MSBA 7002 Lecture 1
Linear Regression

Innovation and Information Management
HKU Business School

[2] Some of the slides, figures, codes are from OpenIntro, Prof. Haipeng Shen, Prof. Mine Cetinkaya-Rundel, Prof. Wei Zhang, Prof. Dan Yang, Prof. Weichen Wang.

Outline

Outline

Multiple Linear Regression

- Simple linear regression: Bivariate - *two variables*: y and x

Multiple Linear Regression

- Simple linear regression: Bivariate - *two variables*: y and x
- Multiple linear regression: *Multiple variables*: y and $x_1, x_2, \cdots$

Multiple Linear Regression Model

- In the multiple regression model, we assume the data follows

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} + \epsilon_i$$

where ϵ_i iid $\sim N(0, \sigma^2)$

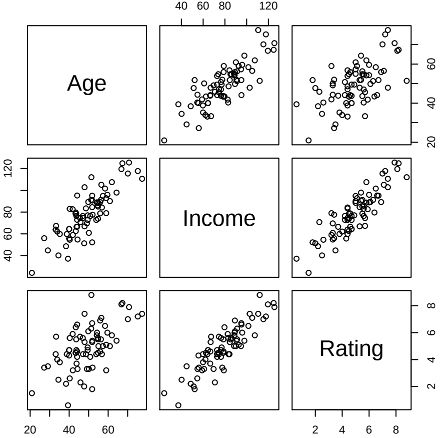
Outline

1. Multiple Linear Regression
   - Model
   - Collinearity
   - Categorical Explanatory Variables

Example: Market Segmentation

- A *marketing* project identified a list of affluent customers for a new phone.
- Should the company target promotion towards the *younger or older* members of this list?
- To answer this question, the marketing firm obtained a sample of 75 consumers and asked them to rate their *"likelihood of purchase"* on a scale of 1 to 10.
- *Age and Income* of consumers were also recorded.

Correlation Among Variables



Correlation

	Age	Income	Rating
Age	1.000	0.828	0.586
Income	0.828	1.000	0.884
Rating	0.586	0.884	1.000

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e - 16$

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e - 16$

- *MRM of Rating, on both variables*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.512374	0.355004	1.443	0.153
Age	-0.071448	0.012576	-5.682	2.65e-07
Income	*0.100591*	0.006491	15.498	$< 2e - 16$

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e-16$

- *MRM of Rating, on both variables*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.512374	0.355004	1.443	0.153
Age	-0.071448	0.012576	-5.682	2.65e-07
Income	*0.100591*	0.006491	15.498	$< 2e-16$

- We need to understand why the slope of *Age* is positive in the simple regression but negative in the multiple regression.
- Given the context, the positive marginal slope is probably more surprising than the negative partial slope.

Collinearity: Highly Correlated X Variables

- MRM allows the use of correlated explanatory variables.
- *Collinearity* occurs when the correlations among the X variables are large.

Collinearity: Highly Correlated X Variables

- MRM allows the use of correlated explanatory variables.
- *Collinearity* occurs when the correlations among the X variables are large.
- As the correlation among these variables grows, it becomes difficult for regression to separate the partial effects of different variables.
  - Highly correlated X variables tend to change together, making it *difficult to estimate* the partial slope.
  - *Difficulties interpreting* the model

Customer Segmentation

- The figure shows regression lines fit within three subsets:

low incomes ($<$ \$45K)

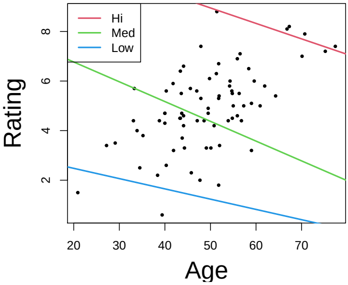
	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	3.30845	3.42190	0.967	0.436
Age	-0.04144	0.10786	-0.384	0.738

moderate incomes (\$70K $\sim$ \$80K)

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	8.36412	2.34772	3.563	0.0026
Age	-0.07978	0.04791	-1.665	0.1153

high incomes ($>$ \$110K)

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	12.07081	1.28999	9.357	0.000235
Age	-0.06243	0.01873	-3.332	0.020727



- The simple regression slopes are negative in each case, as in the *multiple linear regression*.

