# STAT 8020 R Lab 10: Multiple Linear Regression VI

*Whitney*

*September 21, 2020*

## Contents

### Regression with Both Quantitative and Qualitative Predictors: Salaries for Professors Data Set

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

#### Load and plot the data

```
library(carData)
```

```
## Warning: package 'carData' was built under R version 3.6.2
```

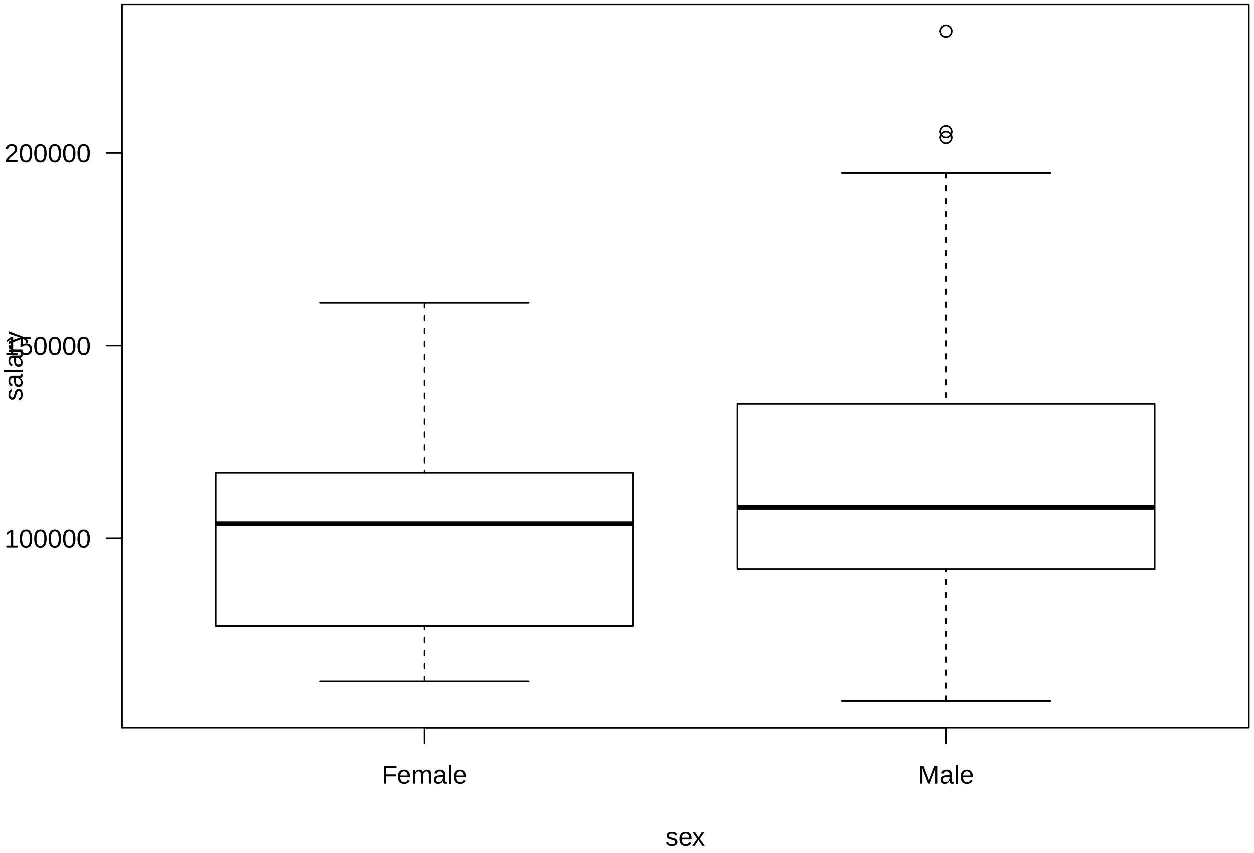
```
data("Salaries")
attach(Salaries)
head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

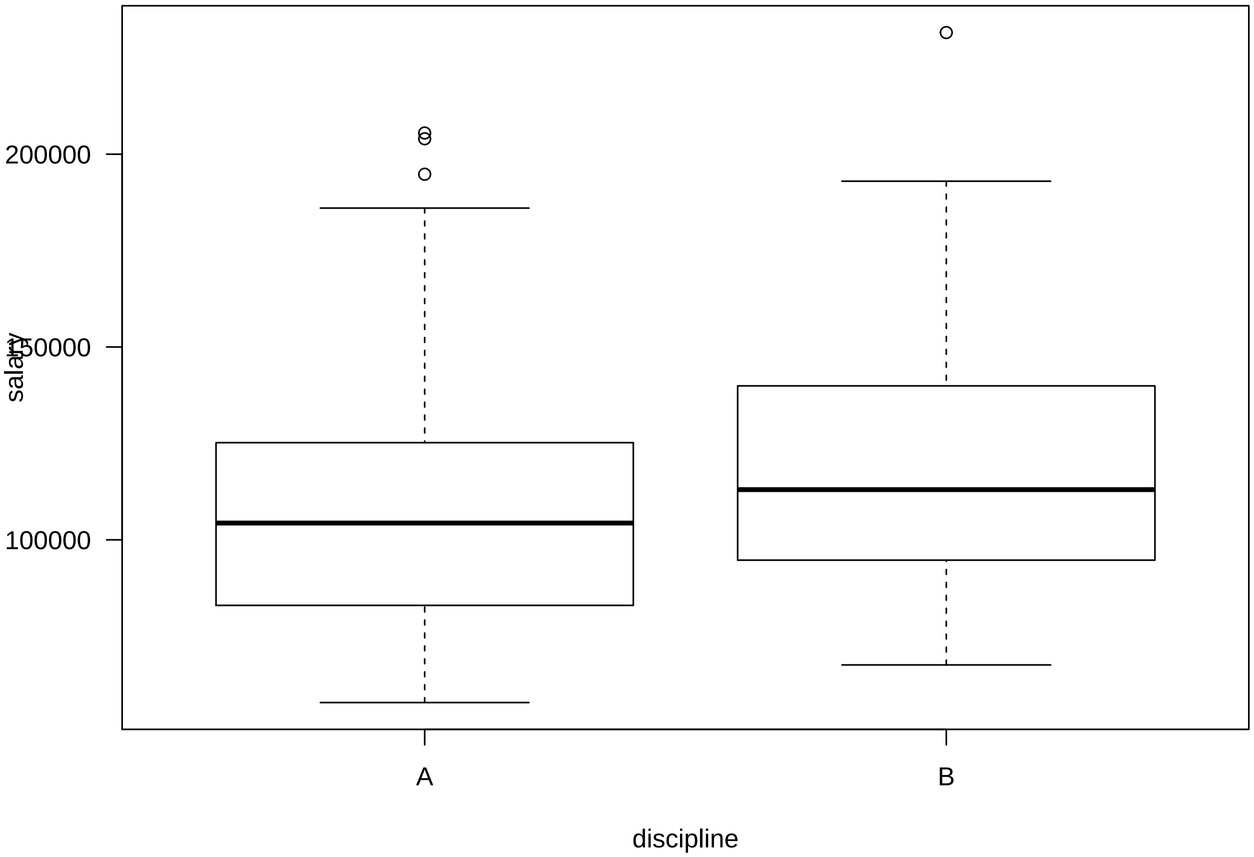
```
summary(Salaries)
```

