Models with categorical factors

Peter Hoff

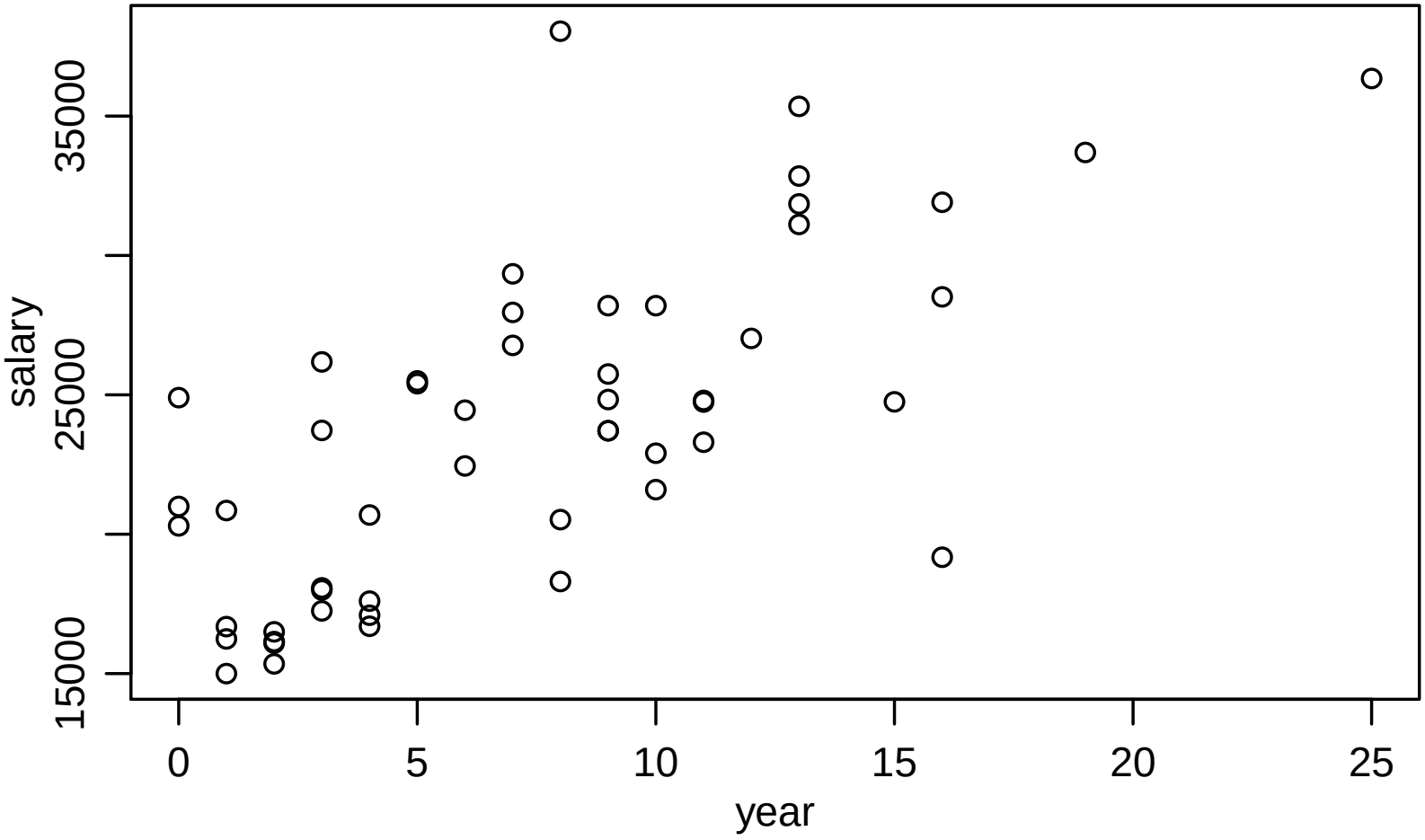
STAT 423

Applied Regression and Analysis of Variance

University of Washington

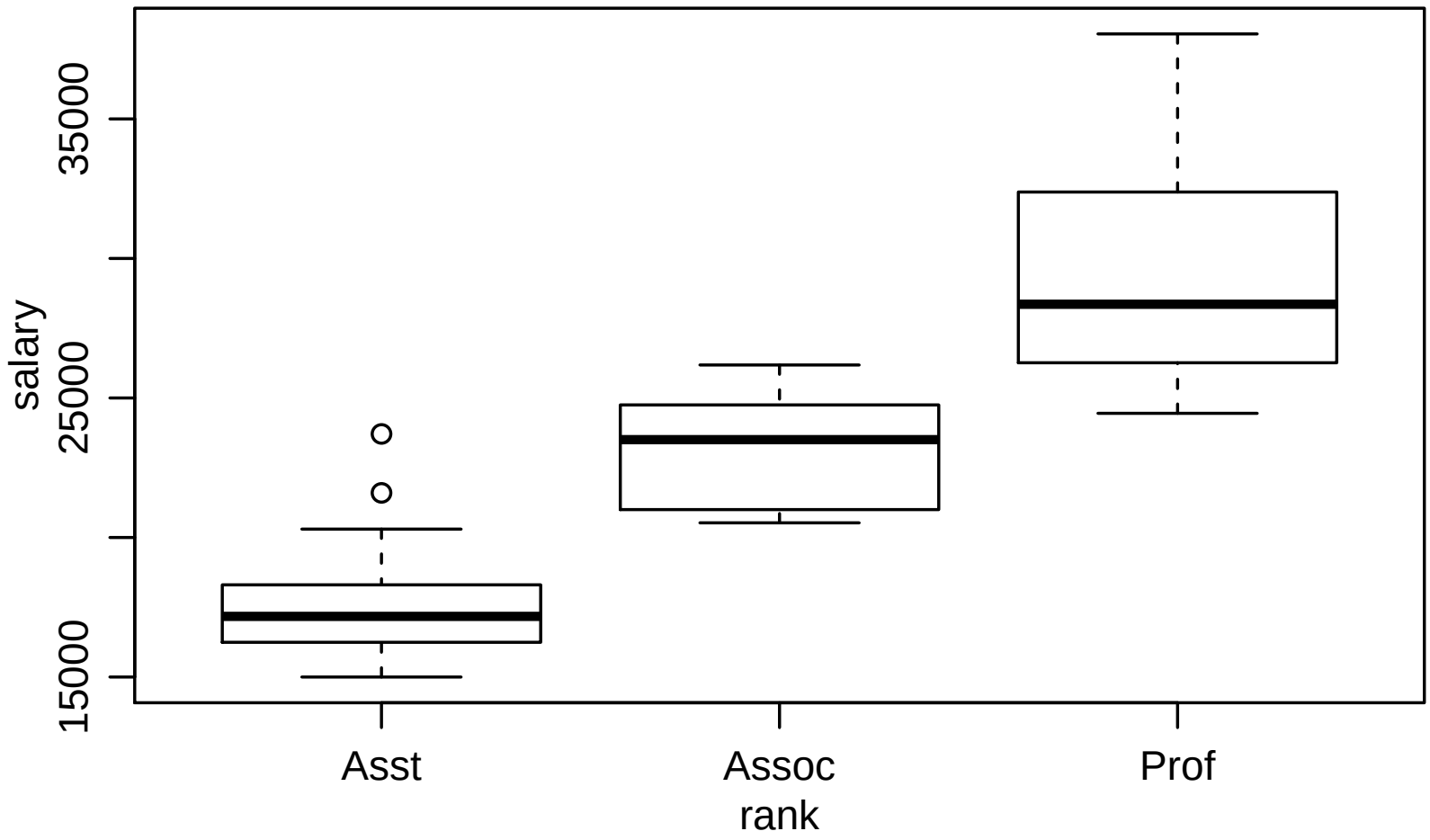
Faculty salary data

```
colnames(salary)

## [1] "degree" "rank"   "sex"    "year"   "ysdeg"  "salary"
```



