AE 11: Harris Trust & Savings Bank

Multiple linear regression

Your Name

2021-02-25

Starting salaries at Harris Trust and Savings Bank

```
library(tidyverse)
library(broom)
library(patchwork)
```

Example: Starting Wages

In the 1970s Harris Trust and Savings Bank was sued for discrimination on the basis of sex. The report from the Department of Labor states, "Prior to filing this case, Treasury retained two statistical experts, Drs. Shafie and Cabral, 'To explore the feasibility of using to determine the existence of an affected class of employees in the workforce of Treasury contractors.' "(Dept of Labor vs. Harris Trust and Savings).

Each side presented a statistical analysis to examine whether if there was sufficient evidence that female employees received lower starting salaries on average than male peers with similar qualifications.

We will take a look at some of the data used for the analyses. The data set contains information on 93 employees from a single job category (skilled, entry-level,clerical) who were hired between 1965 and 1975.

```
wages <- read_csv("data/wages.csv")
```

The variables in the data are

- `Educ`: years of education
- `Exper`: months of experience prior to working at the bank
- `Sex`: sex of employee
- `Senior`: months employed at Harris Trust & Savings Bank
- `Age`: age in months
- `Sal77`: salary as of March 1975
- `Bsal`: annual salary at time of hire

Today we will focus on the relationship between the following variables:

- Response: `Bsal`
- Predictors: `Senior`, `Educ`

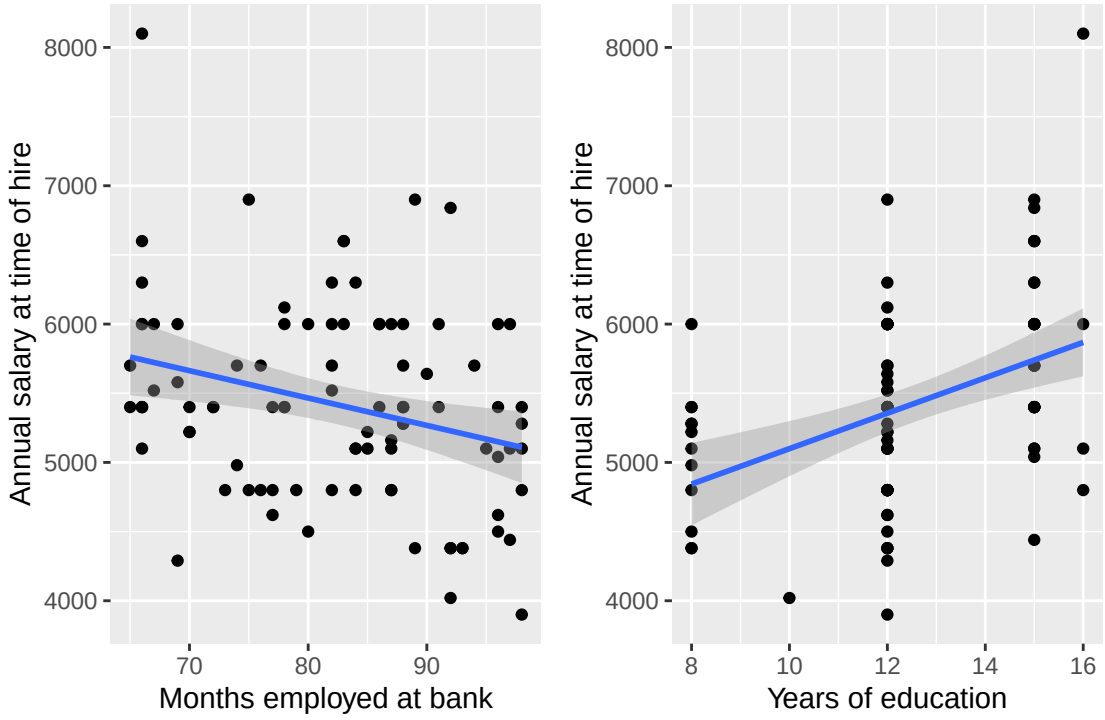
Univariate EDA

```
p1 <- ggplot(data = wages, aes(x = Bsal)) +
  geom_histogram() +
  labs(x = "Annual salary at time of hire")

p2 <- ggplot(data = wages, aes(x = Senior)) +
```

```
  geom_histogram() +
  labs(x = "Months employed at the bank")

p3 <- ggplot(data = wages, aes(x = Educ)) +
  geom_histogram() +
  labs(x = "Years of education")

p1 / (p2 + p3)
```