Customer Segmentation

- The figure shows regression lines fit within three subsets:

low incomes ($< \$45K$)

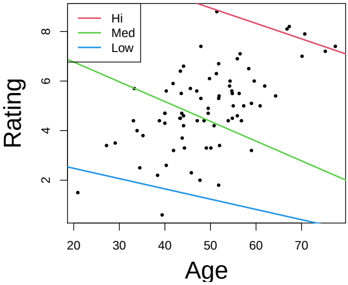
	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	3.30845	3.42190	0.967	0.436
Age	-0.04144	0.10786	-0.384	0.738

moderate incomes ($\$70K \sim \$80K$)

	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	8.36412	2.34772	3.563	0.0026
Age	-0.07978	0.04791	-1.665	0.1153

high incomes ($> \$110K$)

	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	12.07081	1.28999	9.357	0.000235
Age	-0.06243	0.01873	-3.332	0.020727

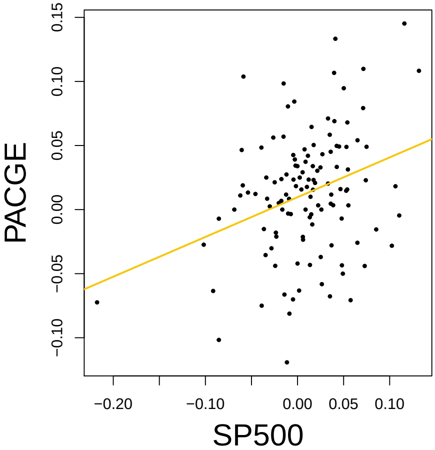


- The simple regression slopes are negative in each case, as in the *multiple linear regression*.
- Based on these results, how should the marketing firm direct their promotional efforts?

The Market Model

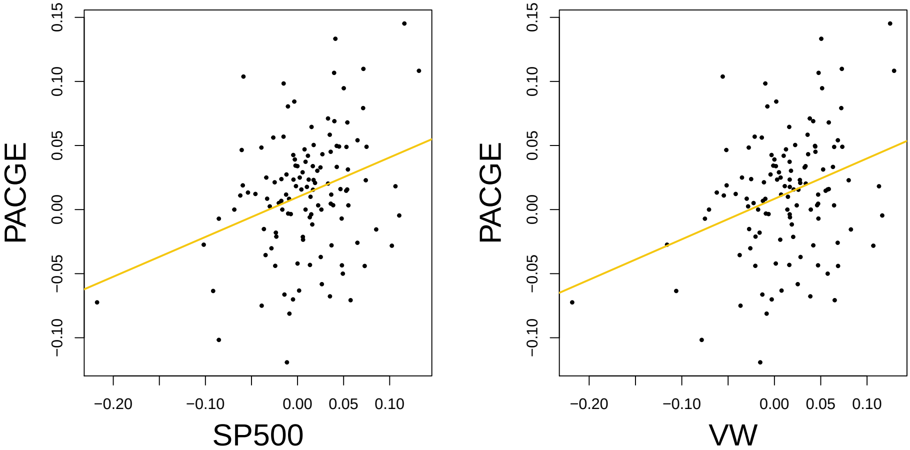
- We consider simple linear regression of
  - exPACGE on exSP500, the excess returns of PACGE and SP500 over TBill30
  - exPACGE on exVW, the excess returns of PACGE and VW over TBill30
- Also, consider multiple linear regression of exPACGE on both exSP500 and exVW

SRM of exPACGE on either the exSP500 or exVW



	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

SRM of exPACGE on either the exSP500 or exVW



	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

- Very similar results.

Regress exPACGE on both exSP500 and exVW

	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

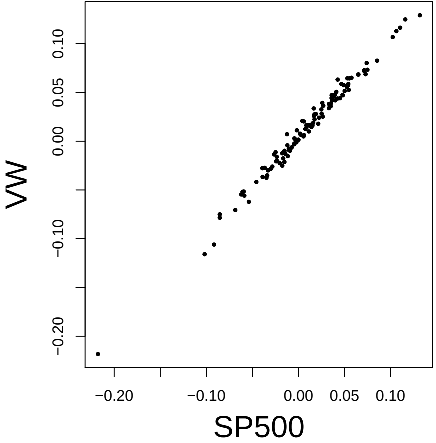
	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	0.005448	0.005119	1.064	0.289
SP500	-0.821098	0.749946	-1.095	0.276
VW	1.111498	0.731784	1.519	0.132
ANOVA	F-statistic	7.513	p-value	0.0008547

