6 Categorical explanatory variables

So far, we have treated the explanatory variables $X_j, j = 1, ..., p$, as continuous (numeric) variables. In many cases we would like to incorporate into the regression a *categorical variable* (in R, this is a variable of type 'factor'), i.e., a variable whose values ('levels' in R) indicate membership in one of several categories. For example: sex ("male", "female"); blood type ("A", "B", "AB", "O"); grape variety for wine ("cabernet sauvignon", "merlot", "pinot noir", "syrah", "tempranillo"). We can do this via coding with binary variables.

Suppose that one of the explanatory variables, say X_p, takes on values in $\{0, 1\}$. Recall that the mean of the response, under the linear model, is

$$\mathbb{E}Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j = \begin{cases} \beta_0 + \sum_{j=1}^{p-1} X_j, & X_p = 0 \\ \beta_0 + \beta_p + \sum_{j=1}^{p-1} X_j, & X_p = 1 \end{cases} .$$

That is, the effect of including a binary variable in the regression model is a shift of the intercept (each of $X_p = 0, X_p = 1$ has its own intercept). Specifically, β_0 is the value of the intercept for the category encoded $X_p = 0$, and $\beta_0 + \beta_p$ is the value of the intercept for the category encoded $X_p = 1$, so that β_p is the difference in the intercept values.

Dummy variables. If we have a categorical variable with only 2 categories, we can use a binary variable to represent it. If we have a categorical variable with more than 2 categories, we can encode it with a collection of corresponding binary variables, commonly called *dummy variables*.

Example. Consider a categorical variable indicating grape variety for wine,"cabernet sauvignon", "merlot", "pinot noir", "syrah", "tempranillo", 5 categories in total. We represent this with a collection of 4 dummy (binary) variables, $X_{p(k)}, k = 1, \ldots, 4$, corresponding to any 4 of the 5original categories; the remaining, left out category, is called the "reference" (or "baseline") category. In this way, only one of the dummy variables equals 1, and all the rest are zero, except when encoding the baseline category, in which case all of the dummies are zero.

	$X_{p(1)}$	$X_{p(2)}$	$X_{p(3)}$	$X_{p(4)}$
Cabernet Sauvignon	0	0	0	0
Merlot	1	0	0	0
Pinot Noir	0	1	0	0
Syrah	0	0	1	0
Tempranillo	0	0	0	1

In the example above, "cabernet sauvignon" was chosen as the baseline category, but this choice is arbitrary (R chooses this automatically, generally according to alphabetical order, which also determines which category is left out as baseline).

In general, if X_p (say) is a categorical variable with K categories (called "levels"), we use $K - 1$ dummy variables in the regression to encode it:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_{p(1)} X_{ip(1)} + \beta_{p(2)} X_{ip(2)} + \cdots + \beta_{p(K-1)} X_{ip(K-1)} + \sum_{j=1}^{p-1} \beta_j X_{ij}$$

This results in a linear mode with $(p - 1) + (K - 1)$ variables + intercept, i.e., $p + K - 1$ variables in total. The "effective" intercept of level k, for $k = 1, \ldots, K - 1$, is $\beta_0 + \beta_{p(k)}$, so that $\beta_{p(k)}$ is the difference in intercepts for the k-th category. For the baseline category the effective intercept is β_0, the general intercept.

Importantly, note that the slopes of the other $(p - 1)$ explanatory variables *remain the same* for all levels of the categorial variable.

