UniFi & IMT

Master II - Data Science and Statistical Learning

Modelling, Statistical Inference and Graphical Models

“Salary” dataset analysis

Prof. G. M. Marchetti

Gianmarco Santoro

TABLE OF CONTENTS

Info about "Salary" dataset

The data file concerns salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue.

All persons in the data hold tenured or tenure track positions; temporary faculty are not included.

The variables include:

Variable	Description
Sex	A factor with levels Male and Female
Rank	A factor with levels Asst, Assoc, and Prof
Year	Number of years in current rank
Degree	Last degree: a factor with levels PhD and MS
YSDeg	Number of years since highest degree was earned
Salary	Academic year salary in US $

Table 1 - Variables description

1. Analysis

1.1 Graphical summaries

Get appropriate graphical summaries of the data and discuss the graphs.

Looking at the data-frame just created, we can notice:

1. The are no missing data, all the 52 measures have all the available variables: X, degree, rank, sex, year, ysdeg, salary. Which are described above.
2. Notice that X is just a progressive number in the dataset, not an explanatory variable as others.

```
summary(salary)
```

X	degree	rank	sex	year	ysdeg	salary
Min. : 1.00	Length:52	Length:52	Min. :1.0	Min. : 0.000	Min. : 1.00	Min. :15000
1st Qu.:13.75	Class :character	Class :character	1st Qu.:1.0	1st Qu.: 3.000	1st Qu.: 6.75	1st Qu.:18247
Median :26.50	Mode :character	Mode :character	Median :1.5	Median : 7.000	Median :15.50	Median :23719
Mean :26.50			Mean :1.5	Mean : 7.481	Mean :16.12	Mean :23798
3rd Qu.:39.25			3rd Qu.:2.0	3rd Qu.:11.000	3rd Qu.:23.25	3rd Qu.:27258
Max. :52.00			Max. :2.0	Max. :25.000	Max. :35.00	Max. :38045

Table 2 - Variables summary [1]

