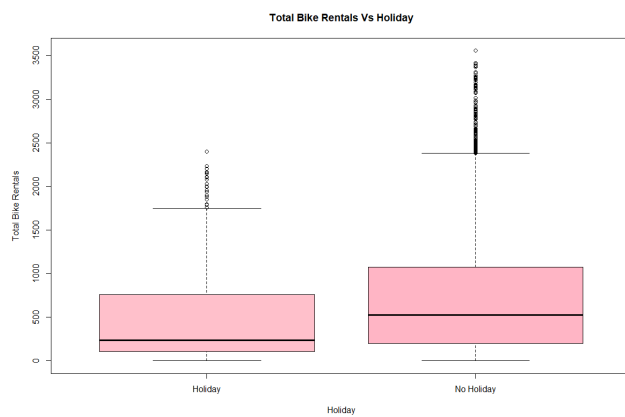
14. **Functional Day** - Ngày làm việc (Yes nếu là ngày làm việc, No nếu ngược lại)

Một vài quan trắc đầu tiên trong bộ dữ liệu được thể hiện trong hình 1.2.1

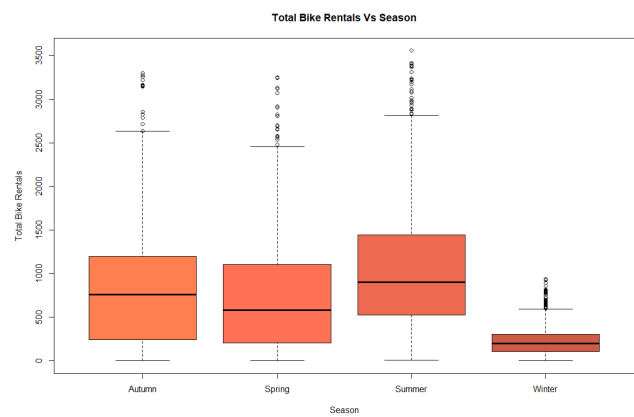
```
> head(bike)
  Date Rented.Bike.Count Hour Temperature..C. Humidity... Wind.speed..m.s.
1 01/12/2017          254     0         -5.2         37         2.2
2 01/12/2017          204     1         -5.5         38         0.8
3 01/12/2017          173     2         -6.0         39         1.0
4 01/12/2017          107     3         -6.2         40         0.9
5 01/12/2017           78     4         -6.0         36         2.3
6 01/12/2017          100     5         -6.4         37         1.5
  Visibility..10m. Dew.point.temperature..C. Solar.Radiation..MJ.m2. Rainfall.mm. Snowfall..cm.
1           2000             -17.6              0              0              0
2           2000             -17.6              0              0              0
3           2000             -17.7              0              0              0
4           2000             -17.6              0              0              0
5           2000             -18.6              0              0              0
6           2000             -18.7              0              0              0
  Seasons      Holiday Functioning.Day
1 Winter No Holiday          Yes
2 Winter No Holiday          Yes
3 Winter No Holiday          Yes
4 Winter No Holiday          Yes
5 Winter No Holiday          Yes
6 Winter No Holiday          Yes
> dim(bike)
[1] 8760 14
```

Hình 1.2.1: Một vài quan trắc đầu tiên và số chiều của bộ dữ liệu

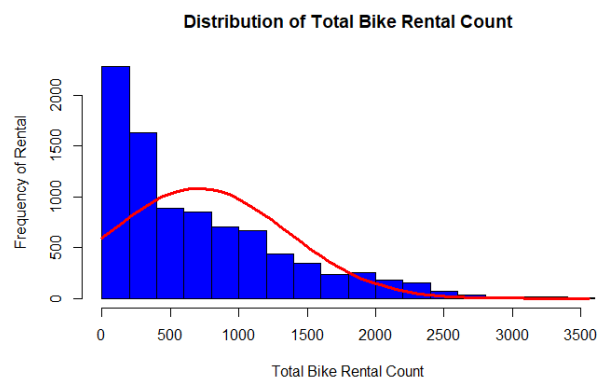
Phân tích và chọn mô hình



(a) ...



(b) Kiểm tra đa cộng tuyến

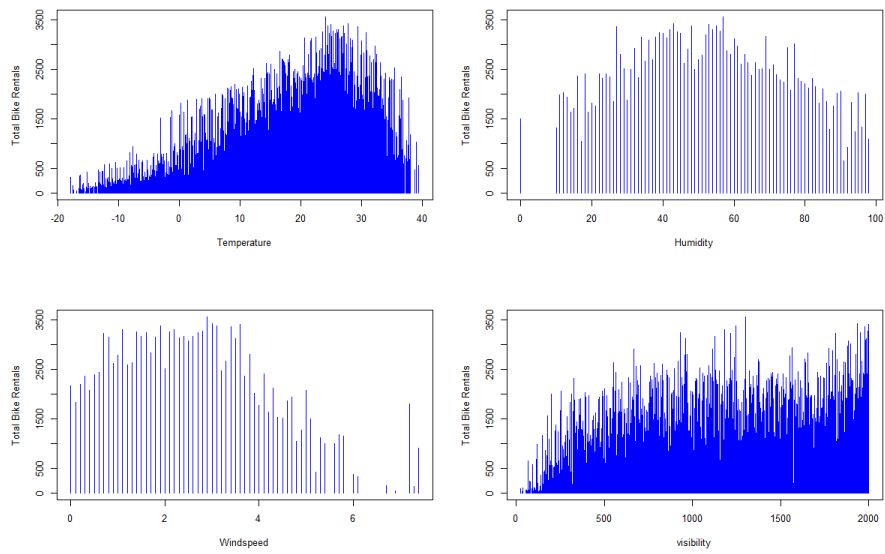


(c) ...

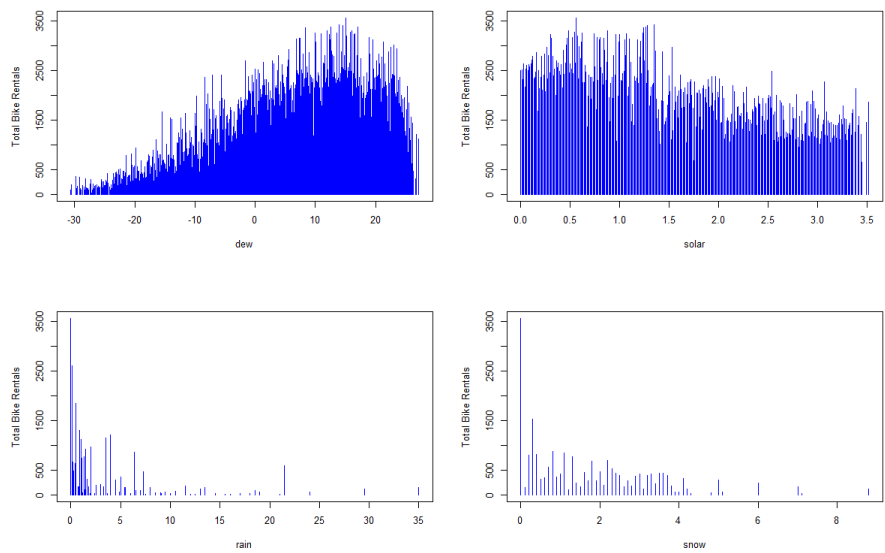
	count	hour	temp	humidity	wind	visibility	dew	solar	rain	snow
count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.28	-0.12	-0.14
hour	0.41	1	0.12	-0.24	0.29	0.1		0.15		0.15
temp	0.54	0.12	1	0.16	-0.04	0.05	0.91	0.35	0.05	-0.22
humidity	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11
wind	0.12	0.29	-0.04	-0.34	1	0.17	-0.18	0.33	-0.17	-0.12
visibility	0.2	0.1	0.05	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12
dew	0.38		0.91	0.54	-0.18	-0.18	1	0.05	0.15	-0.15
solar	0.28	0.15	0.35	-0.46	0.33	0.15	0.05	1	-0.07	-0.07
rain	-0.12		0.05	0.24	-0.17	-0.17	0.15	-0.07	1	
snow	-0.14	0.15	-0.22	0.11	-0.12	-0.12	-0.15	-0.07		1

(d) Kiểm tra đa cộng tuyến

Hình 1.2.2: Một vài quan sát ...



(a) ...



(b) Kiểm tra đa cộng tuyến

Hình 1.2.3: Một vài quan sát ...

Nhận xét và kết luận

Chương 2

Dữ liệu có sẵn

- Chọn mô hình phù hợp nhất giải thích biến phụ thuộc với từng bộ dữ liệu.
- Nêu rõ phương pháp chọn mô hình và lý do chọn phương pháp đó.
- Nói rõ ý nghĩa của mô hình đã chọn.

2.1 Dữ liệu 1

Những thông tin về các giám đốc điều hành các tập đoàn Hoa Kỳ. Bộ dữ liệu gồm 177 quan trắc và 15 biến.

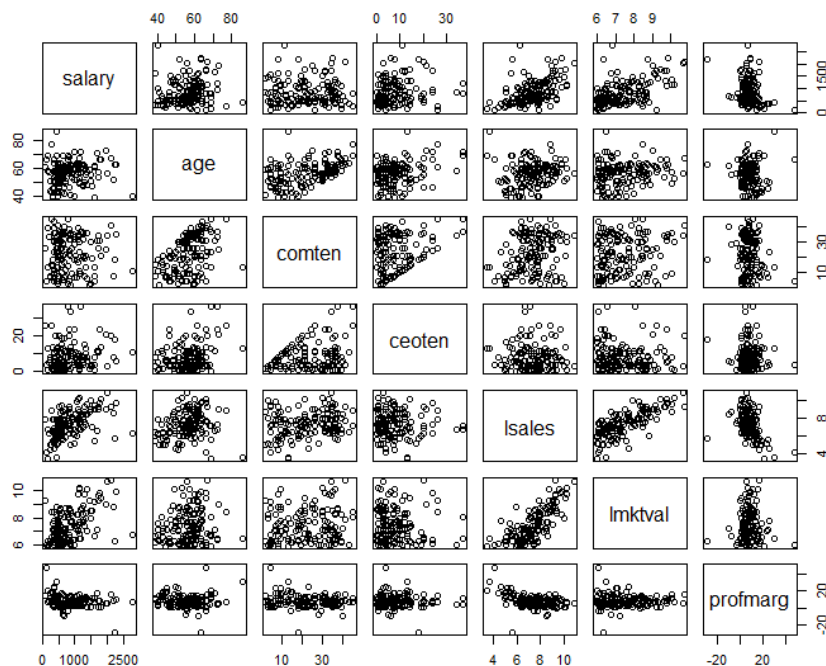
* Phương pháp chọn: Stepwise - tiến; tiêu chuẩn chọn: AIC.

