Regression Model - Final Project

Thanakrit Danny

February 1, 2559 BE

Excutive Summary

From data mtcars, there is signifiant of regression of **mpg** by **am** as regressor. With the Varience Inflation factor technique points out others uncorrelated regressors :- **drat**, **vs** and **gear**. Any way with cacading regressors analysis show that including **drat** and **vs** show significant in regression model.

Single Regression

```
library(ggplot2)
data("mtcars")
dt <- mtcars
#Test single-regression
fit1 <- lm(mpg ~ am, data = dt)
summary(fit1)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

As show in the coefficients summary the base level (**am** = 0; auto) average miles/gallon about 17.14 while the regression show that when **am** = 1 (noauto) average miles/gallon increase +7.244, confirm with the boxplot of both variables (pic 1). Also the residual plot show equally distributed between 0 of both **am** (pic 2).

Multiple regression

To make sure unbias of the model, include others varibles in multiple regression. Start with varience inflation factor.

```
library(car)
fit2 <- lm(mpg ~ ., data = dt)
vif(fit2)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

the vif of **am** = 4.64 while the vif of varibles :- **cyl**, **disp**, **hp**, **wt**, **qsec** and **carb** has far from the **am** means those has correlation with regressor '**am**'; then omit those varibles. Include only :- **drat**, **vs** and **gear** in our model.

Next is to verify the most significant model with cascading the varible and anova test.

```
fit1.1 <- update(fit1, mpg ~ am + drat)
fit1.2 <- update(fit1, mpg ~ am + drat + vs)
fit1.3 <- update(fit1, mpg ~ am + drat + vs + gear)
anova(fit1, fit1.1, fit1.2, fit1.3)
```

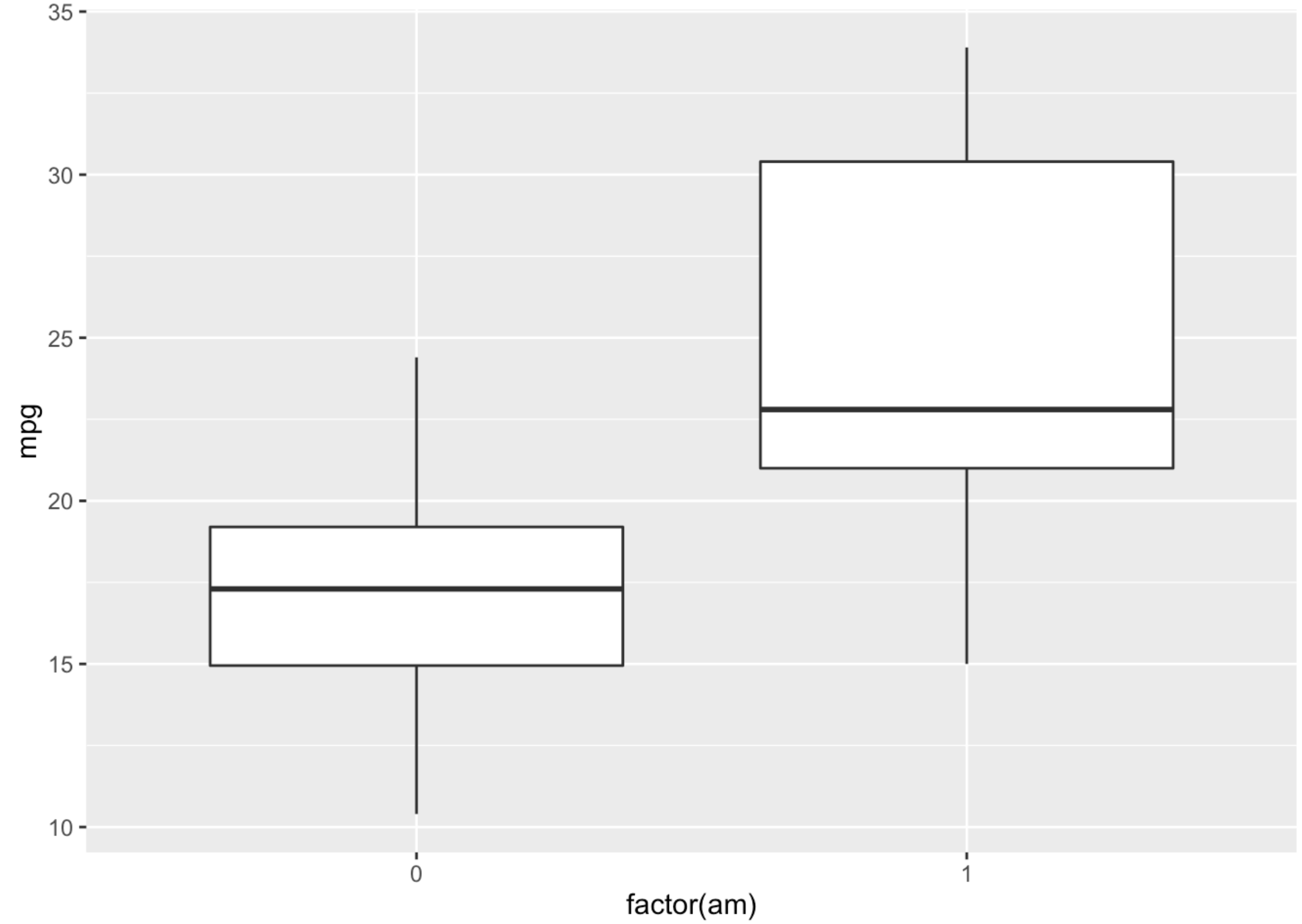
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + drat
## Model 3: mpg ~ am + drat + vs
## Model 4: mpg ~ am + drat + vs + gear
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 573.64  1    147.26 12.0889 0.0017333 **
## 3     28 339.99  1    233.65 19.1814 0.0001611 ***
## 4     27 328.89  1     11.10  0.9113 0.3482507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anyway, including the **gear** the significant of model < 95%, then we add only **drat** and **vs** in our model