Of course, we can include more than one categorical variable in a regression model. In that case, each of the categorical variables will have a baseline level, and the overall intercept will correspond to the combination of all baseline levels. Here is an example analyzed with R.

```
> # install package containing the dataset
> # install.packages("car")
>
> # load dataset
> salaries <- carData::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
>
> # extract variable names and types
> names(salaries)
[1] "rank"          "discipline"    "yrs.since.phd" "yrs.service"   "sex"
[6] "salary"
> str(salaries)
'data.frame': 397 obs. of  6 variables:
 $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ..
 $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
 $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
 $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
 $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
 $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 12
>
> # view dummy coding for the factor 'sex' and 'rank'
> contrasts(salaries$sex)
       Male
Female    0
Male      1
> contrasts(salaries$rank)
          AssocProf Prof
AsstProf          0    0
AssocProf         1    0
Prof              0    1
>
> # frequency table for sex
> table(salaries$sex)
```

```
Female   Male
    39    358
>
> ## regress y = salary on sex + yrs.service
> fm0 <- lm(salary ~ sex + yrs.service, data = salaries)
> summary(fm0)

Call:
lm(formula = salary ~ sex + yrs.service, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-81757 -20614  -3376  16779 101707

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  92356.9     4740.2  19.484  < 2e-16 ***
sexMale       9071.8     4861.6   1.866   0.0628 .
yrs.service    747.6      111.4   6.711 6.74e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28490 on 394 degrees of freedom
Multiple R-squared:  0.1198,Adjusted R-squared:  0.1154
F-statistic: 26.82 on 2 and 394 DF,  p-value: 1.201e-11

>
> # # obtain X matrix
> # head( model.matrix(fm) )
>
> # plot
> par(mar=c(3.2, 3.2, 3.2, 1.2), mgp=c(1.25,.5, 0), las=0)
> is.male <- salaries$sex=='Male'
> plot(salaries$yrs.service, salaries$salary, type='n', xaxt='n', yaxt='n',
ylab = 'salary', xlab = 'yrs.service', cex.lab=1,
main=expression("salary ~ sex + yrs.service"))
> points(salaries$yrs.service[is.male], salaries$salary[is.male])
> points(salaries$yrs.service[!is.male], salaries$salary[!is.male], pch=1, col='red')
> legend('topright', legend=c('Male', 'Female'), col=c("black", "red"),
pch=c(1,1), cex=1, y.intersp=1)
> # add fitted lines for 'Male', 'Female'
> coef(fm0)
(Intercept)     sexMale yrs.service
 92356.9467   9071.8000    747.6121
> beta_0 <- coef(fm0)[1]
> beta_Male <- coef(fm0)[2]
> beta_1 <- coef(fm0)[3]
```

```
> abline(beta_0+beta_Male, beta_1) # Male
> abline(beta_0, beta_1, col='red') # Female
>
> # find predicted (fitted) mean salary for a male with 3 years of service
> predict(fm0, data.frame(sex = 'Male', yrs.service = 3))
       1
103671.6
> # same as
> beta_0 + beta_Male + beta_1*3
(Intercept)
   103671.6
>
```

