Evaluating Which Factors Best Predict Faculty Salary Through Multiple Regression Modelling

Johan Chua

06/08/23

Contents

Introduction **2**

Data Analysis **2**
Main Effects Model 2
Model Validity 3
High Leverage/Influence Points & Collinearity 5
Forwards and Backwards Stepwise Regression 7

Conclusion **8**
Department with Highest Salaries 8
Most Important Prediction Factors 8
Predicting Average Salary Per Faculty Rank 8

Appendix **9**
Summary Output for Forwards and Backwards Stepwise Regression 9

Introduction

This paper aims to derive a multiple linear regression model that can quantify which factors have a significant effect on faculty salaries. The data was taken from a study conducted at the University of California, Berkeley, to investigate issues concerning salary equity among faculty.

Data Analysis

Main Effects Model

We first create a model using all the meaningful variables provided in the dataset in order to explain faculty salaries.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.343777e+06	1.153452e+05	28.9892965	0.0000000
GenderMale	4.186858e+02	2.876510e+02	1.4555341	0.1475309
StartYr	-1.683282e+03	5.871393e+01	-28.6692137	0.0000000
DeptCodeFactor2	-1.013653e+03	5.669643e+02	-1.7878605	0.0757397
DeptCodeFactor3	-8.383473e+02	5.293737e+02	-1.5836588	0.1152966
DeptCodeFactor4	-1.020098e+03	5.854275e+02	-1.7424835	0.0833945
DeptCodeFactor5	-1.587231e+03	5.417324e+02	-2.9299171	0.0038997
DeptCodeFactor6	-4.098710e+02	5.359851e+02	-0.7647059	0.4456019
DeptCodeFactor7	-2.994763e+02	6.751003e+02	-0.4436028	0.6579445
DeptCodeFactor8	-5.171259e+02	5.244185e+02	-0.9860940	0.3256139
Begin.Salary	1.362696e+00	9.266570e-02	14.7055104	0.0000000
Expernc	8.182300e+01	5.105885e+01	1.6025234	0.1110628
RankAsstProf	-1.384658e+03	6.837890e+02	-2.0249780	0.0445750
RankInstruct	-2.105955e+03	1.317262e+03	-1.5987366	0.1119026
RankProfessr	2.900507e+03	6.276126e+02	4.6214920	0.0000079

Table 1: Summary of Model 1 Predictors

Only including statistically significant estimates (0.05 significance level), our inital model is:

$Salary = 3344000 - 1683 * StartYr - 1,014 * DeptCodeFactor2 - 1020 * DeptCodeFactor4 - 1587 DeptCodeFactor5 + 1.361 * Begin.Salary - 1385 * RankAsstProf + 2901 * RankProfessr$

