Regression
Lesson 2b

Kevin Zollicoffer

10/21/2013

Introduction

Regression assignment 2b using R

The complete source for this assigment is available on Github:

`https://github.com/zollie/PASS-Regression-Assignment2b`

Problem 3.5

a

$E(Att) = 28.721 + 1.350Pop - .972Teams - .238Temp$

or

$\hat{Att} = 28.721 + 1.350Pop - .972Teams - .238Temp$

b

Global usefulness test

$H_0 = b_0 = b_1 = b_2 = 0$

$H_a = b_0 \neq 0 \vee b_1 \neq 0 \vee b_2 \neq 0$

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

$R^2 = .914$ $k = 3$ $n = 12$

F-statistic= $\frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{.914/3}{(1-.914)/(12-3-1)} = \frac{.3046687}{.01075} = 28.27907$

$28.279 > 4.07$ therefore we reject H_0

c

$H_0 = b_1 = 0$
$H_a = b_1 < 0$
p-value= $.037/2 = .0185$
$.0185 < .05$ therefore reject H_0

This suggestes the model predicts a non-zero drop in attendence for each additional male major professional sports team to an MLS city.

d

The coefficient for Teams is -0.972, therefore for every additional male professional major sport team, all other things constant, the model predicts a decline of 972,000 attedees for that cities MLS franchise.

e

Presumably, home attendance for a sports franchise is the major source of revenue for a team. These attendees pay a ticket fee to attend the game. Knowing the fee, a potential team, or league can determine the break even point for attendence given the potential location of the team. A model such as this gives a data driven, objective, decision making tool that can help determine whether a potential location is viable for an MLS franchise.

Problem 3.6

```
> smsa <- read.csv("~/R/PASS/Regression/Assignment2b/smsa.csv")
```

a

```
> model <- lm(Mort ~ Edu+Nwt+Jant+Rain+Nox+Hum+Inc, data=smsa)
> summary(model)

Call:
lm(formula = Mort ~ Edu + Nwt + Jant + Rain + Nox + Hum + Inc,
    data = smsa)

Residuals:
    Min      1Q  Median      3Q     Max
-84.380 -22.118   2.907  23.154  77.369
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1006.2441    95.0827  10.583 3.84e-14 ***
Edu          -15.3459     7.2515  -2.116  0.03954 *
Nwt            4.2140     0.6850   6.152 1.47e-07 ***
Jant          -2.1500     0.6593  -3.261  0.00204 **
Rain           1.6238     0.5643   2.878  0.00596 **
Nox           18.5481     5.5065   3.368  0.00150 **
Hum            0.5371     0.9024   0.595  0.55451
Inc           -0.3453     1.3038  -0.265  0.79227
---
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: 35.48 on 48 degrees of freedom
Multiple R-squared:  0.7137,	Adjusted R-squared:  0.6719
F-statistic: 17.09 on 7 and 48 DF,  p-value: 4.183e-11
```

$\hat{Y} = 1006.2441 - 15.3459 Edu + 4.2140 Nwt - 2.15 Jant + 1.6238 Rain + 18.5481 Nox + .5371 Hum - .3453 Inc$

b

$H_0 = Hum = Inc = 0$
$H_a = Hum \neq 0 \vee Inc \neq 0$

```
> model0 <- lm(Mort ~ Hum+Inc, data=smsa)
> summary(model0)

Call:
lm(formula = Mort ~ Hum + Inc, data = smsa)

Residuals:
     Min       1Q   Median       3Q      Max
-118.372  -39.302    1.274   43.194  170.185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1118.9420    98.7427  11.332 8.88e-16 ***
Hum           -0.8753     1.4979  -0.584   0.5615
Inc           -3.7213     1.8264  -2.038   0.0466 *
---
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: 60.35 on 53 degrees of freedom
Multiple R-squared:  0.08512,	Adjusted R-squared:  0.0506
F-statistic: 2.466 on 2 and 53 DF,  p-value: 0.09466
```

```
> k <- 2
> n <- nrow(smsa)
> df2 <- n-k-1
> qf(0.05, k, df2, lower.tail=F)

[1] 3.171626
```

