Multiple Linear Regression Additional Variable Types

Module 6

DATA 5600
Introduction to Regression and Machine Learning for Analytics
Marc Dotson

Module Overview

- Categorical variables
- Interaction variables
- Higher-order variables

Categorical Variables

Salary Data

- Are company pay guides being followed?
- What is the response variable Y, and is it continuous or categorical?
- What are the covariates, and are they continuous or categorical?

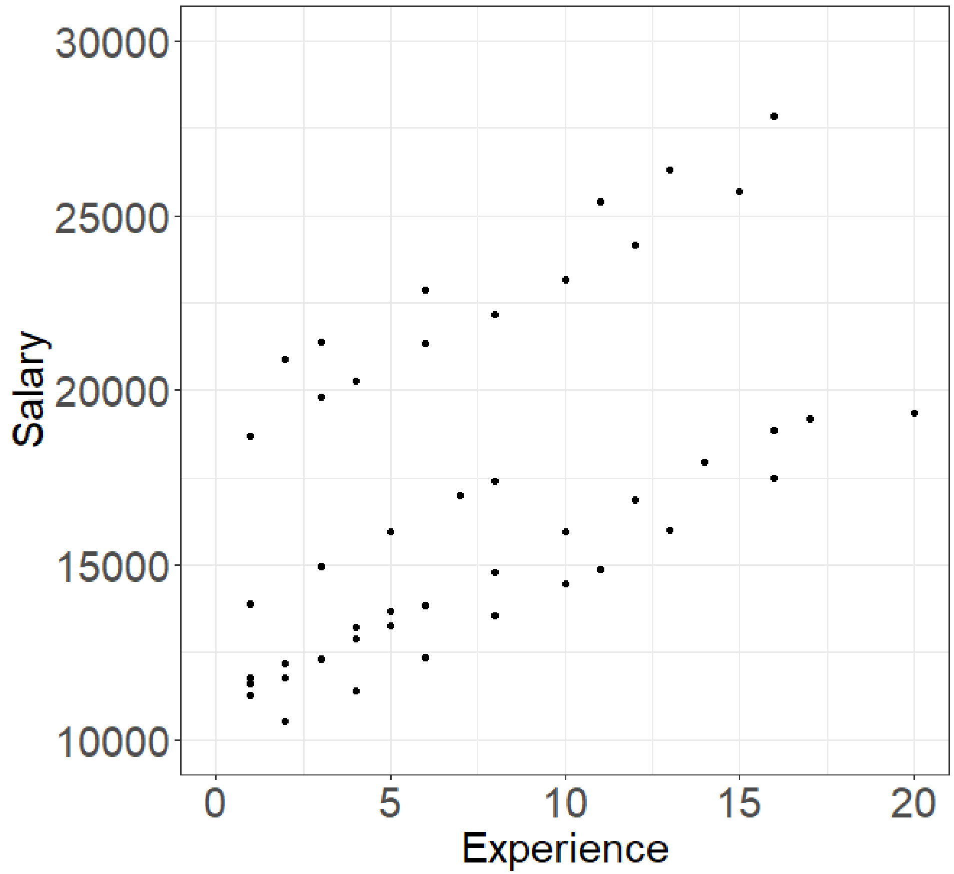
Salary (quarterly)	Experience (in years)	Education	Manager
13876	1	HS	Yes
11608	1	BS+	No
18701	1	BS+	Yes
11283	1	BS	No
	…		
19346	20	HS	No

- How do we deal with categorical variables in regression?

Categorical variables are often called “factors.” Each possible value within the factor is called a “level.”

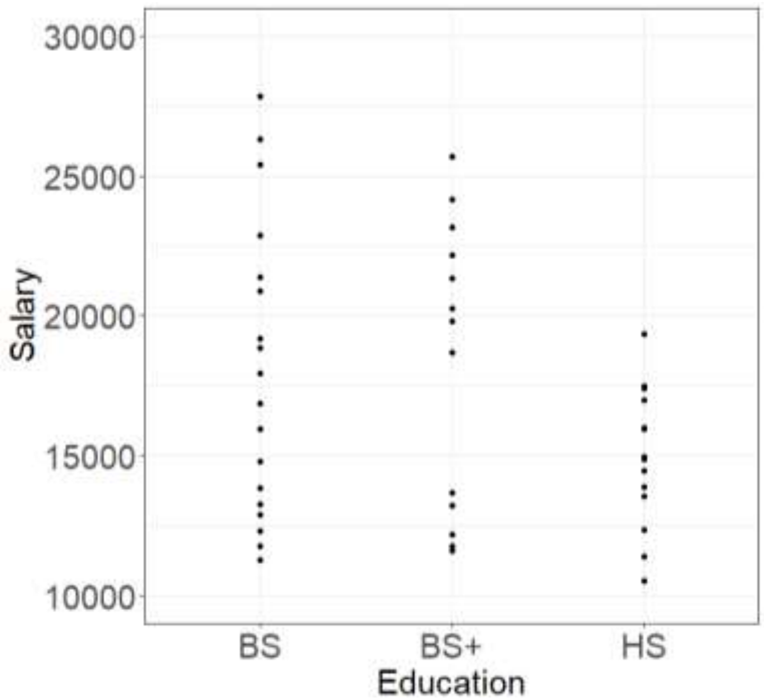
Salary Data

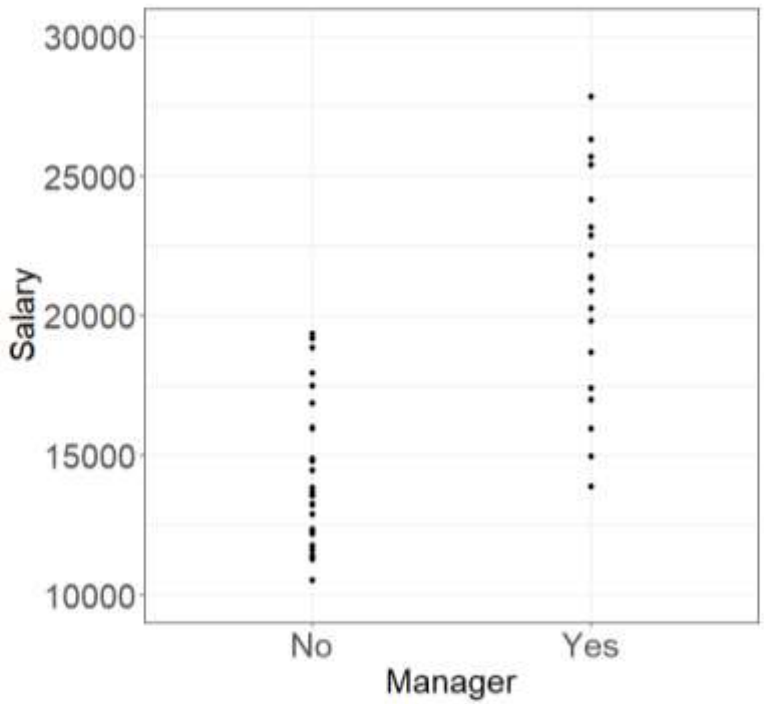
- We know how to visually compare two continuous variables



