# R04 - Regression with Logarithms

HCI/PSYCH 522
Iowa State University

March 22, 2022

# Overview

- Review
  - Simple linear regression (SLR)
  - Regression with a categorical variable
  - Preview of multiple linear regression
- Using logarithms in SLR
  - Logarithm of the dependent variable
  - Logarithm of the independent variable
  - Logarithm of both variables

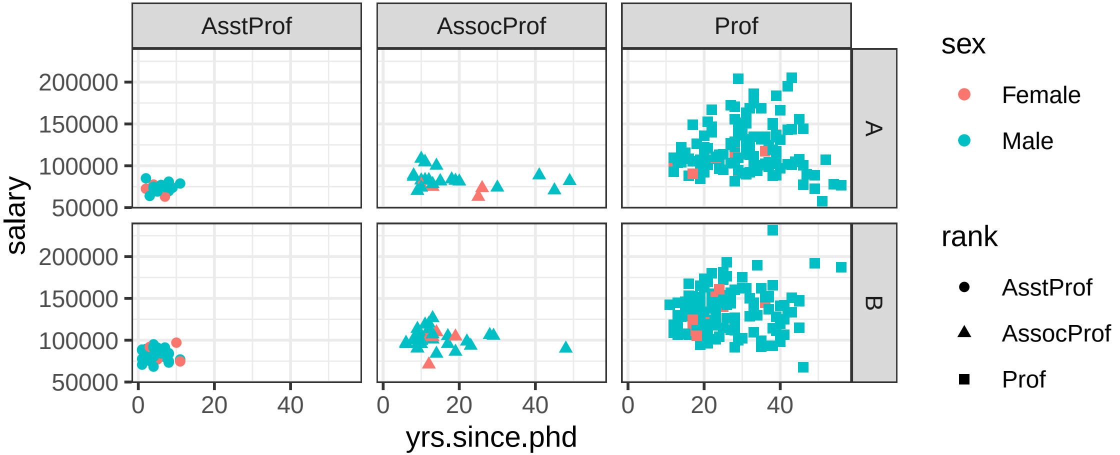
# Understand differences in salary by gender

```
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
```

