Chapter 9

Discussion of Response Models with Factors and Predictors

Contents

We have considered simple models for designed experiments and observational studies where a response variable is modeled as a linear combination of effects due to **factors** or **predictors**, or both. With designed experiments, where only qualitative factors are considered, we get a "pure ANOVA" model. For example, in the experiment comparing survival times of beetles, the potential effects of **insecticide** (with levels A, B, C, and D) and **dose** (with levels 1=low, 2=medium, and 3=high) are included in the model as **factors** because these variables are qualitative. The natural model to consider is a two-way

ANOVA with effects for dose and insecticide and a dose-by-insecticide interaction. If, however, the dose given to each beetle was recorded on a measurement scale, then the dosages can be used to define a predictor variable which can be used as a "regression effect" in a model. That is, the dose or some function of dose can be used as a (quantitative) predictor instead of as a qualitative effect.

For simplicity, assume that the doses are 10, 20, and 30, but the actual levels are irrelevant to the discussion. The simple additive model, or ANCOVA model, assumes that there is a linear relationship between mean survival time and dose, with different intercepts for the four insecticides. If data set includes the survival time (times) for each beetle, the insecticide (insect: an alphanumeric variable, with values A, B, C, and D), and dose, you would fit the ANCOVA model this way

```
dat_beetles <-
  dat_beetles %>%
  mutate(
    insect = factor(insect)
  )
lm_t_i_d <-
  lm(
    times ~ insect + dose
  , data = dat_beetles
  )
```

A more complex model that allows separate regression lines for each insecticide is specified as follows:

```
lm_t_i_d_id <-
  lm(
    times ~ insect + dose + insect:dose
  , data = dat_beetles
  )
```

It is important to recognize that the `factor()` statement defines which variables in the model are treated as **factors**. Each effect of the `factor` data type is treated as a factor. Effects in the model statement that are numeric data types are treated as **predictors**. To treat a measurement variable as a factor (with one level for each distinct observed value of the variable) instead of a predictor, convert that varible type to a factor using `factor()`. Thus, in the survival time experiment, these models

```
dat_beetles <-
  dat_beetles %>%
  mutate(
    insect = factor(insect)
  , dose   = factor(dose)
  )
```

```
# call this (A) for additive
lm_t_i_d <-
  lm(
    times ~ insect + dose
  , data = dat_beetles
  )
# call this (I) for interaction
lm_t_i_d_id <-
  lm(
    times ~ insect + dose + insect:dose
  , data = dat_beetles
  )
```

give the analysis for a two-way ANOVA model without interaction and with interaction, respectively, where both dose and insecticide are treated as factors (since dose and insect are both converted to factors), even though we just defined dose on a measurement scale!

Is there a basic connection between the ANCOVA and separate regression line models for dose and two-way ANOVA models where dose and insecticide are treated as factors? Yes — I mentioned a connection when discussing ANCOVA and I will try now to make the connection more explicit.

For the moment, let us simplify the discussion and assume that only one insecticide was used at three dose levels. The LS estimates of the mean responses from the quadratic model

$$\text{Times} = \beta_0 + \beta_1 \text{ Dose} + \beta_2 \text{ Dose}^2 + \varepsilon$$

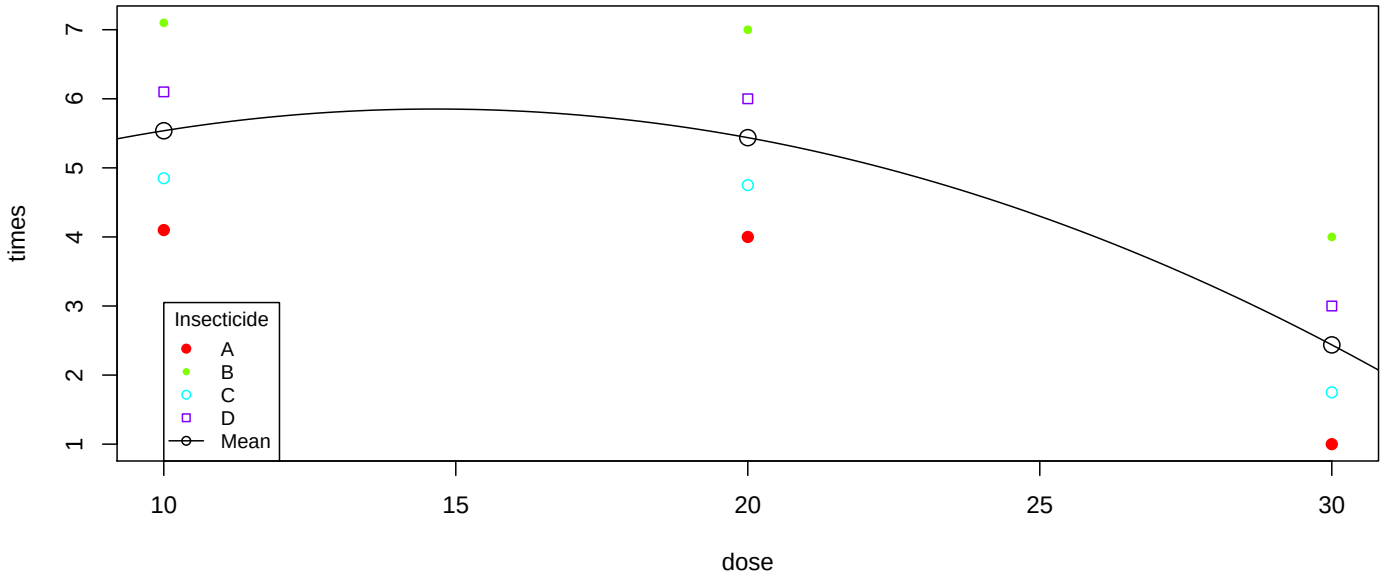
are the observed average survival times at the three dose levels. The LS curve goes through the mean survival time at each dose, as illustrated in the picture below.

If we treat dose as a **factor**, and fit the one-way ANOVA model

$$\text{Times} = \text{Grand Mean} + \text{Dose Effect} + \text{Residual},$$

then the LS estimates of the population mean survival times are the observed mean survival times. The two models are mathematically equivalent, but the parameters have different interpretations. In essence, the one-way ANOVA model places no restrictions on the values of the population means (no *a priori* relation between them) at the three doses, and neither does the quadratic model!

(WHY?)



In a one-way ANOVA, the standard hypothesis of interest is that the dose effects are zero. This can be tested using the one-way ANOVA F-test, or by testing $H_0 : \beta_1 = \beta_2 = 0$ in the quadratic model. With three dosages, the absence of a linear or quadratic effect implies that all the population mean survival times must be equal. An advantage of the polynomial model over the one-way ANOVA is that it provides an easy way to **quantify** how dose impacts the mean survival, and a convenient way to check whether a simple description such as a simple linear regression model is adequate to describe the effect.

More generally, if dose has p levels, then the one-way ANOVA model

$$\text{Times} = \text{Grand Mean} + \text{Dose Effect} + \text{Residual},$$

is equivalent to the $(p-1)^{st}$ degree polynomial

$$\text{Times} = \beta_0 + \beta_1 \text{ Dose} + \beta_2 \text{ Dose}^2 + \cdots + \beta_{p-1} \text{ Dose}^{(p-1)} + \varepsilon$$

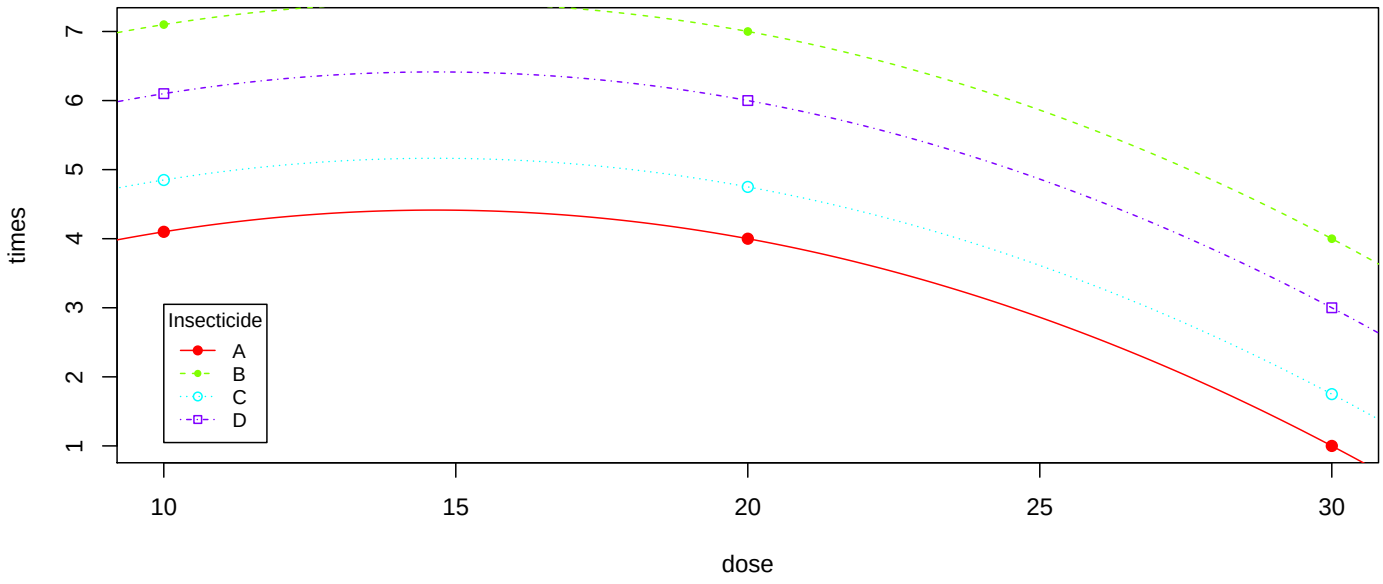
and the one-way ANOVA F-test for no treatment effects is equivalent to testing $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ in this polynomial.

Returning to the original experiment with 4 insecticides and 3 doses, I can show the following two equivalences. First, the two-way additive ANOVA model, with insecticide and dose as factors, i.e., model (A), is mathematically

equivalent to an additive model with insecticide as a factor, and a quadratic effect in dose:

```
dat_beetles <-
  dat_beetles %>%
  mutate(
    insect = factor(insect)
  )
lm_t_i_d_d2 <-
  lm(
    times ~ insect + dose + I(dose^2)
  , data = dat_beetles
  )
```

Thinking of dose^2 as a quadratic term in dose, rather than as an interaction, this model has an additive insecticide effect, but the dose effect is not differentiated across insecticides. That is, the model assumes that the quadratic curves for the four insecticides differ only in level (i.e., different intercepts) and that the coefficients for the dose and dose^2 effects are identical across insecticides. This is an additive model, because the population means plot has parallel profiles. A possible pictorial representation of this model is given below.



Second, the two-way ANOVA interaction model, with insecticide and dose as factors, i.e., model (I), is mathematically equivalent to an interaction model with insecticide as a factor, and a quadratic effect in dose.

```
dat_beetles <-
  dat_beetles %>%
  mutate(
    insect = factor(insect)
  )
lm_t_i_d_d2_id_id2 <-
  lm(
    times ~ insect + dose + I(dose^2) + insect:dose + insect:I(dose^2)
  , data = dat_beetles
  )
```

This model fits separate quadratic relationships for each of the four insecticides, by including interactions between insecticides and the linear and quadratic terms in dose. Because dose has three levels, this model places no restrictions on the mean responses.

To summarize, we have established that

- The additive two-way ANOVA model with insecticide and dose as factors is mathematically identical to an additive model with an insecticide factor and a quadratic effect in dose. The ANCOVA model with a linear effect in dose is a special case of these models, where the quadratic effect is omitted.
- The two-way ANOVA interaction model with insecticide and dose as factors is mathematically identical to a model with an insecticide factor, a quadratic effect in dose, and interactions between the insecticide and the linear and quadratic dose effects. The separate regression lines model with a linear effect in dose is a special case of these models, where the quadratic dose effect and the interaction of the quadratic term with insecticide are omitted.

