Regression Analysis on the Salaries of a University's Electrical Engineering Graduates

Chris Meade and Cordelia Roberts

Santa Barbara, California, United States

1. Introduction

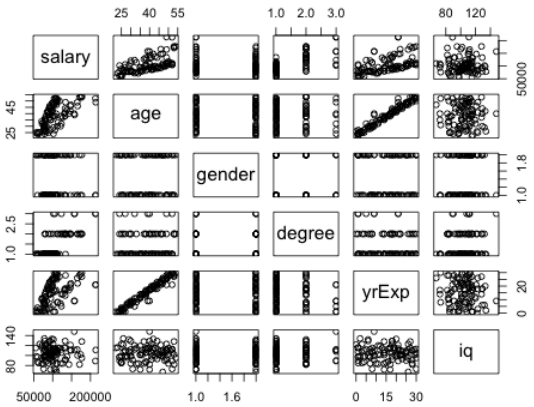
Given data on a single university's electrical engineering graduates, our research question was to determine whether a graduate's age, gender, years of experience, highest degree attained, and IQ can successfully predict their salary. If salaries can be predicted from the given variables, we also wish to attain the best regression model that best fits the data with the highest error reduction.

2. Method

We begin our analysis of the data by conducting some preliminary exploratory examination, looking for any issues in the raw data that may pose a problem. This is performed by examining the head, tail, and summary of our data using built in R functions. From this initial exploration, nothing appears to be problematic. For the sake of diligence, we decide to visually explore our data with some plots.

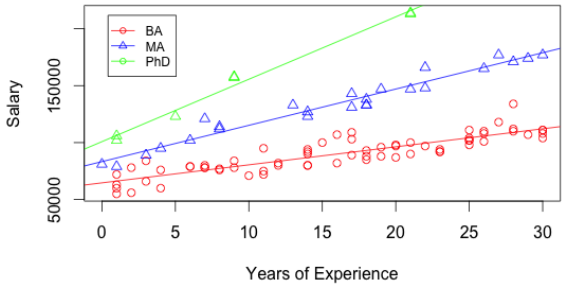
So, we create a scatterplot matrix to examine any possible relationships between the variables. From this, we notice that age and years of experience appear to be highly correlated. Indeed, the correlation coefficient between *age* and $yrExp$ is 0.98, leading us to draw the conclusion that one of these variables can be dropped from our model. Based upon our knowledge of the electrical engineering field, we decided that years of experience is a more appropriate choice for our model. Consequently, we do not include the *age* variable in our model.

Figure 1: Our scatterplot matrix



In several of the scatterplots, we also notice an outlier: the lowest salary is $\$17,500$, far below the second-lowest of $\$55,000$. We summarily remove this outlier. Our scatterplots also suggest that we should test the significance of IQ, as there does not appear to be a linear relationship between it and salary.

We also construct a plot to examine the interaction between highest degree attained and years of experience in predicting salary. From this plot, it appears that there may be some interaction.



With our exploratory analysis complete, we begin to perform regression analysis. We start by fitting a model, $fit1$, with all the predictors except age, which we chose to eliminate due to its correlation with years of experience.

```
fit1 <- lm(salary ~ yrExp + gender + degree + iq)
```

Using the *drop*1() function, we perform a backwards elimination of variables by partial F-test. We conclude that we should drop IQ from the model, since the p-value associated with its F-statistic is not significant.

This leaves us our improved model, $fit2$, which predicts salary only from *yrExp*, *gender*, and *degree*. We believe that this model can be improved further by including the previously discussed interaction between *degree* and *yrExp*. To test our hypothesis, we add the predictor yrExp*degree to a new model, $fit3$, which is otherwise the same as $fit2$.