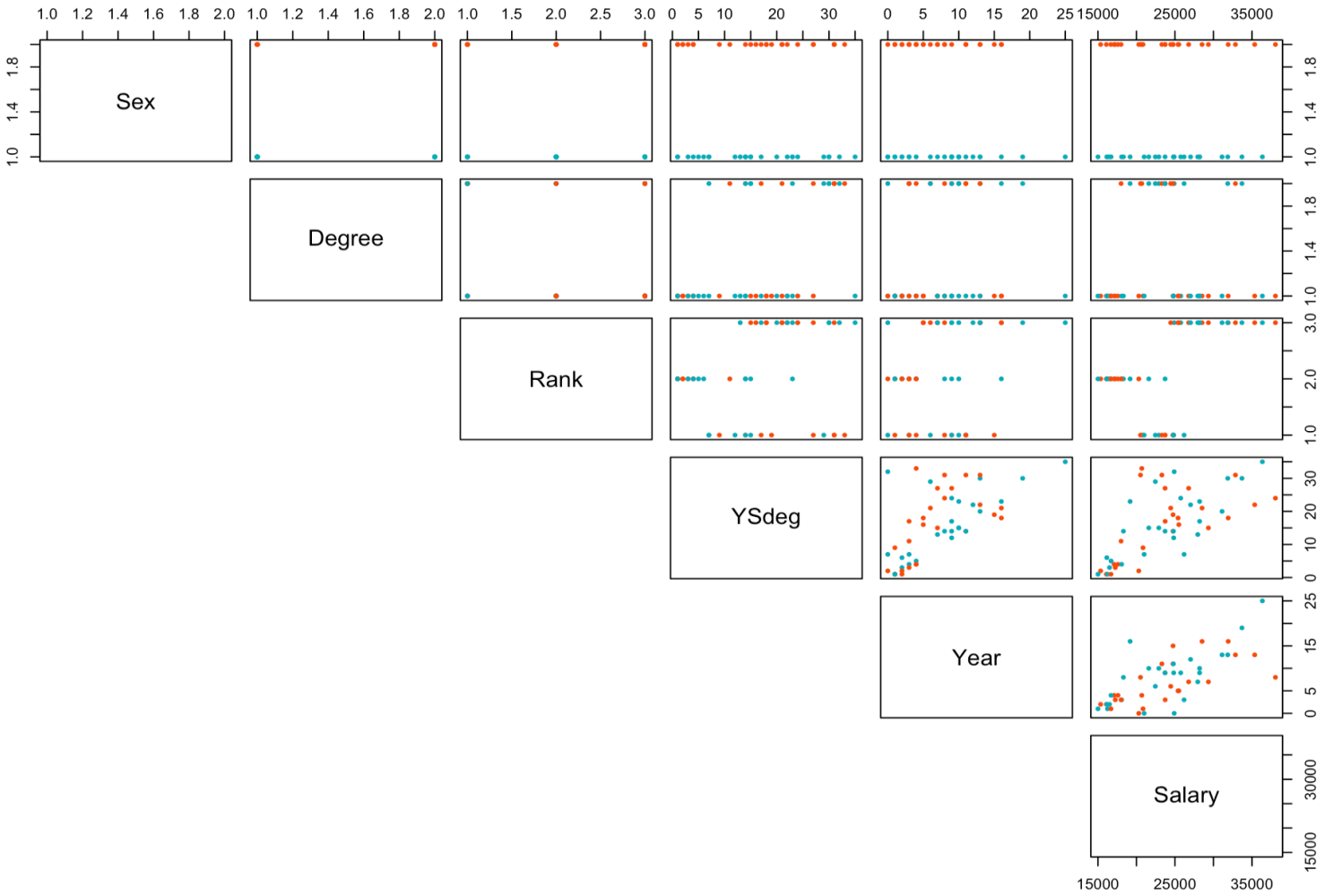
```
head(salary)
```

	X	degree	rank	sex	year	ysdeg	salary
1		Masters	Prof	Male	25	35	36350
2		Masters	Prof	Male	13	22	35350
3		Masters	Prof	Male	10	23	28200
4		Masters	Prof	Female	7	27	26775
5		PhD	Prof	Male	19	30	33696
6		Masters	Prof	Male	16	21	28516

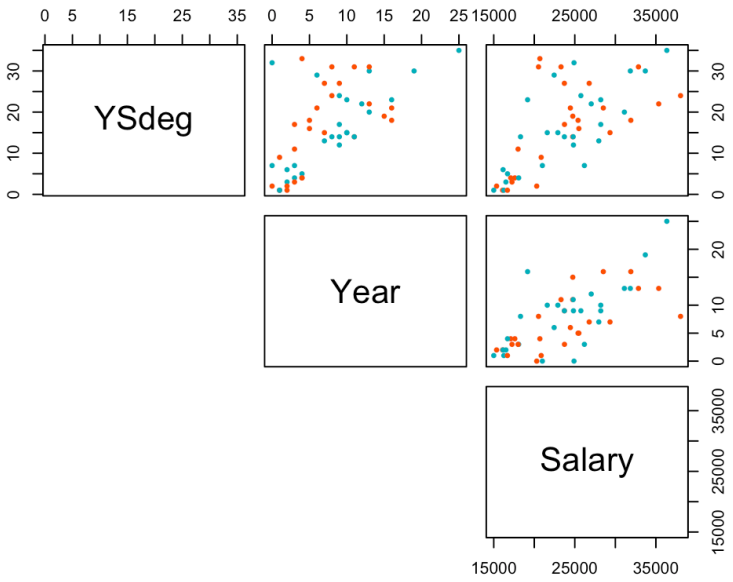
Table 3 - Variables head

In the next pages there are some data displayed.

