Regression
Lesson 2a

Kevin Zollicoffer

10/21/2013

Introduction

Regression assignment 2a using R.

The complete source for this assigment is available on Github:

`https://github.com/zollie/PASS-Regression-Assignment2a`

Problem 3.1

```
> movs <- read.csv("~/R/PASS/Regression/Assignment2a/movies.csv")
```

a

$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 = \hat{b}_0 + \hat{b}_1 Rate + \hat{b}_2 User + b_3 Meta$

or

$E(Box|(Rate, User, Meta)) = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 = \hat{b}_0 + \hat{b}_1 Rate + \hat{b}_2 User + b_3 Meta$

**The are no page numbers in the e-book version of my text book*

b

```
> model <- lm(Box ~ Rate + User + Meta, data=movs)
> summary(model)

Call:
lm(formula = Box ~ Rate + User + Meta, data = movs)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-49.202 -22.749  -3.598   7.266  90.059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -169.0862    92.0925  -1.836  0.08055 .
Rate          35.4962    18.9956   1.869  0.07569 .
User           0.4328     0.1472   2.940  0.00783 **
Meta           1.2462     0.8047   1.549  0.13640
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 39.53 on 21 degrees of freedom
Multiple R-squared:  0.8841,        Adjusted R-squared:  0.8675
F-statistic: 53.37 on 3 and 21 DF,  p-value: 5.35e-10
```

$\hat{Y} = -169.0862 + 35.4962Rate + 0.4328User + 1.2462Meta$

c

For every 1 point increase in user rating on IMDB for a given movie, holding everything else constant, this model predicts an increase in box office receipts of 35.49 million in inflation unadjusted US dollars.

Problem 3.3

a

```
> model2 <- lm(Box ~ Rate+User+Meta+Len+Win+Nom, data=movs)
> summary(model2)

Call:
lm(formula = Box ~ Rate + User + Meta + Len + Win + Nom, data = movs)

Residuals:
    Min      1Q  Median      3Q     Max
-53.161 -22.013  -3.864   9.517  84.574

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -172.28110  106.51894  -1.617   0.1232
Rate          35.34769   22.44744   1.575   0.1327
User           0.38894    0.19304   2.015   0.0591 .
Meta           1.25615    0.89110   1.410   0.1757
Len            0.02473    0.54429   0.045   0.9643
Win           -0.02080    1.33384  -0.016   0.9877
```

```
Nom            0.37261    0.87286   0.427   0.6745
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 42.45 on 18 degrees of freedom
Multiple R-squared:  0.8854,        Adjusted R-squared:  0.8472
F-statistic: 23.18 on 6 and 18 DF,  p-value: 1.505e-07
```

```
> rss2 <- sum(model2$residuals^2)
> rss2
```

```
[1] 32435.31
```

$RSS = 32435.31$

b

```
> summary(model)
```

```
Call:
lm(formula = Box ~ Rate + User + Meta, data = movs)

Residuals:
    Min      1Q  Median      3Q     Max
-49.202 -22.749  -3.598   7.266  90.059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -169.0862    92.0925  -1.836  0.08055 .
Rate          35.4962    18.9956   1.869  0.07569 .
User           0.4328     0.1472   2.940  0.00783 **
Meta           1.2462     0.8047   1.549  0.13640
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 39.53 on 21 degrees of freedom
Multiple R-squared:  0.8841,        Adjusted R-squared:  0.8675
F-statistic: 53.37 on 3 and 21 DF,  p-value: 5.35e-10
```

```
> rss <- sum(model$residuals^2)
> rss
```

```
[1] 32822.96
```

$RSS = 32822.96$

c

Global usefulness test

FYI: I think the denominator degrees of freedom in the question hint are wrong given that k=3 and n=25 of the sample movies data

$TSS = \sum_{i=1}^{n}(Y_i - m_y)^2$
$RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$tss
Fstatistic= $\frac{(TSS-RSS)/k}{RSS/(n-k-1)}$

$H_0 = b_4 = b_5 = b_6 = 0$

$H_a = b_4 \neq 0 \vee b_5 \neq 0 \vee b_6 \neq 0$

significance level is 5% $(1 - .95 = .05)$ for upper tail test

```
> options(scipen=999) # disable scientific notation
> model0 <- lm(Box ~ Len+Win+Nom, data=movs)
> summary(model0)

Call:
lm(formula = Box ~ Len + Win + Nom, data = movs)

Residuals:
     Min       1Q   Median       3Q      Max
-140.226  -23.789   -4.039   29.403  117.709

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.1265    80.7745   0.101 0.920816
Len           1.0093     0.7063   1.429 0.167748
Win           5.0604     1.2733   3.974 0.000691 ***
Nom           1.4002     1.1450   1.223 0.234928
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 61.82 on 21 degrees of freedom
Multiple R-squared:  0.7165,        Adjusted R-squared:  0.676
F-statistic: 17.69 on 3 and 21 DF,  p-value: 0.000005816

> n <- nrow(movs)
> k <- 3
> df2 <- n-k-1
> m_y0 <- mean(model0$fitted.values)
```

```
> tss0 <- sum(sapply(model0$fitted.values, function(v) { (v-m_y0)^2 }))
> rss0 <- sum(model0$residuals^2)
> fstat0 <- ((tss0-rss0)/k)/(rss0/(n-k-1))
> fstat0

[1] 10.69237

> pf <- pf(fstat0, k, df2, lower.tail=F)
> pf

[1] 0.000178813

>
```