Model Validity

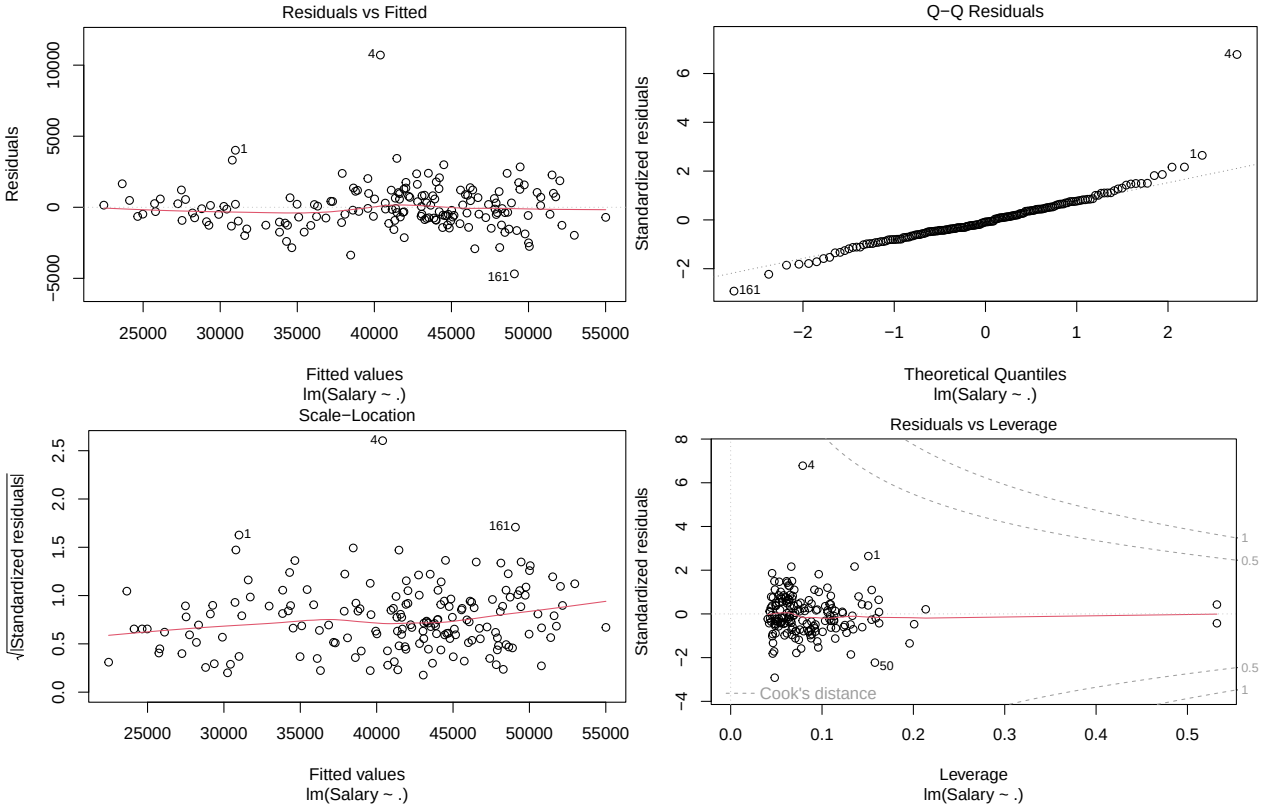
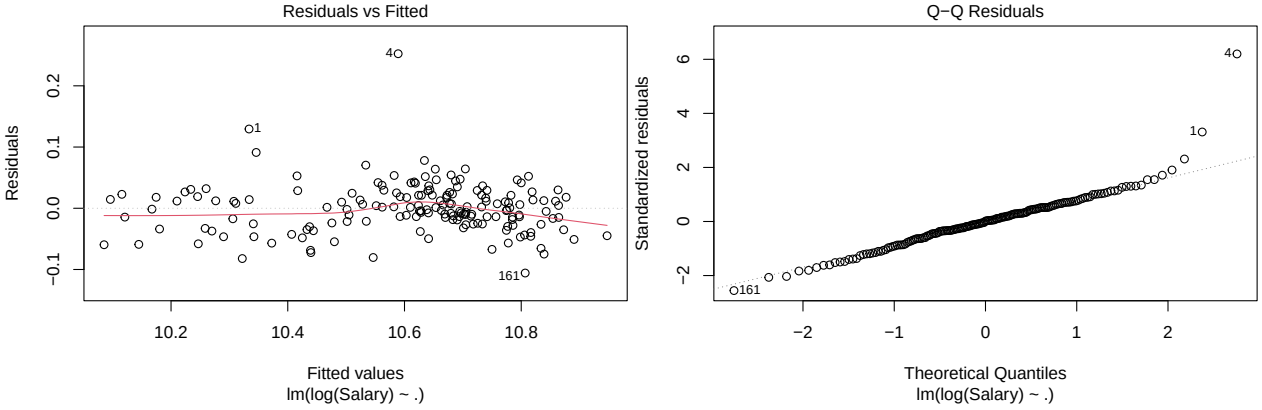