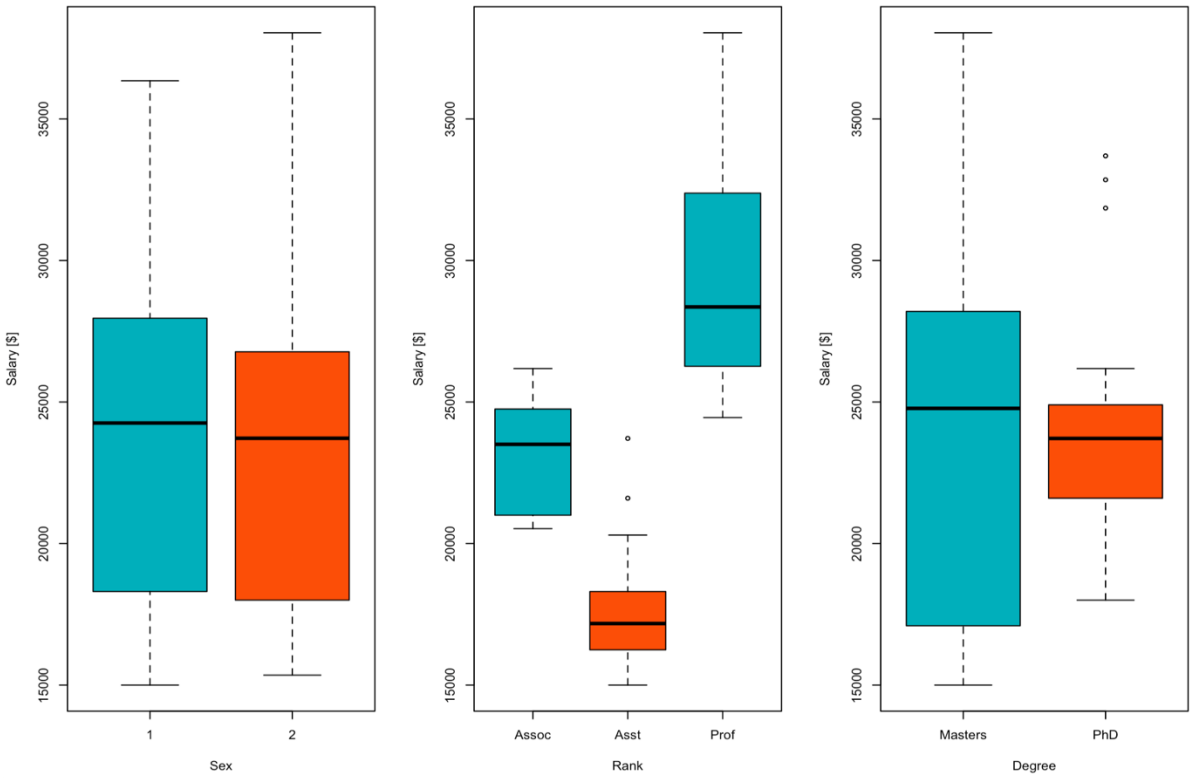
[1] *Sex values: Female = 1, Male = 2.*



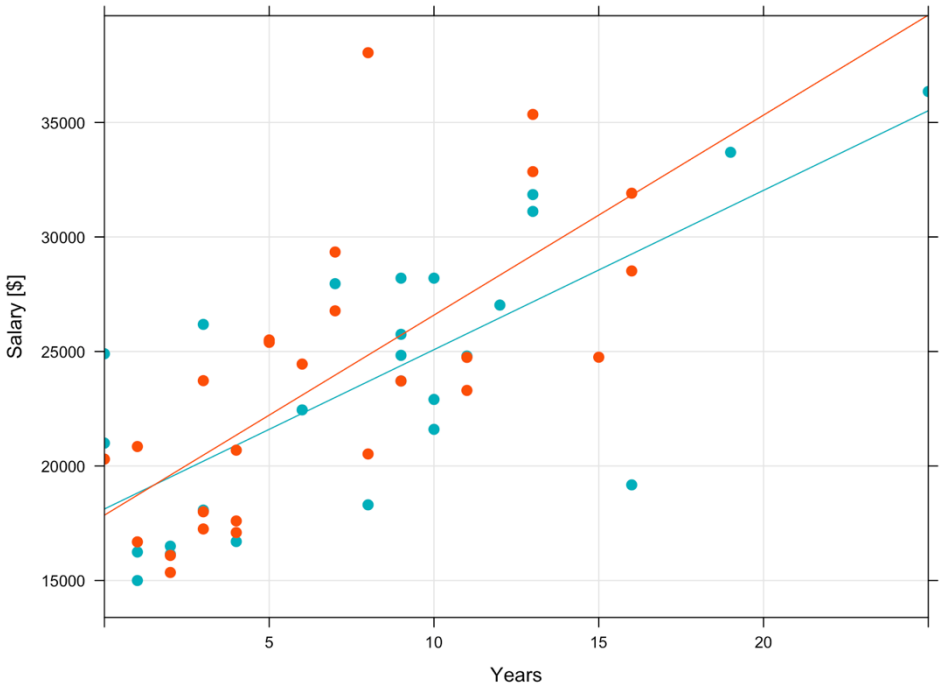
Graph 1 – All variables' scatterplots



Graph 2 – Zoom over some meaning graphs



Graph 3 – Categorical variables boxplots



Graph 4 – Salary vs years, by Sex

I've displayed data by scatter and box plots, where the colors are sex indicator for data points: orange for females and light blue for males.

We can see:

1. Generally, females have lower years of profession and lower rank.
2. The patterns of salary's growth by years look similar for females and males. The regression line indicates that females may have a steeper slope than males in the Years vs Salary scatterplot. However, it is worth noting that this observation could be influenced by the presence of outliers in the data, where two data points, one representing a female and the other a male, appear to deviate significantly from the overall trend:
    a. Masters, Prof, Female, 8 year, 24 ysdeg, 38045 $
    b. PhD, Asst, Male, 16 year, 23 ysdeg, 19175 $

    Further investigation is needed to determine the extent of their impact on the regression results.
3. Female salaries are usually a little lower than male salaries, but looking the whole range, the situation reverses and females can reach higher values.
4. Salary clearly increases with rank, where most Assistants are Female, while the majority of Associated and Prof are Males.
5. Finally, turning to the impact of degree level on salaries, it can be observed that:
    a. Individuals holding a master's degree generally have higher and more variable salaries.
    b. On the other hand, those with a PhD tend to have more consistent or stable salaries.

1.2 Hypothesis test

Test the hypothesis that the mean salary for men and women is the same.

What alternative hypothesis do you think is appropriate?

To test hypothesis that the mean salary for men and women is the same, I can perform a t-test, which can be evaluated using linear regression by fitting an intercept and a dummy variable for sex.