Salary Data

- But, a scatterplot doesn't work very well when one variable is categorical

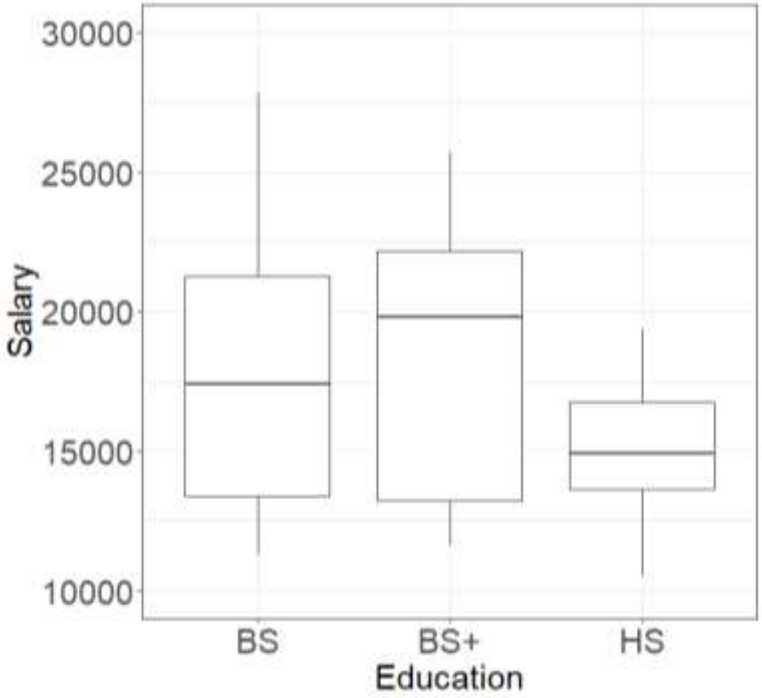


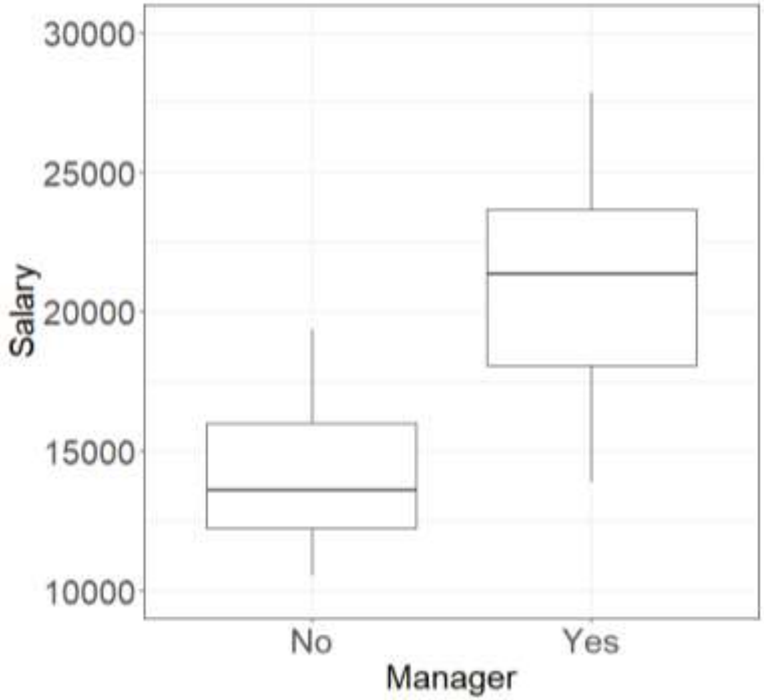


Salary Data: Side-by-Side Boxplots

Categorical

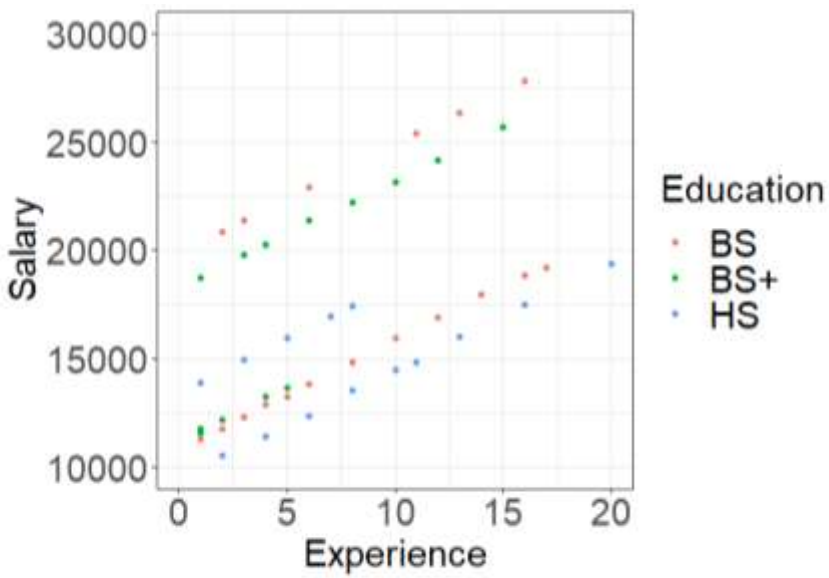
- Instead, we can use boxplots to visualize categorical variables