```
##         rank     discipline yrs.since.phd    yrs.service        sex
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358
##  Prof     :266              Median :21.00   Median :16.00
##                             Mean   :22.31   Mean   :17.61
##                             3rd Qu.:32.00   3rd Qu.:27.00
##                             Max.   :56.00   Max.   :60.00
##      salary
##  Min.   : 57800
##  1st Qu.: 91000
##  Median :107300
##  Mean   :113706
##  3rd Qu.:134185
##  Max.   :231545
```

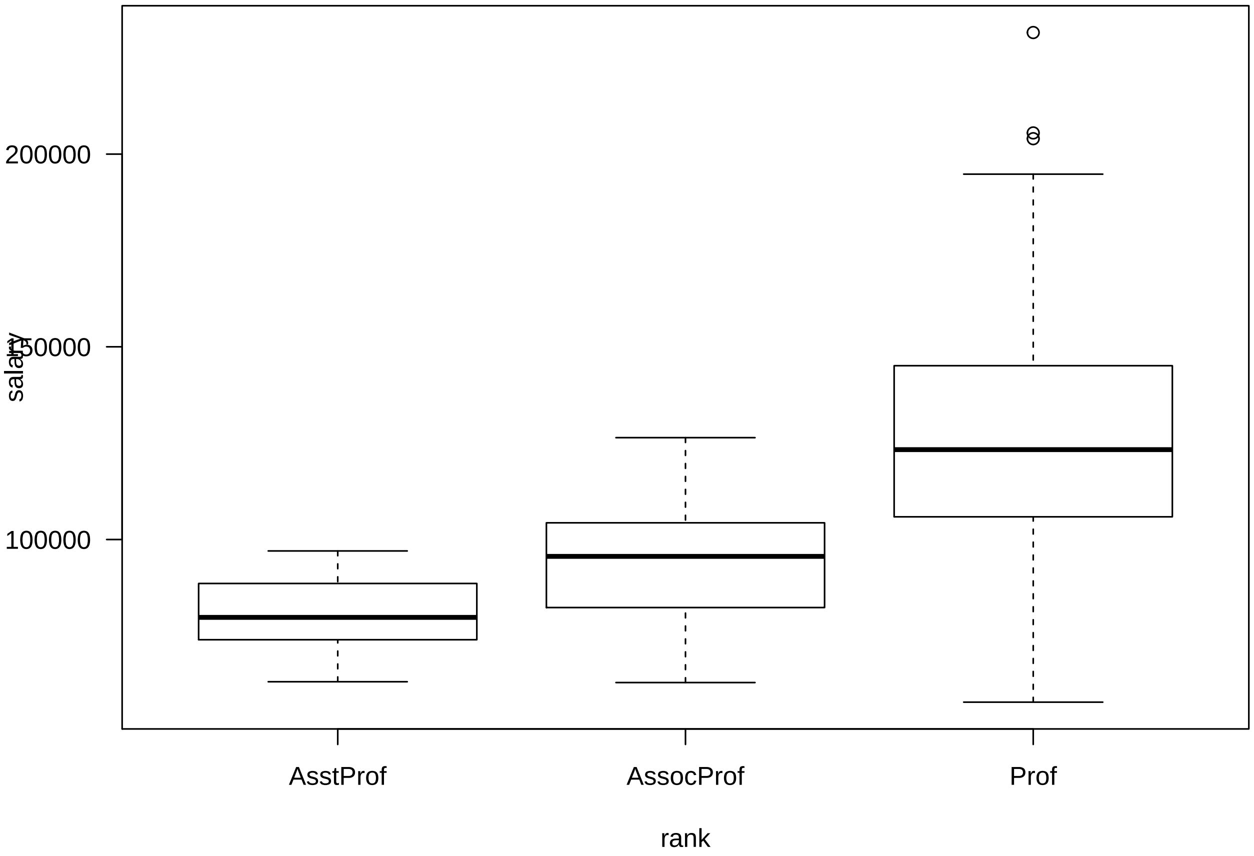
```
boxplot(salary ~ sex, data = Salaries, las = 1)
```