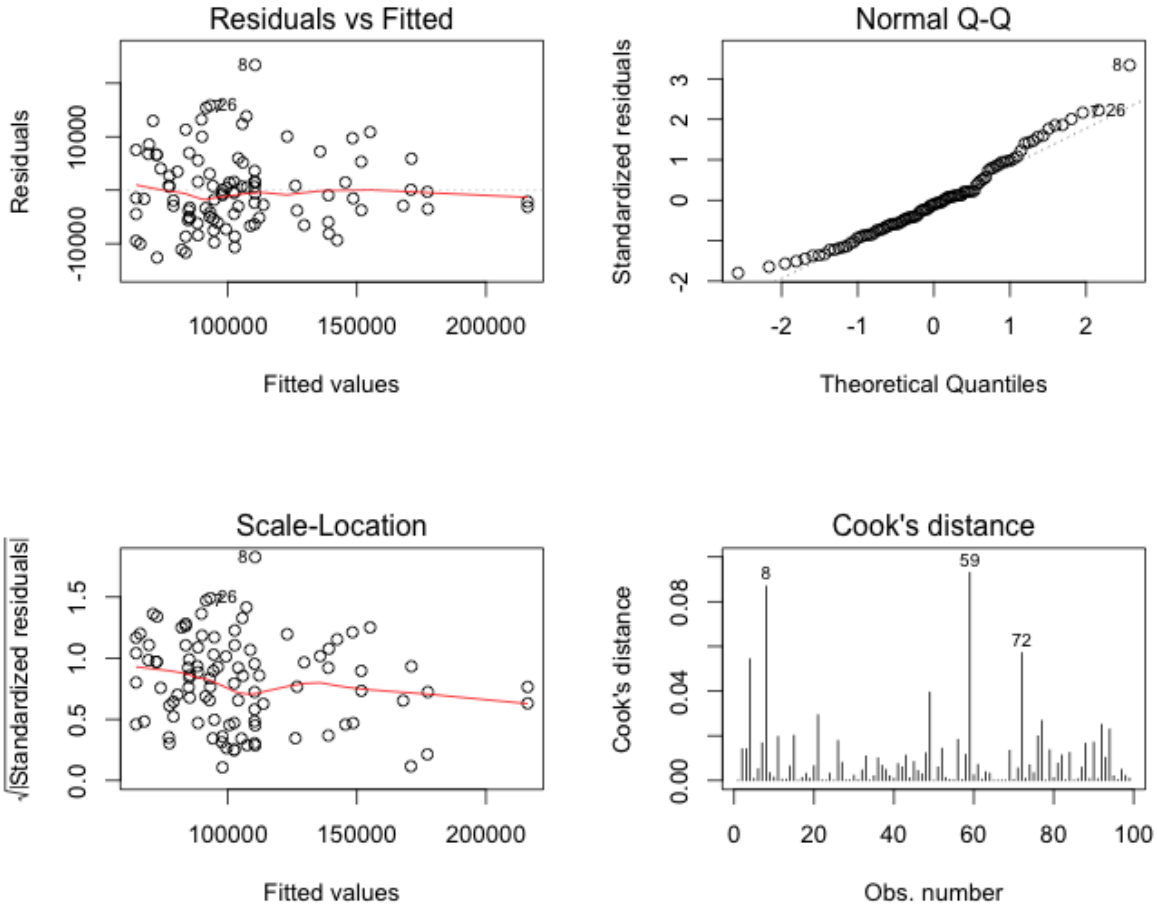
```
fit2 <- lm(salary ~ yrExp + gender + degree)
fit3 <- lm(salary ~  yrExp + gender + degree + (yrExp*degree))
```

Then we run $anova(fit2, fit3)$ to compare these two models. The subsequent output tells us the our new predictor $yrExp * degree$ is significant, since the p-value associated with the F statistic is well below 0.05.