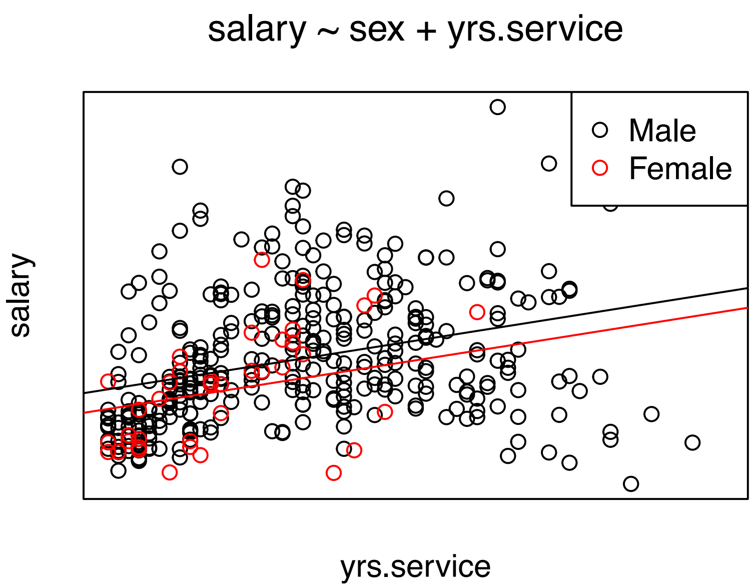

Figure 6: Regression of Salary on Sex + Yrs.Service. Straight lines are the fitted regression lines corresponding to male and female, respectively. Note that, by design, these lines are parallel.

```
> # Important: we fitted an ADDITIVE model -- this is NOT the same as fitting
two separate simple regressions:
> par(mfrow=c(1,2))
> par(mar=c(3.2, 3.2, 3.2, 1.2), mgp=c(1.25,.5, 0), las=0)
> # simple reg for 'Male'
> fm0.male <- lm(salary[is.male] ~ yrs.service[is.male], data = salaries)
> plot(salaries$yrs.service, salaries$salary, type='n', xaxt='n', yaxt='n',
ylab = 'salary', xlab = 'yrs.service', cex.lab=1,
```

```
main=expression("salary ~ yrs.service: Male"), cex.main=.8)
> points(salaries$yrs.service[is.male], salaries$salary[is.male])
> coef(fm0.male)[2]  # coef of "yrs.service[is.male]" is 705.6, compare
to 747.6 in the 'additive' model fm0
yrs.service[is.male]
            705.5634
> abline(fm0.male)
>
> fm0.female <- lm(salary[!is.male] ~ yrs.service[!is.male], data = salaries)
> plot(salaries$yrs.service, salaries$salary, type='n', xaxt='n',
yaxt='n', ylab = 'salary',
xlab = 'yrs.service', cex.lab=1,
main=expression("salary ~ yrs.service: Female"), , cex.main=.8, col='red')
> points(salaries$yrs.service[!is.male], salaries$salary[!is.male], col='red')
> coef(fm0.female)[2]  # coef of "yrs.service[is.male]" is 705.6, compare
to 747.6 in the 'additive' model fm0
yrs.service[!is.male]
               1637.3
> abline(fm0.female, col='red')
```