$2.466 < 3.17$ therefore we do not reject H_0. The coefficients for Hum and Inc may be statistically 0.

c

```
> options(scipen=999) # disable scientific notation
> modelr <- lm(Mort ~ Edu+Nwt+Jant+Rain+Nox, data=smsa)
> summary(modelr)

Call:
lm(formula = Mort ~ Edu + Nwt + Jant + Rain + Nox, data = smsa)

Residuals:
    Min      1Q  Median      3Q     Max
-86.139 -24.728   4.088  21.200  79.659

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 1028.2323    84.9148  12.109 < 0.0000000000000002 ***
Edu          -15.5887     6.4460  -2.418              0.01927 *
Nwt            4.1807     0.6600   6.334          0.000000066 ***
Jant          -2.1313     0.6369  -3.347              0.00156 **
Rain           1.6331     0.5551   2.942              0.00493 **
Nox           18.4132     5.2926   3.479              0.00105 **
---
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: 34.91 on 50 degrees of freedom
Multiple R-squared:  0.7111,        Adjusted R-squared:  0.6822
F-statistic: 24.62 on 5 and 50 DF,  p-value: 0.000000000002044
```

Edu

$H_0 = Edu = 0$ $H_a = Edu \neq 0$
p-value of $0.0000000000000002 < 0.025$ therefore reject H_0

Nwt

$H_0 = Nwt = 0$ $H_a = Nwt \neq 0$
p-value of $0.01927 < 0.025$ therefore reject H_0

Jant

$H_0 = Jant = 0$ $H_a = Jant \neq 0$
p-value of 0.000000066 $<$ 0.025 therefore reject H_0

Rain

$H_0 = Rain = 0$ $H_a = Rain \neq 0$
p-value of 0.00493 $<$ 0.025 therefore reject H_0

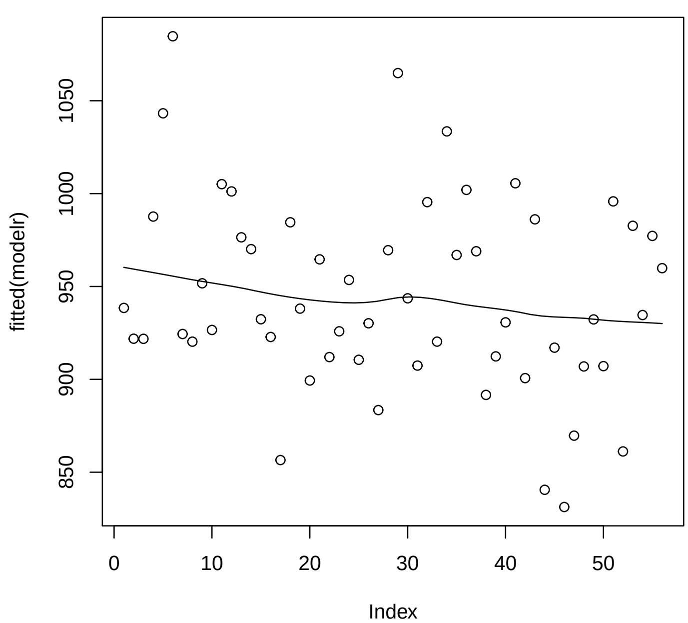
Nox

$H_0 = Nox = 0$ $H_a = Nox \neq 0$
p-value of 0.00493 $<$ 0.025 therefore reject H_0