Figure 1: Model 1 (No Transformation) Diagnostic Plots



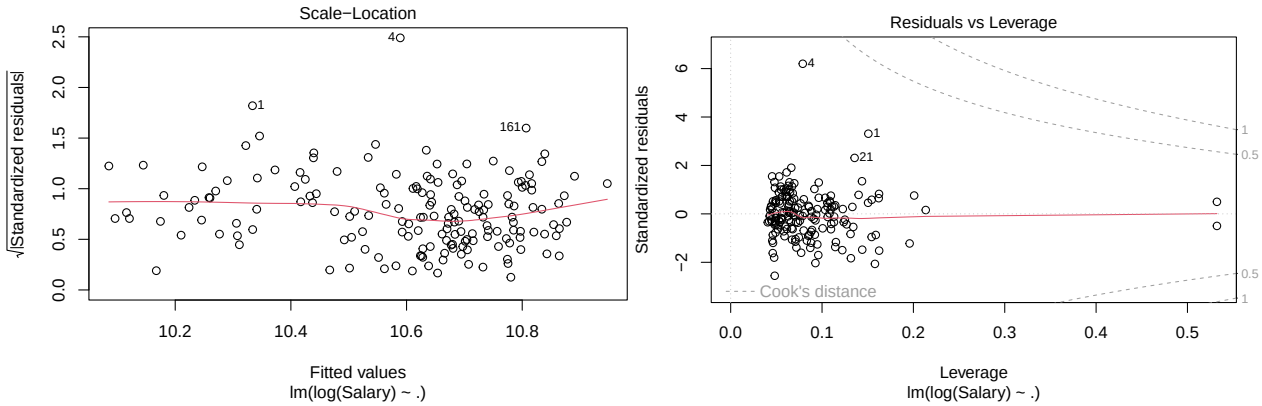


Figure 2: Model 2 (Log Transformation) Diagnostic Plots

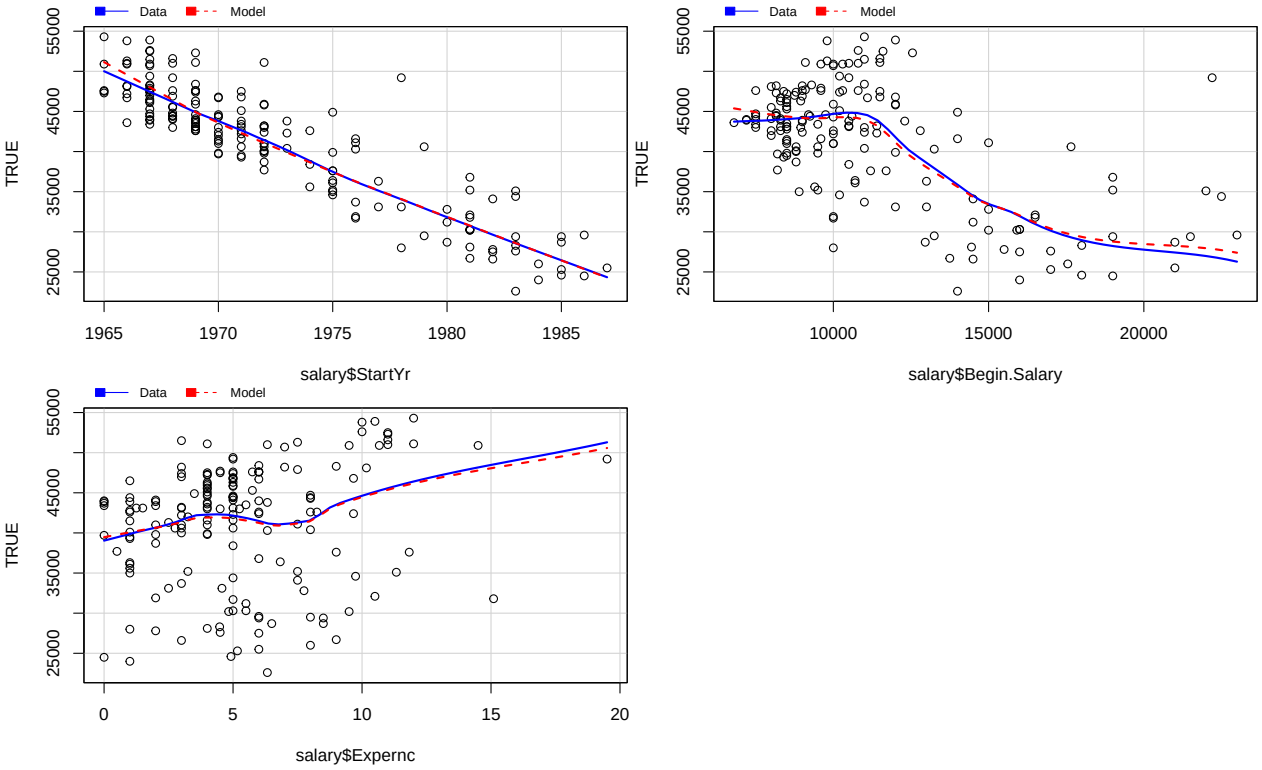


Figure 3: MMP Plots for Model 1

I compared my original model with a log-transformed model (with the hopes it would reduce the effect of the outlier Point 4 and improve normality); however, it ended up producing relatively identical diagnostic plots as the orginal untransformed model, meaning it did not have any effect on model validity. Specifically, the log transform failed to change the relative position of Point 4 to the other points and barely improved normality on the Q-Q plot. Thus, I elected to stick with the original untransformed model stated in (a).

Based on the model's diagnostic and MMP plots, the model is valid.

(1) Residuals vs Fitted: The residuals display no obvious pattern or trend, thus the model satistifes the linearity/structure condition. Additionally, the residuals are scattered randomly and evenly around the horizontal line residuals=0, therefore the model also satisfies the constant variance condition. It is worth noting that Point 4 is an outlier and thus doesn't take away from the model's validity.

