Multiple Regression

Marianne Huebner, Michigan State University

2015-05-01

In many regression models the response variable may be related to more than one explanatory variable. The predictors that have a significant effect can be incorporated in the model

### Example: Baseball

To predict the number of wins for a professional baseball team data for the following variables were collected: *batavg= batting average, rbi=runs batted in, stole= number of stolen bases, strkout=number of struck out, era=earned run average, caught=number caught stealing bases, errors=number of errors.*

baseball=read.csv("http://stt.msu.edu/Academics/ClassPages/uploads/SS15/351-4/baseball\_wins.csv", header=T)

Regression analysis with all predictors:

fit<-lm(wins~batavg+rbi+stole+strkout+caught+error+era, baseball)  
summary(fit)

##   
## Call:  
## lm(formula = wins ~ batavg + rbi + stole + strkout + caught +   
## error + era, data = baseball)  
##   
## Residuals:  
## 1 2 3 4 5 6 7 8   
## -0.95605 1.89557 -0.55961 3.11118 0.47108 -2.26097 -0.90437 -1.46067   
## 9 10 11 12   
## 3.58541 -0.07184 -3.61536 0.76563   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -39.37198 137.26291 -0.287 0.7885   
## batavg 341.61532 391.21354 0.873 0.4318   
## rbi 0.07169 0.04394 1.632 0.1781   
## stole 0.08151 0.11769 0.693 0.5267   
## strkout 0.06472 0.06741 0.960 0.3913   
## caught -0.26967 0.34536 -0.781 0.4785   
## error -0.09166 0.12249 -0.748 0.4958   
## era -15.11384 4.39408 -3.440 0.0263 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.511 on 4 degrees of freedom  
## Multiple R-squared: 0.9472, Adjusted R-squared: 0.8548   
## F-statistic: 10.25 on 7 and 4 DF, p-value: 0.02007

What are the three variables with the most significant pvalues?

sort(summary(fit)$coefficients[,4])[1:3]

## era rbi strkout   
## 0.02630357 0.17810820 0.39133977

Simple linear regression analysis with the two most significant predictors:

fit1<-lm(wins~era, baseball)  
summary(fit1)

##   
## Call:  
## lm(formula = wins ~ era, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.765 -4.277 1.929 4.236 8.694   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 143.500 19.398 7.398 2.32e-05 \*\*\*  
## era -16.469 5.086 -3.238 0.00889 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.752 on 10 degrees of freedom  
## Multiple R-squared: 0.5119, Adjusted R-squared: 0.4631   
## F-statistic: 10.49 on 1 and 10 DF, p-value: 0.008895

fit2<-lm(wins~rbi, baseball)  
summary(fit2)

##   
## Call:  
## lm(formula = wins ~ rbi, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.8393 -2.7850 -0.5008 5.4742 9.3035   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.68908 25.52107 0.497 0.6298   
## rbi 0.10715 0.03989 2.686 0.0229 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.366 on 10 degrees of freedom  
## Multiple R-squared: 0.4191, Adjusted R-squared: 0.361   
## F-statistic: 7.215 on 1 and 10 DF, p-value: 0.02286

The R-squared value for all seven predictors is 0.8548374 while the R-squared for *era* is 0.4630687 and for *rbi* 0.3610048.

A model with the these two variables is

fit3<-lm(wins~era+rbi, baseball)  
summary(fit3)

##   
## Call:  
## lm(formula = wins ~ era + rbi, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8277 -2.3129 0.9134 2.7598 4.1190   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 76.94330 18.54240 4.150 0.002486 \*\*   
## era -15.19631 2.97490 -5.108 0.000638 \*\*\*  
## rbi 0.09683 0.02139 4.526 0.001434 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.932 on 9 degrees of freedom  
## Multiple R-squared: 0.851, Adjusted R-squared: 0.8179   
## F-statistic: 25.71 on 2 and 9 DF, p-value: 0.0001901

This has R-squared value 0.8179163.

