# Model selection and confidence intervals

Homework week 7

*The Solution*

```
pacman::p_load(s20x, tidyverse, gvlma, AICcmodavg, gridExtra, xtable)
```

## Data preparation

### Identification

```
data(attitude)
str(attitude)
```

```
## 'data.frame':    30 obs. of  7 variables:
##  $ rating    : num  43 63 71 61 81 43 58 71 72 67 ...
##  $ complaints: num  51 64 70 63 78 55 67 75 82 61 ...
##  $ privileges: num  30 51 68 45 56 49 42 50 72 45 ...
##  $ learning  : num  39 54 69 47 66 44 56 55 67 47 ...
##  $ raises    : num  61 63 76 54 71 54 66 70 71 62 ...
##  $ critical  : num  92 73 86 84 83 49 68 66 83 80 ...
##  $ advance   : num  45 47 48 35 47 34 35 41 31 41 ...
```

### Assumptions

```
rat.dist.gg <-
  attitude %>%
    ggplot(aes(x=rating)) + theme_bw(16) +
      geom_density(alpha=.2, fill="#FF6666") +
      geom_histogram(aes(y=..density..),
                 binwidth=5,
                 colour="black",
                 fill="lightgreen") +
      labs(x="rating") +
      geom_line(data=data.frame(
                        X=seq(25, 100, 1),
                        Y=dnorm(x=seq(25, 100, 1),
                                mean=mean(attitude$rating),
                                sd=sd(attitude$rating))),
                aes(x=X, y=Y),
                    colour="blue", size=1.1)
rat.QQ.gg <-
  attitude %>%
    ggplot(aes(sample=rating)) + theme_bw(16) +
            stat_qq(size=4, bg="#43a2ca",
                    col="black", pch=21) +
            stat_qq_line(size=1.5, color="blue")
grid.arrange(rat.dist.gg, rat.QQ.gg, nrow = 1)
```

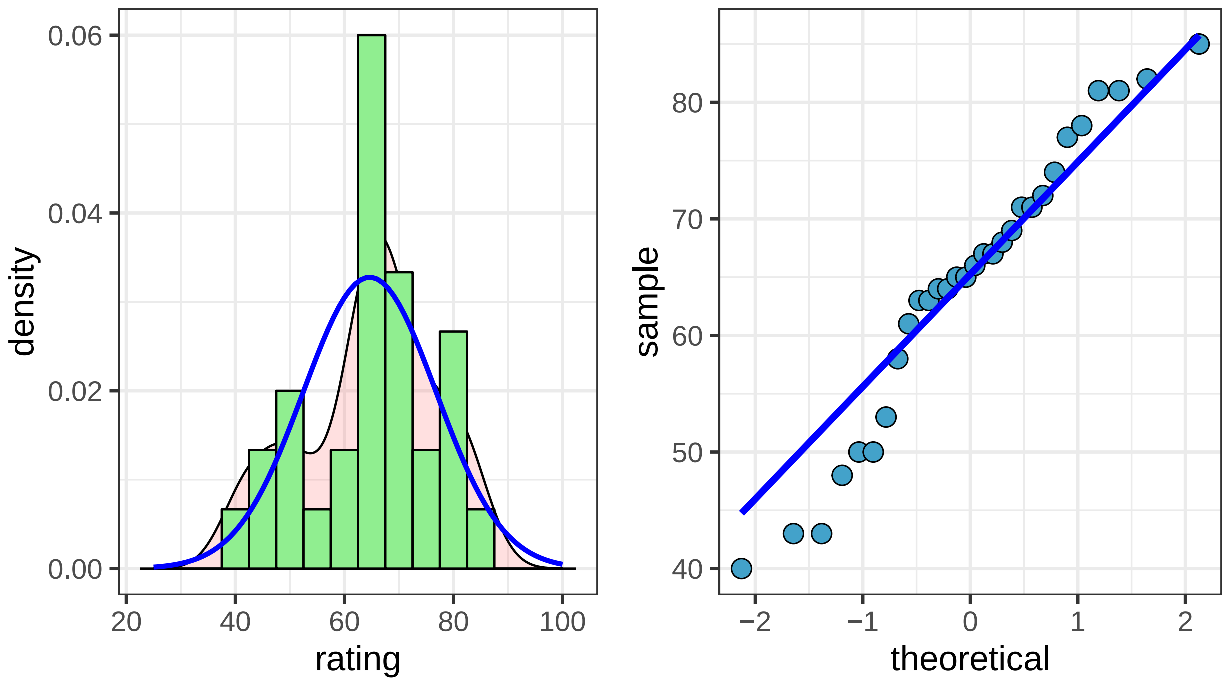


Figure 1: Distribution and Q-Q plot for response variable `rating`.