Recall that response models with **factors** and **predictors** as effects can be fit using the `lm()` procedure, but each factor or interaction involving a factor must be represented in the model using indicator variables or product terms. The number of required indicators or product effects is one less than the number of distinct levels of the factor. For example, to fit the model with "parallel" quadratic curves in dose, you can define (in the `data.frame()`) three indicator

variables for the insecticide effect, say I_1, I_2, and I_3, and fit the model

$$\text{Times} = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 \text{ Dose} + \beta_5 \text{ Dose}^2 + \varepsilon.$$

For the "quadratic interaction model", you must define 6 interaction or product terms between the 3 indicators and the 2 dose terms:

$$\begin{aligned} \text{Times} \;=\; & \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 \text{ Dose} + \beta_5 \text{ Dose}^2 \\ & + \beta_6 I_1 \text{ Dose} + \beta_7 I_2 \text{ Dose} + \beta_8 I_3 \text{ Dose} \\ & + \beta_9 I_1 \text{ Dose}^2 + \beta_{10} I_2 \text{ Dose}^2 + \beta_{11} I_3 \text{ Dose}^2 + \varepsilon. \end{aligned}$$

The ($\beta_6 I_1$ Dose$+\beta_7 I_2$ Dose$+\beta_8 I_3$ Dose) component in the model formally corresponds to the *insect∗dose* interaction, whereas the ($\beta_9 I_1$ Dose$^2+\beta_{10} I_2$ Dose$^2+$ $\beta_{11} I_3$ Dose^2) component is equivalent to the $insect * dose * dose$ interaction (i.e., testing $H_0 : \beta_9 = \beta_{10} = \beta_{11} = 0$).

This discussion is not intended to confuse, but rather to impress upon you the intimate connection between regression and ANOVA, and to convince you of the care that is needed when modelling variation even in simple studies. Researchers are usually faced with more complex modelling problems than we have examined, where many variables might influence the response. If experimentation is possible, a scientist will often control the levels of variables that influence the response but that are not of primary interest. This can result in a manageable experiment with, say, four or fewer qualitative or quantitative variables that are systematically varied in a scientifically meaningful way. In observational studies, where experimentation is not possible, the scientist builds models to assess the effects of interest on the response, adjusting the response for all the uncontrolled variables that might be important. The uncontrolled variables are usually a mixture of factors and predictors. Ideally, the scientist knows what variables to control in an experiment and which to vary, and what variables are important to collect in an observational study.

The level of complexity that I am describing here might be intimidating, but certain basic principles can be applied to many of the studies you will see. Graduate students in statistics often take several courses (5+) in experimental

design, regression analysis, and linear model theory to master the breadth of models, and the subtleties of modelling, that are needed to be a good data analyst. I can only scratch the surface here. I will discuss a reasonably complex study having multiple factors and multiple predictors. The example focuses on strategies for building models, with little attempt to do careful diagnostic analyses. Hopefully, the example will give you an appreciation for statistical modelling, but **please be careful — these tools are dangerous!**

9.1 Some Comments on Building Models

A primary goal in many statistical analyses is to build a model or models to understand the variation in a response. Fortunately, or unfortunately, there is no consensus on how this should be done. Ideally, theory would suggest models to compare, but in many studies the goal is to provide an initial model that will be refined, validated, or refuted, by further experimentation. An extreme view is that the selected model(s) should only include effects that are "**statistically important**", whereas another extreme suggests that all effects that might be "**scientifically important**" should be included.

A difficulty with implementing either approach is that importance is relative to specific goals (i.e., Why are you building the model and what do you plan to use the model for? Is the model a prescription or device to make predictions? Is the model a tool to understand the effect that one or more variables have on a response, after adjusting for uninteresting, but important effects that can not be controlled? etc.) Madigan and Raftery, in the 1994 edition of *The Journal of the American Statistical Association*, comment that "Science is an iterative process in which competing models of reality are compared on the basis of how well they predict what is observed; models that predict much less well than their competitors are discarded." They argue that models should be selected using Occum's razor, a widely accepted norm in scientific investigations whereby the simplest plausible model among all reasonable models, given the data, is preferred. Madigan and Raftery's ideas are fairly consistent with the first extreme, but can be implemented in a variety of ways, depending on how you measure prediction adequacy. They propose a Bayesian approach, based on model averaging and prior beliefs on the plausibility of different models. An alternative method using Mallow's C_p criterion will be discussed later.

A simple compromise between the two extremes might be to start the model building process with the most complex model that is scientifically reasonable, but still interpretable, and systematically eliminate effects using backward elimination. The initial or **maximal** model might include polynomial effects for predictors, main effects and interactions (2 factor, 3 factor, etc.) between fac-

tors, and products or interactions between predictors and factors. This approach might appear to be less than ideal because the importance of effects is assessed using hypothesis tests and no attempt is made to assess the effect of changes on predictions. However, one can show that the average squared error in predictions is essentially reduced by eliminating **insignificant** regression effects from the model, so this approach seems tenable.

It might be sensible to only assess significance of effects specified in the model statement. However, many of these effects consist of several degrees-of-freedom. That is, the effect corresponds to several regression coefficients in the model. (Refer to the discussion following the displayed equations on page 271). The individual regression variables that comprise an effect could also be tested individually. However, if the effect is a factor (with 3+ levels) or an interaction involving a factor, then the interpretation of tests on individual regression coefficients depends on the level of the factor that was selected to be the baseline category. The Type III F-test on the entire effect does not depend on the baseline category. In essence, two researchers can start with different representations of the same mathematical model (i.e., the parameters are defined differently for different choices of baseline categories), use the same algorithm for selecting a model, yet come to different final models for the data.

Statisticians often follow the **hierarchy** principle, which states that a lower order term (be it a factor or a predictor) may be considered for exclusion from a model only if no higher order effects that include the term are present in the model. For example, given an initial model with effects A, B, C, and the $A * B$ interaction, the only candidates for omission at the first step are C and $A * B$. If you follow the hierarchy principle, and test an entire effect rather than test the single degree-of-freedom components that comprise an effect, then the difficulty described above can not occur. The hierarchy principle is most appealing with **pure** ANOVA models (such as the three-factor model in the example below), where all the regression variables are indicators. In ANOVA models, the ANOVA effects are of interest because they imply certain structure on the means. The individual regression variables that define the effects are

not usually a primary interest.

A non-hierarchical backward elimination algorithm where single degree-of-freedom effects are eliminated independently of the other effects in the model is implemented in the `step()` procedure. Recall our discussion of backwards elimination from Chapter 3 earlier this semester.

9.2 Example: The Effect of Sex and Rank on Faculty Salary

The data in this example were collected from the personnel files of faculty at a small college in the 1970s. The data were collected to assess whether women were being discriminated against (consciously or unconsciously) in salary. The sample consists of tenured and tenure-stream faculty only. Temporary faculty were excluded from consideration (because they were already being discriminated against).

The variables below are id (individual identification numbers from 1 to 52), sex (coded 1 for female and 0 for male), rank (coded 1 for Asst. Professor, 2 for Assoc. Professor and 3 for Full Professor), year (number of years in current rank), degree (coded 1 for Doctorate, 0 else), yd (number of years since highest degree was earned), and salary (academic year salary in dollars).

```
library(tidyverse)

# load ada functions
source("ada_functions.R")

#### Example: Faculty salary
dat_faculty <-
  read_table2("http://statacumen.com/teach/ADA2/notes/ADA2_notes_Ch09_faculty.dat") %>%
  mutate(
    sex    = factor(sex   , labels=c("Male", "Female"))
    # ordering the rank variable so Full is the baseline, then descending.
  , rank   = factor(rank  , levels=c(3,2,1), labels=c("Full", "Assoc", "Asst"))
  , degree = factor(degree, labels=c("Other", "Doctorate"))
  )
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   sex = col_double(),
##   rank = col_double(),
##   year = col_double(),
##   degree = col_double(),
##   yd = col_double(),
##   salary = col_double()
## )
```

```
head(dat_faculty)
```

```
## # A tibble: 6 x 7
```

```
##      id sex    rank   year degree       yd salary
##   <dbl> <fct>  <fct> <dbl> <fct>     <dbl>  <dbl>
## 1     1 Male   Full     25 Doctorate    35  36350
## 2     2 Male   Full     13 Doctorate    22  35350
## 3     3 Male   Full     10 Doctorate    23  28200
## 4     4 Female Full      7 Doctorate    27  26775
## 5     5 Male   Full     19 Other        30  33696
## 6     6 Male   Full     16 Doctorate    21  28516
```

```
str(dat_faculty)
```

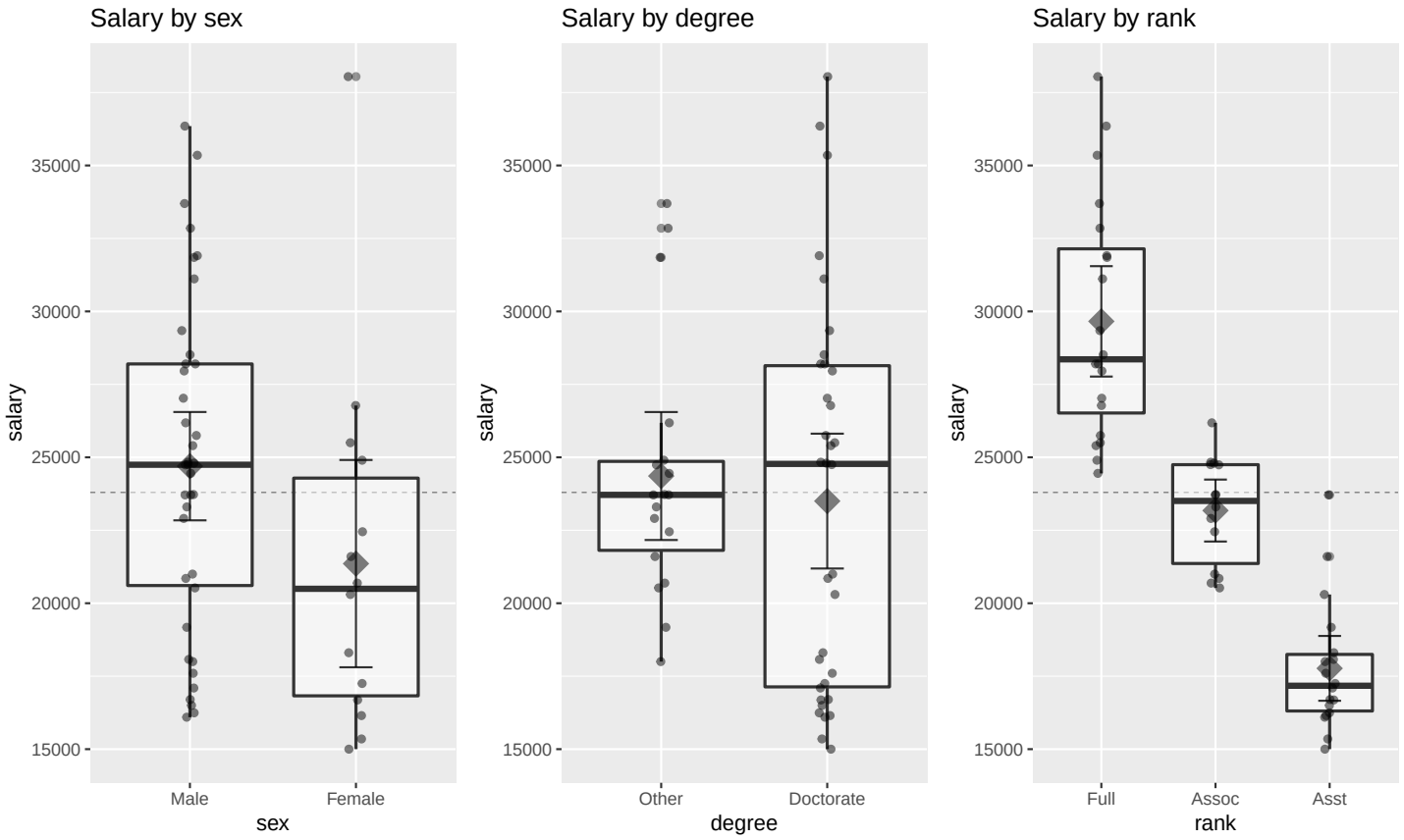
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 52 obs. of  7 variables:
##  $ id    : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ sex   : Factor w/ 2 levels "Male","Female": 1 1 1 2 1 1 2 1 1 1 ...
##  $ rank  : Factor w/ 3 levels "Full","Assoc",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year  : num  25 13 10 7 19 16 0 16 13 13 ...
##  $ degree: Factor w/ 2 levels "Other","Doctorate": 2 2 2 2 1 2 1 2 1 1 ...
##  $ yd    : num  35 22 23 27 30 21 32 18 30 31 ...
##  $ salary: num  36350 35350 28200 26775 33696 ...
```

The data includes two potential predictors of salary (year and yd), and three factors (sex, rank, and degree). A primary statistical interest is whether males and females are compensated equally, on average, after adjusting salary for rank, years in rank, and the other given effects. Furthermore, we wish to know whether an effect due to sex is the same for each rank, or not.

Before answering these questions, let us look at the data. I will initially focus on the effect of the individual factors (sex, rank, and degree) on salary. A series of box-plots is given below. Looking at the boxplots, notice that women tend to earn less than men, that faculty with Doctorates tend to earn more than those without Doctorates (median), and that salary tends to increase with rank.

