DA3

2025-06-20

carseats = read.csv("~/Downloads/Carseats.csv")  
contrasts(factor(carseats$Urban))

## Yes  
## No 0  
## Yes 1

contrasts(factor(carseats$US))

## Yes  
## No 0  
## Yes 1

summary(lm(carseats$Sales ~ carseats$Price + carseats$Urban + carseats$US))

##   
## Call:  
## lm(formula = carseats$Sales ~ carseats$Price + carseats$Urban +   
## carseats$US)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9206 -1.6220 -0.0564 1.5786 7.0581   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*  
## carseats$Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*  
## carseats$UrbanYes -0.021916 0.271650 -0.081 0.936   
## carseats$USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.472 on 396 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335   
## F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

1. According to these, if there was no effect of different predictors (Price was 0, Urban was No and US was No), the sales would be estimately 13 thousands. We can trust this because t value > 2 and Pr(>|t|) is too small for the estimate to change a lot Price has a statistically significant effect on the sales too. Every increase by 1 dollar in price leades to reduction in total sales on 0.054. We can trust this because t value > 2 and Pr(>|t|) is too small for the Price to change a lot Urban is not statistically significant, it probably has no effect on the sales because t value < 2 and Pr(>|t|) is quite big, so it might vary US is statistically significant because t value > 2 and Pr(>|t|) is too small for the estimate to change a lot. As it is linear regression, it is not really suitable for categorical variables, but the location in US improves sales by 1.2
2. Sales = A + B1*Price + B2*Urban + B3\*US Sales =13.0435+(−0.05446)Price +(−0.02192)Urban +1.20057US If Urban and US are 1 (those dummy variables R created), those are substracted/added
3. the Null Hypothesis is that there is no relationship between this predictor and the y. In order to reject/fail to reject those we need to do:

contrasts(factor(carseats$ShelveLoc))

## Good Medium  
## Bad 0 0  
## Good 1 0  
## Medium 0 1

summary(lm(carseats$Sales ~ ., data = carseats))

##   
## Call:  
## lm(formula = carseats$Sales ~ ., data = carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8409 -0.6817 0.0127 0.6468 3.4684   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.7285968 0.6110801 9.375 < 2e-16 \*\*\*  
## X -0.0003284 0.0004538 -0.724 0.470   
## CompPrice 0.0930031 0.0041583 22.366 < 2e-16 \*\*\*  
## Income 0.0156505 0.0018582 8.422 7.3e-16 \*\*\*  
## Advertising 0.1238581 0.0111803 11.078 < 2e-16 \*\*\*  
## Population 0.0002157 0.0003708 0.582 0.561   
## Price -0.0953564 0.0026727 -35.678 < 2e-16 \*\*\*  
## ShelveLocGood 4.8520250 0.1532252 31.666 < 2e-16 \*\*\*  
## ShelveLocMedium 1.9579029 0.1261938 15.515 < 2e-16 \*\*\*  
## Age -0.0461835 0.0031894 -14.480 < 2e-16 \*\*\*  
## Education -0.0224532 0.0198208 -1.133 0.258   
## UrbanYes 0.1278481 0.1132532 1.129 0.260   
## USYes -0.1853717 0.1499447 -1.236 0.217   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.02 on 387 degrees of freedom  
## Multiple R-squared: 0.8736, Adjusted R-squared: 0.8697   
## F-statistic: 222.9 on 12 and 387 DF, p-value: < 2.2e-16

so from the results we have, we can make several conclusions: those with a large t value (>2) are statistically significant, therefore for them we can reject the null hypothesis. Predictors such as CompPrice, Income, Advertising, Price, ShelveLocGood, ShelveLocMedium and Age actully influence the overall Sales.

