Multiple Linear Regression

Assumptions & Special Predictors

Dr. Maria Tackett

09.23.19

STA 210

[Click for PDF of slides]

Announcements

- Lab 04 **due tomorrow at 11:59p**
- HW 02 **due Wednesday, 9/25 at 11:59p**
- Team Feedback #1 **due Wednesday, 9/25 at 11:59p**
  - Please provide honest and constructive feedback. This team feedback will be graded for completion.

Today's agenda

- Math details of multiple linear regression
- Assumptions for multiple linear regression
- Special predictors

R packages

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

Starting wages data

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

Starting wages

```
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, …
```

Regression model

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

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

Math Details

Regression Model

- The multiple linear regression model assumes

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

- For a given observation $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, we can rewrite the previous statement as

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

Estimating σ^2

- For a given observation $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i)$ the residual is

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip})$$

- The estimated value of the regression variance , σ^2, is

$$\hat{\sigma}^2 = \frac{RSS}{n-p-1} = \frac{\sum_{i=1}^n e_i^2}{n-p-1}$$

STA 210

Estimating Coefficients

- One way to estimate the coefficients is by taking partial derivatives of the formula

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \cdots + \beta_p x_{ip})]^2$$

- This produces messy formulas, so instead we can use matrix notation for multiple linear regression and estimate the coefficients using rules from linear algebra.
  - For more details, see Section 1.2 of the textbook and the supplemental notes Matrix Notation for Multiple Linear Regression
  - **Note:** You are <u>not</u> required to know matrix notation for MLR in this class

STA 210

Assumptions

Assumptions

Inference on the regression coefficients and predictions are reliable only when the regression assumptions are reasonably satisfied:

1. **Linearity:** Response variable has a linear relationship with the predictor variables in the model

2. **Constant Variance:** The regression variance is the same for all set of predictor variables $(x_1, \ldots, x_p)$

3. **Normality:** For a given set of predictors $(x_1, \ldots, x_p)$, the response, y, follows a Normal distribution around its mean

4. **Independence:** All observations are independent

Scatterplots

- Look at a scatterplot of the response variable vs. each of the predictor variables in the exploratory data analysis before calculating the regression model
- This is a good way to check for obvious departures from linearity
  - Could be an indication that a higher order term or transformation is needed