# Understand differences in salary by gender

```
summary(Salaries)

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

# Understand differences in salary by gender



# Simple linear regression

The simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

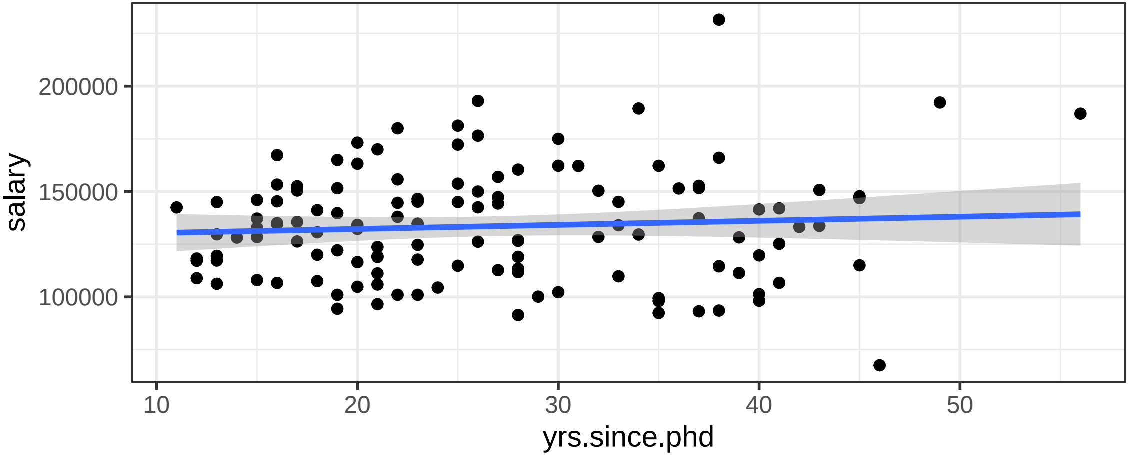
where $Y_i$ and $X_i$ are the dependent and independent variable, respectively, for individual $i$.

To analyze salaries of Male Professors in discipline B at an unknown college in the U.S. from 2008-2009, we will use salary as the dependent variable ($Y$)

- years since PhD as the independent variable ($X$).

In this model, $\beta_1$ is the mean increase in salary for each year since PhD.

# SLR for Salary



# SLR for Salary

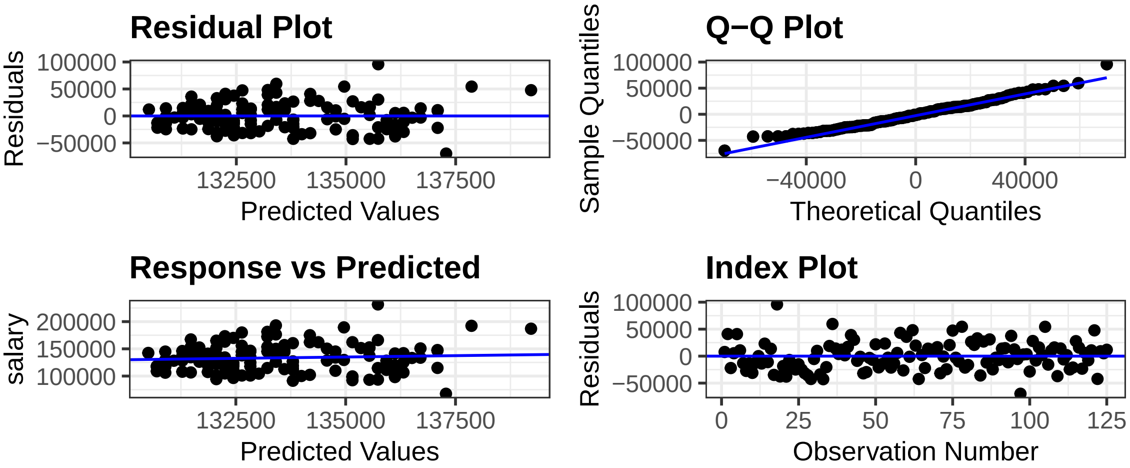
```r
summary(m <- lm(salary ~ yrs.since.phd,
                data = Salaries %>% filter(rank == "Prof", sex == "Male", discipline == "B")))

##
## Call:
## lm(formula = salary ~ yrs.since.phd, data = Salaries %>% filter(rank ==
##     "Prof", sex == "Male", discipline == "B"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -69724 -21138  -1199  15803  95811
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   128378.4     6858.0  18.719   <2e-16 ***
