MA 576 HW 1

Benjamin Draves

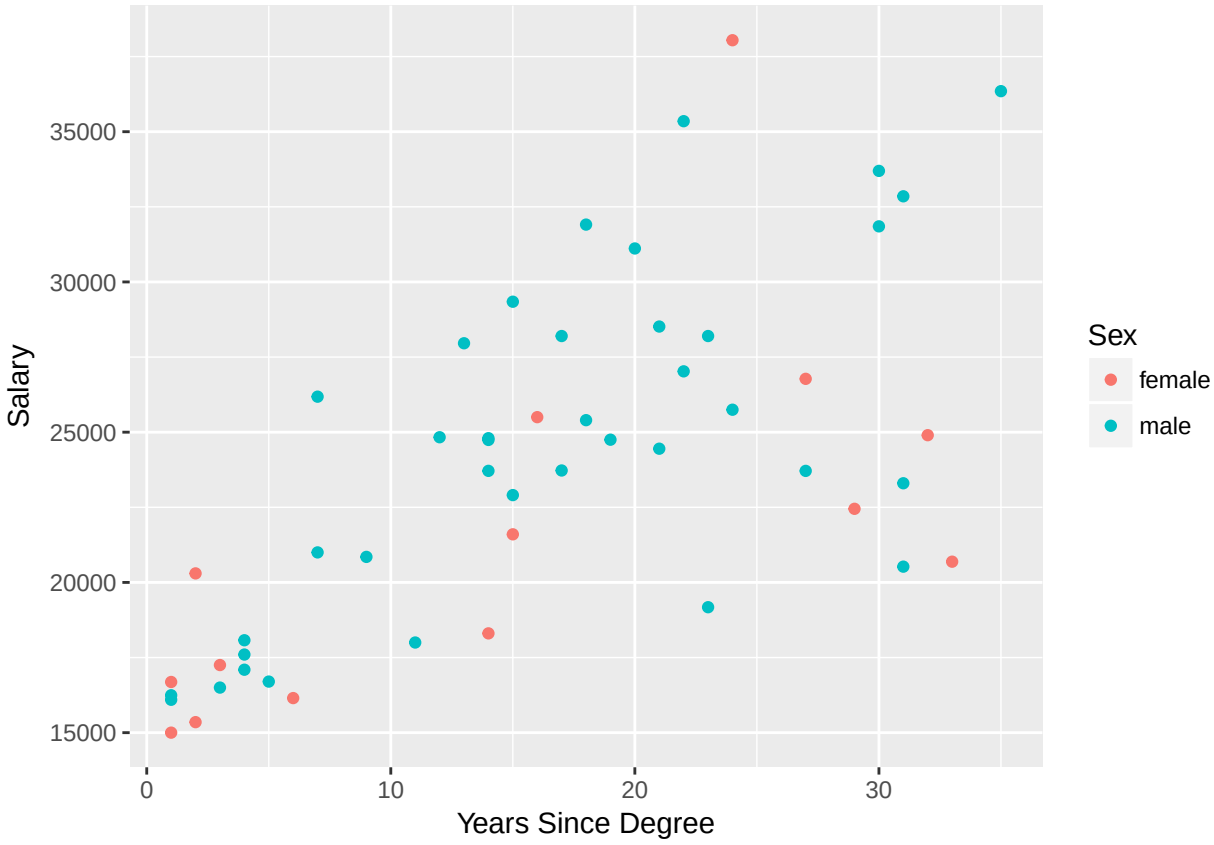
Exercise 3

(a)

```
#load up necessary packages
library(ggplot2)

#read in data
dat = read.csv("~/Desktop/Courses/MA 576/data/salary.csv")

#plot relationship between year from degree and sex
p = ggplot(dat, aes(x = yd, y = sl, color = sx)) +
  geom_point() +
  labs(color = "Sex", x = "Years Since Degree", y = "Salary")
p
```



For the full population, it appears that the average salary increases as the years since degree also increase. This intuitively makes sense as more tenure professors should, in theory, earn a higher salary. While this general trend holds, the variance also increases dramatically as yd increases. Focusing on the two subgroups, female and male, it appears that in general that male professors earn

a similar salary to female professors when yd is near zero. But as yd increases, it appears that male salaries raise faster than their female colleagues. Ignoring other variables, including performance metrics, this plot suggests that there may be bias towards male professors in the promotion process at this institution.

