Multiple Linear Regression

Dr. Maria Tackett

09.18.19

STA 210

[Click for PDF of slides]

Announcements

- HW 01 **due TODAY at 11:59p**
- Reading 03 for Monday
- HW 02 due Wednesday, 9/25 at 11:59p

STA 210

Today's Agenda

- Introducing multiple linear regression

STA 210

R Packages used in the notes

```
library(tidyverse)
library(knitr)
library(broom)
library(Sleuth3) # case 1202 dataset
library(cowplot) # use plot_grid function
```

Multiple Linear Regression

Example: Starting Wages

- In the 1970s Harris Trust and Savings Bank was sued for discrimination on the basis of gender.
- "Prior to filing this case, Treasury retained two statistical experts, Drs. Shafie and Cabral, 'To explore the feasibility of using **multiple regression analysis** to determine the existence of an affected class of employees in the workforce of Treasury contractors.'"- Dept of Labor vs. Harris Trust and Savings
- The defense presented an analysis of the salaries for skilled, entry-level clerical employees as evidence.
- **Question**: Did female employees receive lower starting salaries on average than male employees with similar experience and qualifications?

STA 210

Data

```
glimpse(wages)
```

```
## Observations: 93
## Variables: 6
## $ Bsal   <int> 5040, 6300, 6000, 6000, 6000, 6840, 8100, 6000, 6000, 690
## $ Senior <int> 96, 82, 67, 97, 66, 92, 66, 82, 88, 75, 89, 91, 66, 86, 9
## $ Age    <int> 329, 357, 315, 354, 351, 374, 369, 363, 555, 416, 481, 33
## $ Educ   <int> 15, 15, 15, 12, 12, 15, 16, 12, 12, 15, 12, 15, 15, 15, 1
## $ Exper  <dbl> 14.0, 72.0, 35.5, 24.0, 56.0, 41.5, 54.5, 32.0, 252.0, 13
## $ Female <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,
```

Variables

Explanatory

- **`Educ`:** years of education
- **`Exper`:** months of previous work experience (before hire at bank)
- **`Female`:** 1 if female, 0 if male
- **`Senior`:** months worked at bank since hire
- **`Age`:** age in months

Response

- **`Bsal`:** annual salary at time of hire

Salary comparison

- **Question:** Did female employees receive lower starting salaries on average than male employees with similar experience and qualifications?

Using ANOVA

$$H_0 : \mu_F = \mu_M$$
$$H_a : \mu_F \neq \mu_M$$

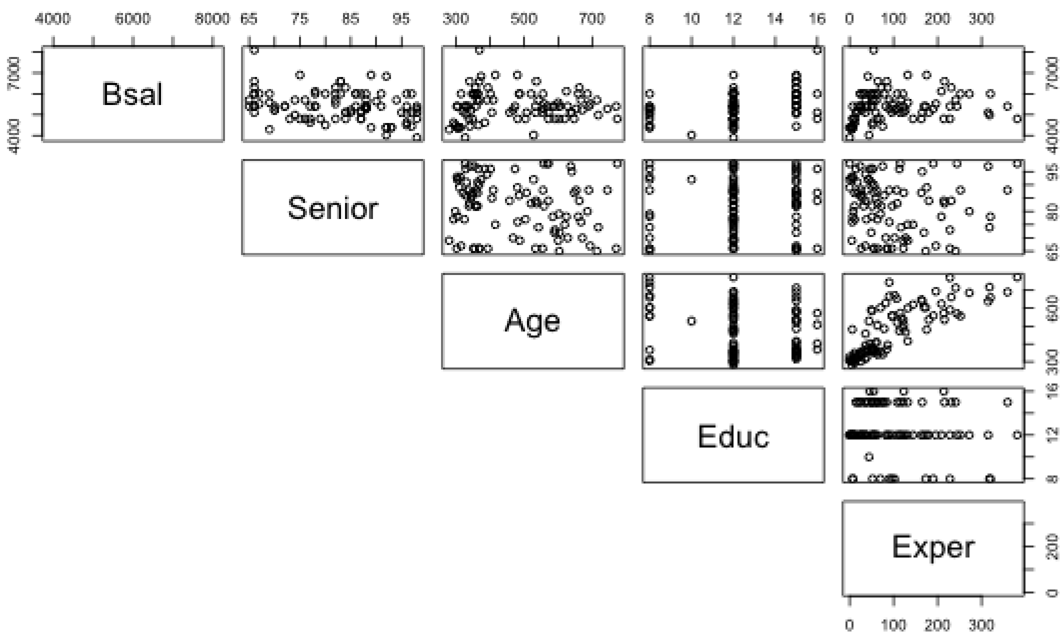
term	df	sumsq	meansq	statistic	p.value
Female	1	14045183	14045183.2	39.597	0
Residuals	91	32278107	354704.5	NA	NA

- What's your conclusion?
- What is a disadvantage to using this method to answer the question?