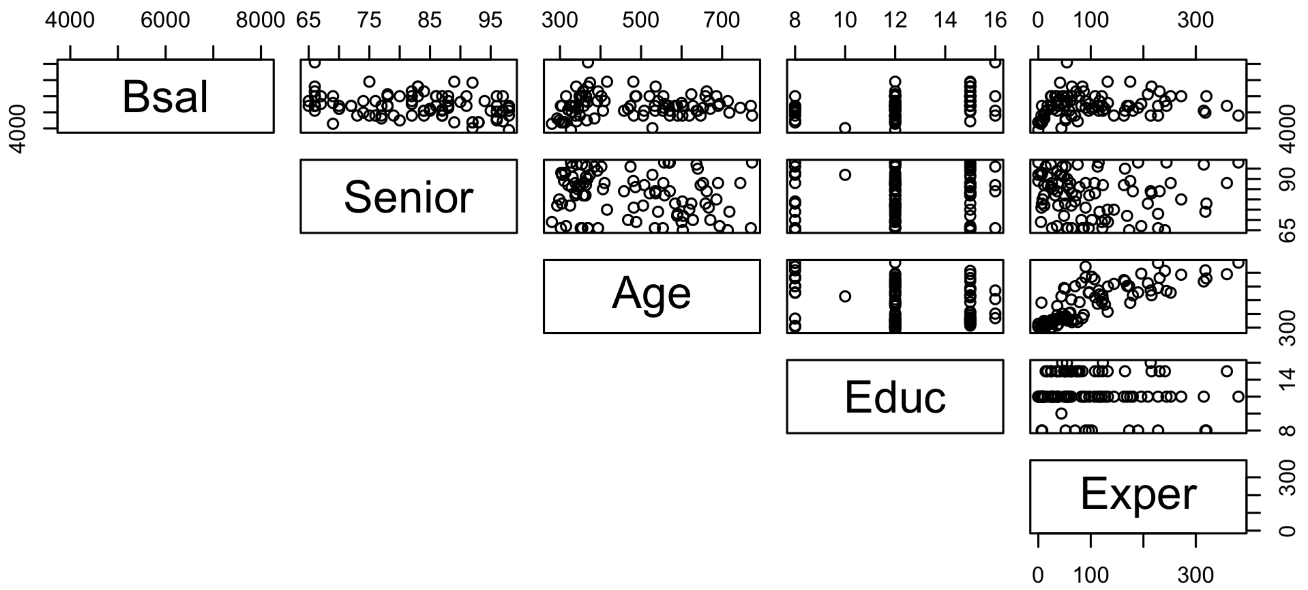
Residual Plots

- **Plot the residuals vs. the predicted values**
  - Can expose issues such at outliers or non-constant variance
  - Should have no systematic pattern
- **Plot the residuals vs. each of the predictors**
  - Can expose issues between the response and a predictor variable that didn't show in the exploratory data analysis
  - Use box plots to plot residuals versus categorical predictor variables
  - Should have no systematic pattern
- **Plot a histogram and QQ-plot of the residuals**
  - Check normality

STA 210

Scatterplots

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

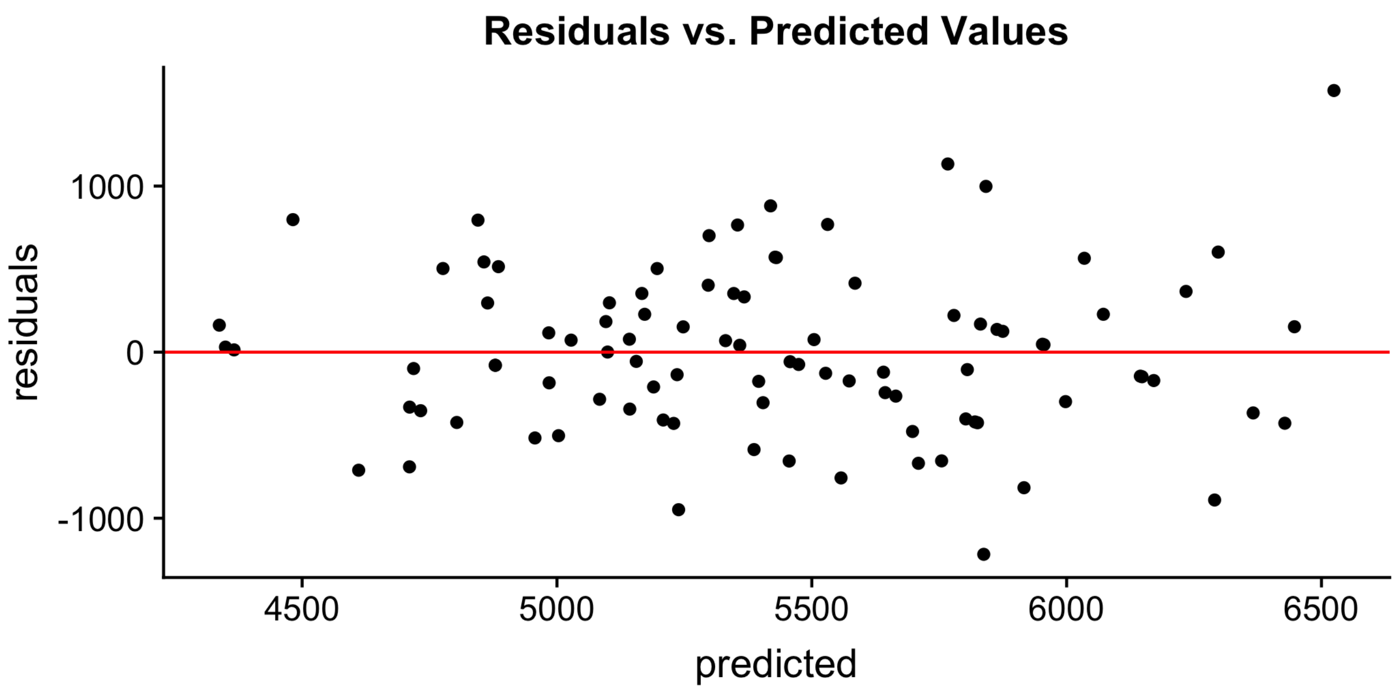


- Only include a 4 - 5 variables in a single pairs plot; otherwise, the scatterplots are too small to be readable