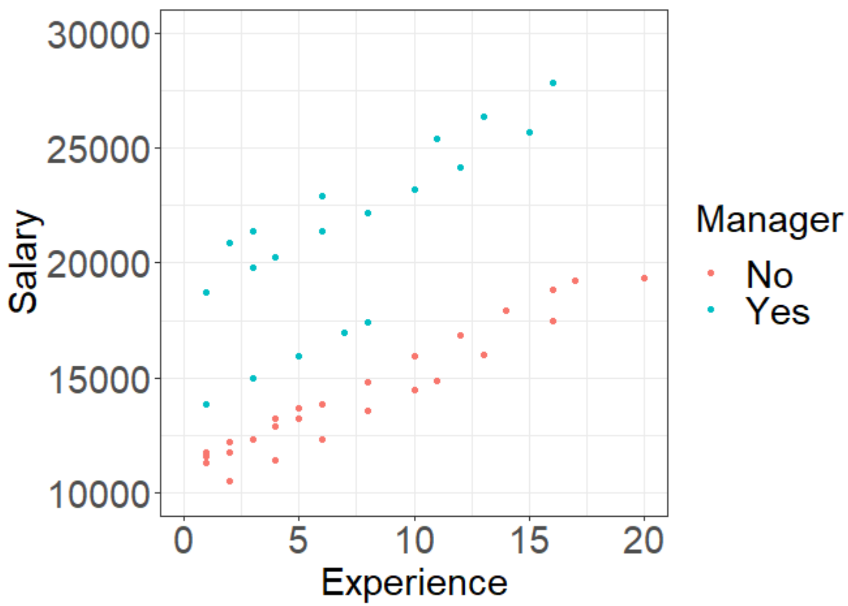




Salary Data: Color-Coded Scatterplots

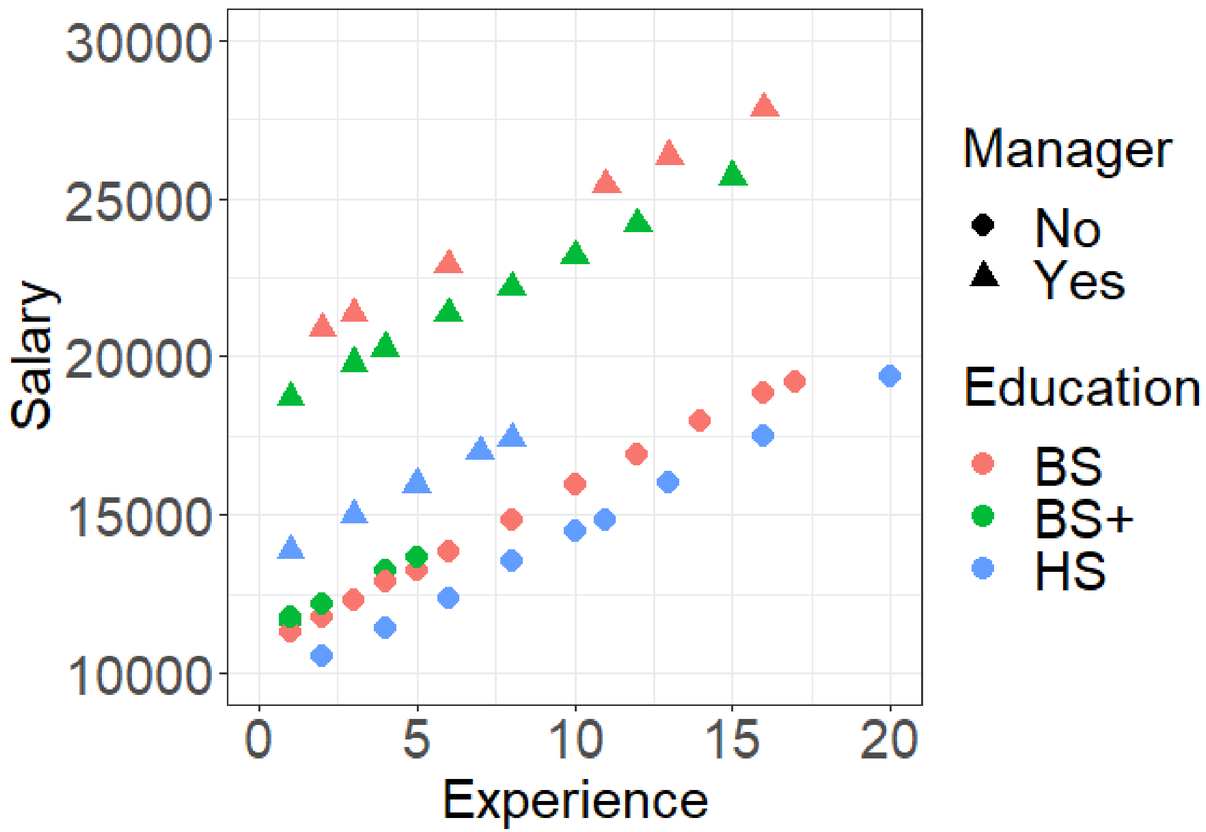
- Alternatively, we could create plots like these





Salary Data: Color- and Shape-Coded Scatterplots

Categorical



Code!

Multiple Linear Regression

Categorical

- We want to use the multiple linear regression model:

$$\text{Salary}_i = \beta_0 + \beta_1 \times \text{Experience}_i + \beta_2 \times \text{Education}_i + \beta_3 \times \text{Manager}_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- But how do we stick categories into a math function?
  - Ex: Education_i = BS+

- Answer: We use indicator variables.

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Note: indicator variables are also called “dummy” variables, and the process of creating an indicator variable from categorical data is called “coding” or “dummy coding,” among other things

Multiple Linear Regression

- A correctly written model:

$$\begin{aligned}\text{Salary}_i = \beta_0 + \beta_1 \times \text{Experience}_i + \\ \beta_2 \times \text{I}(\text{Education}_i = \text{BS}) + \\ \beta_3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ \beta_4 \times \text{I}(\text{Manager}_i = \text{Yes}) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned}$$

What about “HS” and “Not a Manager”?

Multiple Linear Regression

- The design matrix $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{bmatrix} 1 & \text{Exp}_1 & \text{EduBS}_1 & \text{EduBSp}_1 & \text{Man}_1 \\ 1 & \text{Exp}_2 & \text{EduBS}_2 & \text{EduBSp}_2 & \text{Man}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \text{Exp}_n & \text{EduBS}_n & \text{EduBSp}_n & \text{Man}_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 20 & 0 & 0 & 0 \end{bmatrix}$$

In python, the data set you use for modelling would not be the original data. It would be this one with dummy variables.

Salary (quarterly)	Experience (in years)	Education	Manager
13876	1	HS	Yes
11608	1	BS+	No
	…		
19346	20	HS	No

Indicator Variable Notes

- In general, for a categorical predictor variable with q levels, you will need to code $q-1$ indicator variables. If we incorrectly include q indicator variables in the model:
    - the model would be "over-parameterized" – redundant and unnecessary information
    - we cannot compute $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbf{X}'\mathbf{Y}$ since $\boldsymbol{X}'\boldsymbol{X}$ would be singular (non-invertible)

- The level not coded becomes absorbed into the intercept term and is called the "baseline" or "reference" level

Variable Selection Notes

- When performing variable selection, if at least one level of a categorical variable is "significant," you should leave in all other levels of that variable, even if the other levels are not significant.
    - If you remove only some levels of a variable and keep others in, your reference group changes (picks up the categories you removed, and you can artificially change p-values)

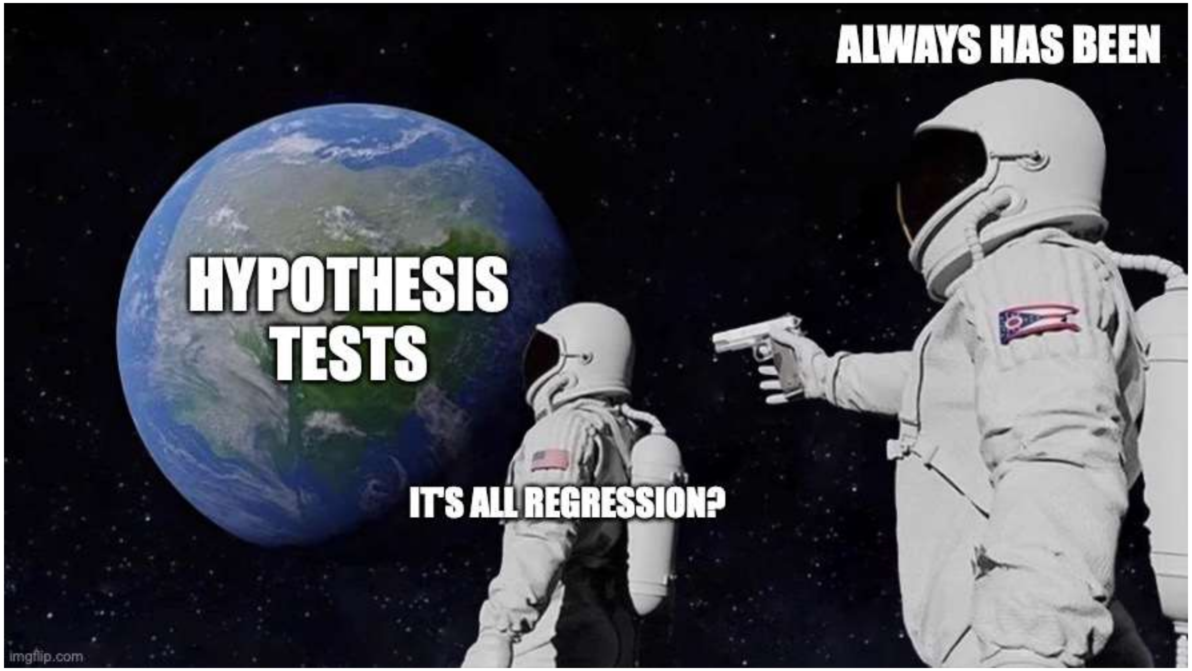
- Cross-validation requires stratification to make sure all available levels show up in both training and testing data.