3. Results

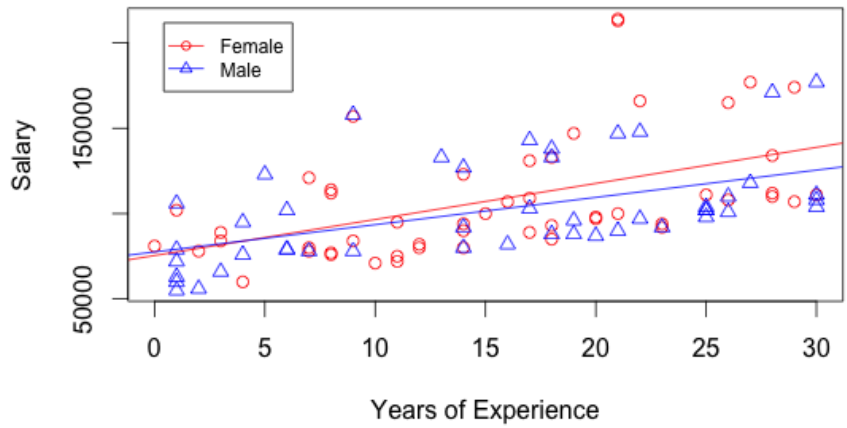
We select $fit3$ to be our final model on the basis that it finds a good balance between goodness of fit and simplicity. It excludes superfluous predictors while very accurately explaining the data. Indeed, his model has an adjusted R^2 value of 0.95, meaning it explains 95% variability of the response data.

In evaluating the goodness of this model, we looked for possible violation of assumptions: linearity, constant variance, independence, and normality of error distribution. The Residuals vs Fitted plot of $fit3$ shows a mostly random scattering, so we can say that variance is constant.The QQ-Plot shows a slight deviation from the line $y = x$ at both ends of the graph, which implies that we me may have an issue of non-normality. This can be rectified through a resampling of the data. The Cook's Distance plot shows all values well below 0.1, so we can conclude from this that no single point is too highly influential. The Scale location plot also has no discernible pattern. Indeed, the only possible violation of the assumptions, if any, is the non-normality exposed by the QQ-Plot.



4. Discussion

From our model, we can draw a few conclusions. Unsuprisingly, more years of experience is correlated with earning a higher salary. Similarly, those with a PhD earn more than those with a MA, who in turn earn more than those with a BA. In addition, there is a reinforcement interaction between *yrExp* and *degree*. Those with a higher degree not only start out at a higher salary, but also see a greater increase in salary for each additional year of experience. Something suprising we found in our model is that women earn more than men on average, which contradicts current socio-economic theory.



We also find that our model has a few limitations. First and foremost, there are only seven PhDs in the data set. The sample size in general is smaller than we would have liked. If it were larger, we believe that we could construct a more accurate model. Also, the datset only contains information about graduates from a single universtiy, so our model may not be able to accurately predict the salaries of electrical engineers who graduated from a differnt school.

5. Future Direction

This model could be improved by studying other predictors that may influence annual salary. We think that the location of the job would be very influential in this regard. For example, an engineer in San Francisco would probably earn more than an engineer in Omaha, simply because the cost of living is so much higher in San Francisco. We also feel that undergraduate GPA may be influential in predicting salary. Those with a higher GPA may be able to land higher paying jobs directly out of school. There also may be a correlation between GPA and highest degree, so in this regard it could very well improve our model.

