Statistical Data Analysis – solved problems

Goals: The text provides a pool of solved problems for labs in the course on Statistical Data Analysis. The exercises help to deepen knowledge gained in parallel Rmd files. At the same time, they serve as illustrative examples of future exam questions.

1 Linear and non-linear regression

Problem 1. *(10 p) You built a linear model that predicts the median value of owner-occupied homes in $1000's in a certain town (medv). The model works with the only independent variable (lstat) that captures the percentage of population with lower (economical) status in the given town. The model was built from a training set based on 506 towns and is this:*

```
lm(formula = medv ~ lstat, data = Boston)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,        Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

(a) *(2 p)* Verbally describe the relationship between lstat and medv. Decide whether lstat affects medv and quantify how. Is it a statistically significant relationship? Why?

The model says that with each additional percentage of people with lower status the median value of homes decreases on average by $950. This relationship is statistically significant, based on the F-statistic as well as the lstat's t-value we can reject the null hypothesis that there is no relationship between lstat and medv.

(b) *(1 p)* How do you understand the meaning of Intercept? Is the value of this coefficient a reliable figure to be interpreted literally? Explain.

The value of Intercept says that the average median value of homes in a town with 0 percentage of people with lower status is around $34,553. This value looks reasonable, however, its true reliability depends on how far the model extrapolates (do we have any towns that at least approach no lower status representation in ou training set?) and how far the model meets the linear regression assumptions (is the relationship between these variables truly linear?).

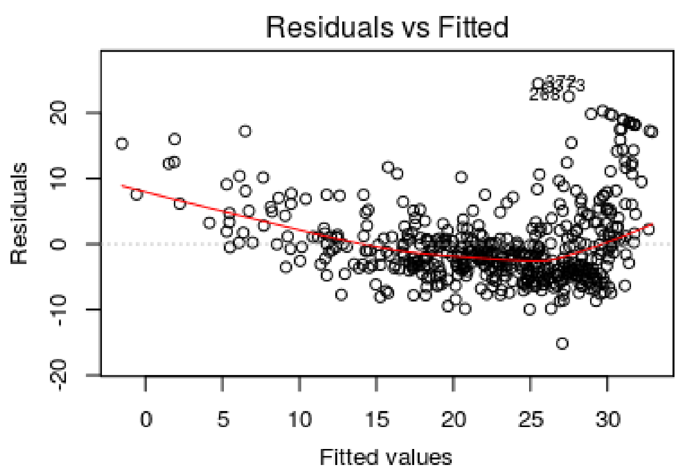
(c) *(1 p)* How much do we improve our median value forecast compared to the simple average forecast that ignores the knowledge of lstat? In other words, how much the knowledge of lstat helps?

The value of R-squared shows that we will reduce the variance of the median home value estimates by about half.