Fun Facts

- A model with only one categorical predictor

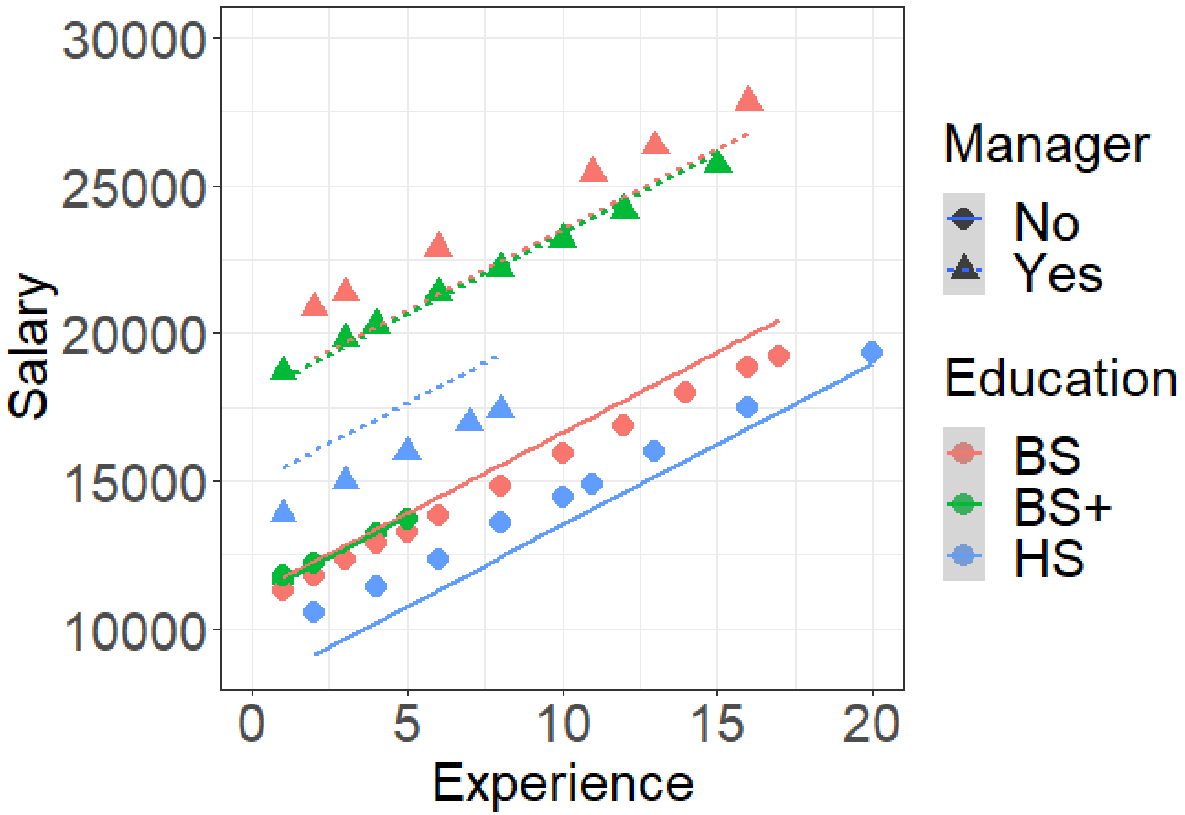
$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \,, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

  - Is equivalent to a two-sample t-test $(H_0 \colon \beta_1 = 0)$
  - Is a special case of the one-way ANOVA model


ALWAYS HAS BEEN
HYPOTHESIS TESTS
IT'S ALL REGRESSION?
imgflip.com

Salary Data Model Fit

$$\widehat{\text{Salary}}_i = 8044.75 + 545.79 \times \text{Experience}_i + 3129.53 \times \text{I}(\text{Education}_i = \text{BS}) + 2999.45 \times \text{I}(\text{Education}_i = \text{BS+}) + 6866.99 \times \text{I}(\text{Manager}_i = \text{Yes})$$



Multiple Linear Regression

Interpret

HS Baseline

	coef	std err	t	P>|t|	[0.025	0.975]
const	8044.7518	392.781	20.482	0.000	7250.911	8838.592
Experience	545.7855	30.912	17.656	0.000	483.311	608.260
Education_BS	3129.5286	370.470	8.447	0.000	2380.780	3878.277
Education_BS+	2999.4451	416.712	7.198	0.000	2157.238	3841.652
Manager_Yes	6866.9856	323.991	21.195	0.000	6212.175	7521.796

BS Baseline

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.117e+04	370.814	30.134	0.000	1.04e+04	1.19e+04
Experience	545.7855	30.912	17.656	0.000	483.311	608.260
Education_HS	-3129.5286	370.470	-8.447	0.000	-3878.277	-2380.780
Education_BS+	-130.0835	398.157	-0.327	0.746	-934.789	674.622
Manager_Yes	6866.9856	323.991	21.195	0.000	6212.175	7521.796

BS+ Baseline

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.104e+04	390.631	28.273	0.000	1.03e+04	1.18e+04
Experience	545.7855	30.912	17.656	0.000	483.311	608.260
Education_HS	-2999.4451	416.712	-7.198	0.000	-3841.652	-2157.238
Education_BS	130.0835	398.157	0.327	0.746	-674.622	934.789
Manager_Yes	6866.9856	323.991	21.195	0.000	6212.175	7521.796

Code!

Interpreting Coefficients

Multiple Linear Regression

- The fitted model is:

$$\begin{aligned}\widehat{\text{Salary}}_i = {} & 8044.75 + \\ & 545.79 \times \text{Experience}_i + \\ & 3129.53 \times \text{I}(\text{Education}_i = \text{BS}) + \\ & 2999.45 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ & 6866.99 \times \text{I}(\text{Manager}_i = \text{Yes})\end{aligned}$$

- How do you interpret the intercept?

- How do you interpret the coefficient for Experience_i?

Multiple Linear Regression

- The fitted model is:

$$
\begin{aligned}
\widehat{\text{Salary}}_i = \; & 8044.75 + \\
& 545.79 \times \text{Experience}_i + \\
& 3129.53 \times \text{I}(\text{Education}_i = \text{BS}) + \\
& 2999.45 \times \text{I}(\text{Education}_i = \text{BS+}) + \\
& 6866.99 \times \text{I}(\text{Manager}_i = \text{Yes})
\end{aligned}
$$

- How do you interpret the coefficient for $\text{I}(\text{Manager}_i \; = \text{Yes})$?

Multiple Linear Regression

- The fitted model is:

$$\widehat{\text{Salary}}_i = 8044.75 + 545.79 \times \text{Experience}_i + 3129.53 \times \text{I}(\text{Education}_i = \text{BS}) + 2999.45 \times \text{I}(\text{Education}_i = \text{BS+}) + 6866.99 \times \text{I}(\text{Manager}_i = \text{Yes})$$

- How do you interpret the coefficient for $\text{I}(\text{Education}_i = \text{BS})$?

- How do you interpret the coefficient for $\text{I}(\text{Education}_i = \text{BS+})$?

- For equal years of experience and managerial levels, how much does average quarterly salary increase with a BS+ compared to a BS degree?

CI for the Slope

- A 95% confidence interval for β_4 (coefficient for $\mathrm{I}(\mathrm{Manager}_i = \mathrm{Yes})$) is

$$6866.99 \pm 2.02 \times 323.99 = (6212.18, 7521.80)$$

- How do you interpret this interval?

