DATA ANALYSIS AND PREDICTION OF U.S. PROFESSORS' SALARIES

Henri & Jennifer

Outline

1. Project goals
2. Data Set
3. First Model
4. Regression Diagnostics
5. Violation of Assumptions
6. Final Selected Model
7. Prediction

Project goals

Find a model that will predict ~~professor's~~ salaries.

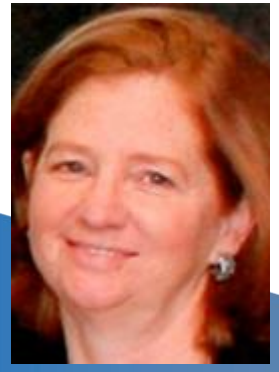
Dr. Ongie's

Dr. Hamilton's

Dr. Yu's

Dr. Sanders'

Dr. Clough's



Dr. Maadooliat's

Dr. Rowe's

Dr. Spiller's

Data set - Salaries

- We used the dataset labeled Salaries from CarData R Package
- Total of 397 rows where each row corresponds to one Professor.

Variable	Type	Description
rank	Categorical	Professor's Rank
		Assistant Professor (10) (17%) Associate Professor (00) (16%) Professor (01) (67%)
discipline	Categorical	Professors type of department
		Type A (0) = Theoretical (54%) Type B (1) = Applied (46%)
yrs.since.phd	Integer	Number of years since Professor has attained Ph.D.
		min = 1 median = 21 mean= 22.31 max = 56
yrs.service	Integer	Number of years Professor has served in department or university
		min = 0 median = 16 mean= 17.61 max = 60
sex	Categorical	Professor's sex
		Male (0) (90%) Female (1) (10%)
salary	Double	Professor's nine-month salary (inflated 1.35 since 2009)
		min = 78,030 median = 144,855 mean= 153,504 max = 312,586

First Model

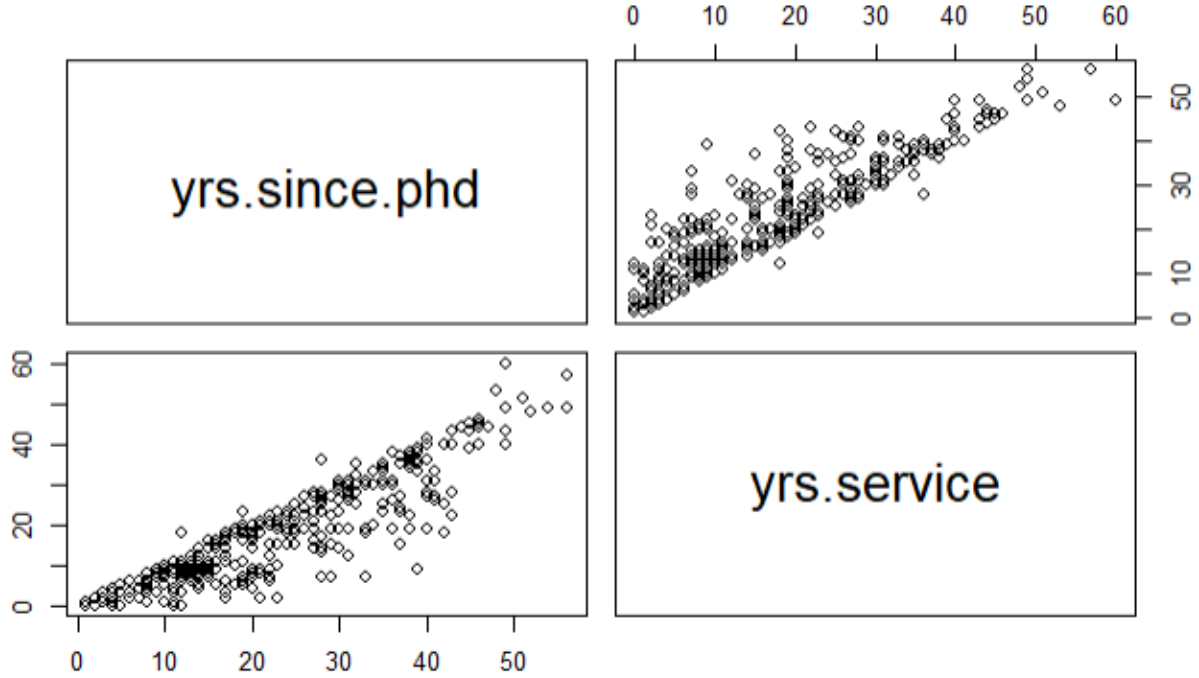
```
Call:
lm(formula = salary ~ discipline * rank + discipline * Salaries2$FRank +
    yrs.service + sex + yrs.since.phd, data = Salaries2)

Residuals:
   Min     1Q Median     3Q    Max
-88114 -17495  -2172  14281 134482

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   109055.9     6862.2  15.892  < 2e-16 ***
discipline1                    25843.7     7792.0   3.317 0.000997 ***
rank1                         -10102.3     8879.3  -1.138 0.255929
Salaries2$FRank1               47169.7     6905.7   6.831 3.28e-11 ***
yrs.service                     -666.0      286.5  -2.325 0.020598 *
sex1                           -6775.2     5233.8  -1.295 0.196256
yrs.since.phd                    734.6      326.0   2.253 0.024798 *
discipline1:rank1              -11762.6    11005.8  -1.069 0.285841
discipline1:Salaries2$FRank1   -6572.4     8606.4  -0.764 0.445528
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30460 on 388 degrees of freedom
Multiple R-squared:  0.4563,    Adjusted R-squared:  0.4451
F-statistic:  40.7 on 8 and 388 DF,  p-value: < 2.2e-16
```