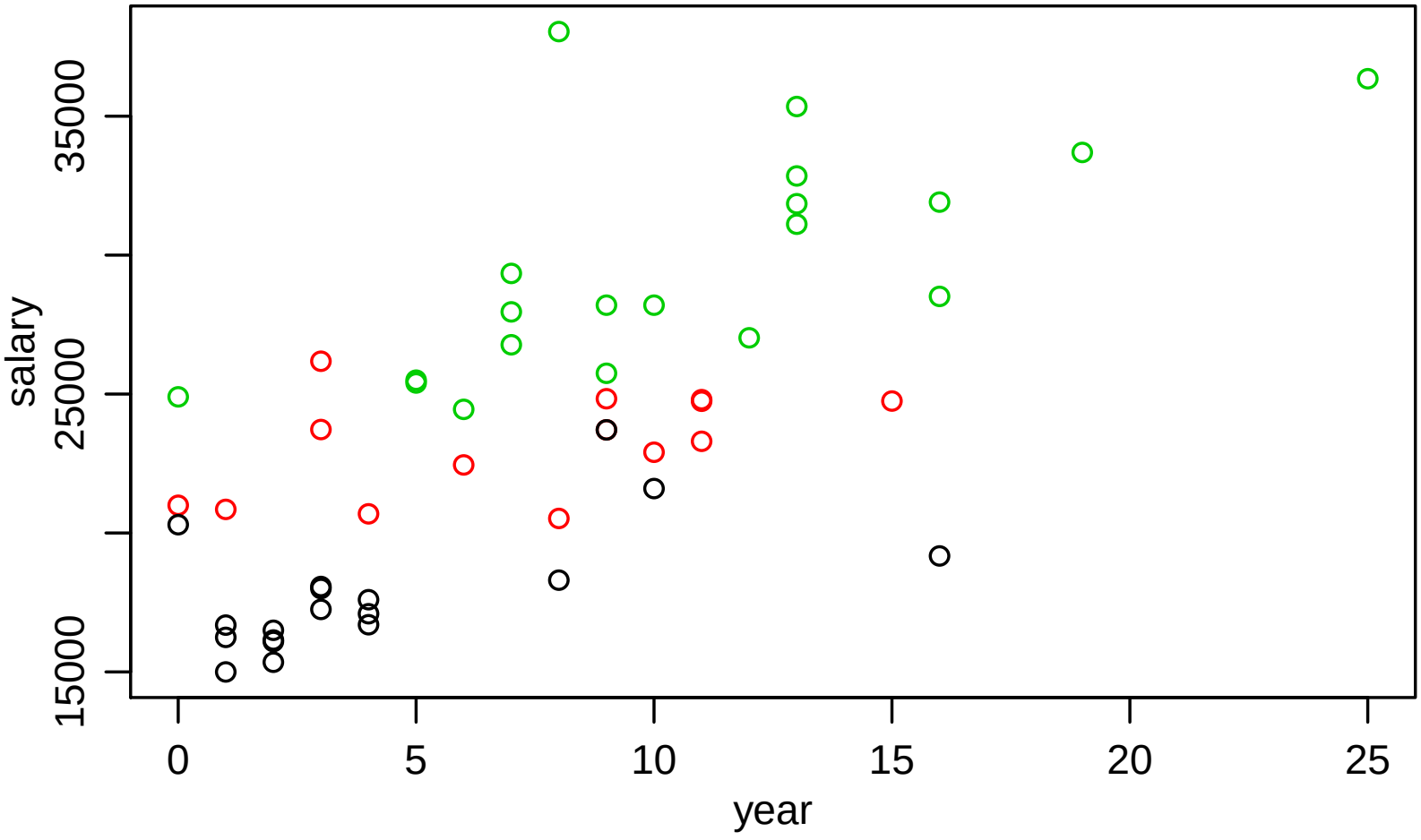

```
summary(lm(salary~year,data=salary))$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 18166.1475  1003.6582 18.099935 1.343049e-23
## year          752.7978   108.4092  6.944039 7.341379e-09
```



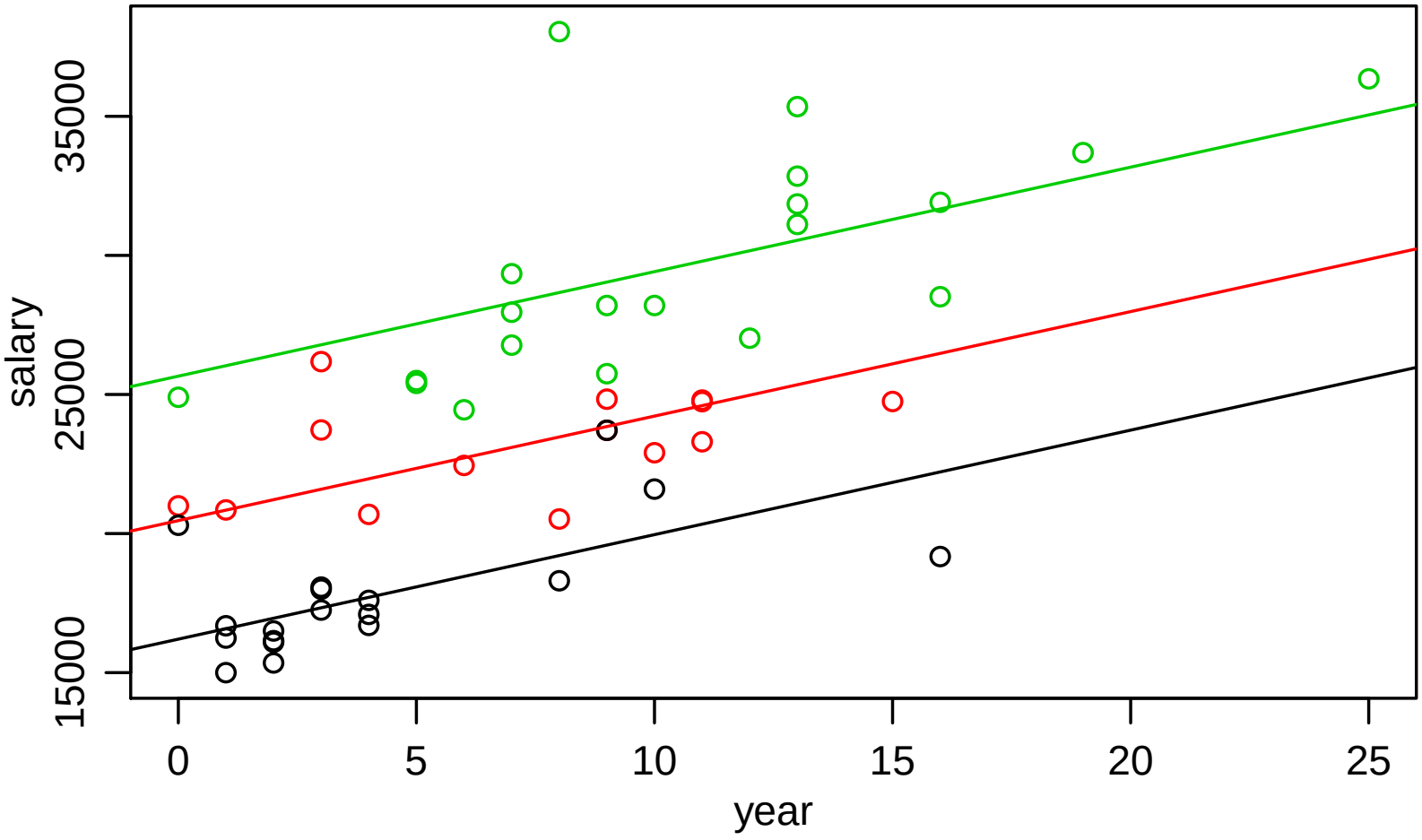

```
summary(lm(salary~rank,data=salary))$coef
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 17768.667   705.4582 25.18741 1.072270e-29
## rankAssoc    5407.262  1066.5525  5.06985 6.085313e-06
## rankProf    11890.283   972.4070 12.22768 1.687464e-16
```




```
summary(lm(salary~year+rank,data=salary))$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 16203.2682  638.67683 25.370058 1.847993e-29
## year          375.6956   70.91772  5.297627 2.904974e-06
## rankAssoc    4262.2847  882.89143  4.827643 1.446223e-05
## rankProf     9454.5232  905.83010 10.437413 6.120546e-14
```




```
summary(lm(salary~year+rank,data=salary))$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 16203.2682  638.67683 25.370058 1.847993e-29
## year          375.6956   70.91772  5.297627 2.904974e-06
## rankAssoc    4262.2847  882.89143  4.827643 1.446223e-05
## rankProf     9454.5232  905.83010 10.437413 6.120546e-14
```

Questions:

Model specification:

- One of the variables is categorical. Is this a linear model?
- If so, how should the design matrix $\mathbf{X}$ be specified?
- How to specify different slopes for each group?