Tìm hiểu và tiền xử lý dữ liệu

Một số biến trong bộ dữ liệu kiểu số có đơn vị tính lớn như: *sales*, *profit*, *lmktval*. Nếu đưa những biến này vào phương trình hồi quy có thể dẫn tới hiện tượng bias do tác động của những biến này lên model lấn át những biến khác còn lại như *age*, *ceoten*.... Nên ta sẽ dùng phương pháp logarit cho 3 biến này trong model tương ứng với 3 biến mới là: *lsales*, *lmktval* và *profmarg*. (1)

Từ biểu đồ dưới ta thấy ba biến định lượng *lsales*, *lmktval* và *profmarg* xảy ra hiện tượng đa cộng tuyến. Tuy nhiên có xảy ra hiện tượng đa cộng tuyến giữa 2 biến *sales* và *profit* (hình 2.1.1).

Tính độ tương quan giữa biến *salary* với lần lượt 2 biến trên ta có:



Hình 2.1.1: Mối tương quan giữa các biến

```
> cor(train[c("salary", "lsales", "lmktval", "profmarg")])
      salary    lsales    lmktval    profmarg
salary 1.0000000 0.4912099 0.51978488 -0.24975911
lsales 0.4912099 1.0000000 0.75006264 -0.42949701
lmktval 0.5197849 0.7500626 1.00000000 0.04471558
profmarg -0.2497591 -0.4294970 0.04471558 1.00000000
```

Hình 2.1.2: Mức độ tương quan giữa biến *lsales* và *promarg*

Xét bảng correlation giữa các biến độc lập với nhau và giữa các biến độc lập với biến phụ thuộc, ta thấy: Giữa hai biến *lmktval* và biến *lsales* có mối tương quan rất cao (≈ 0.75). Tuy nhiên biến *lmktval* lại có mối tương quan cao hơn với biến phụ thuộc *salary*. Mặt khác giữa biến *profmarg* và *lsales* cũng có mối tương quan cao (≈ -0.42). Nên ta loại bỏ biến *lsales* khỏi danh sách các biến được xét. (2)

Từ (1) và (2) ta có mô hình với đầy đủ các biến cần lựa chọn như sau:

$$\begin{aligned} salary = \beta_0 + \beta_1 \times age + \beta_2 \times college + \beta_3 \times grad + \beta_4 \times comten \\ + \beta_5 \times ceoten + \beta_6 \times lmktval + \beta_7 \times profmarg \end{aligned} \quad (2.1.1)$$

Thực hiện phân rã hai biến phân loại gồm *college* và *grad* trước khi thực hiện phương pháp **Stepwise tiến** với **tiêu chuẩn AIC**.

Để đánh giá chất lượng mô hình ta chia tập dữ liệu thành hai phần, training và testing, với tỷ lệ 8 : 2 sau đó tiến hành phương pháp chọn biến trên tập training.

Chọn biến bằng phương pháp Stepwise tiến và tiêu chuẩn AIC

```
[1] "salary" "age" "college" "grad" "comten" "ceoten" "lmktval" "profmarg"
> l0 = lm(formula = train$salary ~ 1, data = train) # non independence variable
> l1 = lm(formula = train$salary ~ ., data = train) # full independence variable
> modbest_Fow = step(l0, scope = list(lower = l0,
+                                     upper = l1), direction = 'forward', k = 2)
Start: AIC=1825.78
train$salary ~ 1
```

	Df	Sum of Sq	RSS	AIC
+ lmktval	1	11242276	42481047	1794.4
+ profmarg	1	993901	52729423	1825.1
+ age	1	833601	52889723	1825.6
+ ceoten	1	816752	52906571	1825.6
+ comten	1	784116	52939207	1825.7
<none>			53723323	1825.8
+ college	1	225711	53497612	1827.2
+ grad	1	1333	53721991	1827.8

```
Step: AIC=1794.44
train$salary ~ lmktval
```

	Df	Sum of Sq	RSS	AIC
+ profmarg	1	1319152	41161895	1792.0
+ ceoten	1	1069953	41411095	1792.8
<none>			42481047	1794.4
+ grad	1	398593	42082454	1795.1
+ comten	1	199305	42281743	1795.8
+ age	1	177509	42303538	1795.8
+ college	1	90861	42390186	1796.1

```
Step: AIC=1791.96
train$salary ~ lmktval + profmarg
```

	Df	Sum of Sq	RSS	AIC
+ ceoten	1	1067048	40094847	1790.2
<none>			41161895	1792.0
+ grad	1	215822	40946074	1793.2
+ age	1	170753	40991143	1793.4
+ college	1	92712	41069183	1793.6
+ comten	1	33866	41128029	1793.8

```
Step: AIC=1790.23
train$salary ~ lmktval + profmarg + ceoten
```

	Df	Sum of Sq	RSS	AIC
<none>			40094847	1790.2
+ grad	1	142580	39952267	1791.7
+ college	1	38627	40056220	1792.1
+ comten	1	28636	40066211	1792.1
+ age	1	1	40094846	1792.2

Hình 2.1.3: Kết quả chọn biến theo phương pháp StepWise tiến với tiêu chuẩn AIC

Tổng quan tiêu chuẩn AIC thì mô hình tốt là mô hình có giá trị AIC nhỏ nhất. Ở mô hình 1, biến *lmktval* được chọn vào mô hình vì có AIC nhỏ nhất trong tất cả các kết hợp với các biến còn lại. Tương tự AIC được tính cho mô hình thêm biến thứ 2, *ceoten*, và biến thứ 3 là *ceoten* (hình 2.1.4).

```
> summary(modbest_Fow)

Call:
lm(formula = train$salary ~ lmktval + profmarg + ceoten, data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-1339.1  -227.0   -72.8   163.7  4351.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -950.696    302.598   -3.142  0.00206 **
lmktval       248.204     38.909    6.379  2.5e-09 ***
profmarg     -13.929      6.544   -2.128  0.03508 *
ceoten        11.714      6.113    1.916  0.05738 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539 on 138 degrees of freedom
Multiple R-squared:  0.2537,    Adjusted R-squared:  0.2375
F-statistic: 15.64 on 3 and 138 DF,  p-value: 8.262e-09
```

Hình 2.1.4: Kết quả hồi quy mô hình với các biến được chọn

Với ba biến được chọn ở trên, mô hình 2.1.1 trở thành mô hình mới:

$$salary = -950.6 + 248.2 * lmktval - 13.9 * profmarg + 11.7 * ceoten \quad (2.1.2)$$

Tuy nhiên ta nhận thấy biến *ceoten* có $p_{value} \geq \alpha$ ($0.05738 \geq 0.05$) nên không có ý nghĩa thống kê trong mô hình. Ta tiến hành bỏ biến *ceoten* và hồi quy mô hình với hai biến còn lại kết quả thu được từ phần mềm R như hình 2.1.5:

```
> new_train = train[c("salary", "lmktval", "profmarg")]
> newModel = lm(formula = new_train$salary ~ ., data = new_train)
> summary(newModel)

Call:
lm(formula = new_train$salary ~ ., data = new_train)

Residuals:
    Min       1Q   Median       3Q      Max
-1127.7  -256.6   -85.3   246.7  4404.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -830.739    298.886   -2.779  0.0062 **
lmktval       245.323     39.252    6.250  4.71e-09 ***
profmarg     -13.944      6.607   -2.111  0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 544.2 on 139 degrees of freedom
Multiple R-squared:  0.2338,    Adjusted R-squared:  0.2228
F-statistic: 21.21 on 2 and 139 DF,  p-value: 9.143e-09
```

Hình 2.1.5: Kết quả hồi quy mô hình với hai biến còn lại

Mô hình thống kê mới:

$$salary = -830.7 + 245.3 * lmktval - 13.9 * profmarg \quad (2.1.3)$$

Trường hợp này hai biến còn lại có ý nghĩa thống kê. Tuy nhiên mô hình được tạo bởi hai biến này chỉ giải thích được 23% sự biến thiên của biến phụ thuộc (hình 2.1.5). Nguyên nhân dẫn tới kết quả thấp là do số lượng data ít, các biến giải thích ít không tạo nên mô hình đặc trưng được.

Test trên tập test và nhận xét kết quả

Thực hiện dự đoán trên tập dữ liệu test từ kết quả mô hình 2.1.3 và dùng chỉ số đánh giá MSE (trung bình bình phương sai số) ta có:

```
> SE = sum((pred_test-y_test) ^2)
> SE
[1] 15893414
> MSE = SE / nrow(test)
> print(MSE)
[1] 454097.5
```

Hình 2.1.6: Chỉ số đo lường kết quả MSE

Kết quả $MSE \approx 454097$ lớn hơn nhiều so với giá trị Mean : 887.5 nên ta có thể thấy hai yếu tố gồm: giá thị trường (*lmktval*) và tỷ lệ phần trăm lợi nhuận (*profmarg*) là chưa đủ để giải thích mức độ tăng giảm của tiền lương của các giám đốc điều hành các tập đoàn Hoa Kỳ.

Để cải thiện kết quả mô hình ta nên tiến hành thu thập thêm dữ liệu và tiến hành lựa chọn biến dựa trên dữ liệu mới này. Bên cạnh đó có thể xem xét tới xem xét tới các nhân tố khác ảnh hưởng tới tiền lương của các giám đốc Hoa kỳ như: Lĩnh vực hoạt động (ngân hàng, hàng không, công nghệ, vận tải...); mức lương trước đó; số năm kinh nghiệm, giới tính,...

2.2 Dữ liệu 2

Bộ dữ liệu ghi lại lịch sử về những ngôi nhà được bán từ 5/2014 đến 5/2015 ở quận King, bang Washington, Hoa Kỳ. Bộ dữ liệu bao gồm 21613 quan trắc, gồm 21 biến.

* Phương pháp chọn: Stepwise - lùi; tiêu chuẩn chọn: BIC.