Regression Diagnostics – Unusual Data

	dfb.1_	dfb.dsc1	dfb.rnk1	dfb.S2$F	dfb.yrs.	dfb.sex1	dfb.d1:1	dfb.d1:S	dffit	cov.r	cook.d	hat
25	0.12	-0.11	-0.10	-0.12	-0.01	0.10	0.08	0.10	0.18	1.07_*	0.00	0.06
35	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.03	1.07_*	0.00	0.05
36	-0.01	0.00	-0.01	0.01	0.00	0.09	0.05	0.00	0.13	1.06_*	0.00	0.05
64	0.01	-0.06	0.00	0.00	-0.01	-0.08	0.04	0.05	-0.12	1.06_*	0.00	0.05
65	0.00	0.00	0.00	0.00	0.00	-0.05	0.15	0.00	0.37	0.93_*	0.02	0.02
105	-0.04	0.03	0.03	0.04	0.00	0.01	-0.02	-0.03	-0.04	1.06_*	0.00	0.04
107	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.06_*	0.00	0.04
108	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	1.06_*	0.00	0.04
109	0.02	-0.02	-0.01	-0.02	0.00	0.00	0.01	0.01	0.02	1.06_*	0.00	0.04
112	-0.02	0.02	0.01	0.02	0.00	0.00	-0.01	-0.02	-0.02	1.06_*	0.00	0.04
113	0.00	0.00	0.04	0.00	0.00	-0.01	-0.03	0.00	0.06	1.07_*	0.00	0.04
115	0.01	0.00	-0.01	0.03	-0.05	0.07	0.01	-0.01	0.09	1.06_*	0.00	0.04
119	0.00	0.00	0.04	0.00	0.00	-0.01	-0.03	0.00	0.06	1.07_*	0.00	0.04
120	0.01	0.00	-0.04	0.00	0.00	-0.04	0.03	0.00	-0.07	1.08_*	0.00	0.06
124	0.46	-0.46	-0.41	-0.55	0.17	0.46	0.35	0.42	0.77_*	0.91_*	0.07	0.06_*
126	-0.09	0.01	0.05	0.01	0.29	0.00	-0.01	-0.06	0.36	0.93_*	0.02	0.02
128	0.00	0.00	-0.01	0.00	0.00	-0.01	0.01	0.00	-0.01	1.08_*	0.00	0.06
130	0.00	0.00	0.03	0.00	0.00	-0.01	-0.02	0.00	0.04	1.07_*	0.00	0.04
131	-0.02	0.01	0.01	0.02	0.00	0.00	-0.01	-0.01	-0.02	1.06_*	0.00	0.04
132	-0.13	0.02	0.07	-0.01	0.39	0.01	-0.01	-0.06	0.45_*	0.94_*	0.03	0.03
133	0.05	-0.05	-0.04	-0.05	0.00	0.04	0.04	0.04	0.07	1.08_*	0.00	0.06
134	0.00	0.00	-0.01	0.00	0.00	-0.01	0.01	0.00	-0.02	1.08_*	0.00	0.06
142	0.01	-0.01	-0.01	-0.01	0.00	0.00	0.01	0.01	0.01	1.06_*	0.00	0.04
154	0.01	-0.06	0.00	0.00	0.00	-0.08	0.04	0.05	-0.12	1.06_*	0.00	0.05
187	0.01	-0.06	0.00	0.00	0.00	-0.08	0.04	0.05	-0.12	1.06_*	0.00	0.05
195	-0.01	0.01	0.00	0.00	0.02	0.00	-0.01	-0.01	0.02	1.10_*	0.00	0.07_*
227	0.02	0.00	0.31	0.00	-0.02	-0.09	-0.26	0.00	0.45_*	0.97	0.03	0.04
232	0.10	-0.10	-0.09	-0.12	0.05	0.10	0.08	0.10	0.18	1.08_*	0.00	0.06_*
236	0.03	0.00	-0.02	0.10	-0.09	-0.03	0.00	-0.07	0.23	0.91_*	0.01	0.01
238	-0.05	0.00	0.28	0.00	0.06	0.26	-0.21	0.00	0.48_*	1.01	0.03	0.06
239	-0.06	0.01	0.04	0.04	0.20	-0.01	-0.01	-0.07	0.31	0.90_*	0.01	0.01
254	0.02	0.00	-0.13	-0.01	-0.01	-0.12	0.10	0.00	-0.22	1.07_*	0.01	0.06
256	-0.02	0.01	0.01	0.01	0.00	0.00	-0.01	-0.01	-0.02	1.06_*	0.00	0.04
259	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.07_*	0.00	0.04
273	0.00	0.00	0.03	0.00	0.00	-0.01	-0.02	0.00	0.04	1.07_*	0.00	0.04
274	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	1.07_*	0.00	0.04
283	-0.30	0.05	0.18	0.02	0.96	0.01	-0.03	-0.19	1.15_*	0.38_*	0.15	0.02
286	-0.09	0.10	0.07	0.15	-0.12	0.01	-0.07	-0.09	-0.18	1.09_*	0.00	0.07_*
290	0.00	0.00	-0.01	0.00	0.00	0.00	0.01	0.00	-0.01	1.07_*	0.00	0.04
299	-0.10	0.02	0.06	0.04	0.31	-0.01	-0.01	-0.09	0.43_*	0.82_*	0.02	0.02
307	0.00	0.00	-0.02	0.00	0.00	0.00	0.01	0.00	-0.02	1.07_*	0.00	0.04
309	0.00	0.00	0.01	0.00	0.00	0.00	-0.01	0.00	0.01	1.07_*	0.00	0.04
317	-0.06	0.32	-0.02	0.03	0.01	0.45	-0.20	-0.29	0.68_*	0.89_*	0.06	0.05
318	-0.18	0.03	0.11	-0.14	0.58	-0.01	-0.02	0.14	0.72_*	0.64_*	0.06	0.02
335	0.01	-0.05	0.01	-0.01	0.01	-0.08	0.03	0.05	-0.11	1.07_*	0.00	0.05
364	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	1.06_*	0.00	0.04
377	0.00	0.00	-0.02	0.00	0.00	0.01	0.01	0.00	-0.02	1.07_*	0.00	0.04
378	0.00	0.00	-0.07	0.00	0.00	0.02	0.06	0.00	-0.10	1.06_*	0.00	0.04
381	0.00	0.00	-0.05	0.00	0.00	0.01	0.04	0.00	-0.07	1.06_*	0.00	0.04