Questions:

Model specification:

- One of the variables is categorical. Is this a linear model?
- If so, how should the design matrix $\mathbf{X}$ be specified?
- How to specify different slopes for each group?

Estimation and inference:

- How do we obtain least squares parameter estimates?
- How do we test for the "effect" of `rank` if it has two parameters?
- How can we obtain confidence intervals?

```r
lm(salary~rank,data=salary)
```

```
##
## Call:
## lm(formula = salary ~ rank, data = salary)
##
## Coefficients:
## (Intercept)    rankAssoc     rankProf
##       17769         5407        11890
```

```r
lm(salary~-1+rank,data=salary)
```

```
##
## Call:
## lm(formula = salary ~ -1 + rank, data = salary)
##
## Coefficients:
##  rankAsst  rankAssoc   rankProf
##     17769      23176      29659
```

Exercise: Identify the relationship between the coefficients.

Set to zero parameterization

```
fit_stz<-lm(salary~rank,data=salary)
summary(fit_stz)
```

```
##
## Call:
## lm(formula = salary ~ rank, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5208.9 -1819.2  -417.8  1586.6  8386.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17768.7      705.5   25.19  < 2e-16 ***
## rankAssoc     5407.3     1066.6    5.07 6.09e-06 ***
## rankProf     11890.3      972.4   12.23  < 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
```

Q: What are the t-stats, p-values evaluating?

Mean value parameterization

```
fit_aov<-lm(salary~-1+rank,data=salary)
summary(fit_aov)

##
## Call:
## lm(formula = salary ~ -1 + rank, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5208.9 -1819.2  -417.8  1586.6  8386.1
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## rankAsst   17768.7      705.5   25.19   <2e-16 ***
## rankAssoc  23175.9      799.9   28.97   <2e-16 ***
## rankProf   29658.9      669.3   44.32   <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:  0.9859,Adjusted R-squared:  0.9851
## F-statistic:  1146 on 3 and 49 DF,  p-value: < 2.2e-16
```

Q: What are the *t*-stats, *p*-values evaluating?

Messed up R^2 in R

$$R^2 = \frac{SSY - RSS}{SSY}$$

```
y<-salary$salary

SSY<-sum( (y-mean(y))^2 )
```

```
(SSY - sum(fit_stz$res^2))/SSY

## [1] 0.7541924

(SSY - sum(fit_aov$res^2))/SSY

## [1] 0.7541924
```

Messed up R^2 in R

$$R^2 = \frac{SSY - RSS}{SSY}$$

```
y<-salary$salary

SSY<-sum( (y-mean(y))^2 )
```

```
(SSY - sum(fit_stz$res^2))/SSY

## [1] 0.7541924

(SSY - sum(fit_aov$res^2))/SSY

## [1] 0.7541924
```

```
(sum(y^2)-sum(fit_aov$res^2))/sum(y^2)

## [1] 0.9859469
```

Set to zero

- t_{Asst} evaluates if $\mu_{Asst} \neq 0$
- t_{Assoc} evaluates if $\mu_{Assoc} - \mu_{Asst} \neq 0$
- t_{Prof} evaluates if $\mu_{Prof} - \mu_{Asst} \neq 0$

Set to zero

- t_{Asst} evaluates if $\mu_{Asst} \neq 0$
- t_{Assoc} evaluates if $\mu_{Assoc} - \mu_{Asst} \neq 0$
- t_{Prof} evaluates if $\mu_{Prof} - \mu_{Asst} \neq 0$

Mean value

- t_{Asst} evaluates if $\mu_{Asst} \neq 0$
- t_{Assoc} evaluates if $\mu_{Assoc} \neq 0$
- t_{Prof} evaluates if $\mu_{Prof} \neq 0$

Set to zero

- t_{Asst} evaluates if $\mu_{Asst} \neq 0$
- t_{Assoc} evaluates if $\mu_{Assoc} - \mu_{Asst} \neq 0$
- t_{Prof} evaluates if $\mu_{Prof} - \mu_{Asst} \neq 0$

Mean value

- t_{Asst} evaluates if $\mu_{Asst} \neq 0$
- t_{Assoc} evaluates if $\mu_{Assoc} \neq 0$
- t_{Prof} evaluates if $\mu_{Prof} \neq 0$

Questions:

- Which parameterization is more useful?
- How would you statistically compare the μ_{Prof} and μ_{Assoc}? (see notes)

Evaluating contrasts

```
fit<-lm(salary ~ rank, data = salary)

iXX<-summary(fit)$cov.unscaled

sigma<-summary(fit)$sigma

sqrt(diag( sigma^2*iXX ))

## (Intercept)   rankAssoc    rankProf
##    705.4582   1066.5525    972.4070
```

Evaluating contrasts

```
fit<-lm(salary ~ rank, data = salary)

iXX<-summary(fit)$cov.unscaled

sigma<-summary(fit)$sigma

sqrt(diag( sigma^2*iXX ))

## (Intercept)   rankAssoc    rankProf
##    705.4582   1066.5525    972.4070
```

```
summary(fit)$coef

##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 17768.667   705.4582 25.18741 1.072270e-29
## rankAssoc    5407.262  1066.5525  5.06985 6.085313e-06
## rankProf    11890.283   972.4070 12.22768 1.687464e-16
```

Evaluating contrasts

```
beta<-fit$coef

a<-c(0,-1,1)

l<-a%*%beta

se_l<- sigma*sqrt( a%*%iXX%*%a )
```

Evaluating contrasts

```
beta<-fit$coef

a<-c(0,-1,1)

l<-a%*%beta

se_l<- sigma*sqrt( a%*%iXX%*%a )
```

```
l

##           [,1]
## [1,] 6483.021

l/se_l

##           [,1]
## [1,] 6.215978

l+c(-1,1)*qt(.975,49)*se_l

## [1] 4387.113 8578.930
```

Comparison with two-sample t-test

```
y_Assoc<-salary$salary[salary$rank=="Assoc"]

y_Prof<-salary$salary[salary$rank=="Prof"]


t.test(y_Prof,y_Assoc)

##
##  Welch Two Sample t-test
##
## data:  y_Prof and y_Assoc
## t = 6.3036, df = 28.278, p-value = 7.769e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4377.244 8588.799
## sample estimates:
## mean of x mean of y
##  29658.95  23175.93
```

Q: Why bother with the regression?

Comparison with two-sample t-test

```
y_Assoc<-salary$salary[salary$rank=="Assoc"]

y_Prof<-salary$salary[salary$rank=="Prof"]


t.test(y_Prof,y_Assoc)

##
##  Welch Two Sample t-test
##
## data:  y_Prof and y_Assoc
## t = 6.3036, df = 28.278, p-value = 7.769e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4377.244 8588.799
## sample estimates:
## mean of x mean of y
##  29658.95  23175.93
```

Q: Why bother with the regression?

A:

- Regression uses a common variance estimate (can be good or bad).
- Regression can control for effects of other variables, allowing for evaluation of differences *conditional* on values of other variables.

Combining numerical and categorical predictors

```
fit<-lm( salary ~ year + rank, data=salary)
```

Q: What is this equivalent to in terms of β's and x's?