Testing the Entire Categorical Variable

For the salary data, suppose we want to test if education has an effect on average salary

$$\begin{aligned}\text{Salary}_i = \beta_0 + \beta_1 \times \text{Experience}_i + \\ \beta_2 \times \text{I}(\text{Education}_i = \text{BS}) + \\ \beta_3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ \beta_4 \times \text{I}(\text{Manager}_i = \text{Yes}) + \epsilon_i, \\ \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned}$$

vs

$$\begin{aligned}\text{Salary}_i = \beta_0 + \beta_1 \times \text{Experience}_i + \\ \beta_4 \times \text{I}(\text{Manager}_i = \text{Yes}) + \epsilon_i, \\ \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned}$$

The hypothesis are

$$H_0: \beta_2 = \beta_3 = 0$$
$$H_a: \text{at least one coefficient is non-zero}$$

$$F = 41.75 \sim F_{5-1-2, 45-5} \rightarrow p - value \approx 0$$

What is the conclusion?

CI and PI: Salary Data

$$\widehat{\text{Salary}}_i = 8044.75 + 545.79 \times \text{Experience}_i + 3129.53 \times \text{I}(\text{Education}_i = \text{BS}) + 2999.45 \times \text{I}(\text{Education}_i = \text{BS+}) + 6866.99 \times \text{I}(\text{Manager}_i = \text{Yes})$$

- The average salary for a manager with a BS education and 10 years experience is

$$\widehat{\text{Salary}}_i = 8044.75 + 545.79 \times 10 + 3129.53 \times 1 + 2999.45 \times 0 + 6866.99 \times 1 = 23{,}499.12$$

- 95% <u>confidence interval</u> for the average salary for a manager with a BS education and 10 years experience is

$$(22829.88, 24168.36)$$

- Interpretation:

- 95% <u>prediction interval</u> for the salary of a manager with a BS education and 10 years experience:

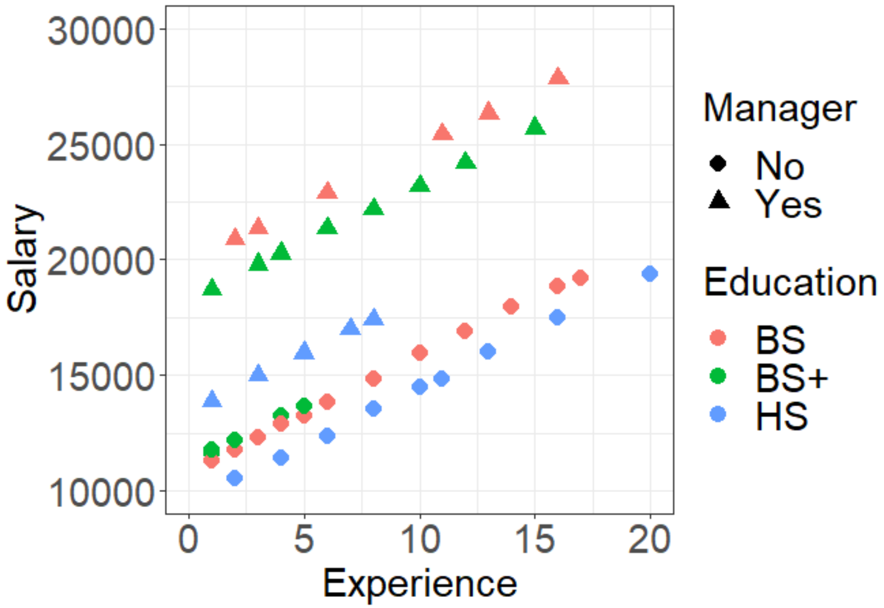
$$(21294.50, 25703.74)$$

- Interpretation:

Interaction Variables

Interactions

- Based on just this picture, if you have a HS degree and become a manager, how much does your salary go up, on average?
- Based on just this picture, if you have a BS degree and become a manager, how much does your salary go up, on average?
- Key observation: How much your salary increases when you become a manager depends on how much education you have.



Interactions

- Interaction: Occurs when the effect of one covariate on the response depends on the value of another covariate.

- Interactions enter the regression model *multiplicatively*.

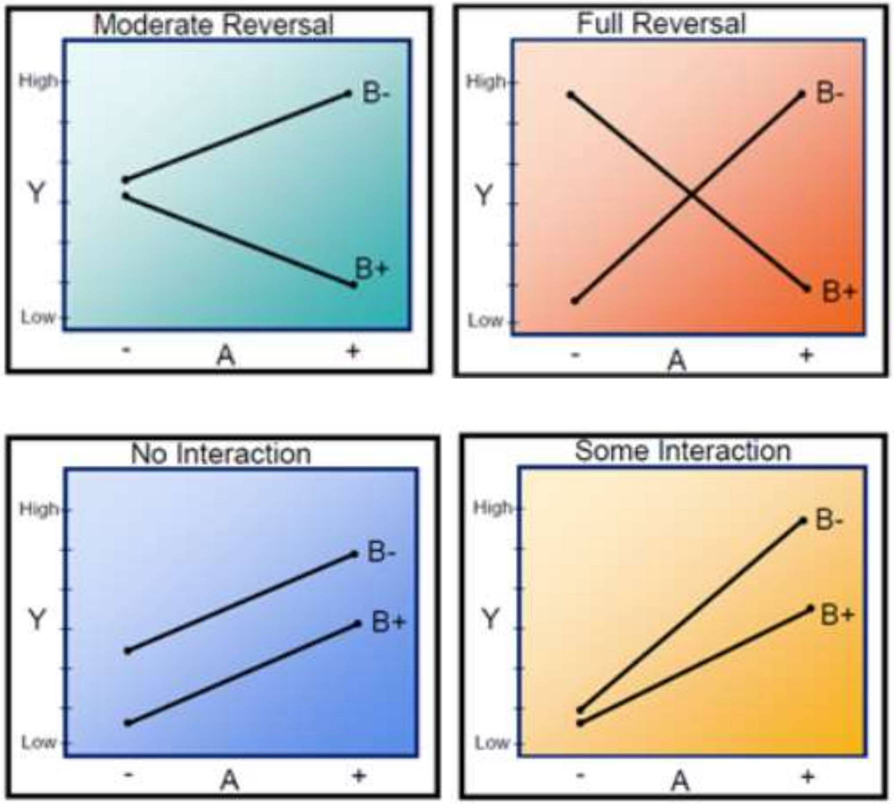
$$y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{Main Effects}} + \underbrace{\beta_3 x_{i1} x_{i2}}_{\text{Interaction Term}} + \epsilon_i \, , \qquad \epsilon_i \; {}^{\text{iid}} \; N(0, \sigma^2)$$

Interactions: Notes

- Two-way interactions (X_1X_2) are good when interpretable
- Higher-order interactions $(X_1X_2X_3)$ become difficult to interpret – you should use these only when clearly interpretable
- If a higher-order interaction $(X_1X_2X_3)$ is used, then the model must also include *all* lower-order interaction terms $(X_1X_2, X_1X_3, X_2X_3, X_1, X_2, X_3)$ - even if lower-order terms are not significant
  - "Good form"
  - Maintain correct interpretation
  - Otherwise, those coefficients are forced to be zero (want a flexible "response surface")
- Meaningful and interpretable interactions are best

