STA 108 Final Project - CDI Data

Name: Abdul Shazif Nawaz

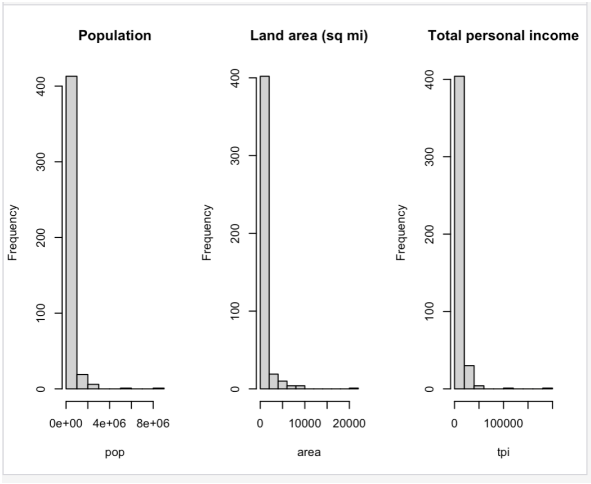
Date: December 5th 2025

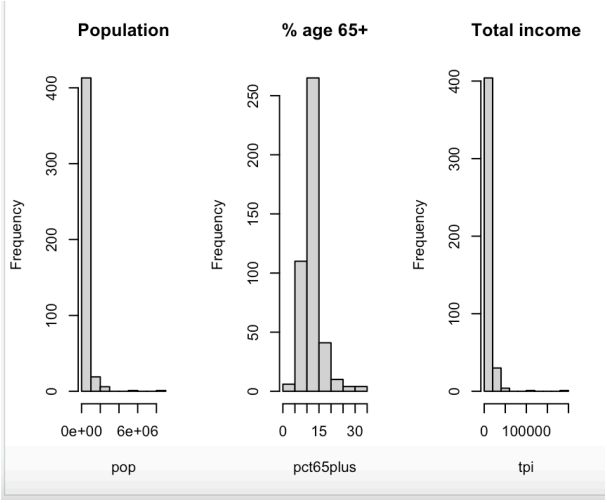
I used the CDI dataset of 440 U.S. counties. The response is number of physicians, with predictors: population (X_1), land area (X_3), total personal income in millions (X_2), and percent of people aged 65 or older (X_4).

Part 1 - Multiple Linear Regression 1
Summaries of predictors in Models A and B

For Model A, the predictors are population (X_1), land area (X_3), and total personal income in millions (X_2). Summary statistics show that both population and total personal income are heavily skewed to the right: most counties are medium-sized, but a few counties have very large populations and incomes. Land area also varies a lot, with some very large counties creating a long right tail in the histogram.

For Model B, the predictors are population (X_1), percent aged 65+ (X_4), and total personal income (X_2). Again, population and income are strongly skewed to the right. In contrast, the distribution of percent aged 65+ is more balanced and mostly falls between about 5% and 25%. This suggests that X_4 acts more like a balanced percentage variable, while X_1 and X_2 are large-scale totals with extreme values.

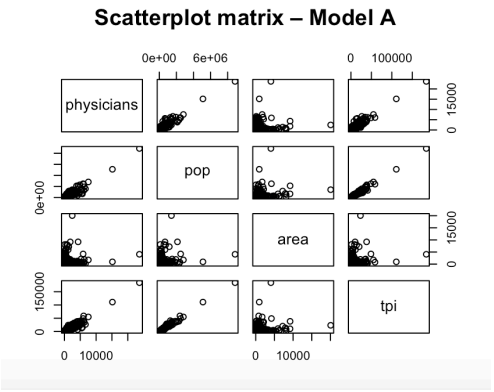


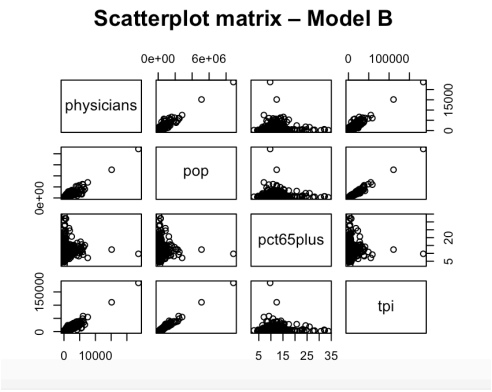


2. Scatterplot matrix & correlation matrix

For Model A, the scatterplot matrix shows strong positive straight-line relationships between physicians and both population and total personal income. The correlation matrix confirms this: physicians have correlations of about 0.94 with population and 0.95 with total personal income. Population and total personal income are also very highly correlated with each other (corr ≈ 0.99), showing a serious potential multicollinearity problem. Land area has only a weak correlation with physicians (≈ 0.08) and with the other predictors (correlations ≈ 0.13 - 0.17).

