Test 2

Acts 170

A dataset on NBA teams and some of their statistics from 2010 are being used. The variables used are:

Wins - # games won in 2010  
PS – average Points Scored by their team  
PA – average Points Allowed scored by their opponents  
FG– average Field Goal percentage by their team  
FGA– average Field Goal Allowed percentage by their opponents  
S– average # of Steals by their team  
SA– average # of Steals Allowed by their opponents  
R– average Rebounds per game by their team  
RA– average Rebounds Allowed per game by their opponents

The first few lines of data look like:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Team** | **Wins** | **PS** | **PA** | **FG** | **FGA** | **S** | **SA** | **R** | **RA** |
| Atlanta | 53 | 101.68 | 97.02 | 46.8 | 46.0 | 7.22 | 6.17 | 41.70 | 41.37 |
| Boston | 50 | 99.22 | 95.56 | 48.3 | 45.1 | 8.54 | 7.04 | 38.59 | 40.07 |
| Charlotte | 44 | 95.28 | 93.81 | 45.3 | 44.8 | 7.70 | 7.48 | 40.82 | 39.63 |
| Chicago | 41 | 97.47 | 99.11 | 45.1 | 44.2 | 6.47 | 7.04 | 44.54 | 42.74 |

**1.)**

> test <- read.csv(choose.files(), header=TRUE)

> attach(test)

> dim(test)

[1] 30 10

a.) Start with a model that predicts Wins with all other variables as explanatory variables. We will call this model1. Do you think multicollinearity is a problem in this dataset? Explain.

> model1 = lm(Wins ~ PS + PA + FG + FGA + S + SA + R + RA)

> summary(model1)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -21.734177 89.350108 -0.243 0.810

PS 2.586712 0.382822 6.757 1.1e-06 \*\*\*

PA -2.794905 0.513028 -5.448 2.1e-05 \*\*\*

FG 0.467988 1.155195 0.405 0.689

FGA 0.464091 1.124596 0.413 0.684

S -0.006644 0.875005 -0.008 0.994

SA -0.454650 1.566086 -0.290 0.774

R 0.890540 0.663876 1.341 0.194

RA 0.162243 0.745372 0.218 0.830

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.235 on 21 degrees of freedom

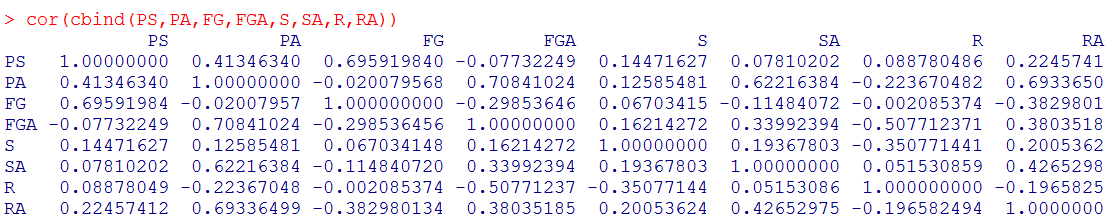
Multiple R-squared: 0.9575, Adjusted R-squared: 0.9414

F-statistic: 59.19 on 8 and 21 DF, p-value: 1.128e-12

We do suspect that we might have some multicollinearity since our F p-value for the model as a whole is very significant, but the individual t p-values are mostly insignificant.

b.) Investigate whether multicollinearity is a problem in this data by finding the correlations and variance inflation factors. Comment appropriately.

> cor(cbind(PS,PA,FG,FGA,S,SA,R,RA))



Whether or not there is multicollinearity is very possible with this output since there are some correlations at .708 (FGA & PA), .696 (FG & PS), .693 (RA & PA), etc.

> library(Rcmdr)

> vif(model1)

PS PA FG FGA S SA R RA

6.797668 15.136656 8.706193 7.891327 1.483505 2.678605 2.810951 7.741816

PA does have severe multicollinearity since it has a variance inflation factor larger than 10.

c.) Perform backward regression. The model that you end up with we will call model2.

> summary(model1)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -21.734177 89.350108 -0.243 0.810

PS 2.586712 0.382822 6.757 1.1e-06 \*\*\*

PA -2.794905 0.513028 -5.448 2.1e-05 \*\*\*

FG 0.467988 1.155195 0.405 0.689

FGA 0.464091 1.124596 0.413 0.684

S -0.006644 0.875005 -0.008 0.994

SA -0.454650 1.566086 -0.290 0.774

R 0.890540 0.663876 1.341 0.194

RA 0.162243 0.745372 0.218 0.830

---

> model1b = lm(Wins ~ PS + PA + FG + FGA + SA + R + RA)

> summary(model1b)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -21.7205 87.2783 -0.249 0.806