```
# plot marginal boxplots

# Plot the data using ggplot
library(ggplot2)
p1 <- ggplot(dat_faculty, aes(x = sex, y = salary, group = sex))
# plot a reference line for the global mean (assuming no groups)
p1 <- p1 + geom_hline(aes(yintercept = mean(salary)),
                      colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p1 <- p1 + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p1 <- p1 + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.5)
# diamond at mean for each group
p1 <- p1 + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 6,
                        alpha = 0.5)
# confidence limits based on normal distribution
p1 <- p1 + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                        width = .2, alpha = 0.8)
p1 <- p1 + labs(title = "Salary by sex")

# Plot the data using ggplot
library(ggplot2)
```

```
p2 <- ggplot(dat_faculty, aes(x = degree, y = salary, group = degree))
# plot a reference line for the global mean (assuming no groups)
p2 <- p2 + geom_hline(aes(yintercept = mean(salary)),
                      colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p2 <- p2 + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p2 <- p2 + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.5)
# diamond at mean for each group
p2 <- p2 + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 6,
                        alpha = 0.5)
# confidence limits based on normal distribution
p2 <- p2 + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                        width = .2, alpha = 0.8)
p2 <- p2 + labs(title = "Salary by degree")

# Plot the data using ggplot
library(ggplot2)
p3 <- ggplot(dat_faculty, aes(x = rank, y = salary, group = rank))
# plot a reference line for the global mean (assuming no groups)
p3 <- p3 + geom_hline(aes(yintercept = mean(salary)),
                      colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p3 <- p3 + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p3 <- p3 + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.5)
# diamond at mean for each group
p3 <- p3 + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 6,
                        alpha = 0.5)
# confidence limits based on normal distribution
p3 <- p3 + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                        width = .2, alpha = 0.8)
p3 <- p3 + labs(title = "Salary by rank")

library(gridExtra)
grid.arrange(grobs = list(p1, p2, p3), nrow = 1)
```



9.2.1 A Three-Way ANOVA on Salary Data

Hopefully, our earlier analyses have cured you of the desire to claim that a sex effect exists before considering whether the differences between male and female salaries might be due to other factors. The output below gives the sample sizes, means, and standard deviations for the 11 combinations of sex, rank, and degree observed in the data. Side-by-side boxplots of the salaries for the 11 combinations are also provided. One combination of the three factors was not observed: female Associate Professors without Doctorates.

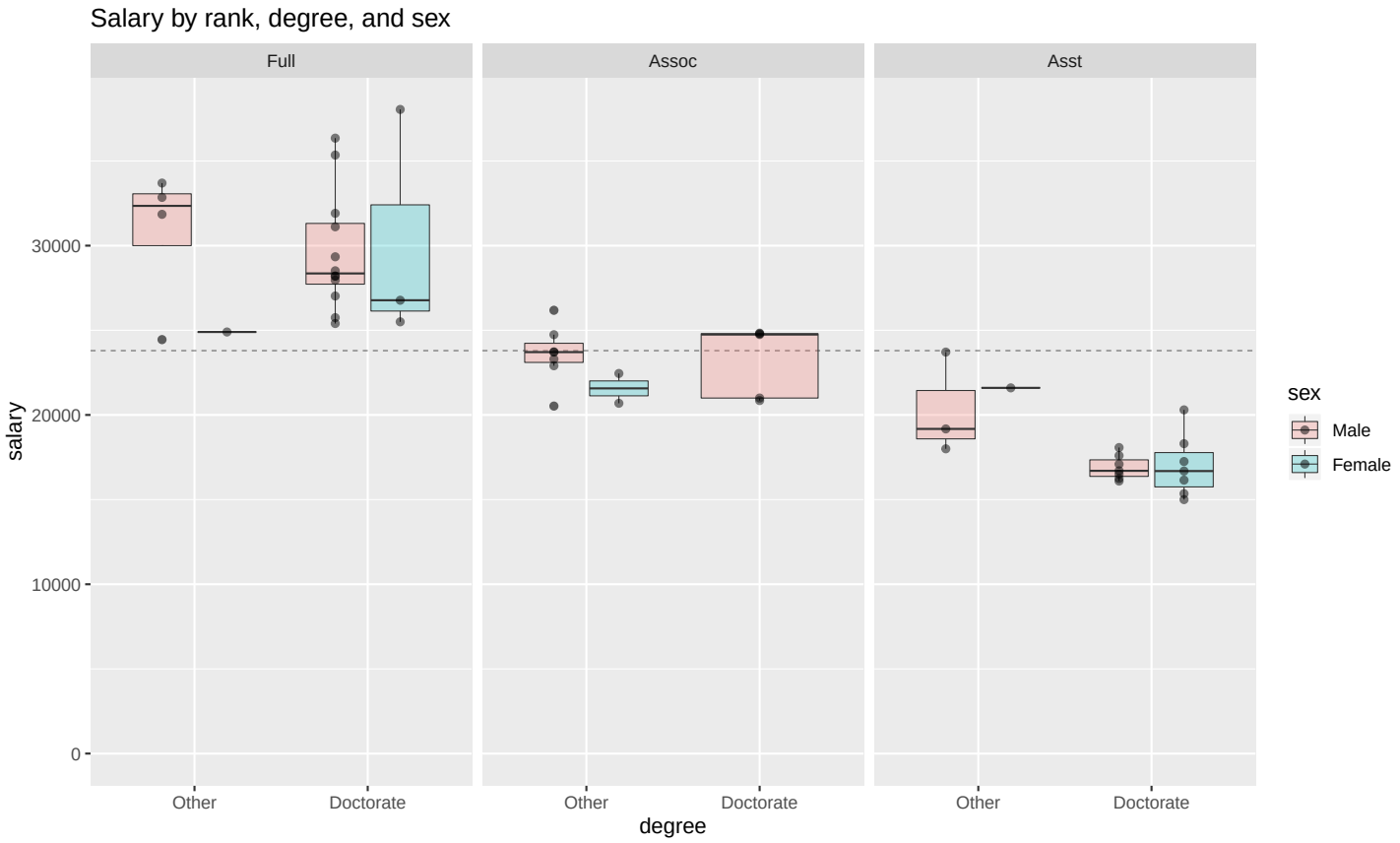
Looking at the summaries, the differences between sexes **within** each combination of rank and degree appear to be fairly small. There is a big difference in the ranks of men and women, with a higher percentage of men in the more advanced ranks. This might explain the differences between male and female salaries, when other factors are ignored.

```
sum_faculty <-
  dat_faculty %>%
  group_by(sex, rank, degree) %>%
  summarize(
    n = n()
  , m = mean(salary)
  , s = sd(salary)
  )
sum_faculty
```

```
## # A tibble: 11 x 6
## # Groups:   sex, rank [6]
##    sex    rank  degree        n      m     s
##    <fct>  <fct> <fct>     <int>  <dbl> <dbl>
##  1 Male   Full  Other         4 30712. 4242.
##  2 Male   Full  Doctorate    12 29593. 3480.
##  3 Male   Assoc Other         7 23585. 1733.
##  4 Male   Assoc Doctorate     5 23246. 2120.
##  5 Male   Asst  Other         3 20296  3017.
##  6 Male   Asst  Doctorate     7 16901.  729.
##  7 Female Full  Other         1 24900   NaN
##  8 Female Full  Doctorate     3 30107. 6904.
##  9 Female Assoc Other         2 21570  1245.
## 10 Female Asst  Other         1 21600   NaN
## 11 Female Asst  Doctorate     7 17006. 1835.
```

```
# plot marginal boxplots

library(ggplot2)
# create position dodge offset for plotting points
pd <- position_dodge(0.75) # 0.75 puts dots up center of boxplots
```

```
p <- ggplot(dat_faculty, aes(x = degree, y = salary, fill = sex))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(aes(yintercept = mean(salary)),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.25 for thin lines
p <- p + geom_boxplot(size = 0.25, alpha = 0.25)
# points for observed data
p <- p + geom_point(position = pd, alpha = 0.5)
p <- p + facet_grid(. ~ rank)
p <- p + scale_y_continuous(limits = c(0, max(dat_faculty$salary)))
p <- p + labs(title = "Salary by rank, degree, and sex")
print(p)
```



I will consider two simple analyses of these data. The first analysis considers the effect of the three factors on salary. The second analysis considers the effect of the predictors. A complete analysis using both factors and predictors is then considered. I am doing the three factor analysis because the most complex pure ANOVA problem we considered this semester has two factors — the analysis is for illustration **only**!!

The full model for a three-factor study includes the three main effects, the three possible two-factor interactions, plus the three-factor interaction. Identifying the factors by S (sex), D (degree) and R (rank), we write the full model

as

$$\begin{aligned}\text{Salary} &= \text{Grand mean} + \text{S effect} + \text{D effect} + \text{R effect} \\ &\quad + \text{S*D interaction} + \text{S*R interaction} + \text{R*D interaction} \\ &\quad + \text{S*D*R interaction} + \text{Residual}.\end{aligned}$$

You should understand what main effects and two-factor interactions measure, but what about the three-factor term? If you look at the two levels of degree separately, then a three-factor interaction is needed if the interaction between sex and rank is different for the two degrees. (i.e., the profile plots are different for the two degrees). Not surprisingly, three-factor interactions are hard to interpret.

I considered a hierarchical backward elimination of effects (see Chapter 3 for details). Individual regression variables are not considered for deletion, unless they correspond to an effect in the model statement. All tests were performed at the 0.10 level, but this hardly matters here.

The first step in the elimination is to fit the full model and check whether the three-factor term is significant. The three-factor term was not significant (in fact, it couldn't be fit because one category had zero observations). After eliminating this effect, I fit the model with all three two-factor terms, and then sequentially deleted the least important effects, one at a time, while still adhering to the hierarchy principle using the AIC criterion from the `step()` function. The final model includes only an effect due to rank. Finally, I compute the `lsmeans()` to compare salary for all pairs of rank.

```
# fit full model
lm_faculty_factor_full <-
  lm(
    salary ~ sex * rank * degree
  , data = dat_faculty
  )

library(car)
Anova(lm_faculty_factor_full, type=3)
```

```
## Error in Anova.III.lm(mod, error, singular.ok = singular.ok, ...):  there are aliased
coefficients in the model
```

Note that there are not enough degrees-of-freedom to estimate all these

effects because we have 0 observations for the Female/Assoc/Doctorate combination.