For other predictors (Population, Education, UrbanYes, USYes) we fail to reject the null hypothesis and we don’t whether they do significantly affect the sales.

summary(lm(carseats$Sales ~ carseats$CompPrice + carseats$Income + carseats$Advertising + carseats$Price + carseats$ShelveLoc + carseats$Age))

##   
## Call:  
## lm(formula = carseats$Sales ~ carseats$CompPrice + carseats$Income +   
## carseats$Advertising + carseats$Price + carseats$ShelveLoc +   
## carseats$Age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7728 -0.6954 0.0282 0.6732 3.3292   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.475226 0.505005 10.84 <2e-16 \*\*\*  
## carseats$CompPrice 0.092571 0.004123 22.45 <2e-16 \*\*\*  
## carseats$Income 0.015785 0.001838 8.59 <2e-16 \*\*\*  
## carseats$Advertising 0.115903 0.007724 15.01 <2e-16 \*\*\*  
## carseats$Price -0.095319 0.002670 -35.70 <2e-16 \*\*\*  
## carseats$ShelveLocGood 4.835675 0.152499 31.71 <2e-16 \*\*\*  
## carseats$ShelveLocMedium 1.951993 0.125375 15.57 <2e-16 \*\*\*  
## carseats$Age -0.046128 0.003177 -14.52 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.019 on 392 degrees of freedom  
## Multiple R-squared: 0.872, Adjusted R-squared: 0.8697   
## F-statistic: 381.4 on 7 and 392 DF, p-value: < 2.2e-16

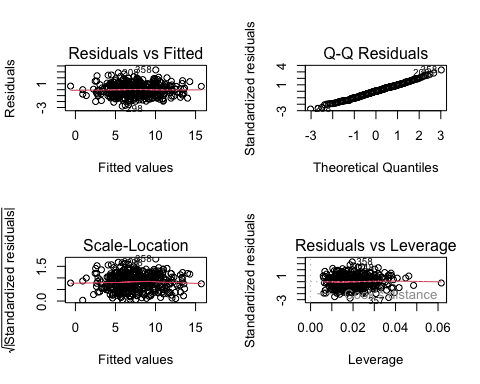
1. If we compare the Multiple R-squared and Adjusted R-squared from a and from e: In a only 23.35-23.93 % of sales variance is explained by those predictors, but in e 86.97-87.2 % of sales variance is explained by predictors. The E model fits data better.

confint(lm(carseats$Sales ~ carseats$CompPrice + carseats$Income + carseats$Advertising + carseats$Price + carseats$ShelveLoc + carseats$Age), level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 4.48236820 6.46808427  
## carseats$CompPrice 0.08446498 0.10067795  
## carseats$Income 0.01217210 0.01939784  
## carseats$Advertising 0.10071856 0.13108825  
## carseats$Price -0.10056844 -0.09006946  
## carseats$ShelveLocGood 4.53585700 5.13549250  
## carseats$ShelveLocMedium 1.70550103 2.19848429  
## carseats$Age -0.05237301 -0.03988204

we can calculate those using formula a +- t value \* std, but there is a built-in function in r for this. From this we can say that if there’s no influence of predictors (each continuous is equal to 0 and categorical are no/bad), the sales in 95% of cases will be something between 4.48 and 6.47. Also, in 95% of cases: One increase at CompPrice will increase the sales by 0.084-0.1 One dollar increase at Income will increase the sales by 0.012 - 0.019 One dollar increase at Advertising will increase the sales by 0.1-0.13  
One dollar increase at Price will decrease the sales by 0.09-0.1 Good ShelveLoc will increase the sales by 4.536 - 5.135 Medium ShelveLoc will increase the sales by 1.706-2.198 One increase of age will decrease the sales by 0.039-0.052

par(mfrow = c(2, 2))  
  
plot(lm(carseats$Sales ~ carseats$CompPrice + carseats$Income + carseats$Advertising + carseats$Price + carseats$ShelveLoc + carseats$Age))



par(mfrow = c(1,1))

model\_e <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age,  
 data = carseats)  
  
r\_stud <- rstudent(model\_e)  
h <- hatvalues(model\_e)  
  
n <- nrow(carseats)  
p <- length(coef(model\_e)) - 1  
  
outl\_cut <- 2 # |r\_stud| > 2  
lev\_cut <- 2\*(p+1)/n # h > 2(p+1)/n  
  
flags <- which(  
 abs(r\_stud) > outl\_cut |  
 h > lev\_cut   
)  
  
flags

## 1 16 35 43 76 101 166 172 175 208 248 285 298 306 311 353 357 358 366 376   
## 1 16 35 43 76 101 166 172 175 208 248 285 298 306 311 353 357 358 366 376

data.frame(  
 Obs = flags,  
 Rstud = r\_stud[flags],  
 Leverage = h[flags]  
)