STA 210

Salary vs. Other Variables

```
pairs(Bsal ~ Senior + Age + Educ + Exper, data=wages,
      lower.panel = NULL)
```



Multiple Regression Model

- We will calculate a multiple linear regression model with the following form:

$$Bsal = \beta_0 + \beta_1 \text{Senior} + \beta_2 \text{Age} + \beta_3 \text{Educ} + \beta_4 \text{Exper} + \beta_5 \text{Female}$$

- Similar to simple linear regression, this model assumes that at each combination of the predictor variables, the values `Bsal` follow a Normal distribution

STA 210

Regression Model

- Recall: The simple linear regression model assumes

$$y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

- Similarly: The multiple linear regression model assumes

$$y|x_1, x_2, \dots, x_p \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \sigma^2)$$

- For a given observation $(x_{i1}, x_{i2} \dots, x_{iP}, y_i)$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

Regression Model

- At any combination of $x's$, the true mean value of y is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- We will use multiple linear regression to estimate the mean y for any combination of $x's$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Regression Output

```
bsal_model <- lm(Bsal ~ Senior + Age + Educ + Exper + Female,
                 data=wages)
kable(tidy(bsal_model),format="html",digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	6277.893	652.271	9.625	0.000
Senior	-22.582	5.296	-4.264	0.000
Age	0.631	0.721	0.876	0.384
Educ	92.306	24.864	3.713	0.000
Exper	0.501	1.055	0.474	0.636
Female1	-767.913	128.970	-5.954	0.000

Interpreting $\hat{\beta}_j$

- An estimated coefficient $\hat{\beta}_j$ is the amount y is expected to change when x_j increases by one unit **holding the values all other predictor variables constant**

- *Example:* The estimated coefficient for `Educ` is 92.31. This means for each additional year of education an employee has, we expect starting salary to increase by about $92.31, holding all other predictor variables constant.

STA 210

Hypothesis Tests for $\hat{\beta}_j$

- We want to test whether a particular coefficient has a value of 0 in the population, given all other variables in the model:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

- The test statistic reported in R is the following:

$$\text{test statistic} = t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

STA 210

Salary

term	estimate	std.error	statistic	p.value
(Intercept)	6277.893	652.271	9.625	0.000
Senior	-22.582	5.296	-4.264	0.000
Age	0.631	0.721	0.876	0.384
Educ	92.306	24.864	3.713	0.000
Exper	0.501	1.055	0.474	0.636
Female1	-767.913	128.970	-5.954	0.000

Given the other variables in the model, are the following significant predictors of salary at time of hire (`Bsal`)?

- Education (`Educ`)
- Experience (`Exper`)

STA 210

Confidence Interval for β_j

The C confidence interval for β_j

$$\hat{\beta}_j \pm t^* SE(\hat{\beta}_j)$$

where t^* follows a t distribution with with $(n-p-1)$ degrees of freedom

- **General Interpretation**: We are C confident that the interval LB to UB contains the population coefficient of x_j. Therefore, for every one unit increase in x_j, we expect y to change LB to UB units, holding all else constant.

STA 210

CI for `Educ`

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	6277.893	652.271	9.625	0.000	4981.434	7574.353
Senior	-22.582	5.296	-4.264	0.000	-33.108	-12.056
Age	0.631	0.721	0.876	0.384	-0.801	2.063
Educ	92.306	24.864	3.713	0.000	42.887	141.725
Exper	0.501	1.055	0.474	0.636	-1.597	2.598
Female1	-767.913	128.970	-5.954	0.000	-1024.255	-511.571

Interpret the 95% confidence interval for the coefficient of `Educ`.

Notes about CI and Hypothesis Tests

- If the sample size is large enough, the test will likely result in rejecting $H_0 : \beta_j = 0$ even x_j has a very small effect on y
  - Consider the practical significance of the result not just the statistical significance
  - Use the confidence interval to draw conclusions instead of p-values
- If the sample size is small, there may not be enough evidence to reject $H_0 : \beta_j = 0$
  - When you fail to reject the null hypothesis, **DON'T** immediately conclude that the variable has no association with the response.
  - There may be a linear association that is just not strong enough to detect given your data, or there may be a non-linear association.

STA 210

Prediction

- We calculate predictions the same as with simple linear regression
- **Example:** Suppose we want to predict the starting wages for a female who is 28 years old with 12 years of education, 11 months seniority and 2 years of prior experience.

$$\begin{aligned}\widehat{bsal} = & 6277.893 - 22.582 \times \text{Senior} + 0.631 \times \text{Age} \\ & + 92.306 \times \text{Educ} + 0.501 \times \text{Exper} - 767.913 \times \text{Female}\end{aligned}$$

```
6277.893 - 22.582 * 11 + 0.631 * 28 + 92.306 * 12 + 0.501 * 24 - 7
```

```
## [1] 6398.942
```

Prediction

- Just like with simple linear regression, we can use the **`predict.lm()`** function in R to calculate the appropriate intervals for our predicted values

- Suppose we want to predict the starting wages for a female who is 28 years old with 12 years of education, 11 months seniority and 2 years of prior experience.

```
x0 <- data.frame(Senior= 11, Age = 28, Educ = 12, Exper = 24, Fema
predict.lm(bsal_model, x0, interval = "prediction")
```

```
##       fit      lwr      upr
## 1 6398.93 4967.054 7830.805
```

Prediction

Suppose we want to predict the mean age for the subset of all females who are 28 years old with 12 years of education, 11 months of seniority and 2 years of prior experience.

- How will the predicted value change?
- How will the interval change?

```
x0 <- data.frame(Senior= 11, Age = 28, Educ = 12, Exper = 24, Fema
predict.lm(bsal_model, x0, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 6398.93 5383.844 7414.016
```

Cautions

- **Do not extrapolate!** Because there are multiple explanatory variables, you can extrapolation in many ways
- The multiple regression model only shows **association, not causality**
  - To prove causality, you must have a carefully designed experiment or carefully account for confounding variables in an observational study

STA 210