

Other Considerations in the Regression Model

Section 3.3

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Outline

1 Qualitative Predictors

2 Interaction Model

Categorical (Qualitative) Predictors

- Using the *mtcars* data set.
- Suppose that we wish to investigate the difference in the *mpg* based on the transmission *am*
- Here the transmission has only two categories, **automatic** and **manual**.

$$am = \begin{cases} 0 & \text{if automatic} \\ 1 & \text{if manual} \end{cases}$$

$$mpg = \beta_0 + \beta_1 am + \epsilon$$

$$mpg = \begin{cases} \beta_0 + \epsilon & \text{if automatic} \\ \beta_0 + \beta_1 + \epsilon & \text{if manual} \end{cases}$$

Dummy Variables

- Notice in \mathbb{R} we have zeroes and ones for the values of `am`.
- These zeros and ones gives us a **dummy variable**, or an indicator variable.

$$x_1 = \begin{cases} 1 & \text{if } i\text{th car has a manual transmission} \\ 0 & \text{if } i\text{th car has an automatic transmission} \end{cases}$$

Dummy Variables

- Notice in \mathbb{R} we have zeroes and ones for the values of `am`.
- These zeros and ones gives us a **dummy variable**, or an indicator variable.

$$x_1 = \begin{cases} 1 & \text{if } i\text{th car has a manual transmission} \\ 0 & \text{if } i\text{th car has an automatic transmission} \end{cases}$$

- This results in the following model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th car has a manual transmission} \\ \beta_0 + \epsilon_i & \text{if } i\text{th car has an automatic transmission} \end{cases}$$

Dummy Variables

- Notice in \mathbb{R} we have zeroes and ones for the values of `am`.
- These zeros and ones gives us a **dummy variable**, or an indicator variable.

$$x_1 = \begin{cases} 1 & \text{if } i\text{th car has a manual transmission} \\ 0 & \text{if } i\text{th car has an automatic transmission} \end{cases}$$

- This results in the following model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th car has a manual transmission} \\ \beta_0 + \epsilon_i & \text{if } i\text{th car has an automatic transmission} \end{cases}$$

- β_0 can be interpreted as the average mpg among automobiles with automatic transmission.
- $\beta_0 + \beta_1$ is the average mpg among automobiles with manual transmission.
- β_1 is the average difference in mpg between automobiles with automatic and manual transmission.

Results

```
> #Use mtcars
> summary(lm(mpg~am,data = mtcars))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	17.147	1.125	15.247	1.13e-15 ***
am	7.245	1.764	4.106	0.000285 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom

Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385

① F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

$$\begin{aligned}\hat{mpg} &= 17.147 + 7.245 \cdot am \\ &= \begin{cases} 17.147 & \text{if automatic transmission} \\ 24.392 & \text{if manual transmission} \end{cases}\end{aligned}$$

Results

```
> #Use mtcars
> summary(lm(mpg~am,data = mtcars))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	17.147	1.125	15.247	1.13e-15	***
am	7.245	1.764	4.106	0.000285	***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom

Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385

F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

- The average `mpg` for an automobile with automatic transmission is 17.147.

Results

```
> #Use mtcars  
> summary(lm(mpg~am,data = mtcars))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	17.147	1.125	15.247	1.13e-15	***
am	7.245	1.764	4.106	0.000285	***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom

Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385

F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

- The average mpg for an automobile with automatic transmission is 17.147.
- Having a manual transmission added 7.245 to the mpg on average.

Results

```
> #Use mtcars
> summary(lm(mpg~am,data = mtcars))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	17.147	1.125	15.247	1.13e-15	***
am	7.245	1.764	4.106	0.000285	***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom

Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385

F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

- The average `mpg` for an automobile with automatic transmission is 17.147.
- Having a manual transmission added 7.245 to the `mpg` on average.
- The average `mpg` for an automobile with manual transmission is $17.147 + 7.245 = 24.392$.