## yrs.since.phd    193.6      242.3   0.799    0.426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26550 on 123 degrees of freedom
## Multiple R-squared:  0.005162,Adjusted R-squared:  -0.002926
## F-statistic: 0.6383 on 1 and 123 DF,  p-value: 0.4259
```

# SLR for Salary

```
confint(m)

##                      2.5 %      97.5 %
## (Intercept)    114803.2971 141953.4711
## yrs.since.phd    -286.0467    673.2097
```

Manuscript statement:
For each year since PhD, the model estimates an mean increase of (-286, 673) dollars.

# Diagnostics



# Regression with a categorical variable

The simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

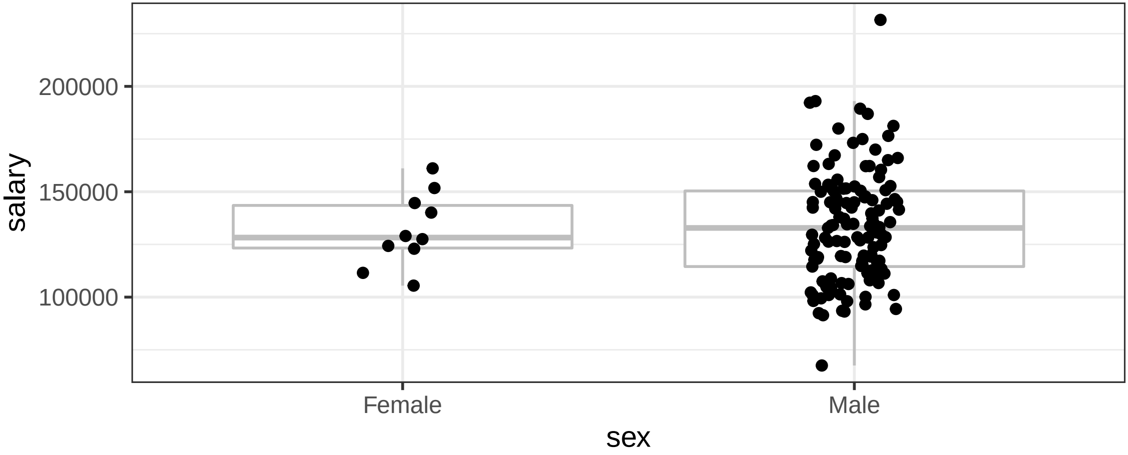
where $Y_i$ and $X_i$ are the dependent and independent variable, respectively, for individual $i$.

To analyze salaries of Professors in discipline B at an unknown college in the U.S. from 2008-2009, we will use salary as the dependent variable $(Y)$

- indicator of being male as the independent variable $(X)$.

In this model, $\beta_1$ is the mean difference in salary between men and women.

# Salary comparison



# Salary comparison