The data are sufficiently symetrical around the mean that a normal (Gaussian) distribution fits them well.

```
mod.sum <- summary(gvlma(lm(rating ~ complaints + privileges + learning,
                         data=attitude )))
```

```
xtable(mod.sum, caption="The full potential model meets assumptions of the linear model.",
       label="tab:gvlma") %>%
  print(comment=FALSE)
```

| | Value | p-value | Decision |
|---|---|---|---|
| Global Stat | 1.68 | 0.80 | Assumptions acceptable. |
| Skewness | 0.00 | 0.97 | Assumptions acceptable. |
| Kurtosis | 1.64 | 0.20 | Assumptions acceptable. |
| Link Function | 0.00 | 0.99 | Assumptions acceptable. |
| Heteroscedasticity | 0.03 | 0.85 | Assumptions acceptable. |

Table 1: The full potential model meets assumptions of the linear model.

## Model fitting and selection

```
car::vif(lm(rating ~ complaints + privileges + learning, data=attitude )) %>%
  t() %>%
  as.data.frame() %>%
  xtable(caption="Low Variable Inflation Factors for each potential
         predictor variables.") %>%
  print(comment=FALSE, include.rownames=FALSE)
```

| complaints | privileges | learning |
|---|---|---|
| 1.81 | 1.54 | 1.65 |

Table 2: Low Variable Inflation Factors for each potential predictor variables.

### Define model set

```
null <- lm(rating ~ 1, attitude)
C <- lm(rating ~ complaints, attitude)
P <- lm(rating ~ privileges, attitude)
L <- lm(rating ~ learning, attitude)
CP <- lm(rating ~ complaints + privileges, attitude)
CL <- lm(rating ~ complaints + learning, attitude)
PL <- lm(rating ~ privileges + learning, attitude)
CPL <- lm(rating ~ complaints + privileges + learning, attitude)

cand.mod.names <- c("null", "C", "P", "L", "CP", "CL", "PL", "CPL")
```

### Model selection

```
cand.mods <- list( )
for(i in 1:length(cand.mod.names)) {
  cand.mods[[i]] <- get(cand.mod.names[i]) }
library(AICcmodavg)
aictab(cand.set = cand.mods,
       modnames = cand.mod.names) %>%
  xtable(caption="Model rankings based on $AIC_{c}$ information criterion.
        Models identified by first letter of predictor variable names.\\label{MS}") %>%
    print(comment=FALSE, include.rownames=FALSE)
```

| Model | K | AICc | Delta AICc | AICc weight | log-Likelihood |
|---|---|---|---|---|---|
| C | 3.00 | 206.69 | 0.00 | 0.39 | -99.88 |
| CL | 4.00 | 206.74 | 0.05 | 0.38 | -98.57 |
| CPL | 5.00 | 208.91 | 2.22 | 0.13 | -98.21 |
| CP | 4.00 | 209.20 | 2.51 | 0.11 | -99.80 |
| L | 3.00 | 226.21 | 19.53 | 0.00 | -109.65 |
| PL | 4.00 | 227.97 | 21.28 | 0.00 | -109.18 |
| P | 3.00 | 234.98 | 28.30 | 0.00 | -114.03 |
| null | 2.00 | 238.51 | 31.83 | 0.00 | -117.04 |

Table 3: Model rankings based on $AIC_c$ information criterion. Models identified by first letter of predictor variable names.

All models that include `complaints` can be considered competitive in $AIC_c$-based model selection (Table 3).

### Model averaging

```
terms <- c("(Intercept)", "complaints", "learning", "privileges")
        av.params <- as.data.frame(array(NA,c(length(terms),4)))
        colnames(av.params)<-c("term","estimate","ciL","ciU")
```

```
      for(i in 1:length(terms)) {
       av <- modavg(parm = paste(terms[i]),
                    cand.set = cand.mods,
                    modnames = cand.mod.names)
           av.params[i,1] <- terms[i]
           av.params[i,2] <- round(av$Mod.avg.beta, 2)
           av.params[i,3] <- round(av$Lower.CL, 3)
           av.params[i,4] <- round(av$Upper.CL, 3) }
av.params %>%
  xtable(caption="Averaged regression coefficients on top-ranked models
         from $AIC_{c}$-based model selection (Table \\ref{MS}).
         \\label{MA}") %>%
      print(comment=FALSE, include.rownames=FALSE)
```

| term | estimate | ciL | ciU |
|---|---|---|---|
| (Intercept) | 12.39 | -1.90 | 26.68 |
| complaints | 0.71 | 0.46 | 0.95 |
| learning | 0.22 | -0.05 | 0.48 |
| privileges | -0.08 | -0.34 | 0.18 |

Table 4: Averaged regression coefficients on top-ranked models from $AIC_c$-based model selection (Table 3).

