STAT 151A Homework 1 Question 4 Partial Solutions

Billy Fang

The following is a very rough demonstration of how to do certain parts of number four, and is not meant to be an example of a student's proper submission.

Reading the data

```
autompg <- read.table("auto-mpg.data.txt")
colnames(autompg) <- c("mpg", "cyl", "disp", "hp", "wt", "acc", "year", "orig", "name")

# remove rows with missing horsepower
missing <- which(autompg$hp == "?")
autompg <- autompg[-missing,]
```

```
# see the class (numeric, factor, etc.) of each column
sapply(autompg, class)
```

```
##       mpg       cyl      disp        hp        wt       acc      year
## "numeric" "integer" "numeric"  "factor" "numeric" "numeric" "integer"
##      orig      name
## "integer"  "factor"
```

```
# check what the values/levels look like
unique(autompg$cyl)
levels(autompg$hp)
unique(autompg$year)
unique(autompg$orig)
```

(a)

Omitted.

(b) - (c), Model 1

I will consider two models. In both cases I will treat horespower as a quantitative variable rather than a categorical variable, and I will treat origin as a categorical variable rather than an quantitative one.

In this first model, I will treat model year and cylinders as numeric.

```
autompg$hp <- as.numeric(as.character(autompg$hp))
autompg$orig <- as.factor(autompg$orig)
```

```
y <- autompg$mpg
X <- autompg[, c("disp", "hp", "wt", "acc", "year", "cyl")]
X <- cbind(intercept=1, X)

orig.levels <- levels(autompg$orig)
```

```r
orig.dum <- numeric()
for (i in 2:length(orig.levels)) {
  orig.dum <- cbind(orig.dum, autompg$orig == orig.levels[i])
  colnames(orig.dum)[i-1] <- paste0("orig", orig.levels[i])
}
X <- cbind(X, orig.dum)
X <- as.matrix(X)
```

```r
# compute least squares estimate manually and compare with lm()
beta.hat <- solve(t(X) %*% X, t(X) %*% y)
fit <- lm(mpg ~ disp + hp + wt + acc + year + cyl + orig, data=autompg)
t(beta.hat)
```

```
##      intercept       disp          hp           wt        acc      year
## [1,]  -17.9546 0.02397864 -0.01818346 -0.006710384 0.07910304 0.7770269
##             cyl    orig2    orig3
## [1,] -0.4897094 2.630002 2.853228
```

```r
coef(fit)
```

```
##   (Intercept)          disp            hp            wt           acc
## -17.954602067   0.023978644  -0.018183464  -0.006710384   0.079103036
##          year           cyl         orig2         orig3
##   0.777026939  -0.489709424   2.630002360   2.853228228
```

```r
# compute and check RSS
y.hat <- X %*% beta.hat
RSS <- sum((y.hat - y)^2)
RSS
```

```
## [1] 4187.392
```

```r
sum(residuals(fit)^2)
```

```
## [1] 4187.392
```

```r
# compute and check SSReg
y.bar <- mean(y)
SSReg <- sum((y.hat - y.bar)^2)
SSReg
```

```
## [1] 19631.6
```

```r
sum((fitted(fit) - y.bar)^2)
```

```
## [1] 19631.6
```

```r
# compute TSS and check RSS + SSReg = TSS
TSS <- sum((y - y.bar)^2)
RSS + SSReg
```

```
## [1] 23818.99
```

```r
TSS
```

```
## [1] 23818.99
```