```
boxplot(salary ~ discipline, data = Salaries, las = 1)
```



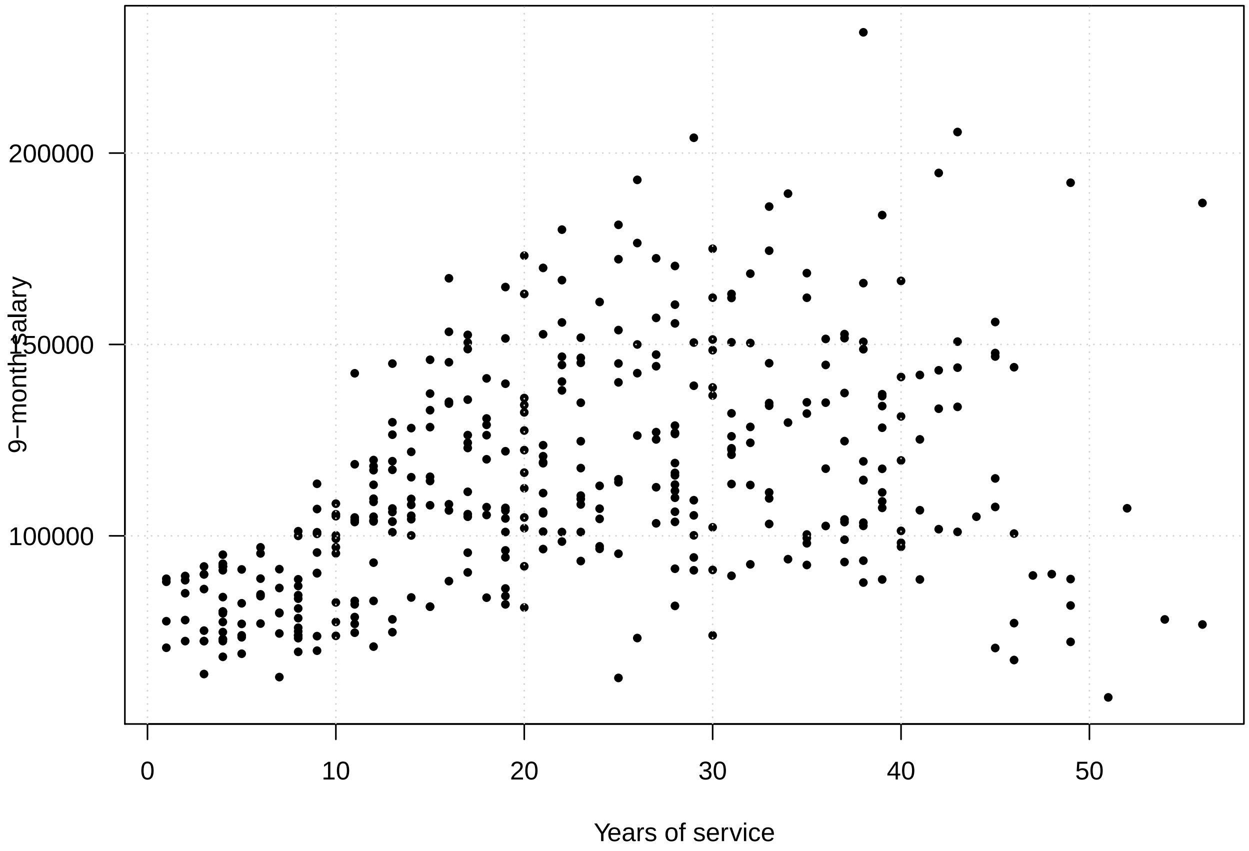
```
boxplot(salary ~ rank, data = Salaries, las = 1)
```



```r
xtabs(~ sex + rank + discipline, data = Salaries)
```

```
## , , discipline = A
##
##         rank
## sex      AsstProf AssocProf Prof
##   Female        6         4    8
##   Male         18        22  123
##
## , , discipline = B
##
##         rank
## sex      AsstProf AssocProf Prof
##   Female        5         6   10
##   Male         38        32  125
```

```r
plot(yrs.since.phd, salary, las = 1, pch = 16, cex = 0.75,
     xlab = "Years of service", ylab = "9-month salary")
grid()
```



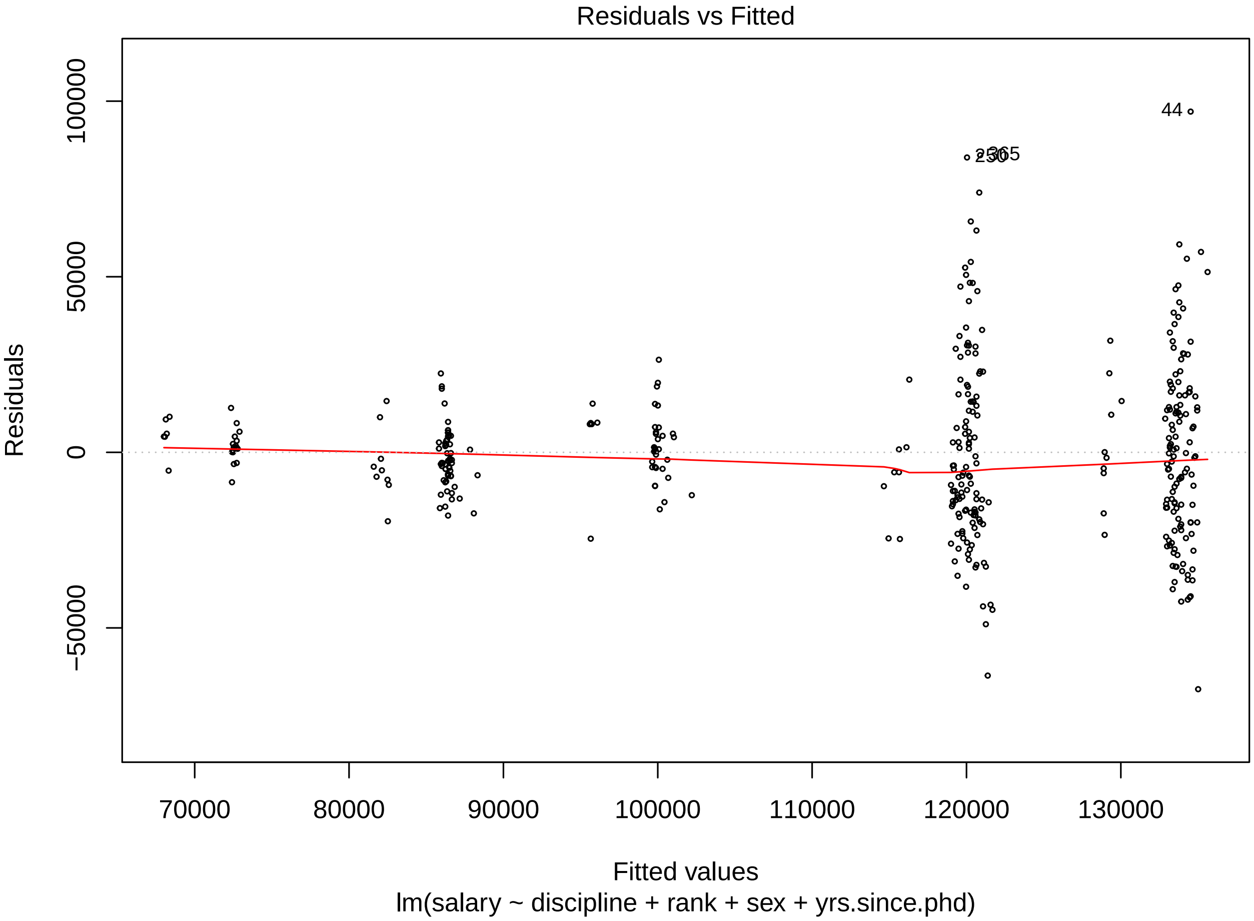
### Model fitting