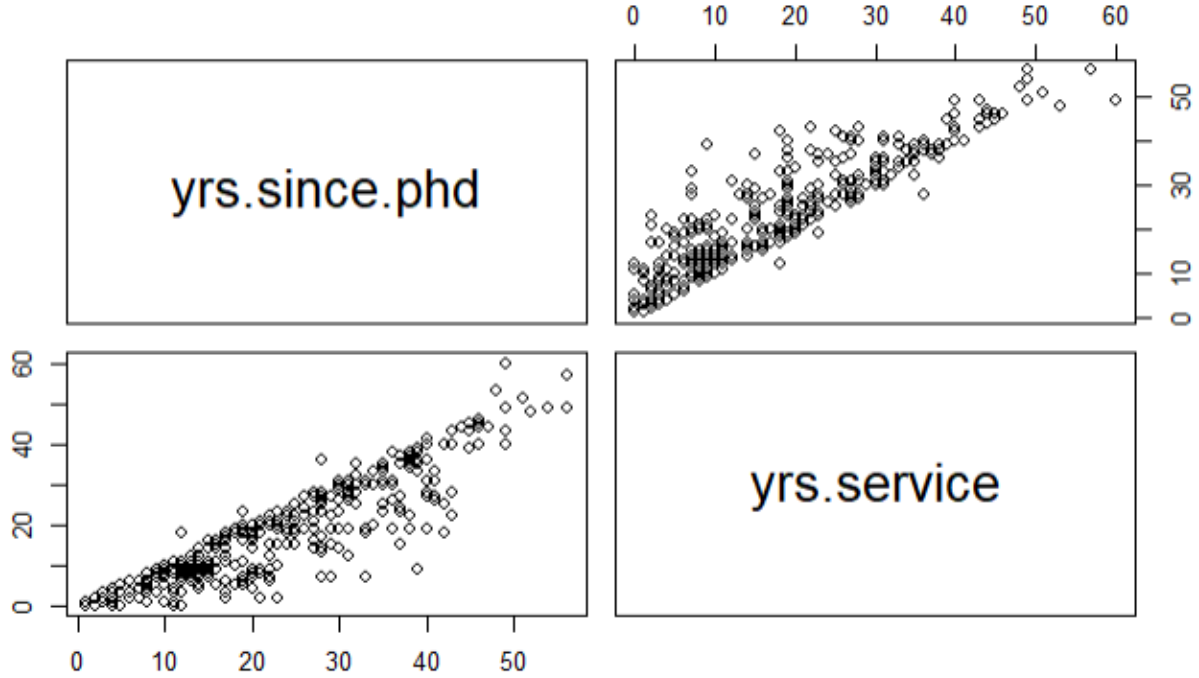
Need to Fix	To fix…
Unusual Data	Remove some influential points
Multicollinearity	
Normality	
Non-Constant Variance	
Non-Linearity	