For Model B, physicians again correlate strongly with population (≈ 0.94) and total personal income (≈ 0.95). Percent aged 65+ has almost no correlation with population, income, or physicians (correlations around 0 to –0.03). So, Model B still has strong collinearity between population and income but doesn't add extra collinearity from the age variable. In both models, the main collinearity concern is the very strong correlation between population and total personal income.





3. Fit each model, report equation, interpret coefficients

Model A

Fitted Equation

$\hat{Y}$ = -13.32 + 0.0008366(pop) - 0.06552(area) + 0.09413(tpi)

Interpretation

- pop: +0.0008366 physicians per person (≈ +84 per 100,000 people).
- area: –0.0655 physicians per sq mile (small negative effect).
- tpi: +0.094 physicians per $1M income (≈ +9.4 per $100M).

Model B

Fitted Equation

$\hat{Y}$ = -172.6 + 0.0005457(pop) + 8.807(pct65plus) + 0.1066(tpi)

Interpretation (very short)

- pop: +0.0005457 physicians per person (≈ +55 per 100,000 people).
- pct65plus: +8.81 physicians per +1% elderly.
- tpi: +0.1066 physicians per $1M (≈ +10.7 per $100M).

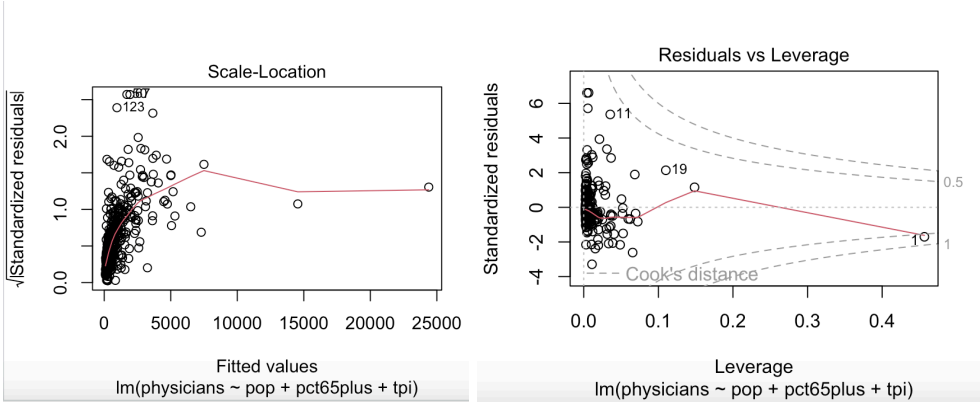
4. Compare R^2 and adjusted R^2

Model A:
R^2 = 0.903
Adjusted R^2 = 0.902

Model B:
R^2 = 0.9
Adjusted R^2 = 0.899
Model A has R^2 ≈ 0.903 and adjusted R^2 ≈ 0.902. Model B has R^2 ≈ 0.900 and adjusted R^2 ≈ 0.899. Both

models explain around 90% of the variation in the number of physicians across counties, which shows a very good fit. Model A has slightly higher adjusted R^2, so by this measure Model A provides a slightly better fit than Model B. However, the difference is very small, so both models perform similarly in terms of explained variation.

5. Diagnostics (residual vs fitted, residual vs predictors, QQ plot)



For both Model A and Model B, the residuals vs fitted plots show some signs of non-constant variance: residuals tend to spread out more for counties with larger fitted physician counts. This is expected because a few very large counties have extreme values for population and income.

The residuals vs predictor plots (residuals vs population, area or percent 65+, and income) don't show a strong curved pattern, though there is more spread at high population and high income levels. The normal Q-Q plots show clear deviations from the straight line in the tails, showing that residuals are somewhat heavy-tailed compared to a normal distribution.