## Obs Rstud Leverage  
## 1 1 2.1832477 0.01508106  
## 16 16 2.6383858 0.02033736  
## 35 35 -2.1002723 0.01373969  
## 43 43 -0.6233534 0.04849447  
## 76 76 0.9445543 0.04625617  
## 101 101 -2.5152208 0.01133121  
## 166 166 -0.9907049 0.04166477  
## 172 172 2.0925478 0.03026818  
## 175 175 0.6063223 0.04480914  
## 208 208 2.8389720 0.01860086  
## 248 248 2.3467108 0.03185608  
## 285 285 2.4743676 0.02403057  
## 298 298 -2.7688874 0.01862259  
## 306 306 0.5905636 0.04081196  
## 311 311 -0.2192045 0.06154635  
## 353 353 2.0002368 0.01966238  
## 357 357 -2.6621881 0.03677450  
## 358 358 3.3407504 0.01963874  
## 366 366 2.4402064 0.02244374  
## 376 376 2.2304099 0.01043779

from this we can see that there are some outliers and hatvalues.

R LAB 3.6.1

library(MASS)  
library(ISLR)

3.6.2

names(Boston)

## [1] "crim" "zn" "indus" "chas" "nox" "rm" "age"   
## [8] "dis" "rad" "tax" "ptratio" "black" "lstat" "medv"

attach(Boston)  
model = lm(medv~lstat)  
summary(model)

##   
## Call:  
## lm(formula = medv ~ lstat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.168 -3.990 -1.318 2.034 24.500   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.55384 0.56263 61.41 <2e-16 \*\*\*  
## lstat -0.95005 0.03873 -24.53 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.216 on 504 degrees of freedom  
## Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432   
## F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

coef(model)

## (Intercept) lstat   
## 34.5538409 -0.9500494

confint(model)

## 2.5 % 97.5 %  
## (Intercept) 33.448457 35.6592247  
## lstat -1.026148 -0.8739505

predict(model, data.frame(lstat=c(5, 10 ,15)),  
 interval ="confidence")

## fit lwr upr  
## 1 29.80359 29.00741 30.59978  
## 2 25.05335 24.47413 25.63256  
## 3 20.30310 19.73159 20.87461

by default leve is 95%; The “predict” returns the three columns; fit: the point estimate y lwr: lower bound of the 95% for the mean lstat. upr: upper bound of the 95%.

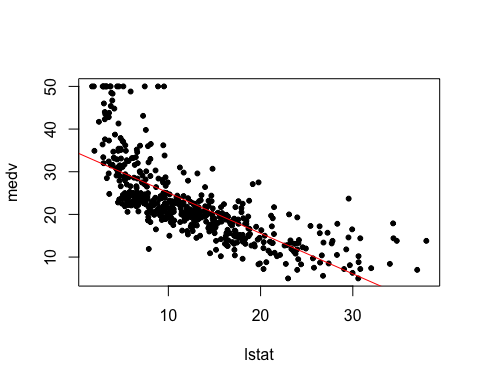
predict(model, data.frame(lstat=c(5, 10 ,15)),  
 interval ="prediction")

## fit lwr upr  
## 1 29.80359 17.565675 42.04151  
## 2 25.05335 12.827626 37.27907  
## 3 20.30310 8.077742 32.52846

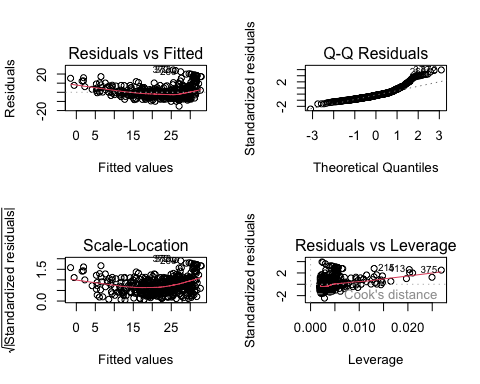
by default level is 95%; The “predict” returns the three columns; fit: the point estimate y lwr: lower bound of the 95% for the mean lstat. upr: upper bound of the 95%.