Regression Diagnostics – Multicollinearity



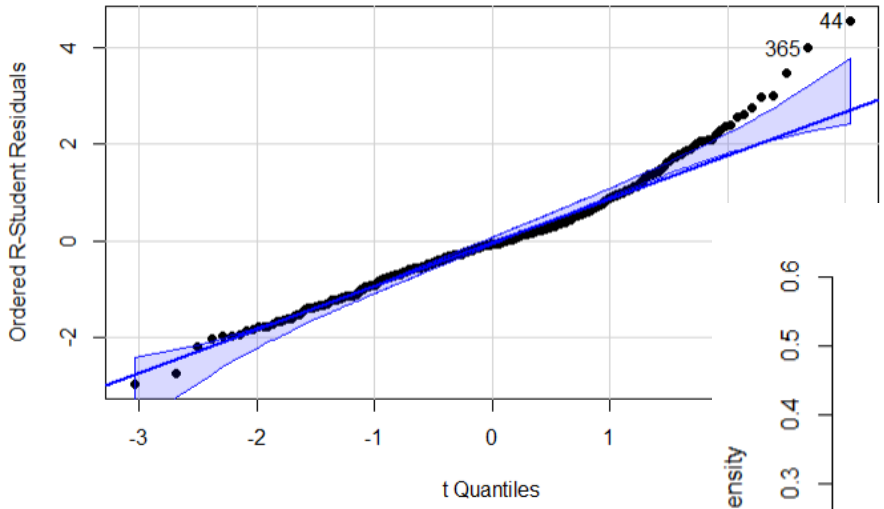
Need to Fix	To fix…
Unusual Data	Remove some influential points
Multicollinearity	???
Normality	
Non-Constant Variance	
Non-Linearity	

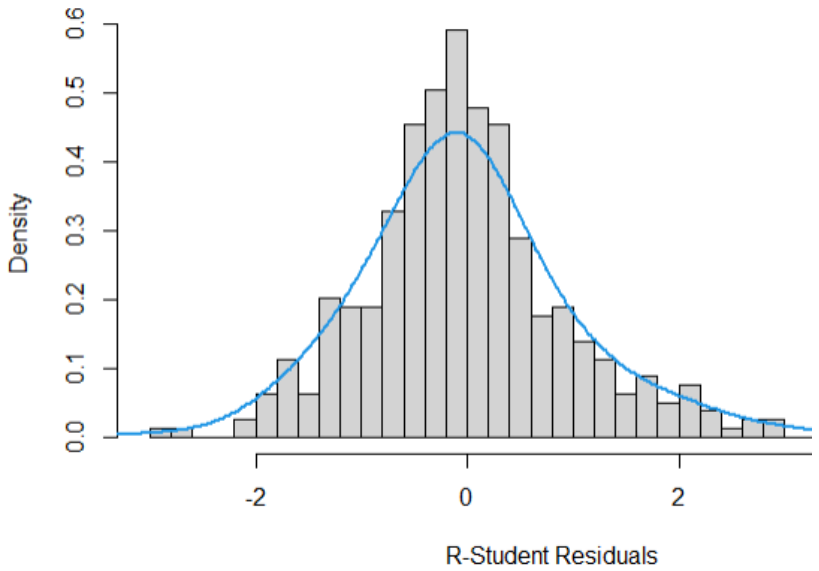
Regression Diagnostics – Multicollinearity



Need to Fix	To fix…
Unusual Data	Remove some influential points
Multicollinearity	Choose best predictor
Normality	
Non-Constant Variance	
Non-Linearity	

