



**Applied Regression Modeling:  
A Business Approach**  
**Chapter 4: Regression Model Building I**  
**Sections 4.2–4.3**

by Iain Pardoe

## 4.2 Interactions

### Interactions

SALES1 scatterplot

Data and model  
Interaction model  
results

Interaction  
hypothesis test

SALES2 data

SALES2 scatterplot

Interaction model  
results

No interaction  
model results

Interactions in  
practice

## 4.3 Qualitative predictors

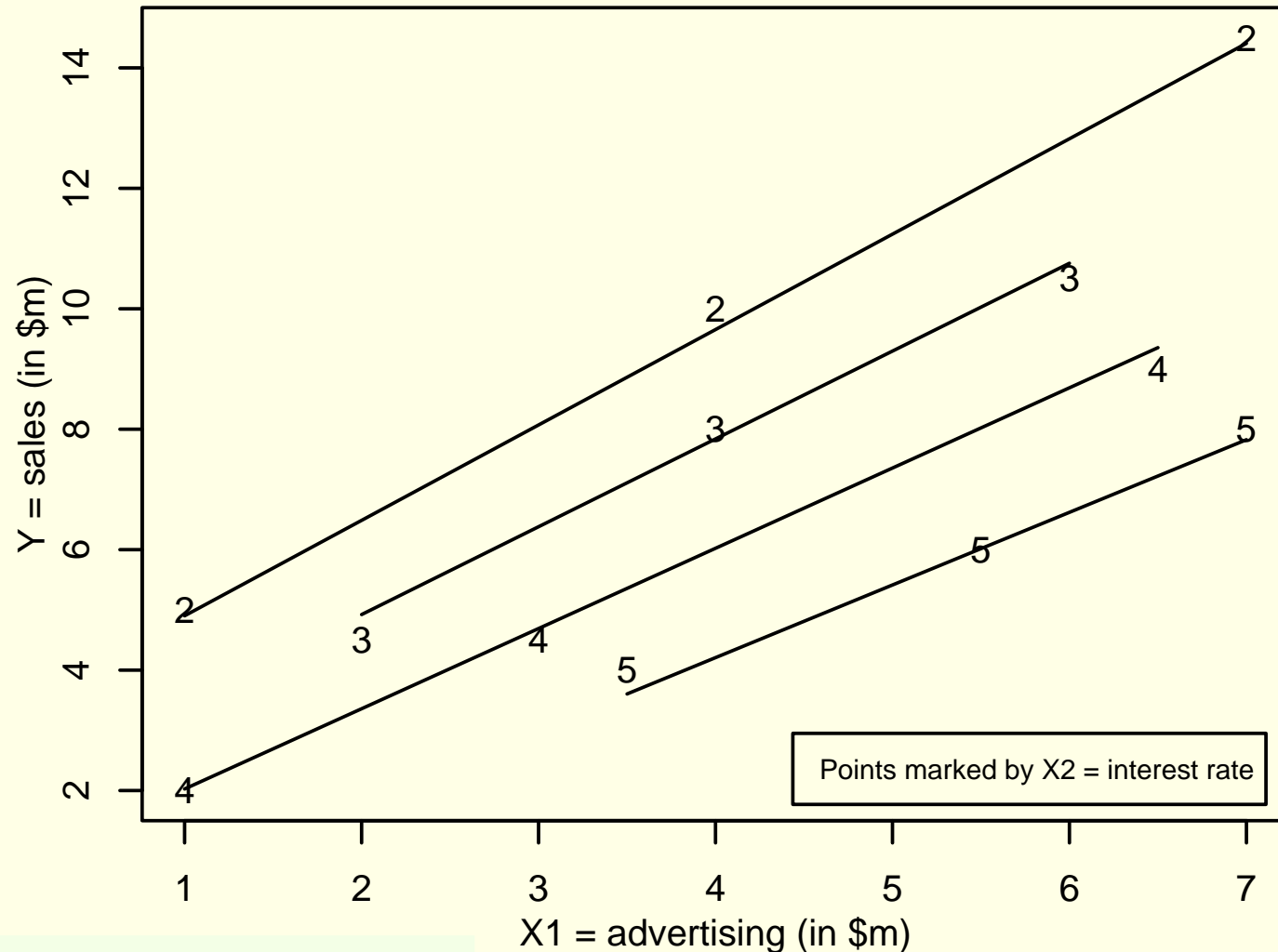
### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Suppose the association between one predictor ( $X_1$ ) and the response variable ( $Y$ ) varies according to the value of another predictor ( $X_2$ ).
- Can model this *interaction* by including the term  $X_1X_2$  in our model:
  - value of  $X_1X_2$  for each sample observation is the corresponding values of  $X_1$  and  $X_2$  multiplied together.
- **SALES1** data:  $Y$  = annual sales (in \$m),  
 $X_1$  = annual spending on advertising (in \$m),  
 $X_2$  = the prevailing interest rate (in %).
- Suppose sales tend to increase by \$1.58m for each additional \$1m we spend on advertising when the prevailing interest rate is 2%, but the increase is only \$1.21m when the interest rate is 5%.