```
m1 <- lm(salary ~ discipline + rank + sex + yrs.since.phd, data = Salaries)
X <- model.matrix(m1)
summary(m1)
```

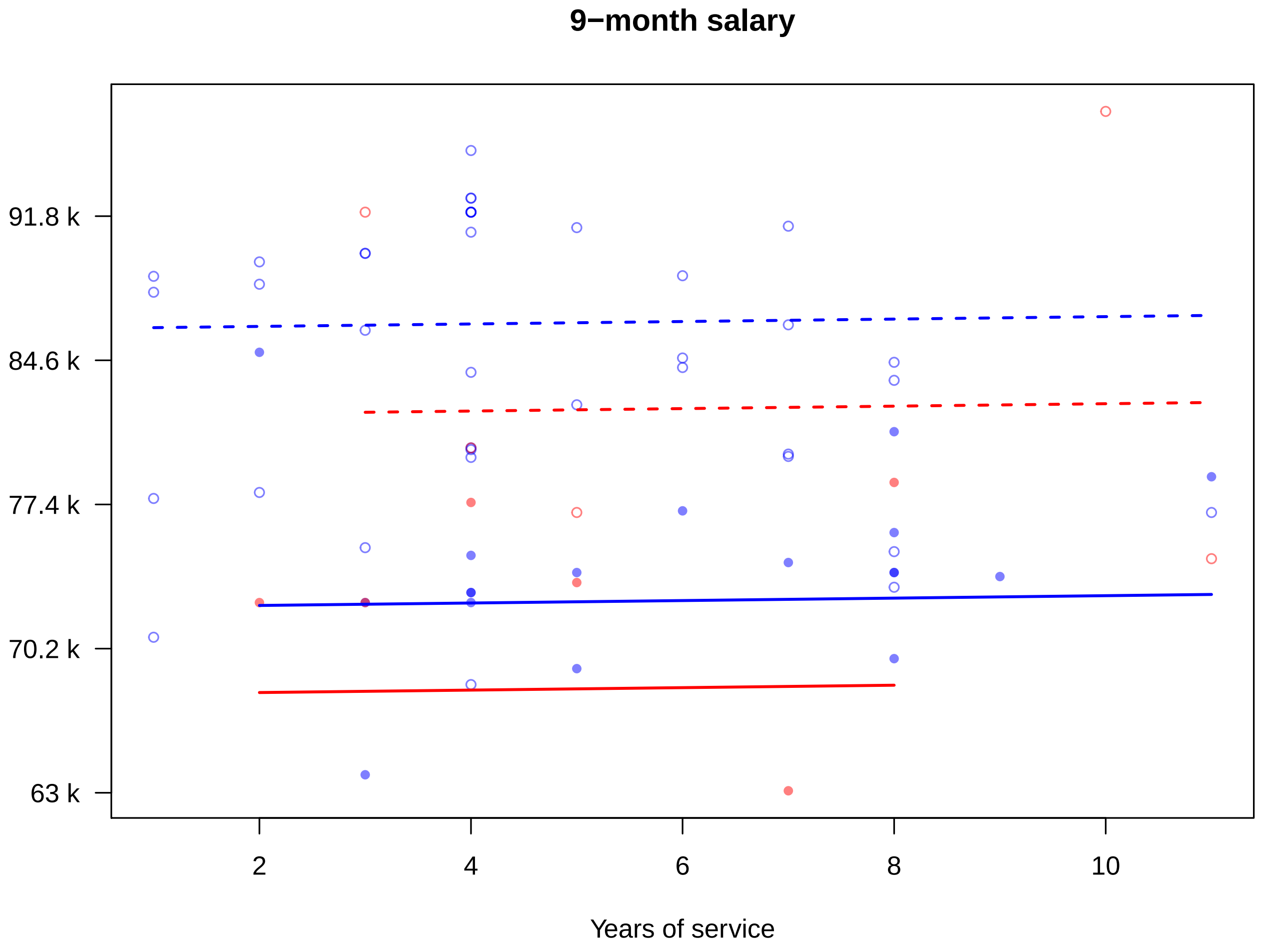
```
##
## Call:
## lm(formula = salary ~ discipline + rank + sex + yrs.since.phd,
##     data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67451 -13860  -1549  10716  97023
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67884.32    4536.89  14.963  < 2e-16 ***
## disciplineB   13937.47    2346.53   5.940 6.32e-09 ***
## rankAssocProf 13104.15    4167.31   3.145  0.00179 **
## rankProf      46032.55    4240.12  10.856  < 2e-16 ***
## sexMale        4349.37    3875.39   1.122  0.26242
## yrs.since.phd    61.01     127.01   0.480  0.63124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 22660 on 391 degrees of freedom
## Multiple R-squared:  0.4472, Adjusted R-squared:  0.4401
## F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

```r
plot(m1, which = 1, cex = 0.4)
```