Thus, I've created a linear regression model with the dependent variable Salary and the independent variable Sex, of the kind:

$$Y = \beta_0 + \beta_1 M + \varepsilon$$

```
model0 <- lm(formula = salary ~ sex, data = salary)
```

Finally, the estimated coefficients of the linear regression model, including the intercept and slope, and their standard errors, T-values, and P-values:

```
summary(model0) $ coef
confint(model0)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21357.143	1545.326	13.82047	1.052424e-18
sexMale	3339.647	1807.716	1.84744	7.060394e-02

Table 4 - Salary~Sex

We can observe that the P-values gives back a level about 0.07, which is in between form statistically significant and non-significant, so the difference over Male and Female is not peculiarly significant, as a first analysis.

1.3 Further tests

1.3.1 Rank and sex

Obtain a test of the hypothesis that salary adjusted for years in current rank, highest degree, and years since highest degree is the same for each of the three ranks, versus the alternative that the salaries are not the same. Test to see if the sex differential in salary is the same in each rank (i.e., test interaction).

The first hypothesis test asks to investigate that Rank has no effect on Salary (effect of Rank is null), this can be measured by an analysis of variance test.

```
model1 <- lm(salary ~ year + ysdeg + degree, salary)
model2 <- update(model1, ~. + factor(rank))
anova(model1, model2)
```

```
Analysis of Variance Table

Model 1: salary ~ year + ysdeg + degree
Model 2: salary ~ year + ysdeg + degree + factor(rank)
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1     48 672102002
2     46 267993336  2 404108665 34.682 6.544e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 5 – ANOVA test on Rank

The tiny P-value for comparing model 1 to model 2 shows that there is a Rank effect, so there is strong evidence against H_0, $P - value \ll 0.01$, thus rank is highly significant variable.