Leverage and influence checks identify a small number of highly influential counties (those with very large populations and incomes). These points are influential but not obviously wrong outliers; they represent major city areas. Overall, the checks suggest mild violations of the ideal linear model assumptions but no major problems. Both models are reasonably good, though one should be careful in interpreting predictions for the largest counties.

6. Add two-factor interactions, refit, compare adjusted R^2

For Model A, adding all two-way interactions among population, area, and income increases adjusted R^2 from about 0.902 to about 0.905. For Model B, adding interactions among population, percent aged 65+, and income increases adjusted R^2 from about 0.899 to about 0.902. In both cases, the interactions provide a small improvement in model fit.

This suggests that there may be some joint effects, so for example, the impact of income on physician counts may depend somewhat on county size. However, the improvement in adjusted R^2 is modest, and the resulting models are more complex and harder to understand. For this reason, one might report both the simpler additive models and the interaction models, noting that interactions improve fit slightly but at the cost of being harder to interpret.

Part 2 - Multiple Linear Regression 2

1. **Partial R^2 for each extra variable X3, X4, X5**

The partial R^2 values show that, after controlling for population and total income, beds (X5) explains about 55% of the remaining variation in physicians, while area (X3) adds about 3% and %65+ (X4) adds less than 1%.

X5 (beds) clearly adds the most explanatory power beyond {X1, X2}.

2. **Partial F-test for the best single variable (X5) at alpha = 0.01**

Hypotheses

H0: $\beta 5 = 0$ (beds does NOT improve the model beyond X_1, X_2)

Ha: $\beta 5 \neq 0$ (beds DOES improve the model)

Test Statistic

From the ANOVA table:

- F = 541.18
- df1 = 1
- df2 = 436
- p-value < 2.2 × 10^-16

Conclusion (alpha = 0.01)

Because the p-value is far below 0.01, we reject H_0.

Adding beds (X5) significantly improves the model beyond {population, income}.

Beds is the strongest additional predictor.

3. **Partial R^2 for pairs beyond {X1, X2}**

Among the pairs, (X4, X5) = (%65+, beds) has the largest partial R^2 (≈0.56).

Interpretation: after accounting for population and income, adding %65+ and beds together explains about 56% of the remaining variation in physician counts.

4. **Partial F-test for the best pair (X4 & X5)**

Hypotheses

H0: $\beta_4 = \beta_5 = 0$

Ha: At least one of $\beta 4, \beta 5 \neq 0$

Test Statistic

- F = 281.67
- df1 = 2
- df2 = 435
- p-value < 2.2 × 10^-16

Conclusion (alpha = 0.01)

Since the p-value is far below 0.01, we reject H0.

Adding the pair (%65+, beds) significantly improves the model beyond {population, income}.

This confirms that the best pair (X_4, X_5) meaningfully increases explanatory power.

Part 3 - Discussion

Across all models we looked at, population (X1), total personal income (X2), and hospital beds (X5) consistently showed up as the most important predictors of how many physicians are in a county. Population and income both have strong direct connections with physician counts, and they were highly correlated with each other, showing that large and wealthy counties tend to have more doctors. However, the partial R^2 analysis showed that after accounting for population and income, beds (X5) added the most extra explanatory power, over 50% of what was left, making it the single most important extra predictor. This makes sense: the number of hospital beds shows healthcare infrastructure and capacity, which is strongly linked to how many physicians are needed to work in those facilities.

Interaction terms gave only small improvements in adjusted R^2 values for both Model A and Model B. While these interaction models captured small joint effects among predictors, they didn't really change the overall story. The relationships between population, income, beds, and physician counts appear mostly additive rather than multiplicative. Because being easy to understand is important, and the interactions didn't really improve the models much, the simpler additive models are still better for explaining things.

Several important statistical ideas were used throughout this analysis. The multiple linear regression framework let us look at several predictors at the same time while controlling for others. Partial R^2 measured how much extra variation each new variable explained beyond the starting model. Partial F-tests formally tested whether adding new predictors really improved the model. Multicollinearity was checked through correlation matrices and was especially noticeable between population and income. Model diagnostics like residual vs. fitted plots and Q–Q plots were used to check linearity, normality, and equal variance assumptions.