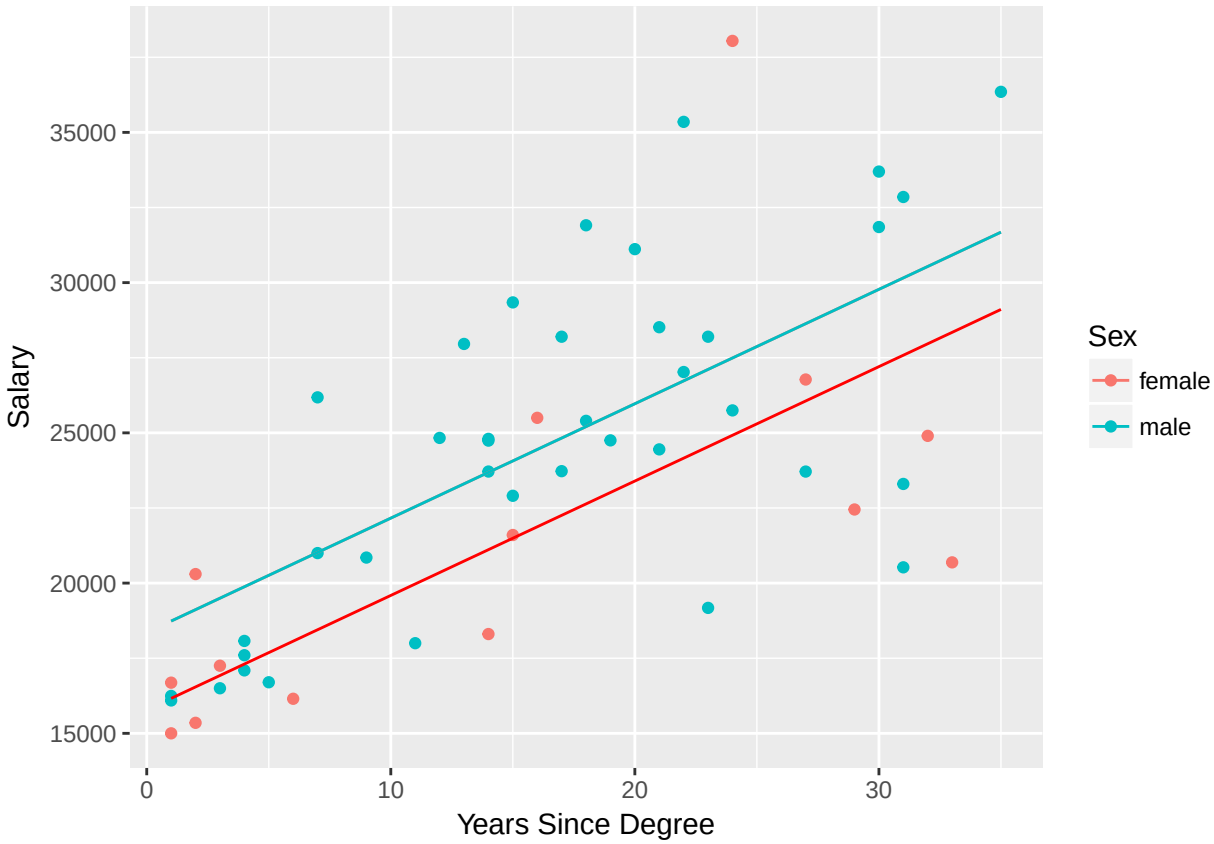
(b)

```
#build a linear model for this relationship
m = lm(sl ~ yd + sx, data = dat)
summary(m)
```

```
##
## Call:
## lm(formula = sl ~ yd + sx, data = dat)
##
## 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) 15782.70    1438.33  10.973 8.44e-15 ***
## yd            380.69      59.11   6.440 4.88e-08 ***
## sxmale       2572.53    1349.08   1.907   0.0624 .
## ---
## 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
```