Regress exPACGE on both exSP500 and exVW

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.005448	0.005119	1.064	0.289
SP500	-0.821098	0.749946	-1.095	0.276
VW	1.111498	0.731784	1.519	0.132
ANOVA	F-statistic	7.513	p-value	0.0008547

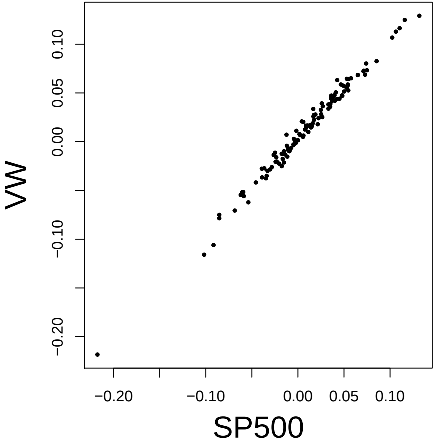


Regress exPACGE on both exSP500 and exVW

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.005448	0.005119	1.064	0.289
SP500	-0.821098	0.749946	-1.095	0.276
VW	1.111498	0.731784	1.519	0.132
ANOVA	F-statistic	7.513	p-value	0.0008547



- Huge Collinearity!!!

The F Test and Correlated Predictors

- *Seemingly contradiction* between
  - Overall F Ratio in the ANOVA Table
  - Individual p-value (T test) for each regression coefficient

The F Test and Correlated Predictors

- *Seemingly contradiction* between
  - Overall F Ratio in the ANOVA Table
  - Individual p-value (T test) for each regression coefficient
- The overall F Ratio comes in handy when the explanatory variables in a regression are correlated.
  - *Overall F Ratio*: whether at least one of the X variables is significant;
  - *Individual T test*: whether each individual X variable is significant, having included the other ones.
- When the predictors are highly correlated (i.e. *high collinearity*), they may contradict each other.

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  where R_k^2 is *R^2 from regressing x_k on the other x's.*
- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated,

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  where R_k^2 is *R^2 from regressing x_k on the other x's.*
- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated, VIF $= 1$.
- If the x's are correlated,

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

where R_k^2 is *R^2 from regressing x_k on the other x's.*

- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated, VIF $= 1$.
- If the x's are correlated, VIF can be much larger than 1.

VIF Results

- For market example

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.005448	0.005119	1.064	0.289	
SP500	-0.821098	0.749946	-1.095	*0.276*	74.29672
VW	1.111498	0.731784	1.519	*0.132*	74.29672

- For Customer Segmentation

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.512374	0.355004	1.443	0.153	
Age	-0.071448	0.012576	-5.682	*2.65e-07*	3.188591
Income	0.100591	0.006491	15.498	$< 2e - 16$	3.188591

VIF Results

- For market example

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.005448	0.005119	1.064	0.289	
SP500	-0.821098	0.749946	-1.095	*0.276*	74.29672
VW	1.111498	0.731784	1.519	*0.132*	74.29672

- For Customer Segmentation

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.512374	0.355004	1.443	0.153	
Age	-0.071448	0.012576	-5.682	*2.65e-07*	3.188591
Income	0.100591	0.006491	15.498	$< 2e - 16$	3.188591

- The VIF answers a very handy question when an explanatory variable is not statistically significant:
  - Is this explanatory variable simply not useful, or is it just redundant?

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's
- The *F Ratio* detects statistical significance that can be disguised by collinearity.
  - The F ratio allows you to look at the importance of several factors simultaneously.
  - When predictors are collinear, the F test reveals their net effect, rather than trying to separate their effects as a t ratio does.

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's
- The *F Ratio* detects statistical significance that can be disguised by collinearity.
  - The F ratio allows you to look at the importance of several factors simultaneously.
  - When predictors are collinear, the F test reveals their net effect, rather than trying to separate their effects as a t ratio does.
- *VIF measures* the impact of collinearity on the coefficients of specific explanatory variables.

Summary: Collinearity

- Collinearity does *not violate* any assumption of the MRM, but it does make regression harder to interpret.
  - In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.