```r
summary(m <- lm(salary ~ sex,
                data = Salaries %>% filter(rank == "Prof", discipline == "B")))

##
## Call:
## lm(formula = salary ~ sex, data = Salaries %>% filter(rank ==
##     "Prof", discipline == "B"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65959 -18970  -1257  16670  98027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   131836       8223  16.033   <2e-16 ***
## sexMale         1682       8546   0.197    0.844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26000 on 133 degrees of freedom
## Multiple R-squared:  0.0002913,Adjusted R-squared:  -0.007225
## F-statistic: 0.03875 on 1 and 133 DF,  p-value: 0.8442
```

# Salary comparison

```
confint(m)

##                   2.5 %    97.5 %
## (Intercept) 115571.52 148100.88
## sexMale     -15220.59  18584.91
```

Manuscript statement:
Difference in mean salary between men and women is estimated to be between (-15,19) thousand dollars more for men.

# Improved model

This is a bit unsatisfactory because this is only for

- Professors in
- Discipline B and
- doesn't account for years since PhD.

We can run a multiple regression model that includes

- sex,
- rank,
- discipline, and
- years since PhD.

This model will provide a comparison of the effect of sex on salary after *adjusting* for rank, discipline, and years since PhD.

# Multiple regression model

The simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots, \sigma^2)$$

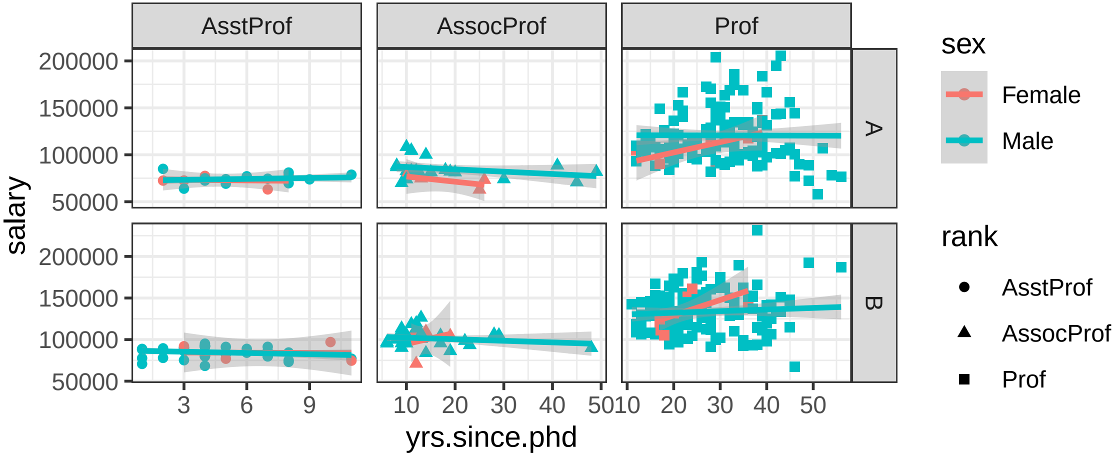
where $Y_i$ and $X_{i,j}$ are the dependent and independent variable(s), respectively, for individual $i$.

To analyze salaries of Professors in discipline B at an unknown college in the U.S. from 2008-2009, we will use salary as the dependent variable $(Y)$

- sex $(X1)$,
- rank $(X2$ and $X3)$,
- discipline $(X4)$, and
- years since PhD $(X5)$

as independent variables. In this model, $\beta_1$ is the mean difference in salary between men and women after adjusting for rank, discipline, and years since PhD.

# Multiple regression



# Salary comparison

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

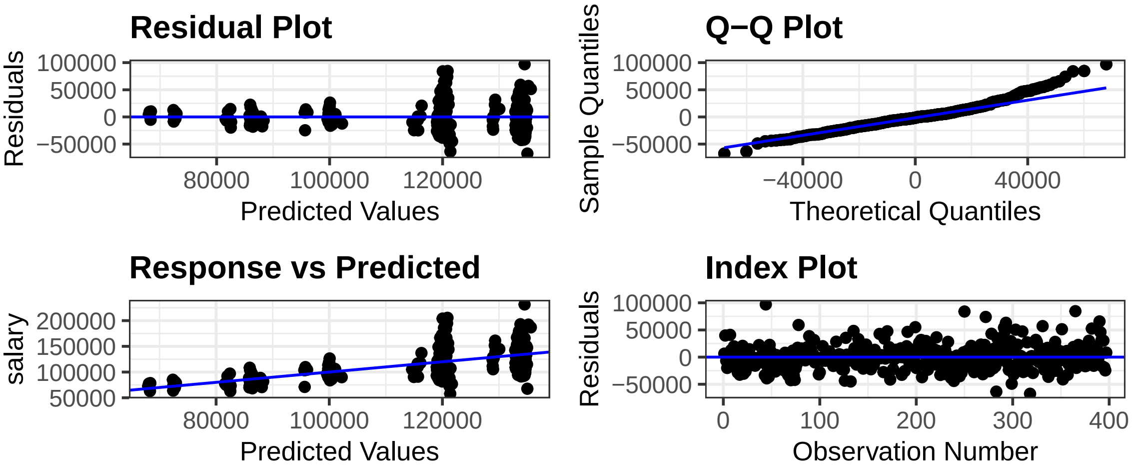
```
##
## Call:
## lm(formula = salary ~ sex + rank + discipline + yrs.since.phd,
##     data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67451 -13860  -1549  10716  97023
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67884.32    4536.89  14.963  < 2e-16 ***
## sexMale        4349.37    3875.39   1.122  0.26242
## rankAssocProf 13104.15    4167.31   3.145  0.00179 **
## rankProf      46032.55    4240.12  10.856  < 2e-16 ***
## disciplineB   13937.47    2346.53   5.940 6.32e-09 ***
## 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
```

# Salary comparison

```
confint(m)