After fitting the model, the $\hat{\beta}$ vector can be written as $(\hat{\beta}_0, \hat{\beta}_{yd}, \hat{\beta}_{male}) = (15782.70, 380.69, 2572.53)$. Here, we can interpret the $\hat{\beta}_0$ coefficient as an estimate for the starting salary of a professor at this institution. That is, immediately after earning their degree, an incoming professor could expect to earn a \$15782.70 salary to start. The $\hat{\beta}_{yd}$ estimate can be intepretted as the expected raise in income each year for a professor at the institution. With ever year of experience past earning their degree, a professor could expect a raise of \$380.69 each year. Lastly, the $\hat{\beta}_{male}$ variable can be interpreted as the expected difference between a female and male professor's salary at this institution who earned their degree at the same time. That is, on average, a male professor in this sample will make \$2572.53 dollars more than a female colleague who earned their degrees at the same time.

```
Slmale = function(x) (15782.70 + 2572.53) + x*(380.69)
Slfemale = function(x) (15782.70) + x*(380.69)
p + stat_function(fun = Slmale) + stat_function(fun = Slfemale, col = "red")
```



(c)

Here, we look to test if there is a significant difference between male and female salaries. In our model, this problem reduces to testing $H_0 : \hat{\beta}_{Male} = 0$ versus $H_A : \hat{\beta}_{Male} \neq 0$. Using the standard errors reported above, we construct a 95% confidence interval, Ω_0 for the estimator. Seeing that $n = 52$ we use the Z distribution for proper scaling values.

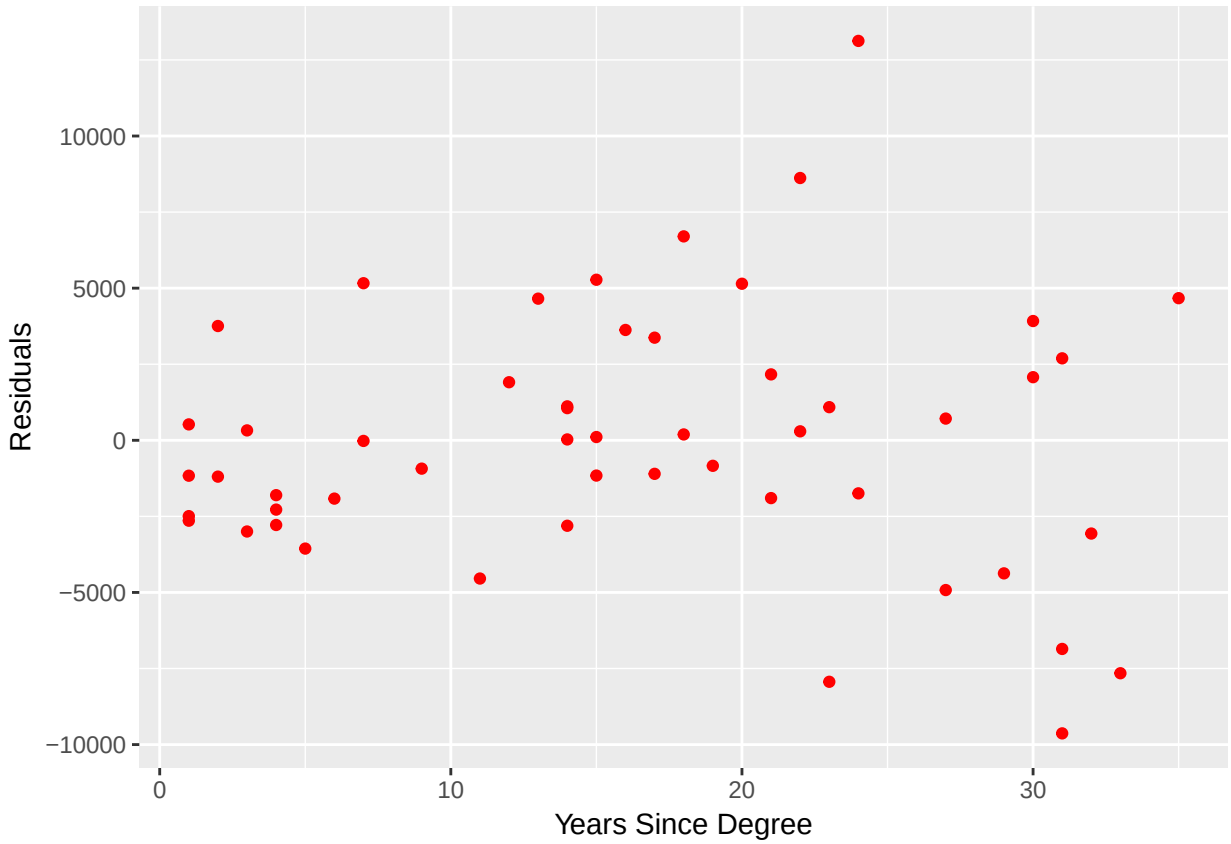
```
l = 2572.53 - qnorm(.975) * 1349.08
r = 2572.53 + qnorm(.975) * 1349.08
c(l,r)
```

```
## [1]  -71.61821 5216.67821
```

Seeing that $0 \in \Omega_0$ we fail to reject H_0. We do not have sufficient evidence to suggest the presence of wage discrimination based on sex.