Summary: Collinearity

- Collinearity does *not violate* any assumption of the MRM, but it does make regression harder to interpret.
  - In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.
- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

R^2 vs. adjusted R^2

- When <u>any</u> variable is added to the model, R^2 *increases*.

R^2 vs. adjusted R^2

- When any variable is added to the model, R^2 *increases*.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adj R^2 does not increase.

R^2 vs. adjusted R^2

- When any variable is added to the model, R^2 *increases*.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adj R^2 does not increase.
- R^2

$$R^2 = 1 - \frac{SSE}{TSS}$$

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{SSE/(n-K-1)}{TSS/(n-1)}$$

where n is the number of cases and K is the number of predictors

R^2 vs. adjusted R^2

- When any variable is added to the model, *R^2 increases*.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adj R^2 does not increase.
- R^2

$$R^2 = 1 - \frac{SSE}{TSS}$$

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{SSE/(n-K-1)}{TSS/(n-1)}$$

  where n is the number of cases and K is the number of predictors
- Because K is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors
- Therefore, we can choose models with higher R^2_{adj} over others.

R^2 vs. adjusted R^2

```
> summary(lm(PACGE~SP500+VW,data = stock))

Call:
lm(formula = PACGE ~ SP500 + VW, data = stock)

Residuals:
      Min        1Q    Median        3Q       Max
-0.117084 -0.025683  0.001373  0.029422  0.112175

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.005448   0.005119   1.064    0.289
SP500       -0.821098   0.749946  -1.095    0.276
VW           1.111498   0.731784   1.519    0.132

Residual standard error: 0.04591 on 116 degrees of freedom
Multiple R-squared:  0.1147,    Adjusted R-squared:  0.09942
F-statistic: 7.513 on 2 and 116 DF,  p-value: 0.0008547

>
> x3 <- rnorm(length(SP500))
> summary(lm(PACGE~SP500+VW+x3,data = stock))

Call:
lm(formula = PACGE ~ SP500 + VW + x3, data = stock)

Residuals:
      Min        1Q    Median        3Q       Max
-0.117041 -0.023896  0.004667  0.030164  0.108113

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006487   0.005151   1.259    0.210
SP500       -0.711646   0.750875  -0.948    0.345
VW           0.988744   0.733957   1.347    0.181
x3          -0.005476   0.003898  -1.405    0.163

Residual standard error: 0.04571 on 115 degrees of freedom
Multiple R-squared:  0.1296,    Adjusted R-squared:  0.1069
F-statistic: 5.708 on 3 and 115 DF,  p-value: 0.001117
```

Outline

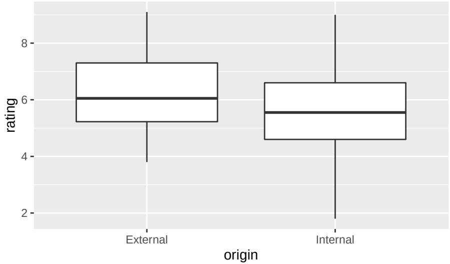
1. Multiple Linear Regression
   - Model
   - Collinearity
   - Categorical Explanatory Variables

Example: Employee Performance Study

- "Which of two prospective job candidates should we hire for a position that pays $80,000: the internal manager or the externally recruited manager?"
- Data set:
  - 150 managers: 88 internal and 62 external
  - *Manager Rating* is an evaluation score of the employee in their current job, indicating the "value" of the employee to the firm.
  - *Origin* is a categorical variable that identifies the managers as either External or Internal to indicate from where they were hired.
  - *Salary* is the starting salary of the employee when they were hired. It indicates what sort of job the person was initially hired to do. In the context of this example, it does not measure how well they did that job. That's measured by the rating variable.

Two-Sample Comparison: Manager Rating vs Origin

- *Origin*: a categorical variable.




```
        Welch Two Sample t-test

data:  rating by origin
t = 3.0484, df = 140.49, p-value = 0.00275
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 0.2517995 1.1810451
sample estimates:
mean in group External mean in group Internal
              6.320968               5.604545
```

- We can recognize a significant difference between the means via two-sample t-test.