(2) Normal Q-Q: Overall, the residuals did not systematically stray from the dashed line, suggesting that the model satisfies the normality condition/the residuals follow a normal distribution. The only points that did not obey normality were Point 161 and Point 4. However, these two points (that were on opposite sides of the plot) weren't enough to consitute a noteworthy trend.

(3) Scale-Location: There is no clear trend to the residuals (they are randomly scattered) which indicates that the model satisifies the constant variance condition.

(4) Residuals vs Leverage: Using a high leverage cutoff of leverage > 0.1754386 and a high influence cutoffs of standardized residual $<$ -4 or standardized residual > 4, we find that there are no bad leverage points.

(5) MMP Plots: In all of the individual plots, we see that the red y_hat trend lines closely follow their respective blue loess lines. This means that the fitted model matches the trend for each individual predictor well. In other words, with respect to the trend, each of the variables are a good fit.

High Leverage/Influence Points & Collinearity

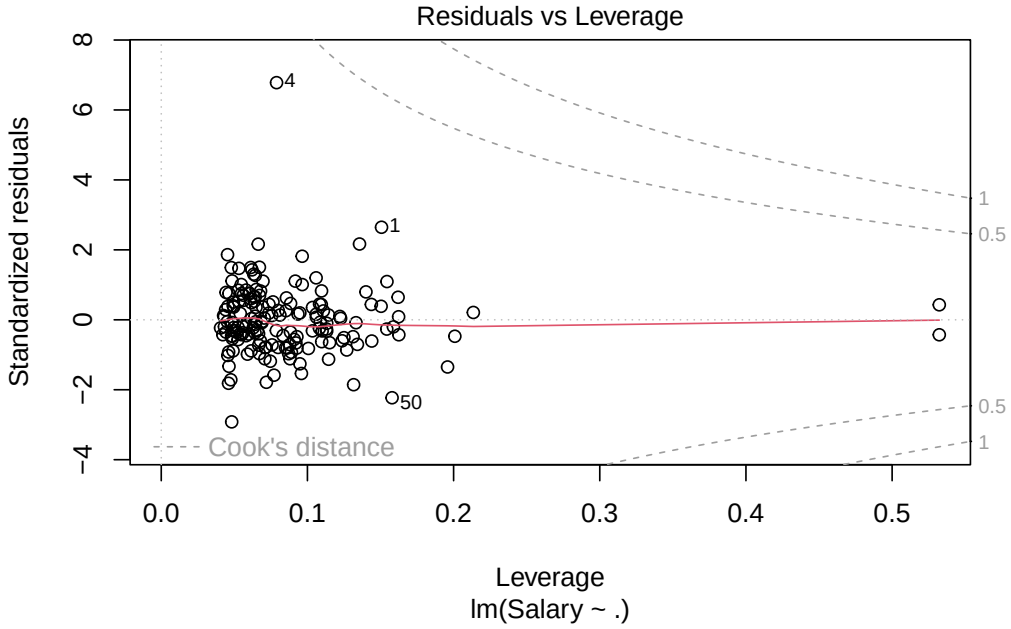


Figure 4: Residuals vs Leverage Plot for Model 1

High.Leverage	High.Influence
5	4
7	4
20	4
22	4
143	4

Table 2: High Leverage and High Influence Residuals

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.238137	1	1.112716
StartYr	7.291878	1	2.700348
DeptCodeFactor	1.549109	7	1.031757
Begin.Salary	6.956330	1	2.637485
Expernc	1.754507	1	1.324578
Rank	2.901984	3	1.194307