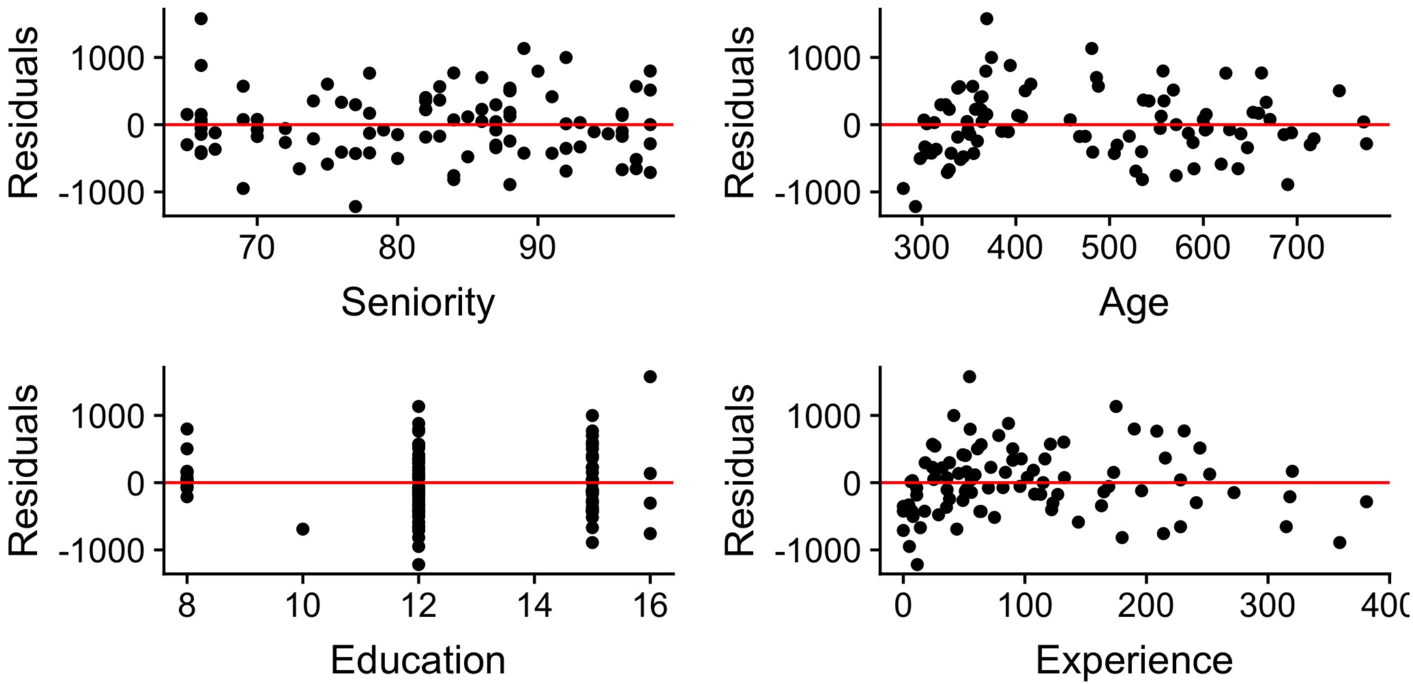
STA 210

Residuals vs. Predicted Values

```
wages <- wages %>%
  mutate(predicted = predict.lm(bsal_model), residuals = resid(bsa
ggplot(data=wages,aes(x=predicted, y=residuals)) +
  geom_point() +
  geom_hline(yintercept=0,color="red") +
  labs(title="Residuals vs. Predicted Values")
```