(d) *(2 p)* Calculate/estimate 95% confidence interval for β_{lstat}. What is this interval good for?

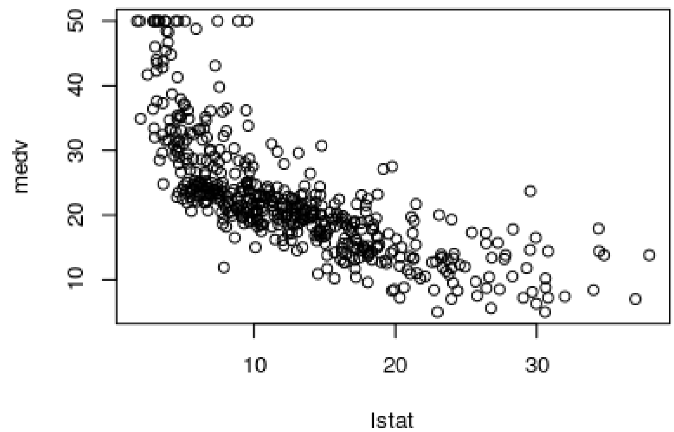
A rough estimate could be [-0.95005-2*0.03873,-0.95005+2*0.03873]=[-1.02751,-0.87259]. A more precise estimate puts $|t_{\alpha/2,m-2}| = |t_{0.025,504}|$ instead of 2 into the formula above, however, the value 1.964682 is close to the rough estimate. This confidence interval has about 95% chance to contain the true value of β_{lstat}. This interval helps us to assume on the strength of relationship between lstat and medv, the interval does not contain 0, the relationship could be considered significant.

(e) *(2 p)* Look at the model residual plot in the figure below (it plots differences between the actual and predicted values of the dependent variable). What conclusions can be drawn from the figure?

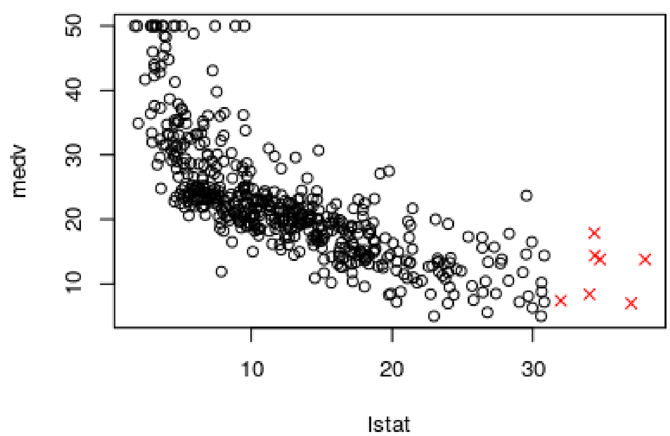


The plot shows that the assumption of linearity has not been met. The residuals should follow the normal distribution for all the values of lstat, they are heavily skewed in the plot. We should conclude that the relationship is non-linear and introduce new terms into the regression formula ($lstat^2$ is a good idea to start with).

(f) *(2 p)* Explain the concept of influential observations. Denote a couple of influential points in the scatter plot below and explain how would you find them.



An influential observation is an observation whose deletion from the dataset would noticeably change the model parameter estimates. It could either be an outlier (a data point that differs significantly from other observations) or a high-leverage point (an observation made at extreme values of independent variables). The most influential observations can be found in the figure below.



Problem 2. *(10 p) You are a mechanical locksmith and you are trying to find out how the shaft machining error is related to the machine tool parameter setting. You have compiled a multivariate linear model. The model expresses the relationship between the production error (the difference between the ideal shaft diameter and the actual shaft diameter, ProdError) and the setting of ten different continuous machine parameters (P1-P10). Below is the output you received:*