```
summary(lm_faculty_factor_full)
```

```
##
## Call:
## lm(formula = salary ~ sex * rank * degree, data = dat_faculty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6261.5 -1453.0  -225.9  1349.7  7938.3
##
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         30711.5     1485.0  20.681  < 2e-16
## sexFemale                           -5811.5     3320.7  -1.750 0.087581
## rankAssoc                           -7126.9     1861.6  -3.828 0.000433
## rankAsst                           -10415.5     2268.4  -4.591 4.13e-05
## degreeDoctorate                     -1118.8     1714.8  -0.652 0.517774
## sexFemale:rankAssoc                  3796.9     4086.3   0.929 0.358229
## sexFemale:rankAsst                   7115.5     4773.7   1.491 0.143734
## sexFemale:degreeDoctorate            6325.4     3834.4   1.650 0.106653
## rankAssoc:degreeDoctorate             780.4     2442.3   0.320 0.750952
## rankAsst:degreeDoctorate            -2276.1     2672.3  -0.852 0.399304
## sexFemale:rankAssoc:degreeDoctorate      NA         NA      NA       NA
## sexFemale:rankAsst:degreeDoctorate  -7524.8     5383.7  -1.398 0.169720
##
## (Intercept)                        ***
## sexFemale                          .
## rankAssoc                          ***
## rankAsst                           ***
## degreeDoctorate
## sexFemale:rankAssoc
## sexFemale:rankAsst
## sexFemale:degreeDoctorate
## rankAssoc:degreeDoctorate
## rankAsst:degreeDoctorate
## sexFemale:rankAssoc:degreeDoctorate
## sexFemale:rankAsst:degreeDoctorate
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2970 on 41 degrees of freedom
## Multiple R-squared:  0.7975,Adjusted R-squared:  0.7481
## F-statistic: 16.14 on 10 and 41 DF,  p-value: 2.989e-11
```

Let's use backward selection to choose a reduced model.

```
## AIC
# option: test="F" includes additional information
```

```
#           for parameter estimate tests that we're familiar with
# option: for BIC, include k=log(nrow( [data.frame name] ))
lm_faculty_factor_red_AIC <-
  step(
    lm_faculty_factor_full
  , direction="backward"
  , test="F"
  )
```

```
## Start:  AIC=841.26
## salary ~ sex * rank * degree
##
##                   Df Sum of Sq       RSS    AIC F value Pr(>F)
## <none>                         361677227 841.26
## - sex:rank:degree  1  17233177 378910404 841.68  1.9536 0.1697
```

Because the full model can not be fit, the `step()` procedure does not work. Below we remove the three-way interaction, then the `step()` procedure will do the rest of the work for us.

Remove the three-way interaction, then use `step()` to perform backward selection based on AIC.

```
# model reduction using update() and subtracting (removing) model terms
lm_faculty_factor_red <- lm_faculty_factor_full

# remove variable
lm_faculty_factor_red <-
  update(
    lm_faculty_factor_red
  , ~ . - sex:rank:degree
  )
Anova(lm_faculty_factor_red, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: salary
##                 Sum Sq Df  F value    Pr(>F)
## (Intercept) 3932650421  1 435.9113 < 2.2e-16 ***
## sex           11227674  1   1.2445 0.2709438
## rank         196652264  2  10.8989 0.0001539 ***
## degree          421614  1   0.0467 0.8298945
## sex:rank       2701493  2   0.1497 0.8614045
## sex:degree     7661926  1   0.8493 0.3620198
## rank:degree   33433415  2   1.8529 0.1693627
## Residuals    378910404 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# AIC backward selection
```

```r
lm_faculty_factor_red_AIC <-
  step(
    lm_faculty_factor_red
  , direction="backward"
  , test="F"
  )
```

```
## Start:  AIC=841.68
## salary ~ sex + rank + degree + sex:rank + sex:degree + rank:degree
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - sex:rank     2   2701493 381611896 838.05  0.1497 0.8614
## - sex:degree   1   7661926 386572329 840.72  0.8493 0.3620
## <none>                     378910404 841.68
## - rank:degree  2  33433415 412343819 842.08  1.8529 0.1694
##
## Step:  AIC=838.05
## salary ~ sex + rank + degree + sex:degree + rank:degree
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - sex:degree   1  12335789 393947686 837.71  1.4223 0.2394
## <none>                     381611896 838.05
## - rank:degree  2  32435968 414047864 838.29  1.8699 0.1662
##
## Step:  AIC=837.71
## salary ~ sex + rank + degree + rank:degree
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - sex          1   3009036 396956722 836.10  0.3437 0.5606
## - rank:degree  2  27067985 421015671 837.16  1.5460 0.2242
## <none>                     393947686 837.71
##
## Step:  AIC=836.1
## salary ~ rank + degree + rank:degree
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - rank:degree  2  31019255 427975976 836.01  1.7973 0.1772
## <none>                     396956722 836.10
##
## Step:  AIC=836.01
## salary ~ rank + degree
##
##          Df  Sum of Sq        RSS    AIC F value   Pr(>F)
## - degree  1   10970082  438946058 835.33  1.2304   0.2729
## <none>                  427975976 836.01
## - rank    2 1349072233 1777048209 906.04 75.6532 1.45e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Step:  AIC=835.33
## salary ~ rank
##
##        Df  Sum of Sq        RSS    AIC F value    Pr(>F)
## <none>                438946058 835.33
## - rank  2 1346783800 1785729858 904.30  75.171 1.174e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# all are significant, stop.
# final model: salary ~ rank
lm_faculty_factor_final <- lm_faculty_factor_red_AIC

library(car)
Anova(lm_faculty_factor_final, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: salary
##                 Sum Sq Df  F value    Pr(>F)
## (Intercept) 1.7593e+10  1 1963.932 < 2.2e-16 ***
## rank        1.3468e+09  2   75.171 1.174e-15 ***
## Residuals   4.3895e+08 49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_faculty_factor_final)
```

```
##
## Call:
## lm(formula = salary ~ rank, data = dat_faculty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5209.0 -1819.2  -417.8  1586.6  8386.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29659.0      669.3  44.316  < 2e-16 ***
## rankAssoc    -6483.0     1043.0  -6.216 1.09e-07 ***
## rankAsst    -11890.3      972.4 -12.228  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2993 on 49 degrees of freedom
## Multiple R-squared:  0.7542,Adjusted R-squared:  0.7442
## F-statistic: 75.17 on 2 and 49 DF,  p-value: 1.174e-15
```

All ranks are different with salaries increasing with rank.

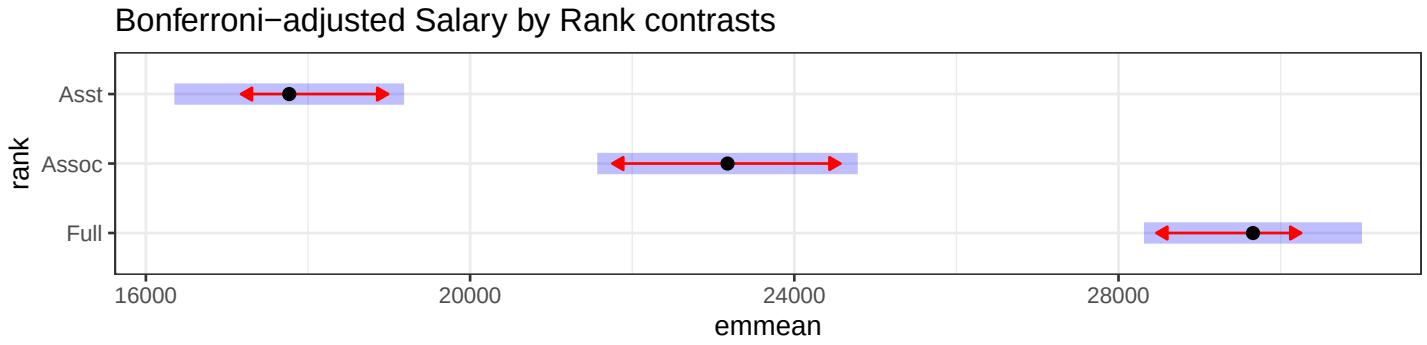
```
# Contrasts to perform pairwise comparisons
cont_f <-
  emmeans::emmeans(
    lm_faculty_factor_final
  , specs = "rank"
  )

# Means and CIs
confint(cont_f, adjust = "bonferroni")
```

```
##  rank  emmean  SE df lower.CL upper.CL
##  Full   29659 669 49    28000    31318
##  Assoc  23176 800 49    21193    25159
##  Asst   17769 705 49    16020    19517
##
## Confidence level used: 0.95
## Conf-level adjustment: bonferroni method for 3 estimates
```

```
# Pairwise comparisons
cont_f %>% pairs(adjust = "bonf")   # adjust = "tukey" is default
```

```
##  contrast     estimate   SE df t.ratio p.value
##  Full - Assoc     6483 1043 49  6.216  <.0001
##  Full - Asst     11890  972 49 12.228  <.0001
##  Assoc - Asst     5407 1067 49  5.070  <.0001
##
## P value adjustment: bonferroni method for 3 tests
```

```
# Plot means and contrasts
p <- plot(cont_f, comparisons = TRUE, adjust = "bonf")  # adjust = "tukey" is default
p <- p + labs(title = "Bonferroni-adjusted Salary by Rank contrasts")
p <- p + theme_bw()
print(p)
```



This analysis suggests that sex is not predictive of salary, once other factors are taken into account. In particular, faculty rank appears to be the sole important effect, in the sense that once salaries are adjusted for rank no other factors explain a significant amount of the unexplained variation in salaries.

As noted earlier, *the analysis was meant to illustrate a three-factor*

ANOVA and backward selection. The analysis is likely flawed, because it ignores the effects of year and year since degree on salary.

9.2.2 Using Year and Year Since Degree to Predict Salary

Plots of the salary against years in rank and years since degree show fairly strong associations with salary. The variability in salaries appears to be increasing with year and with year since degree, which might be expected. You might think to transform the salaries to a log scale to eliminate this effect, but doing so has little impact on the conclusions (not shown).

```
library(ggplot2)

p1 <- ggplot(dat_faculty, aes(x = year, y = salary, colour = rank, shape = sex, size = degree))
p1 <- p1 + scale_size_discrete(range=c(3,5))
## Warning:  Using size for a discrete variable is not advised.
p1 <- p1 + geom_point(alpha = 0.5)
p1 <- p1 + labs(title = "Salary by year")
p1 <- p1 + theme(legend.position = "bottom", legend.direction="vertical")
#print(p1)

p2 <- ggplot(dat_faculty, aes(x = yd, y = salary, colour = rank, shape = sex, size = degree))
p2 <- p2 + scale_size_discrete(range=c(3,5))
## Warning:  Using size for a discrete variable is not advised.
p2 <- p2 + geom_point(alpha = 0.5)
p2 <- p2 + labs(title = "Salary by yd")
p2 <- p2 + theme(legend.position = "bottom", legend.direction="vertical")
#print(p2)

library(gridExtra)
grid.arrange(grobs = list(p1, p2), nrow = 1)
```

As a point of comparison with the three-factor ANOVA, I fit a multiple regression model with year and years since degree as predictors of salary. These two predictors are important for explaining the variation in salaries, but together they explain much less of the variation (58%) than rank does on its own (75%).