PS 2.5860 0.3631 7.122 3.85e-07 \*\*\*

PA -2.7935 0.4653 -6.003 4.84e-06 \*\*\*

FG 0.4677 1.1281 0.415 0.682

FGA 0.4621 1.0669 0.433 0.669

SA -0.4595 1.3978 -0.329 0.745

R 0.8917 0.6303 1.415 0.171

RA 0.1612 0.7151 0.225 0.824

---

> model1c = lm(Wins ~ PS + PA + FG + FGA + SA + R)

> summary(model1c)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.4696 48.1529 -0.114 0.911

PS 2.6279 0.3055 8.603 1.21e-08 \*\*\*

PA -2.7257 0.3477 -7.839 6.06e-08 \*\*\*

FG 0.2583 0.6266 0.412 0.684

FGA 0.3075 0.8004 0.384 0.704

SA -0.4633 1.3686 -0.339 0.738

R 0.8024 0.4799 1.672 0.108

---

> model1d = lm(Wins ~ PS + PA + FG + FGA + R)

> summary(model1d)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.0316 47.2283 -0.128 0.899

PS 2.6729 0.2699 9.904 5.93e-10 \*\*\*

PA -2.8042 0.2541 -11.036 6.93e-11 \*\*\*

FG 0.2162 0.6027 0.359 0.723

FGA 0.3910 0.7473 0.523 0.606

R 0.7708 0.4620 1.669 0.108

---

> model1e = lm(Wins ~ PS + PA + FGA + R)

> summary(model1e)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1258 40.6643 0.052 0.959

PS 2.7428 0.1834 14.955 5.62e-14 \*\*\*

PA -2.8314 0.2383 -11.880 8.90e-12 \*\*\*

FGA 0.3760 0.7330 0.513 0.612

R 0.7279 0.4384 1.660 0.109

---

> model1f = lm(Wins ~ PS + PA + R)

> summary(model1f)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 19.3802 22.5297 0.860 0.398

PS 2.6901 0.1498 17.962 3.55e-16 \*\*\*

PA -2.7322 0.1375 -19.877 < 2e-16 \*\*\*

R 0.6183 0.3773 1.639 0.113

---

> model2 = lm(Wins ~ PS + PA)

> summary(model2)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 46.6301 15.6679 2.976 0.00609 \*\*

PS 2.7402 0.1511 18.133 < 2e-16 \*\*\*

PA -2.7969 0.1357 -20.609 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.031 on 27 degrees of freedom

Multiple R-squared: 0.9521, Adjusted R-squared: 0.9485

F-statistic: 268.1 on 2 and 27 DF, p-value: < 2.2e-16

d.) Is model2 significant? How do you know? State your null hypothesis, alternative hypothesis and all components of the decision-making rule.

H0: β1 = β2 = 0  
H1: at least 1 β ≠ 0

F p-value <2.2e-16 which is < alpha = 0.05  
So, we do Reject Ho and we do believe that model2 is significant as a whole at predicting Wins.

e.) Make plots of S, SA, R, and RA to see if any of them need a transformation. Replace out any variables you transform into model1, then do stepwise regression, and call this model3.

> layout(matrix(c(1,2,3,4,5,6,7,8,9,10), byrow=TRUE, ncol=5) )

> plot.new()

> hist(S)

> hist(SA)

> hist(R)

> hist(RA)

> hist(Wins)

> plot(S, Wins)

> text(S, Wins, labels=row.names(test), pos=1)

> plot(SA, Wins)

> plot(R, Wins)

> plot(RA, Wins)



We see in the histogram of RA that it is the most skewed and looks skewed to the right. None of the scatterplots here seem curved, so we do not need a squared term.

> par(mfrow=c(1,2))

> hist(RA)

> lnRA = log(RA)

> hist(lnRA)



We see that the log version is more symmetric and improved from the original RA.

> model3 = lm(Wins ~ PS + PA + FG + FGA + S + SA + R + lnRA)

> step(model3)

