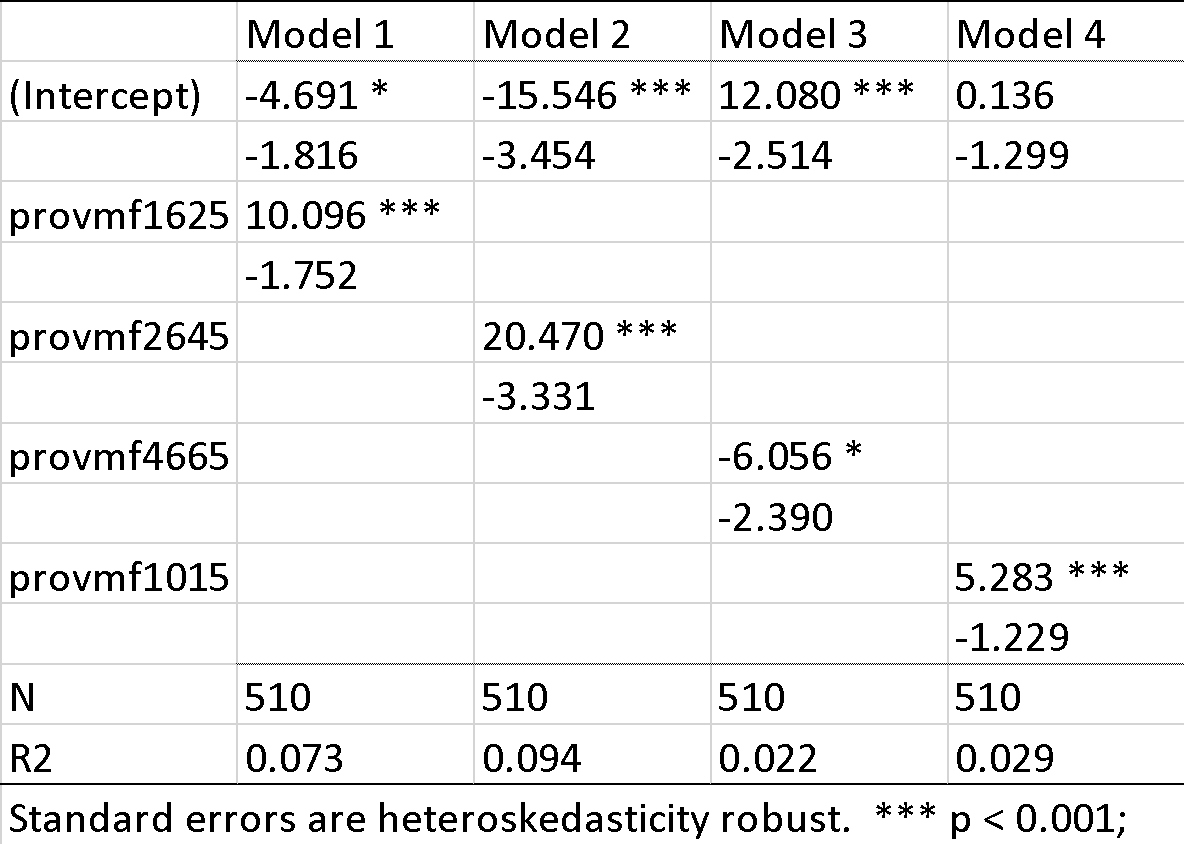
1. Research question

What factors can have an effect on the crime rate in the 55 provinces in China between 1998 to 2004

* 1. How to choose the specific age group

First, intuitively, the age group from 26 to 45 has a relatively larger population and labor force.

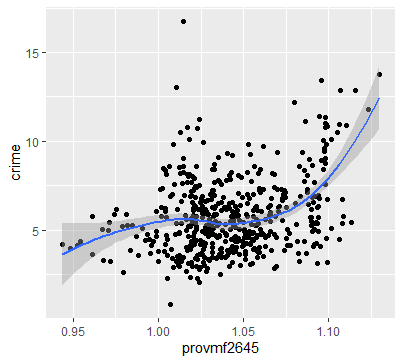
Second, we do linear regression to analyze the effect of sex ratio on crime rate by different age groups and we find that:



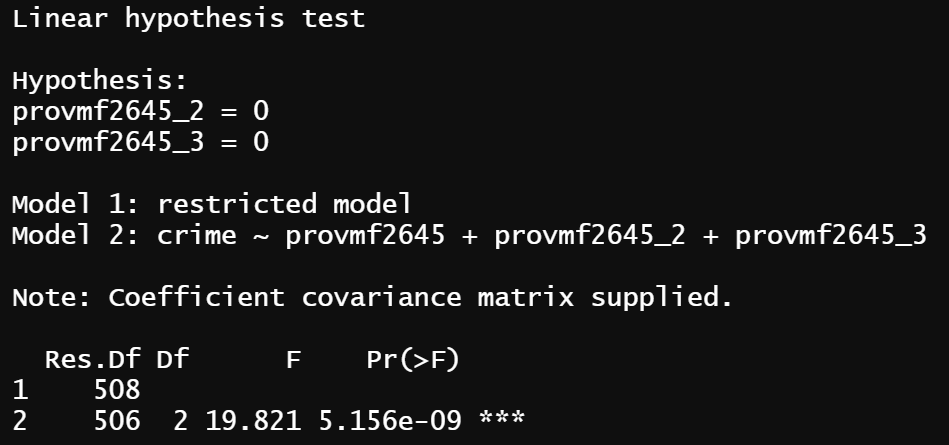
It indicates that the sex ratio of the population between 26 to 45 has the most significant effect.

1. Nonlinear Regression for this age group

It might be nonlinear according to the dot plot



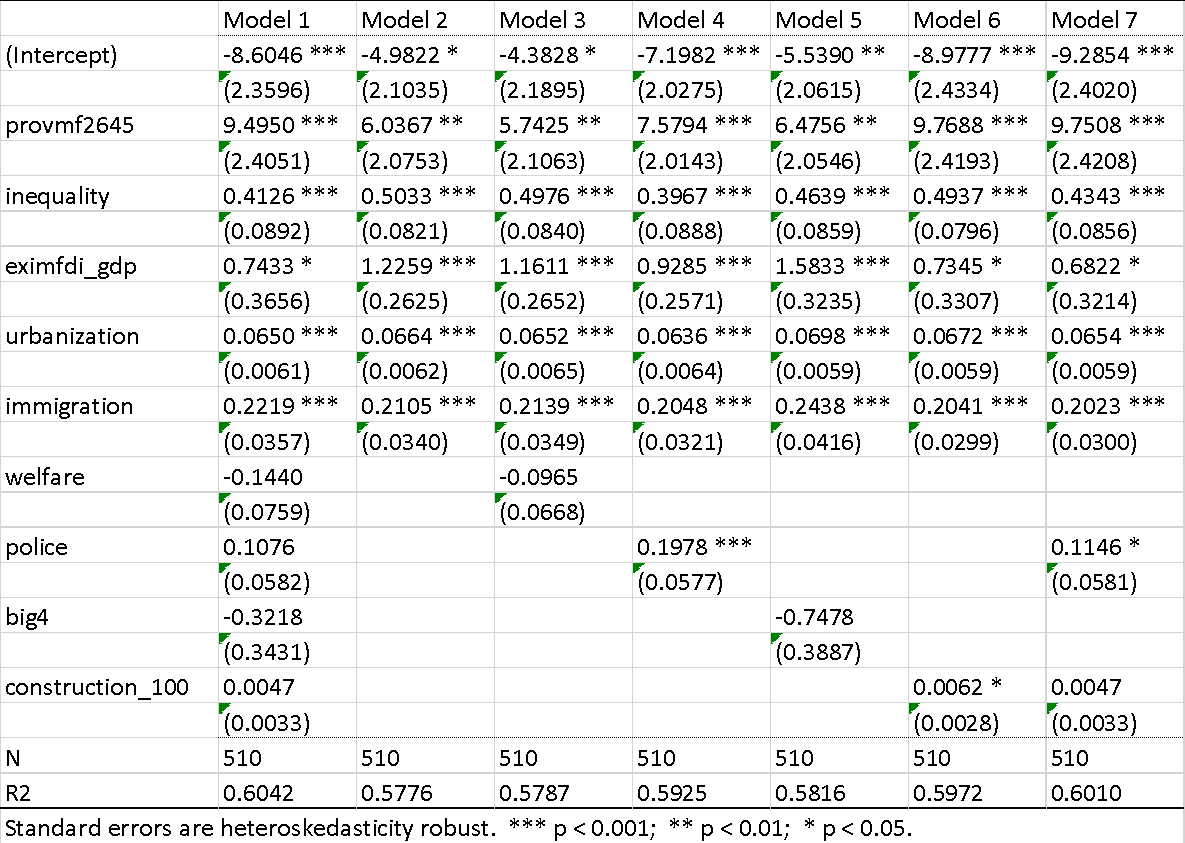




Therefore, it must be a nonlinear relationship but both polynomials and ln are possible.