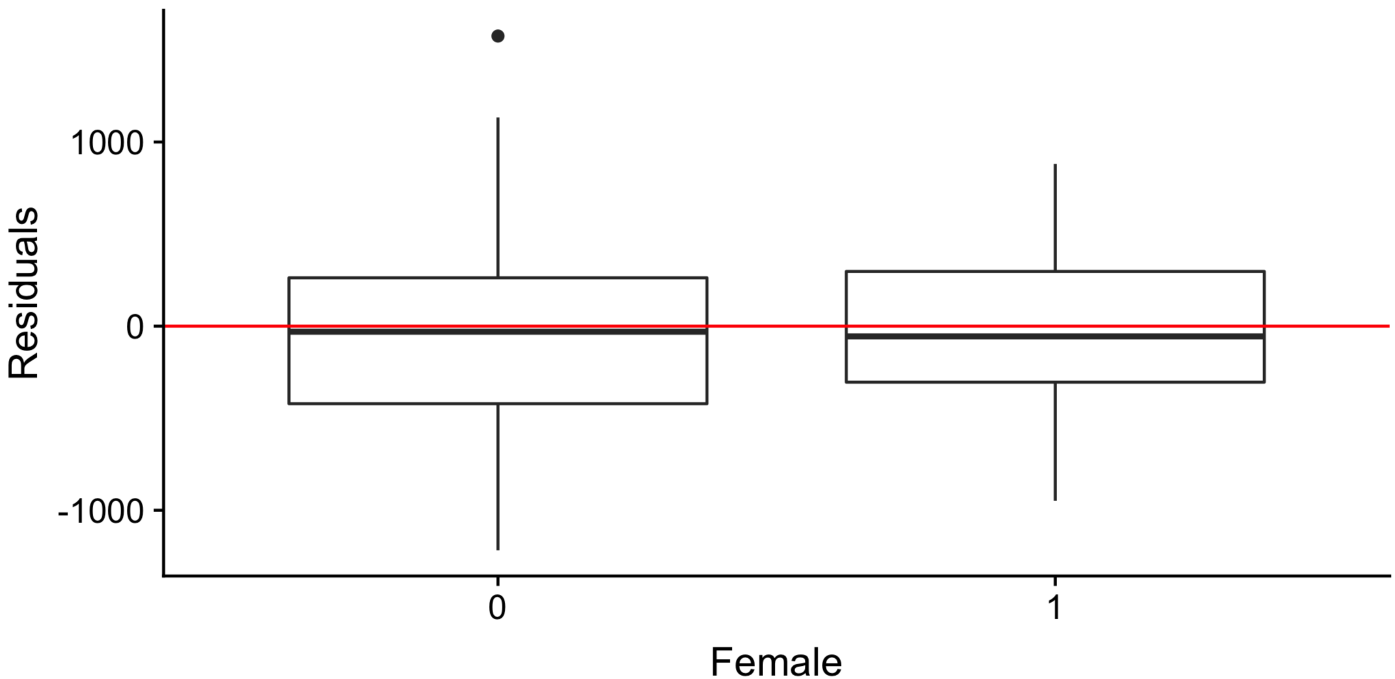


Residuals vs. Predictors

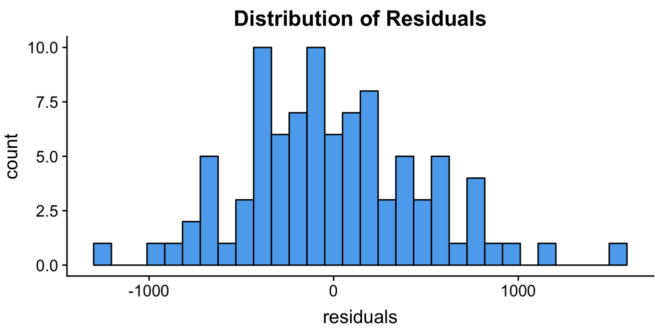


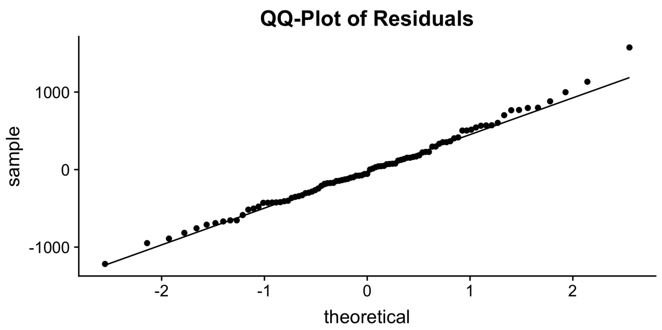
Residuals vs. Predictors

```
ggplot(data=wages,aes(x=Female,y=residuals)) +
  geom_boxplot() +
  geom_hline(yintercept=0,color="red") +
  labs(x = "Female",
       y="Residuals")
```



Normality of Residuals





Special Predictors

STA 210

Interpreting the Intercept

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

- Interpret the intercept.
- Is this interpretation meaningful? Why or why not?

STA 210

Mean-Centered Variables