Combining numerical and categorical predictors

```
fit<-lm( salary ~ year + rank, data=salary)
```

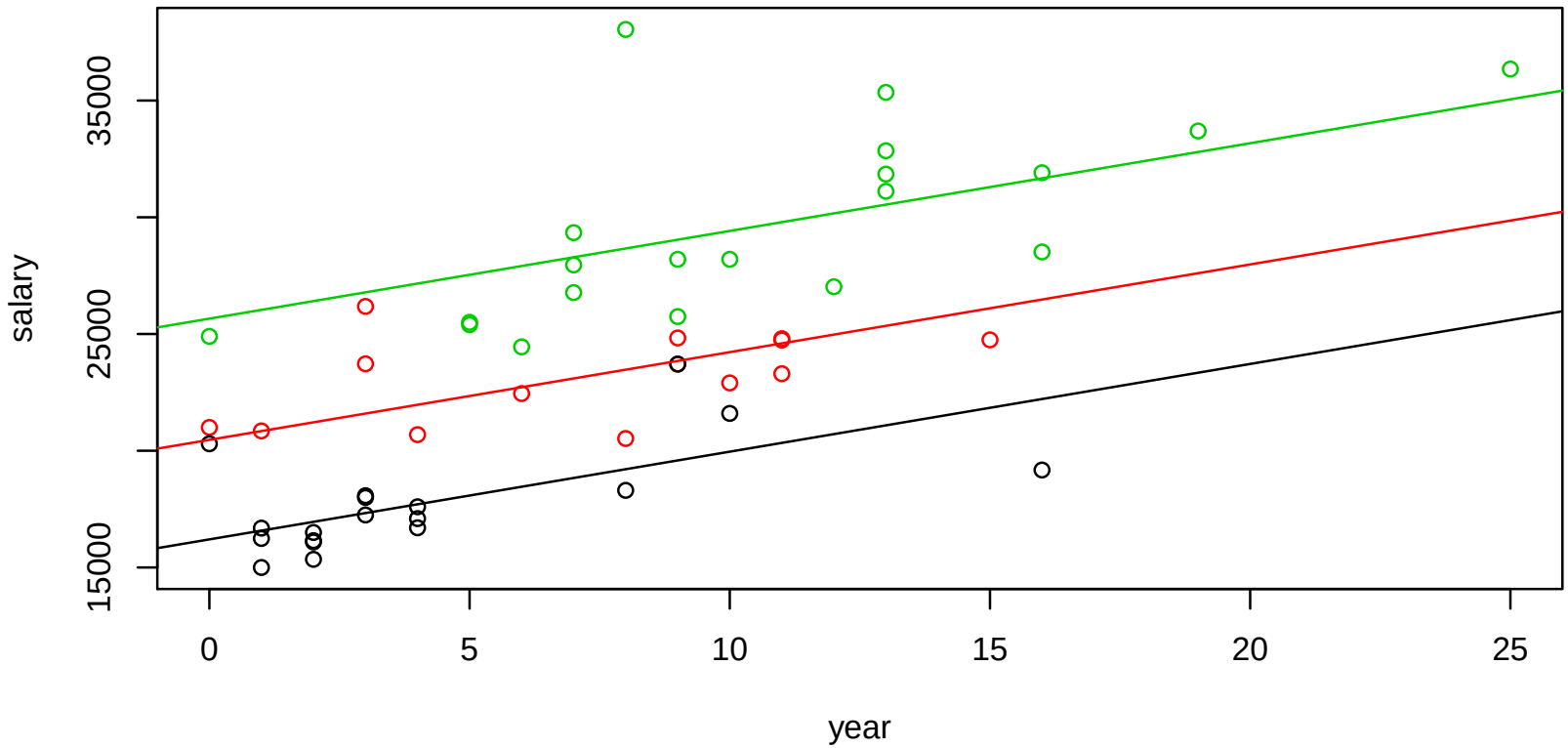
Q: What is this equivalent to in terms of β's and x's?

$$\mathsf{E}[y|x_1, x_2, x_3] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where

- $x_1 =$ year
- $x_2 =$ indicator of Associate Prof
- $x_3 =$ indicator of Full Prof

```
plot(salary~year,col=rank,data=salary)

abline(fit$coef[1],              fit$coef[2],col=1)
abline(fit$coef[1]+fit$coef[3],fit$coef[2],col=2)
abline(fit$coef[1]+fit$coef[4],fit$coef[2],col=3)
```



Conditional evaluation of contrasts

Contrasts between ranks without controlling for year:

```
summary(lm(salary~rank,data=salary))$coef

##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 17768.667   705.4582 25.18741 1.072270e-29
## rankAssoc    5407.262  1066.5525  5.06985 6.085313e-06
## rankProf    11890.283   972.4070 12.22768 1.687464e-16
```

Conditional evaluation of contrasts

Contrasts between ranks without controlling for year:

```
summary(lm(salary~rank,data=salary))$coef

##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 17768.667   705.4582 25.18741 1.072270e-29
## rankAssoc    5407.262  1066.5525  5.06985 6.085313e-06
## rankProf    11890.283   972.4070 12.22768 1.687464e-16
```

Contrasts between ranks controlling for year:

```
summary(lm(salary~year+rank,data=salary))$coef

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 16203.2682  638.67683 25.370058 1.847993e-29
## year          375.6956   70.91772  5.297627 2.904974e-06
## rankAssoc    4262.2847  882.89143  4.827643 1.446223e-05
## rankProf     9454.5232  905.83010 10.437413 6.120546e-14
```

```r
VB<-summary(fit)$cov.unscaled * summary(fit)$sigma^2

beta<-fit$coef

a<-c(0,0,-1,1)

l<-a%*%beta

se_l<- sqrt( a%*%VB%*%a )
```

```r
l
```

```
##           [,1]
## [1,] 5192.239
```

```r
l/se_l
```

```
##           [,1]
## [1,] 5.955544
```

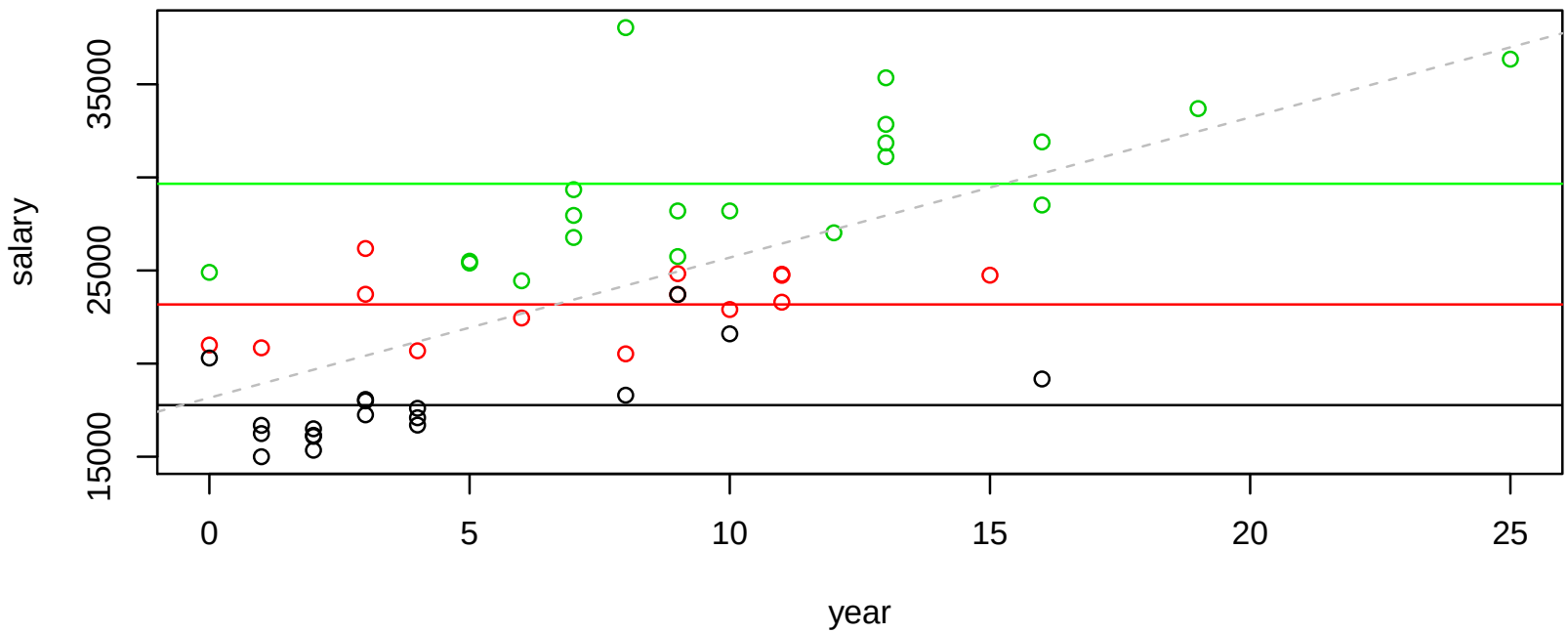
```r
l+c(-1,1)*qt(.975,48)*se_l
```

```
## [1] 3439.301 6945.176
```