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$
\begin{aligned}
y_{i|x=External} &= \mu_{External} + \epsilon_i \\
y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i
\end{aligned}
$$

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$
\begin{aligned}
y_{i|x=External} &= \mu_{External} + \epsilon_i \\
y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i
\end{aligned}
$$

- *In regression*
  - 'External' as the base
  - x_1 be the indicator function of being 'Internal', $I(Origin = Internal)$
  - $\beta_0 = \mu_{External}$
  - $\beta_1 = \mu_{Internal} - \mu_{External}$

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$\begin{aligned} y_{i|x=External} &= \mu_{External} + \epsilon_i \\ y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i \end{aligned}$$

- *In regression*
  - 'External' as the base
  - x_1 be the indicator function of being 'Internal', $I(Origin = Internal)$
  - $\beta_0 = \mu_{External}$
  - $\beta_1 = \mu_{Internal} - \mu_{External}$
- ANOVA model is the same as

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$
\begin{aligned}
y_{i|x=External} &= \mu_{External} + \epsilon_i \\
y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i
\end{aligned}
$$

- *In regression*
  - 'External' as the base
  - x_1 be the indicator function of being 'Internal', $I(Origin = Internal)$
  - $\beta_0 = \mu_{External}$
  - $\beta_1 = \mu_{Internal} - \mu_{External}$
- ANOVA model is the same as

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

- These two tests are *equivalent*

$$H_0 : \mu_{Internal} = \mu_{External} \text{ and } H_0 : \beta_1 = 0$$

Regress Manager Rating on Origin

```
Call:
lm(formula = rating ~ origin)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8045 -1.0169 -0.1045  0.9790  3.3955

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.3210     0.1839  34.372  < 2e-16 ***
originInternal  -0.7164     0.2401  -2.984  0.00333 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.448 on 148 degrees of freedom
Multiple R-squared:  0.05675,	Adjusted R-squared:  0.05037
F-statistic: 8.904 on 1 and 148 DF,  p-value: 0.00333
```

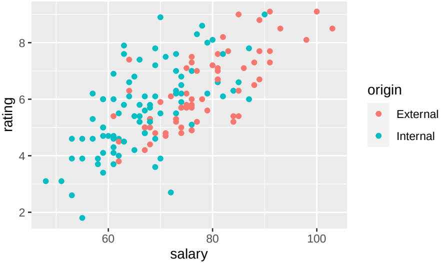
- The difference in the rating (-0.72) between internal and external managers is significant since the p-value $= .003 < .05$.
- In terms of regression, *Origin* explains significant variation in *Manager Rating*.

Regress Manager Rating on Origin

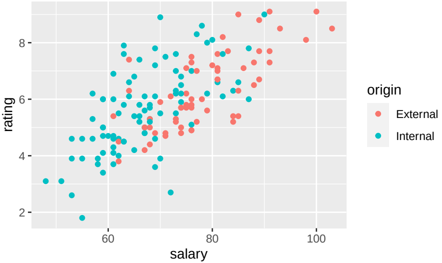
```
Call:
lm(formula = rating ~ origin)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8045 -1.0169 -0.1045  0.9790  3.3955 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.3210     0.1839  34.372  < 2e-16 ***
originInternal  -0.7164     0.2401  -2.984  0.00333 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.448 on 148 degrees of freedom
Multiple R-squared:  0.05675,	Adjusted R-squared:  0.05037 
F-statistic: 8.904 on 1 and 148 DF,  p-value: 0.00333
```

- The difference in the rating (-0.72) between internal and external managers is significant since the p-value $= .003 < .05$.
- In terms of regression, *Origin* explains significant variation in *Manager Rating*.
- Before we claim that the external candidate should be hired, is there a possible confounding variable, another explanation for the difference in rating?
- Let's explore the relationship between *Manager Rating and Salary*.