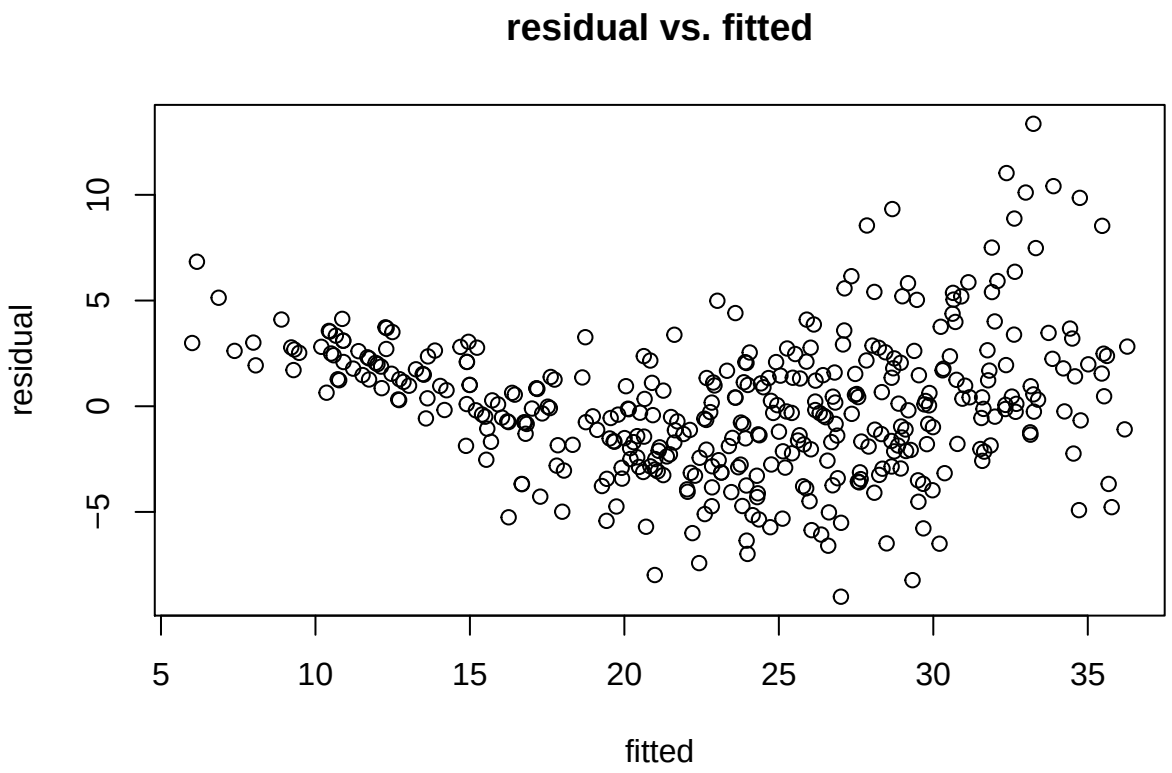
```r
# compute and check r^2
r2 <- SSReg/TSS
r2
```

```
## [1] 0.8241995
```

```
summary(fit)$r.squared
```

```
## [1] 0.8241995
```

```
plot(y.hat, y - y.hat, xlab="fitted", ylab="residual", main="residual vs. fitted")
```



(b) - (c), Model 2

Let us instead consider model year and cylinders as factors.

```
autompg$year <- factor(autompg$year)
autompg$cyl <- factor(autompg$cyl)

X <- autompg[, c("disp", "hp", "wt", "acc")]
X <- cbind(intercept=1, X)

year.levels <- levels(autompg$year)
year.dum <- numeric()
for (i in 2:length(year.levels)) {
  year.dum <- cbind(year.dum, autompg$year == year.levels[i])
  colnames(year.dum)[i-1] <- paste0("year", year.levels[i])
}

cyl.levels <- levels(autompg$cyl)
cyl.dum <- numeric()
for (i in 2:length(cyl.levels)) {
  cyl.dum <- cbind(cyl.dum, autompg$cyl == cyl.levels[i])
  colnames(cyl.dum)[i-1] <- paste0("cyl", cyl.levels[i])
```

```
}

# orig.dum was already constructed
X <- cbind(X, year.dum, cyl.dum, orig.dum)
X <- as.matrix(X)
```

Then we repeat the same calculations of RSS, etc. as before.

```
# compute least squares estimate manually and compare with lm()
beta.hat <- solve(t(X) %*% X, t(X) %*% y)
fit <- lm(mpg ~ disp + hp + wt + acc + year + cyl + orig, data=autompg)
t(beta.hat)
```

```
##      intercept       disp          hp           wt         acc    year71
## [1,]  30.91684 0.01182459 -0.03923228 -0.005180179 0.003607983 0.9104285
##          year72     year73   year74    year75  year76   year77   year78
## [1,] -0.4903062 -0.5528934 1.241998 0.8704016 1.49666 2.998697 2.973778
##        year79   year80   year81   year82     cyl4     cyl5     cyl6
## [1,] 4.896176 9.058932 6.458158 7.837585 6.939922 6.637731 4.297314
##          cyl8    orig2    orig3
## [1,] 6.366813 1.693285 2.292927
```

```
coef(fit)
```

```
##  (Intercept)         disp           hp           wt          acc
## 30.916841489  0.011824592 -0.039232282 -0.005180179  0.003607983
##       year71       year72       year73       year74       year75
##  0.910428513 -0.490306154 -0.552893391  1.241997594  0.870401578
##       year76       year77       year78       year79       year80
##  1.496659785  2.998696745  2.973778349  4.896176328  9.058931568
##       year81       year82         cyl4         cyl5         cyl6
##  6.458158033  7.837584958  6.939921560  6.637730992  4.297313906
##         cyl8        orig2        orig3
##  6.366812930  1.693285334  2.292926778
```

```
# compute and check RSS
y.hat <- X %*% beta.hat
RSS <- sum((y.hat - y)^2)
RSS
```

```
## [1] 2992.061
```

```
sum(residuals(fit)^2)
```

```
## [1] 2992.061
```

```
# compute and check SSReg
y.bar <- mean(y)
SSReg <- sum((y.hat - y.bar)^2)
SSReg
```

```
## [1] 20826.93
```

```
sum((fitted(fit) - y.bar)^2)
```

```
## [1] 20826.93
```

```
# compute TSS and check RSS + SSReg = TSS
TSS <- sum((y - y.bar)^2)
```

```
RSS + SSReg
```

```
## [1] 23818.99
```

```
TSS
```

```
## [1] 23818.99
```