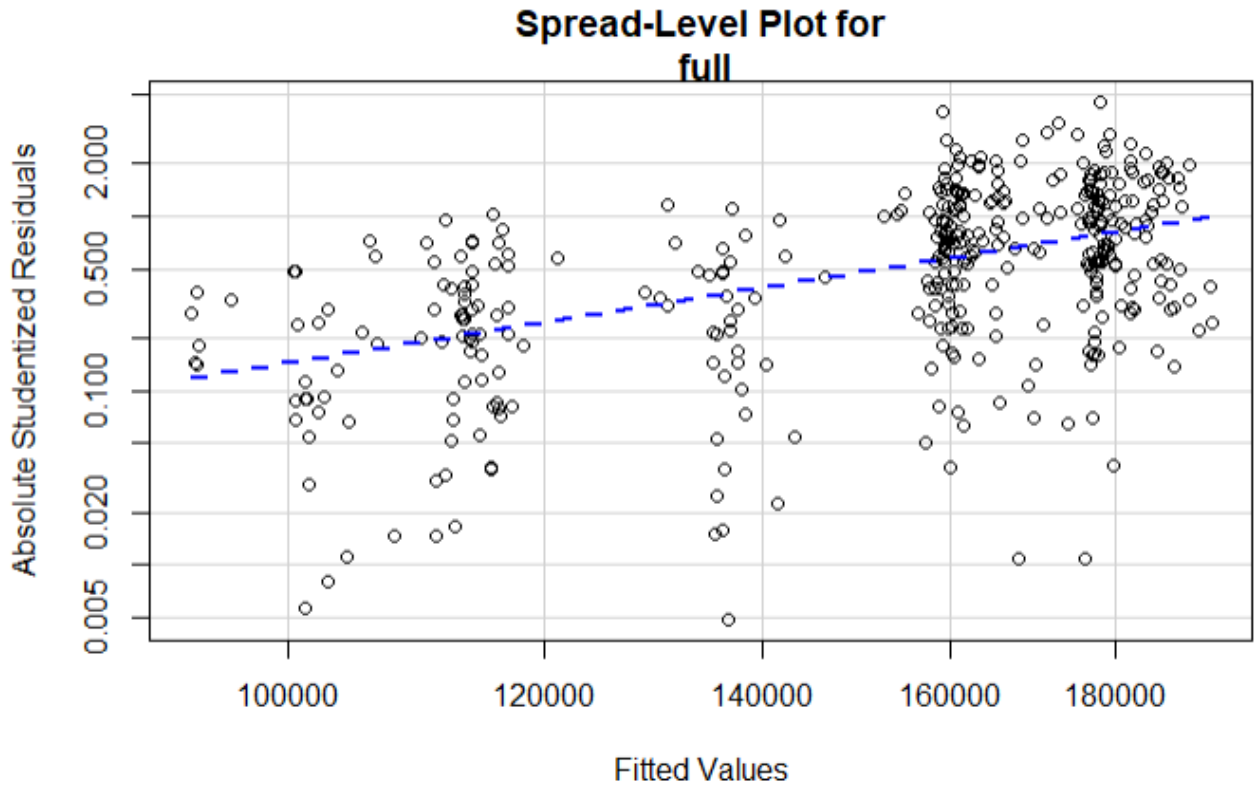
Violation of Assumptions - Normality





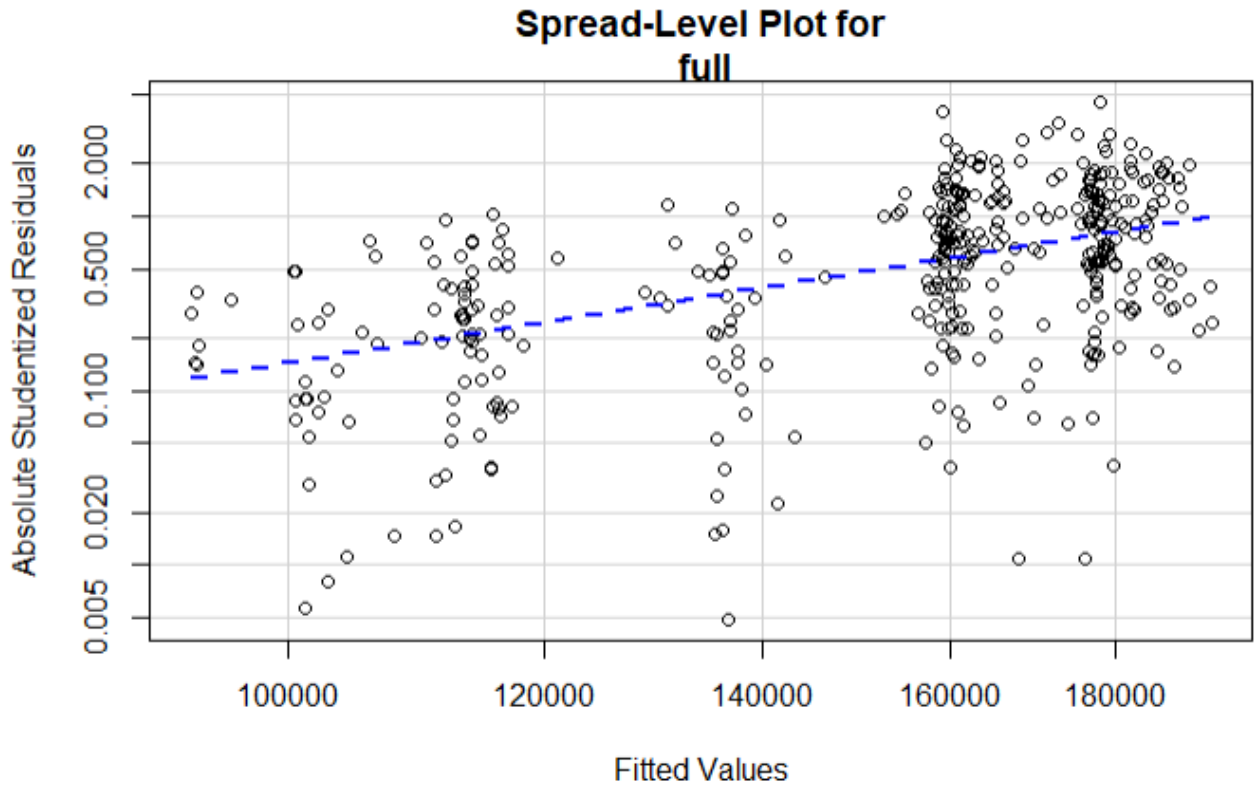
Need to Fix	To fix…
Unusual Data	Remove some influential points
Multicollinearity	Choose best predictor
Normality	none
Non-Constant Variance	
Non-Linearity	