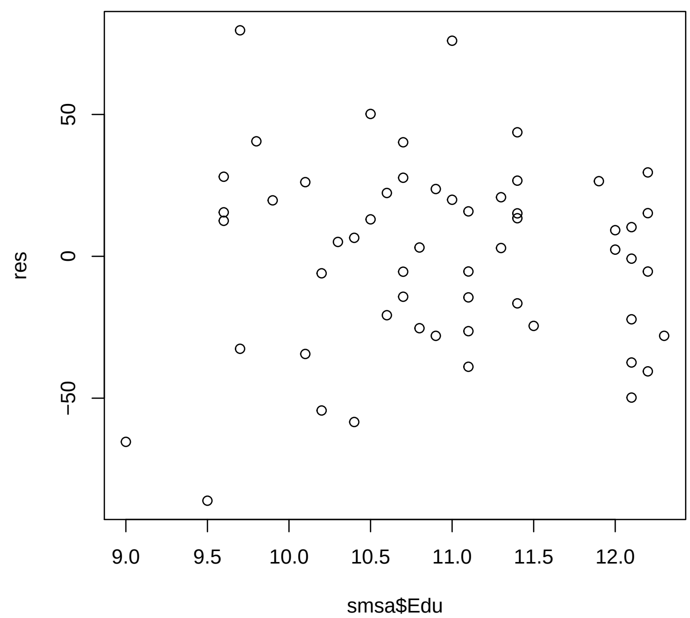
d

Random error assumptions

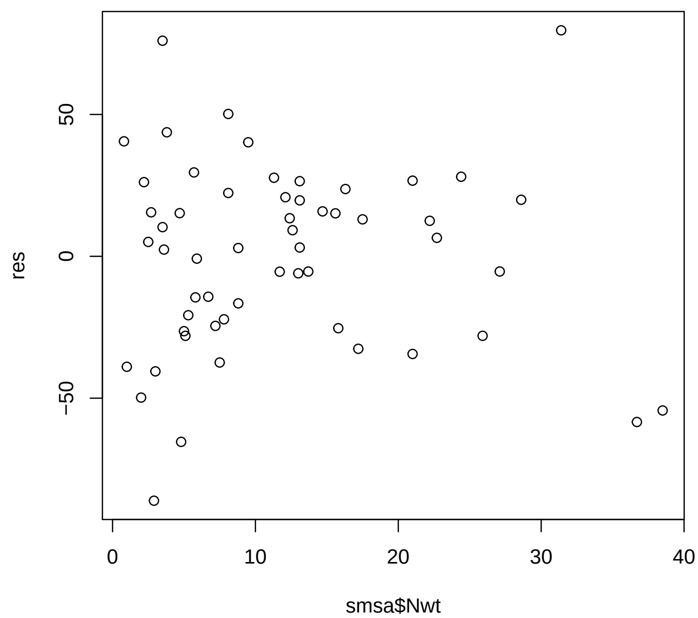
```
> scatter.smooth(fitted(modelr))
```



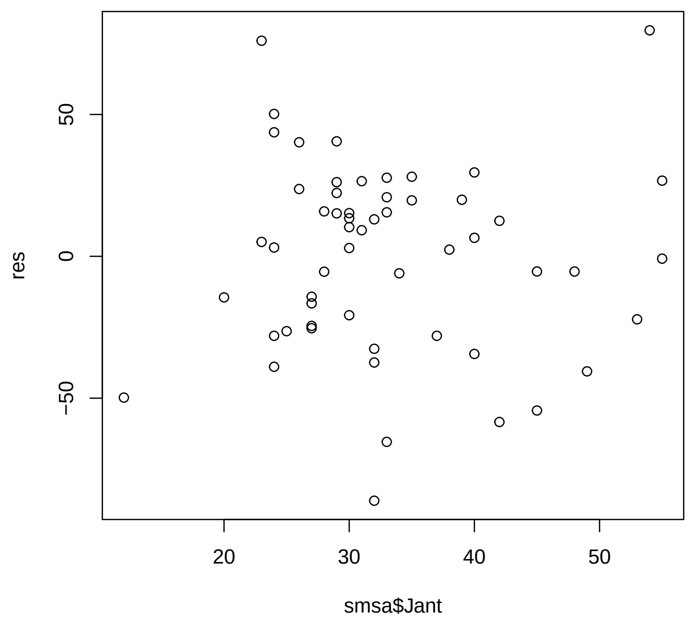

```
> res <- residuals(modelr)
> fitted <- predict(modelr)
> plot(smsa$Edu, res)
```