(d)

```
d = data.frame(yd = dat$yd, res = m$residuals)
p2 = ggplot(d, aes(yd, res)) +
  geom_point(color = "red") +
  labs(x = "Years Since Degree", y = "Residuals")
p2
```



The residuals appear to depend on the years since degree variable. As yd increases, on average the residuals decrease. This suggests that the model is overestimating the salaries earned by professors who have more than 20 years since earning their degree. Generally, there appears to be a quadratic trend between yd and these residuals.

(e)

```
#build model with interaction term
m2 = lm(sl ~ yd*sx, data = dat)
summary(m2)
```

```
##
## Call:
## lm(formula = sl ~ yd * sx, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10192.5  -2053.4   -395.8   2101.0  13732.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16732.13    1824.82   9.169    4e-12 ***
## yd            315.85      96.65   3.268    0.002 **
## sxmale        971.75    2319.91   0.419    0.677
```

```
## yd:sxmale     103.94     122.37   0.849    0.400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4311 on 48 degrees of freedom
## Multiple R-squared:  0.5006, Adjusted R-squared:  0.4693
## F-statistic: 16.04 on 3 and 48 DF,  p-value: 2.35e-07
```

Here our model can be written as $Sl = \beta_0 + \beta_1 Yd + \beta_2 1_{\{Sex=Male\}} + \beta_3 (Yd \times 1_{\{Sex=Male\}}) + \epsilon$. Hence for the two nested models describing male and female salaries we have