```
summary(lm(ProdError ~ P1+P2+P3+P4+P5+P6+P7+P8+P9+P10),data=d)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.05270    0.09576  -0.550   0.5835
X1           0.01298    0.08924   0.145   0.8847
```

```
X2           0.01596    0.10939   0.146   0.8843
X3          -0.02865    0.09079  -0.316   0.7531
X4           0.04611    0.09548   0.483   0.6303
X5           0.14151    0.09343   1.515   0.1334
X6          -0.02375    0.10277  -0.231   0.8178
X7           0.25522    0.10516   2.427   0.0172 *
X8           0.06672    0.08972   0.744   0.4590
X9           0.09949    0.10171   0.978   0.3306
X10         -0.04003    0.09317  -0.430   0.6685
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9039 on 89 degrees of freedom
Multiple R-squared:  0.1145,       Adjusted R-squared:  0.01502
F-statistic: 1.151 on 10 and 89 DF,  p-value: 0.3346
```

(a) *(2 p)* Decide whether at least one of the machine parameters (independent variables) is useful for estimating a manufacturing error (ProdError). In other words, formally decide whether you can decline $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{10} = 0$. Justify correctly.

The reasoning should be based on the F-statistic and its corresponding p-value. The null hypothesis cannot be rejected, the model does not seem to be useful. The reasoning that stems from the statistics reached for the individual variables could be misleading due to multiple comparisons. For 10 variables, truly valid H_0 and $\alpha = 0.05$, there is only $0.95^10 = 0.6$ probability that there will be no type I error in the individual coefficient tests, 40% of trials will find at least one falsely significant coefficient.

(b) *(2 p)* Let us compare the full model constructed above with the intercept model and with the model that employs only the variable $P7$ identified as the most relevant. Let us compare them with F-test through an ANOVA run. Interpret the ANOVA table below.