- To have a meaningful interpretation of the intercept, use **mean-centered** predictor variables in the model (quantitative predictors only)

- A mean-centered variable is calculated by subtracting the mean from each value of the variable, i.e.

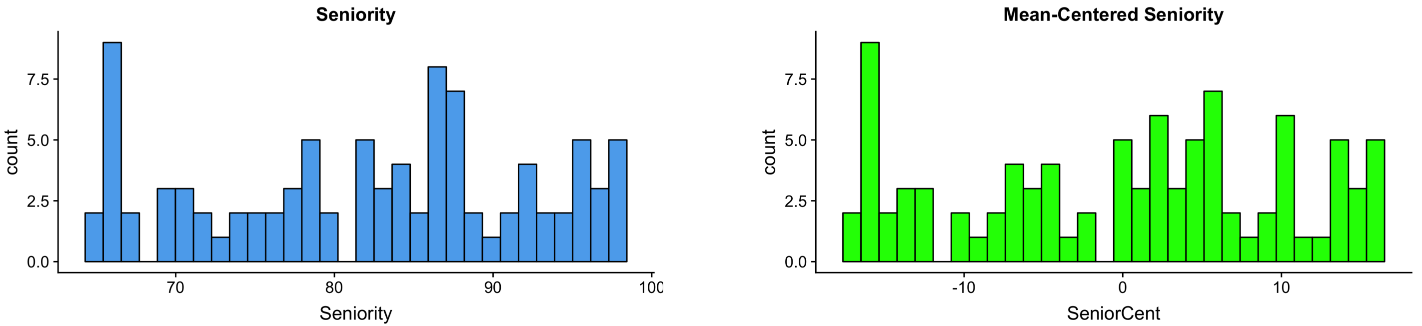
$$x_{ip} - \bar{x}_{.p}$$

- Now the intercept is interpreted as the expected value of the response at the mean value of all quantitative predictors

STA 210

Salary: Mean-Centered Variables

```
wages <- wages %>%
  mutate(SeniorCent = Senior - mean(Senior),
         AgeCent = Age-mean(Age),
         EducCent = Educ - mean(Educ),
         ExperCent = Exper - mean(Exper))
```