Interaction Plots

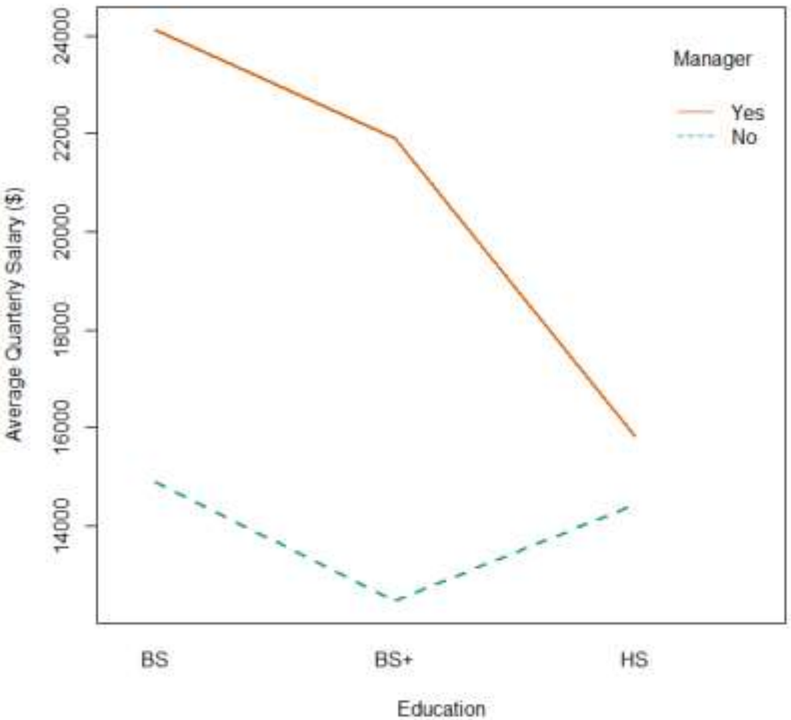
- Interaction plots can be used to help determine if there is an interaction (the effect of X_1 on y depends on X_2)
- Parallel lines indicate no interaction
- Nonparallel lines indicate an interaction. The more nonparallel the lines, the stronger the interaction

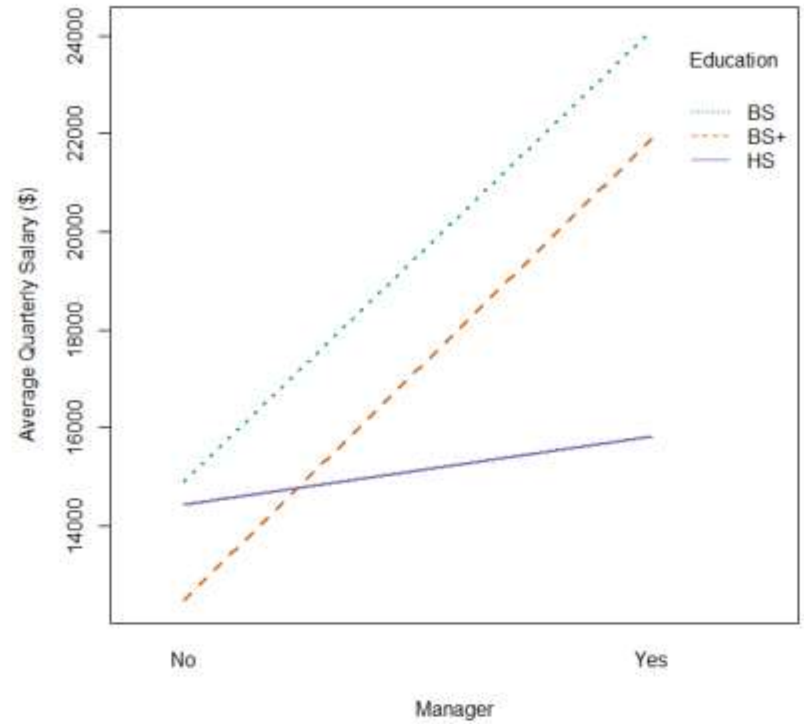


Only works well with categorical data

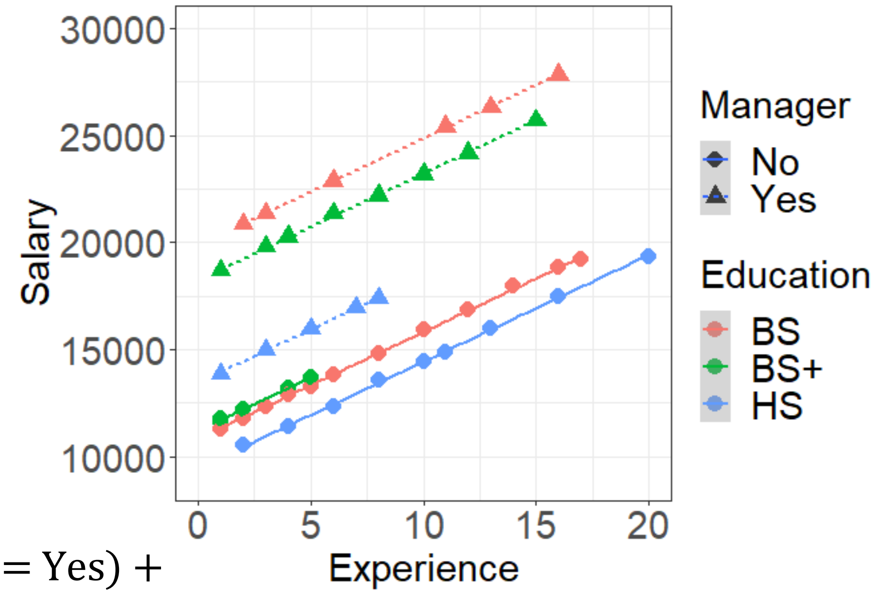
Figures taken from: https://www.datasciencecentral.com/profiles/blogs/the-significance-of-interaction-plots-in-statistics

Interaction Plots: Salary Data





Salary Model with Interactions



General/Theoretical Model:

$$\begin{aligned}\text{Salary}_i = \beta_0 + \beta_1 \times \text{Experience}_i + &\\ &\beta_2 \times \text{I}(\text{Education}_i = \text{BS}) + \\ &\beta_3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ &\beta_4 \times \text{I}(\text{Manager}_i = \text{Yes}) + \\ &\beta_5 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + \\ &\beta_6 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes}) + \epsilon_i,\end{aligned}$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Fitted model:

$$\begin{aligned}\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + &\\ &1384.3 \times \text{I}(\text{Education}_i = \text{BS}) + \\ &1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ &3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) + \\ &5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + \\ &3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})\end{aligned}$$

Salary Model with Interactions

$$\begin{aligned}\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + \\ 1384.3 \times \text{I}(\text{Education}_i = \text{BS}) + \\ 1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ 3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) + \\ 5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + \\ 3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})\end{aligned}$$

- How would you test if the interaction between education and manager position is “significant”?
  - Perform an F-test (see earlier notes in this module)
  - $H_0: \beta_5 = \beta_6 = 0$ vs. H_a: at least one coefficient is non-zero
  - In this case, $F = 4776.7 \rightarrow p-\text{value} \approx 0$
    - Conclude:

Code!

Interaction Interpretations

Interpreting Interactions

- Let X_1 and X_2 be continuous:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \,, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Holding X_2 constant, as X_1 increases by 1, how much does Y change, on average?

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$
$$y_i^* = \beta_0 + \beta_1 (x_{i1} + 1) + \beta_2 x_{i2} + \beta_3 (x_{i1} + 1) x_{i2} + \epsilon_i$$

$$y_i^* - y_i = \beta_0 + \beta_1 (x_{i1} + 1) + \beta_2 x_{i2} + \beta_3 (x_{i1} + 1) x_{i2} + \epsilon_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i)$$

$$\Delta y_i = \beta_1 + \beta_3 x_{i2} \qquad \neq \qquad \beta_1$$

- So, the effect of X_1 on Y depends on X_2.

Continuous-Continuous Interactions

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \,, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Interpretation of the interaction effect:
  - Holding X_2 constant, if X_1 increases by 1 unit, then we expect an average change of $\beta_1 + \beta_3 x_{i2}$ in Y.
  - Similarly, holding X_1 constant, if X_2 increases by 1 unit, then we expect an average change of $\beta_2 + \beta_3 x_{i1}$ in Y
- Interpretation of the main effects*:
  - If X_1 increases by 1 unit and $\boldsymbol{x_{i2} = 0}$, then we expect an average change of β_1 in Y.
  - If X_2 increases by 1 unit and $\boldsymbol{x_{i1} = 0}$, then we expect an average change of β_2 in Y.