Tìm hiểu dữ liệu

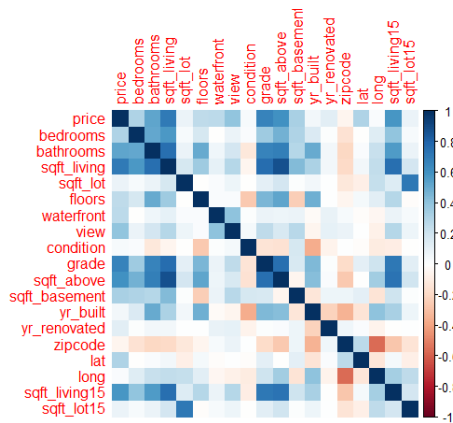
```
> mydata <- read.csv("data2.csv")
> head(mydata)
```

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
1	7129300520	10/13/2014	221900	3	1.00	1180	5650	1	0	0
2	6414100192	12/9/2014	538000	3	2.25	2570	7242	2	0	0
3	5631500400	2/25/2015	180000	2	1.00	770	10000	1	0	0
4	2487200875	12/9/2014	604000	4	3.00	1960	5000	1	0	0
5	1954400510	2/18/2015	510000	3	2.00	1680	8080	1	0	0
6	7237550310	5/12/2014	1230000	4	4.50	5420	101930	1	0	0

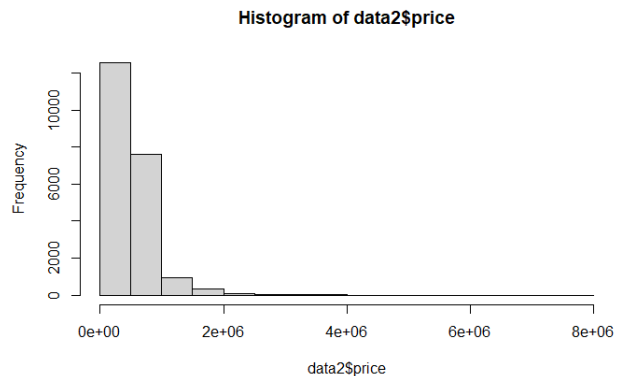
	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long
1	3	7	1180	0	1955	0	98178	47.5112	-122.257
2	3	7	2170	400	1951	1991	98125	47.7210	-122.319
3	3	6	770	0	1933	0	98028	47.7379	-122.233
4	5	7	1050	910	1965	0	98136	47.5208	-122.393
5	3	8	1680	0	1987	0	98074	47.6168	-122.045
6	3	11	3890	1530	2001	0	98053	47.6561	-122.005

	sqft_living15	sqft_lot15
1	1340	5650
2	1690	7639
3	2720	8062
4	1360	5000
5	1800	7503
6	4760	101930

(a) Một số quan trắc đầu tiên



(b) Hệ số tương quan giữa các biến



(c) Phân bố của biến phụ thuộc

Hình 2.2.1: Một số quan sát ban đầu của bộ dữ liệu

Bộ dữ liệu cung cấp gồm 21 biến, trong đó biến **id** và **date** được loại bỏ khỏi dữ liệu trước khi tiến hành phân tích, vì nhóm em nghĩ các biến này chỉ để ghi lại chỉ số và thời gian mua bán, không mang nhiều ý nghĩa thống kê.

Quan sát ban đầu cho thấy: các biến độc lập **bathrooms** (số phòng tắm), **sqft_living**

(diện tích căn nhà), **grade** (điểm số đánh giá), **sqft_ above** (diện tích ngoài tầng hầm), **sqft_ living15** (diện tích ngôi nhà vào năm 2015) có mối tương quan cao với biến phụ thuộc **Price** - giá nhà; biến phụ thuộc **Price** phân bố không đều, bị lệch hẳn về một phía và giá trị chủ yếu từ 0 đến 2 000 000.

Phân tích, chọn mô hình

```
> # Create full model
> mod_full_1 = lm(price ~ ., data2) #full model
> summary(mod_full_1)
```

Call:
lm(formula = price ~ ., data = data2)

Residuals:

Min	1Q	Median	3Q	Max
-1291631	-99089	-9569	77778	4330096

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.564e+06	2.933e+06	2.238	0.02523 *
bedrooms	-3.556e+04	1.901e+03	-18.707	< 2e-16 ***
bathrooms	4.128e+04	3.268e+03	12.632	< 2e-16 ***
sqft_living	1.496e+02	4.397e+00	34.033	< 2e-16 ***
sqft_lot	1.289e-01	4.792e-02	2.690	0.00714 **
floors	6.474e+03	3.602e+03	1.797	0.07229 .
waterfront	5.833e+05	1.736e+04	33.593	< 2e-16 ***
view	5.278e+04	2.141e+03	24.652	< 2e-16 ***
condition	2.679e+04	2.353e+03	11.387	< 2e-16 ***
grade	9.701e+04	2.161e+03	44.894	< 2e-16 ***
sqft_above	3.129e+01	4.361e+00	7.174	7.53e-13 ***
sqft_basement	NA	NA	NA	NA
yr_built	-2.628e+03	7.272e+01	-36.135	< 2e-16 ***
yr_renovated	1.983e+01	3.656e+00	5.425	5.87e-08 ***
zipcode	-5.819e+02	3.299e+01	-17.635	< 2e-16 ***
lat	6.022e+05	1.074e+04	56.071	< 2e-16 ***
long	-2.156e+05	1.316e+04	-16.385	< 2e-16 ***
sqft_living15	2.116e+01	3.451e+00	6.131	8.88e-10 ***
sqft_lot15	-3.907e-01	7.334e-02	-5.327	1.01e-07 ***

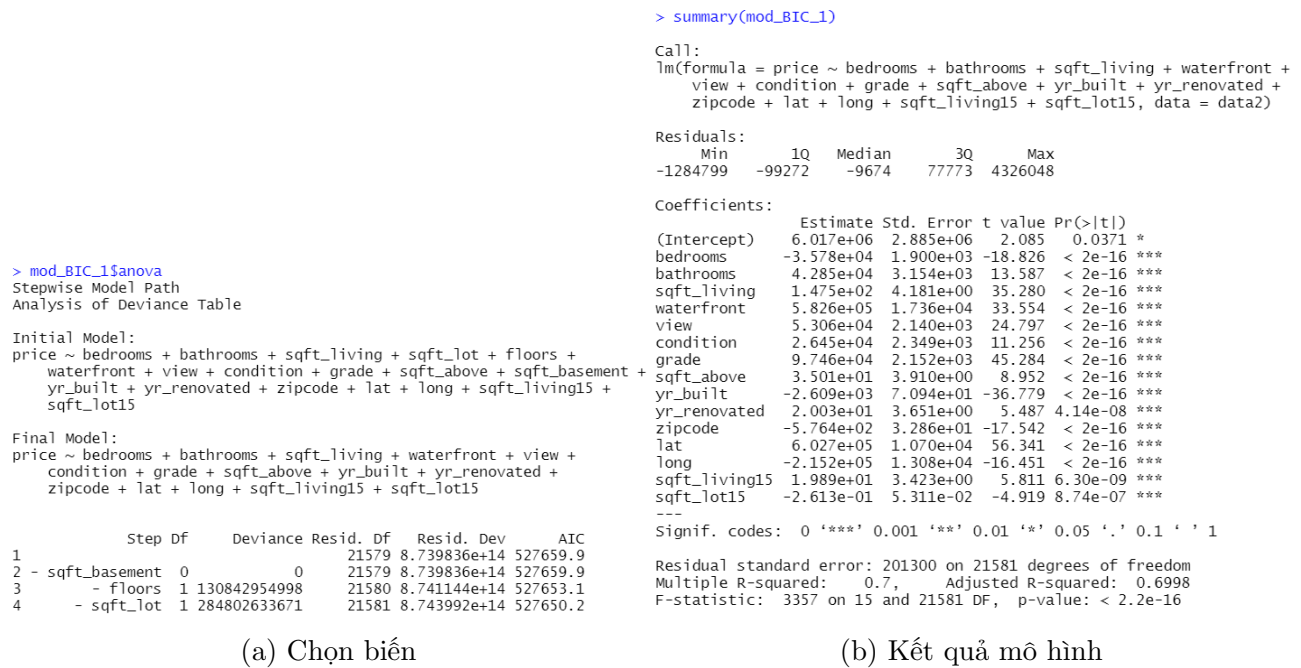
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 201300 on 21579 degrees of freedom
Multiple R-squared: 0.7001, Adjusted R-squared: 0.6999
F-statistic: 2964 on 17 and 21579 DF, p-value: < 2.2e-16

Hình 2.2.2: Mô hình hồi quy đầy đủ ban đầu

Bộ dữ liệu (sau khi loại bỏ id và date) có 18 biến giải thích, do đó nhóm em chọn phương pháp lùi (**stepwise - backward**) cho bộ dữ liệu này. Trong mô hình hồi quy đầy đủ (Hình 2.2.2), đa số các biến giải thích đều có ý nghĩa thống kê, do đó tiến hành phương pháp lùi (loại biến dần dần) sẽ tiết kiệm thời gian hơn so với các phương pháp còn lại. Tiêu chuẩn BIC có xu hướng chọn các mô hình ít phức tạp hơn so với tiêu chuẩn

AIC, đặc biệt khi số lượng quan trắc lớn.



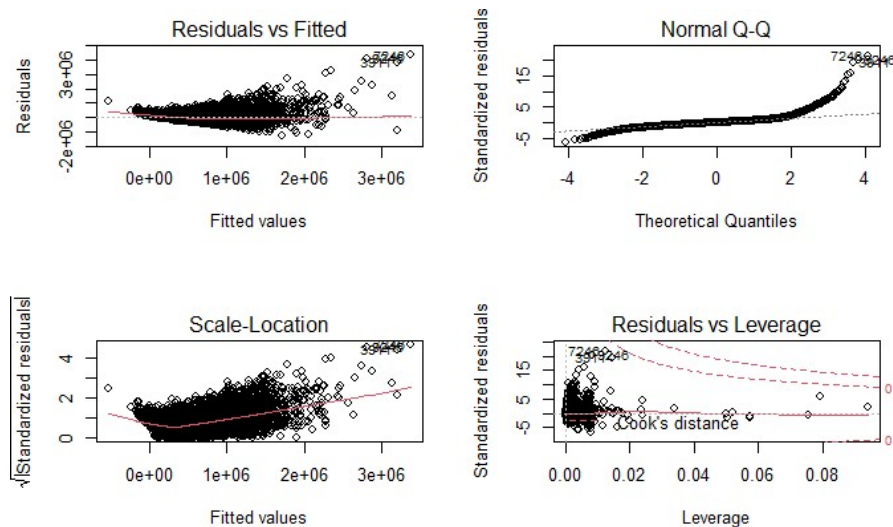
(a) Chọn biến

(b) Kết quả mô hình

Hình 2.2.3: Mô hình khi chọn bằng tiêu chuẩn BIC

Bằng phương pháp lùi và tiêu chuẩn BIC (Hình 2.2.3), các biến **sqft_basement**, **floors**, **sqft_lot** đã bị loại bỏ khỏi mô hình. Mô hình được chọn có $R^2 = 0.7$, $R^2_{adj} = 0.69$, các tham số ước lượng của mô hình đều có ý nghĩa thống kê.