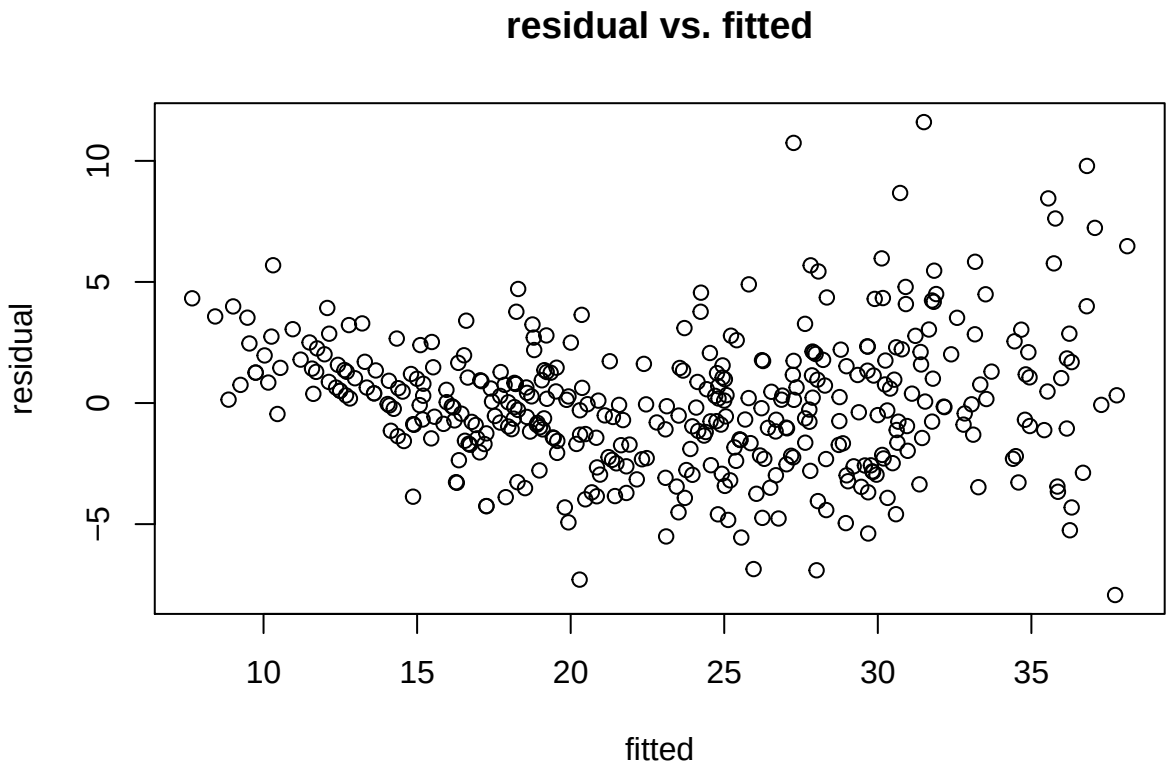
```
# compute and check r^2
r2 <- SSReg/TSS
r2
```

```
## [1] 0.8743834
```

```
summary(fit)$r.squared
```

```
## [1] 0.8743834
```

```
plot(y.hat, y - y.hat, xlab="fitted", ylab="residual", main="residual vs. fitted")
```



(d)

Omitted.