Violation of Assumptions – Non-Constant Variance



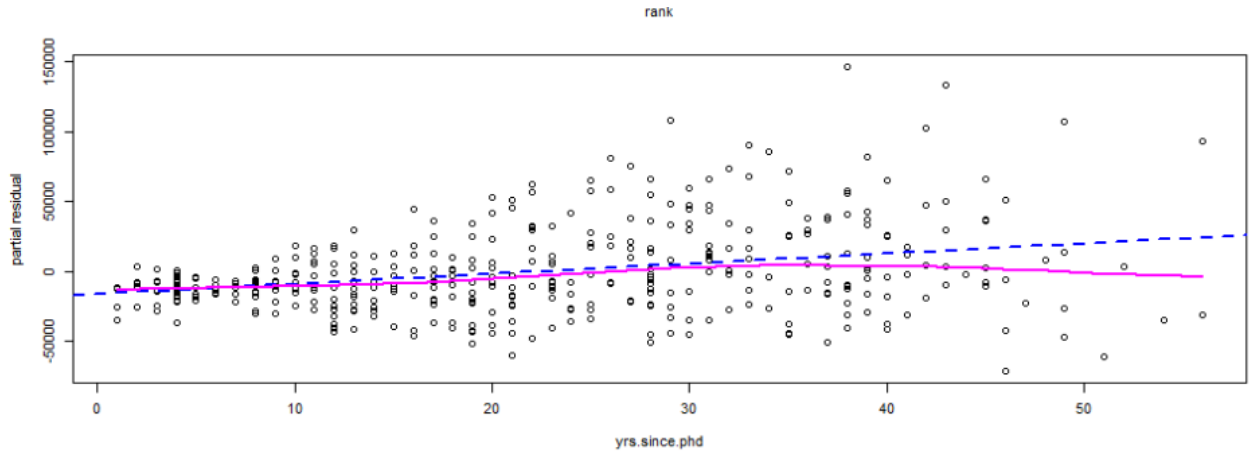
Need to Fix	To fix…
Unusual Data	Remove some influential points
Multicollinearity	Choose best predictor
Normality	none
Non-Constant Variance	???
Non-Linearity	

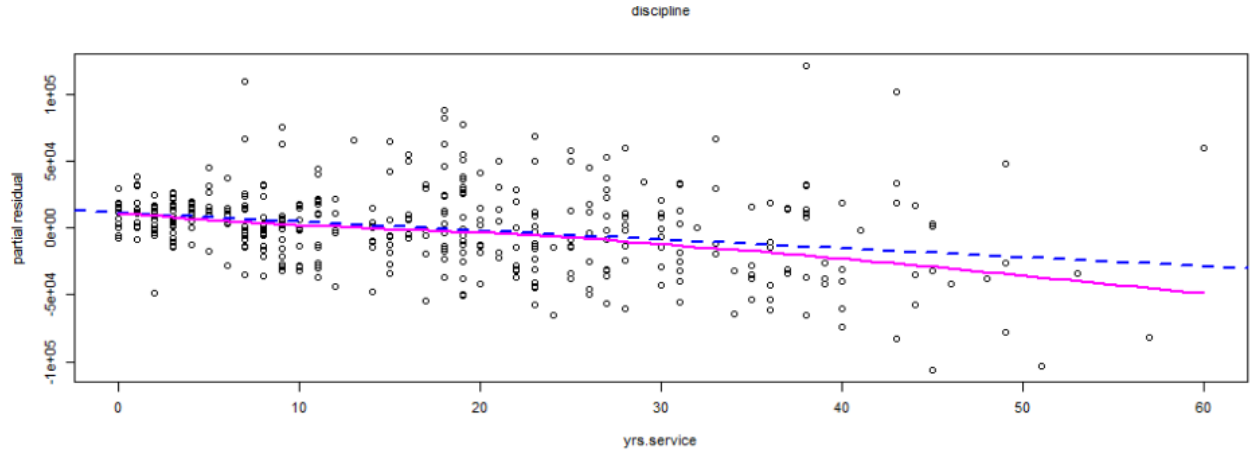
Violation of Assumptions – Non-Constant Variance