```
lm.const<-lm(ProdError ~ 1,data=d) # the intercept model
lm.sel<-lm(ProdError ~ P7,data=d) # the P7 model
anova(lm.const,lm.sel,lm.all)
Analysis of Variance Table

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     99 82.114
2     98 76.076  1    6.0384 7.3911 0.007879 **
3     89 72.711  9    3.3647 0.4576 0.899016
```

ANOVA easily compares nested models where the independent variables of a simpler model make a subset of the independent variables of a more complex model. We order the models from the most simple to the most complex and ANOVA compares all the pairs of neighboring models. The ANOVA table suggests that lm.sel outperforms lm.const while lm.all does not further improve lm.sel. This conclusion is in contradiction with the conclusion in the previous answer. The contradiction arises from a methodological fault that we did. We used the same dataset to select $P7$ as the best variable and to test whether it performs well. This approach suffers from bias and could be misleading.

(c) *(2 p)* The dataset under consideration contains 100 samples. How do the type I error and type II error in the individual coefficient tests change with increasing number of samples if we maintain a constant level of significance α?

Type I error is a controlled parameter and its probability remains unchanged with the α value unchanged. However, the power of the test will increase, so the type II error will decrease. At the same time, the robustness of RSS, R^2 and consequently the F-test power will increase as well.

(d) *(4 p)* Describe in detail the way in which you would validate your models over the samples that you currently have. You can create additional auxiliary models. Describe the validation method, define the error function, and specify with which baseline you will compare the calculated error.

Let us assume that we want to compare lm.const, lm.sel and lm.all. Let us assume that our sample set is small and thus the hold-out method that splits the sample set on training and testing set is inappropriate (we need to use as many training samples as possible, the same holds for testing set). Then, a good option seems to be to run 10-fold cross-validation. We will always train our models on 9 folds and test them on the remaining one. We will gradually shift the testing fold. The dependent variable is continuous, we can use the root mean square error (RMSE) or mean absolute percentage error (MAPE). The error will always be calculated over the testing fold and averaged over the folds. If we repeat 10-fold cross-validation multiple times, we can statistically test whether performances of the individual models truly differ.

Watch out. Feature selection is a part of training process. It cannot be done only once before cross-validation, it must be repeated again and again for each split. Consequently, we will have 10 different lm.sel models to test, the set of relevant variables included into the model may change over folds as well as their regression coefficients. These 10 models will serve to estimate the performance of the final lm.sel model. Only the final model (to be reported and deployed) could be based on all the available samples, and will thus certainly employ the variable $P7$.

Problem 3. *(10 p) There is a cubic spline with one knot ξ given by the formula:* $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)^3_+$.

(a) *(1 p)* Define the basis function $(x - \xi)^3_+$.

The definition is: $(x - \xi)^3$ for $x > \xi$, otherwise 0.

(b) *(1 p)* How many degrees of freedom does the given cubic spline have? Why?

A cubic spline with K knots has $K + 4$ parameters or degrees of freedom. Our spline has one knot and thus it has 5 independent parameters/degrees of freedom. The number of parameters can be seen from the formula above too, there are $\beta_0, \ldots, \beta_4$ there.

(c) *(1 p)* What are the properties of the cubic spline at the knot?

The spline is continuous at the knot and it has a continuous first and second derivative there. The properties follow from the general properties for d-degree splines.

(d) *(2 p)* Write down a cubic spline with one knot as a piecewise polynomial. Note: you will only change the form of notation, name the parameters differently from the spline parameters above.

$f_1(x) = a_1 + b_1x + c_1x^2 + d_1x^3$ for $x < \xi$

$f_2(x) = a_2 + b_2x + c_2x^2 + d_2x^3$ for $x \geq \xi$

(e) *(3 p)* Express the piecewise polynomial parameters using the cubic spline parameters $\beta_0, \beta_1, \ldots, \beta_4$.

The procedure is straightforward: the spline must match f_1 before the knot and f_2 after the knot. For the first polynomial it is trivial, because the basis function is zero before the first knot: $a_1 = \beta_0$, $b_1 = \beta_1$, $\ldots$, $d_1 = \beta_3$. For the second polynomial it holds: $a_2 + b_2x + c_2x^2 + d_2x^3 = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \beta_4(x-\xi)^3$. By developing the last term we get: $\beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \beta_4(x^3 - 3x^2\xi + 3x\xi^2 - \xi^3) = (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3$, of which follows: $a_2 = \beta_0 - \beta_4\xi^3$, $b_2 = \beta_1 + 3\beta_4\xi^2$, $c2 = \beta_2 - 3\beta_4\xi$, $d2 = \beta_3 + \beta_4$.

(f) *(2 p)* Proof that the piecewise cubic polynomial found in the previous two steps maintains the knot properties of a cubic spline.

Continuity $f_1(\xi) = f_2(\xi)$ can be proven by substituting for coefficients a, b, c, d: $f_1(\xi) = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3$, $f_2(\xi) = \beta_0 - \beta_4\xi^3 + (\beta_1 + 3\beta_4\xi^2)\xi + (\beta_2 - 3\beta_4\xi)\xi^2 + (\beta_3 + \beta_4)\xi^3 = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3 = f_1(\xi)$.

Continuity of the first derivative $f_1'(\xi) = f_2'(\xi)$ can be confirmed by substituting for the coefficients and deriving: $f_1'(\xi) = \beta_1 + 2\beta_2\xi + 3\beta_3\xi^2 = f_2'(\xi)$.

Continuity of the second derivative $f_1''(\xi) = 2\beta_2 + 6\beta_3\xi = f_2''(\xi)$.

2 Linear regression and ANOVA

Problem 4. *(10 p) You are analyzing salary data from an unknown university stored in a data frame df. You want to find the key factors that actually influence the professors' wages. Besides the target variable salary you deal with the following set of independent variables: rank . . . a factor with levels AssocProf, AsstProf, Prof; discipline . . . a factor with levels A ("theoretical" departments) or B ("applied" departments); yrs.since.phd . . . a numerical variable that gives the number of years since PhD completion; yrs.service . . . a numerical variable that gives the number of years of service and sex . . . a factor with levels Female and Male.*

(a) *(2 p)* Explain in which way you would best decide whether *sex* influences *salary* with the aid of linear regression. Below there are two ultimate sample *lm* calls. Interpret both of them, decide whether any of them could be used to answer the role of *sex*. If they are not applicable, propose your own *lm* call.