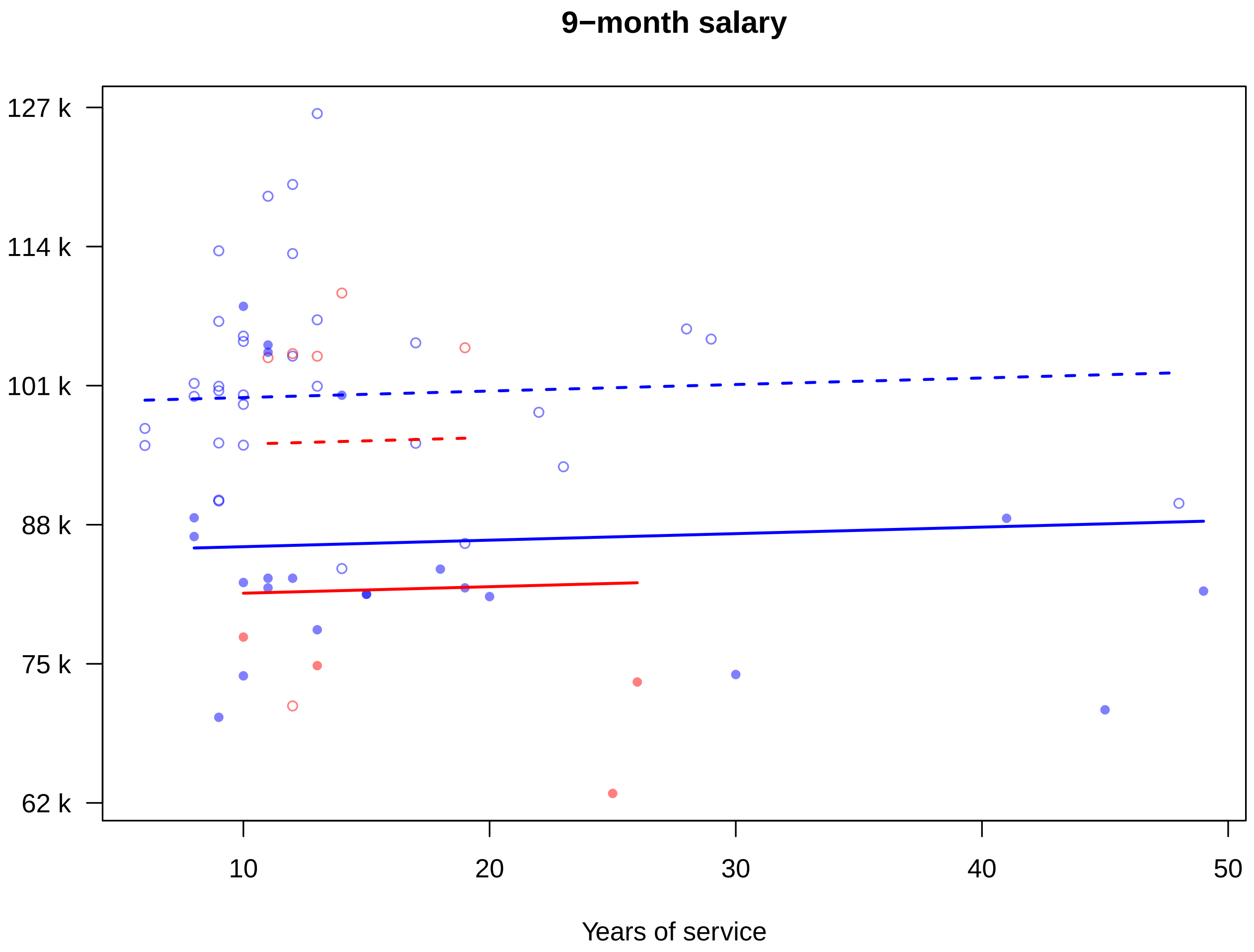


```r
yr.range <- tapply(yrs.since.phd, list(discipline, sex, rank), range)
sex.col <- ifelse(sex == "Male", "blue", "red")
dis.col <- ifelse(discipline == "A", 16, 1)

beta0 <- m1$coefficients[1]
betaDisp <- m1$coefficients[2]
betaAssoc <- m1$coefficients[3]
betaProf <- m1$coefficients[4]
betaMale <- m1$coefficients[5]
beta1 <- m1$coefficients[6]

library(scales)
# Plot the model fits by rank
assistant <- which(rank == "AsstProf")

plot(yrs.since.phd[assistant], salary[assistant],
     pch = dis.col[assistant], cex = 0.8,
     col = alpha(sex.col[assistant], 0.5),
     yaxt = "n", xlab = "Years of service",
     main = "9-month salary", ylab = "")
axis(2, at = seq(63000, 99000, len = 6),
     labels = paste(seq(63000, 99000, len = 6)/ 1000, "k"),
     las = 1)

segments(yr.range[[1]][1], beta0 + yr.range[[1]][1] * beta1,
         yr.range[[1]][2], beta0 + yr.range[[1]][2] * beta1,
         col = "red", lwd = 1.8)
segments(yr.range[[2]][1], beta0 + betaDisp + yr.range[[2]][1] * beta1,
         yr.range[[2]][2], beta0 + betaDisp + yr.range[[2]][2] * beta1,
         col = "red", lty = 2, lwd = 1.8)
segments(yr.range[[3]][1], beta0 + betaMale + yr.range[[3]][1] * beta1,
         yr.range[[3]][2], beta0 + betaMale + yr.range[[3]][2] * beta1,
         col = "blue", lwd = 1.8)
segments(yr.range[[4]][1], beta0 + betaDisp + betaMale + yr.range[[4]][1] * beta1,
         yr.range[[4]][2], beta0 + betaDisp + betaMale + yr.range[[4]][2] * beta1,
         col = "blue", lty = 2, lwd = 1.8)
```