The second one, ask to test Sex by Rank interaction.

```
model3 <- update(model1, ~. + sex)
model4 <- update(model3, ~. + sex : factor(rank))
anova(model1, model3, model4)
```

```
Analysis of Variance Table

Model 1: salary ~ year + ysdeg + degree
Model 2: salary ~ year + ysdeg + degree + sex
Model 3: salary ~ year + ysdeg + degree + sex + sex:factor(rank)
  Res.Df       RSS Df Sum of Sq       F    Pr(>F)
1     48 672102002
2     47 658649047  1  13452955  2.3319    0.1341
3     43 248070090  4 410578957 17.7922 1.098e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 6 – ANOVA test on Sex~Rank

These tests exhibit that Sex effect is irrelevant, showing high P-value, while the Sex differential depends on Rank.

So, let's test the differential in salary of Sex-Rank interaction.

```
model5 <- lm(salary ~ year + ysdeg + degree + factor(rank) + sex : factor(rank), salary)
summary(model5)
```

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 22064.74    2522.76   8.746 4.27e-11 ***
year                          522.14     105.53   4.948 1.20e-05 ***
ysdeg                        -148.61      86.83  -1.711  0.09421 .
degreePhD                    1501.51    1029.80   1.458  0.15209
factor(rank)Asst            -5617.29    2526.52  -2.223  0.03150 *
factor(rank)Prof             7432.32    2163.46   3.435  0.00132 **
factor(rank)Assoc:sexMale    -942.58    2194.85  -0.429  0.66974
factor(rank)Asst:sexMale     -444.30    1153.54  -0.385  0.70202
factor(rank)Prof:sexMale    -2954.54    1609.28  -1.836  0.07329 .
```

Table 7 – Test on dummies Rank & Sex~Rank

These results are differences compared to an “Associated Female Professor” as a reference.

We can see that the most significant variables are, in order of importance, year (very high) and rank (high). Whitin the latter shows higher significance in being Prof (very high) and Asst (high) instead of being Assoc.

And, again, it’s hard to say that sex variable makes any statistically significant effect, apart from Prof – Male which could be a little significant.

1.3.2 Difference in salary male-female

Assuming no interactions between sex and the other predictors, obtain a 95% confidence interval for the difference in salary between males and females.

```
model6 <- lm(salary ~ ., salary)
summary(model6) $ coef
confint(model6)
```

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	21060.11017	3044.29467	6.9178948	1.508306e-08	14924.7374	27195.4829
X	27.34488	63.75547	0.4289024	6.700869e-01	-101.1458	155.8356
degreePhD	1438.04146	1034.55034	1.3900159	1.715167e-01	-646.9577	3523.0407
rankAsst	-5498.57346	1251.92956	-4.3920789	6.963734e-05	-8021.6717	-2975.4752
rankProf	6127.86392	1240.56564	4.9395725	1.176869e-05	3627.6682	8628.0597
sexMale	-1089.63740	951.05696	-1.1457120	2.581076e-01	-3006.3668	827.0920
year	503.23124	114.52094	4.3942289	6.916087e-05	272.4294	734.0330
ysdeg	-118.31432	79.54898	-1.4873142	1.440636e-01	-278.6348	42.0061

Table 8 – 95% conf. interval

The R-output shows the 95% confidence intervals for the estimated coefficients of the linear regression model, which means that we can be 95% confident that the true coefficient for sexMale falls within this interval.

Since this confidence interval includes zero, we cannot conclude that there is strong evidence that the coefficient for Sex variable is not significantly different from zero, that means it has a good chance of finding no difference or correlation in data. Thus, we cannot reject the null hypothesis that the true coefficient for sexMale is zero.

This is also highlighted by high P-value in the statistical test, meaning that results could have occurred under the null hypothesis of no relationship between variables.

It is also important to note that the sign of the estimated coefficient indicates the direction of the relationship between the predictor variable and the response variable. So, in case of the coefficient for sexMale is negative, indicating that males could have lower salaries than females, after adjusting for the other predictor variables in the model.

1.4 Sex based salary discrimination discussion

Finkelstein (1980), in a discussion of the use of regression in discrimination cases, wrote, "[a] variable may reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of the position or status, the variable may be 'tainted.'" Thus, for example, if discrimination is at work in promotion of faculty to higher ranks, using rank to adjust salaries before comparing the sexes may not be acceptable to the courts.