Start: AIC=77.66

Wins ~ PS + PA + FG + FGA + S + SA + R + lnRA

Df Sum of Sq RSS AIC

- S 1 0.00 219.20 75.664

- SA 1 0.83 220.02 75.776

- lnRA 1 1.00 220.20 75.800

- FGA 1 2.34 221.54 75.982

- FG 1 2.42 221.62 75.993

<none> 219.19 77.663

- R 1 20.81 240.00 78.383

- PA 1 320.32 539.52 102.684

- PS 1 480.24 699.43 110.472

Step: AIC=75.66

Wins ~ PS + PA + FG + FGA + SA + R + lnRA

Df Sum of Sq RSS AIC

- lnRA 1 1.01 220.21 73.801

- SA 1 1.06 220.25 73.808

- FG 1 2.42 221.62 73.993

- FGA 1 2.42 221.62 73.993

<none> 219.20 75.664

- R 1 22.44 241.63 76.587

- PA 1 368.56 587.76 103.254

- PS 1 511.61 730.81 109.789

Step: AIC=73.8

Wins ~ PS + PA + FG + FGA + SA + R

Df Sum of Sq RSS AIC

- SA 1 1.10 221.31 71.951

- FGA 1 1.41 221.62 71.993

- FG 1 1.63 221.84 72.022

<none> 220.21 73.801

- R 1 26.77 246.98 75.243

- PA 1 588.41 808.62 110.824

- PS 1 708.61 928.82 114.982

Step: AIC=71.95

Wins ~ PS + PA + FG + FGA + R

Df Sum of Sq RSS AIC

- FG 1 1.19 222.49 70.111

- FGA 1 2.52 223.83 70.291

<none> 221.31 71.951

- R 1 25.67 246.98 73.243

- PS 1 904.58 1125.88 118.754

- PA 1 1123.04 1344.35 124.074

Step: AIC=70.11

Wins ~ PS + PA + FGA + R

Df Sum of Sq RSS AIC

- FGA 1 2.34 224.84 68.425

<none> 222.49 70.111

- R 1 24.54 247.03 71.250

- PA 1 1256.15 1478.64 124.930

- PS 1 1990.38 2212.87 137.025

Step: AIC=68.43

Wins ~ PS + PA + R

Df Sum of Sq RSS AIC

<none> 224.8 68.425

- R 1 23.2 248.1 69.374

- PS 1 2790.0 3014.8 144.303

- PA 1 3416.7 3641.6 149.969

Call:

lm(formula = Wins ~ PS + PA + R)

Coefficients:

(Intercept) PS PA R

19.3802 2.6901 -2.7322 0.6183

> model3 = lm(Wins ~ PS + PA + R)

> summary(model3)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 19.3802 22.5297 0.860 0.398

PS 2.6901 0.1498 17.962 3.55e-16 \*\*\*

PA -2.7322 0.1375 -19.877 < 2e-16 \*\*\*

R 0.6183 0.3773 1.639 0.113

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.941 on 26 degrees of freedom

Multiple R-squared: 0.9565, Adjusted R-squared: 0.9515

F-statistic: 190.8 on 3 and 26 DF, p-value: < 2.2e-16

f.) Thoroughly compare the goodness of fits models 1, 2, and 3.

> extractAIC(model1)

[1] 9.00000 77.73222

> extractAIC(model2)

[1] 3.00000 69.37383

> extractAIC(model3)

[1] 4.00000 68.42519

Model1 Model2 Model3

s 3.235 3.031 2.941

R² .9575 .9521 .9565

R²a  .9414 .9485 .9515

F p-value 1.128e-12 <2.2e-16 <2.2e-16

AIC 77.732 69.374 68.425

I highlighted the best values in red. Model3 seems to be the best, although there is not much difference in any of the values for any of the models.

g.) Remove all outliers and high leverage points from model2, call this model4. Note any significant changes in the model.

> rs = rstandard(model2)

> rs[order(rs)]

27 9 18 26 23 22 29 30

-1.750566634 -1.745659289 -1.437851062 -1.289126500 -1.085044672 -0.971653023 -0.917887062 -0.623723553

3 2 20 1 21 15 11 25

-0.474487768 -0.427626679 -0.410470590 -0.309303736 -0.208777415 -0.188660910 -0.161818583 -0.114385873

24 8 16 17 7 5 10 28

-0.045116556 -0.004880798 0.050172925 0.285196285 0.335234930 0.657816074 0.743475952 0.872165821

19 13 14 4 12 6

0.988039931 1.001133795 1.162742817 1.516847322 1.940043923 2.207837749

> lev = hatvalues(model2)

> lev[order(lev)]

19 10 6 4 21 11 14 23 26 1

0.04460250 0.04485004 0.04520200 0.05182218 0.05244393 0.05404782 0.05555259 0.06073736 0.06798041 0.06817986

13 16 2 27 30 20 29 25 28 5

0.06910814 0.06979215 0.07377597 0.07695761 0.08286451 0.08355289 0.08420508 0.08643767 0.09424525 0.10225096

15 7 12 22 8 3 17 18 24 9

0.10425228 0.10856050 0.10884525 0.12204406 0.12464335 0.12488578 0.18515438 0.21423719 0.23012453 0.30864376

> 9/30

[1] 0.3

> model4 = lm(Wins ~ PS + PA, subset=-c(6,9))

> summary(model4)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 31.5156 16.0742 1.961 0.0612 .

PS 2.7666 0.1373 20.153 <2e-16 \*\*\*

PA -2.6729 0.1306 -20.472 <2e-16 \*\*\*

---

Residual standard error: 2.666 on 25 degrees of freedom

Multiple R-squared: 0.9626, Adjusted R-squared: 0.9596

F-statistic: 321.8 on 2 and 25 DF, p-value: < 2.2e-16

We see that R² and R²a both increased a little (from .9521 to .9626 and from .9485 to .9596), s went down from 3.031 to 2.66 and all p-values are still the same at virtually 0. The model has improved in all aspects here.

h.) Is the last variable in your model4 significant? State your null hypothesis, alternative hypothesis and all components of the decision-making rule.