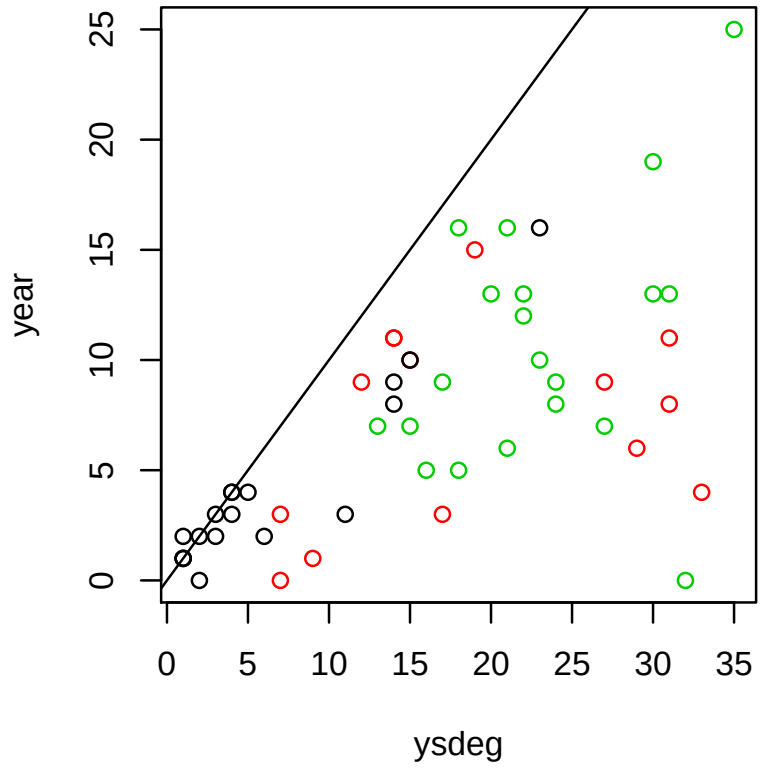
Discuss

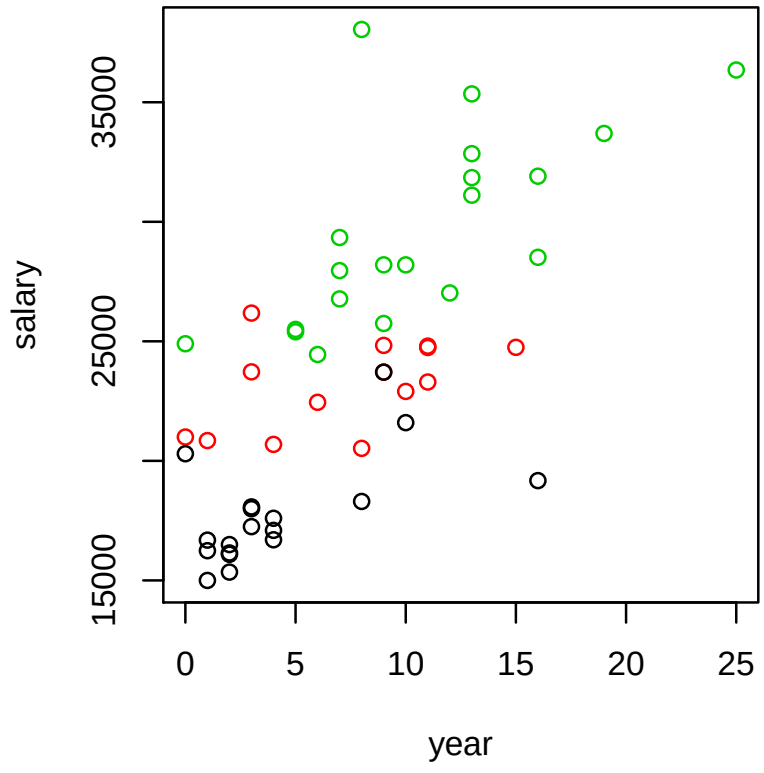
After adding `year` to the model,

- contrast estimates went down;
- standard errors went down;
- t-statistics were about the same.



Adding a third variable





```r
summary(lm(salary~year+ysdeg+rank,data=salary))$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 16317.46287  667.85860 24.4325114 2.320894e-28
## year          400.46009   81.49863  4.9137034 1.125556e-05
## ysdeg         -34.32314   54.54184 -0.6292993 5.322002e-01
## rankAssoc    4619.12028 1054.02128  4.3823786 6.547685e-05
## rankProf     9864.30333 1120.27016  8.8052897 1.645743e-11
```

```r
summary(lm(salary~ysdeg+rank,data=salary))$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17166.46499  785.40068 21.856952 1.350626e-26
## ysdeg          95.08447   58.14789  1.635218 1.085453e-01
## rankAssoc    4209.65030 1279.19805  3.290851 1.877161e-03
## rankProf    10310.29631 1359.38547  7.584527 9.399039e-10
```

```r
summary(lm(salary~year+ysdeg,data=salary))$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 16555.6897 1052.38960 15.731522 8.671060e-21
## year          489.2945  129.56024  3.776579 4.307559e-04
## ysdeg         222.2513   69.80376  3.183945 2.525343e-03
```

`lm(salary~year+ysdeg+rank)`
`ysdeg` not significant

`lm(salary~year+ysdeg+rank)`

`ysdeg` not significant

It is basically a noisy combination of `rank` and `year`

`lm(salary~year+ysdeg+rank)`
`ysdeg` not significant
It is basically a noisy combination of `rank` and `year`

`lm(salary~ysdeg+rank)`
`ysdeg` not significant

`lm(salary~year+ysdeg+rank)`
`ysdeg` not significant
It is basically a noisy combination of `rank` and `year`

`lm(salary~ysdeg+rank)`
`ysdeg` not significant
It is basically a noisy version of `rank`

`lm(salary~year+ysdeg+rank)`
`ysdeg` not significant
It is basically a noisy combination of `rank` and `year`

`lm(salary~ysdeg+rank)`
`ysdeg` not significant
It is basically a noisy version of `rank`

`lm(salary~year + ysdeg)`
`ysdeg` is significant

`lm(salary~year+ysdeg+rank)`
`ysdeg` not significant
It is basically a noisy combination of `rank` and `year`

`lm(salary~ysdeg+rank)`
`ysdeg` not significant
It is basically a noisy version of `rank`

`lm(salary~year + ysdeg)`
`ysdeg` is significant
It contains some information about `rank`

Preliminary model selection

The process of choosing predictors to include is called **model selection**

Preliminary model selection

The process of choosing predictors to include is called **model selection**

- `year` and `rank`?
- `year` and `ysdeg`?