$$Sl_{Female} = \beta_0 + \beta_1 Yd + \epsilon$$

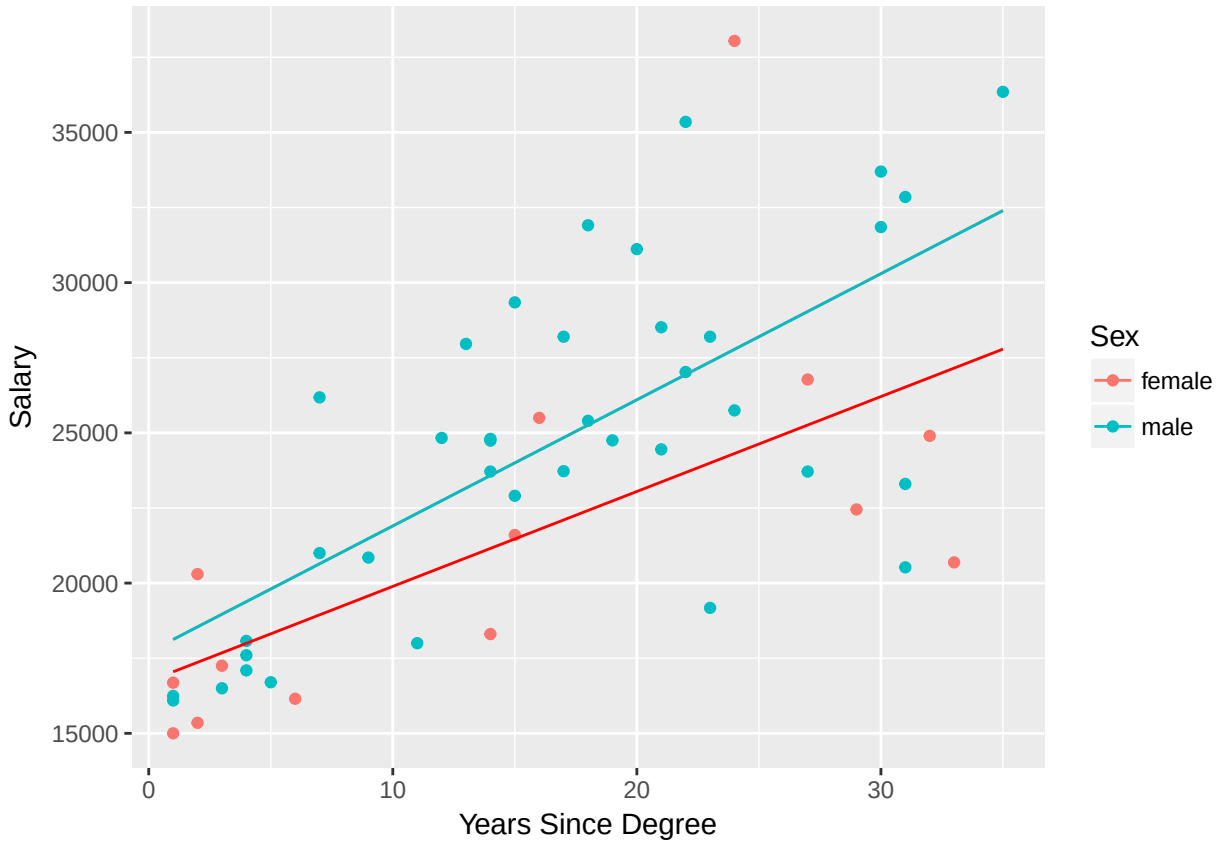
and

$$Sl_{Male} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Yd + \epsilon$$

In this context, β_2 is the expected difference in male and female starting salaries while β_3 is the difference between the effect of Yd between male and female professors. That is, β_3 measures if there is a difference in the effect of gaining an additional year of experience between female and male professors.

Including this term, we see that $\hat{\beta}_3 = 103.94$ suggesting that as professors get older, on average male professor's salaries rise faster than their female counterparts. Notice, however, after including the interaction term that $\hat{\beta}_2 = 971.75$ compared to the original model estimate of $\hat{\beta}_2 = 2572.53$. That is if any bias exists, the majority of it exists in the promotion process or the salaries of more tenured professors. These results would suggest that on average male professors make more than female professors as well as are granted higher pay raises than female professors but upon considering the standard error estimates (as well as the validity of the assumption that $\epsilon \sim N(0, \sigma^2)$) it is hard to determine if these effects are not just noise inherent in our sample. Therefore we can not say with certainty that this trend is attributable to sexual bias as compared to random chance.

```
#define regression functions by group
Slmale = function(x) (16732.13 + 971.75) + x*(315.85 + 103.94)
Slfemale = function(x) (16732.13) + x*(315.85)
p + stat_function(fun = Slmale) + stat_function(fun = Slfemale, col = "red")
```



(f)

```
#include quadratic term for yd
d = cbind(dat, sqyd = dat$yd^2)
m3 = lm(sl ~ yd+sx + sqyd, data = d)
summary(m3)
```