##                     2.5 %     97.5 %
## (Intercept)   58964.5651 76804.0734
## sexMale       -3269.8493 11968.5812
## rankAssocProf  4911.0049 21297.3001
## rankProf      37696.2618 54368.8354
## disciplineB    9324.0682 18550.8744
## yrs.since.phd  -188.6961   310.7186
```

Manuscript statement:
Difference in mean salary between men and women is estimated to be between (-3,12) thousand dollars more for men after adjusting for rank, discipline, and years since PhD.

# Diagnostics



# Logarithms in regression

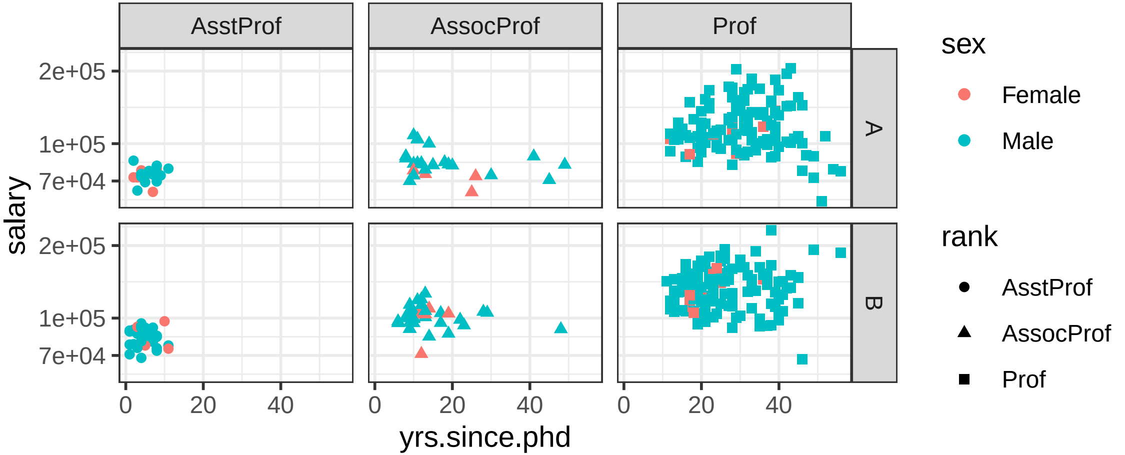
When running a regression, you [the data analyst] has a choice of whether to

- take logarithms of the dependent variable and
- take logarithms of any numeric independent variables.

Suggestions for when to take logarithms:

- You can only take logarithms if the variable is strictly positive.
- If the variable is non-negative (but has zeroes), you can take the logarithm of the variable after adding the smallest non-zero value to all observations.
- Consider taking logarithms if the maximum value divided by the minimum value is greater than 10.

# Understand differences in salary by gender



# Simple linear regression

The simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

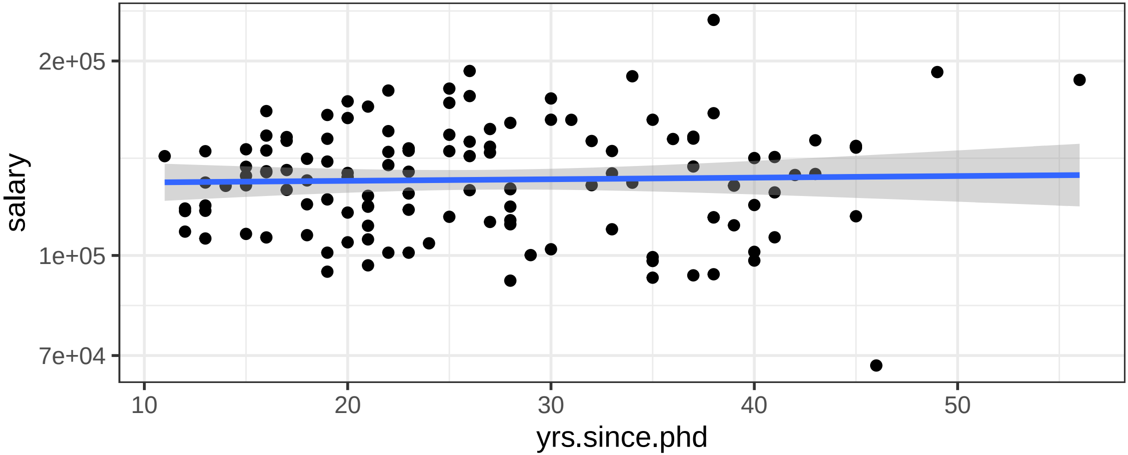
where $Y_i$ and $X_i$ are the dependent and independent variable, respectively, for individual $i$.

To analyze salaries of Male Professors in discipline B at an unknown college in the U.S. from 2008-2009, we will use log salary as the dependent variable $(Y)$

- years since PhD as the independent variable $(X)$.

In this model, $100(e^{\beta_1} - 1)$ will be the percent change in median salary per year since PhD.

# SLR for Salary



# SLR for Salary