Scatterplot of Manager Rating vs. Salary

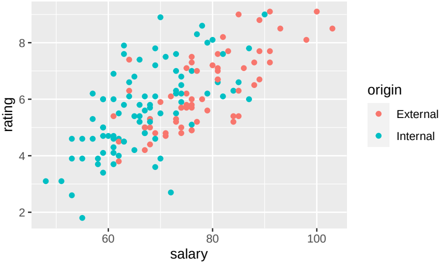


Scatterplot of Manager Rating vs. Salary



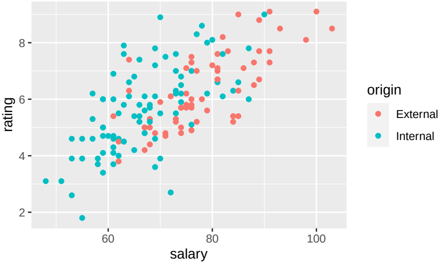
- (a) Salary is correlated with Manager Rating, and (b) that external managers were hired at higher salaries

Scatterplot of Manager Rating vs. Salary



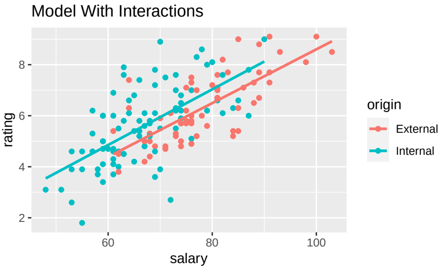
- (a) Salary is correlated with Manager Rating, and (b) that external managers were hired at higher salaries
- This combination indicates *confounding*: not only are we comparing internal vs. external managers; we are comparing internal managers hired into lower salary jobs with external managers placed into higher salary jobs.

Scatterplot of Manager Rating vs. Salary



- (a) Salary is correlated with Manager Rating, and (b) that external managers were hired at higher salaries
- This combination indicates *confounding*: not only are we comparing internal vs. external managers; we are comparing internal managers hired into lower salary jobs with external managers placed into higher salary jobs.
- *Easy fix*: compare only those whose starting salary near $80K. But that leaves too few data points for a reasonable comparison.

Separate Regressions of Manager Rating on Salary



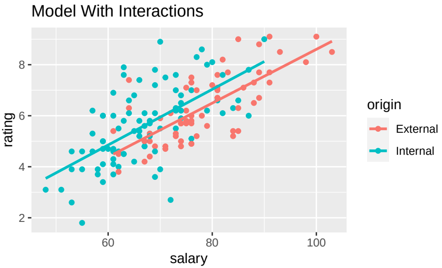
Internal

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.69352	0.94925	-1.784	0.0779
salary	0.10909	0.01407	7.756	1.65e-11

External

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.9369	0.9862	-1.964	0.0542
salary	0.1054	0.0125	8.432	9.01e-12

Separate Regressions of Manager Rating on Salary



Internal

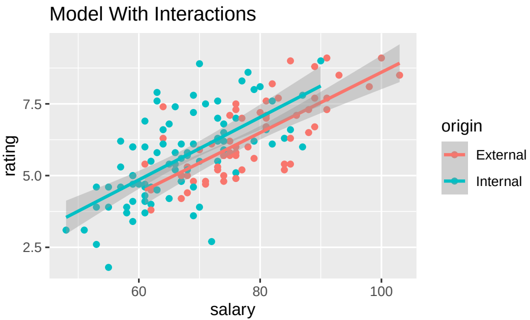
	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.69352	0.94925	-1.784	0.0779
salary	0.10909	0.01407	7.756	1.65e-11

External

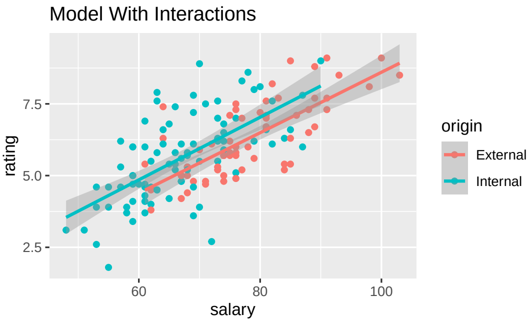
	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.9369	0.9862	-1.964	0.0542
salary	0.1054	0.0125	8.432	9.01e-12

- At any given salary, internal managers get higher average ratings!
- In regression, *confounding* is a form of *collinearity*.
  - *Salary* is related to *Origin* which was the variable used to explain *Rating*.
  - With *Salary* added, the effect of *Origin* changes sign. Now internal managers look better.

Are the Two Fits Significantly Different?



Are the Two Fits Significantly Different?



- The two confidence bands overlap, which make the comparison indecisive.
- A more powerful idea is to combine these two separate simple regressions into one multiple regression that will allow us to compare these fits.