So the mean at both “confidence” and at “prediction” is the same (~25), however range is different. 24.47 - 25.63 at confidence and 12.83-37.28 at prediction

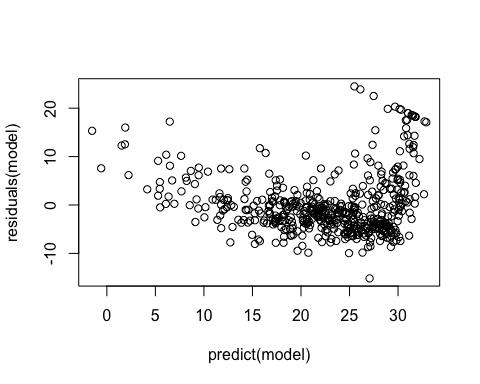
plot(lstat ,medv ,pch = 20)  
abline(model, col ="red")



par(mfrow = c(2,2))  
plot(model)



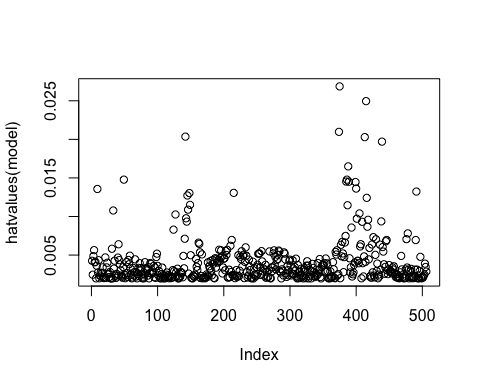
plot(predict (model), residuals (model))

 Residuals tend to be positive for smaller fitted values and negative for larger ones, so it is another evidence of non-linearity.

More points on the right

A few extreme points (>15) appear as outliers.

plot(hatvalues (model))

 points around 400 are extreme hat values which can influence the overall regression pattern.

which.max(hatvalues (model))

## 375   
## 375

3.6.3 Multiple Linear Regression

lm.fit = lm(medv ~ lstat+age, data = Boston)   
summary(lm.fit)

##   
## Call:  
## lm(formula = medv ~ lstat + age, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.981 -3.978 -1.283 1.968 23.158   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.22276 0.73085 45.458 < 2e-16 \*\*\*  
## lstat -1.03207 0.04819 -21.416 < 2e-16 \*\*\*  
## age 0.03454 0.01223 2.826 0.00491 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.173 on 503 degrees of freedom  
## Multiple R-squared: 0.5513, Adjusted R-squared: 0.5495   
## F-statistic: 309 on 2 and 503 DF, p-value: < 2.2e-16

lm.fit=lm(medv~.,data=Boston)  
summary (lm.fit)

##   
## Call:  
## lm(formula = medv ~ ., data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.595 -2.730 -0.518 1.777 26.199   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.646e+01 5.103e+00 7.144 3.28e-12 \*\*\*  
## crim -1.080e-01 3.286e-02 -3.287 0.001087 \*\*   
## zn 4.642e-02 1.373e-02 3.382 0.000778 \*\*\*  
## indus 2.056e-02 6.150e-02 0.334 0.738288   
## chas 2.687e+00 8.616e-01 3.118 0.001925 \*\*   
## nox -1.777e+01 3.820e+00 -4.651 4.25e-06 \*\*\*  
## rm 3.810e+00 4.179e-01 9.116 < 2e-16 \*\*\*  
## age 6.922e-04 1.321e-02 0.052 0.958229   
## dis -1.476e+00 1.995e-01 -7.398 6.01e-13 \*\*\*  
## rad 3.060e-01 6.635e-02 4.613 5.07e-06 \*\*\*  
## tax -1.233e-02 3.760e-03 -3.280 0.001112 \*\*   
## ptratio -9.527e-01 1.308e-01 -7.283 1.31e-12 \*\*\*  
## black 9.312e-03 2.686e-03 3.467 0.000573 \*\*\*  
## lstat -5.248e-01 5.072e-02 -10.347 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.745 on 492 degrees of freedom  
## Multiple R-squared: 0.7406, Adjusted R-squared: 0.7338   
## F-statistic: 108.1 on 13 and 492 DF, p-value: < 2.2e-16