```
summary(m <- lm(log(salary) ~ yrs.since.phd,
                data = Salaries %>% filter(rank == "Prof", sex == "Male", discipline == "B")))
```

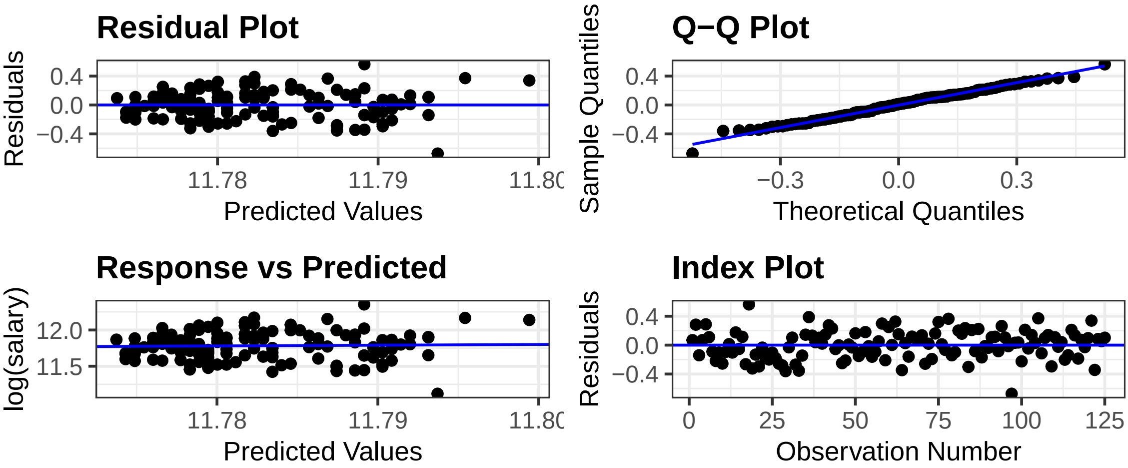
```
##
## Call:
## lm(formula = log(salary) ~ yrs.since.phd, data = Salaries %>%
##     filter(rank == "Prof", sex == "Male", discipline == "B"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67295 -0.14082  0.01135  0.13517  0.56338
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.177e+01  5.138e-02 229.007   <2e-16 ***
## yrs.since.phd 5.704e-04  1.816e-03   0.314    0.754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.199 on 123 degrees of freedom
## Multiple R-squared:  0.0008019,Adjusted R-squared:  -0.007322
## F-statistic: 0.09872 on 1 and 123 DF,  p-value: 0.7539
```

# SLR for Salary

```
confint(m)

##                      2.5 %      97.5 %
## (Intercept)   11.665758432 11.86918421
## yrs.since.phd -0.003023258  0.00416408

exp(confint(m))

##                      2.5 %       97.5 %
## (Intercept)   1.165130e+05 1.427977e+05
## yrs.since.phd 9.969813e-01 1.004173e+00
```

Manuscript statement:
For each year since PhD, the model estimates an increase of (-100, -100)% in median salary.

# Diagnostics



# Regression with a categorical variable

The simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

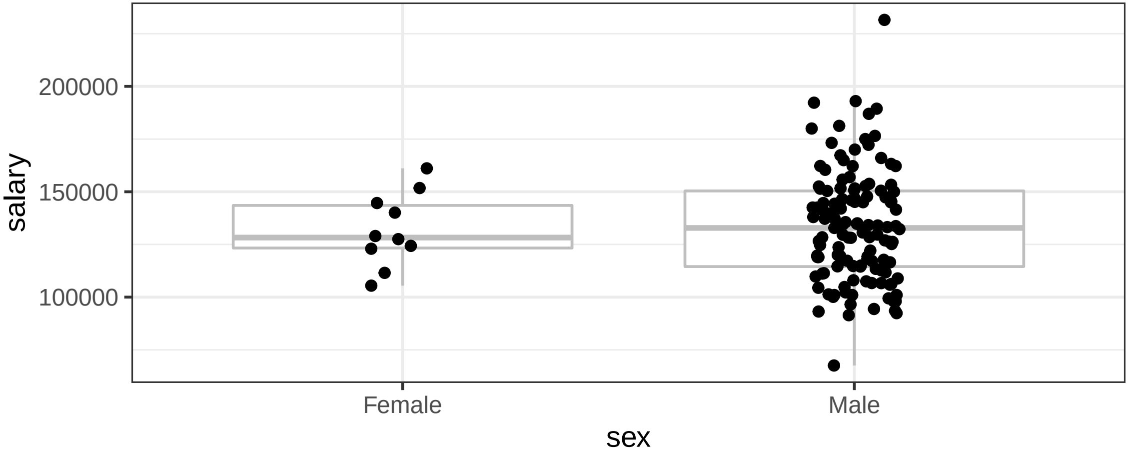
where $Y_i$ and $X_i$ are the dependent and independent variable, respectively, for individual $i$.

To analyze salaries of Professors in discipline B at an unknown college in the U.S. from 2008-2009, we will use log salary as the dependent variable $(Y)$

- indicator of being male as the independent variable $(X)$.

In this model, $100(e^{\beta_1} - 1)$ will be the percent difference in median salary of men compared to women.

# Salary comparison



# Salary comparison