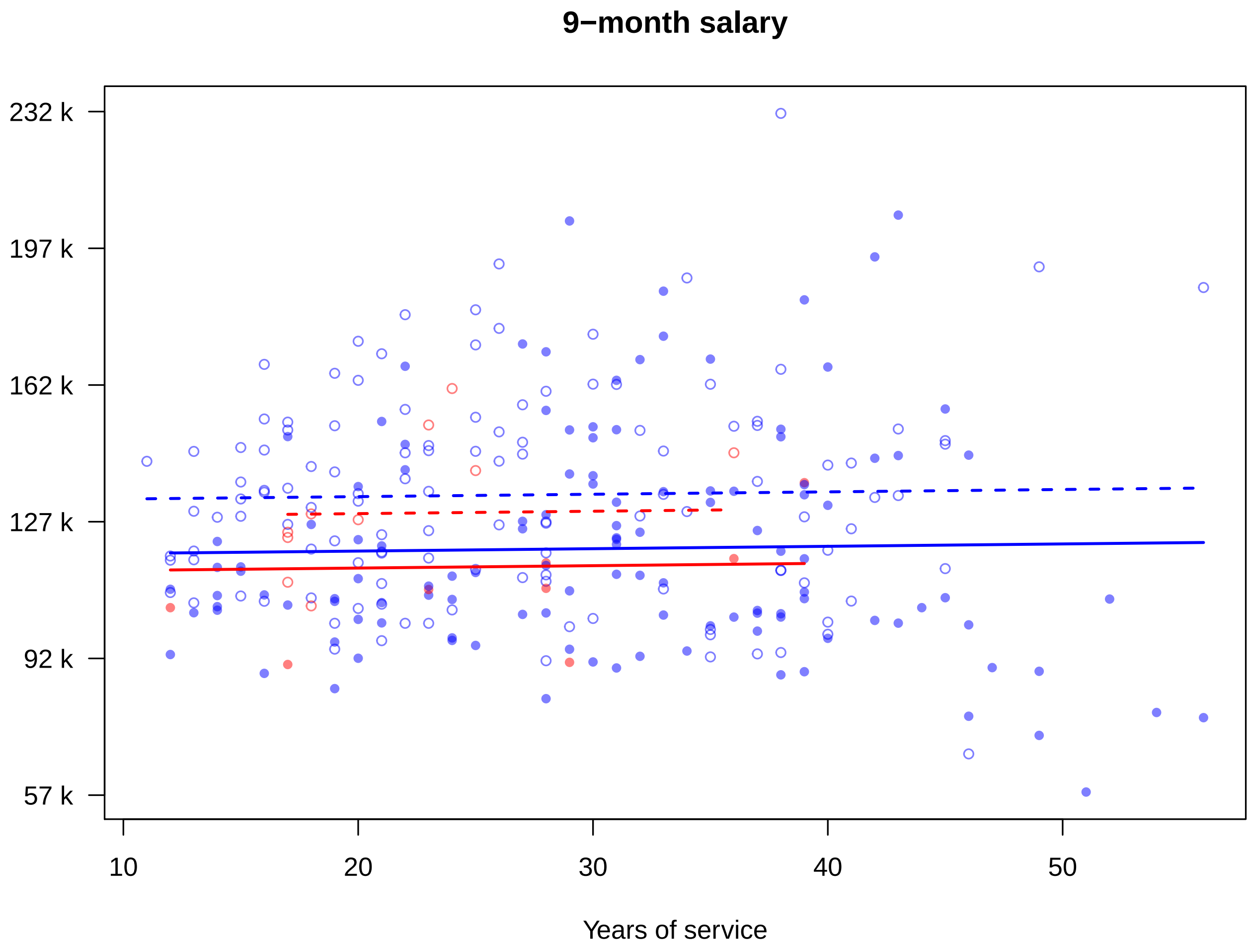
```
assoc <- which(rank == "AssocProf")
plot(yrs.since.phd[assoc], salary[assoc],
     pch = dis.col[assoc], cex = 0.8,
     col = alpha(sex.col[assoc], 0.5),
     yaxt = "n", xlab = "Years of service",
     main = "9-month salary", ylab = "")
```

```
axis(2, at = seq(62000, 127000, len = 6),
     labels = paste(seq(62000, 127000, len = 6)/ 1000, "k"),
     las = 1)

segments(yr.range[[5]][1], beta0 + betaAssoc + yr.range[[5]][1] * beta1,
         yr.range[[5]][2], beta0 + betaAssoc + yr.range[[5]][2] * beta1,
         col = "red", lwd = 1.8)
segments(yr.range[[6]][1], beta0 + betaDisp + betaAssoc + yr.range[[6]][1] * beta1,
         yr.range[[6]][2], beta0 + betaDisp + betaAssoc + yr.range[[6]][2] * beta1,
         col = "red", lty = 2, lwd = 1.8)
segments(yr.range[[7]][1], beta0 + betaAssoc + betaMale + yr.range[[7]][1] * beta1,
         yr.range[[7]][2], beta0 + betaAssoc + betaMale + yr.range[[7]][2] * beta1,
         col = "blue", lwd = 1.8)
segments(yr.range[[8]][1], beta0 + betaDisp + betaAssoc + betaMale + yr.range[[8]][1] * beta1,
         yr.range[[8]][2], beta0 + betaDisp + betaAssoc + betaMale + yr.range[[8]][2] * beta1,
         col = "blue", lty = 2, lwd = 1.8)
```



```
prof <- which(rank == "Prof")
plot(yrs.since.phd[prof], salary[prof],
     pch = dis.col[prof], cex = 0.8,
     col = alpha(sex.col[prof], 0.5),
     yaxt = "n", xlab = "Years of service",
     main = "9-month salary", ylab = "")
```

```
axis(2, at = seq(57000, 232000, len = 6),
     labels = paste(seq(57000, 232000, len = 6)/ 1000, "k"),
     las = 1)

segments(yr.range[[9]][1], beta0 + betaProf + yr.range[[9]][1] * beta1,
         yr.range[[9]][2], beta0 + betaProf + yr.range[[9]][2] * beta1,
         col = "red", lwd = 1.8)
segments(yr.range[[10]][1], beta0 + betaDisp + betaProf + yr.range[[10]][1] * beta1,
         yr.range[[10]][2], beta0 + betaDisp + betaProf + yr.range[[10]][2] * beta1,
         col = "red", lty = 2, lwd = 1.8)
segments(yr.range[[11]][1], beta0 + betaProf + betaMale + yr.range[[11]][1] * beta1,
         yr.range[[11]][2], beta0 + betaProf + betaMale + yr.range[[11]][2] * beta1,
         col = "blue", lwd = 1.8)
segments(yr.range[[12]][1], beta0 + betaDisp + betaProf + betaMale + yr.range[[12]][1] * beta1,
         yr.range[[12]][2], beta0 + betaDisp + betaProf + betaMale + yr.range[[12]][2] * beta1,
         col = "blue", lty = 2, lwd = 1.8)
```