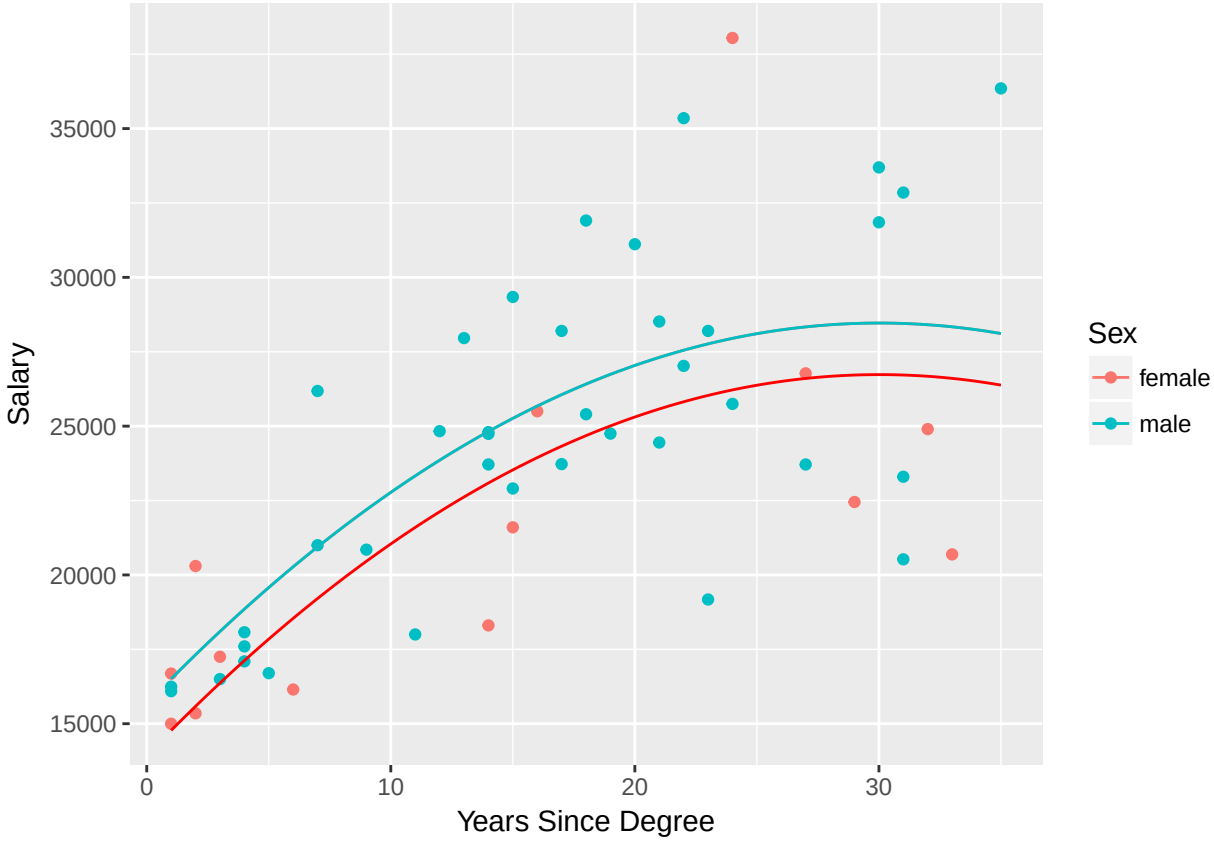
```
##
## Call:
## lm(formula = sl ~ yd + sx + sqyd, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8589.1 -2234.0  -339.2  1968.2 11828.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13937.47    1601.90   8.701 1.96e-11 ***
## yd            852.49     215.23   3.961 0.000247 ***
## sxmale       1732.47    1346.89   1.286 0.204518
## sqyd          -14.20       6.25  -2.272 0.027577 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4127 on 48 degrees of freedom
## Multiple R-squared:  0.5423, Adjusted R-squared:  0.5137
## F-statistic: 18.96 on 3 and 48 DF,  p-value: 3.004e-08
```

By including a quadratic yd term, our fit appears to dramatically improve (see below). This also matches our intuition as we do not expect for salaries to grow linearly throughout a professor's tenure. We begin by noting that $(\hat{\beta}_0, \hat{\beta}_{yd}, \hat{\beta}_{yd^2}) = (13937.47, 852.49, -14.20)$. We can interpret these estimates as we have above. We expect a first year professor (since earning their degree) to earn \$13937.47, and with each year past earning their degree, yd, to recieve a $852.49yd - 14.20yd^2$ raise.

Now focusing on the sex variable, we note that by including the quadratic term that $\hat{\beta}_{Male} = 1732.47$. That is the model estimates that male professors make more than their female colleagues. Upon investigation of the standard errors corresponding to this estimate, however, we do not have evidence to suggest that this difference is not just due to noise in the sample.

```
#plot two models
Slmale = function(x) (13937.47 + 1732.47) + 852.49*x + x^2*(-14.20)
Slfemale = function(x) (13937.47) + x*(852.49) + x^2*(-14.20)
p + stat_function(fun = Slmale) + stat_function(fun = Slfemale, col = "red")
```



(g)

This analysis looks to investigate if there is a difference in salary values between male and female professors at a higher education institution. There are $n = 52$ datapoints in the dataset summarizing

professor's attributes from the 1980s. Accounting for years since earning their degree and their sex, we find that the years since earning their degree explains a great deal of the trend in professor's salaries. This matches our intuition as we expect with more experience, professors earn higher salaries. In addition, we have evidence to suggest that there is "plateau" point, where professor's salaries begin to become constant with respect to their experience. This trend could be attributable to the tenure structure, where tenure professors in general earn the same salary.

With regards to salary difference between the sexes, we find that the difference in male and female salaries negligible. While in this dataset, our model does in fact estimate that male professors earn a higher wage, we cannot attribute this feature to true underlying differences in the salary structure between males and females working for this institution. Further investigation, with ideally a larger sample of professors (only 14 women were included in this sample), would uncover if this difference can be attributed to random chance or if there is in fact a legitimate bias in the salary structure at this institution.