```r
summary(m <- lm(log(salary) ~ sex,
                data = Salaries %>% filter(rank == "Prof", discipline == "B")))
```

```
##
## Call:
## lm(formula = log(salary) ~ sex, data = Salaries %>% filter(rank ==
##     "Prof", discipline == "B"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66186 -0.13387  0.00992  0.13703  0.56991
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.781363   0.061510  191.54   <2e-16 ***
## sexMale      0.001253   0.063923    0.02    0.984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1945 on 133 degrees of freedom
## Multiple R-squared:  2.891e-06,Adjusted R-squared:  -0.007516
## F-statistic: 0.0003845 on 1 and 133 DF,  p-value: 0.9844
```

# Salary comparison

```
confint(m)

##                  2.5 %     97.5 %
## (Intercept) 11.659700 11.9030269
## sexMale     -0.125183  0.1276897
```

Manuscript statement:
Median salary is estimated to be (-12, 14)% larger for men compared to women.

This is a bit unsatisfactory because this is only for

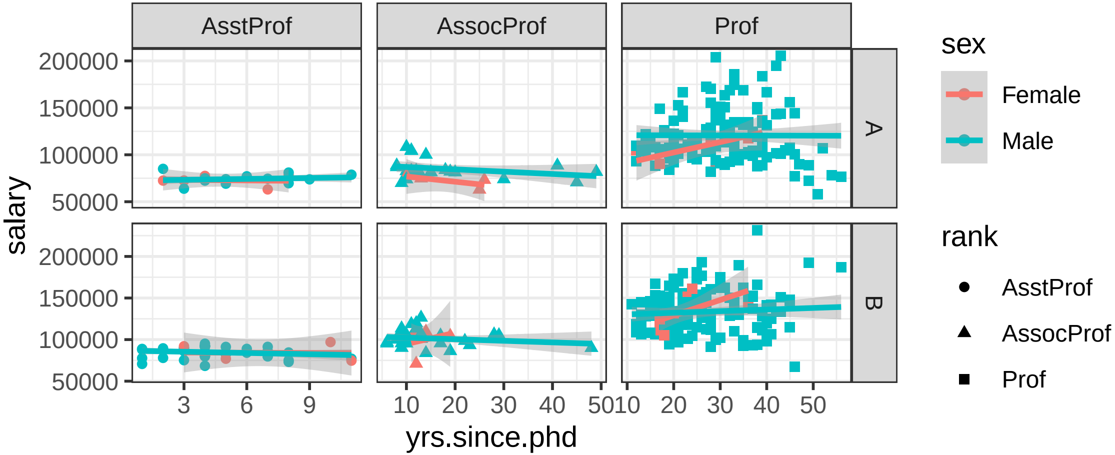
- Professors in
- Discipline B and
- doesn't account for years since PhD.

We can run a multiple regression model that includes

- sex,
- rank,
- discipline, and
- years since PhD.

This model will provide a comparison of the effect of sex on salary after *adjusting* for rank, discipline, and years since PhD.

# Multiple regression



# Multiple regression model

The simple linear regression model is

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots, \sigma^2)$$

where $Y_i$ and $X_{i,j}$ are the dependent and independent variable(s), respectively, for individual $i$.

To analyze salaries of Professors in discipline B at an unknown college in the U.S. from 2008-2009, we will use log salary as the dependent variable ($Y$)

- sex ($X1$),
- rank ($X2$ and $X3$),
- discipline ($X4$), and
- years since PhD ($X5$)

as independent variables. In this model, $100(e^{\beta_1} - 1)$ will be the percent difference in median salary of men compared to women after adjusting for rank, discipline, and years since PhD.

# Salary comparison