Is it possible to find a third variable, which together with *era* and *rbi*, improves R-squared?

fitx<-lm(wins~era+rbi+batavg, baseball); summary(fitx)

##   
## Call:  
## lm(formula = wins ~ era + rbi + batavg, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.209 -2.207 1.196 2.281 3.508   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19.52678 48.09116 0.406 0.695359   
## era -15.49761 2.88179 -5.378 0.000663 \*\*\*  
## rbi 0.07731 0.02563 3.017 0.016640 \*   
## batavg 276.89733 215.25556 1.286 0.234292   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.796 on 8 degrees of freedom  
## Multiple R-squared: 0.8766, Adjusted R-squared: 0.8303   
## F-statistic: 18.94 on 3 and 8 DF, p-value: 0.0005425

fitx<-lm(wins~era+rbi+stole, baseball); summary(fitx)

##   
## Call:  
## lm(formula = wins ~ era + rbi + stole, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.216 -2.866 1.325 2.794 4.121   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 57.73864 37.98620 1.520 0.1670   
## era -13.63972 4.07296 -3.349 0.0101 \*  
## rbi 0.11137 0.03330 3.345 0.0102 \*  
## stole 0.02702 0.04607 0.587 0.5737   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.084 on 8 degrees of freedom  
## Multiple R-squared: 0.8572, Adjusted R-squared: 0.8036   
## F-statistic: 16 on 3 and 8 DF, p-value: 0.000964

fitx<-lm(wins~era+rbi+strkout, baseball); summary(fitx)

##   
## Call:  
## lm(formula = wins ~ era + rbi + strkout, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.5152 -1.5633 0.5279 2.7919 3.8388   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.81645 29.36369 2.241 0.05530 .   
## era -15.29591 3.11270 -4.914 0.00117 \*\*  
## rbi 0.09949 0.02295 4.334 0.00250 \*\*  
## strkout 0.01054 0.02092 0.504 0.62782   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.106 on 8 degrees of freedom  
## Multiple R-squared: 0.8556, Adjusted R-squared: 0.8015   
## F-statistic: 15.8 on 3 and 8 DF, p-value: 0.001006

fitx<-lm(wins~era+rbi+caught, baseball); summary(fitx)

##   
## Call:  
## lm(formula = wins ~ era + rbi + caught, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.5475 -2.2543 0.7666 3.0950 3.6140   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.75294 33.63351 1.896 0.09462 .   
## era -14.32264 3.60430 -3.974 0.00410 \*\*  
## rbi 0.10776 0.03193 3.375 0.00971 \*\*  
## caught 0.04793 0.09986 0.480 0.64410   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.112 on 8 degrees of freedom  
## Multiple R-squared: 0.8552, Adjusted R-squared: 0.8009   
## F-statistic: 15.75 on 3 and 8 DF, p-value: 0.001017

fitx<-lm(wins~era+rbi+error, baseball); summary(fitx)

##   
## Call:  
## lm(formula = wins ~ era + rbi + error, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8798 -1.2882 -0.1748 1.4314 4.9845   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 89.85360 13.83885 6.493 0.000190 \*\*\*  
## era -12.66064 2.26918 -5.579 0.000523 \*\*\*  
## rbi 0.09731 0.01524 6.387 0.000212 \*\*\*  
## error -0.18080 0.05792 -3.121 0.014201 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.8 on 8 degrees of freedom  
## Multiple R-squared: 0.9328, Adjusted R-squared: 0.9076   
## F-statistic: 37.03 on 3 and 8 DF, p-value: 4.874e-05

The multiple regression model with the highest R-squared includes the variables *era, rbi, error* with adjusted R-squared 0.9076367.

Confidence intervals and Anova table is given as follows:

fitx<-lm(wins~era+rbi+error, baseball);  
confint(fitx)

## 2.5 % 97.5 %  
## (Intercept) 57.94115703 121.76603898  
## era -17.89338352 -7.42788675  
## rbi 0.06217374 0.13244343  
## error -0.31436969 -0.04722466

anova(fitx)

## Analysis of Variance Table  
##   
## Response: wins  
## Df Sum Sq Mean Sq F value Pr(>F)   
## era 1 478.10 478.10 60.9624 5.2e-05 \*\*\*  
## rbi 1 316.76 316.76 40.3901 0.0002194 \*\*\*  
## error 1 76.41 76.41 9.7425 0.0142009 \*   
## Residuals 8 62.74 7.84   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To test the utility of the full model against the null model the hypotheses are