2. Econometric models and data:

* 1. We add more variables.

We change the unit of construction from 10,000 square meters to 1,000,000 square meters. 

Though intuitively they are relevant to the crime rate, they are not statistically significant.

* 1. **Variable screening,**

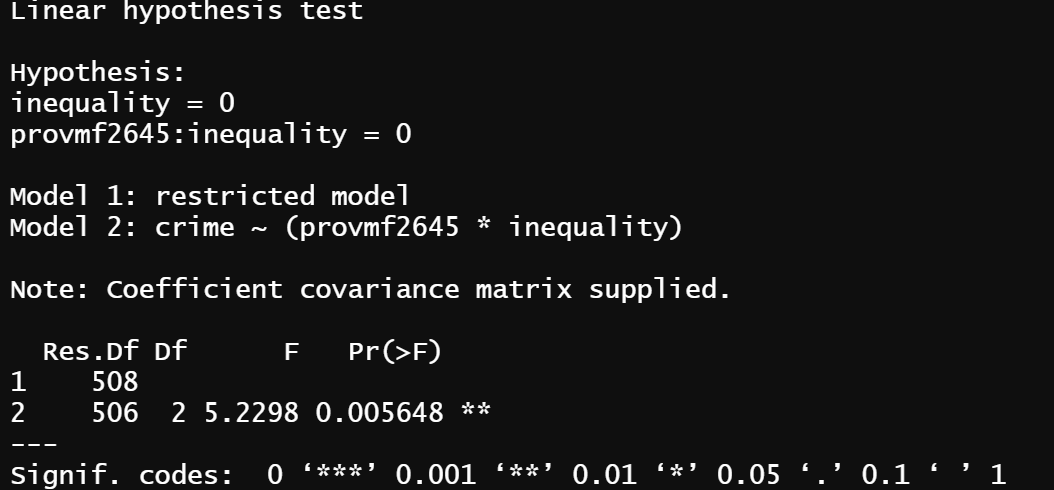
However, it is notable that when removing all these 4 variables (*welfare, police, big4, construction\_100*), the coefficient of *provmf2645* decreases significantly, which depicts that there is an omitted variable bias if we omit all these 4 variables. Thus, we need to screen the variables one by one and find out which variable should be chosen.

When we add the variable *police*, comparing model 4 to model 2, the coefficient of *provmf2645* soared by 25.56%, the coefficient of *inequality* declined by 42.08%, and the coefficient of *eximfdi\_gdp* decreased by 19.78%, while other coefficients almost remain the same. Since the coefficient of *police* becomes statistically significant, it is necessary to include *police* in one of the selected variables. In the same way, comparing model 5 with model 6, after adding the variable construction\_100, the coefficient of *provmf2645* increases by 61.82%, the coefficient of *eximfdi\_gdp* decreases by 32.69%, and at the same time the coefficient of it turns statistically significant, so it is a must to add it. if ignoring police and construction\_100 will cause omitted variables bias.

Generally, we add two more variables which are *police* and *construction\_100.*

* 1. Add interaction to linear regression

We add *provmf2645\*inequality* to the regression model. In model 1, the coefficients are not statistically significant. F test should be one to test whether there is an interaction relationship between these 2 variables which leads to imperfect multicollinearity. The value is significant which indicates there is an interactive relationship between these 2 variables.

(Does the effect of sex ratio of 26-45 on crime rate depend on inequality?)