```r
summary(m <- lm(log(salary) ~ sex + rank + discipline + yrs.since.phd,
                data = Salaries))
```

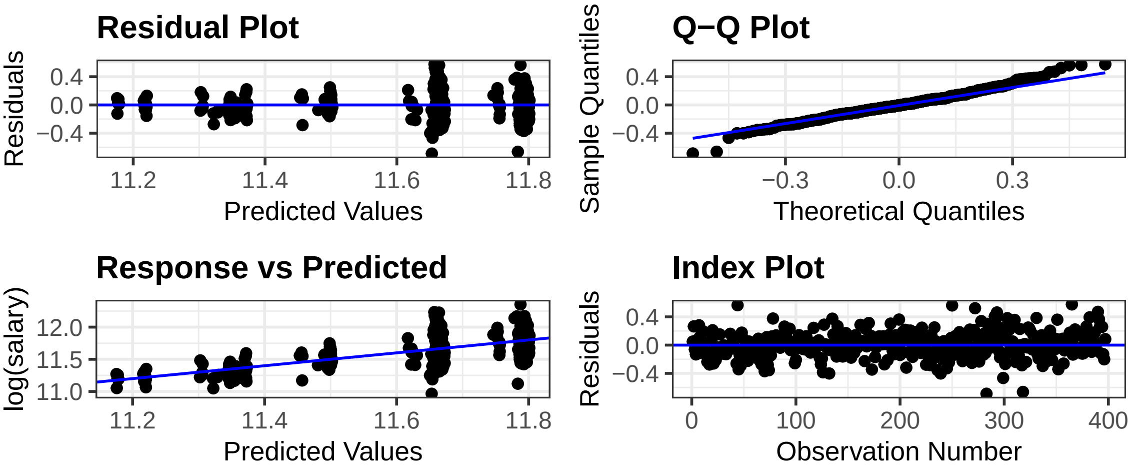
```
##
## Call:
## lm(formula = log(salary) ~ sex + rank + discipline + yrs.since.phd,
##     data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68837 -0.11190 -0.00583  0.09518  0.57604
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.1795836  0.0363782 307.316  < 2e-16 ***
## sexMale        0.0421082  0.0310741   1.355    0.176
## rankAssocProf  0.1553606  0.0334148   4.649 4.56e-06 ***
## rankProf       0.4571986  0.0339986  13.448  < 2e-16 ***
## disciplineB    0.1280259  0.0188152   6.804 3.82e-11 ***
## yrs.since.phd -0.0005054  0.0010184  -0.496    0.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1817 on 391 degrees of freedom
## Multiple R-squared:  0.5183,Adjusted R-squared:  0.5122
## F-statistic: 84.15 on 5 and 391 DF,  p-value: < 2.2e-16
```

# Salary comparison

```r
confint(m)

##                      2.5 %      97.5 %
## (Intercept)   11.108062322 11.25110489
## sexMale       -0.018984950  0.10320139
## rankAssocProf  0.089665510  0.22105578
## rankProf       0.390355769  0.52404150
## disciplineB    0.091034246  0.16501757
## yrs.since.phd -0.002507598  0.00149686
```

Manuscript statement:
Percentage difference in median salary between men and women is estimated to be between (-2, 11)% more for men compared to women after adjusting for rank, discipline, and years since PhD.

# Diagnostics



# Summary

- Consider (natural) logarithms when the variable
  - is strictly positive,
  - is non-negative (add smallest non-zero value to all observations), and
  - has a ratio (max/min) over 10.
- Interpretation:
  - When independent variable is logged, $100(e^{\beta} - 1)$ is the percent change in median response.
  - When dependent variable is logged, ...
  - When both are logged, ...

More details in my SLR using Logarithms video.