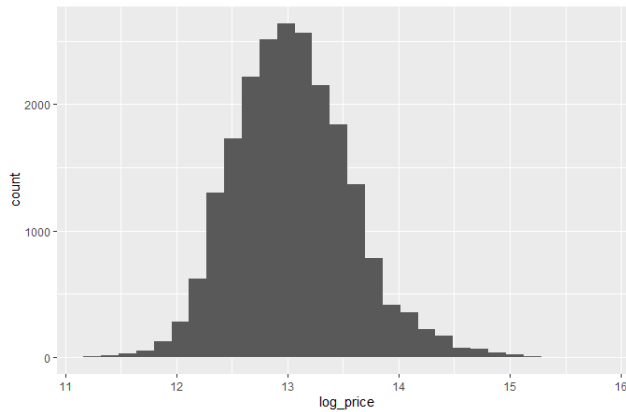
Ta tiến hành kiểm tra xem mô hình này có thỏa mãn các giả thiết của mô hình hồi quy hay không.



Hình 2.2.4: Các biểu đồ kiểm định mô hình

Dựa vào hình 2.2.4, phương sai của sai số không phải là hằng số, kì vọng của sai số bằng 0; sai số có vẻ tuân theo phân phối chuẩn nhưng phần đuôi trên bị lệch khá nhiều.

Kết hợp với nhận xét ban đầu, về việc biến **Price** phân bố không đều, nhóm em tiến hành biến đổi biến này thành $\log(\text{Price})$.



Hình 2.2.5: Phân bố của biến **Price** sau khi biến đổi

Sau khi biến đổi, ta tiến hành hồi quy cho: **mô hình 1** mô hình có 15 biến đã chọn bằng tiêu chuẩn BIC trước đó, và **mô hình 2** mô hình đầy đủ rồi áp dụng tiêu chuẩn BIC để chọn biến.

> summary(mod_2)

```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + sqft_living +
  waterfront + view + condition + grade + sqft_above + yr_built +
  yr_renovated + zipcode + lat + long + sqft_living15 + sqft_lot15,
  data = data2)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.72685	-0.16385	0.00299	0.16386	1.18219

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.436e+01	3.645e+00	-3.940	8.18e-05 ***
bedrooms	-1.351e-02	2.400e-03	-5.629	1.83e-08 ***
bathrooms	8.720e-02	3.984e-03	21.891	< 2e-16 ***
sqft_living	1.238e-04	5.282e-06	23.444	< 2e-16 ***
waterfront	3.702e-01	2.193e-02	16.881	< 2e-16 ***
view	6.195e-02	2.703e-03	22.919	< 2e-16 ***
condition	5.984e-02	2.968e-03	20.163	< 2e-16 ***
grade	1.643e-01	2.719e-03	60.449	< 2e-16 ***
sqft_above	2.582e-05	4.939e-06	5.228	1.73e-07 ***
yr_built	-3.126e-03	8.960e-05	-34.882	< 2e-16 ***
yr_renovated	4.008e-05	4.612e-06	8.690	< 2e-16 ***
zipcode	-5.816e-04	4.150e-05	-14.014	< 2e-16 ***
lat	1.414e+00	1.351e-02	104.612	< 2e-16 ***
long	-1.741e-01	1.652e-02	-10.537	< 2e-16 ***
sqft_living15	8.802e-05	4.324e-06	20.355	< 2e-16 ***
sqft_lot15	1.512e-07	6.709e-08	2.254	0.0242 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2543 on 21581 degrees of freedom
Multiple R-squared: 0.767, Adjusted R-squared: 0.7668
F-statistic: 4736 on 15 and 21581 DF, p-value: < 2.2e-16

(a) Mô hình 1

> summary(mod_BIC_2)

```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + sqft_living +
  sqft_lot + floors + waterfront + view + condition + grade +
  yr_built + yr_renovated + zipcode + lat + long + sqft_living15,
  data = data2)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.7953	-0.1615	0.0037	0.1590	1.1735

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.932e+00	3.639e+00	-1.905	0.0568 .
bedrooms	-1.174e-02	2.382e-03	-4.930	8.27e-07 ***
bathrooms	7.137e-02	4.047e-03	17.634	< 2e-16 ***
sqft_living	1.403e-04	4.197e-06	33.431	< 2e-16 ***
sqft_lot	3.426e-07	4.355e-08	7.868	3.78e-15 ***
floors	6.979e-02	4.049e-03	17.234	< 2e-16 ***
waterfront	3.686e-01	2.176e-02	16.937	< 2e-16 ***
view	6.148e-02	2.649e-03	23.205	< 2e-16 ***
condition	6.352e-02	2.941e-03	21.594	< 2e-16 ***
grade	1.591e-01	2.682e-03	59.299	< 2e-16 ***
yr_built	-3.419e-03	9.120e-05	-37.494	< 2e-16 ***
yr_renovated	3.650e-05	4.585e-06	7.962	1.78e-15 ***
zipcode	-6.441e-04	4.137e-05	-15.569	< 2e-16 ***
lat	1.404e+00	1.337e-02	104.988	< 2e-16 ***
long	-1.715e-01	1.619e-02	-10.590	< 2e-16 ***
sqft_living15	9.566e-05	4.278e-06	22.359	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2524 on 21581 degrees of freedom
Multiple R-squared: 0.7703, Adjusted R-squared: 0.7702
F-statistic: 4826 on 15 and 21581 DF, p-value: < 2.2e-16

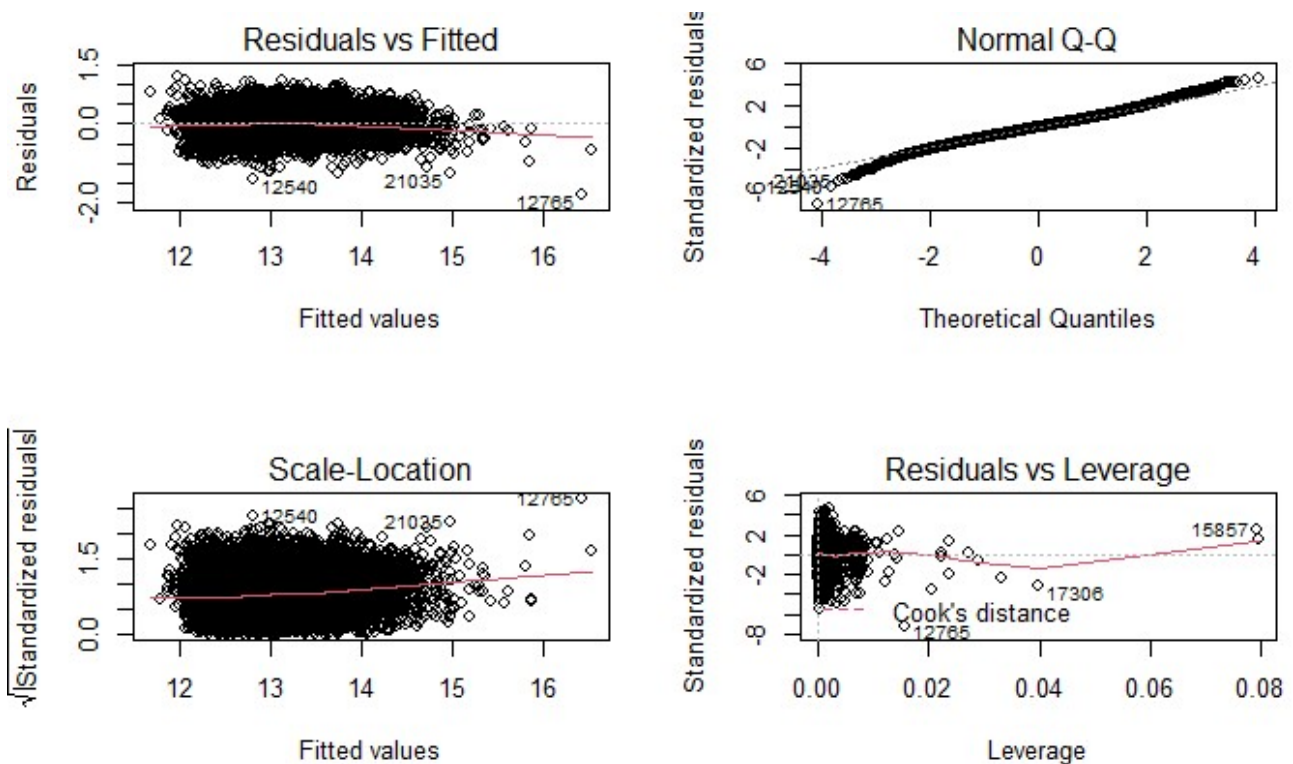
(b) Mô hình 2

Hình 2.2.6: Kết quả khi biến đổi **Price** thành $\log(\text{Price})$

Cả hai mô hình đều gồm 15 biến giải thích, mô hình 2 đã loại bỏ các biến **sqft_basement**, **sqft_above**, **sqft_lot15** khác với 3 biến đã loại trước khi biến đổi **Price**.

Nhóm em chọn **mô hình 2** là mô hình cuối cùng, vì: mô hình 2 có hệ số xác định lớn hơn ($R^2 = 77.03\%$), các biến liên quan đến diện tích tầng hầm (**sqft_basement**, **sqft_above**) đã được bao gồm trong **sqft_living**, diện tích khu đất vào năm 2015 cũng không mang nhiều ý nghĩa thống kê trong mô hình 1 nên có thể loại bỏ.

Kiểm tra giả thiết mô hình 2: phương sai của sai số không thay đổi, kì vọng bằng 0 và đã tuân theo phân phối chuẩn, chưa phát hiện hiện tượng đa cộng tuyến trong mô hình (các chỉ số $VIF < 5$) (Hình 2.2.7).



