Exam 2

Roberto Campos

4/8/2020

1. The file HARRIS shows the values of the following variables for 93 employees of Harris Bank Chicago in 1977:  
   •Salary= Annual beginning (or starting) salary,in dollars  •Educate= Years of schooling at the time of hire  
   •Exper= Number of months of previous work experienceat time of hire  
   •Months= Number of months after January 1, 1969, that the individual was hired  
   Construct an initial regression model for the starting salaries of employees of Harris Bank Chicago  
   using only the variables above. Base all of your responses below on the initial model only!

a. Write down the regression equation for the initial model.

**Salary = 3179.74 + 139.618 \* Educate + 1.481 \* Exper + 20.633 \* Months**

b. Test the overall utility of the model. Explain.

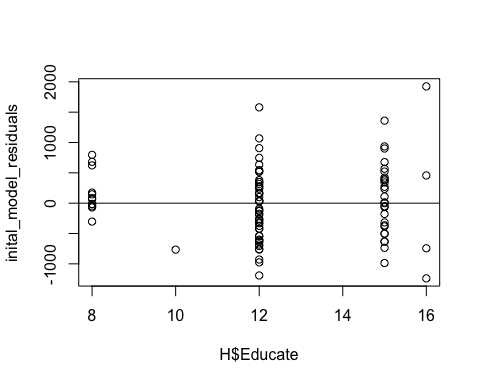
The p-value is actually very low meaning that we can assume there is a linear relationship between the dependent variable and independent variables.

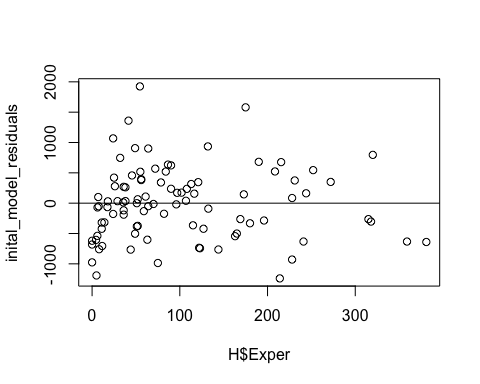
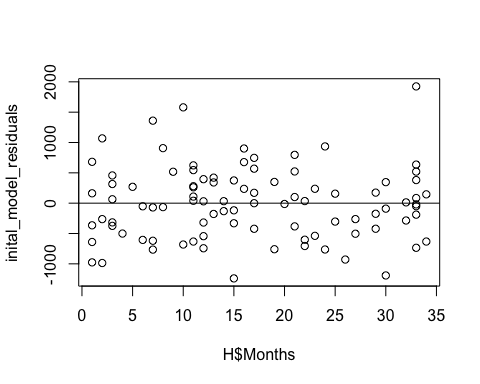
The F-statistic is greater than 1, this can help us assume that we can reject the null hypothesis.

##   
## Call:  
## lm(formula = H$Salary ~ H$Educate + H$Exper + H$Months, data = H)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1240.0 -421.6 -0.3 350.1 1924.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3179.744 383.485 8.292 1.09e-12 \*\*\*  
## H$Educate 139.618 27.716 5.037 2.45e-06 \*\*\*  
## H$Exper 1.481 0.697 2.124 0.03641 \*   
## H$Months 20.633 6.155 3.352 0.00118 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 602.8 on 89 degrees of freedom  
## Multiple R-squared: 0.3019, Adjusted R-squared: 0.2784   
## F-statistic: 12.83 on 3 and 89 DF, p-value: 4.847e-07

c. Are the required conditions satisfied? You do not need to discuss outliers or influential points. (Briefly summarize your conclusions and provide relevant output for support.)

We can see that there is constant spread in variables Months and Exper, they are centered on 0 throught the range of fitted values. Educate seems to have constant sprean on education levels higher than 12 but less spread on education levels that are lower. Althought there is some outliers, from looking at the residual plots we can assume that the resiuals are normally distributed satisfying the assumption.



d. Interpret the slopes of each of the independent variables carefully and using the appropriate units.

**Salary = 3179.74 + 139.618 \* Educate + 1.481 \* Exper + 20.633 \* Months**

The mean salary increases about $139.6 for each year of education. (fixed exper and # on months)

For each additional month of experience the mean salary will increase by about $1.48. (fixed educ and # of months)

For each additional month the mean salary will increase $20.6 per month (fixed educ and exper)

e. Is there evidence, at the 5% level of significance, for a difference in salaries, on average, for male and female employees at the bank? Explain.

Yes, The p - value for Male shows that there is significance.

##   
## Call:  
## lm(formula = H$Salary ~ H$Educate + H$Exper + H$Months + H$Male)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1238.66 -352.62 -24.76 280.08 1569.23   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3526.4221 327.7254 10.760 < 2e-16 \*\*\*  
## H$Educate 90.0203 24.6936 3.645 0.000451 \*\*\*  
## H$Exper 1.2690 0.5877 2.159 0.033562 \*   
## H$Months 23.4062 5.2009 4.500 2.07e-05 \*\*\*  
## H$Male 722.4607 117.8216 6.132 2.41e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 507.4 on 88 degrees of freedom  
## Multiple R-squared: 0.5109, Adjusted R-squared: 0.4886   
## F-statistic: 22.98 on 4 and 88 DF, p-value: 5.072e-13

f. Interpret the coefficient for the dummy variable Male.

The mean salary with the same education, experience, and months is $722.46 more for males.

g. Predict, with 95% confidence, the mean starting salary for male employees with 12 years of education, 10 years (i.e.,120 months) of experience, and Months = 36 (i.e., hired in January of 1972). Do the same for females with 12 years of education, 10 years of experience, and Months = 36. Does this support your conclusion in part (d)? Explain.