```
# interaction model
lm_s_y_yd_yyd <-
  lm(
    salary ~ year * yd
  , data = dat_faculty
  )
summary(lm_s_y_yd_yyd)
```

```
##
## Call:
## lm(formula = salary ~ year * yd, data = dat_faculty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10368.5  -2361.5   -505.7   2363.1  12211.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16287.391   1395.049  11.675 1.25e-15 ***
## year          561.155    275.243   2.039  0.04700 *
## yd            235.415     83.266   2.827  0.00683 **
## year:yd        -3.089     10.412  -0.297  0.76796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3958 on 48 degrees of freedom
## Multiple R-squared:  0.579,Adjusted R-squared:  0.5527
## F-statistic:    22 on 3 and 48 DF,  p-value: 4.17e-09
```

```
# interaction is not significant
lm_s_y_yd <-
  lm(
    salary ~ year + yd
  , data = dat_faculty
  )
summary(lm_s_y_yd)
```

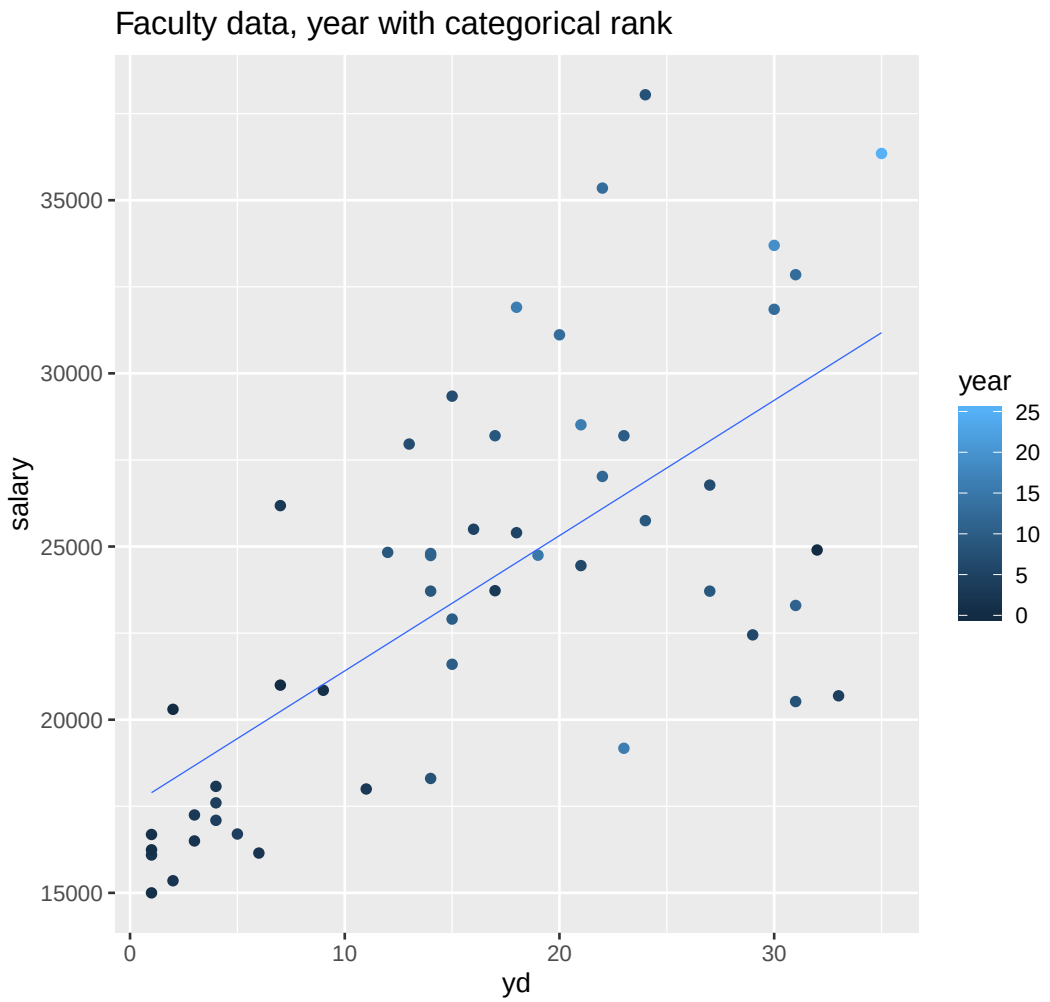
```
##
## Call:
## lm(formula = salary ~ year + yd, data = dat_faculty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -10321.2  -2347.2   -332.7   2298.8  12240.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16555.7     1052.4  15.732  < 2e-16 ***
## year           489.3      129.6   3.777 0.000431 ***
## yd             222.2       69.8   3.184 0.002525 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3921 on 49 degrees of freedom
## Multiple R-squared:  0.5782,Adjusted R-squared:  0.561
## F-statistic: 33.58 on 2 and 49 DF,  p-value: 6.532e-10
```

```
# Put predicted values into the dataset for plotting
dat_faculty$pred <- predict(lm_s_y_yd)
```

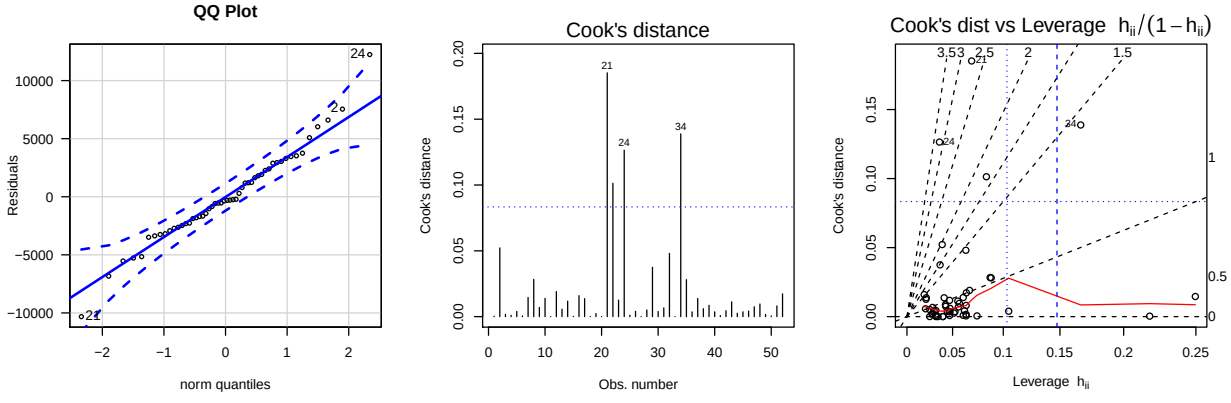
Parallel lines look reasonable. There are a few extreme salaries for Full professor rank; the effect they will have is to shift up the entire regression line and inflate the variance since they are in the middle of the range of the year variable with low leverage.

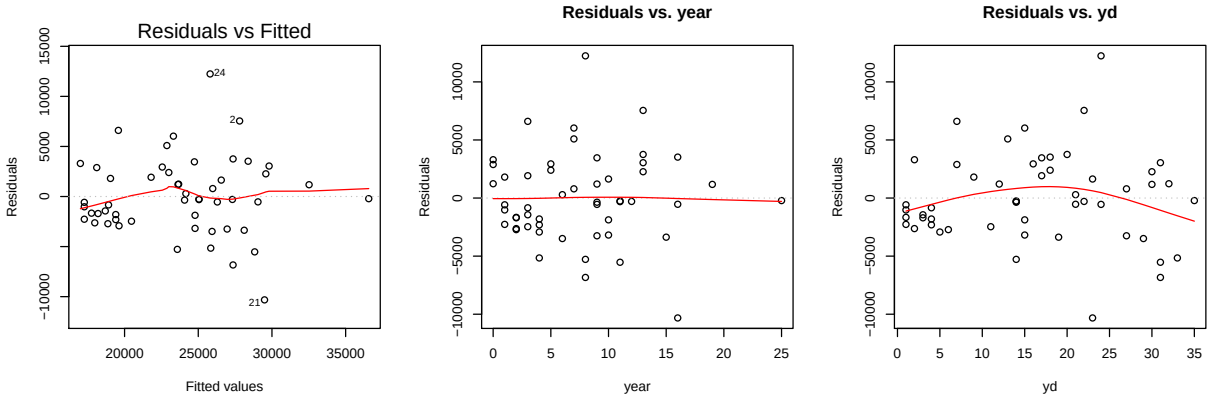
```
library(ggplot2)
p <- ggplot(dat_faculty, aes(x = yd, y = salary, colour = year))
p <- p + geom_point()
p <- p + geom_smooth(method = lm, se = FALSE, size = 1/4)
#p <- p + geom_line(aes(y = pred), size = 1)
p <- p + labs(
    title = "Faculty data, year with categorical rank"
  , caption = paste0( "Solid lines are regression lines fit to each group separately (interaction model).\n"
                    , "Dashed line is regression line from fitted model (additive equal slope model).")
  )
print(p)
```



Solid lines are regression lines fit to each group separately (interaction model).
Dashed line is regression line from fitted model (additive equal slope model).

```
# plot diagnostics
lm_diag_plots(lm_s_y_yd, sw_plot_set = "simple")
```




Residuals vs Fitted
Residuals
Fitted values
24
2
21
Residuals vs. year
Residuals
year
Residuals vs. yd
Residuals
yd

9.2.3 Using Factors and Predictors to Model Salaries

The plots we looked at helped us to understand the data. In particular, the plot of salary against years in rank, using rank as a plotting symbol, suggests that a combination of predictors and factors will likely be better for modelling faculty salaries than either of the two models that we proposed up to this point.

There is no evidence of non-linearity in the plots of salary against the predictors, so I will not consider transforming years since degree, years in rank, or salary. Note that the increasing variability in salaries for increasing years in rank and increasing years since degree is partly due to differences in the relationships across ranks. The non-constant variance should be less of a concern in any model that includes rank and either years in rank or years since degree as effects.

I started the model building process with a **maximal** or full model with the five main effects plus the 10 possible interactions between two effects, regardless of whether the effects were factors or predictors. Notationally, this model is written as follows:

$$\begin{aligned}\text{Salary} = {} & \text{Grand mean} + \text{S effect} + \text{D effect} + \text{R effect} + \text{YEAR effect} + \text{YD effect} \\ & + \text{S * D interaction} + \text{S * R interaction} + \text{S * YEAR interaction} + \text{S * YD interaction} \\ & + \text{D * R interaction} + \text{D * YEAR interaction} + \text{D * YD interaction} \\ & + \text{R * YEAR interaction} + \text{R * YD interaction} + \text{YEAR * YD interaction} + \text{Residual},\end{aligned}$$

where the year and year since degree effects (YD) are linear terms (as in the multiple regression model we considered). To check whether any important effects might have been omitted, I added individual three-factor terms to this model. All of the three factor terms were insignificant (not shown), so I believe that my choice for the "maximal" model is sensible.

The output below gives the fit to the maximal model, and subsequent fits, using the hierarchy principle. Only selected summaries are provided.