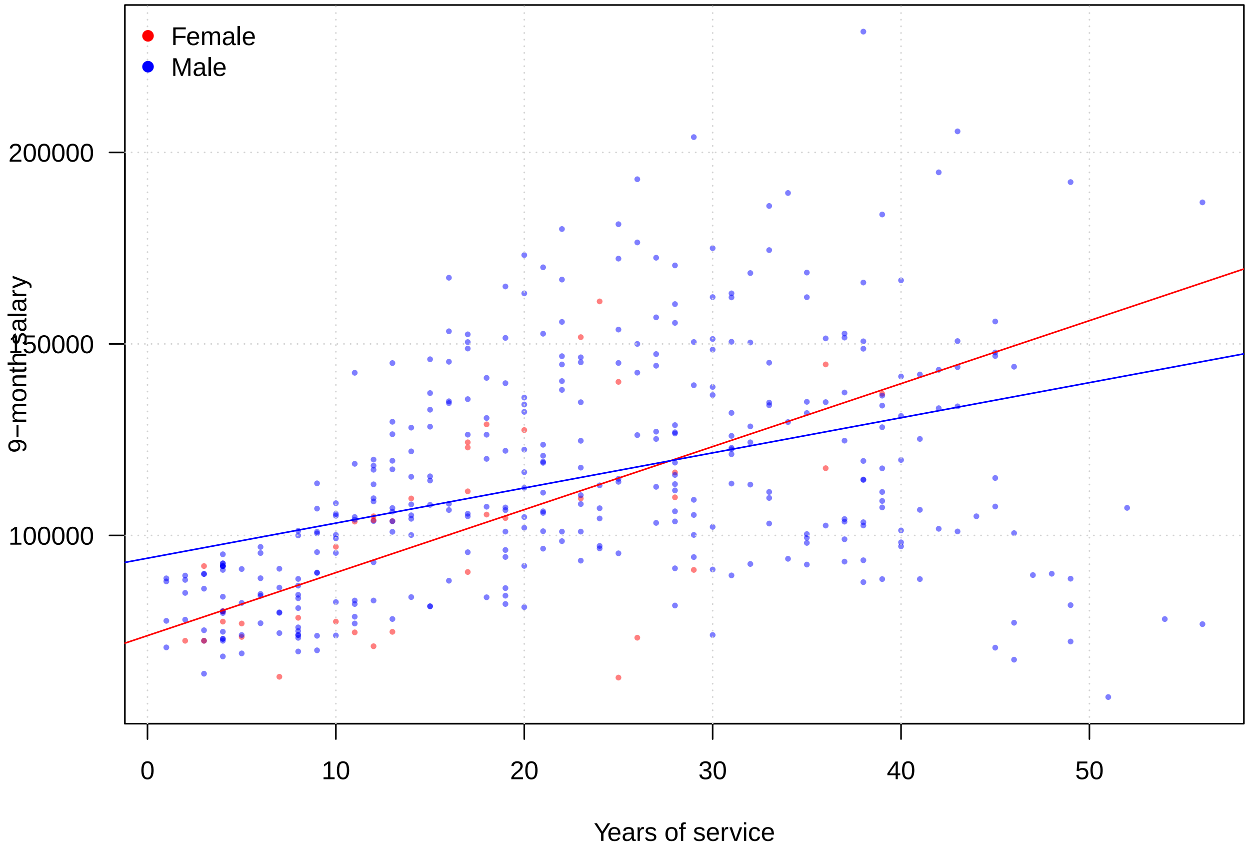


```
m2 <- lm(salary ~ sex * yrs.since.phd)
summary(m2)
```

```
##
## Call:
## lm(formula = salary ~ sex * yrs.since.phd)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -83012 -19442  -2988  15059 102652
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             73840.8     8696.7   8.491 4.27e-16 ***
## sexMale                 20209.6     9179.2   2.202 0.028269 *
## yrs.since.phd            1644.9      454.6   3.618 0.000335 ***
## sexMale:yrs.since.phd    -728.0      468.0  -1.555 0.120665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27420 on 393 degrees of freedom
## Multiple R-squared:  0.1867, Adjusted R-squared:  0.1805
## F-statistic: 30.07 on 3 and 393 DF,  p-value: < 2.2e-16
```

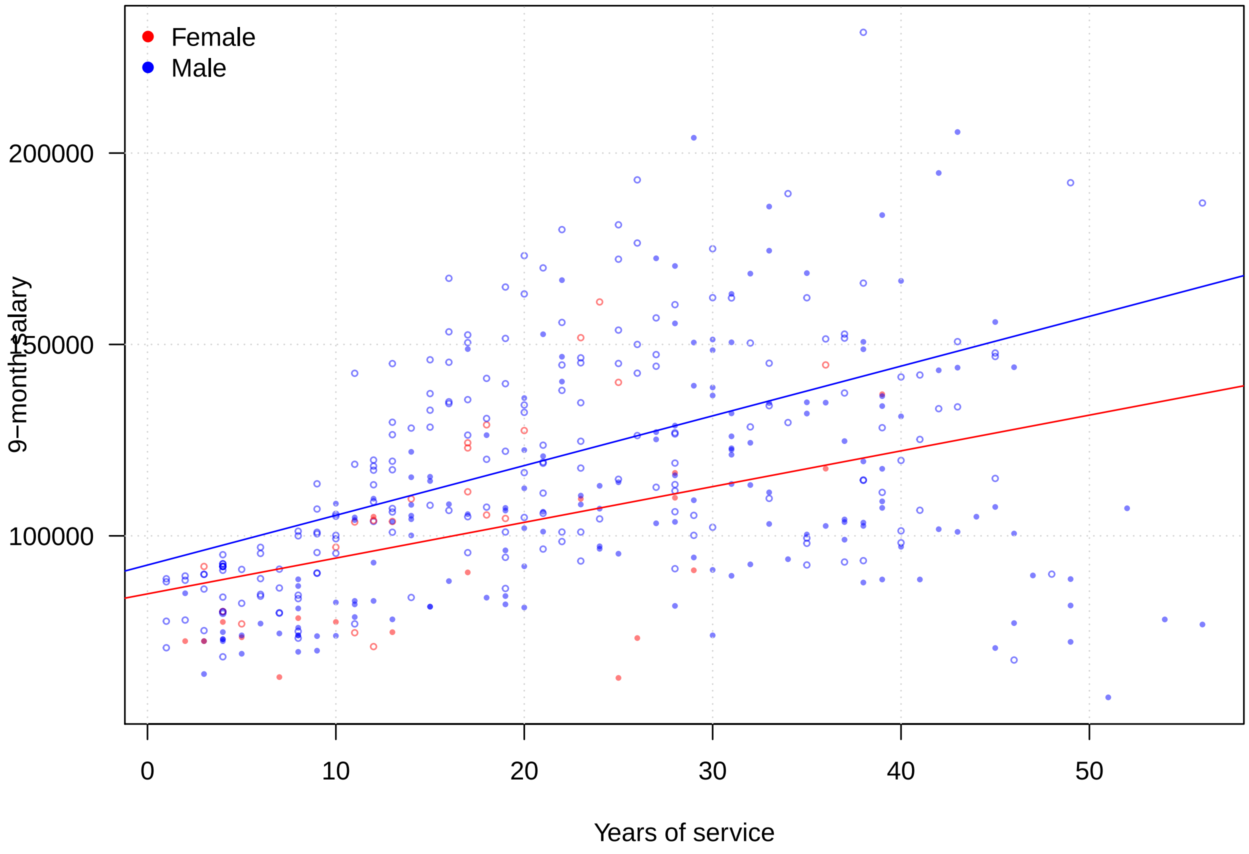
```r
coeff <- m2$coefficients
plot(yrs.since.phd, salary, las = 1, pch = 16, cex = 0.5, col = alpha(sex.col, 0.5),
     xlab = "Years of service", ylab = "9-month salary")
grid()
abline(coeff[1], coeff[3], col = "red")
abline(coeff[1] + coeff[2], coeff[3] + coeff[4],
       col = "blue")
legend("topleft", legend = c("Female", "Male"),
       pch = 16, col = c("red", "blue"),
       bty = "n")
```



```r
m3 <- lm(salary ~ discipline * yrs.since.phd)
summary(m3)
```