Fit two mean functions, one including Sex, Year, YSdeg and Degree, and the second adding Rank. Summarize and compare the results of leaving out rank effects on inferences concerning differential in pay by sex.

Fitting two mean functions:

Initially including variables sex, year, ysdeg, degree:

```
model7 <- lm(salary ~ sex + year + ysdeg + degree, salary)
summary(model7) $ coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15897.0274	1259.86617	12.6180286	1.059158e-16
sexMale	1286.5443	1313.08854	0.9797849	3.322090e-01
year	351.9686	142.48087	2.4702865	1.718541e-02
ysdeg	339.3990	80.62097	4.2098109	1.143695e-04
degreePhD	-3299.3488	1302.51952	-2.5330514	1.470396e-02

Table 9 – Function including sex, year, ysdeg, degree

And finally adding rank:

```
summary(model8 <- update(model7, ~. + rank)) $ coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22204.7816	1451.25158	15.300436	1.891948e-19
sexMale	-1166.3731	925.56888	-1.260169	2.141043e-01
year	476.3090	94.91357	5.018345	8.653790e-06
ysdeg	-124.5743	77.48628	-1.607695	1.148967e-01
degreePhD	1388.6133	1018.74688	1.363060	1.796454e-01
rankAsst	-5292.3608	1145.39802	-4.620543	3.216343e-05
rankProf	5826.4032	1012.93301	5.752012	7.278088e-07

Table 10 – Function including sex, year, ysdeg, degree, rank

In both cases Year is very significant variable, while Sex is not significant.

Moreover, the instability in the sign of Sex variable could be an indication of collinearity between Sex and the added variable, which occurs when two or more predictor variables in a regression model are highly correlated with each other, causing instability in the estimates of their regression coefficients. In such cases, the sign of the regression coefficients can change when different combinations of predictor variables are included in the model, such this. Alternatively, the change in sign could be due to a confounding variable that is associated with both the original variable and the outcome.

A further examination in the relationships between the variables in the model is needed to understand the reason for the change in estimate value and to interpret the results of the model appropriately.