```
# fit full model with two-way interactions
lm_faculty_full <-
  lm(
    salary ~ (sex + rank + degree + year + yd)^2
  , data = dat_faculty
  )
library(car)
Anova(lm_faculty_full, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: salary
##                Sum Sq Df F value  Pr(>F)
## (Intercept)  22605087  1  3.6916 0.06392 .
## sex           4092995  1  0.6684 0.41984
## rank          5731837  2  0.4680 0.63059
## degree        4137628  1  0.6757 0.41735
## year          2022246  1  0.3302 0.56966
## yd            3190911  1  0.5211 0.47578
## sex:rank       932237  2  0.0761 0.92688
## sex:degree    7164815  1  1.1701 0.28773
## sex:year      7194388  1  1.1749 0.28676
## sex:yd        2024210  1  0.3306 0.56947
## rank:degree  13021265  2  1.0632 0.35759
## rank:year     1571933  2  0.1284 0.88001
## rank:yd       9822382  2  0.8020 0.45750
## degree:year   4510249  1  0.7366 0.39735
## degree:yd     6407880  1  1.0465 0.31424
## year:yd         50921  1  0.0083 0.92793
## Residuals   189825454 31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# This time I use BIC for model reduction by specifying k=
#   (compare this to model result using AIC --
#      too many nonsignificant parameters left in model)

## BIC
# option: test="F" includes additional information
#           for parameter estimate tests that we're familiar with
# option: for BIC, include k=log(nrow( [data.frame name] ))
lm_faculty_red_BIC <-
  step(
    lm_faculty_full
  , direction="backward"
  , test="F"
  , k=log(nrow(dat_faculty))
  )
```

```
## Start:  AIC=868.72
## salary ~ (sex + rank + degree + year + yd)^2
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - sex:rank     2    932237 190757690 861.07  0.0761 0.9269
## - rank:year    2   1571933 191397386 861.24  0.1284 0.8800
## - rank:yd      2   9822382 199647836 863.44  0.8020 0.4575
## - rank:degree  2  13021265 202846719 864.26  1.0632 0.3576
## - year:yd      1     50921 189876375 864.78  0.0083 0.9279
## - sex:yd       1   2024210 191849663 865.32  0.3306 0.5695
```

```
## - degree:year  1   4510249 194335703 865.99  0.7366 0.3974
## - degree:yd    1   6407880 196233334 866.49  1.0465 0.3142
## - sex:degree   1   7164815 196990268 866.69  1.1701 0.2877
## - sex:year     1   7194388 197019841 866.70  1.1749 0.2868
## <none>                     189825454 868.72
##
## Step:  AIC=861.07
## salary ~ sex + rank + degree + year + yd + sex:degree + sex:year +
##     sex:yd + rank:degree + rank:year + rank:yd + degree:year +
##     degree:yd + year:yd
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - rank:year    2   4480611 195238301 854.37  0.3876 0.6818
## - rank:yd      2  14587933 205345624 857.00  1.2618 0.2964
## - year:yd      1     25889 190783580 857.12  0.0045 0.9470
## - rank:degree  2  16365099 207122790 857.45  1.4155 0.2572
## - sex:yd       1   3293276 194050966 858.01  0.5697 0.4557
## - degree:year  1   4428068 195185758 858.31  0.7660 0.3878
## - degree:yd    1   6525075 197282766 858.87  1.1288 0.2957
## - sex:year     1  10462381 201220071 859.89  1.8099 0.1877
## - sex:degree   1  10654937 201412628 859.94  1.8432 0.1838
## <none>                     190757690 861.07
##
## Step:  AIC=854.37
## salary ~ sex + rank + degree + year + yd + sex:degree + sex:year +
##     sex:yd + rank:degree + rank:yd + degree:year + degree:yd +
##     year:yd
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - year:yd      1    582367 195820669 850.58  0.1044 0.7485
## - rank:degree  2  18612514 213850816 851.21  1.6683 0.2032
## - sex:yd       1   3008739 198247041 851.22  0.5394 0.4676
## - rank:yd      2  20258184 215496486 851.60  1.8158 0.1777
## - degree:year  1   7497925 202736226 852.38  1.3441 0.2542
## - degree:yd    1   8179958 203418259 852.56  1.4664 0.2340
## - sex:degree   1  12500896 207739197 853.65  2.2410 0.1434
## - sex:year     1  12669105 207907406 853.69  2.2712 0.1408
## <none>                     195238301 854.37
##
## Step:  AIC=850.58
## salary ~ sex + rank + degree + year + yd + sex:degree + sex:year +
##     sex:yd + rank:degree + rank:yd + degree:year + degree:yd
##
##               Df Sum of Sq       RSS    AIC F value  Pr(>F)
## - sex:yd       1   2456466 198277134 847.27  0.4516 0.50587
## - rank:degree  2  21836322 217656990 848.17  2.0072 0.14912
## - degree:year  1   7414066 203234734 848.56  1.3630 0.25069
## - degree:yd    1   9232872 205053541 849.02  1.6974 0.20090
```

```
## - sex:degree   1  12831931 208652600 849.93  2.3590 0.13330
## - sex:year     1  13646799 209467467 850.13  2.5089 0.12196
## <none>                     195820669 850.58
## - rank:yd      2  41051000 236871669 852.57  3.7734 0.03253 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=847.27
## salary ~ sex + rank + degree + year + yd + sex:degree + sex:year +
##     rank:degree + rank:yd + degree:year + degree:yd
##
##               Df Sum of Sq       RSS    AIC F value  Pr(>F)
## - rank:degree  2  21157939 219435073 844.64  1.9741 0.15324
## - degree:year  1   8497324 206774458 845.50  1.5857 0.21583
## - degree:yd    1   9463400 207740534 845.75  1.7659 0.19202
## - sex:degree   1  10394382 208671516 845.98  1.9397 0.17202
## <none>                     198277134 847.27
## - sex:year     1  22789419 221066553 848.98  4.2527 0.04626 *
## - rank:yd      2  42516602 240793736 849.47  3.9670 0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=844.64
## salary ~ sex + rank + degree + year + yd + sex:degree + sex:year +
##     rank:yd + degree:year + degree:yd
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - degree:yd    1    361929 219797002 840.78  0.0643 0.8011
## - degree:year  1    855102 220290175 840.89  0.1520 0.6988
## - sex:degree   1   1616150 221051223 841.07  0.2872 0.5950
## - rank:yd      2  24391011 243826084 842.22  2.1675 0.1281
## - sex:year     1  10569795 230004869 843.14  1.8786 0.1783
## <none>                     219435073 844.64
##
## Step:  AIC=840.78
## salary ~ sex + rank + degree + year + yd + sex:degree + sex:year +
##     rank:yd + degree:year
##
##               Df Sum of Sq       RSS    AIC F value  Pr(>F)
## - sex:degree   1   3112507 222909509 837.56  0.5664 0.45609
## - degree:year  1   4414318 224211320 837.86  0.8033 0.37546
## - rank:yd      2  24695126 244492128 838.41  2.2471 0.11889
## - sex:year     1  16645026 236442028 840.62  3.0292 0.08947 .
## <none>                     219797002 840.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=837.56
```

```
## salary ~ sex + rank + degree + year + yd + sex:year + rank:yd +
##     degree:year
##
##               Df Sum of Sq       RSS    AIC F value Pr(>F)
## - degree:year  1   2585275 225494784 834.21  0.4755 0.4943
## - rank:yd      2  25367664 248277174 835.26  2.3330 0.1098
## - sex:year     1  14770974 237680484 836.94  2.7168 0.1069
## <none>                     222909509 837.56
##
## Step:  AIC=834.21
## salary ~ sex + rank + degree + year + yd + sex:year + rank:yd
##
##            Df Sum of Sq       RSS    AIC F value Pr(>F)
## - rank:yd   2  24905278 250400062 831.75  2.3194 0.1108
## - degree    1   8902098 234396882 832.27  1.6581 0.2049
## - sex:year  1  14134386 239629170 833.42  2.6326 0.1122
## <none>                  225494784 834.21
##
## Step:  AIC=831.75
## salary ~ sex + rank + degree + year + yd + sex:year
##
##            Df Sum of Sq       RSS    AIC F value    Pr(>F)
## - sex:year  1   8458303 258858365 829.53  1.4863   0.22929
## - degree    1  11217823 261617885 830.08  1.9712   0.16734
## - yd        1  16309342 266709404 831.08  2.8659   0.09755 .
## <none>                  250400062 831.75
## - rank      2 406263292 656663354 873.98 35.6941 6.144e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=829.53
## salary ~ sex + rank + degree + year + yd
##
##          Df Sum of Sq       RSS    AIC F value    Pr(>F)
## - sex     1   9134971 267993336 827.38  1.5880    0.2141
## - degree  1  10687589 269545954 827.68  1.8579    0.1796
## - yd      1  14868158 273726523 828.48  2.5847    0.1149
## <none>                258858365 829.53
## - year    1 144867403 403725768 848.69 25.1838 8.654e-06 ***
## - rank    2 399790682 658649047 870.19 34.7499 7.485e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=827.38
## salary ~ rank + degree + year + yd
##
##          Df Sum of Sq       RSS    AIC F value    Pr(>F)
## - degree  1   6684984 274678320 824.71  1.1475    0.2897
```

```
## - yd      1   7871680 275865016 824.93  1.3511    0.2511
## <none>                267993336 827.38
## - year    1 147642871 415636208 846.25 25.3423 7.839e-06 ***
## - rank    2 404108665 672102002 867.29 34.6818 6.544e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=824.71
## salary ~ rank + year + yd
##
##        Df Sum of Sq       RSS    AIC F value    Pr(>F)
## - yd    1   2314414 276992734 821.19   0.396    0.5322
## <none>              274678320 824.71
## - year  1 141105647 415783967 842.32  24.145 1.126e-05 ***
## - rank  2 478539101 753217421 869.26  40.941 5.067e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=821.19
## salary ~ rank + year
##
##        Df Sum of Sq       RSS    AIC F value    Pr(>F)
## <none>              276992734 821.19
## - year  1 161953324 438946058 841.18  28.065 2.905e-06 ***
## - rank  2 632056217 909048951 875.09  54.764 4.103e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `add1()` function will indicate whether a variable from the "full" model should be added to the current model. In our case, our BIC-backward selected model appears adequate.

```
add1(
    lm_faculty_red_BIC
  , . ~ (sex + rank + degree + year + yd)^2
  , test="F"
  )
```

```
## Single term additions
##
## Model:
## salary ~ rank + year
##           Df Sum of Sq       RSS    AIC F value Pr(>F)
## <none>                 276992734 813.39
## sex        1   2304648 274688086 814.95  0.3943 0.5331
## degree     1   1127718 275865016 815.18  0.1921 0.6632
## yd         1   2314414 274678320 814.95  0.3960 0.5322
## rank:year  2  15215454 261777280 814.45  1.3368 0.2727
```

No variables are suggested for addition, though sex is the first contender

with a p-value = 0.5331.

Let's look carefully at our resulting model.

```
# all are significant, stop.
# final model: salary ~ year + rank
lm_faculty_final <- lm_faculty_red_BIC

library(car)
Anova(lm_faculty_final, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: salary
##                 Sum Sq Df F value    Pr(>F)
## (Intercept) 4422688839  1 766.407 < 2.2e-16 ***
## rank         632056217  2  54.764 4.103e-13 ***
## year         161953324  1  28.065 2.905e-06 ***
## Residuals    276992734 48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_faculty_final)
```