Results

```
> #Use mtcars  
> summary(lm(mpg~am,data = mtcars))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	17.147	1.125	15.247	1.13e-15	***
am	7.245	1.764	4.106	0.000285	***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom

Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385

F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

- The average `mpg` for an automobile with automatic transmission is 17.147.
- Having a manual transmission added 7.245 to the `mpg` on average.
- The average `mpg` for an automobile with manual transmission is $17.147 + 7.245 = 24.392$.
- Note could have done 1 and -1 as a code instead.

More Than Two Levels

Suppose we want use the number of cylinders as predictors. We have already indicated that this is categorical.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } i\text{th car has 6 cylinders} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if the } i\text{th car has 8 cylinders} \\ \beta_0 + \epsilon_i & \text{if the } i\text{th car has 4 cylinders} \end{cases}$$

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th car has 6 cylinders} \\ 0 & \text{if } i\text{th car does not have 6 cylinders} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th car has 8 cylinders} \\ 0 & \text{if } i\text{th car does not have 8 cylinders} \end{cases}$$

Summary

```
> cyl.fact = as.factor(mtcars$cyl)
> summary(lm(mtcars$mpg~cyl.fact))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	26.6636	0.9718	27.437	< 2e-16	***
cyl.fact6	-6.9208	1.5583	-4.441	0.000119	***
cyl.fact8	-11.5636	1.2986	-8.905	8.57e-10	***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom
Multiple R-squared: 0.7325, Adjusted R-squared: 0.714
F-statistic: 39.7 on 2 and 29 DF, p-value: 4.979e-09

$$\hat{mpg} = \begin{cases} 26.4436 & \text{if } cyl = 4 \\ 26.4436 - 6.9208 & \text{if } cyl = 6 \\ 26.4436 - 11.5636 & \text{if } cyl = 8 \end{cases}$$

$$\hat{mpg} = \begin{cases} 26.4436 & \text{if } cyl = 4 \\ 19.7428 & \text{if } cyl = 6 \\ 15.1 & \text{if } cyl = 8 \end{cases}$$

Example from Textbook

```
library(ISLR)
summary(lm(Balance~Ethnicity,data = Credit))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	531.00	46.32	11.464	<2e-16 ***
EthnicityAsian	-18.69	65.02	-0.287	0.774
EthnicityCaucasian	-12.50	56.68	-0.221	0.826

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.9 on 397 degrees of freedom

Multiple R-squared: 0.0002188, Adjusted R-squared: -0.004818

F-statistic: 0.04344 on 2 and 397 DF, p-value: 0.9575

$H_0: \beta_1 = \beta_2 = 0$

H_A : At least

From pp 85 - 86.

$$\hat{\text{Balance}} = \begin{cases} 531.00 & \text{if African American} \\ 531 - 18.69 & \text{if Asian} \\ 531 - 12.50 & \text{if Caucasian} \end{cases}$$

one $\beta_j \neq 0$

Lab Questions

Use the model from the previous slide to answer the questions.

1. What is the estimated average balance for a person that is Asian?

a) 531

b) -18.69

c) -12.50

d) 512.31



2. Do we have evidence that there is a difference in balance based on ethnicity?

a) Yes

b) No



Two Important Assumptions

1. The **additive** assumptions means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors.
2. The **linear** assumptions means that the change in the response Y due to a one-unit change in X_j is constants, regardless of the value of X_j .

Recall Stock Price Data: Best Subset of Predictors

```
stock2.lm <- lm(Stock_Index_Price~Interest_Rate+Unemployment_Rate,  
               data = stock_price)  
summary(stock2.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1798.4	899.2	2.000	0.05861	.
Interest_Rate	345.5	111.4	3.103	0.00539	**
Unemployment_Rate	-250.1	117.9	-2.121	0.04601	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.56 on 21 degrees of freedom
Multiple R-squared: 0.8976, Adjusted R-squared: 0.8879
F-statistic: 92.07 on 2 and 21 DF, p-value: 4.043e-11

$$\hat{Stock_Index_Price} = 1798.4 + 345.5 \times Interest_Rate - 250.1 \times Unemployment_Rate$$

Removing The Additive Assumption

- In our *stock price* data, we conclude that both *unemployment rate* and *interest rate* seem to be associated with *stock index price*.
- The linear model that we formed assumes that the effect on the stock index price of increasing one of percent of the interest rate is independent of the unemployment rate.
- For example the linear models states that the average effect on *stock index price* of a one percent increase in *interest rate* is always β_1 regardless of the *unemployment rate*.
- However, do we think that interest rate is independent of unemployment rate?
- Thus we can use an **interaction** term between interest rate and unemployment rate.
- The simplest method to construct an interaction term is to multiply two predictors together.

Model With Interaction Term

For our example we can have the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$\begin{aligned} \text{Stock_Index_Price} &= \beta_0 + \beta_1 \times \text{Interest_Rate} + \beta_2 \times \text{Unemployment_Rate} \\ &+ \beta_3 \times \text{Interest_Rate} \times \text{Unemployment_Rate} + \epsilon \end{aligned}$$

$$\begin{aligned} &= \beta_0 + \beta_1 \times \text{Interest_Rate} \\ &+ (\beta_2 + \beta_3 \times \text{Interest_Rate}) \times \text{Unemployment_Rate} + \epsilon \end{aligned}$$

We can interpret β_3 as the increase in the effectiveness of *unemployment rate* for a one unit increase in *interest rate* (or vice-versa).