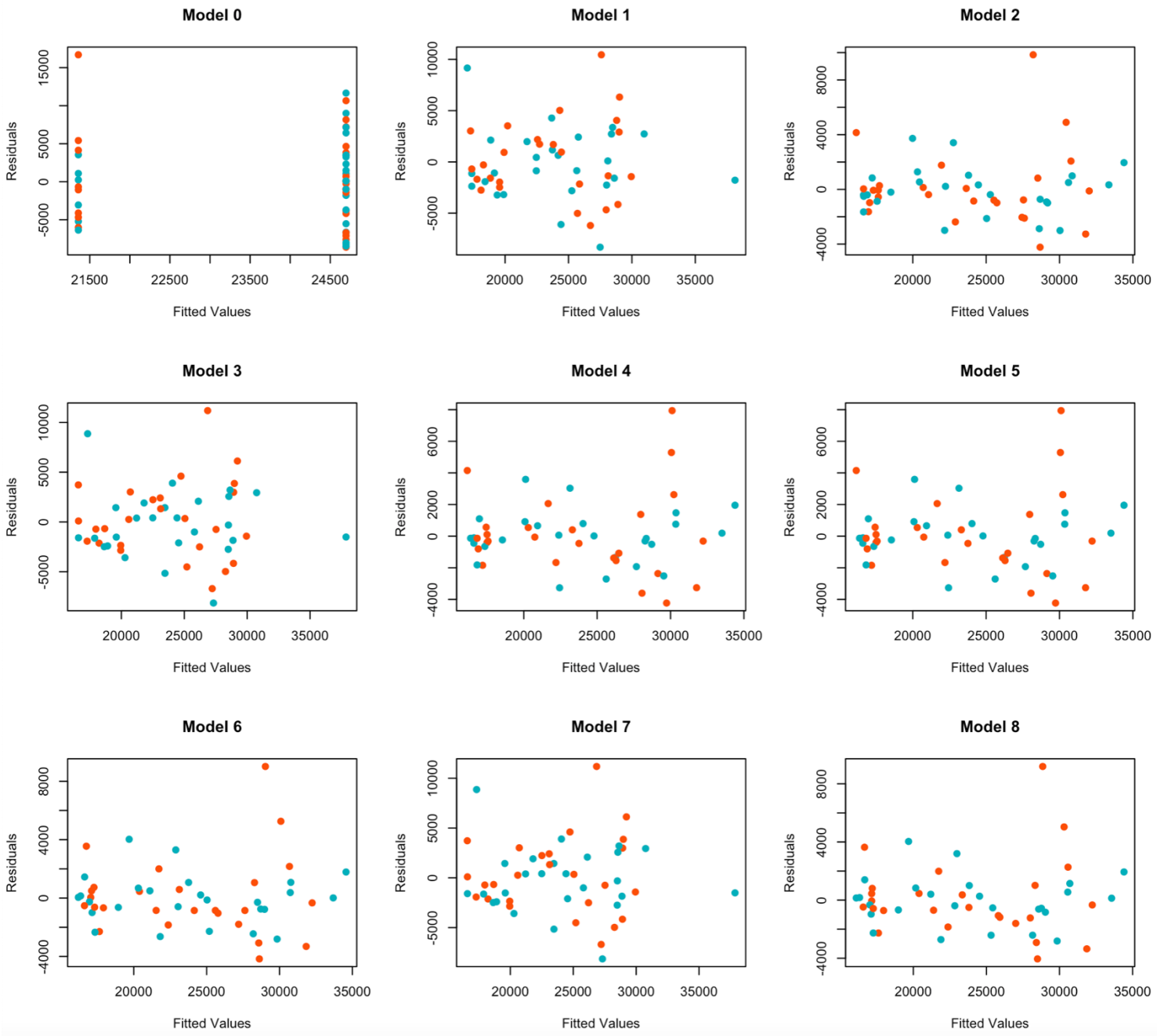
In the first fit, YSDeg and Degree appear to be relevant but their significance changes, along with the signs, in the second fit, where their P-values show them as irrelevant. This could be caused by interaction with Rank variable, so those models could not be perfectly additive.

Rank variable, as already shown above, is a high significant variable.

Thus, it is possible to say that the model with Rank is the best fitting one, where we can try to reduce the model, deleting sex, ysdeg and degree variables.

1.5 Extra

I've finally displayed Residuals vs Fitted Values plots has a further visual metric to see how models perform.



Graph 5 – Residuals vs Fitted

In just few words, again we can see how models with a better fit have smaller residuals, the outlier deviates from 0 and variance of residuals appears arguably not constant.