Table 3: VIF for Model 1 Predictors

(1) High Leverage Points: These are defined by the criteria leverage > 0.1754386 (from $\frac{2(p+1)}{n}$). Points 5, 7, 20, 22 and 143 are high leverage.

(2) High Influence Points: These are defined by the criteria standardized residual $<$ -4 or standardized residual > 4 for large data sets. Point 4 is high influence as it's standardized residual is 6+.

(3) Bad Leverage Points: Since none of the high leverage points are high influence, there are no bad leverage points.

(4) Collinearity: These are variables with a high GVIF ($GVIF^{\frac{1}{2df}} > 2.2$). StartYr and Begin.Salary are both highly correlated with other parameters in the model.

Forwards and Backwards Stepwise Regression

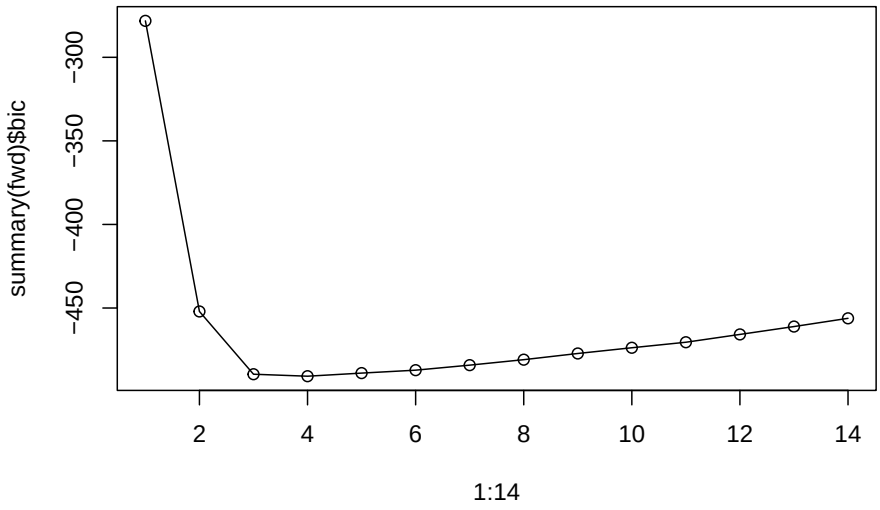


Figure 5: BIC for Forwards Stepwise Regression

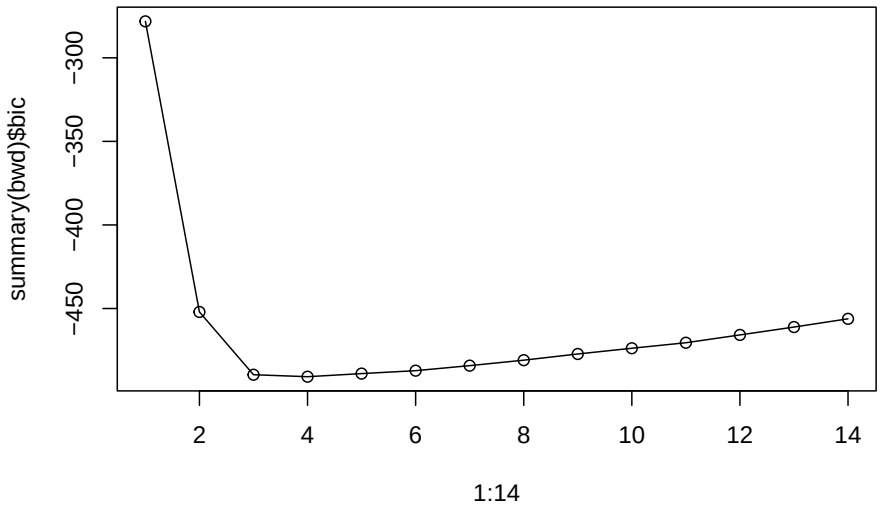


Figure 6: BIC for Backwards Stepwise Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3479021.06952	9.796062e+04	35.5144870	0.0000000
StartYr	-1752.16899	4.984539e+01	-35.1520758	0.0000000
DeptCodeFactor2	-1046.82912	5.703434e+02	-1.8354365	0.0683211
DeptCodeFactor3	-890.86054	5.314353e+02	-1.6763291	0.0956513
DeptCodeFactor4	-1199.25797	5.845004e+02	-2.0517660	0.0418431
DeptCodeFactor5	-1528.92780	5.429653e+02	-2.8158851	0.0054845
DeptCodeFactor6	-485.99443	5.392237e+02	-0.9012854	0.3688084
DeptCodeFactor7	-454.89420	6.769803e+02	-0.6719460	0.5025995
DeptCodeFactor8	-606.88605	5.271133e+02	-1.1513389	0.2513321
Begin.Salary	1.48522	7.120940e-02	20.8570668	0.0000000
RankAsstProf	-1493.30137	6.864051e+02	-2.1755394	0.0310758
RankInstruct	-2251.64575	1.305999e+03	-1.7240793	0.0866494
RankProfessr	2904.73711	6.229640e+02	4.6627686	0.0000066

Table 4: Summary of Final Model Predictors

Both models produced the same set of best model for each number of variables. Both the forward and backwards stepwise regressions report that the model with the lowest BIC (a goodness of fit criteria that rewards the addition of successful fitting variables while punishing model complexity) was the model with 4 variables.

Thus, using the BIC goodness of fit criterion, the best model with the lowest BIC was the model containing the variables StartYr, DeptCodeFactor, Begin.Salary and Rank:

$Salary = 3479000 - 1752 * StartYr - 1047 * DeptCodeFactor2 - 890.9 * DeptCodeFactor3 - 1199 * DeptCodeFactor4 - 1529 * DeptCodeFactor5 + 1.485 * Begin.Salary - 1493 * RankAsstProf - 2252 * RankInstruct + 2905 * RankProfessr$

Conclusion

Department with Highest Salaries

From the model (Table 4), the only DeptCodeFactor parameters that are statistically significant (at a 0.05 significance level) are DeptCodeFactor2, DeptCodeFactor3, DeptCodeFactor4 and DeptCodeFactor5. Since all those parameters have negative estimates, we can conclude that they have, on average, lower salaries than the baseline department, Dept 1. For the other departments (6, 7 and 8), whose estimates were not statistically significant, we can conclude that it is not unlikely for their estimates to be zero, and thus we can say they have the same mean salaries as Dept 1.

Thus, there are two tiers of salaries. Departments 1, 6, 7, 8 have roughly the same average salaries (tied for highest of the departments according to the model); and Departments 3, 2, 4 and 5 have the lowest salaries (in descending order/with department 5 having the lowest mean salaries of all departments).

Most Important Prediction Factors

From the model (Table 4), since all of the parameters are statistically significant in this model, all of the factors included in this model are important: Start Yr, DeptCodeFactor, Begin.Salary, and Rank. Additionally, it is important to note that since 4 out of 7 parameters within the DeptCodeFactor variable were statistically significant, we deem that the DeptCodeFactor variable as a whole is significant and can be considered an important factor in determining salary.

Predicting Average Salary Per Faculty Rank

fit	lwr	upr
42808.32	42166.03	43450.6

Table 5: Confidence Interval for Professor

fit	lwr	upr
38410.28	37092.06	39728.5

Table 6: Confidence Interval for Assistant Professor

The average faculty in this data set started in 1972, works in Dept 8, had a beginning salary of $11289, had 5.243 years of experience, and was a male. Using our final model, we calculate a 95% confidence interval for an average faculty who has rank "Professor" and for "Assistant Professor".