* If you have interaction terms in the model, focus only on interpreting that interaction effect (not the main effects)

Continuous-Continuous Interactions

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i , \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Interpretation of the interaction effect:
  - Holding X_2 constant, if X_1 increases by 1 unit, then we expect an average change of $\beta_1 + \beta_3 x_{i2}$ in Y.
  - Similarly, holding X_1 constant, if X_2 increases by 1 unit, then we expect an average change of $\beta_2 + \beta_3 x_{i1}$ in Y
- Interpretation of the main effects*:

  Not necessarily meaningful by itself – may not even be possible for x_{i2} to equal 0

  - If X_1 increases by 1 unit and $\boldsymbol{x_{i2} = 0}$, then we expect an average change of β_1 in Y.
  - If X_2 increases by 1 unit and $\boldsymbol{x_{i1} = 0}$, then we expect an average change of β_2 in Y.

* If you have interaction terms in the model, focus only on interpreting that interaction effect (not the main effects)

Interactions

- Types of interactions:
  - Continuous-Continuous
  - Continuous-Categorical
  - Categorical-Categorical

Continuous-Continuous Interactions

- $\widehat{\text{Salary}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Experience}_i + \hat{\beta}_2 \times \text{Age}_i + \hat{\beta}_3 \times \text{Experience}_i \times \text{Age}_i$

- The effect of **Experience** on average quarterly Salary depends on Age:

$$\hat{\beta}_1 + \hat{\beta}_3 \times \text{Age}_i$$

- The effect of **Age** on average quarterly Salary depends on Experience:

Continuous-Continuous Interactions

- $\widehat{\text{Salary}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Experience}_i + \hat{\beta}_2 \times \text{Age}_i + \hat{\beta}_3 \times \text{Experience}_i \times \text{Age}_i$

- The effect of **Experience** on average quarterly Salary depends on Age:

$$\hat{\beta}_1 + \hat{\beta}_3 \times \text{Age}_i$$

- The effect of **Age** on average quarterly Salary depends on Experience:

$$\hat{\beta}_2 + \hat{\beta}_3 \times \text{Experience}_i$$

Continuous-Categorical Interactions

$$\widehat{\text{Salary}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Experience}_i + \hat{\beta}_2 \times \text{I}(\text{Manager}_{\,i} = \text{Yes}) + \hat{\beta}_3 \times \text{Experience}_i \times \text{I}(\text{Manager}_i = \text{Yes})$$

- The effect of **Experience** on average quarterly Salary depends on Manager status:

  for managers		for non-managers

- The effect of being a **Manager** on average quarterly Salary depends on Experience:

Continuous-Categorical Interactions

$$\widehat{\text{Salary}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Experience}_i + \hat{\beta}_2 \times \text{I}(\text{Manager}_i = \text{Yes}) + \hat{\beta}_3 \times \text{Experience}_i \times \text{I}(\text{Manager}_i = \text{Yes})$$

- The effect of **Experience** on average quarterly Salary depends on Manager status:

  $\hat{\beta}_1 + \hat{\beta}_3$ for managers

  $\hat{\beta}_1$ for non-managers

- The effect of being a **Manager** on average quarterly Salary depends on Experience:

  $\hat{\beta}_2 + \hat{\beta}_3 \times \text{Experience}_i$

Categorical-Categorical Interactions

$$\widehat{\text{Salary}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{I}(\text{Education}_i = \text{BS}) + \hat{\beta}_2 \times \text{I}(\text{Education}_i = \text{BS+}) + \hat{\beta}_3 \times \text{I}(\text{Manager}_{\,i} = \text{Yes}) +$$
$$\hat{\beta}_4 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) +$$
$$\hat{\beta}_5 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})$$

- The effect of **Education** on average quarterly Salary depends on Manager status:

for managers with a BS education

for non-managers with a BS education

for managers with a BS+ education

for non-managers with a BS+ education

- The effect of being a **Manager** on average quarterly Salary depends on Education:

for a BS education

for a BS+ education

Categorical-Categorical Interactions

$$\widehat{\text{Salary}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{I}(\text{Education}_i = \text{BS}) + \hat{\beta}_2 \times \text{I}(\text{Education}_i = \text{BS+}) + \hat{\beta}_3 \times \text{I}(\text{Manager}_{\ i} = \text{Yes}) +$$
$$\hat{\beta}_4 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) +$$
$$\hat{\beta}_5 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})$$

- The effect of **Education** on average quarterly Salary depends on Manager status:

  $\hat{\beta}_1 + \hat{\beta}_4$ for managers with a BS education

  $\hat{\beta}_1$ for non-managers with a BS education

  $\hat{\beta}_2 + \hat{\beta}_5$ for managers with a BS+ education

  $\hat{\beta}_2$ for non-managers with a BS+ education

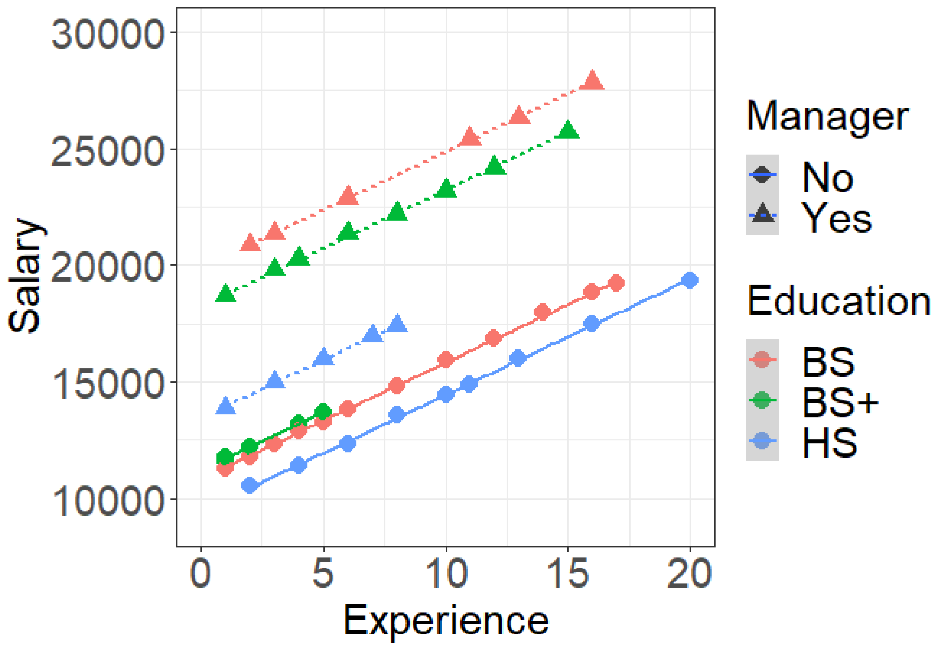
- The effect of being a **Manager** on average quarterly Salary depends on Education:

  $\hat{\beta}_3 + \hat{\beta}_4$ for a BS education

  $\hat{\beta}_3 + \hat{\beta}_5$ for a BS+ education

Salary Data Model with Interactions

Final fitted model: $\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i +$
$1384.3 \times \text{I}(\text{Education}_i = \text{BS}) +$
$1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) +$
$3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) +$
$5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) +$
$3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})$



Salary Data Model with Interactions

$$
\begin{aligned}
\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + \\
& 1384.3 \times \text{I}(\text{Education}_i = \text{BS}) + 1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\
& 3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) + \\
& 5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + \\
& 3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})
\end{aligned}
$$

- What effect does becoming a manager have on average quarterly salary if you have a HS education?
- What effect does becoming a manager have on average quarterly salary if you have a BS education?
- What effect does becoming a manager have on average quarterly salary if you have a BS+ education?