2. R code

```
# ----
# Modelling, Statistical Inference and Graphical Models - by Prof. G.M.Marchetti
# What's in this script? --> Data set analysis by Gianmarco Santoro



# ----
# Info about "Salary" data set:
# The data file concerns salary and other characteristics of all faculty in a
# small Midwestern college collected in the early 1980s for presentation in legal
# proceedings for which discrimination against women in salary was at issue.

# All persons in the data hold tenured or tenure track positions;
# temporary faculty are not included.

# The variables include:
# - degree, a factor with levels PhD and MS;
# - rank,   a factor with levels Asst, Assoc, and Prof;
# - sex,    a factor with levels Male and Female;
# - Year,   years in current rank;
# - ysdeg,  years since highest degree, and
# - salary, academic year salary in dollars.



# ----
# 0) Load the data

require(lattice)

salary <- read.csv("salary.csv")
salary.plot <- salary

# salary$sex <- as.numeric(as.factor(c("Female", "Male")))
salary.plot$sex <- as.numeric(as.factor(c("Female", "Male")))

# salary$degree <- as.numeric(as.factor(c("Masters", "PhD")))
# salary$rank <- as.numeric(as.character(c("Asst", "Assoc", "Prof")))

salary
summary(salary)
head(salary)



# ----
# 1) Get appropriate graphical summaries of the data and discuss the graphs.
```

```
# Colors
my_cols <- c("#00AFBB","#FC4E07")

# Scatterplots
plotvalues <- data.frame(#"Sex"     = salary.plot$sex,
                         #"Degree"  = salary.plot$degree,
                         #"Rank"    = salary.plot$rank,
                         "YSdeg"   = salary.plot$ysdeg,
                         "Year"    = salary.plot$year,
                         "Salary"  = salary.plot$salary)

plot(plotvalues,
     cex = 0.5,
     pch = 19,
     col = my_cols[salary.plot$sex],
     lower.panel = NULL)


# Boxplots for categorical variables
par(mfrow=c(1, 3))

boxplot(salary ~ sex,
        salary,
        xlab= "Sex",
        ylab="Salary [$]",
        # main="Sex",
        col = my_cols[salary.plot$sex])

boxplot(salary ~ rank,
        salary,
        xlab= "Rank",
        ylab="Salary [$]",
        col = my_cols[salary.plot$sex])

boxplot(salary ~ degree,
        salary,
        xlab= "Degree",
        ylab="Salary [$]",
        col = my_cols[salary.plot$sex])

# Plots
xyplot(salary ~ year, # | sex,
       data  = salary.plot,
       group = sex,
       xlim  = range(salary$year),
       xlab  = "Years",
```

```
       ylab  = "Salary [$]",
       cex   = 1.0,
       pch   = 19,
       col   = my_cols[salary.plot$sex],
       type  = c("p", "g", "r"))  # g = mesh, r = regr lin

# xyplot(salary ~ ysdeg,
#        data  = salary.plot,
#        group = sex,
#        xlim  = range(salary$ysdeg),
#        xlab  = "YSDeg",
#        ylab  = "Salary [$]",
#        cex   = 1.0,
#        pch   = 19,
#        col   = my_cols[salary.plot$sex],
#        type  = c("p", "g", "r"))



# ----
# 2) Test the hypothesis that the mean salary for men and women is the same.
#    What alternative hypothesis do you think is appropriate?

# Test
model0 <- lm(formula = salary ~ sex, data = salary)
summary(model0) $ coef
confint(model0)



# ----
# 3a) Obtain a test of the hypothesis that salary adjusted for years in current rank,
#    highest degree, and years since highest degree is the same for each of the three
#    ranks, versus the alternative that the salaries are not the same. Test to see
#    if the sex differential in salary is the same in each rank (i.e. test interaction).

# test 1
model1 <- lm(salary ~ year + ysdeg + degree, salary)
model2 <- update(model1, ~. + factor(rank))
anova(model1, model2)

# Test 2
model3 <- update(model1, ~. + sex)
model4 <- update(model3, ~. + sex : factor(rank))
anova(model1, model3, model4)

# Test sex-rank interactions
model5 <- lm(salary ~ year + ysdeg + degree + factor(rank) + sex : factor(rank), salary)
summary(model5)
```

```
# confint(model5)


# 3b) Assuming no interactions between sex and the other predictors, obtain a 95%
#    confidence interval for the difference in salary between males and females.

model6 <- lm(salary ~ ., salary)
summary(model6) $ coef
confint(model6)["sexMale", , drop = FALSE]


# ----
# 4) Finkelstein (1980), in a discussion of the use of regression in discrimination
#    cases, wrote, “[a] variable may reflect a position or status bestowed by the
#    employer, in which case if there is discrimination in the award of the position
#    or status, the variable may be ‘tainted.’ ” Thus, for example, if discrimination
#    is at work in promotion of faculty to higher ranks, using rank to adjust salaries
#    before comparing the sexes may not be acceptable to the courts.


model7 <- lm(salary ~ sex + year + ysdeg + degree, salary)
summary(model7) $ coef

summary(model8 <- update(model7, ~. + rank)) $ coef


# ----
# Residuals & plot
par(mfrow = c(3, 3))

models <- list(model0, model1, model2, model3, model4, model5, model6, model7, model8)

for (i in 1:length(models)) {

  model <- models[[i]]
  resid_i <- resid(model)

  plot(fitted(model),
       resid_i,
       col = my_cols[salary.plot$sex],
       cex = 1.0,
       pch = 19,
       main = paste("Model", i-1),
       xlab = "Fitted Values",
       ylab = "Residuals")}
```