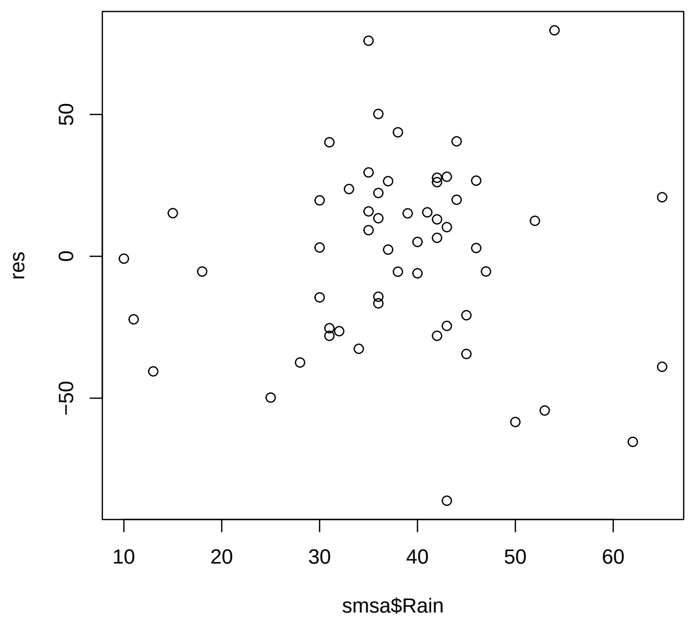

```
> plot(smsa$Nwt, res)
```



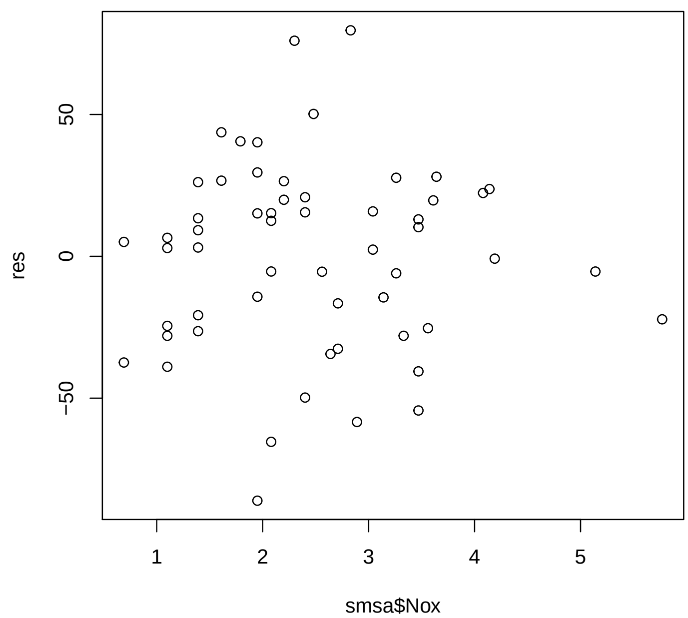

```
> plot(smsa$Jant, res)
```



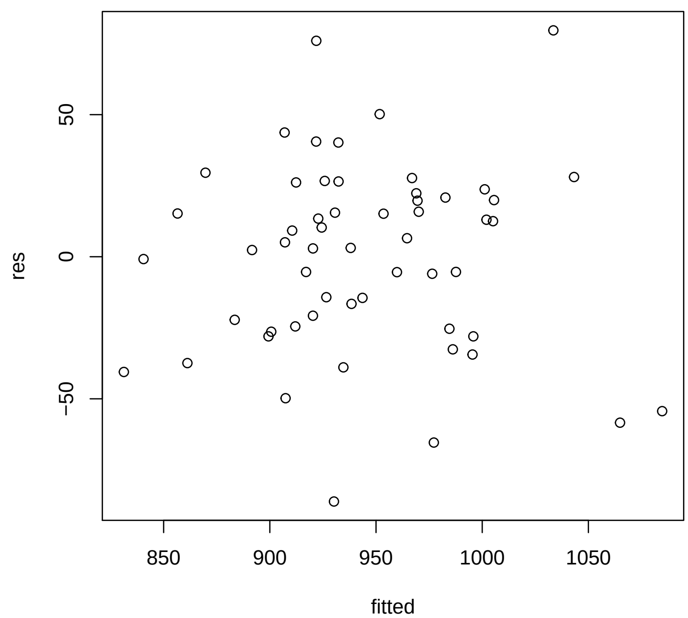

```
> plot(smsa$Rain, res)
```



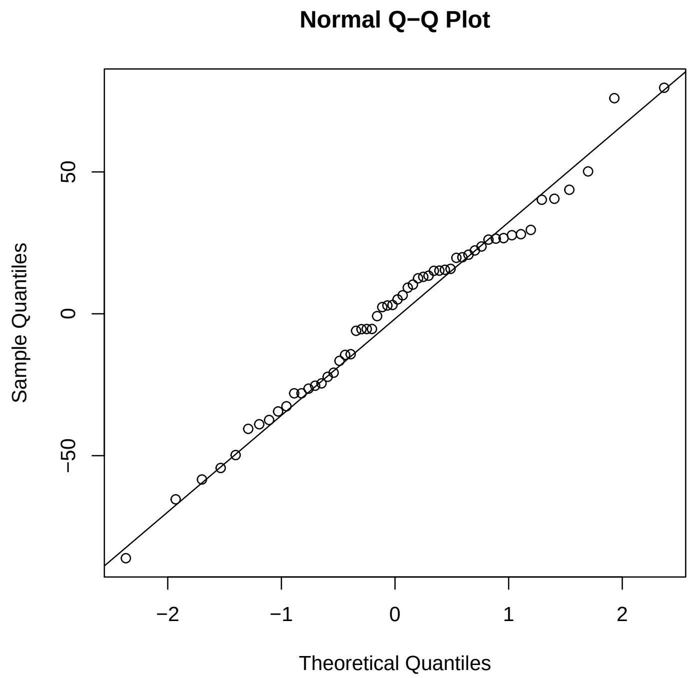

```
> plot(smsa$Nox, res)
```



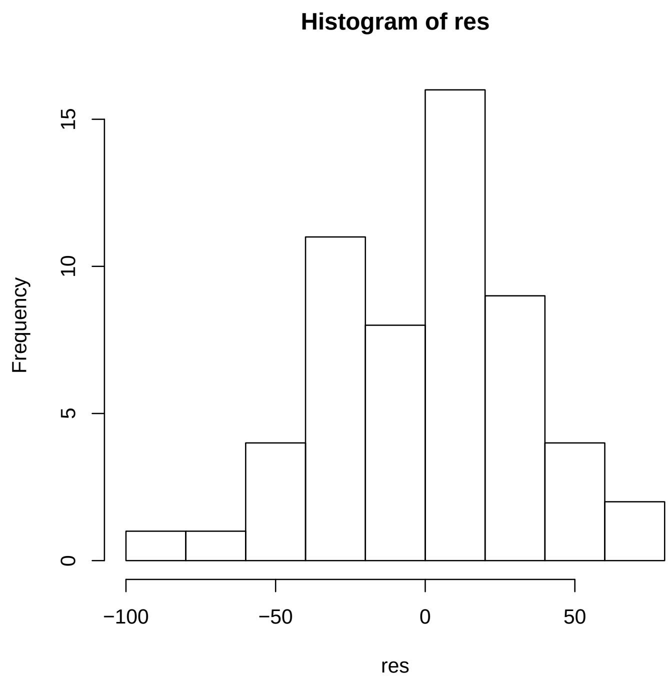

```
> plot(fitted, res)
```




```
> qqnorm(res)
> qqline(res)
```




```
> hist(res, freq=T)
```



e

$$\hat{Y} = 1028.2323 - 15.5887Edu + 4.1807Nwt - 2.1313Jant + 1.6331Rain + 18.4132Nox$$

The signs of the estimated regression parameters make sense in this context in a few ways. Life expectancy increeases with the level of education (the rate of mortality declines), the rate of mortality incresases as the amount of NO gas in the atmosphere increases. Mortality increases with amount of rain as well, perhaps due to accidents.

f

```
> nd <- data.frame(Edu=10,Nwt=15,Jant=35,Rain=40,Nox=2)
> predict(modelr, newdata=nd, interval="confidence")

       fit      lwr      upr
1 962.6092 946.2731 978.9454
```

g

```
> predict(modelr, newdata=nd, interval="prediction")
```

```
       fit      lwr      upr
1 962.6092 890.6053 1034.613
```