Regress Manager Rating on both Salary and Origin

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.100459	0.768140	-2.734	0.00702
originInternal	0.514966	0.209029	2.464	0.01491
salary	0.107478	0.009649	11.139	$<$ 2e-16

- x_1 dummy variable of being 'Internal', $I(Origin = Internal)$
- Notice that we only require one dummy variable to distinguish internal from external managers.

Regress Manager Rating on both Salary and Origin

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.100459	0.768140	-2.734	0.00702
originInternal	0.514966	0.209029	2.464	0.01491
salary	0.107478	0.009649	11.139	$<$ 2e-16

- x_1 dummy variable of being 'Internal', $I(Origin = Internal)$
- Notice that we only require one dummy variable to distinguish internal from external managers.
- This enables two *parallel* lines for two kinds of managers.
  - Origin = External
    Manager Rating = -2.100459 + 0.107478 Salary
  - Origin = Internal
    Manager Rating = -2.100459 + 0.107478 Salary + 0.514966
- The coefficient of the dummy variable is the difference between the intercepts.

Model with Parallel Lines

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.100459	0.768140	-2.734	0.00702
originInternal	0.514966	0.209029	2.464	0.01491
salary	0.107478	0.009649	11.139	$<$ 2e-16

- The difference between the intercepts is significantly different from 0, since 0.0149, the p-value for Origin[Internal], is less than 0.05.
- Thus, if we assume the slopes are equal, a model using a categorical predictor implies that *controlling* for initial salary, internal managers rate significantly higher.
- How can we check the assumption that the slopes are parallel?

Model with Interaction: Different Slopes

- Beyond just looking at the plot, we can fit a model that allows the slopes to differ.
- This model gives an estimate of the difference between the slopes.
- This estimate is known as an *interaction*.
- An interaction between a dummy variable and a numerical variable measures the difference between the slopes of the numerical variable in the two groups.

	Estimate	Std. Error	t value	$Pr(>|t|)$
(Intercept)	-1.936941	1.156482	-1.675	0.0961
originInternal	0.243417	1.447230	0.168	0.8667
salary	0.105391	0.014657	7.191	3.09e-11
originInternal:salary	0.003702	0.019520	0.190	0.8499

- *Interaction* variable – product of the dummy variable and Salary:

originInternal:salary	=salary	if Origin = Internal
	=0	if Origin = External

- Origin = External
  Manager Rating = -1.94 + 0.11 Salary
- Origin = Internal
  Manager Rating = (-1.94+0.24) + (0.11+0.0037) Salary
  = -1.69 + 0.11 Salary

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.936941	1.156482	-1.675	0.0961
originInternal	0.243417	1.447230	0.168	0.8667
salary	0.105391	0.014657	7.191	3.09e-11
originInternal:salary	0.003702	0.019520	0.190	0.8499

- *Interaction* variable – product of the dummy variable and Salary:

originInternal:salary	=salary	if Origin = Internal
	=0	if Origin = External

- Origin = External
  Manager Rating = -1.94 + 0.11 Salary
- Origin = Internal
  Manager Rating = (-1.94+0.24) + (0.11+0.0037) Salary
  = -1.69 + 0.11 Salary
- These equations *match* the simple regressions fit to the two groups separately.
  The interaction is *not significant* because its p-value is large.

Principle of Marginality

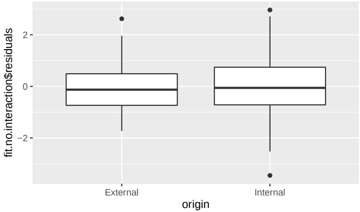
- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.

Principle of Marginality

- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.
- *Origin* became insignificant when *Salary*$*$*Origin* was added, which is due to collinearity.

Principle of Marginality

- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.
- *Origin* became insignificant when *Salary*∗*Origin* was added, which is due to collinearity.
- The assumption of equal error variance should also be checked by comparing boxplots of the residuals grouped by the levels of the categorical variable.



Summary

- Categorical variables model the differences between groups using regression, while taking account of other variables.
- In a model with a categorical variable, the *coefficients of the categorical terms* indicate *differences between parallel lines*.
- In a model that includes interactions, the *coefficients of the interaction* measure the *differences in the slopes* between the groups.
- Significant categorical variable $\Rightarrow$ different intercepts
- Significant interaction $\Rightarrow$ different slopes