The real-world meaning of the findings is simple: counties with more people, higher total income, and more hospital infrastructure tend to have a lot more physicians. The percent of people aged 65+ had a smaller and less consistent effect, suggesting that while what people need matters, structural and economic factors play a bigger role in determining physician availability.

Diagnostics showed mild violations of assumptions. Residual plots showed some unequal spread, especially for counties with very large populations or incomes, and Q-Q plots showed heavy-tailed residuals, suggesting that a few

really large counties affect the distribution. These issues suggest that transformations like using logs for population or income or nonlinear models could improve fit. Also, future models could include per-capita measures (like physicians per 100,000 people) to better handle the wide range in county sizes.

Overall, the analysis suggests that while population and income are basic drivers of physician counts, the availability of hospital beds is the strongest predictor once core factors are accounted for. The models work well, explaining roughly 90% of the variation, but could be improved with transformations or more flexible methods.

Appendix

Appendix A - Part 1 Code and Output

1. Numerical / graphical summaries for predictors

```
> # Model A predictors: pop, area, tpi
> summary(cdi[, c("pop", "area", "tpi")])
      pop               area              tpi
 Min.   : 100043   Min.   :   15.0   Min.   :  1141
 1st Qu.: 139027   1st Qu.:  451.2   1st Qu.:  2311
 Median : 217280   Median :  656.5   Median :  3857
 Mean   : 393011   Mean   : 1041.4   Mean   :  7869
 3rd Qu.: 436064   3rd Qu.:  946.8   3rd Qu.:  8654
 Max.   :8863164   Max.   :20062.0   Max.   :184230
> par(mfrow = c(1, 3))
> hist(cdi$pop,  main = "Population",            xlab = "pop")
> hist(cdi$area, main = "Land area (sq mi)",     xlab = "area")
> hist(cdi$tpi,  main = "Total personal income", xlab = "tpi")
> # Model B predictors: pop, pct65plus, tpi
> summary(cdi[, c("pop", "pct65plus", "tpi")])
      pop            pct65plus          tpi
 Min.   : 100043   Min.   : 3.000   Min.   :  1141
 1st Qu.: 139027   1st Qu.: 9.875   1st Qu.:  2311
 Median : 217280   Median :11.750   Median :  3857
 Mean   : 393011   Mean   :12.170   Mean   :  7869
 3rd Qu.: 436064   3rd Qu.:13.625   3rd Qu.:  8654
 Max.   :8863164   Max.   :33.800   Max.   :184230
>
> par(mfrow = c(1, 3))
> hist(cdi$pop,       main = "Population",        xlab = "pop")
> hist(cdi$pct65plus, main = "% age 65+",         xlab = "pct65plus")
> hist(cdi$tpi,       main = "Total income",      xlab = "tpi")
>
> par(mfrow = c(1, 1))
>
```

2. Scatterplot matrix & correlation matrix

```
> # Model A: physicians, pop, area, tpi
> pairs(cdi[, c("physicians", "pop", "area", "tpi")],
+       main = "Scatterplot matrix – Model A")
> corA <- cor(cdi[, c("physicians", "pop", "area", "tpi")])
> corA
           physicians       pop       area       tpi
physicians 1.00000000 0.9402486 0.07807466 0.9481106
pop        0.94024859 1.0000000 0.17308335 0.9867476
area       0.07807466 0.1730834 1.00000000 0.1270743
tpi        0.94811057 0.9867476 0.12707426 1.0000000
> # Model B: physicians, pop, pct65plus, tpi
> pairs(cdi[, c("physicians", "pop", "pct65plus", "tpi")],
+       main = "Scatterplot matrix – Model B")
> corB <- cor(cdi[, c("physicians", "pop", "pct65plus", "tpi")])
> corB
            physicians         pop   pct65plus         tpi
physicians  1.00000000  0.94024859 -0.00312863  0.94811057
pop         0.94024859  1.00000000 -0.02903739  0.98674763
pct65plus  -0.00312863 -0.02903739  1.00000000 -0.02273315
tpi         0.94811057  0.98674763 -0.02273315  1.00000000
>
```