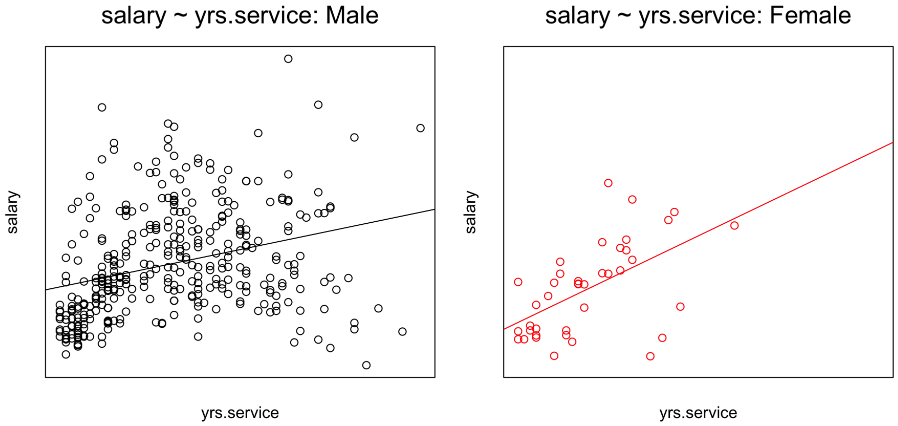

Figure 7: Regression of Salary on Yrs.Service, fitted separately for males and females. Straight lines are the fitted simple regression lines. Here these two lines are not parallel.

```
> # Now add ``rank" to the regression:

> # add another factor: regress y = salary on sex + rank + yrs.service
> # frequency table for rank*sex
> table(salaries$rank, salaries$sex)

            Female Male
  AsstProf      11   56
  AssocProf     10   54
```

```
  Prof          18  248
>
> fm1 <- lm(salary ~ sex + rank + yrs.service, data = salaries)
> summary(fm1)$coef
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   76612.810  4426.0007 17.309715 2.847735e-50
sexMale        5468.708  4035.3366  1.355205 1.761327e-01
rankAssocProf 14702.856  4266.5563  3.446071 6.303299e-04
rankProf      48980.224  3991.8299 12.270118 1.635066e-29
yrs.service    -171.792   115.2707 -1.490335 1.369404e-01
> # the general intercept 76612.810 is the 'effective' intercept for
female assistant professor
> # the 'effective' intercept for a female professor will
be 76612.810 + 48980.224

> # let's write down the fitted model
> # Extract coefficients
> coef(fm1)
  (Intercept)       sexMale rankAssocProf      rankProf   yrs.service
    76612.810      5468.708     14702.856     48980.224      -171.792

> # so the estimate model is

> # Y_hat = 76612.8 + 5468.7 * sexMale + 14702.86 * rankAssocProf +
48980.2 * rankProf -171.792 * yrs.service

> # E.g. the (effective) intercept for a female assistant Professor is 76612.8

> # E.g. the predicted salary of (=the estimate of the mean salary for) a
male prof with 3 years of service is:

 Y_hat = 76612.8 + 5468.7 + 48980.2 * rankProf -171.792 * 3
```

Important remarks. (1) This model allows a different intercept for each of the $2 * 3 = 6$ categories, but, importantly, it imposes an *additive* structure: the change in the intercept term (hence, in the predicted value of Y) when we compare 'Assistant Professor' to 'Professor' is the same whether we're looking at a male or a female (similarly, the change in the intercept when moving from 'Male' to 'Female' is the same when considering an 'Assistant Professor' or a 'Professor'). (2) Also, note that the LS estimates for the general intercept β_0, and for the coefficient of the dummy variable "sexMale" and of the continuous variable "yrs.service", are different from the values in the previously fitted model (without "rank"); in fact, the coefficient of "yrs.service" even changes sign!

7 Interactions

An *interaction* term (variable) for two explanatory variables, say X_p and X_j, is a new explanatory variable given by $X_{\text{new}} = X_p X_j$. Including interaction variables *allows the effect of one variable to depend on the* *value of another variable*. Suppose we start with a regression model with 3 explanatory variables,

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}.$$

In this model, the interpretation of the third coefficient (for example) is as follows: β_3 is the increase in the expectation (population mean) of Y per unit increase in X_3, if we hold the values of X_1, X_2 fixed. I.e., this is the increase in the mean response value 'conditional' on $X_1 = x_1, X_2 = x_2$, and note that here this increase-per-unit is *the same* no matter the specific values x_1, x_2 that we condition on.

Now let's add an interaction between X_2 and X_3, i.e., add the new variable $X_{i4} = X_{i2}X_{i3}$:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4}.$$

If X_2 and X_3 are both binary, the interaction has an effect on the effective intercept only:

$$\text{intercept}_i = \begin{cases} \beta_0, & \text{if } X_{i2} = 0, X_{i3} = 0 \\ \beta_0 + \beta_2, & \text{if } X_{i2} = 1, X_{i3} = 0 \\ \beta_0 + \beta_3, & \text{if } X_{i2} = 0, X_{i3} = 1 \\ \beta_0 + \beta_2 + \beta_3 + \beta_4, & \text{if } X_{i2} = 1, X_{i3} = 1 \end{cases}$$

If X_2 is binary and X_3 is continuous:

$$\text{slope of } X_{i3} = \begin{cases} \beta_3, & \text{if } X_{i2} = 0 \\ \beta_3 + \beta_4, & \text{if } X_{i2} = 1 \end{cases}$$

Finally, if X_2 and X_3 are both continuous:

$$\begin{array}{ll} \text{slope of } X_{i2} = \beta_2 + \beta_4 u, & \text{if } X_{i3} = u \\ \text{slope of } X_{i3} = \beta_3 + \beta_4 v, & \text{if } X_{i2} = v \end{array}$$

To demonstrate this, let's return to the salaries dataset. Y = salary. X_1 = yrs.service. X_2 = sex.