6. R Code

```
[language=R]
# Chris Meade and Cordelia Roberts
# Class Project Appendix (R Code)
# PSTAT 126
# March 8, 2016

# Import libraries for plots
install.packages("ggplot2")
library(ggplot2, GGally)

# Import our data into R
income = read.table('/Users/cordeliaroberts/Documents/PSTAT126/Class Project/income.txt', header=TRU
attach(income)

# Exploratory analysis
summary(income)
head(income)
tail(income)

# Scatterplot matrix
pairs(income)

# We notice from the scatterplot matrix that
# age and yrExp appear highly correlated.
cor(age,yrExp)
# Indeed, the correlation coefficient is approximately
# 0.98. We can drop one of these variables

##plotting numeric data with regression
ggplot(income) +
geom_jitter(aes(age,salary), colour="blue") + geom_smooth(aes(age,salary), method=lm, se=FALSE) +
geom_jitter(aes(yrExp,salary), colour="green") + geom_smooth(aes(yrExp,salary), method=lm, se=FALSE)
geom_jitter(aes(iq,salary), colour='red') + geom_smooth(aes(iq,salary), method=lm, se=FALSE) +
```

```
labs(x = "Percentage cover (%)", y = "Salary ($)")

##Examining possible relationships between variables with some scatterplot matrices
pairs(income)
ggpairs(income, colour='Red', alpha=0.4)



# In several of the scatterplots, we notice an outlier:
# the lowest salary is $17,500, far below the second-lowest
# of $55,000. We remove this outlier

newincome<-subset(income, salary != 17500)

detach()
attach(newincome)

# An exploratory plot to check for possible interaction between
# gender and yrExp
plot(yrExp[gender=='F'],salary[gender=='F'],pch=1,xlim=range(yrExp),ylim=range(salary),col='red',yla
points(yrExp[gender=='M'],salary[gender=='M'],pch=2,xlim=range(yrExp),ylim=range(salary), col='blue'
abline(lm(salary[gender=='F']~yrExp[gender=='F']), col='red')
abline(lm(salary[gender=='M']~yrExp[gender=='M']), col = 'blue')
legend('topleft', inset=.05,cex=.75,pch=1:2,lty=c(1,1),col=c('red','blue'),legend=c('Female','Male')

# A similar plot to look at interactions between degree and yrExp
plot(yrExp[degree=='BA'],salary[degree=='BA'],pch=1,xlim=range(yrExp),ylim=range(salary),col='red',y
points(yrExp[degree=='MA'],salary[degree=='MA'],pch=2,xlim=range(yrExp),ylim=range(salary), col='blu
points(yrExp[degree=='PhD'],salary[degree=='PhD'],pch=2,xlim=range(yrExp),ylim=range(salary), col='g
abline(lm(salary[degree=='BA']~yrExp[degree=='BA']), col='red')
abline(lm(salary[degree=='MA']~yrExp[degree=='MA']), col = 'blue')
abline(lm(salary[degree=='PhD']~yrExp[degree=='PhD']), col = 'green')
legend('topleft', inset=.05,cex=.75,pch=1:2,lty=c(1,1),col=c('red','blue','green'),legend=c('BA','MA

# Testing multiple regression models, backwards elimination by partial f test
```

```
# Start by fitting model with all predictors except age
fit1 <- lm(salary~yrExp+gender+iq+degree)
summary(fit1)
drop1(fit1, test='F')

# Conclude drop IQ from AIC, since p-value associated
# with F-test is not significant

fit2 <- lm(salary~yrExp+gender+degree)
summary(fit2)

# Can we include this model by adding the interaction
# between degree and yrExp?
fit3 <- lm(salary~yrExp+gender+degree+yrExp*degree)
summary(fit3)

# Hypothesis test
anova(fit2,fit3)
# Conclude fit3 is a better model

# Run diagnostics to see if this model fits assumptions
plot(fit3)

# fit3 appears to fit the assumptions of linearity,
# normality, and constant variance. Since we do not
# know the order in which the data was collected, we
# do not include a graph of residuals vs. index to
# assess independence
```