Yes, the predicted salaries given the constraints give different salaries. The salary for male was $6324 The salary for female was $5601

## [1] "Male"

## fit lwr upr  
## 1 6324.03 5275.884 7372.176

## [1] "Female"

## fit lwr upr  
## 1 5601.569 4567.061 6636.077

h. Come up with the best model of starting salary, based on the variables in the dataset, that you think you can justify, and justify your choice

Looking at the summary, model with degree 5 has a better Adjusted R squared then our model in part **e.** The Residual Standard Error is also lower. Running an AIC test shows that our model with degree 5 is better, than our model in part **e.** AIC score of model with degree 5 was 1416.911, for model1 in part e. the AIC score was 1429.441.

(Both use the independent variables Educate, Exper, Months, and Male.)

##   
## Call:  
## lm(formula = x ~ poly(a, 5, raw = TRUE) + poly(b, 5, raw = TRUE) +   
## poly(c, 5, raw = TRUE) + poly(d, 5, raw = TRUE))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -954.39 -282.90 -2.59 205.22 1467.20   
##   
## Coefficients: (5 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.040e+05 9.468e+04 2.155 0.034309 \*   
## poly(a, 5, raw = TRUE)1 -6.774e+04 3.218e+04 -2.105 0.038583 \*   
## poly(a, 5, raw = TRUE)2 8.351e+03 3.991e+03 2.092 0.039695 \*   
## poly(a, 5, raw = TRUE)3 -4.468e+02 2.152e+02 -2.076 0.041216 \*   
## poly(a, 5, raw = TRUE)4 8.798e+00 4.276e+00 2.057 0.043031 \*   
## poly(a, 5, raw = TRUE)5 NA NA NA NA   
## poly(b, 5, raw = TRUE)1 4.156e+01 1.084e+01 3.834 0.000256 \*\*\*  
## poly(b, 5, raw = TRUE)2 -5.981e-01 2.069e-01 -2.891 0.004992 \*\*   
## poly(b, 5, raw = TRUE)3 3.750e-03 1.535e-03 2.442 0.016897 \*   
## poly(b, 5, raw = TRUE)4 -1.043e-05 4.806e-06 -2.170 0.033100 \*   
## poly(b, 5, raw = TRUE)5 1.047e-08 5.327e-09 1.966 0.052934 .   
## poly(c, 5, raw = TRUE)1 -2.319e+01 2.127e+02 -0.109 0.913498   
## poly(c, 5, raw = TRUE)2 1.555e+01 3.713e+01 0.419 0.676497   
## poly(c, 5, raw = TRUE)3 -1.418e+00 2.672e+00 -0.531 0.597230   
## poly(c, 5, raw = TRUE)4 4.810e-02 8.460e-02 0.569 0.571315   
## poly(c, 5, raw = TRUE)5 -5.529e-04 9.738e-04 -0.568 0.571843   
## poly(d, 5, raw = TRUE)1 7.042e+02 1.132e+02 6.223 2.36e-08 \*\*\*  
## poly(d, 5, raw = TRUE)2 NA NA NA NA   
## poly(d, 5, raw = TRUE)3 NA NA NA NA   
## poly(d, 5, raw = TRUE)4 NA NA NA NA   
## poly(d, 5, raw = TRUE)5 NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 450.5 on 77 degrees of freedom  
## Multiple R-squared: 0.6626, Adjusted R-squared: 0.5968   
## F-statistic: 10.08 on 15 and 77 DF, p-value: 1.007e-12

## df AIC  
## model1 6 1429.441  
## model\_degree5 17 1416.911

1. As part of a multiple regression model for personal income, a survey is administered to randomly selected individuals. As well as several quantitative variables, a qualitative variable for education (the highest degree earned) is included in the survey and coded as follows: 1 = less than a high school degree,2 = high school degree,3 = bachelors degree,4 = masters degree and 5 = doctorateor above.

a. How many dummy variables must be coded for the education variable?

4 dummy variables must be encoded.

b. Describe (define) dummy variables that you would create to represent the education variable in a regression model. Give them names and describe the way you would code them for use in the model.

Dummy variables are used to turn categorical data into numeric values. There would be 4 columns, if the highest degree earned was for example a bachelors degree, there would be a 1 in that column and 0 in the other 3 columns (less-hs,hs, masters).

c. Can all of the dummy variables equal 1 simultaneously for a person included in the survey? If so, what is the interpretation?

No, only for the highest degree earned. All others will have 0.

d. Can all of the dummy variables equal 0 simultaneously for a person included in the survey? If so, what is the interpretation?

Yes, the column that is left out would be the column test against.

1. Suppose you have built a multiple regression model to predict house prices in a community that uses three variables: house size, distance from downtown, and typical commute time as independent variables. You then randomly selected 50 homes in the community and determined the values for each variable for every house in your sample. For this model, provide the correct degrees of freedom for the following sums of squares as they would appear in the Statgraphics output. (Bespecific, I want numbers, not formulas!)

a. sst: df= **n - 1 = 49**

b. sse: df= **n - k -1 = 46**

c. ssr: df= **k = 3**

d. State in symbolic form the hypotheses for the test of the model.

salary = b1 \* housesize + b2 \* distanceFromTOwn + b3 \* typicalCommute  
**H0:** b1=b2=b3=0  
**H(alternative):** b1 != 0 or b2 !=0 or b3 != 0 ‘atleast one of the coefficients is not equal to 0’

e. What is the test statistic associated with the hypothesis test described in part (b)? Write out the equation for this test statistic.

F = MSR/MSE = (SSR/k)/(SSE/(n-k-1))

**4**.

A close up of a whiteboard

Description automatically generated