Regress Y = salary on X_1 = yrs.service:

$$Y_i = \beta_0 + \beta_1 \times \text{ yrs.service } + \epsilon_i.$$

```
> # regress salary on yrs.service
> fm0 <- lm(salary ~ yrs.service, data=salaries)
> summary(fm0)

Call:
lm(formula = salary ~ yrs.service, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-81933 -20511  -3776  16417 101947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99974.7     2416.6   41.37  < 2e-16 ***
```

```
yrs.service    779.6      110.4    7.06 7.53e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28580 on 395 degrees of freedom
Multiple R-squared:  0.1121,Adjusted R-squared:  0.1098
F-statistic: 49.85 on 1 and 395 DF,  p-value: 7.529e-12

> fm0$coefficients
(Intercept) yrs.service
 99974.6529    779.5691
>
> # plot
> # par(mfrow=c(1,3))
> par(mar=c(3.2, 3.2, 2.2, 1.2), mgp=c(1.25, .5, 0), las=0)
> plot(salaries$yrs.service, salaries$salary, type='n', xaxt='n', yaxt='n',
ylab = 'salary', xlab = 'yrs.service', cex.lab=1, main=
expression('salary ~ yrs.service'), cex.main=1.5)
> points(salaries$yrs.service, salaries$salary)
> abline(fm0, col='black', lwd=1.5)
>
```

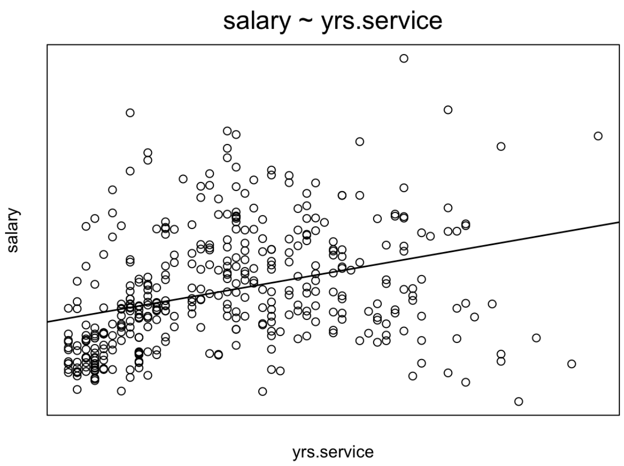

Figure 8: Simple regression of Salary on Years of service

Now, regress Y = salary on X_1 = yrs.service and X_2 = sex, no interaction:

$$Y_i = \beta_0 + \beta_1 \times \text{yrs.service} + \beta_2 \times (\text{sex} = \text{Male}) + \epsilon_i$$

```
> # regress salary on yrs.service + sex, no interaction
> fm1 <- lm(salary ~ yrs.service + sex, data=salaries)
> summary(fm1)
```

```
Call:
lm(formula = salary ~ yrs.service + sex, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-81757 -20614  -3376  16779 101707

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  92356.9     4740.2  19.484  < 2e-16 ***
yrs.service    747.6      111.4   6.711 6.74e-11 ***
sexMale       9071.8     4861.6   1.866   0.0628 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28490 on 394 degrees of freedom
Multiple R-squared:  0.1198,Adjusted R-squared:  0.1154
F-statistic: 26.82 on 2 and 394 DF,  p-value: 1.201e-11

> fm1$coefficients
(Intercept) yrs.service     sexMale
 92356.9467    747.6121   9071.8000
> # plot
> plot(salaries$yrs.service, salaries$salary, type='n', xaxt='n',
yaxt='n', ylab = 'salary', xlab = 'yrs.service', cex.lab=1,
main=expression('salary ~ yrs.service + sex'), cex.main=1.5)
> points(salaries$yrs.service, salaries$salary)
> abline(92356.9467, 747.6121, col='red', lwd=1.5)
> abline(92356.9467 + 9071.8000, 747.6121, col='black', lwd=1.5)
>
> # prediced mean salary:
> # Male: 92356.9467 + 9071.8000 + 747.6121 * yrs.service
> # Female: 92356.9467 + 747.6121 * yrs.service
>
```