H0: β2 = 0  
H1: β2 ≠ 0

t p-value <2.2e-16 which is < alpha = 0.05  
So, we Reject Ho and we do believe that PA (Points Allowed) is significant at predicting Wins.

i.) Plot the residuals and the fitted values of model4. Comment on it appropriately.

> plot(residuals(model4) ~ fitted.values(model4), main="Residuals vs. Fitted Values")

I actually do not think this looks like a random pattern. It seems to have a curved or quadratic pattern to it. Perhaps we are missing a squared term of a variable in our model. I also do not think that there is homoscedasticity (equal variance) in the plot as the middle values seem to have smaller variance than the ends of the plot.

j.) Comment on whether you feel your model4 suffers from bias due to omitted variables. Explain why or why not you think this.

I would normally not think this since the R² values are so high (along with our significant F p-value) so this model is explaining almost all variation that there is to be explained in predicting Wins. However, the possible pattern present in the residual plot above may denote an omitted variable in that a squared term may be missing.

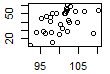
k.) Make a scatterplot matrix of all the variables in model1. Pick the one variable that you feel looks most like it needs a squared term. Insert this term into model2 and see if it is significant.

> plots=data.frame(PS,PA,FG,FGA,S,SA,R,RA,Wins)

> pairs(plots,upper.panel=NULL)



None of the plots look terribly curved. But, the most likely candidate to me is PS.



> PS2 = PS\*PS

> model2c = lm(Wins ~ PS + PA + PS2)

> summary(model2c)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -284.99441 264.45789 -1.078 0.2911

PS 9.22383 5.16368 1.786 0.0857 .

PA -2.74006 0.14170 -19.337 <2e-16 \*\*\*

PS2 -0.03219 0.02563 -1.256 0.2202

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.999 on 26 degrees of freedom

Multiple R-squared: 0.9548, Adjusted R-squared: 0.9496

F-statistic: 183.1 on 3 and 26 DF, p-value: < 2.2e-16

This squared term is not significant since its p-value = .2202.

l.) Change PS to a categorical variable that divides the teams into low scorers (98 points or less), medium scorers (104 to over 98), and high scorers (over 104). Run a one-factor ANOVA model and see whether your new categorical variable for points scored is a significant predictor.

> PSlow= (1\*(PS<=98))

> PSmed= (1\*(PS>98 & PS<=104))

> PShigh = (1\*(104<PS))

> check = data.frame(PS,PSlow, PSmed,PShigh)

> fix(check)

OR

> categories=cut(PS,br=c(92,98,104,111))

> table(categories)

categories

(92,98] (98,104] (104,111]

9 16 5

> model5 = lm(Wins ~ categories)

> summary(model5)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.222 4.251 7.815 2.1e-08 \*\*\*

categories(98,104] 10.840 5.314 2.040 0.0512 .

categories(104,111] 11.978 7.113 1.684 0.1037

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.75 on 27 degrees of freedom

Multiple R-squared: 0.1513, Adjusted R-squared: 0.08841

F-statistic: 2.406 on 2 and 27 DF, p-value: 0.1092

OR

> model5 = lm(Wins ~ PSmed + PShigh)

> summary(model5)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.222 4.251 7.815 2.1e-08 \*\*\*

PSmed 10.840 5.314 2.040 0.0512 .

PShigh 11.978 7.113 1.684 0.1037

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.75 on 27 degrees of freedom

Multiple R-squared: 0.1513, Adjusted R-squared: 0.08841

F-statistic: 2.406 on 2 and 27 DF, p-value: 0.1092

This model as a whole is not significant since the p-value = .1092.

m.) Use Tukey to see whether there are any significant differences in the amount of Points Allowed (PA) according to the levels in your new categorical variable.

> TukeyHSD(aov(PA ~ categories))

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = PA ~ categories)

$categories

diff lwr upr p adj

(98,104]-(92,98] 1.284097 -3.0242471 5.592442 0.7427324

(104,111]-(92,98] 6.246222 0.4788311 12.013613 0.0317891

(104,111]-(98,104] 4.962125 -0.3355621 10.259812 0.0696377

The answer here depends on your choice of p-value. If we choose alpha = 0.05, then the low scoring teams and the high scoring teams do have a significant difference in the number of Points Allowed during a game (with the high scorers allowing more points in a game by 6.24 on average).

If you picked alpha = .10, then the medium and high scorers are also different.

If you picked alpha = .01, then there are no differences between the scorers and the amount of points they allow.