Preliminary model selection

The process of choosing predictors to include is called **model selection**

- `year` and `rank`?
- `year` and `ysdeg`?

```
summary( lm(salary~year+ysdeg,data=salary) )$sigma

## [1] 3920.687

summary( lm(salary~year+rank,data=salary) )$sigma

## [1] 2402.224
```

Preliminary model selection

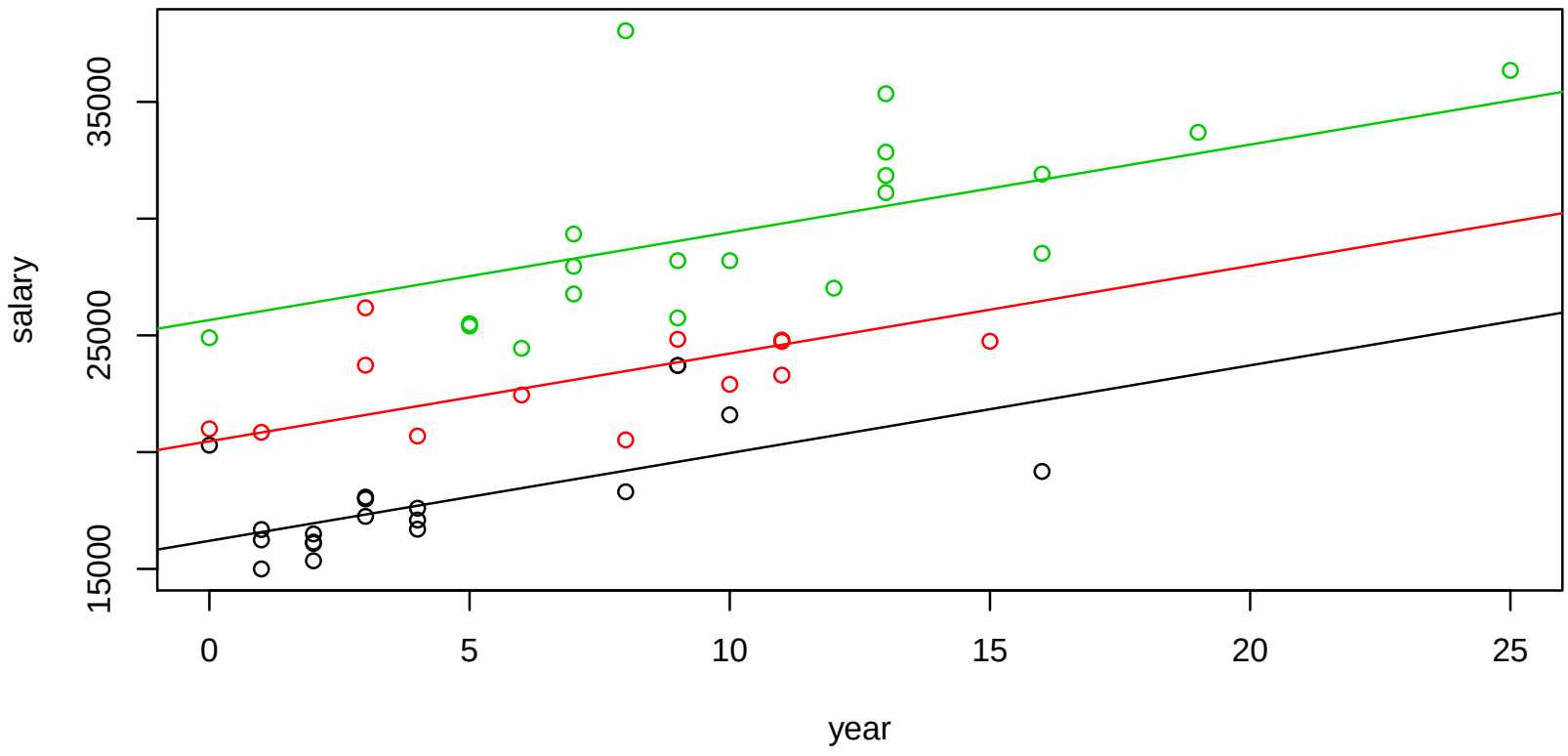
The process of choosing predictors to include is called **model selection**

- `year` and `rank`?
- `year` and `ysdeg`?

```
summary( lm(salary~year+ysdeg,data=salary) )$sigma

## [1] 3920.687

summary( lm(salary~year+rank,data=salary) )$sigma

## [1] 2402.224
```

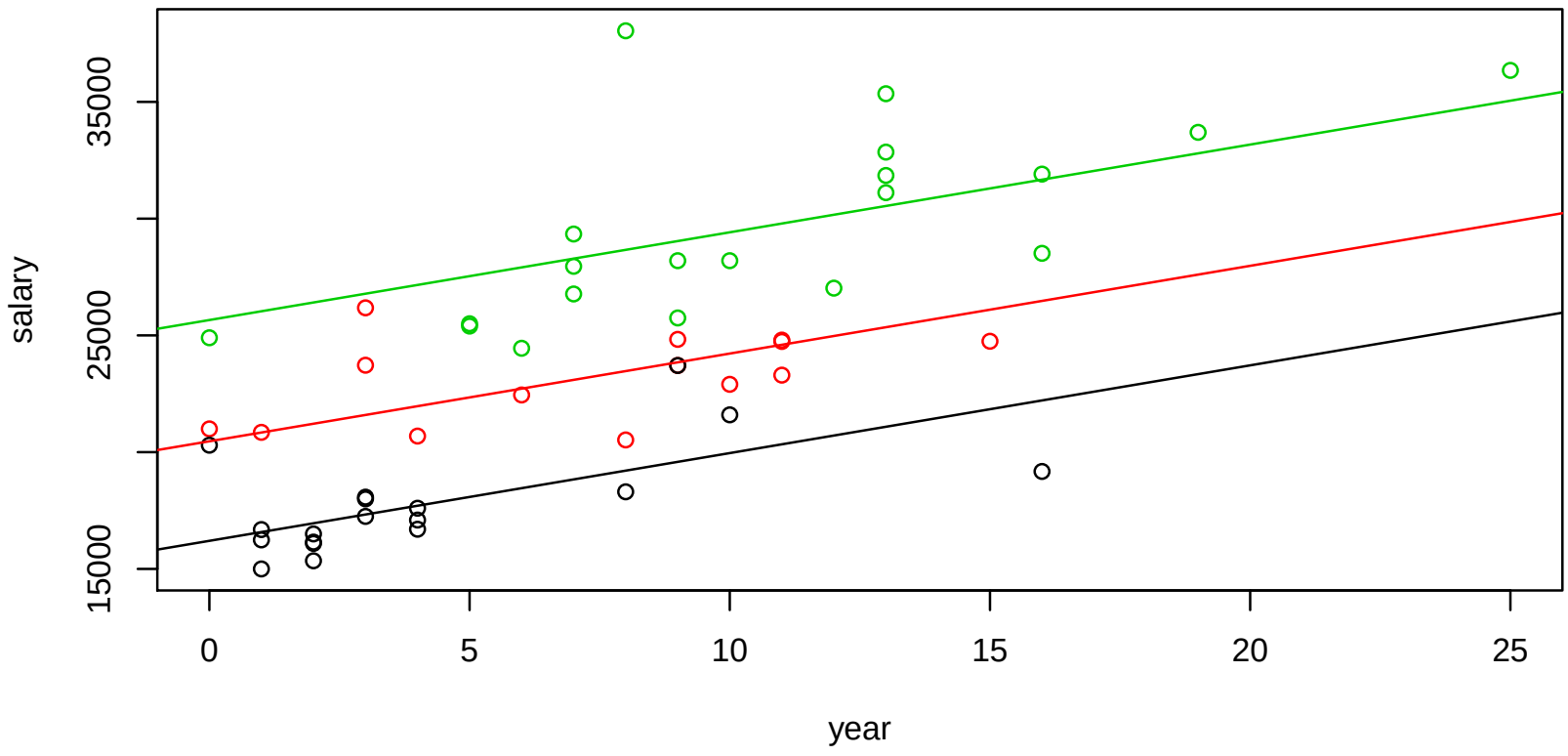
`year` and `rank` seem like a better choice.

Main effects model



The model we have selected is a **main effects model**.