We find that

- 95% Confidence Interval for Professor: (42166.03, 43450.6)
- 95% Confidence Interval for Assistant Professor: (37092.06, 39728.5)

Appendix

Summary Output for Forwards and Backwards Stepwise Regression

```
# forward stepwise regression
which(summary(fwd)$bic == min(summary(fwd)$bic))
```

```
## [1] 4
```

```
summary(fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., data = salary, nvmax = 14, method = "forward")
## 14 Variables  (and intercept)
##                  Forced in Forced out
## GenderMale           FALSE      FALSE
## StartYr              FALSE      FALSE
## DeptCodeFactor2      FALSE      FALSE
## DeptCodeFactor3      FALSE      FALSE
## DeptCodeFactor4      FALSE      FALSE
## DeptCodeFactor5      FALSE      FALSE
## DeptCodeFactor6      FALSE      FALSE
## DeptCodeFactor7      FALSE      FALSE
## DeptCodeFactor8      FALSE      FALSE
## Begin.Salary         FALSE      FALSE
## Expernc              FALSE      FALSE
## RankAsstProf         FALSE      FALSE
## RankInstruct         FALSE      FALSE
## RankProfessr         FALSE      FALSE
## 1 subsets of each size up to 14
## Selection Algorithm: forward
##           GenderMale StartYr DeptCodeFactor2 DeptCodeFactor3 DeptCodeFactor4
## 1  ( 1 )  " "        "*"     " "             " "             " "
## 2  ( 1 )  " "        "*"     " "             " "             " "
## 3  ( 1 )  " "        "*"     " "             " "             " "
## 4  ( 1 )  " "        "*"     " "             " "             " "
## 5  ( 1 )  " "        "*"     " "             " "             " "
## 6  ( 1 )  " "        "*"     " "             " "             " "
## 7  ( 1 )  "*"        "*"     " "             " "             " "
## 8  ( 1 )  "*"        "*"     " "             " "             " "
## 9  ( 1 )  "*"        "*"     "*"             " "             " "
## 10  ( 1 ) "*"        "*"     "*"             " "             "*"
## 11  ( 1 ) "*"        "*"     "*"             "*"             "*"
## 12  ( 1 ) "*"        "*"     "*"             "*"             "*"
## 13  ( 1 ) "*"        "*"     "*"             "*"             "*"
## 14  ( 1 ) "*"        "*"     "*"             "*"             "*"
##           DeptCodeFactor5 DeptCodeFactor6 DeptCodeFactor7 DeptCodeFactor8
## 1  ( 1 )  " "             " "             " "             " "
## 2  ( 1 )  " "             " "             " "             " "
## 3  ( 1 )  " "             " "             " "             " "
## 4  ( 1 )  "*"             " "             " "             " "
## 5  ( 1 )  "*"             " "             " "             " "
## 6  ( 1 )  "*"             " "             " "             " "
## 7  ( 1 )  "*"             " "             " "             " "
## 8  ( 1 )  "*"             " "             " "             " "
```

```
## 9  ( 1 )  "*"             " "             " "             " "
## 10  ( 1 ) "*"             " "             " "             " "
## 11  ( 1 ) "*"             " "             " "             " "
## 12  ( 1 ) "*"             " "             " "             "*"
## 13  ( 1 ) "*"             "*"             " "             "*"
## 14  ( 1 ) "*"             "*"             "*"             "*"
##           Begin.Salary Expernc RankAsstProf RankInstruct RankProfessr
## 1  ( 1 )  " "          " "     " "           " "          " "
## 2  ( 1 )  "*"          " "     " "           " "          " "
## 3  ( 1 )  "*"          " "     " "           " "          "*"
## 4  ( 1 )  "*"          " "     " "           " "          "*"
## 5  ( 1 )  "*"          "*"     " "           " "          "*"
## 6  ( 1 )  "*"          "*"     "*"           " "          "*"
## 7  ( 1 )  "*"          "*"     "*"           " "          "*"
## 8  ( 1 )  "*"          "*"     "*"           "*"          "*"
## 9  ( 1 )  "*"          "*"     "*"           "*"          "*"
## 10  ( 1 ) "*"          "*"     "*"           "*"          "*"
## 11  ( 1 ) "*"          "*"     "*"           "*"          "*"
## 12  ( 1 ) "*"          "*"     "*"           "*"          "*"
## 13  ( 1 ) "*"          "*"     "*"           "*"          "*"
## 14  ( 1 ) "*"          "*"     "*"           "*"          "*"
```