```
Call 1:
lm(salary ~ sex, data = df)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   101002       4809  21.001  < 2e-16 ***
sexMale        14088       5065   2.782  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30030 on 395 degrees of freedom
Multiple R-squared:  0.01921,	Adjusted R-squared:  0.01673
F-statistic: 7.738 on 1 and 395 DF,  p-value: 0.005667

Call 2:
lm(salary ~ rank + discipline + yrs.since.phd + yrs.service + sex, data = df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    65955.2     4588.6  14.374  < 2e-16 ***
rankAssocProf  12907.6     4145.3   3.114  0.00198 **
rankProf       45066.0     4237.5  10.635  < 2e-16 ***
disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
yrs.since.phd    535.1      241.0   2.220  0.02698 *
yrs.service     -489.5      211.9  -2.310  0.02143 *
sexMale         4783.5     3858.7   1.240  0.21584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22540 on 390 degrees of freedom
Multiple R-squared:  0.4547,	Adjusted R-squared:  0.4463
F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

The first call suggests that males have higher salaries and the difference is significant. However, we have to assume that the individual independent variables are related and deal with the multivariate model that contains all the predictors.

The actual role of sex can be assumed from the coefficient that is related to sex variable in the Call 2 and its p-value. There, the absolute value of the regression coefficient adjoined to sex is much smaller than in Call 1 and the difference between males and females seems to be insignificant. We would need more data to decide with more power, however, the significance of sex in Call 1 was obviously caused by the fact that males have higher ranks and longer careers. This could easily be checked by e.g. *aov(yrs.since.phd sex,df)*:

```
             Df Sum Sq Mean Sq F value  Pr(>F)
sex           1   1456  1455.9   8.942 0.00296 **
Residuals   395  64310   162.8
```

(b) *(2 p)* Now you will make the following call: *summary(aov(salary∼sex,df))*. Explain what you will learn from the call.

One-way ANOVA (one independent variable and one dependent variable) could be considered equivalent to single variate *lm* calls. Consequently, we will get nearly the same message as we got in previous *lm* Call 1 (the same F-statistic and the same p-value). The main difference is in the form of the output (the coefficients are not reported here). In particular:

```
              Df    Sum Sq   Mean Sq F value  Pr(>F)
sex            1 6.980e+09 6.980e+09   7.738 0.00567 **
Residuals    395 3.563e+11 9.021e+08
```

(c) *(2 p)* Now you will make the following call: *summary(aov(salary ∼rank + discipline + yrs.since.phd + yrs.service + sex,df))*. Explain what you will learn from the call.

Unlike the previous case, this multivariate ANOVA call will not match the previous multivariate *lm* call. The reason is that both the methods treat sum of squares in regression differently. In *lm*, we search for main effects of the individual variables, their influence is considered in parallel (Type III sum of squares). In *aov*, we evaluate the individual predictors sequentially, in the order of their appearance in the formula (Type I sum of squares). For example, if we enter *rank* first, its sums of squares are computed ignoring *discipline* and other variables. Therefore, any variance in *salary* that is shared by *rank* and *discipline* will be attributed solely to *rank*. The sums of squares for *discipline* will then be computed excluding any variance that has already been attributed to *rank*. Consequently, in this type of call we may learn what is the contribution of a new variable to the existing model based on all the variables that precede it in the formula.

The above described *aov* call in fact performs an analysis of covariance (ANCOVA) which blends ANOVA and regression. It evaluates whether the means of a dependent variable (DV) are equal across levels of a categorical independent variable (IV) often called a treatment, while statistically controlling for the effects of other (continuous) variables that are not of primary interest, known as covariates (CV). In particular:

```
               Df    Sum Sq   Mean Sq F value    Pr(>F)
rank            2 1.432e+11 7.162e+10 140.979  < 2e-16 ***
discipline      1 1.843e+10 1.843e+10  36.280 3.95e-09 ***
yrs.since.phd   1 1.656e+08 1.656e+08   0.326   0.5683
yrs.service     1 2.576e+09 2.576e+09   5.072   0.0249 *
sex             1 7.807e+08 7.807e+08   1.537   0.2158
Residuals     390 1.981e+11 5.080e+08
```

(d) *(2 p)* Explain in which way you would best decide whether *sex* influences *salary* with the aid of ANOVA. Show a particular R call or calls (*aov* and *anova* commands).