Bivariate EDA

```
p4 <- ggplot(data = wages, aes(x = Senior, y = Bsal)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Months employed at bank",
       y = "Annual salary at time of hire")

p5 <- ggplot(data = wages, aes(x = Educ, y = Bsal)) +
  geom_point() +
  geom_smooth(method = "lm") +
   labs(x = "Years of education",
       y = "Annual salary at time of hire")

p4 + p5
```



Simple linear regression

```
senior_mod <- lm(Bsal ~ Senior, data  = wages)
tidy(senior_mod, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   7048.     576.       12.2  6.50e-21   5903.    8193.
## 2 Senior         -19.8      6.95     -2.85 5.48e- 3    -33.6     -5.97
```

```
educ_mod <- lm(Bsal ~ Educ, data  = wages)
tidy(educ_mod, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)    3819.     377.      10.1  1.45e-16   3069.      4568.
## 2 Educ            128.      29.7      4.31 4.08e- 5     69.1      187.
```

Why multiple linear regression?

We would like to use the employees' years of education and time employed at the bank to understand variation in their starting salaries. Why do we want to do this fitting a multiple linear regression model instead of separate simple linear regression models for each predictor?

- One variable may be a confounding variable. Once I account for the relationship between the response and one predictor, the effect of the other predictor may change/ disappear.
- This is a more realistic view of the relationship between response and predictor variables.

Fit the model

Fit the linear regression model and output the results. Include `conf.int = TRUE` in the `tidy` function to display the confidence interval for each coefficient.

```
wages_model <- lm(Bsal ~ Senior + Educ, data = wages)
tidy(wages_model, conf.int = T)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   5520.     609.        9.06 2.57e-14   4310.    6731.
## 2 Senior         -21.6      6.26     -3.45 8.68e- 4    -34.0     -9.13
## 3 Educ           134.      28.1       4.76 7.31e- 6     78.0    190.
```

Interpreting coefficients

1. Interpret the coefficient of `Educ` in the context of the data.- POLL

For every additional year in education, we expect the starting salary to increase by $133.88, on average, holding months employed constant.

For every additional year in education, we expect the starting salary to increase by $133.88, on average, holding all other variables constant.

2. Interpret the 95% confidence interval for `Senior` in the context of the data.

-33.99521 -9.129636

slope / coefficient interchangeable

We're 95% confident that the true coefficient of Senior is between -33.995 to -9.130.

We're 95% confident that for each additional month employed, the starting salary is expected to be lower by $9.13 to $34.00, on average, holding the years of education constant.

I may want to include interaction if I think the effect of education on starting salary differs by sex.

$$Bsal = \beta_0 + \beta_1 Educ + \beta_2 Sex + \beta_3 Educ \times Sex + \epsilon, \epsilon \sim N(0, \sigma_\epsilon^2)$$

Heart Rate, Groups: Control, Pet, Friend

$$heart = \beta_0 + \beta_1 Pet + \beta_2 Friend + \epsilon$$

Hypothesis testing

Does `Educ` help explain some of the variability in the starting salary after accounting for `Senior`? Let's use a hypothesis test to answer this question.

1. State the null and alternative hypotheses in words and mathematical notation. - POLL

$$H_0 : \beta_{educ} = 0$$
$$H_a : \beta_{educ} \neq 0$$

Null: There is no linear relationship between education and starting salary, after accounting for months employed at the bank.

Alternative: There is linear relationship between education and starting salary, after accounting for months employed at the bank.

2. What is the test statistic? What does this value mean?

4.761525

We're looking at the distribution of $\hat{\beta}_{educ}$ in a model that also includes Senior.

Given the null is true, the mean of this distribution is 0, the observed coefficient of 133.88 is about 4.76 standard errors above the mean.

The observed coefficient of 133.88 is about 4.76 standard errors above the hypothesized mean of 0.

3. What distribution was used to calculate the p-value? - POLL

t distribution with n - p - 1 degrees of freedom

= t distribution with 93 - 2 - 1 = 90 degrees of freedom

Simple Linear Regression is special case of MLR with p = 1

n - p - 1 = n - 1 -1 = n - 2

4. State your conclusion in the context of the data.

7.309074e-06

0.04. Using a significance level of $\alpha = 0.05$, we reject...

Our p-value is very small, so we reject the null hypothesis. The data provide sufficient evidence that there is a linear relationship between education and starting salary, after adjusting for months employed at the bank.