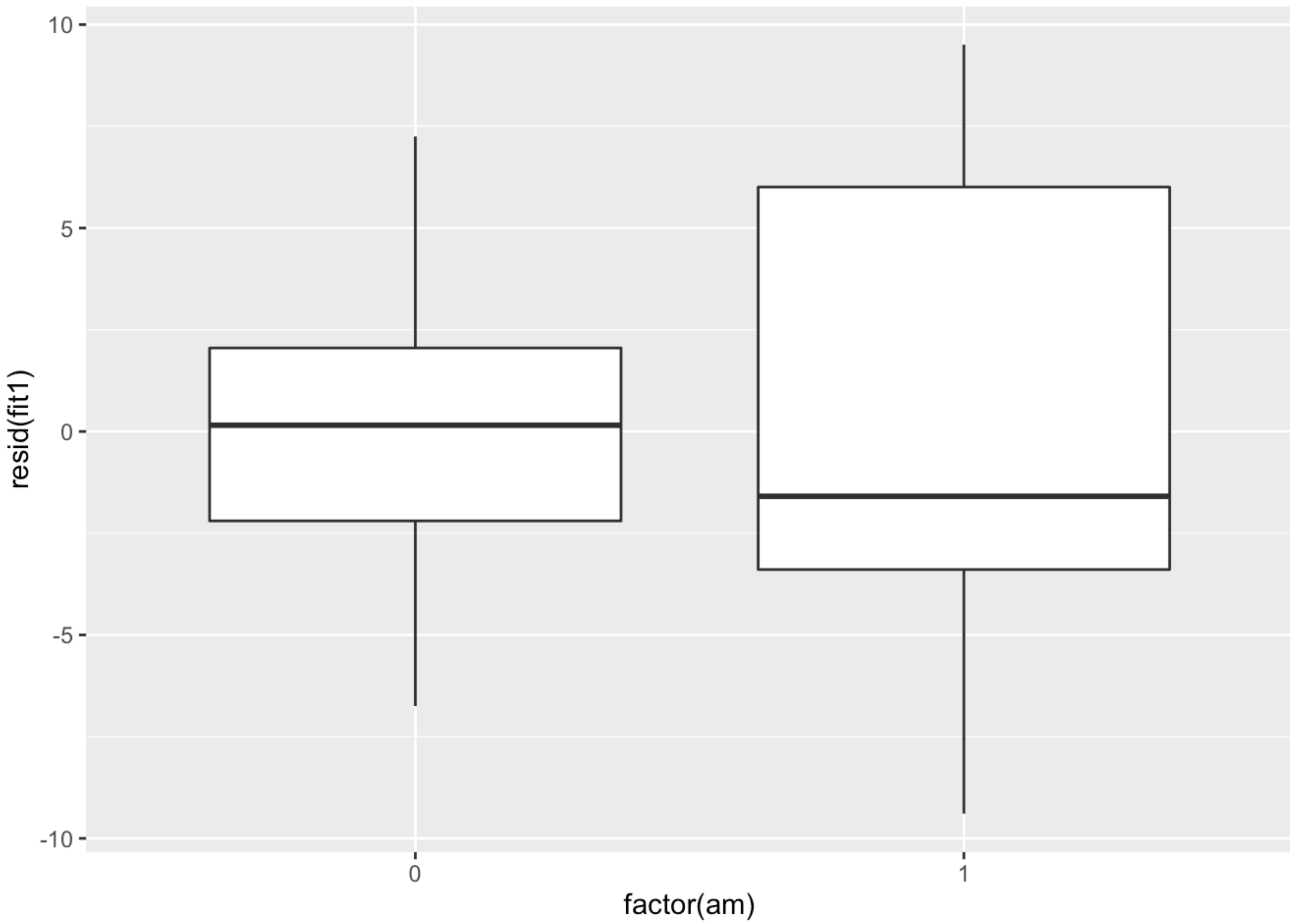
Re-confirm with the histogram of residual plot of fit1.2 model (pic 3), show close to normal distribution with mean = 0.