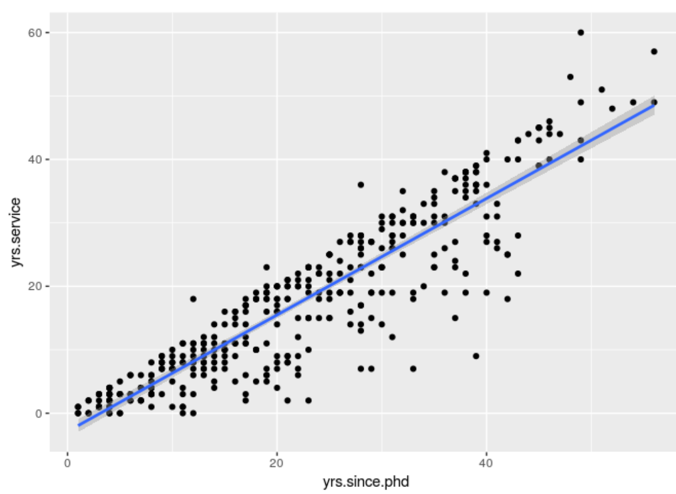
We have already shown that one-way ANOVA could be considered equivalent to single variate *lm* call and it may oversimplify in multivariate tasks because of ignoring confounders. The best calls could be: *summary(aov(salary ∼rank + discipline + yrs.since.phd + yrs.service + sex,df))*, in fact any *aov* call that puts *sex* last in the full multivariate model. Another option is: *anova(lm(salary ∼rank + discipline + yrs.since.phd + yrs.service,df),lm(salary ∼rank + discipline + yrs.since.phd + yrs.service + sex,df))*. The outcome of both the calls is exactly the same wrt the role of *sex* under discussion.

However, this approach studies sequential effect of *sex* wrt the existing model. The question that we answer with these calls is different from the original question whether *sex* influences *salary*. It is much better answered by the previous call *lm(salary ∼rank + discipline + yrs.since.phd + yrs.service + sex, data = df)*.

(e) *(2 p)* Discuss the relationship between *yrs.service* and *yrs.since.phd*. Is there any issue to be checked? Consider both the real meaning of these two variables and *lm* calls above. If so, propose a solution.

Obviously, these two variables are closely related and necessarily correlated. In general, the more years from PhD, the longer service. This correlation could be strong and may cause collinearity problems in the multivariate model. The issue may affect estimates regarding the individual predictors. In the full multivariate model (*lm* Call 1) it is unexpected and suspicious that salary decreases with the length of service. We would expect exactly the opposite and collinearity could be the reason.

A simple solution is to calculate correlation (it is 0.91 in our data) and/or draw a scatter plot for the two variables:



Knowing the collinearity, we would possibly remove one of the two variables from the model (if we do so and remove *yrs.since.phd* in the first call and *yrs.service* in the second call, any of the two variables comes out insignificant then):

```
lm(salary ~ rank + discipline + yrs.since.phd + sex, data = df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   67884.32    4536.89  14.963  < 2e-16 ***
rankAssocProf 13104.15    4167.31   3.145  0.00179 **
rankProf      46032.55    4240.12  10.856  < 2e-16 ***
disciplineB   13937.47    2346.53   5.940 6.32e-09 ***
yrs.since.phd    61.01     127.01   0.480  0.63124
sexMale        4349.37    3875.39   1.122  0.26242
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22660 on 391 degrees of freedom
Multiple R-squared:  0.4472,	Adjusted R-squared:  0.4401
F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

```
lm(salary ~ rank + discipline + yrs.service + sex, data = df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   68351.67    4482.20  15.250  < 2e-16 ***
rankAssocProf 14560.40    4098.32   3.553 0.000428 ***
rankProf      49159.64    3834.49  12.820  < 2e-16 ***
disciplineB   13473.38    2315.50   5.819 1.24e-08 ***
yrs.service     -88.78     111.64  -0.795 0.426958
sexMale        4771.25    3878.00   1.230 0.219311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22650 on 391 degrees of freedom
Multiple R-squared:  0.4478,	Adjusted R-squared:  0.4407
F-statistic: 63.41 on 5 and 391 DF,  p-value: < 2.2e-16
```

If we look at the model scores and compare the simplified models with lm Call 2, it is obvious that shrinkage is beneficial (no decrease in Adjusted R-squared, slightly tighter coefficient estimates).