7. R Output

```
[language = R]
> # Chris Meade and Cordelia Roberts
> # Class Project Appendix (R Code)
> # PSTAT 126
> # March 8, 2016
>
> # Import libraries for plots
> library(ggplot2, GGally)
Warning message:
package ggplot2 was built under R version 3.2.4
>
> # Import our data into R
> income = read.table('/Users/chrismeade/Desktop/income.txt', header=TRUE)
> attach(income)
>
> # Exploratory analysis
> summary(income)
salary            age        gender degree       yrExp             iq
Min.   : 55000   Min.   :22.00   F:52   BA :67   Min.   : 0.00   Min.   : 67.0
1st Qu.: 81500   1st Qu.:30.00   M:47   MA :25   1st Qu.: 8.00   1st Qu.: 95.0
Median : 98000   Median :39.00          PhD: 7   Median :17.00   Median :106.0
Mean   :105323   Mean   :38.61                   Mean   :15.44   Mean   :104.2
3rd Qu.:116000   3rd Qu.:46.50                   3rd Qu.:22.50   3rd Qu.:114.5
Max.   :214000   Max.   :54.00                   Max.   :30.00   Max.   :148.0
> head(income)
salary age gender degree yrExp  iq
1 110000  50      F     BA    28  84
2  78000  26      F     BA     2 134
3 133000  41      F     MA    18  92
4 121000  30      F     MA     7  84
5  80000  30      F     BA     7  98
6  80000  36      F     BA    14 101
> tail(income)
salary age gender degree yrExp  iq
94 118000  50      M     BA    27 115
95  88000  42      M     BA    19 116
96 112000  51      F     BA    28 110
97 110000  47      M     BA    26  67
98  80000  35      F     BA    12 119
99  76000  31      F     BA     8  73
>
> # Scatterplot matrix
> pairs(income)
>
> # We notice from the scatterplot matrix that
> # age and yrExp appear highly correlated.
> cor(age,yrExp)
[1] 0.9770619
> # Indeed, the correlation coefficient is approximately
> # 0.98. We can drop one of these variables
>
> ##plotting numeric data with regression
> ggplot(income) +
+   geom_jitter(aes(age,salary), colour="blue") + geom_smooth(aes(age,salary), method=lm, se=FALSE)
+   geom_jitter(aes(yrExp,salary), colour="green") + geom_smooth(aes(yrExp,salary), method=lm, se=FA
```

```
+   geom_jitter(aes(iq,salary), colour='red') + geom_smooth(aes(iq,salary), method=lm, se=FALSE) +
+   labs(x = "Percentage cover (%)", y = "Salary ($)")
>
> ##Examining possible relationships between variables with some scatterplot matrices
> pairs(income)
> ggpairs(income, colour='Red', alpha=0.4)
Error: could not find function "ggpairs"
>
>
> # In several of the scatterplots, we notice an outlier:
> # the lowest salary is $17,500, far below the second-lowest
> # of $55,000. We remove this outlier
>
> newincome<-subset(income, salary != 17500)
>
> detach()
> attach(newincome)
>
> # An exploratory plot to check for possible interaction between
> # gender and yrExp
> plot(yrExp[gender=='F'],salary[gender=='F'],pch=1,xlim=range(yrExp),ylim=range(salary),col='red',y
> points(yrExp[gender=='M'],salary[gender=='M'],pch=2,xlim=range(yrExp),ylim=range(salary), col='blu
> abline(lm(salary[gender=='F']~yrExp[gender=='F']), col='red')
> abline(lm(salary[gender=='M']~yrExp[gender=='M']), col = 'blue')
> legend('topleft', inset=.05,cex=.75,pch=1:2,lty=c(1,1),col=c('red','blue'),legend=c('Female','Male
>
> # A similar plot to look at interactions between degree and yrExp
> plot(yrExp[degree=='BA'],salary[degree=='BA'],pch=1,xlim=range(yrExp),ylim=range(salary),col='red'
> points(yrExp[degree=='MA'],salary[degree=='MA'],pch=2,xlim=range(yrExp),ylim=range(salary), col='b
> points(yrExp[degree=='PhD'],salary[degree=='PhD'],pch=2,xlim=range(yrExp),ylim=range(salary), col=
> abline(lm(salary[degree=='BA']~yrExp[degree=='BA']), col='red')
> abline(lm(salary[degree=='MA']~yrExp[degree=='MA']), col = 'blue')
> abline(lm(salary[degree=='PhD']~yrExp[degree=='PhD']), col = 'green')
> legend('topleft', inset=.05,cex=.75,pch=1:2,lty=c(1,1),col=c('red','blue','green'),legend=c('BA','