summary(lm.fit)$r.sq

## [1] 0.7406427

summary(lm.fit)$sigma

## [1] 4.745298

library(car)

## Loading required package: carData

vif(lm.fit)

## crim zn indus chas nox rm age dis   
## 1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945   
## rad tax ptratio black lstat   
## 7.484496 9.008554 1.799084 1.348521 2.941491

lm.fit1=lm(medv~.-age ,data=Boston )  
summary (lm.fit1)

##   
## Call:  
## lm(formula = medv ~ . - age, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.6054 -2.7313 -0.5188 1.7601 26.2243   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.436927 5.080119 7.172 2.72e-12 \*\*\*  
## crim -0.108006 0.032832 -3.290 0.001075 \*\*   
## zn 0.046334 0.013613 3.404 0.000719 \*\*\*  
## indus 0.020562 0.061433 0.335 0.737989   
## chas 2.689026 0.859598 3.128 0.001863 \*\*   
## nox -17.713540 3.679308 -4.814 1.97e-06 \*\*\*  
## rm 3.814394 0.408480 9.338 < 2e-16 \*\*\*  
## dis -1.478612 0.190611 -7.757 5.03e-14 \*\*\*  
## rad 0.305786 0.066089 4.627 4.75e-06 \*\*\*  
## tax -0.012329 0.003755 -3.283 0.001099 \*\*   
## ptratio -0.952211 0.130294 -7.308 1.10e-12 \*\*\*  
## black 0.009321 0.002678 3.481 0.000544 \*\*\*  
## lstat -0.523852 0.047625 -10.999 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.74 on 493 degrees of freedom  
## Multiple R-squared: 0.7406, Adjusted R-squared: 0.7343   
## F-statistic: 117.3 on 12 and 493 DF, p-value: < 2.2e-16

3.6.4 Interaction Terms

summary (lm(medv~lstat\*age ,data=Boston))

##   
## Call:  
## lm(formula = medv ~ lstat \* age, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.806 -4.045 -1.333 2.085 27.552   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.0885359 1.4698355 24.553 < 2e-16 \*\*\*  
## lstat -1.3921168 0.1674555 -8.313 8.78e-16 \*\*\*  
## age -0.0007209 0.0198792 -0.036 0.9711   
## lstat:age 0.0041560 0.0018518 2.244 0.0252 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.149 on 502 degrees of freedom  
## Multiple R-squared: 0.5557, Adjusted R-squared: 0.5531   
## F-statistic: 209.3 on 3 and 502 DF, p-value: < 2.2e-16

we see that this interaction is actually not helping, but we can find better interactions.

summary (lm(medv~lstat\*rm ,data=Boston))

##   
## Call:  
## lm(formula = medv ~ lstat \* rm, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.2349 -2.6897 -0.6158 1.9663 31.6141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -29.12452 3.34250 -8.713 <2e-16 \*\*\*  
## lstat 2.19398 0.20570 10.666 <2e-16 \*\*\*  
## rm 9.70126 0.50023 19.393 <2e-16 \*\*\*  
## lstat:rm -0.48494 0.03459 -14.018 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.701 on 502 degrees of freedom  
## Multiple R-squared: 0.7402, Adjusted R-squared: 0.7387   
## F-statistic: 476.9 on 3 and 502 DF, p-value: < 2.2e-16