Summary

```
> #Stock Price data
> #Model with Interaction term
> stock.int = lm(Stock_Index_Price~Interest_Rate*Unemployment_Rate)
> summary(stock.int)
```

Call:
lm(formula = Stock_Index_Price ~ Interest_Rate * Unemployment_Rate)
Residuals:

Min	1Q	Median	3Q	Max
-156.009	-40.238	-8.873	52.131	122.073

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2522.85	2634.04	0.958	0.350
Interest_Rate	-32.49	1293.06	-0.025	0.980
Unemployment_Rate	-380.76	461.09	-0.826	0.419
Interest_Rate:Unemployment_Rate	68.54	233.53	0.293	0.772

Residual standard error: 72.15 on 20 degrees of freedom
Multiple R-squared: 0.8981, Adjusted R-squared: 0.8828
F-statistic: 58.74 on 3 and 20 DF, p-value: 4.266e-10

$P\text{-value} = 0.980$ $H_0: \beta_1 = 0$, given $\beta_2 \neq 0$ and $\beta_3 \neq 0$

Interpretation

$$\text{Stock_Index_Price} \approx 2522.85 - 32.49 \times \text{Interest_Rate} - 308.76 \times \text{Unemployment_Rate} \\ + 68.54 \times (\text{Interest_Rate} \times \text{Unemployment_Rate})$$

$$= 2522.85 - 32.49 \times \text{Interest_Rate} \\ + (63.54 \times \text{Interest_Rate} - 308.76) \times \text{Unemployment_Rate}$$

Interpretation

$$\begin{aligned} \text{Stock_Index_Price} &\approx 2522.85 - 32.49 \times \text{Interest_Rate} - \overset{386}{\cancel{308.76}} \times \text{Unemployment_Rate} \\ &\quad + 68.54 \times (\text{Interest_Rate} \times \text{Unemployment_Rate}) \end{aligned}$$

$$\begin{aligned} &= 2522.85 - 32.49 \times \text{Interest_Rate} \\ &\quad + (63.54 \times \text{Interest_Rate} - \overset{380.74}{\cancel{308.76}}) \times \text{Unemployment_Rate} \end{aligned}$$

- Increasing the unemployment rate by 1% will increase the stock index price by $63.54 \times \text{Interest_Rate} - \overset{380.74}{\cancel{308.76}}$.

Interpretation

$$\text{Stock_Index_Price} \approx 2522.85 - 32.49 \times \text{Interest_Rate} - 308.76 \times \text{Unemployment_Rate} \\ + 68.54 \times (\text{Interest_Rate} \times \text{Unemployment_Rate})$$

$$= 2522.85 - 32.49 \times \text{Interest_Rate} \\ + (63.54 \times \text{Interest_Rate} - 308.76) \times \text{Unemployment_Rate}$$

- Increasing the unemployment rate by 1% will increase the stock index price by $63.54 \times \text{Interest_Rate} - 308.76$.
- However, notice the p-values when testing $H_0 : \beta_j = 0$. Since all are greater than 0.05, this means that at least one of these terms are not needed in the model.

Using The `step()` Function

```
> step(stock.int)
Start:  AIC=209
Stock_Index_Price ~ Interest_Rate * Unemployment_Rate
```

	Df	Sum of Sq	RSS	AIC
- Interest_Rate:Unemployment_Rate	1	448.39	104559	207.11
<none>			104110	209.00

```
Step:  AIC=207.11
Stock_Index_Price ~ Interest_Rate + Unemployment_Rate
```

	Df	Sum of Sq	RSS	AIC
<none>			104559	207.11
- Unemployment_Rate	1	22394	126953	209.76
- Interest_Rate	1	47932	152491	214.16

```
Call:
lm(formula = Stock_Index_Price ~ Interest_Rate + Unemployment_Rate)
```

```
Coefficients:
(Intercept)      Interest_Rate  Unemployment_Rate
  1798.4           345.5           -250.1
```

Roller Coaster Model

- Speed = top speed of a roller coaster
- Type = 2 if steel 1 if wood
- Height = tallest point of the roller coaster
- Model:

$$\begin{aligned}\text{Speed} &\approx \beta_0 + \beta_1 \times \text{Height} + \begin{cases} \beta_2 & \text{if steel} \\ 0 & \text{if wood} \end{cases} \\ &= \beta_1 \times \text{Height} + \begin{cases} \beta_0 + \beta_2 & \text{if steel} \\ \beta_0 & \text{if wood} \end{cases}\end{aligned}$$