>
> # Testing multiple regression models, backwards elimination by partial f test
> # Start by fitting model with all predictors except age
> fit1 <- lm(salary~yrExp+gender+iq+degree)
> summary(fit1)

Call:
lm(formula = salary ~ yrExp + gender + iq + degree)

Residuals:
Min     1Q Median     3Q    Max
-34346  -7619  -1508   7885  33339

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 56210.679   8586.108   6.547 3.19e-09 ***
yrExp        2207.320    135.211  16.325  < 2e-16 ***
genderM     -4992.159   2382.328  -2.095   0.0388 *
iq              8.011     77.722   0.103   0.9181
degreeMA    43947.349   2775.966  15.831  < 2e-16 ***
degreePhD   77207.287   4800.426  16.083  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 11820 on 93 degrees of freedom
Multiple R-squared:  0.8729,Adjusted R-squared:  0.866
F-statistic: 127.7 on 5 and 93 DF,  p-value: < 2.2e-16

> drop1(fit1, test='F')
Single term deletions

Model:
salary ~ yrExp + gender + iq + degree
Df  Sum of Sq        RSS    AIC  F value  Pr(>F)
<none>               1.2997e+10 1862.6
yrExp   1 3.7244e+10 5.0241e+10 1994.5 266.5074 < 2e-16 ***
gender  1 6.1366e+08 1.3610e+10 1865.2   4.3911 0.03884 *
iq      1 1.4846e+06 1.2998e+10 1860.6   0.0106 0.91813
degree  2 6.1052e+10 7.4049e+10 2030.8 218.4337 < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> # Conclude drop IQ from AIC, since p-value associated
> # with F-test is not significant
>
> fit2 <- lm(salary~yrExp+gender+degree)
> summary(fit2)

Call:
lm(formula = salary ~ yrExp + gender + degree)

Residuals:
Min     1Q Median     3Q    Max
-34508  -7624  -1524   7898  33362

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  57045.4     2836.7  20.110   <2e-16 ***
yrExp         2206.5      134.3  16.433   <2e-16 ***
genderM      -4982.4     2367.9  -2.104    0.038 *
degreeMA     43964.4     2756.4  15.950   <2e-16 ***
degreePhD    77256.0     4751.9  16.258   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11760 on 94 degrees of freedom
Multiple R-squared:  0.8729,Adjusted R-squared:  0.8675
F-statistic: 161.3 on 4 and 94 DF,  p-value: < 2.2e-16

>
> # Can we include this model by adding the interaction
> # between degree and yrExp?
> fit3 <- lm(salary~yrExp+gender+degree+yrExp*degree)
> summary(fit3)

Call:
lm(formula = salary ~ yrExp + gender + degree + yrExp * degree)

Residuals:
Min       1Q   Median       3Q      Max
```

```
-12601.3  -4990.4   -693.2   3832.7  23422.7

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     66271.94    1942.95  34.109  < 2e-16 ***
yrExp            1582.33      99.18  15.955  < 2e-16 ***
genderM         -3390.04    1463.73  -2.316   0.0228 *
degreeMA        18535.01    3434.26   5.397 5.24e-07 ***
degreePhD       37050.27    4681.18   7.915 5.42e-12 ***
yrExp:degreeMA   1613.84     190.37   8.477 3.63e-13 ***
yrExp:degreePhD  3789.53     364.43  10.399  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7204 on 92 degrees of freedom
Multiple R-squared:  0.9533,Adjusted R-squared:  0.9503
F-statistic:   313 on 6 and 92 DF,  p-value: < 2.2e-16

>
> # Hypothesis test
> anova(fit2,fit3)
Analysis of Variance Table

Model 1: salary ~ yrExp + gender + degree
Model 2: salary ~ yrExp + gender + degree + yrExp * degree
Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1     94 1.2998e+10
2     92 4.7742e+09  2 8.224e+09 79.239 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> # Conclude fit3 is a better model
>
> # Run diagnostics to see if this model fits assumptions
> plot(fit3)
>
> # fit3 appears to fit the assumptions of linearity,
> # normality, and constant variance. Since we do not
> # know the order in which the data was collected, we
> # do not include a graph of residuals vs. index to
> # assess independence
```