(a) Các biểu đồ kiểm định

```
> vif(mod_BIC_2)
```

bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
1.649892	3.283149	5.032386	1.102209	1.618698	1.202269
view	condition	grade	yr_built	yr_renovated	zipcode
1.397185	1.241008	3.356002	2.432360	1.150352	1.661301
lat	long	sqft_living15			
1.163852	1.759544	2.912665			

(b) Kiểm tra đa cộng tuyến

Hình 2.2.7: Kết quả khi biến đổi thành $\log(\text{Price})$

Vậy **mô hình cuối cùng được chọn** có các hệ số ước lượng như hình 2.2.8.

```
> coef(mod_BIC_2)
(Intercept) bedrooms bathrooms sqft_living sqft_lot floors
-6.932157e+00 -1.174353e-02 7.137346e-02 1.403104e-04 3.426024e-07 6.978707e-02
waterfront view condition grade yr_built yr_renovated
3.685686e-01 6.147550e-02 6.351646e-02 1.590506e-01 -3.419313e-03 3.650388e-05
zipcode lat long sqft_living15
-6.441469e-04 1.404181e+00 -1.714684e-01 9.565513e-05
```

Hình 2.2.8: Hệ số mô hình được chọn

$$\begin{aligned}\log(\text{Price}) = & -6.93 - 0.011 \times \text{bedrooms} + 0.071 \times \text{bathrooms} + 1.403 \times 10^{-4} \times \text{sqft_living} \\ & + 3.426 \times 10^{-7} \times \text{sqft_lot} + 0.069 \times \text{floors} + 0.36 \times \text{waterfront} + 0.061 \times \text{view} \\ & + 0.063 \times \text{condition} + 0.159 \times \text{grade} - 3.4196 \times 10^{-3} \times \text{yr_built} \\ & + 3.650 \times 10^{-5} \times \text{yr_renovated} - 6.441 \times 10^{-4} \times \text{zipcode} + 1.404 \times \text{lat} \\ & - 0.171 \times \text{long} + 9.565.171 \times 10^{-5} \times \text{sqft_living15}\end{aligned}$$

Kết luận

Có 77.06% sự biến thiên của giá nhà ở quận King được giải thích bởi 15 biến độc lập, trong đó các yếu tố ảnh hưởng nhiều nhất gồm *số phòng ngủ, số phòng tắm, diện tích nhà, số tầng, hướng nhà ra bờ sông, tình trạng ngôi nhà (mới/cũ), điểm theo phân loại của quận, vị trí (kinh độ - vĩ độ), năm xây dựng*.

Giá trị của một căn nhà **không bị ảnh hưởng nhiều** bởi các yếu tố: diện tích tầng hầm, diện tích khu đất, diện tích ngoài tầng hầm, năm sửa chữa căn nhà, zipcode (mã vùng) của ngôi nhà. Diện tích của căn nhà cũng có ảnh hưởng, tuy nhiên sự ảnh hưởng là không nhiều.

Số phòng ngủ có mối tương quan nghịch với giá nhà, vì khi số phòng ngủ tăng lên, nhưng các yếu tố còn lại không thay đổi, thì diện tích của mỗi phòng ngủ sẽ giảm đi, gây cảm giác chật chội.

Nhìn vào các kết quả hình 2.2.7, vẫn thấy có nhiều điểm ngoại lai (**outlier**), hướng nghiên cứu tiếp theo có thể loại bỏ những điểm này ra khỏi bộ dữ liệu, tiến hành quan sát riêng để rút ra thêm các kết luận khác (nếu có).

2.3 Dữ liệu 3

Bộ dữ liệu ghi lại tỷ lệ tai nạn, gồm 39 quan trắc được thực hiện trên vài đoạn đường cao tốc ở tiểu bang Minnesota vùng Trung Tây của Hoa Kỳ.

* **Phương pháp chọn: Stepwise từng bước; Tiêu chuẩn chọn: BIC.**

Tìm hiểu dữ liệu

Bộ dữ liệu gồm 1 biến phụ thuộc và 13 biến giải thích sau:

- Y : tỷ lệ % tai nạn trên đoạn đường khảo sát.
- X_1 : chiều dài đoạn đường (dặm).
- X_2 : lượng giao thông trung bình hàng ngày (nghìn xe).
- X_3 : tỷ lệ % xe tải trên tổng số.
- X_4 : tốc độ giới hạn cho phép (dặm/giờ).
- X_5 : chiều rộng làn đường (bước chân).
- X_6 : chiều rộng làn đường khẩn cấp (bước chân).
- X_7 : số làn đường thay đổi tự do trên đoạn đường cao tốc.
- X_8 : số làn đường thay đổi (báo hiệu) trên đoạn đường cao tốc.
- X_9 : số cửa vào đoạn đường cao tốc.
- X_{10} : tổng số làn đường (trên hai chiều của đường cao tốc).
- X_{11} : 1 nếu là tuyến đường liên thông xa lộ và cao tốc, 0 nếu ngược lại.
- X_{12} : 1 nếu là tuyến đường lớn của cao tốc, 0 nếu ngược lại.
- X_{13} : 1 nếu là tuyến đường cao tốc chính, 0 nếu ngược lại.

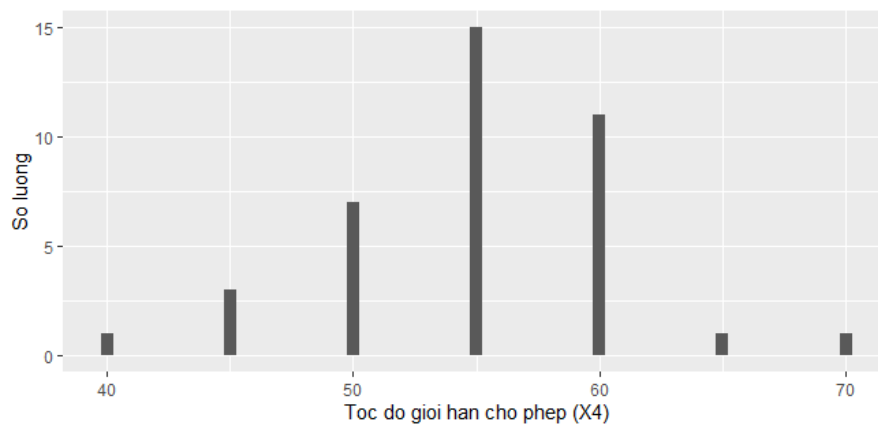
Một vài quan trắc đầu tiên trong bộ dữ liệu được thể hiện trong hình 2.3.1.

```
> head(data3)
  x_i.1 x_i.2 x_i.3 x_i.4 x_i.5 x_i.6 x_i.7 x_i.8 x_i.9 x_i.10 x_i.11 x_i.12 x_i.13 y_i
1  4.99   69    8   55   12   10  1.20  0.00  4.6    8    1    0    0  4.58
2 16.11   73    8   60   12   10  1.43  0.00  4.4    4    1    0    0  2.86
3  9.75   49   10   60   12   10  1.54  0.00  4.7    4    1    0    0  3.02
4  1.65   61   13   65   12   10  0.94  0.00  3.8    6    1    0    0  2.29
5 20.01   28   12   70   12   10  0.65  0.00  2.2    4    1    0    0  1.61
6  5.97   30    6   55   12   10  0.34  1.84 24.8    4    0    1    0  6.87
```

Hình 2.3.1: Một vài quan trắc đầu tiên

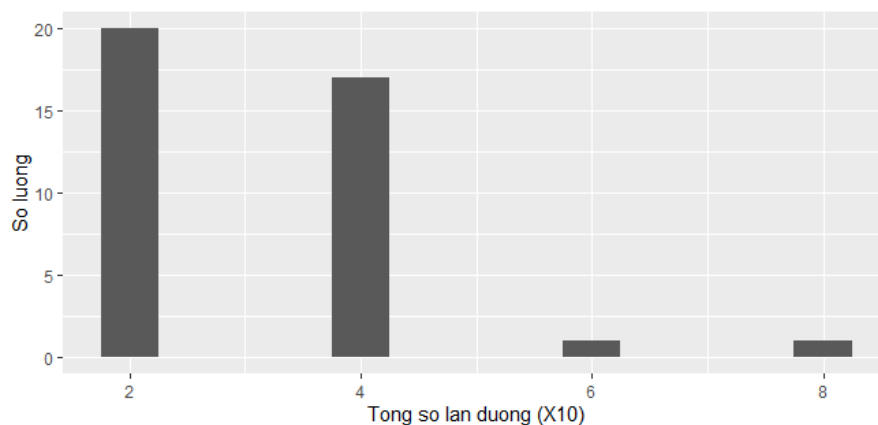
Một số phân bố theo biến:

- X_4 : Có 33 trong 39 quan trắc có tốc độ tối đa là 50, 55 và 60 (hình 2.3.2).



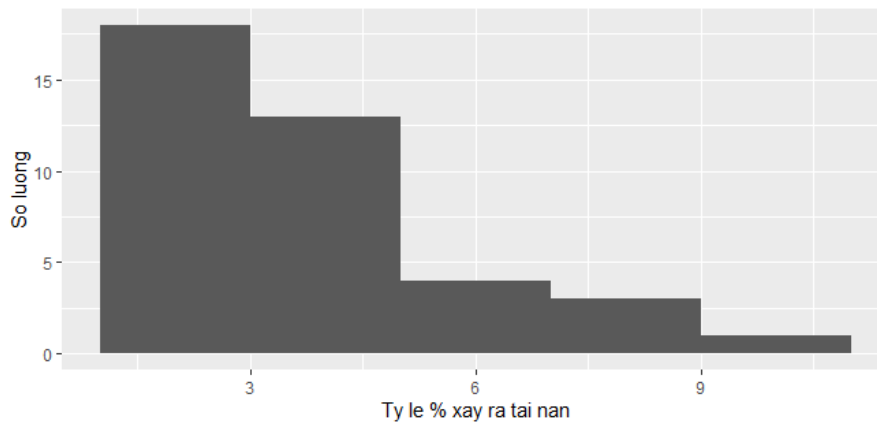
Hình 2.3.2: Phân bố theo tốc độ giới hạn cho phép (X_4) (dặm/giờ)

- X_{10} : Có 32 trong 39 quan trắc có tổng số làn đường là 2 hoặc 4 (hình 2.3.3).



Hình 2.3.3: Phân bố theo tổng số làn đường (X_{10})

- Y : Phần lớn tỷ lệ tai nạn là 1 – 5% (hình 2.3.4).