Salary: Mean-Centered Variables

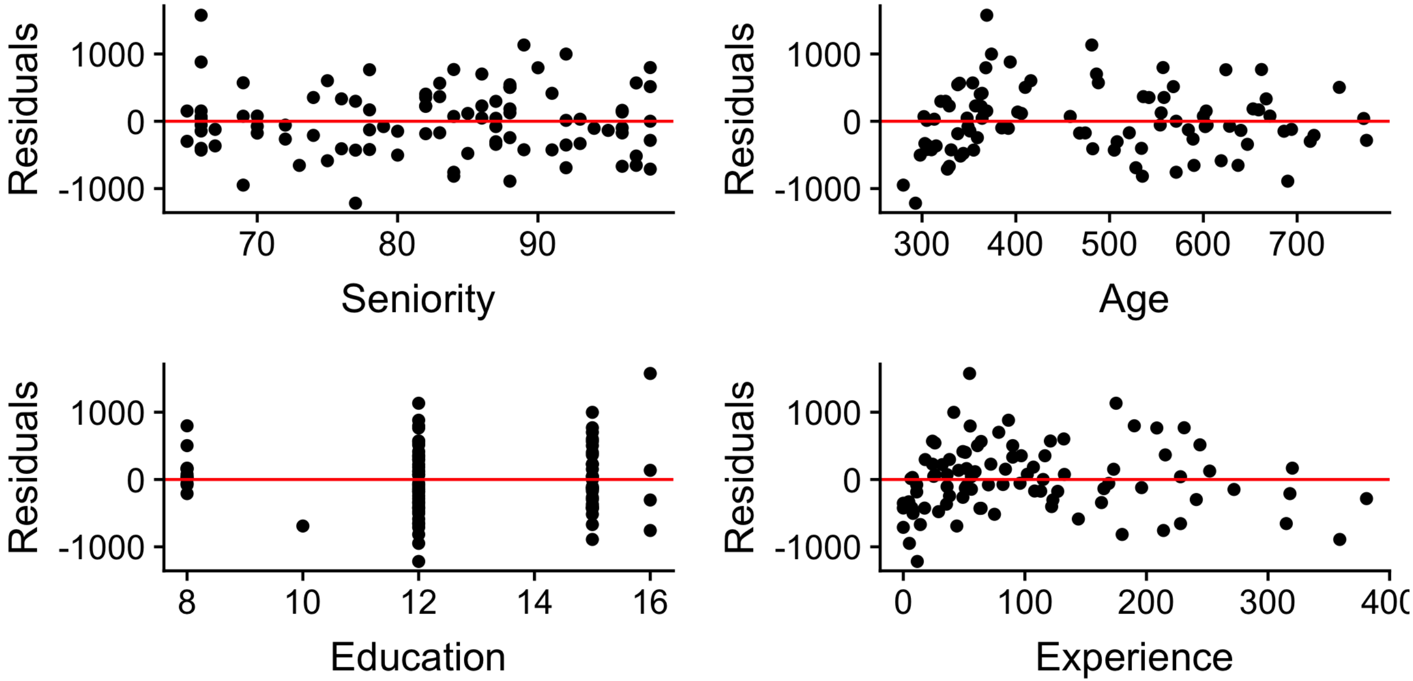
Application Exercise: Fit the regression model using the mean-centered variables.

- How did the model stay the same?
- How did the model change?

Quadratic Terms

- Sometimes the response variable may have a quadratic relationship with one or more predictor variables
    - You can see this in a plot of the residuals vs. a predictor variable
    - Include quadratic terms in the model to capture the relationship
- **Good Practice:** Also include all lower order terms even if they are not significant.
    - This helps with interpretation
- You can show quadratic relationships by plotting the predicted mean response for different values of the predictors variable
- Note: The same ideas apply for higher-order polynomial terms

Below are plots of the residuals versus each quantitative predictor variable.



Which variables (if any) appear to have a quadratic relationship with `Bsal`?

Indicator (dummy) variables

- Suppose there is a categorical variable with k levels (categories)
- Make k indicator variables (also known as dummy variables)
- Use $k-1$ of the indicator variables in the model
  - Can't uniquely estimate all k variables at once if the intercept is in the model
- Level that doesn't have a variable in the model is called the **baseline**
- Coefficients interpreted as the change in the mean of the response over the baseline

Indicator variables when $k = 2$

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5924.007	99.659	59.443	0.000	5725.925	6122.090
Female1	-767.913	128.970	-5.954	0.000	-1024.255	-511.571
SeniorCent	-22.582	5.296	-4.264	0.000	-33.108	-12.056
AgeCent	0.631	0.721	0.876	0.384	-0.801	2.063
EducCent	92.306	24.864	3.713	0.000	42.887	141.725
ExperCent	0.501	1.055	0.474	0.636	-1.597	2.598

- Write the model equation used to predict the starting salary for males.
- Write the model equation used to predict the starting salary for males.