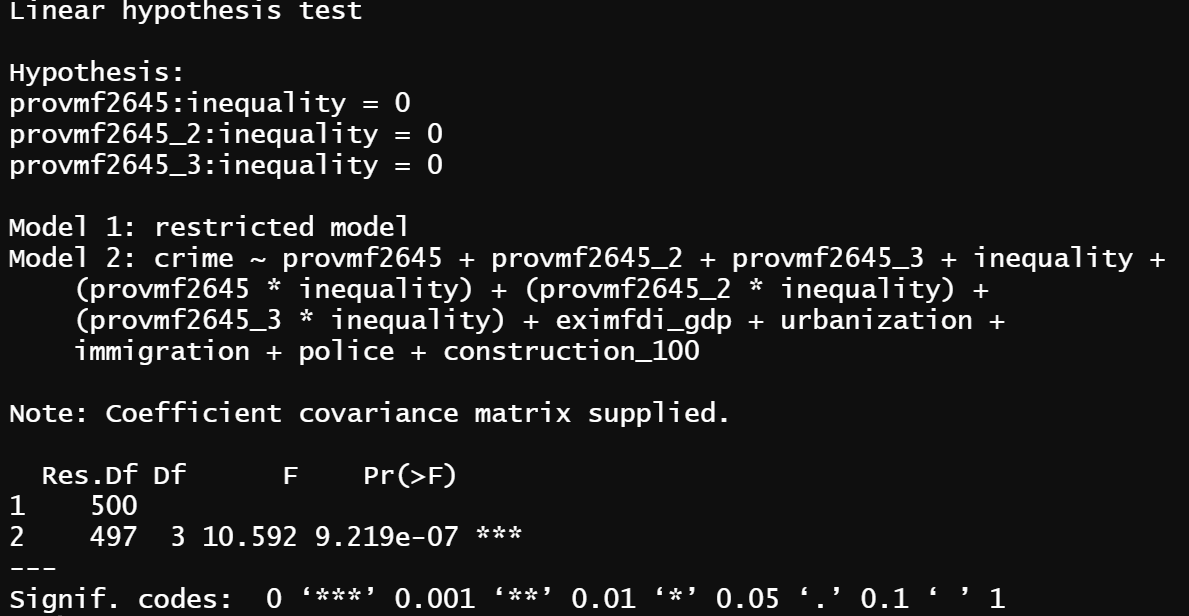
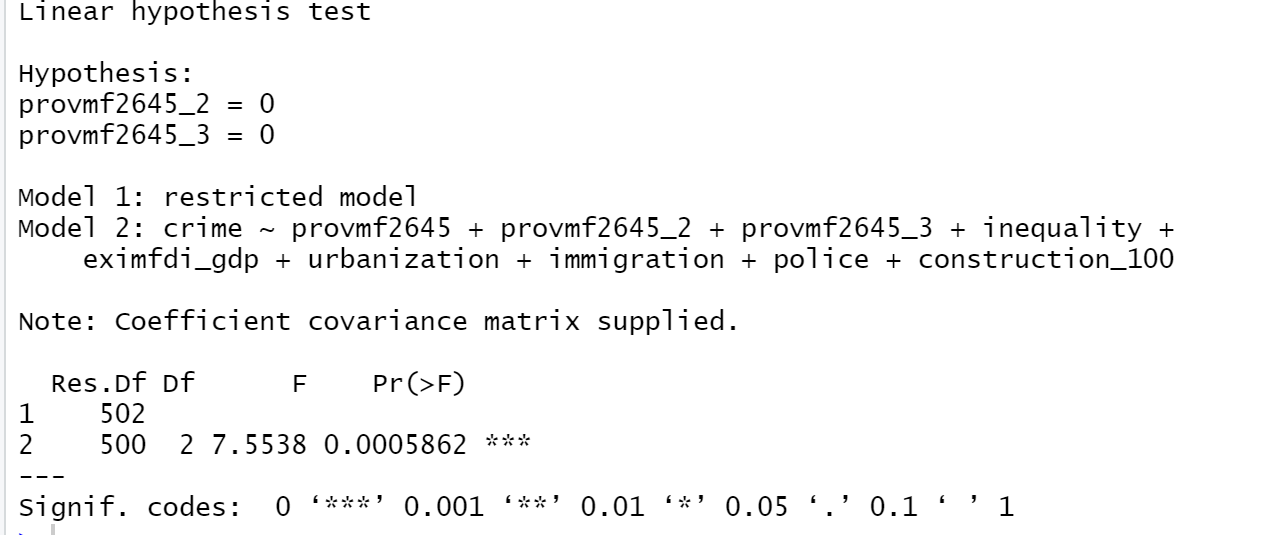
|  |  |  |
| --- | --- | --- |
|  | Model 1 | Model 2 |
| (Intercept) | 2.811 | 9.626 |
|  | (11.666) | (8.825) |
| provmf2645 | 3.520 | -8.219 |
|  | (11.132) | (8.404) |
| inequality | -7.651 | -6.907 \* |
|  | (4.508) | (3.321) |
| provmf2645:inequality | 7.080 | 6.990 \* |
|  | (4.291) | (3.161) |
| eximfdi\_gdp |  | 0.723 \* |
|  |  | (0.222) |
| urbanization |  | 0.060 \*\*\* |
|  |  | (0.006) |
| immigration |  | 0.234 \*\*\* |
|  |  | (0.039) |
| police |  | 0.119 \* |
|  |  | (0.052) |
| construction\_100 |  | 0.004 |
|  |  | (0.001) |
| N | 510 | 510 |
| R2 | 0.104 | 0.605 |
| Standard errors are heteroskedasticity robust. \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05. | | |
|  |  |  |