Hình 2.3.4: Phân bố theo tỷ lệ % tai nạn (Y)

Trung bình của tổng tỷ lệ tai nạn theo các loại tuyến đường (hình 2.3.5) cho thấy loại tuyến đường cao tốc chính có tỷ lệ tai nạn cao nhất.

```
> aggregate(y_i ~ x_i.11 + x_i.12 + x_i.13, data3, mean)
  x_i.11 x_i.12 x_i.13      y_i
1      0      0      0 3.585000
2      1      0      0 2.872000
3      0      1      0 3.608421
4      0      0      1 4.870000
```

Hình 2.3.5: Trung bình của tổng tỷ lệ tai nạn theo các loại tuyến đường

Trung bình của tổng tỷ lệ % tai nạn theo các mức tốc độ giới hạn cho phép (hình 2.3.6) cho thấy giới hạn tốc độ cho phép trên đường cao tốc càng thấp thì xảy ra tai nạn càng nhiều, tỷ lệ tai nạn giảm dần khi giới hạn tốc độ cho phép tăng.

```
> aggregate(y_i ~ x_i.4, data3, mean)
  x_i.4      y_i
1    40 9.230000
2    45 7.283333
3    50 4.055714
4    55 3.985333
5    60 2.750000
6    65 2.290000
7    70 1.610000
```

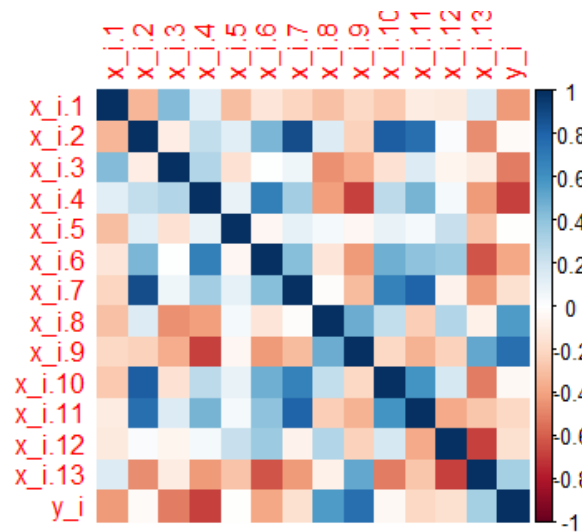
Hình 2.3.6: Trung bình của tổng tỷ lệ % tai nạn theo các mức tốc độ giới hạn cho phép

Trung bình của tổng tỷ lệ % tai nạn theo tổng số làn đường (hình 2.3.7) cho thấy trên đoạn đường có 8 làn đường có tỷ lệ xảy ra tai nạn cao nhất, kể đến là đoạn đường có 2 làn.

```
> aggregate(y_i ~ x_i.10, data3, mean)
  x_i.10    y_i
1      2 4.000500
2      4 3.912941
3      6 2.290000
4      8 4.580000
```

Hình 2.3.7: Trung bình của tổng tỷ lệ % tai nạn theo tổng số làn đường

Ma trận ở hình 2.3.8 thể hiện độ tương quan giữa các biến, cho thấy tốc độ giới hạn cho phép (X_4) có tương quan nghịch và số cửa đoạn đường cao tốc (X_9) có tương quan thuận đối với tỷ lệ % tai nạn (Y).



Hình 2.3.8: Ma trận tương quan giữa các biến

Phân tích, chọn mô hình

Đầu tiên, ta xét mô hình đầy đủ có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \epsilon \quad (2.3.1)$$

Mô hình hồi quy đầy đủ có các thông số ở hình 2.3.9, ta thấy được gần như tất cả 13 biến đều không có ý nghĩa thống kê. Ta tiến hành kiểm tra hiện tượng đa cộng tuyến có trong mô hình này sử dụng phương pháp tính hệ số VIF. Kết quả ở hình 2.3.10 cho thấy hiện tượng đa cộng tuyến xảy ra nặng nề giữa các biến, có 7/13 biến giải thích vượt ngưỡng chấp nhận được với hệ số VIF là 5 theo quy ước chung.

```
> summary(mod_full)

Call:
lm(formula = y_i ~ ., data = data3)

Residuals:
    Min       1Q   Median       3Q      Max
-2.00773 -0.63409 -0.04212  0.63969  2.53722

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.7129031   6.9126865   1.984  0.0584 .
x_i.1       -0.0589293   0.0314673  -1.873  0.0728 .
x_i.2       -0.0054182   0.0337952  -0.160  0.8739
x_i.3       -0.1106588   0.1134557  -0.975  0.3387
x_i.4       -0.1266860   0.0817554  -1.550  0.1338
x_i.5       -0.1196817   0.5985717  -0.200  0.8431
x_i.6        0.0183357   0.1628311   0.113  0.9112
x_i.7       -0.3882033   1.1811637  -0.329  0.7451
x_i.8        0.7087845   0.5245588   1.351  0.1887
x_i.9        0.0654378   0.0427391   1.531  0.1383
x_i.10       0.0006672   0.2864299   0.002  0.9982
x_i.11       0.5033821   1.7304348   0.291  0.7735
x_i.12      -0.9602033   1.1124585  -0.863  0.3963
x_i.13      -0.5605308   0.9784518  -0.573  0.5718
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.202 on 25 degrees of freedom
Multiple R-squared:  0.7589,    Adjusted R-squared:  0.6335
F-statistic: 6.053 on 13 and 25 DF,  p-value: 6.176e-05
```

Hình 2.3.9: Mô hình hồi quy đầy đủ ban đầu

```
> vif(mod_full)
      x_i.1      x_i.2      x_i.3      x_i.4      x_i.5      x_i.6      x_i.7      x_i.8      x_i.9
1.588934 10.400300  1.875948  6.011025  1.957449  6.426287  6.226508  2.901934  4.169496
      x_i.10     x_i.11     x_i.12     x_i.13
3.993268  9.029566  8.341780  5.739884
```

Hình 2.3.10: Hiện tượng đa cộng tuyến giữa các biến trong mô hình

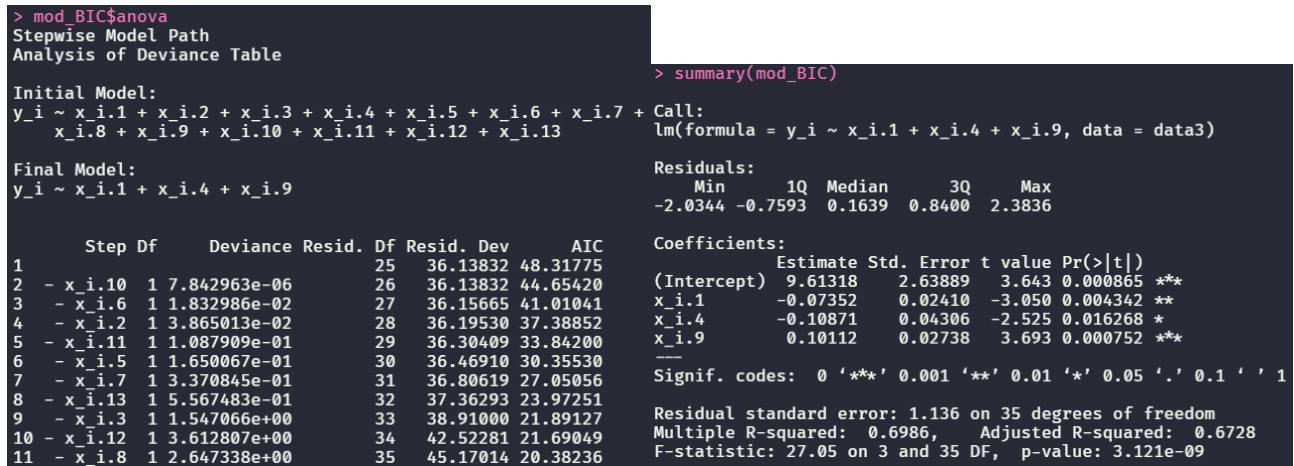
Vì số lượng biến giải thích khá ít, chỉ có 13 biến và có hiện tượng đa cộng tuyến, nên nhóm em sử dụng phương pháp hồi quy Stepwise từng bước để dễ dàng thêm bớt các biến khi chọn mô hình. Đối với tiêu chuẩn đánh giá mô hình, vì bộ dữ liệu này có cỡ mẫu nhỏ, chỉ có 39 quan trắc, nên nhóm em dùng tiêu chuẩn BIC cho cỡ mẫu $n = 39$.

Dùng phần mềm R cho phương pháp Stepwise tiến lùi và tiêu chuẩn BIC, ta có kết quả ở hình 2.3.11, mô hình lựa chọn có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_9 X_9 + \epsilon \quad (2.3.2)$$

Trong quá trình chọn mô hình, đa số các biến đã bị loại bỏ hết chỉ trừ 3 biến X_1 , X_4 , và X_9 lần lượt giải thích cho chiều dài đoạn đường, tốc độ giới hạn cho phép và số cửa vào đoạn đường cao tốc. Mô hình 2.3.2 có hệ số xác định $R^2 = 0.6986$ và hệ số hiệu

chỉnh $R_{adj}^2 = 0.6728$, các tham số ước lượng của mô hình đều có ý nghĩa thống kê.



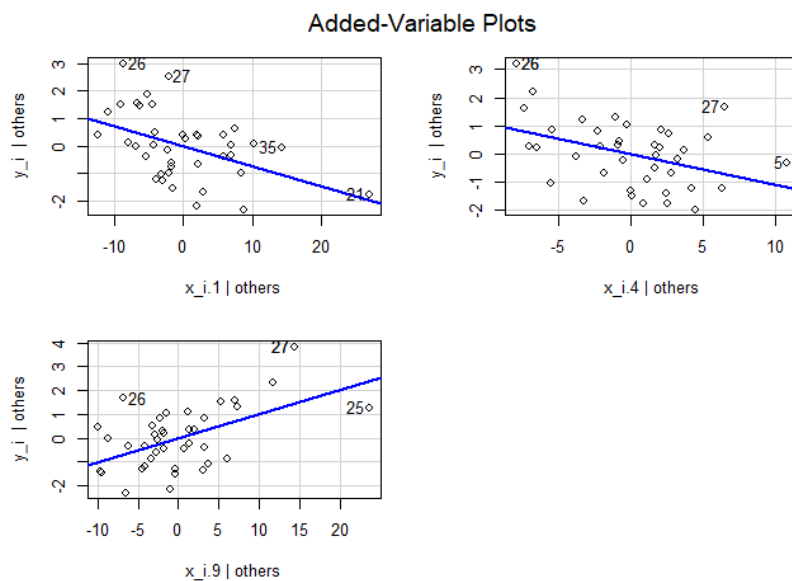
(a) Chọn biến

(b) Kết quả mô hình

Hình 2.3.11: Chọn mô hình với tiêu chuẩn BIC

Mô hình 2.3.2 giải thích được 69.86% sự biến thiên của tỷ lệ % tai nạn được giải thích bởi 3 biến độc lập. Các hệ số của mô hình lần lượt là: $\hat{\beta}_0 = 9.613$, $\hat{\beta}_1 = -0.073$, $\hat{\beta}_4 = -0.109$, $\hat{\beta}_9 = 0.101$.