Summary

```
> rollercoaster$Type=factor(rollercoaster$Type)
> roller.lm = lm(Speed~Height+Type,data = rollercoaster)
> summary(roller.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	29.294434	1.563647	18.735	< 2e-16 ***
Height	0.240109	0.008658	27.733	< 2e-16 ***
Type2	-6.114395	1.486380	-4.114	7.54e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.768 on 110 degrees of freedom

Multiple R-squared: 0.8749, Adjusted R-squared: 0.8727

F-statistic: 384.8 on 2 and 110 DF, p-value: < 2.2e-16

$$\hat{\text{Speed}} = \begin{cases} 29.294 + 0.24 * \text{height} & \text{if wood} \\ 23.18004 + 0.24 * \text{height} & \text{if steel} \end{cases}$$

Lab Questions

3. From the output is at least one of the predictors associated with the speed of the roller-coaster?

☒ a) Yes

b) No

4. From the output which predictor(s) can be used in the model, given that the other predictor is in the model?

a) Height

b) Type

☒ c) Both height and type

d) Neither height nor type

5. What is the estimate average speed for a wooden roller-coaster with height of 120 ft.?

a) 29.29

b) -6.11

c) 0.24

☒ d) 58.10

Interaction Terms Between Quantitative and Categorical Predictors

- There maybe an interaction between height and type of roller coaster.
- With the interaction term the model becomes:

$$\begin{aligned}\text{Speed} &\approx \beta_0 + \beta_1 \times \text{height} + \begin{cases} \beta_2 + \beta_3 \times \text{height}, & \text{if steel} \\ 0 & \text{if wood} \end{cases} \\ &= \begin{cases} \overset{\text{y-intercept}}{(\beta_0 + \beta_2)} + \overset{\text{slope}}{(\beta_1 + \beta_3)} \times \text{height}, & \text{if } \underline{\text{steel}} \\ \underset{\text{y-intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} \times \text{height}, & \text{if } \underline{\underline{\text{wood}}} \end{cases}\end{aligned}$$

Summary

```
> roller.int = lm(Speed~Height*Type,data = rollercoaster)
> summary(roller.int)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	30.35998	3.99872	7.592	1.14e-11	***
Height	0.22985	0.03645	6.306	6.27e-09	***
Type2	-7.25747	4.21806	-1.721	0.0882	.
Height:Type2	0.01088	0.03753	0.290	0.7726	<i>p-value > 0.1</i>

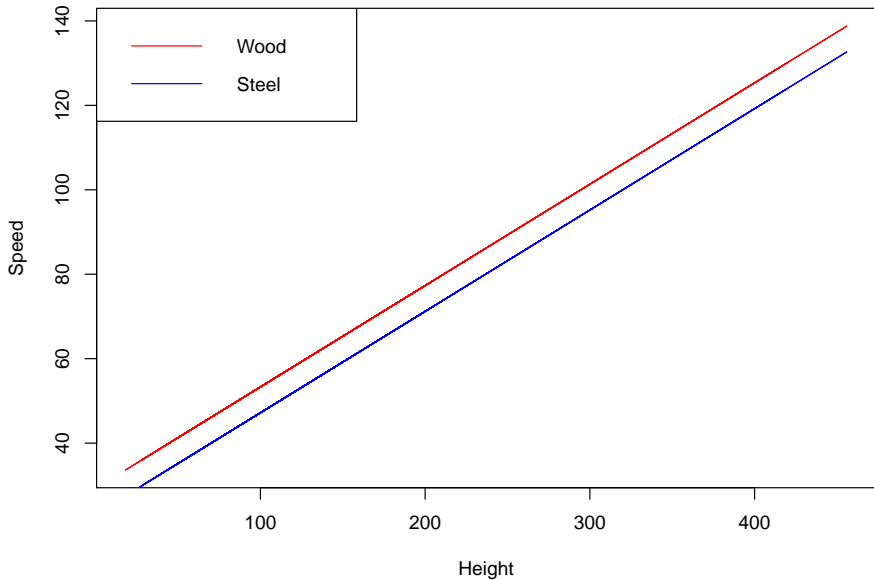
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.796 on 109 degrees of freedom

Multiple R-squared: 0.875, Adjusted R-squared: 0.8716

F-statistic: 254.4 on 3 and 109 DF, p-value: < 2.2e-16

No Interaction Term



Two Separate Regression Lines