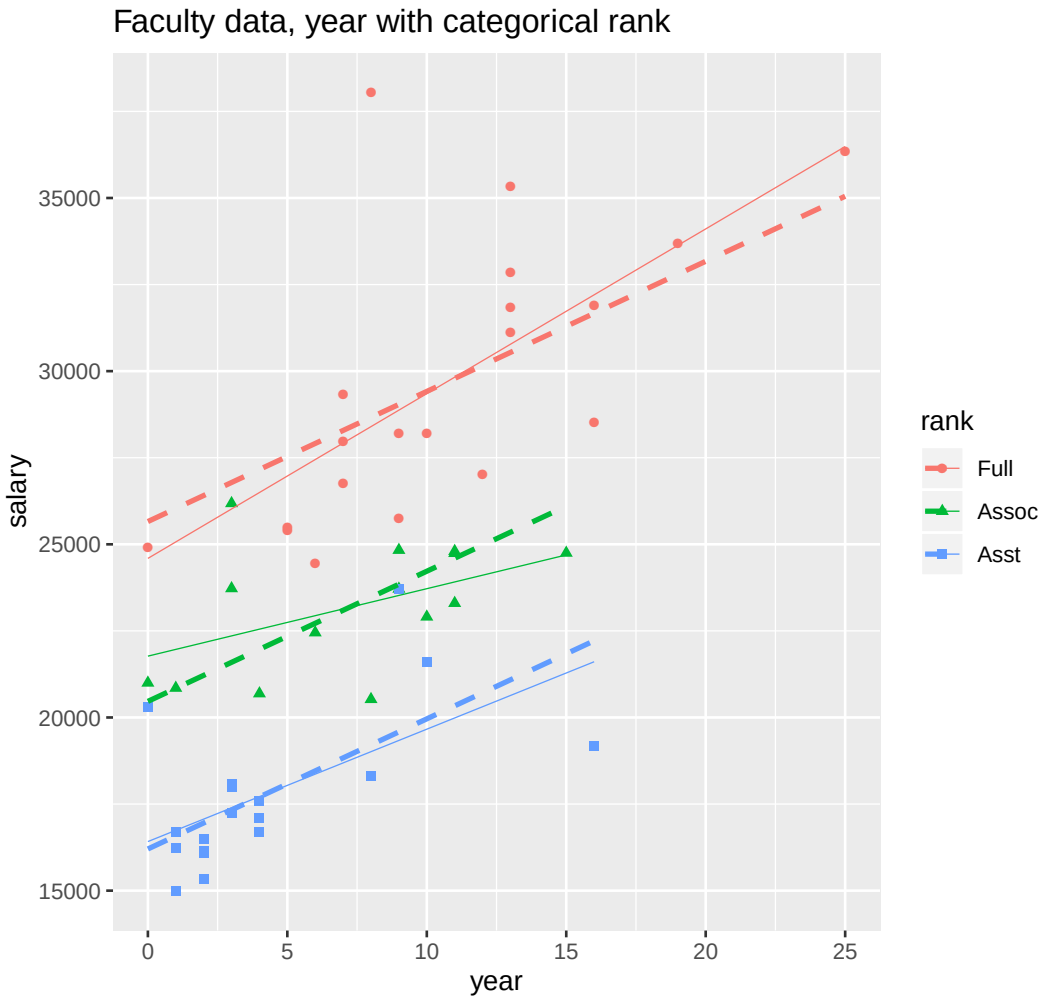
```
##
## Call:
## lm(formula = salary ~ rank + year, data = dat_faculty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3462.0 -1302.8  -299.2   783.5  9381.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25657.79     926.81  27.684  < 2e-16 ***
## rankAssoc   -5192.24     871.83  -5.956 2.93e-07 ***
## rankAsst    -9454.52     905.83 -10.437 6.12e-14 ***
## year          375.70      70.92   5.298 2.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2402 on 48 degrees of freedom
## Multiple R-squared:  0.8449,Adjusted R-squared:  0.8352
## F-statistic: 87.15 on 3 and 48 DF,  p-value: < 2.2e-16
```

```
# Put predicted values into the dataset for plotting
dat_faculty$pred <- predict(lm_faculty_final)
```

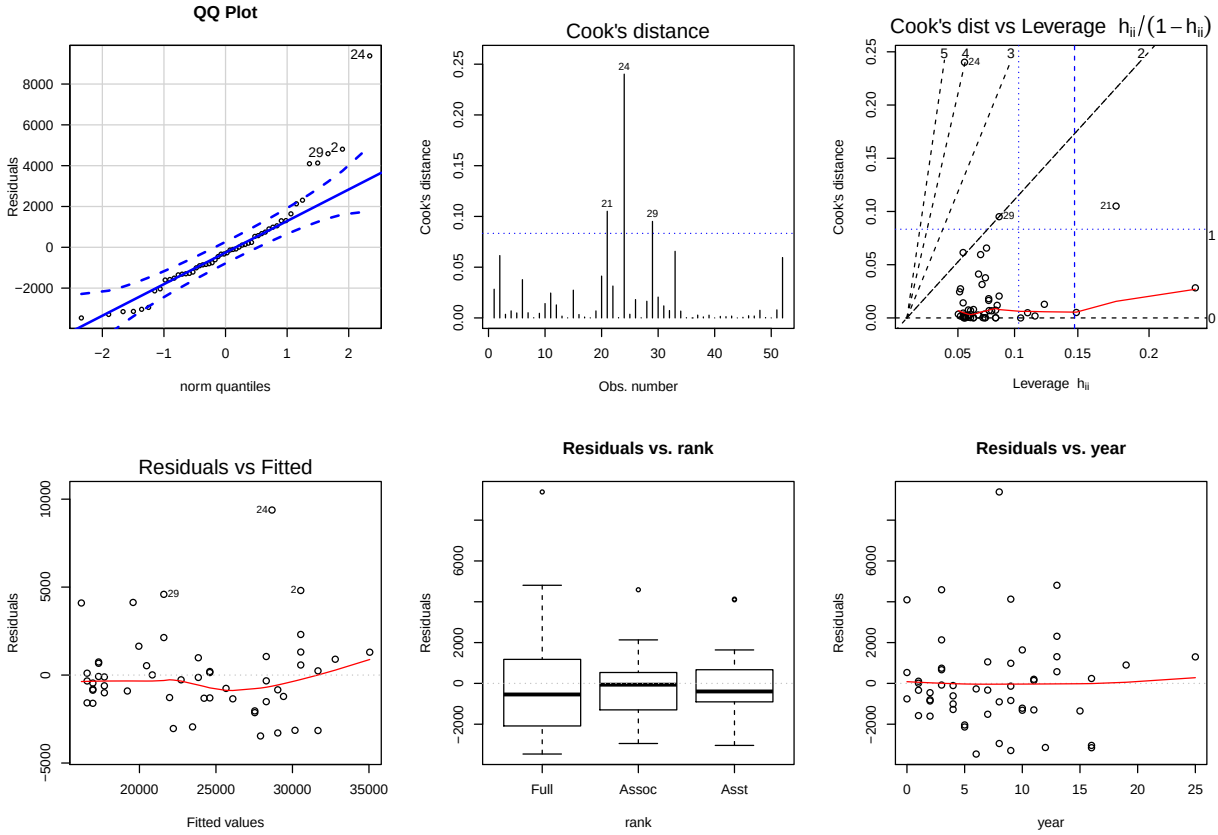
Parallel lines look reasonable. There are a few extreme salaries for Full professor rank; the effect they will have is to shift up the entire regression line and inflate the variance since they are in the middle of the range of the year

variable with low leverage.

```
library(ggplot2)
p <- ggplot(dat_faculty, aes(x = year, y = salary, colour = rank, shape = rank))
p <- p + geom_point()
p <- p + geom_smooth(method = lm, se = FALSE, size = 1/4)
p <- p + geom_line(aes(y = pred), linetype = 2, size = 1)
p <- p + labs(
    title = "Faculty data, year with categorical rank"
  , caption = paste0( "Solid lines are regression lines fit to each group separately (interaction model).\n"
                    , "Dashed line is regression line from fitted model (additive equal slope model).")
  )
print(p)
```




```
# plot diagnistics
lm_diag_plots(lm_faculty_final, sw_plot_set = "simple")
```



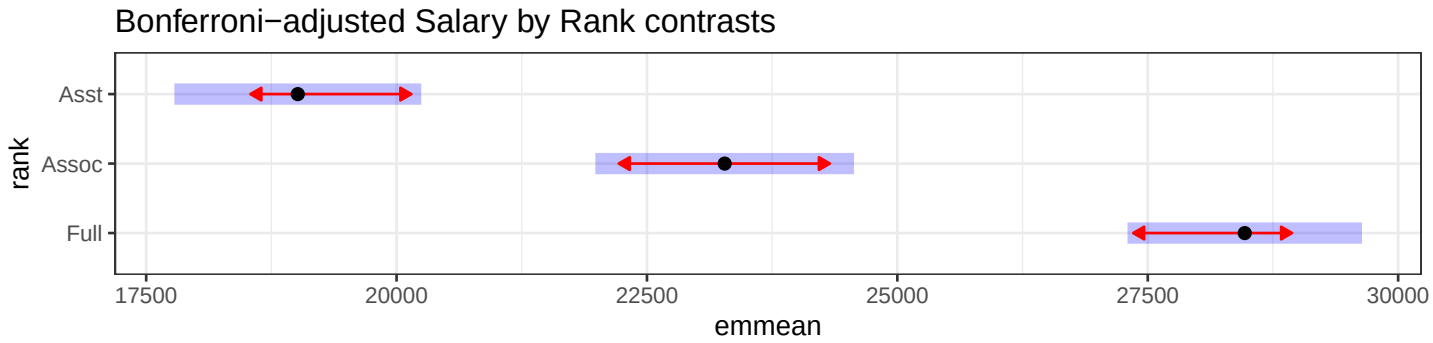

```
# Contrasts to perform pairwise comparisons
cont_f <-
  emmeans::emmeans(
    lm_faculty_final
  , specs = "rank"
  )

# Means and CIs
confint(cont_f, adjust = "bonferroni")
```

```
##  rank  emmean  SE df lower.CL upper.CL
##  Full   28468 582 48    27024    29913
##  Assoc  23276 642 48    21683    24869
##  Asst   19014 613 48    17493    20535
##
## Confidence level used: 0.95
## Conf-level adjustment: bonferroni method for 3 estimates
```

```
# Pairwise comparisons
cont_f %>% pairs(adjust = "bonf")   # adjust = "tukey" is default
```

```
##  contrast     estimate  SE df t.ratio p.value
##  Full - Assoc     5192 872 48  5.956  <.0001
##  Full - Asst      9455 906 48 10.437  <.0001
##  Assoc - Asst     4262 883 48  4.828  <.0001
##
## P value adjustment: bonferroni method for 3 tests
```

```
# Plot means and contrasts
p <- plot(cont_f, comparisons = TRUE, adjust = "bonf")  # adjust = "tukey" is default
p <- p + labs(title = "Bonferroni-adjusted Salary by Rank contrasts")
p <- p + theme_bw()
print(p)
```



9.2.4 Discussion of the Salary Analysis

The selected model is a simple ANCOVA model with a rank effect and a linear effect due to years in rank. Note that the maximal model has 20 single df effects with an $R^2 = 0.89$ while the selected model has 3 single df effects with $R^2 = 0.84$.

Looking at the parameter estimates table, all of the single df effects in the selected model are significant. The baseline group is Full Professors, with rank=3. Predicted salaries for the different ranks are given by:

$$\begin{aligned}
\text{Full:} \quad & \widehat{\text{salary}} = 25658 + 375.70\ \text{year} \\
\text{Assoc:} \quad & \widehat{\text{salary}} = 25658 - 5192 + 375.70\ \text{year} = 20466 + 375.70\ \text{year} \\
\text{Assis:} \quad & \widehat{\text{salary}} = 25658 - 9454 + 375.70\ \text{year} = 16204 + 375.70\ \text{year}
\end{aligned}$$

Do you remember how to interpret the **lsmeans**, and the p-values for comparing **lsmeans**?

You might be tempted to conclude that rank and years in rank are the only effects that are predictive of salaries, and that differences in salaries by sex are insignificant, once these effects have been taken into account. However, you must be careful because you have not done a diagnostic analysis. The following two issues are also important to consider.

A sex effect may exist even though there is insufficient evidence to support it based on these data. (**Lack of power corrupts; and absolute lack of power corrupts absolutely!**) If we are interested in the possibility of a sex effect, I think that we would do better by focusing on how large the effect might be, and whether it is important. A simple way to check is by constructing a confidence interval for the sex effect, based on a simple additive model that includes sex plus the effects that were selected as statistically significant, rank and year in rank. I am choosing this model because the omitted effects are hopefully small, and because the regression coefficient for a sex indicator is easy to interpret in an additive model. Other models might be considered for comparison. Summary output from this model is given below.