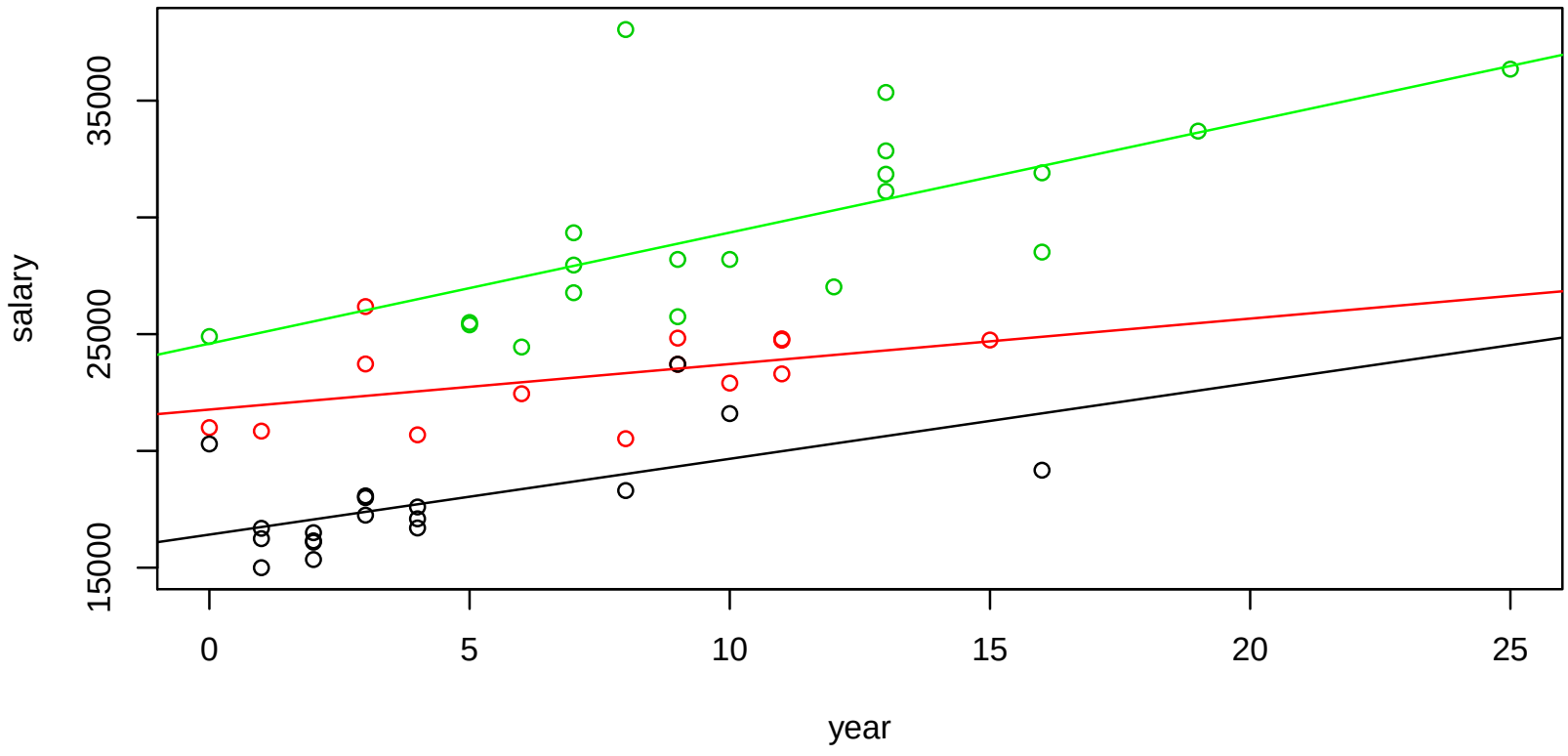
Main effects model



The model we have selected is a **main effects model**.

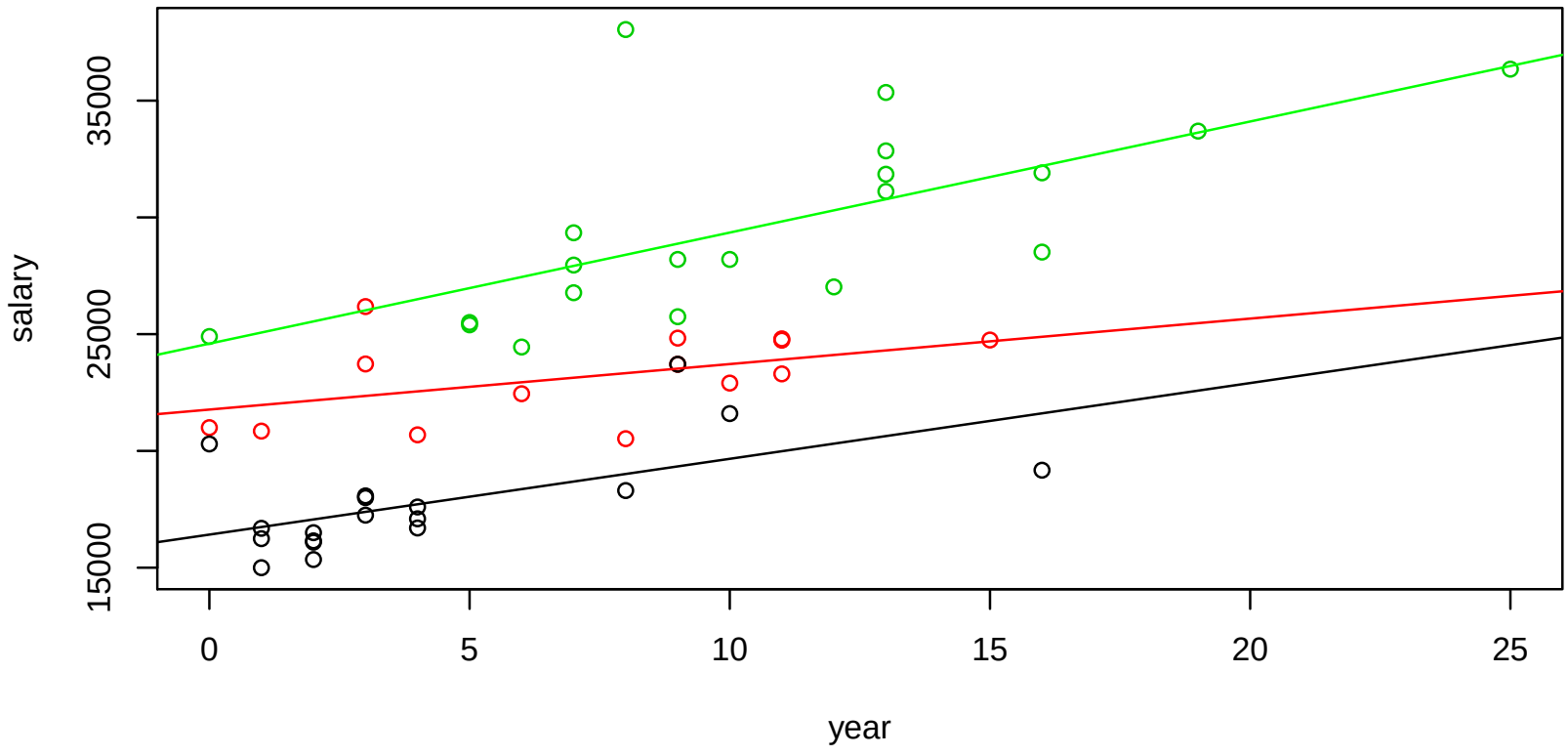
It assumes effects of one variable are constant across levels of the other.

Interaction model



Alternatively, we may want to consider an **interaction model**.

Interaction model



Alternatively, we may want to consider an **interaction model**.

This allows effects of one variable to vary across levels of the other.

```r
fit_int<-lm( salary ~ year + rank + year:rank,data=salary)

summary(fit_int)$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    16416.5723   816.0186 20.1178895 1.967510e-24
## year             324.5027   141.9312  2.2863379 2.688729e-02
## rankAssoc       5354.2430  1492.5574  3.5872945 8.063338e-04
## rankProf        8176.4105  1418.1287  5.7656336 6.493300e-07
## year:rankAssoc  -129.7345   205.7747 -0.6304686 5.315079e-01
## year:rankProf    151.1750   171.7437  0.8802364 3.833070e-01
```

```r
fit_int<-lm( salary ~ year + rank + year:rank,data=salary)

summary(fit_int)$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    16416.5723   816.0186 20.1178895 1.967510e-24
## year             324.5027   141.9312  2.2863379 2.688729e-02
## rankAssoc       5354.2430  1492.5574  3.5872945 8.063338e-04
## rankProf        8176.4105  1418.1287  5.7656336 6.493300e-07
## year:rankAssoc  -129.7345   205.7747 -0.6304686 5.315079e-01
## year:rankProf    151.1750   171.7437  0.8802364 3.833070e-01
```

Quiz 1: Give the design matrix $\mathbf{X}$ that corresponds to this model.

```r
fit_int<-lm( salary ~ year + rank + year:rank,data=salary)

summary(fit_int)$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    16416.5723   816.0186 20.1178895 1.967510e-24
## year             324.5027   141.9312  2.2863379 2.688729e-02
## rankAssoc       5354.2430  1492.5574  3.5872945 8.063338e-04
## rankProf        8176.4105  1418.1287  5.7656336 6.493300e-07
## year:rankAssoc  -129.7345   205.7747 -0.6304686 5.315079e-01
## year:rankProf    151.1750   171.7437  0.8802364 3.833070e-01
```

Quiz 1: Give the design matrix $\mathbf{X}$ that corresponds to this model.