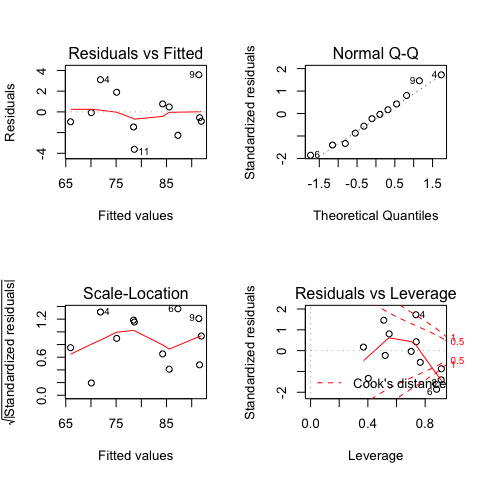
The ANOVA table is given by

library(knitr)  
fitx<-lm(wins~era+rbi+error, baseball);  
xx<-anova(fitx)  
msr<-sum(xx[1:3,2])/sum(xx[1:3,1])  
fvalue<-msr/xx[4,3]  
df1<-sum(xx[1:3,1])  
df2<-sum(xx[4,1])  
pval<-1-pf(fvalue,df1,df2)  
yy<-rbind(c(df1,sum(xx[1:3,2]),msr, fvalue, pval ), c(xx[4,1], xx[4,2], xx[4,3],NA, NA))  
colnames(yy)<-c("DF","SS", "MS", "F Value", "P(>F)")  
rownames(yy)<-c("regression", "residual")  
kable(yy, format='markdown')

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DF | SS | MS | F Value | P(>F) |
| regression | 3 | 871.26013 | 290.420044 | 37.03164 | 4.87e-05 |
| residual | 8 | 62.73987 | 7.842484 | NA | NA |

A residual plot and a normal probability plot for residuals assesses the validity of model

fit.lm<-lm(wins~era+rbi+error, baseball)  
par(mfrow=c(2,2))  
plot(fit)



par(mfrow=c(1,1))

If one wants to test the model with *era* against a model with *era, rbi,* and *error*, then the full model has predictors and the partial model has predictors. The test statistic for testing (partial model) versus (full model) is

with and degrees of freedom.

The ANOVA table comparing these two models is as follows:

fitx<-lm(wins~era+rbi+error, baseball); #p=3  
fity<-lm(wins~era, baseball); #l=1  
kable(anova(fitx)) # full model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| era | 1 | 478.09649 | 478.096490 | 60.962384 | 0.0000520 |
| rbi | 1 | 316.75845 | 316.758450 | 40.390069 | 0.0002194 |
| error | 1 | 76.40519 | 76.405192 | 9.742474 | 0.0142009 |
| Residuals | 8 | 62.73987 | 7.842484 | NA | NA |

kable(anova(fity)) # partial model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| era | 1 | 478.0965 | 478.09649 | 10.48679 | 0.0088945 |
| Residuals | 10 | 455.9035 | 45.59035 | NA | NA |

kable(anova(fity, fitx)) #ANOVA table for comparing full and partial model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| 10 | 455.90351 | NA | NA | NA | NA |
| 8 | 62.73987 | 2 | 393.1636 | 25.06627 | 0.0003587 |

xx<-anova(fitx)  
ssr<-sum(xx[2:3,2])  
msr<-ssr/sum(xx[2:3,1])  
fvalue<-msr/xx[4,3]  
df1<-sum(xx[2:3,1]) # p-l  
df2<-sum(xx[4,1]) # n-(p+1)  
pval<-1-pf(fvalue,df1,df2)  
yy<-rbind(c(df1,sum(xx[2:3,2]),msr,fvalue, pval), c(xx[4,1], xx[4,2],xx[4,3],NA,NA))  
colnames(yy)<-c("DF","SS", "MS", "F Value", "P(>F)")  
rownames(yy)<-c("regression", "residual")  
kable(yy) #ANOVA table for comparing full and partial model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DF | SS | MS | F Value | P(>F) |
| regression | 2 | 393.16364 | 196.581821 | 25.06627 | 0.0003587 |
| residual | 8 | 62.73987 | 7.842484 | NA | NA |