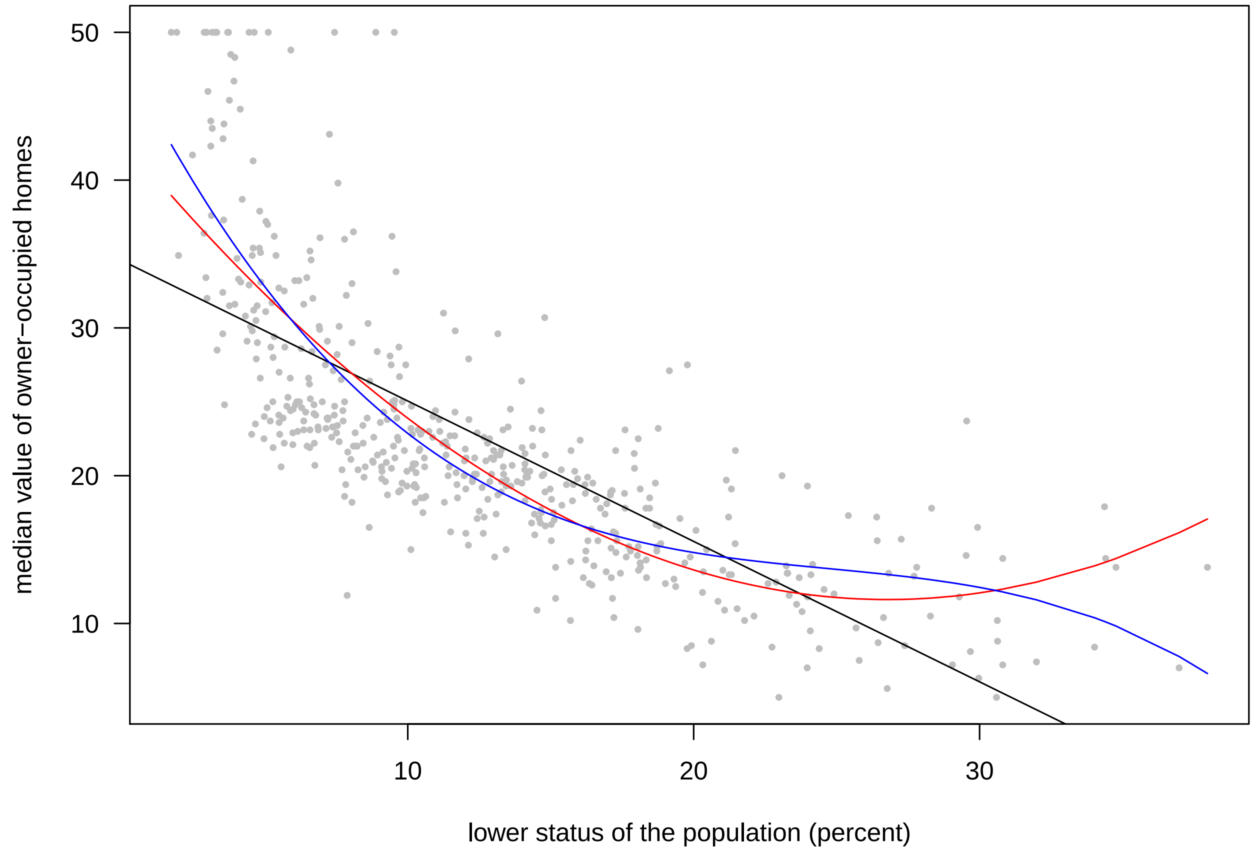
```
##
## Call:
## lm(formula = salary ~ discipline * yrs.since.phd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84580 -16974  -3620  15733  92072
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               84845.4     4283.9  19.806  < 2e-16 ***
## disciplineB                7530.0     5492.2   1.371   0.1711
## yrs.since.phd               933.9      150.0   6.225 1.24e-09 ***
## disciplineB:yrs.since.phd   365.3      211.0   1.731   0.0842 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26400 on 393 degrees of freedom
## Multiple R-squared:  0.2458, Adjusted R-squared:  0.2401
## F-statistic:  42.7 on 3 and 393 DF,  p-value: < 2.2e-16
```

```r
coeff <- m3$coefficients
plot(yrs.since.phd, salary, las = 1, pch = dis.col, cex = 0.5, col = alpha(sex.col, 0.5),
     xlab = "Years of service", ylab = "9-month salary")
grid()
abline(coeff[1], coeff[3], col = "red")
abline(coeff[1] + coeff[2], coeff[3] + coeff[4],
       col = "blue")
legend("toplef", legend = c("Female", "Male"),
       pch = 16, col = c("red", "blue"),
       bty = "n")
```



### Polynomial regression: Housing Values in Suburbs of Boston

```
library(MASS)
data(Boston)

plot(Boston$lstat, Boston$medv, col = "gray", pch = 16,
     cex = 0.6, las = 1, xlab = "lower status of the population (percent)", ylab = "median value of owne

m1 <- lm(medv ~ lstat, data = Boston)
abline(m1)

m2 <- lm(medv ~ lstat + I(lstat^2), data = Boston)
lines(sort(Boston$lstat), m2$fitted.values[order(Boston$lstat)], col = "red")

m3 <- lm(medv ~ lstat + I(lstat^2)+ I(lstat^3), data = Boston)
lines(sort(Boston$lstat), m3$fitted.values[order(Boston$lstat)], col = "blue")
```



```
anova(m2, m3)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ lstat + I(lstat^2)
## Model 2: medv ~ lstat + I(lstat^2) + I(lstat^3)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    503 15347
## 2    502 14616  1    731.76 25.134 7.428e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m2new <- lm(medv ~ poly(lstat, 2), data = Boston)
m3new <- lm(medv ~ poly(lstat, 3), data = Boston)
summary(m3new)
```

```
##
## Call:
## lm(formula = medv ~ poly(lstat, 3), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5441  -3.7122  -0.5145   2.4846  26.4153
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        22.5328     0.2399  93.937  < 2e-16 ***
## poly(lstat, 3)1 -152.4595     5.3958 -28.255  < 2e-16 ***
## poly(lstat, 3)2   64.2272     5.3958  11.903  < 2e-16 ***
## poly(lstat, 3)3  -27.0511     5.3958  -5.013 7.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.396 on 502 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.6558
## F-statistic: 321.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
anova(m2new, m3new)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ poly(lstat, 2)
## Model 2: medv ~ poly(lstat, 3)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    503 15347
## 2    502 14616  1    731.76 25.134 7.428e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```