3. Fit each model, report equation, interpret coefficients

```
> modelA <- lm(physicians ~ pop + area + tpi, data = cdi)
> summary(modelA)

Call:
lm(formula = physicians ~ pop + area + tpi, data = cdi)

Residuals:
    Min      1Q  Median      3Q     Max
-1855.6  -215.2   -74.6    79.0  3689.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.332e+01  3.537e+01  -0.377 0.706719
pop          8.366e-04  2.867e-04   2.918 0.003701 **
area        -6.552e-02  1.821e-02  -3.597 0.000358 ***
tpi          9.413e-02  1.330e-02   7.078 5.89e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560.4 on 436 degrees of freedom
Multiple R-squared:  0.9026,    Adjusted R-squared:  0.902
F-statistic:  1347 on 3 and 436 DF,  p-value: < 2.2e-16
```

```
> modelB <- lm(physicians ~ pop + pct65plus + tpi, data = cdi)
> summary(modelB)

Call:
lm(formula = physicians ~ pop + pct65plus + tpi, data = cdi)

Residuals:
    Min      1Q  Median      3Q     Max
-1856.1  -208.0   -73.5    43.6  3739.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.726e+02  8.959e+01  -1.926   0.0547 .
pop          5.457e-04  2.775e-04   1.966   0.0499 *
pct65plus    8.807e+00  6.791e+00   1.297   0.1954
tpi          1.066e-01  1.296e-02   8.223  2.3e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 567.5 on 436 degrees of freedom
Multiple R-squared:  0.9001,    Adjusted R-squared:  0.8995
F-statistic:  1310 on 3 and 436 DF,  p-value: < 2.2e-16
```

5. Diagnostics (residual vs fitted, residual vs predictors, QQ plot)

```
> # Model A
> plot(modelA, which = 1)  # residuals vs fitted
> plot(modelA, which = 2)  # normal Q-Q
> plot(modelA, which = 3)  # scale-location
> plot(modelA, which = 5)  # residuals vs leverage
> # Model B
> plot(modelB, which = 1)
> plot(modelB, which = 2)
> plot(modelB, which = 3)
> plot(modelB, which = 5)
>
```

6. Add two-factor interactions, refit, compare adjusted R^2

```
> modelA_int <- lm(physicians ~ pop*area + pop*tpi + area*tpi, data = cdi)
> summary(modelA_int)

Call:
lm(formula = physicians ~ pop * area + pop * tpi + area * tpi,
    data = cdi)

Residuals:
    Min      1Q  Median      3Q     Max
-1950.2  -198.0   -61.1    76.6  3578.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.826e+01  4.727e+01  -1.232  0.21848
pop          7.252e-04  3.259e-04   2.225  0.02657 *
area        -6.421e-02  3.014e-02  -2.131  0.03369 *
tpi          1.087e-01  1.450e-02   7.496 3.76e-13 ***
pop:area     6.173e-07  2.058e-07   2.999  0.00287 **
pop:tpi      1.696e-09  1.041e-09   1.630  0.10392
area:tpi    -3.706e-05  1.152e-05  -3.217  0.00139 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 551.4 on 433 degrees of freedom
Multiple R-squared:  0.9064,    Adjusted R-squared:  0.9051
F-statistic: 698.7 on 6 and 433 DF,  p-value: < 2.2e-16
```

```
> # Model B with interactions
> modelB_int <- lm(physicians ~ pop*pct65plus + pop*tpi + pct65plus*tpi, data = cdi)
> summary(modelB_int)

Call:
lm(formula = physicians ~ pop * pct65plus + pop * tpi + pct65plus *
    tpi, data = cdi)

Residuals:
    Min      1Q  Median      3Q     Max
-1950.9  -184.8   -49.3    60.2  3709.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.292e+01  1.245e+02   0.746  0.45599
pop           -1.351e-03  1.198e-03  -1.127  0.26017
pct65plus     -1.623e+01  9.672e+00  -1.678  0.09408 .
tpi            1.582e-01  5.332e-02   2.967  0.00317 **
pop:pct65plus  1.719e-04  9.494e-05   1.811  0.07090 .
pop:tpi       -9.003e-11  6.818e-10  -0.132  0.89502
pct65plus:tpi -4.555e-03  4.195e-03  -1.086  0.27808
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 559.2 on 433 degrees of freedom
Multiple R-squared:  0.9037,    Adjusted R-squared:  0.9024
F-statistic: 677.2 on 6 and 433 DF,  p-value: < 2.2e-16

> # Compare adjusted R^2
> summary(modelA)$adj.r.squared
[1] 0.9019733
> summary(modelA_int)$adj.r.squared
[1] 0.9050816
> summary(modelB)$adj.r.squared
[1] 0.8994517
> summary(modelB_int)$adj.r.squared
[1] 0.902366
>
```