Indicator variables when $k > 2$

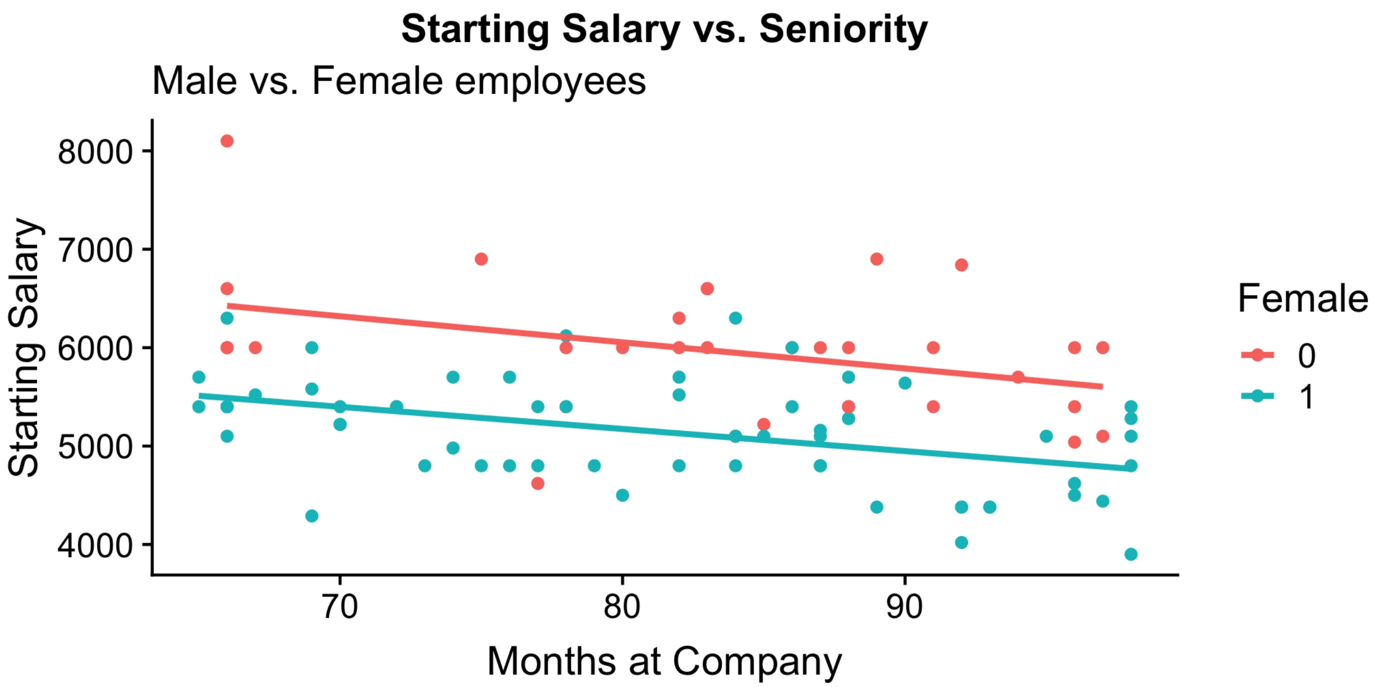
Application Exercise: Fit a regression model with Education treated as a categorical variable.

- What is the baseline for Education?
- Interpret the coefficient for `EducCat16`.
- What is your conclusion from the p-value of `EducCat16`?
- Write the model equation for those with 8 years of education.
- Write the model equation for those with 16 years of education.

Interaction Terms

- **Case:** Relationship of the predictor variable with the response depends on the value of another predictor variable
  - This is an **interaction effect**
- Create a new interaction variable that is one predictor variable times the other in the interaction
- **Good Practice:** When including an interaction term, also include the associated **main effects** (each predictor variable on its own) even if they are not statistically significant

Interaction effects



Do you think there is a significant interaction effect between `Female` and `Senior`? Why or why not?

Before next class

- Review Reading 03 on special predictors
- Reading 04 on transformations