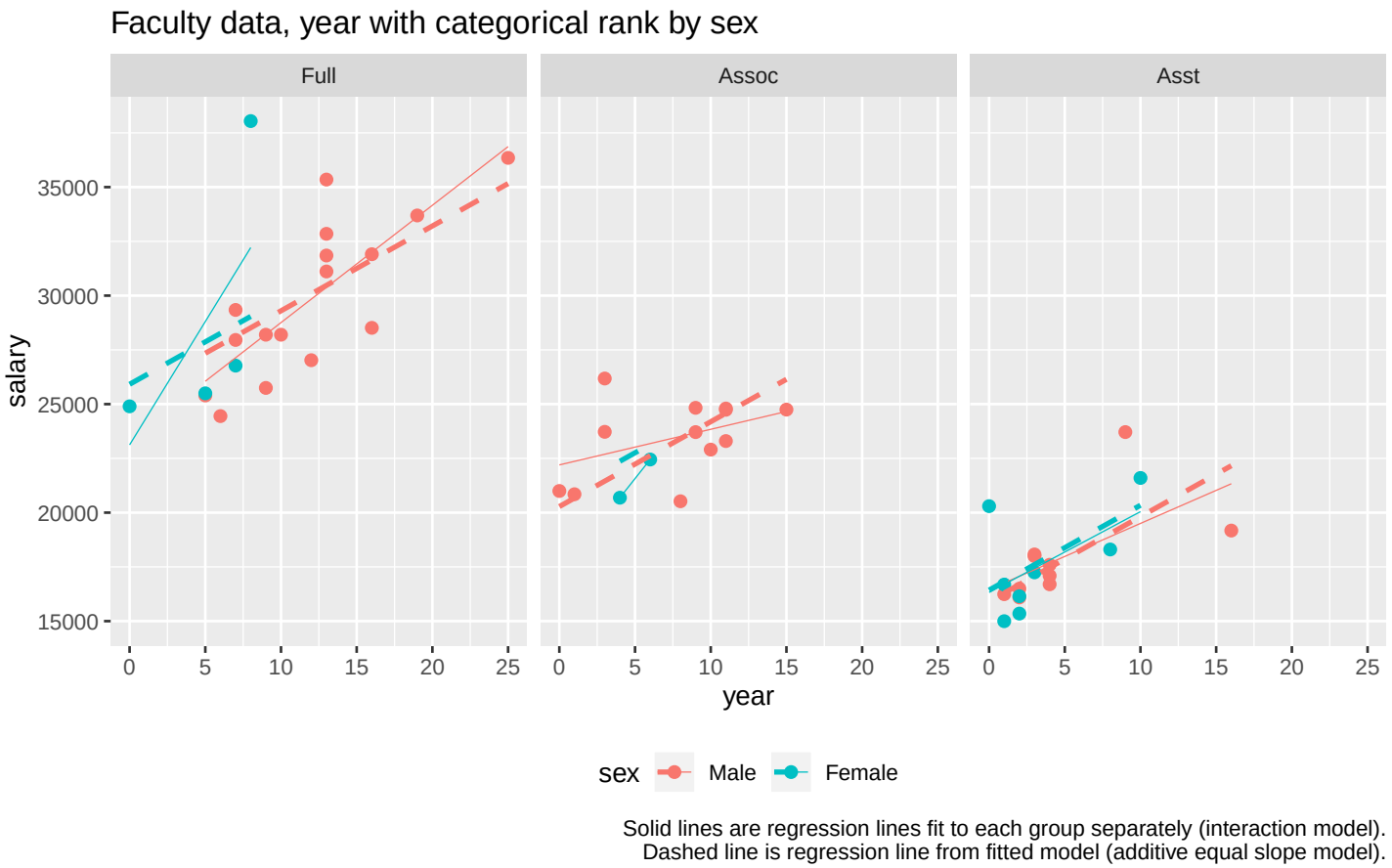
```
# add sex to the model
lm_faculty_final_sex <-
  update(
    lm_faculty_final
  , . ~ . + sex
  )
summary(lm_faculty_final_sex)
```

```
##
## Call:
## lm(formula = salary ~ rank + year + sex, data = dat_faculty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3286.3 -1311.8  -178.4   939.1  9002.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25390.65    1025.14  24.768  < 2e-16 ***
## rankAssoc   -5109.93     887.12  -5.760 6.20e-07 ***
## rankAsst    -9483.84     912.79 -10.390 9.19e-14 ***
## year          390.94      75.38   5.186 4.47e-06 ***
## sexFemale     524.15     834.69   0.628    0.533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2418 on 47 degrees of freedom
## Multiple R-squared:  0.8462,Adjusted R-squared:  0.8331
## F-statistic: 64.64 on 4 and 47 DF,  p-value: < 2.2e-16
```

```
# Put predicted values into the dataset for plotting
dat_faculty$pred <- predict(lm_faculty_final_sex)
```

Parallel lines look reasonable. There are a few extreme salaries for Full professor rank; the effect they will have is to shift up the entire regression line and inflate the variance since they are in the middle of the range of the year variable with low leverage.

```
library(ggplot2)
p <- ggplot(dat_faculty, aes(x = year, y = salary, colour = sex))
p <- p + geom_point(size = 2)
p <- p + geom_smooth(method = lm, se = FALSE, size = 1/4)
p <- p + geom_line(aes(y = pred), linetype = 2, size = 1)
p <- p + facet_wrap(~ rank, nrow = 1)
p <- p + theme(legend.position="bottom")
p <- p + labs(
    title = "Faculty data, year with categorical rank by sex"
  , caption = paste0( "Solid lines are regression lines fit to each group separately (interaction model).\n"
                    , "Dashed line is regression line from fitted model (additive equal slope model).")
  )
print(p)
```



Men are the baseline group for the sex effect, so the predicted salaries for men are 524 dollars **less** than that for women, adjusting for rank and year. A rough 95% CI for the sex differential is the estimated sex coefficient plus or minus two standard errors, or $524 \pm 2 * (835)$, or -1146 to 2194 dollars. The range of plausible values for the sex effect would appear to contain values of practical importance, so further analysis is warranted here.

Another concern, and potentially a more important issue, was raised by M. O. Finkelstein in a 1980 discussion in the **Columbia Law Review** on the use of regression in discrimination cases: "... **[a] variable may reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of the position or status, the variable may be 'tainted'.**" *Thus, if women are unfairly held back from promotion through the faculty ranks, then using faculty rank to adjust salary before comparing sexes may not be acceptable to the courts.* This suggests that an analysis comparing sexes but ignoring rank effects might be justifiable. What happens if this is done?

```
lm_faculty_sex_yd <-
  lm(
    salary ~ sex + yd
  , data = dat_faculty
  )
library(car)
Anova(lm_faculty_sex_yd, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: salary
##                 Sum Sq Df  F value    Pr(>F)
## (Intercept) 4275963832  1 231.4448 < 2.2e-16 ***
## sex           67178787  1   3.6362   0.06241 .
## yd           766344185  1  41.4799 4.883e-08 ***
## Residuals    905279453 49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_faculty_sex_yd)
```

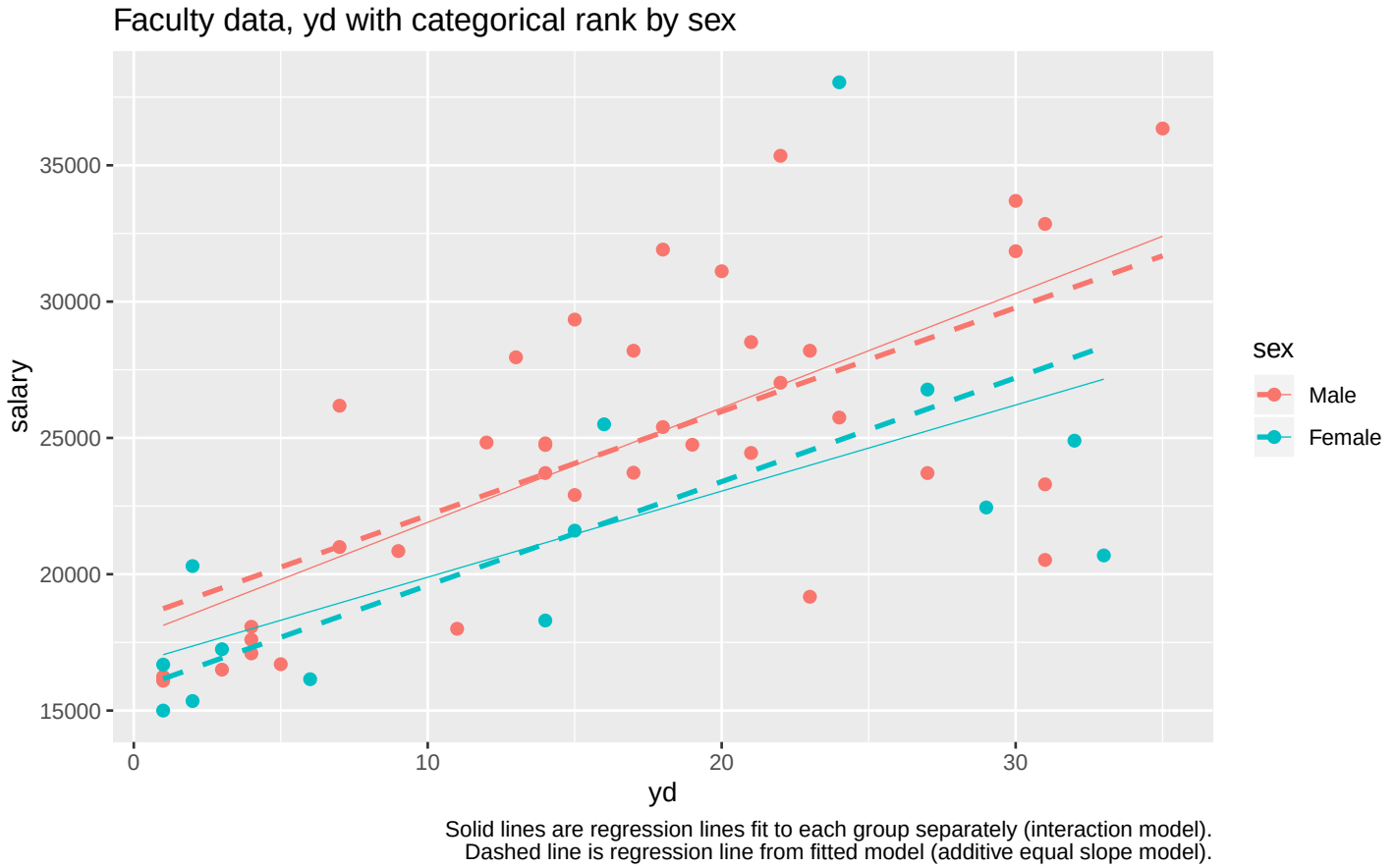
```
##
## Call:
## lm(formula = salary ~ sex + yd, data = dat_faculty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9631.7 -2529.4     3.5  2298.0 13125.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18355.23    1206.52  15.213  < 2e-16 ***
## sexFemale   -2572.53    1349.08  -1.907   0.0624 .
## yd            380.69      59.11   6.440 4.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4298 on 49 degrees of freedom
## Multiple R-squared:  0.493,Adjusted R-squared:  0.4724
## F-statistic: 23.83 on 2 and 49 DF,  p-value: 5.911e-08
```

```
# Put predicted values into the dataset for plotting
dat_faculty$pred <- predict(lm_faculty_sex_yd)
```

Parallel lines look reasonable. There are a few extreme salaries for Full professor rank; the effect they will have is to shift up the entire regression line and inflate the variance since they are in the middle of the range of the year variable with low leverage.

```
library(ggplot2)
p <- ggplot(dat_faculty, aes(x = yd, y = salary, colour = sex))
```

```
p <- p + geom_point(size = 2)
p <- p + geom_smooth(method = lm, se = FALSE, size = 1/4)
p <- p + geom_line(aes(y = pred), linetype = 2, size = 1)
p <- p + labs(
    title = "Faculty data, yd with categorical rank by sex"
  , caption = paste0( "Solid lines are regression lines fit to each group separately (interaction model).\n"
                    , "Dashed line is regression line from fitted model (additive equal slope model).")
  )
print(p)
```



Similar result as before, insufficient evidence between sexes (due to large proportion of variability in salary explained by yd [which I'm using in place of year since year is paired with rank]). Furthermore (not shown), there is insufficient evidence for a sex:yd interaction. However, rank and sex are (potentially) confounded. This data can not resolve this question. Instead, data on promotions would help resolve this issue.