Need to Fix	To fix…
Unusual Data	Remove some influential points
Multicollinearity	Choose best predictor
Normality	none
Non-Constant Variance	Transform y to y^{-3}
Non-Linearity	

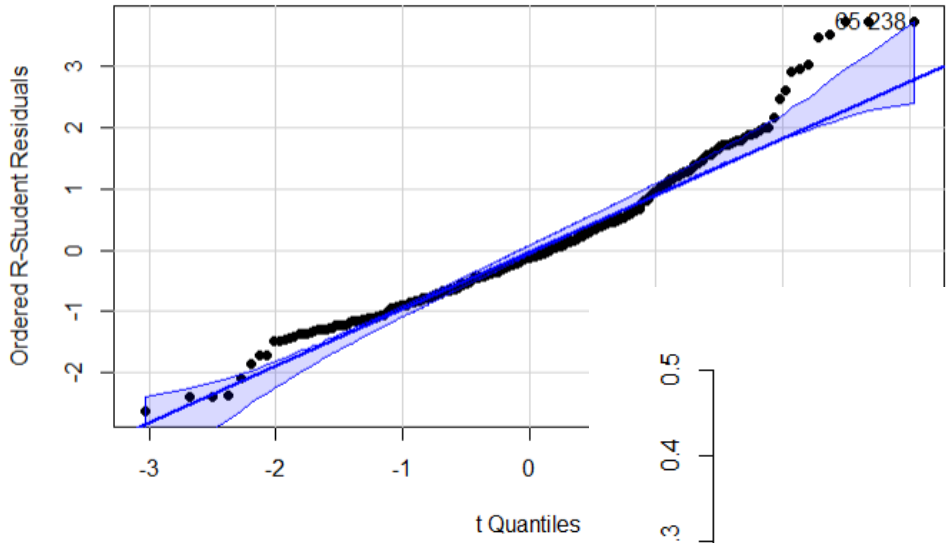
Violation of Assumptions – Non-Linearity