Mối tương quan giữa từng biến giải thích trong mô hình và biến phụ thuộc có quan hệ tuyến tính được biểu diễn trong hình 2.3.12. Hiện tượng đa cộng tuyến giữa các biến cũng không còn tồn tại trong mô hình được biểu diễn trong hình 2.3.13.

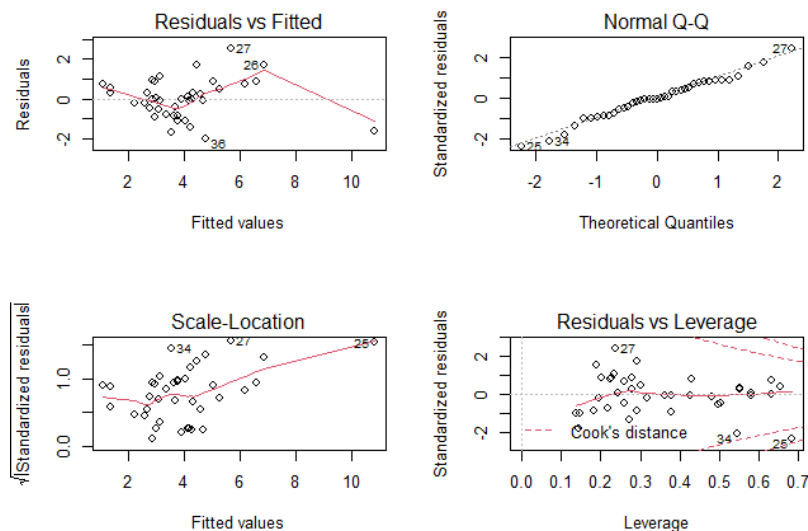


Hình 2.3.12: Mối tương quan giữa từng biến giải thích và biến phụ thuộc

```
> vif(mod_BIC)
      x_i.1      x_i.4      x_i.9
1.044222 1.867700 1.917150
```

Hình 2.3.13: Hiện tượng đa cộng tuyến giữa các biến trong mô hình được chọn

Tuy nhiên, biểu đồ phần dư ở hình 2.3.14 cho thấy mối liên quan giữa biến phụ thuộc và các biến giải thích không tuân theo hàm tuyến tính. Nhưng quan sát thấy có một số giá trị ngoại lai (*outlier*) tồn tại trong dữ liệu, nhóm em sử dụng phương pháp kiểm tra là tính dao động phần dư (*residuals*) và chuẩn hóa dữ liệu sao cho có trung bình 0 và phương sai 1, rồi từ đó tìm đối tượng nào có dao động phần dư chuẩn hóa cao hơn $|2|$.



Hình 2.3.14: Các biểu đồ của mô hình đầy đủ

Dùng phần mềm R tính toán, ta có kết quả ở hình 2.3.15, xác định được quan trắc thứ 26 và 27 là các giá trị ngoại lai.

```
> data_res = residuals(mod_BIC)
> data_crit = 2*sd(data_res)
> data_outlier = ifelse(abs(data_res>data_crit),1,0); data_outlier
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  0  0
33 34 35 36 37 38 39
0  0  0  0  0  0  0
```

Hình 2.3.15: Kiểm tra các giá trị ngoại lai trong mô hình đầy đủ

Ta thử loại bỏ các biến này và tiến hành chọn lại mô hình với phương pháp Stepwise

và tiêu chuẩn BIC, ta có kết quả từ phần mềm R ở hình 2.3.16. Mô hình lựa chọn thứ hai đã được thêm một biến X_8 là số làn đường thay đổi (báo hiệu) trên đoạn đường cao tốc, mô hình này có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_8 X_8 + \beta_9 X_9 + \epsilon \quad (2.3.3)$$

```
> summary(new_mod_full)
```

Call:
lm(formula = y_i ~ ., data = new_data)

Residuals:

	Min	1Q	Median	3Q	Max
	-1.53969	-0.54568	-0.05226	0.61573	1.73194

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.556e+01	5.880e+00	2.646	0.0145 *
x_i.1	-4.372e-02	2.658e-02	-1.645	0.1136
x_i.2	-5.889e-05	2.807e-02	-0.002	0.9983
x_i.3	-1.154e-01	9.466e-02	-1.219	0.2353
x_i.4	-8.676e-02	7.006e-02	-1.238	0.2281
x_i.5	-4.560e-01	5.047e-01	-0.903	0.3756
x_i.6	-1.074e-01	1.395e-01	-0.770	0.4493
x_i.7	-3.986e-01	9.793e-01	-0.407	0.6877
x_i.8	3.788e-01	4.492e-01	0.843	0.4078
x_i.9	8.194e-02	3.785e-02	2.165	0.0410 *
x_i.10	1.091e-01	2.397e-01	0.455	0.6533
x_i.11	4.498e-01	1.438e+00	0.313	0.7572
x_i.12	-5.255e-01	9.307e-01	-0.565	0.5778
x_i.13	-1.085e+00	8.237e-01	-1.317	0.2008

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9967 on 23 degrees of freedom
Multiple R-squared: 0.7878, Adjusted R-squared: 0.6678
F-statistic: 6.566 on 13 and 23 DF, p-value: 4.983e-05

```
> summary(new_mod_BIC)
```

Call:
lm(formula = y_i ~ x_i.1 + x_i.4 + x_i.8 + x_i.9, data = new_data)

Residuals:

	Min	1Q	Median	3Q	Max
	-1.5208	-0.6412	-0.1565	0.6998	1.5248

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.71314	2.41591	3.607	0.00104 **
x_i.1	-0.05079	0.02103	-2.415	0.02164 *
x_i.4	-0.09801	0.03904	-2.510	0.01731 *
x_i.8	0.56615	0.29097	1.946	0.06051 .
x_i.9	0.07244	0.02612	2.773	0.00918 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.946 on 32 degrees of freedom
Multiple R-squared: 0.734, Adjusted R-squared: 0.7007
F-statistic: 22.07 on 4 and 32 DF, p-value: 8.008e-09

(a) Mô hình hồi quy đầy đủ

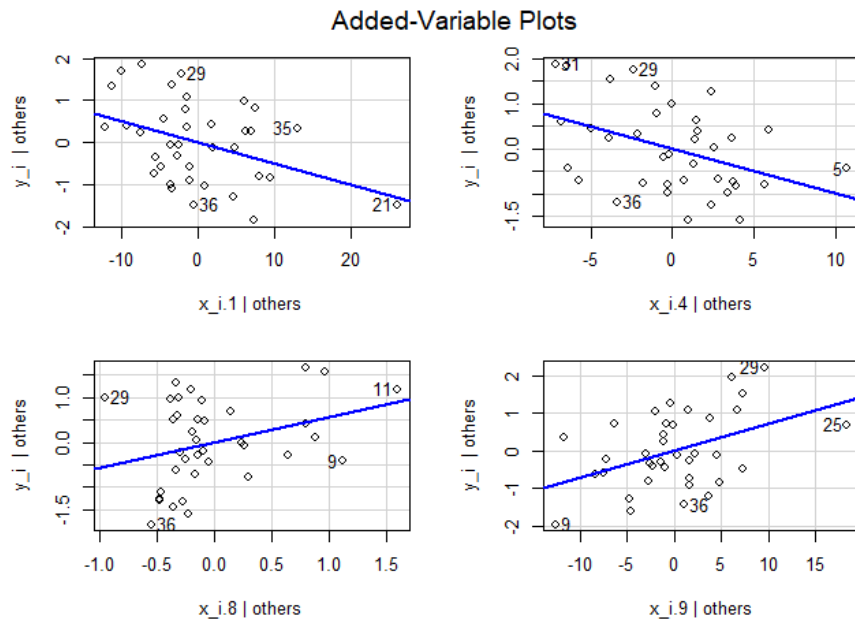
(b) Mô hình lựa chọn mới với tiêu chuẩn BIC

Hình 2.3.16: Mô hình đầy đủ và lựa chọn sau khi loại quan trắc 26, 27

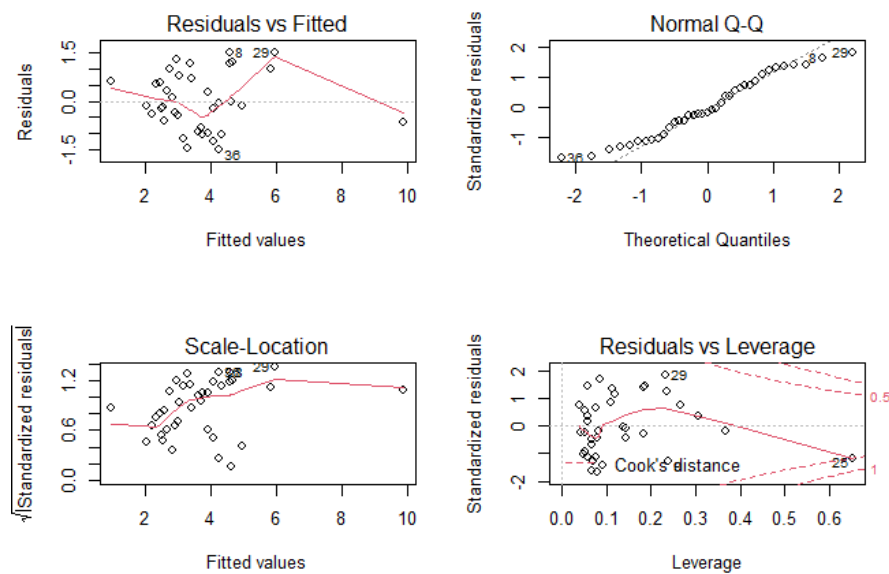
Ta nhận thấy tỷ lệ phần trăm sự biến thiên giải thích được của biến phụ thuộc:

- Đối với mô hình đầy đủ, có cải thiện từ 75.89% thành 78.78% và hệ số R^2 hiệu chỉnh cũng tăng tương đối từ 0.6335 lên 0.6678.
- Đối với mô hình mới 2.3.3, có cải thiện đáng kể từ 69.86% thành 73.4% và hệ số R^2 hiệu chỉnh cũng tăng tương đối từ 0.6728 lên 0.7007.