Quiz 2: Describe what each t-statistic is evaluating.

Quiz 3: Statistically compare the slopes of `Assoc` and `Prof`.

Contrast in slopes

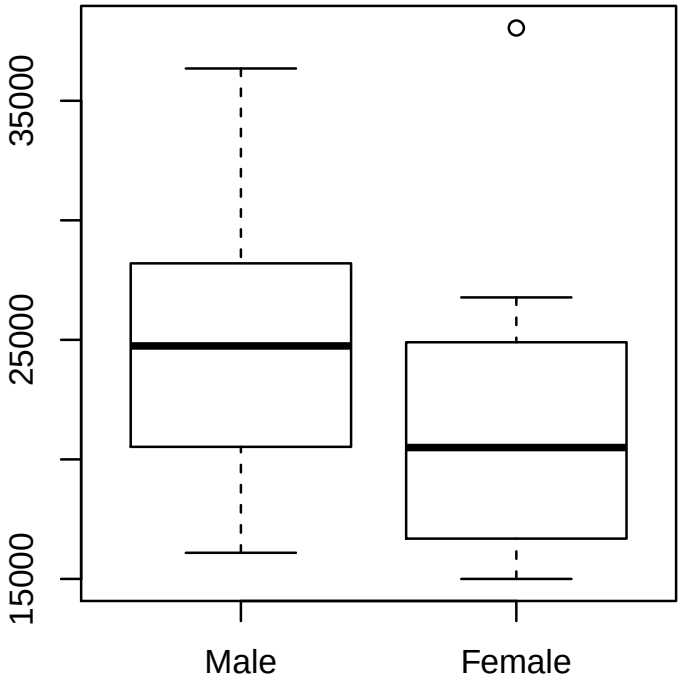
```
VB<-summary(fit_int)$cov.unscaled * summary(fit_int)$sigma^2

fit_int$coef

##    (Intercept)           year      rankAssoc       rankProf year:rankAssoc
##     16416.5723       324.5027      5354.2430      8176.4105      -129.7345
##  year:rankProf
##       151.1750

a<-c(0,0,0,0,-1,1)

l<-a%*%fit_int$coef

se_l<- sqrt( a%*%VB%*%a )
```

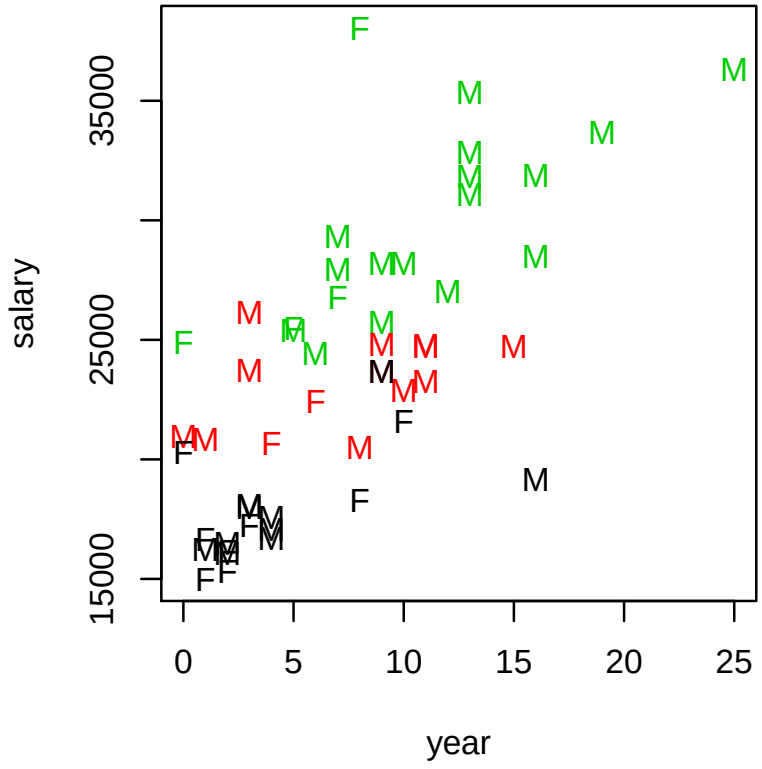
```
l

##           [,1]
## [1,] 280.9095

l/se_l

##           [,1]
## [1,] 1.581486

l+c(-1,1)*qt(.975,46)*se_l

## [1] -76.6285 638.4475
```

Further analysis

```
colnames(salary)

## [1] "degree" "rank"   "sex"    "year"   "ysdeg"  "salary"
```

The data were orginally used to evaluate salary discrimination.





```r
t.test(salary~sex,data=salary,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  salary by sex
## t = 1.8474, df = 50, p-value = 0.0706
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -291.257 6970.550
## sample estimates:
##   mean in group Male mean in group Female
##             24696.79             21357.14
```

How to compare salaries of males to those of females?

- Compare all males to all females?
  (marginal sample means)
- Compare Asst males to Asst females, and for other ranks?
  (conditional sample means)
- Condition on other variables?

Model 1:

```
summary(lm( salary ~ sex ,data=salary ))$coef

##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 24696.789   937.9776 26.32983 5.761530e-31
## sexFemale   -3339.647  1807.7156 -1.84744 7.060394e-02
```

Marginal means suggest the possibility of discrimination.

Model 2:

```
summary(lm( salary ~ sex + rank ,data=salary ))$coef

##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 18155.0909   830.5142 21.8600619 1.342254e-26
## sexFemale    -869.4545   980.5011 -0.8867451 3.796376e-01
## rankAssoc    5145.0455  1109.0352  4.6392083 2.721784e-05
## rankProf    11677.7500  1003.5749 11.6361519 1.412938e-15
```

- No strong evidence of discrimination *within* ranks.
- However, department could discriminate by not promoting women.

Model 2:

```
summary(lm( salary ~ sex + rank ,data=salary ))$coef

##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 18155.0909   830.5142 21.8600619 1.342254e-26
## sexFemale    -869.4545   980.5011 -0.8867451 3.796376e-01
## rankAssoc    5145.0455  1109.0352  4.6392083 2.721784e-05
## rankProf    11677.7500  1003.5749 11.6361519 1.412938e-15
```

- No strong evidence of discrimination *within* ranks.
- However, department could discriminate by not promoting women.

We probably do not want to condition on this variable:

- A department could discriminate in promotion, but not salary.
- There would be an indirect causal effect of sex on salary.
- There would be a marginal effect of sex on salary.
- There would not be a conditional effect.

Model 3:

```
summary(lm( salary ~ sex + year,data=salary ))$coef

##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 18065.4054  1247.7738 14.4781095 2.500540e-19
## sexFemale     201.4668  1455.1450  0.1384514 8.904511e-01
## year          759.0138   118.3363  6.4140410 5.366076e-08
```

Review the plot: High value of `year` positively correlates with `male` and `salary`

Model 4:

```
summary(lm( salary ~ sex + year + sex:year,data=salary ))$coef

##                  Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    18222.5835  1308.6316 13.9249143 1.767461e-18
## sexFemale       -570.7543  2297.2398 -0.2484522 8.048445e-01
## year             741.0236   126.2311  5.8703727 3.952977e-07
## sexFemale:year   169.0535   386.9542  0.4368824 6.641558e-01
```

Little evidence of different `year` effect by `sex`.

Caution: These data are not a random sample.

- we do not have information on faculty who left the department;
- This is a type of data censoring, and could affect conclusions.