Salary Data Model with Interactions

$$\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + 1384.3 \times \text{I}(\text{Education}_i = \text{BS}) + 1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) + 3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) + 5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + 3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})$$

- What effect does becoming a manager have on average quarterly salary if you have a HS education? Average quarterly salary would increase by $3,988.8
- What effect does becoming a manager have on average quarterly salary if you have a BS education?
  Average quarterly salary would increase by $3,988.8 + $5,049.3 = $9,038.1
- What effect does becoming a manager have on average quarterly salary if you have a BS+ education?
  Average quarterly salary would increase by $3,988.8 + $3,051.8 = $7,040.6

Salary Data Model with Interactions

$$
\begin{aligned}
\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + \\
& 1384.3 \times \mathrm{I}(\text{Education}_i = \text{BS}) + 1741.3 \times \mathrm{I}(\text{Education}_i = \text{BS+}) + \\
& 3988.8 \times \mathrm{I}(\text{Manager}_i = \text{Yes}) + \\
& 5049.3 \times \mathrm{I}(\text{Education}_i = \text{BS})\mathrm{I}(\text{Manager}_i = \text{Yes}) + \\
& 3051.8 \times \mathrm{I}(\text{Education}_i = \text{BS+})\mathrm{I}(\text{Manager}_i = \text{Yes})
\end{aligned}
$$

- For managers, what is the effect on average quarterly salary of having a BS+ education vs. a BS education?

Salary Data Model with Interactions

$$\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + \\ 1384.3 \times \text{I}(\text{Education}_i = \text{BS}) + 1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ 3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) + \\ 5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + \\ 3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})$$

- For managers, what is the effect on average quarterly salary of having a BS+ education vs. a BS education?

$$(1741.3 + 3051.8) - \\ (1384.3 + 5049.3) = -1640.5$$

Average quarterly salary would decrease by $1,640.5

Salary Data Model with Interactions

$$\begin{aligned}\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + \\ 1384.3 \times \text{I}(\text{Education}_i = \text{BS}) + 1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ 3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) + \\ 5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + \\ 3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})\end{aligned}$$

- What is the average salary for a non-manager with a BS education and 24 years experience?

- What is the average salary for a manager with a BS education and 24 years experience?

Salary Data Model with Interactions

$$\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times \text{Experience}_i + \\ 1384.3 \times \text{I}(\text{Education}_i = \text{BS}) + 1741.3 \times \text{I}(\text{Education}_i = \text{BS+}) + \\ 3988.8 \times \text{I}(\text{Manager}_i = \text{Yes}) + \\ 5049.3 \times \text{I}(\text{Education}_i = \text{BS})\text{I}(\text{Manager}_i = \text{Yes}) + \\ 3051.8 \times \text{I}(\text{Education}_i = \text{BS+})\text{I}(\text{Manager}_i = \text{Yes})$$

- What is the average salary for a non-manager with a BS education and 24 years experience?

$$\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times 24 + 1384.3 \times 1 = \$22{,}804.30$$

- What is the average salary for a manager with a BS education and 24 years experience?

$$\widehat{\text{Salary}}_i = 9458.4 + 498.4 \times 24 + 1384.3 \times 1 + 3988.8 \times 1 + 5049.3 \times 1 = \$31{,}842.40$$

Interactions: Notes

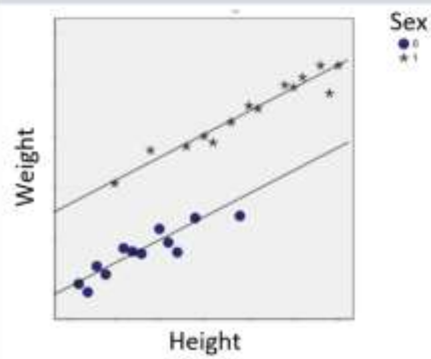
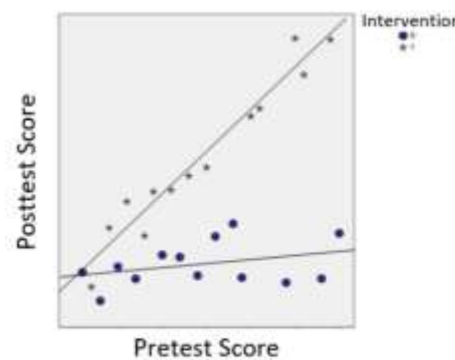
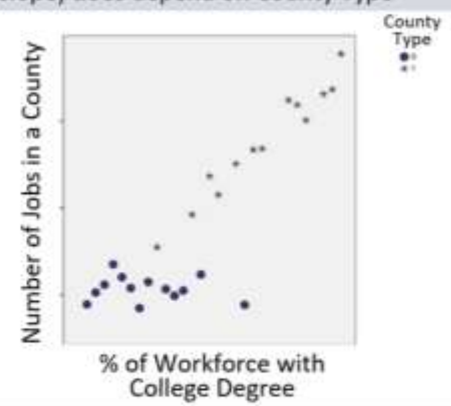
- Interactions vs. Multicollinearity (*do not confuse these*)

- There is an <u>interaction</u> between predictors X_1 and X_2 if:
  - the *effect* of X_1 on Y depends on X_2

- There is <u>multicollinearity</u> involving predictors X_1 and X_2 if:
  - X_1 is linearly related to X_2 (no mention of Y)

- Since X_1 and X_1X_2 are likely to be collinear, standardizing may help reduce the collinearity, but we only need to do this if we need X_1 as a stand-alone independent variable (it will not affect the test on the interaction term)

Interactions vs Multicollinearity

	Multicollinearity Only	Interaction Only	Multicollinearity and Interaction
	Y = Weight X1 = Height X2 = Sex (0,1)	Y = Posttest Score X1 = Pretest Score X2 = Intervention (none, teaching)	Y = Number of Jobs in a County X1 = % of Workforce with College Degree X2 = County Type (Rural, Urban)
Correlation Info:	Height **is** correlated with Sex	Pretest Score is **not** correlated with subsequent Intervention	% of Workforce with College Degree **is** correlated with County Type
Interaction Info:	The *effect* of Height on Weight (slope) does **not** depend on Sex	The *effect* of Pretest Score on Posttest Score (slope) **does** depend on Intervention	The *effect* of % of Workforce with College Degree on Number of Jobs in a County (slope) **does** depend on County Type
			

Examples taken from: https://www.theanalysisfactor.com/interaction-association/

Higher-Order Variables

Polynomials

- If the effect of X_1 on Y appears quadratic (or higher-order…), add predictor $X_1 X_1 = X_1^2$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \epsilon_i \,, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Same approach applies as for interaction terms
  - Include lower-order terms
  - Can standardize to reduce multicollinearity (not critical)
  - Coefficient interpretation is important: if X_1 increases by 1 unit, and X_2 is held constant, then we expect an average change in Y of $\beta_1 + \beta_3(2x_{i1} + 1)$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \epsilon_i$$
$$y_i^* = \beta_0 + \beta_1 (x_{i1} + 1) + \beta_2 x_{i2} + \beta_3 (x_{i1} + 1)^2 + \epsilon_i$$

$$y_i^* - y_i = \beta_1 + \beta_3(2x_{i1} + 1)$$