According to model 2, the coefficient of provmf26-45 is not statistically significant, it can not explain anything. so model2 is not good, we need to try other models(combined with previous research we try nonlinear models).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| (Intercept) | -85.705 | -882.140 \*\*\* | -23809.635 \*\*\* | -96.351 \* | -3017.449 \*\* |
|  | (48.370) | (251.702) | (6748.297) | (48.387) | (1026.686) |
| provmf2645 | 175.947 \* | 1691.407 \*\*\* | 67810.940 \*\*\* | 176.669 \* | 8659.903 \*\* |
|  | (92.267) | (478.983) | (19441.926) | (92.695) | (2979.779) |
| provmf2645\_2 | -88.948 \* | -809.309 \*\*\* | -64323.460 \*\*\* | -80.019 \* | -8284.060 \*\* |
|  | (44.379) | (227.780) | (18660.447) | (44.425) | (2880.674) |
| inequality | -7.477 \* | 308.386 \*\*\* | 8198.242 \*\* | 0.428 \*\*\* | 0.457 \*\*\* |
|  | (3.323) | (98.053) | (2621.679) | (0.094) | (0.094) |
| eximfdi\_gdp | 0.755 \* | 0.823 \* | 0.794 \* | 0.708 \* | 0.712 \* |
|  | (0.222) | (0.221) | (0.218) | (0.222) | (0.220) |
| urbanization | 0.061 \*\*\* | 0.061 \*\*\* | 0.061 \*\*\* | 0.066 \*\*\* | 0.068 \*\*\* |
|  | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) |
| immigration | 0.241 \*\*\* | 0.261 \*\*\* | 0.233 \*\*\* | 0.206 \*\*\* | 0.182 \*\*\* |
|  | (0.039) | (0.039) | (0.039) | (0.036) | (0.037) |
| police | 0.131 \* | 0.115 \* | 0.104 | 0.125 \* | 0.130 \* |
|  | (0.052) | (0.052) | (0.051) | (0.052) | (0.052) |
| construction\_100 | 0.004 | 0.005 | 0.005 | 0.005 | 0.004 |
|  | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| provmf2645:inequality | 7.525 \* | -593.038 \*\*\* | -23324.490 \*\* |  |  |
|  | (3.162) | (186.354) | (7552.507) |  |  |
| provmf2645\_2:inequality |  | 285.273 \*\*\* | 22101.093 \*\* |  |  |
|  |  | (88.507) | (7248.446) |  |  |
| provmf2645\_3 |  |  | 20322.432 \*\*\* |  | 2642.023 \*\* |
|  |  |  | (5966.737) |  | (927.581) |
| provmf2645\_3:inequality |  |  | -6974.254 \*\* |  |  |
|  |  |  | (2317.576) |  |  |
| N | 510 | 510 | 510 | 510 | 510 |
| R2 | 0.608 | 0.616 | 0.629 | 0.604 | 0.610 |
| Standard errors are heteroskedasticity robust. \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05. | | | | | |

Then, we add interaction variables for higher degrees, which model 2 with 2nd degree interactions, model 3 with 3rd interactions, model 4 and model 5 with 2nd and 3rd degree respectively but no interaction variables.

Comparing these 5 models, it is manifest that model 2 and model 5 are a little better. Notably, we wonder if there is an interaction relationship, so we do a joint hypothesis test.



It is statistically significant, which indicates the interaction variables should exist, so model 2 is the best model.

3.Results and Inference

The 2nd degree interactions regression is the best model we get

**crime ~ provmf2645 + provm2645\_2 + inequality + provm2645\*inequality + provmf2645\_2\*inequality +eximfdi\_gdp + urbanization + immigration + police + construction\_100**