Appendix B - Part 2 Code and Output

Setup:

```
> cdi <- read.table("CDI.txt", header = FALSE)
> names(cdi) <- c("id","county","state","area","pop","pct18_34","pct65plus",
+                 "physicians","beds","crimes","pct_hs","pct_ba","pct_poverty",
+                 "pct_unemp","pci","tpi","region")
>
> base <- lm(physicians ~ pop + tpi, data = cdi)
> summary(base)   # baseline model with X1,X2 only

Call:
lm(formula = physicians ~ pop + tpi, data = cdi)

Residuals:
    Min      1Q  Median      3Q     Max
-1849.1  -198.3   -71.4    39.7  3755.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.444e+01  3.283e+01  -1.963   0.0503 .
pop          5.310e-04  2.775e-04   1.914   0.0563 .
tpi          1.072e-01  1.297e-02   8.269 1.64e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568 on 437 degrees of freedom
Multiple R-squared:  0.8998,    Adjusted R-squared:  0.8993
F-statistic:  1961 on 2 and 437 DF,  p-value: < 2.2e-16
```

1. **Partial R^2 for each extra variable X3, X4, X5**

```
>
> m_X3 <- lm(physicians ~ pop + tpi + area, data = cdi)
> m_X4 <- lm(physicians ~ pop + tpi + pct65plus, data = cdi)
> m_X5 <- lm(physicians ~ pop + tpi + beds, data = cdi)
> SSE_base <- sum(resid(base)^2)
> SSE_X3   <- sum(resid(m_X3)^2)
> SSE_X4   <- sum(resid(m_X4)^2)
> SSE_X5   <- sum(resid(m_X5)^2)
> pR2_X3 <- (SSE_base - SSE_X3)/SSE_base
> pR2_X4 <- (SSE_base - SSE_X4)/SSE_base
> pR2_X5 <- (SSE_base - SSE_X5)/SSE_base
> pR2_X3; pR2_X4; pR2_X5
[1] 0.02882495
[1] 0.003842367
[1] 0.5538182
```

2. **Partial F-test for the best single variable (X5) at alpha = 0.01**

```
> anova(base, m_X5)
Analysis of Variance Table

Model 1: physicians ~ pop + tpi
Model 2: physicians ~ pop + tpi + beds
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1    437 140967081
2    436  62896949  1  78070132 541.18 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

3. **Partial R^2 for pairs beyond {X1, X2}**

```
> m_X3X4 <- lm(physicians ~ pop + tpi + area + pct65plus, data = cdi)
> m_X3X5 <- lm(physicians ~ pop + tpi + area + beds,      data = cdi)
> m_X4X5 <- lm(physicians ~ pop + tpi + pct65plus + beds, data = cdi)
> SSE_X3X4 <- sum(resid(m_X3X4)^2)
> SSE_X3X5 <- sum(resid(m_X3X5)^2)
> SSE_X4X5 <- sum(resid(m_X4X5)^2)
>
> pR2_X3X4 <- (SSE_base - SSE_X3X4)/SSE_base
> pR2_X3X5 <- (SSE_base - SSE_X3X5)/SSE_base
> pR2_X4X5 <- (SSE_base - SSE_X4X5)/SSE_base
> pR2_X3X4; pR2_X3X5; pR2_X4X5
[1] 0.03314181
[1] 0.5558232
[1] 0.5642756
```

4. Partial F-test for the best pair (X4 & X5)

```
> anova(base, m_X4X5)
Analysis of Variance Table

Model 1: physicians ~ pop + tpi
Model 2: physicians ~ pop + tpi + pct65plus + beds
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1    437 140967081
2    435  61422794  2  79544288 281.67 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```