pf $< .05$ therefore we reject H_0. It is plausible that $H_a = b_4 \neq 0 \vee b_5 \neq 0 \vee b_6 \neq 0$

d

```
> summary(model)

Call:
lm(formula = Box ~ Rate + User + Meta, data = movs)

Residuals:
    Min      1Q  Median      3Q     Max
-49.202 -22.749  -3.598   7.266  90.059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -169.0862    92.0925  -1.836  0.08055 .
Rate          35.4962    18.9956   1.869  0.07569 .
User           0.4328     0.1472   2.940  0.00783 **
Meta           1.2462     0.8047   1.549  0.13640
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 39.53 on 21 degrees of freedom
Multiple R-squared:  0.8841,        Adjusted R-squared:  0.8675
F-statistic: 53.37 on 3 and 21 DF,  p-value: 0.000000000535

> summary(model2)

Call:
lm(formula = Box ~ Rate + User + Meta + Len + Win + Nom, data = movs)

Residuals:
    Min      1Q  Median      3Q     Max
-53.161 -22.013  -3.864   9.517  84.574
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -172.28110  106.51894  -1.617   0.1232
Rate          35.34769   22.44744   1.575   0.1327
User           0.38894    0.19304   2.015   0.0591 .
Meta           1.25615    0.89110   1.410   0.1757
Len            0.02473    0.54429   0.045   0.9643
Win           -0.02080    1.33384  -0.016   0.9877
Nom            0.37261    0.87286   0.427   0.6745
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 42.45 on 18 degrees of freedom
Multiple R-squared:  0.8854,        Adjusted R-squared:  0.8472
F-statistic: 23.18 on 6 and 18 DF,  p-value: 0.0000001505
```

```
> anova(model, model2)

Analysis of Variance Table

Model 1: Box ~ Rate + User + Meta
Model 2: Box ~ Rate + User + Meta + Len + Win + Nom
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     21 32823
2     18 32435  3    387.66 0.0717 0.9744
```

The F-statistic of .0717 $<$ 3.16 and the p-value of .9744 $>$ .05 therefore we do not reject H_0. There appears to be strong evidence that $H_0 = b_4 = b_5 = b_6 = 0$ holds.

e

```
> summary(model)

Call:
lm(formula = Box ~ Rate + User + Meta, data = movs)

Residuals:
    Min      1Q  Median      3Q     Max
-49.202 -22.749  -3.598   7.266  90.059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -169.0862    92.0925  -1.836  0.08055 .
Rate          35.4962    18.9956   1.869  0.07569 .
User           0.4328     0.1472   2.940  0.00783 **
```

```
Meta           1.2462     0.8047   1.549  0.13640
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 39.53 on 21 degrees of freedom
Multiple R-squared:  0.8841,        Adjusted R-squared:  0.8675
F-statistic: 53.37 on 3 and 21 DF,  p-value: 0.000000000535
```

```
> summary(model2)

Call:
lm(formula = Box ~ Rate + User + Meta + Len + Win + Nom, data = movs)

Residuals:
    Min      1Q  Median      3Q     Max
-53.161 -22.013  -3.864   9.517  84.574

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -172.28110  106.51894  -1.617   0.1232
Rate          35.34769   22.44744   1.575   0.1327
User           0.38894    0.19304   2.015   0.0591 .
Meta           1.25615    0.89110   1.410   0.1757
Len            0.02473    0.54429   0.045   0.9643
Win           -0.02080    1.33384  -0.016   0.9877
Nom            0.37261    0.87286   0.427   0.6745
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 42.45 on 18 degrees of freedom
Multiple R-squared:  0.8854,        Adjusted R-squared:  0.8472
F-statistic: 23.18 on 6 and 18 DF,  p-value: 0.0000001505
```

For the reduced model $s = 39.53, R^2 = .8841, adjusted R^2 = .8675$

For the complete model $s = 42.45, R^2 = .8854, adjusted R^2 = .8472$

s is lower for the reduced model corroborating the findings related to the extra predictors in the full model in *c* and *d*. R^2 is higher for the complete model but this appears mostly do to overfitting. Adjusted R^2 is lower for the complete model lending support to this view.