Figure 9: Regression of "yrs.service" on "sex", no interaction

Finally, regress Y = salary on X_1 = yrs.service and X_2 = sex, with interaction:

$$Y_i = \beta_0 + \beta_1 \times \text{yrs.service} + \beta_2 \times (\text{sex = Male}) + \epsilon_i + \beta_3 \times \text{yrs.service} \times (\text{sex=Male}) + \epsilon_i.$$

```
> # Regress salary on yrs.service + sex, with interaction
> fm2 <- lm(salary ~ yrs.service + sex + yrs.service:sex, data=salaries)
> summary(fm2)

Call:
lm(formula = salary ~ yrs.service + sex + yrs.service:sex, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-80381 -20258  -3727  16353 102536

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          82068.5     7568.7  10.843  < 2e-16 ***
yrs.service           1637.3      523.0   3.130  0.00188 **
sexMale              20128.6     7991.1   2.519  0.01217 *
yrs.service:sexMale   -931.7      535.2  -1.741  0.08251 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28420 on 393 degrees of freedom
Multiple R-squared:  0.1266,Adjusted R-squared:  0.1199
F-statistic: 18.98 on 3 and 393 DF,  p-value: 1.622e-11
```

```
> fm2$coefficients
        (Intercept)          yrs.service            sexMale yrs.service:sexMale
         82068.5087            1637.2997         20128.6258           -931.7363
> plot(salaries$yrs.service, salaries$salary, type='n', xaxt='n',
yaxt='n', ylab = 'salary', xlab = 'yrs.service', cex.lab=1, main=
'salary ~ yrs.service + sex + yrs.service:sex', cex.main=1.5)
> points(salaries$yrs.service, salaries$salary)
> abline(82068.5087 + 20128.6258, 1637.2997-931.7363, lwd=1.5) #males
> abline(82068.5087 , 1637.2997, lwd=1.5, col='red') #females
```

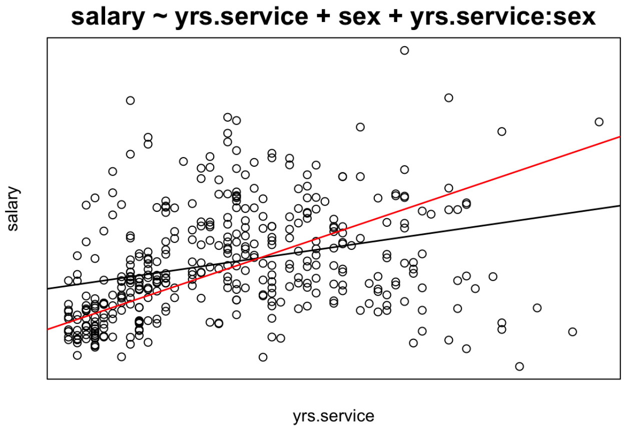

Figure 10: Regression of Salary on Years of service and Sex, with interaction

How to evaluate the significance of interaction term, i.e. whether the effect of 'yrs.service' differs significantly between males and females or not? Look at the p-value for the interaction term. In the output below this is ≈ 0.083, so, e.g., it is significant at $\alpha = 0.1$ level.

```
> # regress salary on yrs.service + sex, with interaction
> fm2 <- lm(salary ~ yrs.service + sex + yrs.service:sex, data=salaries)
> summary(fm2)

Call:
lm(formula = salary ~ yrs.service + sex + yrs.service:sex, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-80381 -20258  -3727  16353 102536

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          82068.5     7568.7  10.843  < 2e-16 ***
```

```
yrs.service           1637.3      523.0   3.130  0.00188 **
sexMale              20128.6     7991.1   2.519  0.01217 *
yrs.service:sexMale   -931.7      535.2  -1.741  0.08251 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28420 on 393 degrees of freedom
Multiple R-squared:  0.1266,Adjusted R-squared:  0.1199
F-statistic: 18.98 on 3 and 393 DF,  p-value: 1.622e-11
```