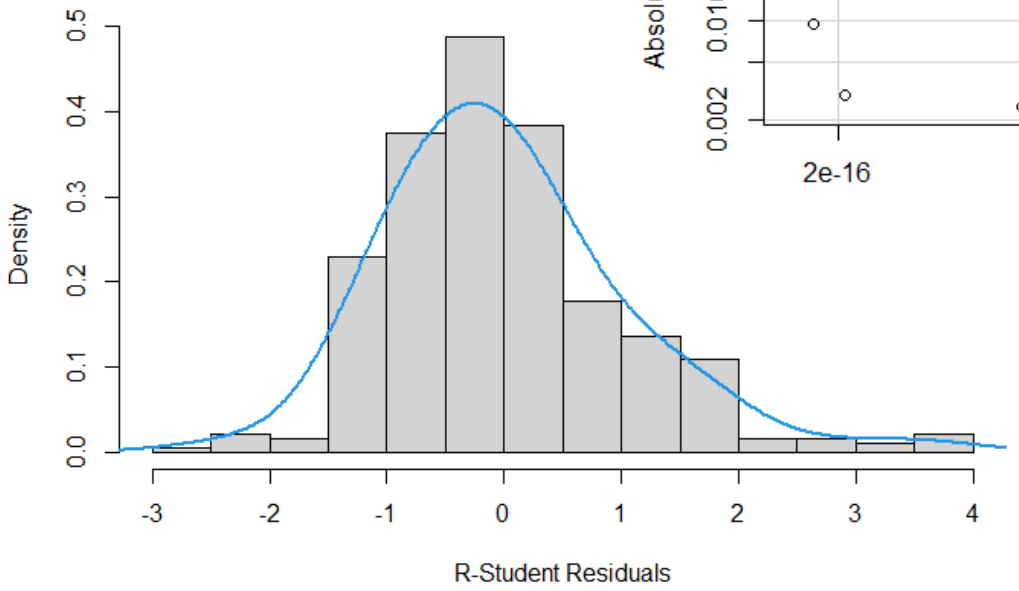


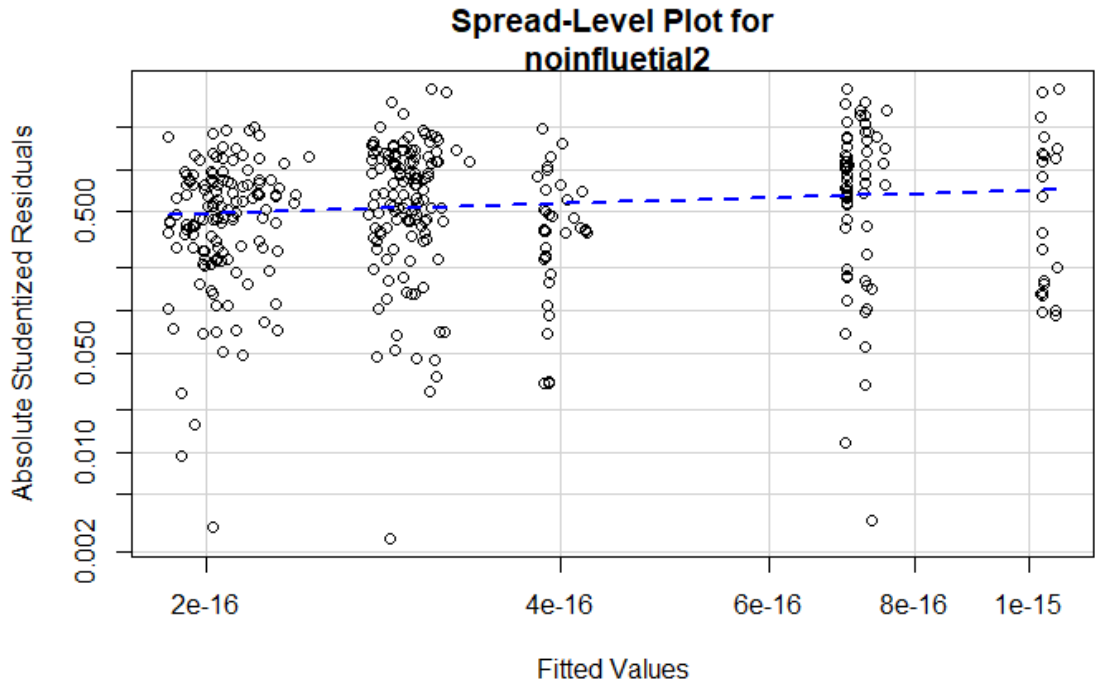


Need to Fix	To fix…
Unusual Data	Remove some influential points
Multicollinearity	Choose best predictor
Normality	none
Non-Constant Variance	Transform y to y^{-3}
Non-Linearity	none

Fixing Initial Model







Final Selected Model

```
lm(formula = (salary)^(-3) ~ discipline * rank + discipline *
    FRank + yrs.service * rank + yrs.service * FRank + sex, data = Salaries_noIP)

Residuals:
      Min        1Q    Median        3Q       Max
-4.184e-16 -1.037e-16 -1.819e-17  7.561e-17  5.836e-16

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.513e-16  4.629e-17  14.071  < 2e-16 ***
discipline1       -3.221e-16  4.097e-17  -7.861 4.11e-14 ***
rank1              4.290e-16  6.257e-17   6.856 2.94e-11 ***
FRank1            -3.709e-16  5.179e-17  -7.161 4.28e-12 ***
yrs.service        6.488e-18  2.795e-18   2.321   0.0208 *
sex1               1.715e-17  2.660e-17   0.645   0.5194
discipline1:rank1  9.666e-18  5.621e-17   0.172   0.8636
discipline1:FRank1 2.371e-16  4.491e-17   5.280 2.20e-07 ***
rank1:yrs.service -4.133e-17  1.308e-17  -3.161   0.0017 **
FRank1:yrs.service -6.004e-18  2.911e-18  -2.062   0.0399 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.477e-16 on 374 degrees of freedom
Multiple R-squared:  0.7185,    Adjusted R-squared:  0.7118
F-statistic: 106.1 on 9 and 374 DF,  p-value: < 2.2e-16
```

Violations	issues
Unusual Data	yes
Multicollinearity	no
Normality	no
Non-Constant Variance	no
Non-Linearity	no

Prediction

Description: df [8 × 10]

	Professor <chr>	Rank = Asst Prof <fctr>	Rank = Full Prof <fctr>	discipline <fctr>	yrs.since.phd <dbl>	yrs.service <dbl>	Gender = Male <fctr>	fit <dbl>	Lower <dbl>	Upper <dbl>
1	Yu	1	0	0	4	2	0			
2	Clough	0	1	0	37	37	1			
3	Hamilton	0	0	0	10	8	1			
4	Ongie	1	0	0	6	2	0			
5	Maadooliat	0	0	0	11	9	0			
6	Rowe	0	1	0	24	8	0			
7	Sanders	0	1	0	18	7	1			
8	Spiller	0	0	0	17	14	1			

?

References

1. https://www.aier.org/cost-of-living-calculator/
2. Library(CarData); View(Salaries)
3. Professor Information
    1. On Marquette faculty page, their personal website, or linked in

Questions?