Dù vậy, các biến trong mô hình lựa chọn mới 2.3.3 lại kém có ý nghĩa thống kê hơn mô hình lựa chọn cũ. Nếu chúng ta dựa trên tỷ lệ phần trăm giải thích được cho mô hình thì mô hình mới vẫn là một lựa chọn không tồi. Các biến trong mô hình vẫn đảm bảo mối quan hệ tuyến tính với biến phụ thuộc và không xảy ra hiện tượng đa cộng tuyến theo chỉ số VIF (hình 2.3.17, 2.3.19), tuy nhiên, biểu đồ phần dư cũng không thay đổi nhiều so với mô hình cũ (hình 2.3.18).



Hình 2.3.17: Mối tương quan giữa từng biến giải tích và biến phụ thuộc



Hình 2.3.18: Các biểu đồ của mô hình lựa chọn mới

```
> vif(new_mod_BIC)
  x_i.1  x_i.4  x_i.8  x_i.9
1.088244 2.039255 1.413318 2.312290
```

Hình 2.3.19: Hiện tượng đa cộng tuyến giữa các biến trong mô hình được chọn

Kết luận

Khi kiểm tra các điều kiện ý nghĩa của mô hình:

- Vấn đề đa cộng tuyến trong cả hai mô hình lựa chọn đều được đảm bảo là không xảy ra.
- Tuy nhiên, phần dư ϵ trong cả hai mô hình đều không tuân theo phân phối chuẩn, kỳ vọng không bằng 0 và phương sai không là một hằng số.

Vậy **mô hình không có ý nghĩa**, tuy rằng mô hình lựa chọn cuối cùng đã có thể giải thích 73.4% phương sai của biến phụ thuộc Y , nói cách khác, có 73.4% phần trăm sự biến thiên của tỷ lệ tai nạn (Y) có thể được giải thích bởi chiều dài đoạn đường ($X1$), tốc độ giới hạn cho phép ($X4$), số làn đường thay đổi tự do trên đoạn đường cao tốc và số cửa vào đường cao tốc.

Mặt khác, dù giải pháp loại bỏ giá trị ngoại lai là cần thiết, nhưng vì dữ liệu quá ít, lý do vì sao bộ dữ liệu có những giá trị ngoại lai này vẫn chưa thể giải thích được chúng có thật sự là giá trị ngoại lai. Vì vậy, chúng ta cần nhiều dữ liệu hơn để mô hình có thể cho kết quả hồi quy tốt và chính xác hơn.

2.4 Dữ liệu 4

Bộ dữ liệu ghi lại những yếu tố có thể ảnh hưởng đến lương (\$ giờ) của người lao động ở Anh năm 1976

Tìm hiểu dữ liệu

Bộ dữ liệu gồm 13 biến sau:

- *wage*: Số lượng trung bình một giờ
- *educ*: Số năm học
-

Vẽ dữ liệu

Phân tích dữ liệu

Dựa vào độ tương quan ta nhận thấy được ba biến *lwage*, *tenursq*, *expersq* trên 90 nên ta sẽ bỏ ba biến này ra khỏi bộ dữ liệu. Nên ta sẽ tiến hành kiểm định thử xem ta có thể bỏ được 3 biến này được hay không.

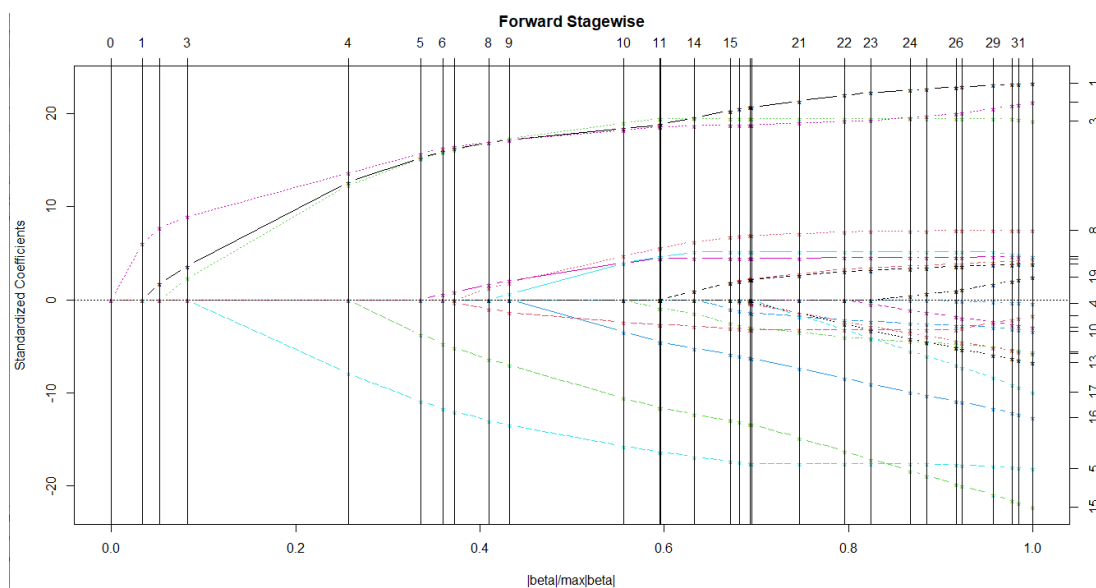
Chọn mô hình

Hướng tiếp cận 1: Chọn tất cả

Số lượng biến	Biến dự đoán
1	<i>profocc</i>
2	<i>educ</i> , <i>tenure</i>
3	<i>educ</i> , <i>tenure</i> , <i>female</i>
4	<i>educ</i> , <i>tenure</i> , <i>female</i> , <i>profocc</i>
5	<i>educ</i> , <i>tenure</i> , <i>femanle</i> , <i>profocc</i> , <i>trade</i>
6	<i>educ</i> , <i>tenure</i> , <i>female</i> , <i>profocc</i> , <i>trade</i> , <i>west</i>

Số lượng biến	Biến dự đoán
7	educ, tenure, female, profocc, trade, west, services
8	educ, tenure, female, profocc, trade, west, services, smsa
9	educ, tenure, female, profocc, trade, west, services, smsa, married
10	educ, tenure, female, profocc, trade, west, services, smsa, married, northcen
11	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv
12	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu
13	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen
14	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen, numdep
15	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen, numdep, exper
16	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen, numdep, exper, south
17	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen, numdep, exper, construc, clerocc
18	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen, numdep, exper, construc, clerocc, south
19	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen, numdep, exper, construc, clerocc, south, servocc
20	educ, tenure, female, profocc, trade, west, services, smsa, married, ndurman, profserv, trcommpu, northcen, numdep, exper, construc, clerocc, south, servocc, nonwhite

Dựa vào bảng ta thấy được mô hình có chỉ số BIC tốt nhất là mô hình 8 biến. Tuy nhiên mô hình có hệ số xác định hiệu chỉnh tốt nhất là mô hình có 15 biến và mô hình có hệ số C_p tốt nhất là mô hình 19 biến dự đoán



Hướng tiếp cận 2:: Phương pháp tiến dựa trên AIC

Sau khi chạy code R phương pháp lùi dựa theo tiêu chí AIC thì mô hình chọn được là mô hình gồm 11 biến

$$\text{wage} = \text{educ} + \beta_0 \text{tenure} + \text{female} + \text{married} + \text{smsa} + \text{northcen} + \text{west} + \text{ndurman} + \text{trcomm} + \text{pu} + \text{trade} + \text{ser}$$

Hướng tiếp cận 3: Phương pháp Stagewise

Dựa vào hình và bảng trên phương pháp Stagewise đưa ra đề xuất model gồm 11 biến

$$\text{wage} = \text{educ} + \text{tenure} + \text{female} + \text{married} + \text{smsa} + \text{west} + \text{trade} + \text{services} + \text{profocc} + \text{servocc}$$

Dựa vào cả ba hướng tiếp cận, ta sẽ xây dựng 5 mô hình hồi quy lại theo các phương pháp chọn biến

Dựa vào kết quả hồi quy của R chạy thì ta chọn mô hình 8 biến theo phương pháp chọn tất cả dựa vào chuẩn BIC. Tuy nhiên, các hệ số hồi quy vẫn chưa đạt được mức ý nghĩa thống kê nên ta tiến hành lấy log để xem thử

Listing 2.1: R output

```
> stepAIC(model1, direction = "backward")
Start:  AIC=1082.43
wage ~ (educ + exper + tenure + nonwhite + female + married +
        numdep + smsa + northcen + south + west + construc + ndurman +
        trcommpu + trade + services + profserv + profocc + clerocc +
        servocc + lwage + expersq + tenursq) - lwage - expersq -
        tenursq
```

	Df	Sum of Sq	RSS	AIC
- nonwhite	1	0.110	3802.2	1080.4
- servocc	1	1.962	3804.0	1080.7
- clerocc	1	3.244	3805.3	1080.9
- south	1	6.361	3808.4	1081.3
- construc	1	6.748	3808.8	1081.4
- exper	1	10.983	3813.0	1081.9
- numdep	1	12.966	3815.0	1082.2
- west	1	14.370	3816.4	1082.4
<none>			3802.1	1082.4
- married	1	16.721	3818.8	1082.7
- northcen	1	19.548	3821.6	1083.1
- trcommpu	1	25.582	3827.6	1084.0
- ndurman	1	27.832	3829.9	1084.3
- profserv	1	37.002	3839.1	1085.5
- smsa	1	47.870	3849.9	1087.0
- services	1	95.313	3897.4	1093.5
- profocc	1	208.622	4010.7	1108.5
- trade	1	211.656	4013.7	1108.9
- female	1	240.625	4042.7	1112.7
- tenure	1	249.799	4051.9	1113.9
- educ	1	300.257	4102.3	1120.4

Step: AIC=1080.44

```
wage ~ educ + exper + tenure + female + married + numdep + smsa +
        northcen + south + west + construc + ndurman + trcommpu +
        trade + services + profserv + profocc + clerocc + servocc
```

	Df	Sum of Sq	RSS	AIC
- servocc	1	1.894	3804.1	1078.7
- clerocc	1	3.348	3805.5	1078.9
- south	1	6.517	3808.7	1079.3
- construc	1	6.796	3809.0	1079.4
- exper	1	10.961	3813.1	1080.0
- numdep	1	12.858	3815.0	1080.2
- west	1	14.343	3816.5	1080.4
<none>			3802.2	1080.4
- married	1	17.080	3819.3	1080.8
- northcen	1	19.507	3821.7	1081.1
- trcommpu	1	25.474	3827.6	1082.0