As the slope term of the regression equation denotes the strength of the modelled relationship, the estimated coefficient for each slope term in a regression model can be interpreted as a measure of that term's relative importance to the response variable, or *effect size*.

### Plot confidence intervals

```
av.params %>%
  filter(terms != "(Intercept)") %>%
    ggplot() + theme_bw(16) +
      coord_flip() +
      geom_hline(yintercept = 0) +
      geom_errorbar(aes(x=term,
                        ymin=ciL,
                        ymax=ciU),
                        width=0.1, size=1,
                    color="#377eb8") +
      geom_point(aes(x=term,
                     y=estimate),
                 size=4, pch=21, stroke=2,
                 bg="#377eb8", color="white")
```

## Conclusions

$AIC_c$-based model selection indicated that three variables, `privileges`, `learning`, and `complaints` were associated with employee `ratings` of job satisfaction (Table 3). In comparing model-averaged regression coefficients (Fig. 2), only `complaints` had a non-zero relationship with `rating`, which was positive. 95% CIs for `learning` and `privileges` overlapped zero; these terms had positive and negative trends with `rating`, respectively. Thus, employee ratings appear to be most strongly determined by how well eomployees felt the company handled employee complaints.

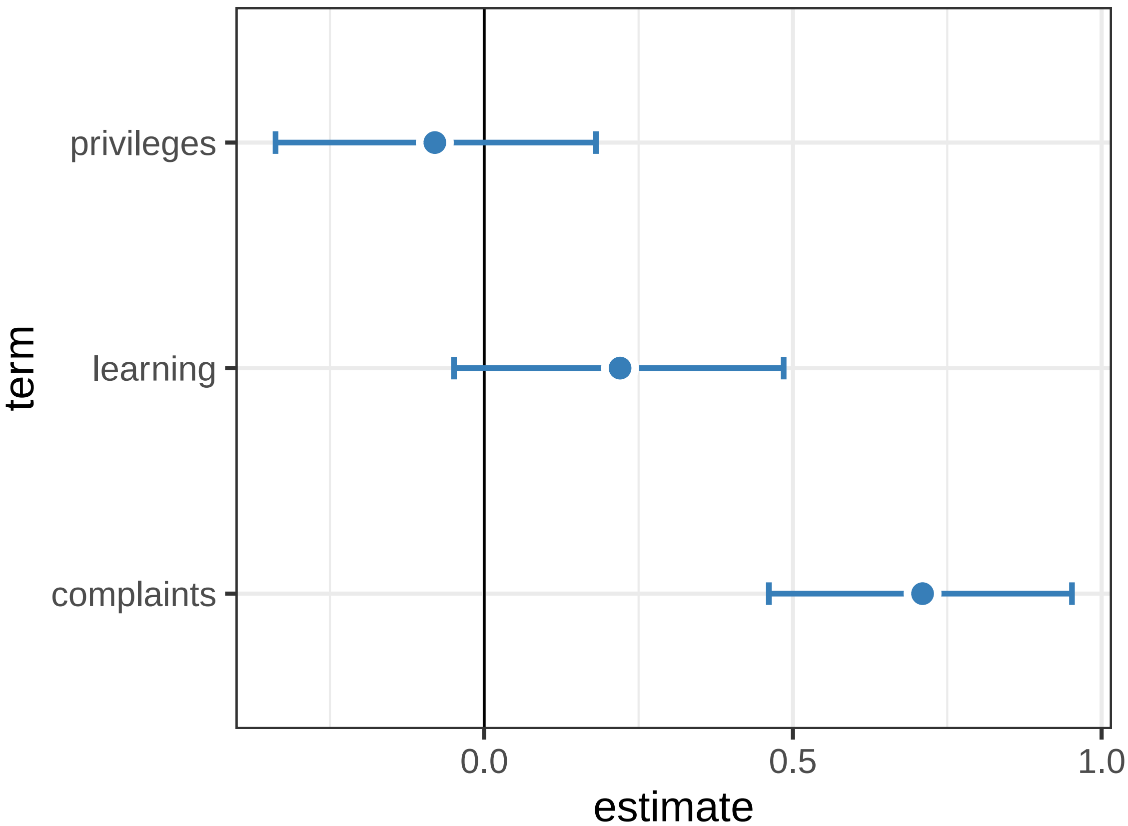


Figure 2: Model-averaged 95% confidence intervals with regression coefficient estimates for terms in top-ranked models (Table 3).