Appendix

pic 1 : Boxplot of mgp ~ am



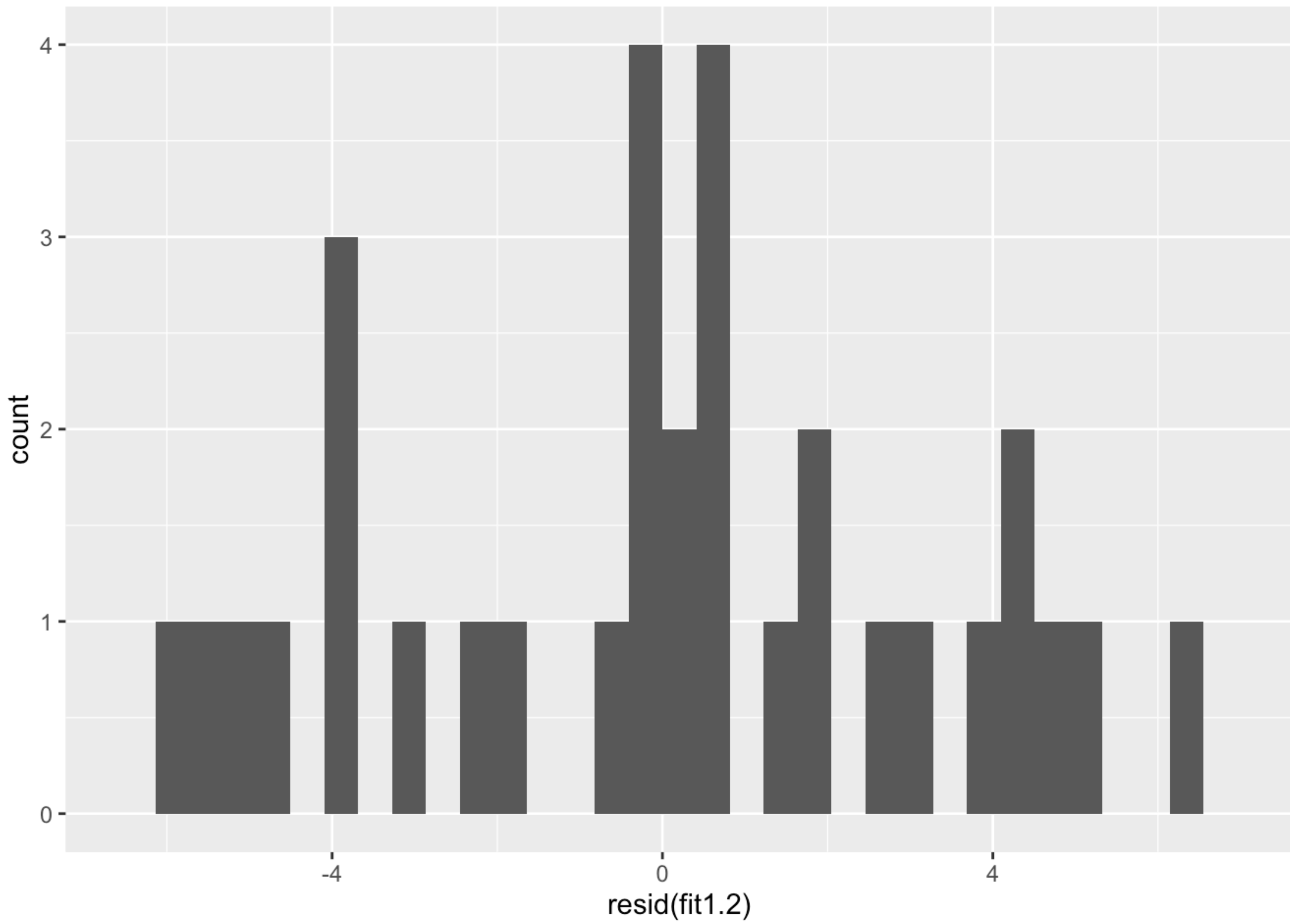
pic 2 : Boxplot residual of mgp ~ **am**



pic 3 : Histogram residual of fit1.2

```
i <- ggplot(data = mtcars) +
  geom_histogram(aes(resid(fit1.2)))
i
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



count
4
3
2
1
0
-4
0
4
resid(fit1.2)