```
# backward stepwise regresion
which(summary(bwd)$bic == min(summary(bwd)$bic))
```

```
## [1] 4
```

```
summary(bwd)
```

```
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., data = salary, nvmax = 14, method = "backward")
## 14 Variables  (and intercept)
##                  Forced in Forced out
## GenderMale           FALSE      FALSE
## StartYr              FALSE      FALSE
## DeptCodeFactor2      FALSE      FALSE
## DeptCodeFactor3      FALSE      FALSE
## DeptCodeFactor4      FALSE      FALSE
## DeptCodeFactor5      FALSE      FALSE
## DeptCodeFactor6      FALSE      FALSE
## DeptCodeFactor7      FALSE      FALSE
## DeptCodeFactor8      FALSE      FALSE
## Begin.Salary         FALSE      FALSE
## Expernc              FALSE      FALSE
## RankAsstProf         FALSE      FALSE
## RankInstruct         FALSE      FALSE
## RankProfessr         FALSE      FALSE
## 1 subsets of each size up to 14
## Selection Algorithm: backward
##           GenderMale StartYr DeptCodeFactor2 DeptCodeFactor3 DeptCodeFactor4
## 1  ( 1 )  " "        "*"     " "             " "             " "
## 2  ( 1 )  " "        "*"     " "             " "             " "
## 3  ( 1 )  " "        "*"     " "             " "             " "
## 4  ( 1 )  " "        "*"     " "             " "             " "
## 5  ( 1 )  " "        "*"     " "             " "             " "
```

```
## 6  ( 1 )  " "         "*"      " "              " "              " "
## 7  ( 1 )  "*"         "*"      " "              " "              " "
## 8  ( 1 )  "*"         "*"      " "              " "              " "
## 9  ( 1 )  "*"         "*"      "*"              " "              " "
## 10  ( 1 ) "*"         "*"      "*"              " "              "*"
## 11  ( 1 ) "*"         "*"      "*"              "*"              "*"
## 12  ( 1 ) "*"         "*"      "*"              "*"              "*"
## 13  ( 1 ) "*"         "*"      "*"              "*"              "*"
## 14  ( 1 ) "*"         "*"      "*"              "*"              "*"
##           DeptCodeFactor5 DeptCodeFactor6 DeptCodeFactor7 DeptCodeFactor8
## 1  ( 1 )  " "             " "             " "             " "
## 2  ( 1 )  " "             " "             " "             " "
## 3  ( 1 )  " "             " "             " "             " "
## 4  ( 1 )  "*"             " "             " "             " "
## 5  ( 1 )  "*"             " "             " "             " "
## 6  ( 1 )  "*"             " "             " "             " "
## 7  ( 1 )  "*"             " "             " "             " "
## 8  ( 1 )  "*"             " "             " "             " "
## 9  ( 1 )  "*"             " "             " "             " "
## 10  ( 1 ) "*"             " "             " "             " "
## 11  ( 1 ) "*"             " "             " "             " "
## 12  ( 1 ) "*"             " "             " "             "*"
## 13  ( 1 ) "*"             "*"             " "             "*"
## 14  ( 1 ) "*"             "*"             "*"             "*"
##           Begin.Salary Expernc RankAsstProf RankInstruct RankProfessr
## 1  ( 1 )  " "          " "     " "          " "          " "
## 2  ( 1 )  "*"          " "     " "          " "          " "
## 3  ( 1 )  "*"          " "     " "          " "          "*"
## 4  ( 1 )  "*"          " "     " "          " "          "*"
## 5  ( 1 )  "*"          "*"     " "          " "          "*"
## 6  ( 1 )  "*"          "*"     "*"          " "          "*"
## 7  ( 1 )  "*"          "*"     "*"          " "          "*"
## 8  ( 1 )  "*"          "*"     "*"          "*"          "*"
## 9  ( 1 )  "*"          "*"     "*"          "*"          "*"
## 10  ( 1 ) "*"          "*"     "*"          "*"          "*"
## 11  ( 1 ) "*"          "*"     "*"          "*"          "*"
## 12  ( 1 ) "*"          "*"     "*"          "*"          "*"
## 13  ( 1 ) "*"          "*"     "*"          "*"          "*"
## 14  ( 1 ) "*"          "*"     "*"          "*"          "*"
```