# SALES1 scatterplot

Each line represents the association between sales and advertising for a fixed interest rate.



What are the slopes of the upper/lower lines?

## 4.2 Interactions

Interactions

SALES1 scatterplot

Data and model

Interaction model results

Interaction hypothesis test

SALES2 data

SALES2 scatterplot

Interaction model results

No interaction model results

Interactions in practice

## 4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

$Y$ (sales)	4.0	6.0	8.0	2.0	4.5	9.0	4.5	8.0	10.5	5.0	10.0	14.5
$X_1$ (advert)	3.5	5.5	7.0	1.0	3.0	6.5	2.0	4.0	6.0	1.0	4.0	7.0
$X_2$ (interest)	5	5	5	4	4	4	3	3	3	2	2	2

- Interaction model:

$$E(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2.$$

- When  $X_2 = 2$ ,  $E(Y) = (b_0 + 2b_2) + (b_1 + 2b_3)X_1$ ;
  - When  $X_2 = 3$ ,  $E(Y) = (b_0 + 3b_2) + (b_1 + 3b_3)X_1$ ;
  - When  $X_2 = 4$ ,  $E(Y) = (b_0 + 4b_2) + (b_1 + 4b_3)X_1$ ;
  - When  $X_2 = 5$ ,  $E(Y) = (b_0 + 5b_2) + (b_1 + 5b_3)X_1$ .
- In other words, each “line” has different slopes (as long as  $b_3 \neq 0$ ).

# Interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.997 <sup>a</sup>	0.994	0.992	0.3075

<sup>a</sup> Predictors: (Intercept), X1, X2, X1X2.

## Parameters<sup>a</sup>

Model	Estimate	Std. Error	t-stat	Pr(>  t )
1 (Intercept)	5.941	0.662	8.979	0.000
X1	1.836	0.135	13.611	0.000
X2	-1.312	0.197	-6.669	0.000
X1X2	-0.126	0.039	-3.261	0.012

<sup>a</sup> Response variable: Y.

- $$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_1 X_2$$
$$= 5.941 + 1.836 X_1 - 1.312 X_2 - 0.126 X_1 X_2.$$

4.2 Interactions

Interactions

SALES1 scatterplot

Data and model

Interaction model results

Interaction

hypothesis test

SALES2 data

SALES2 scatterplot

Interaction model results

No interaction model results

Interactions in practice

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

# Interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.997 <sup>a</sup>	0.994	0.992	0.3075

<sup>a</sup> Predictors: (Intercept), X1, X2, X1X2.

## Parameters<sup>a</sup>

Model	Estimate	Std. Error	t-stat	Pr(>  t )
1 (Intercept)	5.941	0.662	8.979	0.000
X1	1.836	0.135	13.611	0.000
X2	-1.312	0.197	-6.669	0.000
X1X2	-0.126	0.039	-3.261	0.012

<sup>a</sup> Response variable: Y.

- $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_1 X_2$   
 $= 5.941 + 1.836 X_1 - 1.312 X_2 - 0.126 X_1 X_2.$
- When  $X_2 = 2$ ,  $\hat{Y} = 3.32 + 1.58 X_1.$

4.2 Interactions

Interactions

SALES1 scatterplot

Data and model

Interaction model results

Interaction

hypothesis test

SALES2 data

SALES2 scatterplot

Interaction model results

No interaction model results

Interactions in practice

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

# Interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.997 <sup>a</sup>	0.994	0.992	0.3075

<sup>a</sup> Predictors: (Intercept), X1, X2, X1X2.

## Parameters<sup>a</sup>

Model	Estimate	Std. Error	t-stat	Pr(>  t )
1 (Intercept)	5.941	0.662	8.979	0.000
X1	1.836	0.135	13.611	0.000
X2	-1.312	0.197	-6.669	0.000
X1X2	-0.126	0.039	-3.261	0.012

<sup>a</sup> Response variable: Y.

- $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_1 X_2$   
 $= 5.941 + 1.836 X_1 - 1.312 X_2 - 0.126 X_1 X_2.$
- When  $X_2 = 2$ ,  $\hat{Y} = 3.32 + 1.58 X_1.$
- When  $X_2 = 3/4/5$ ,  $\hat{Y} = ?$

- 4.2 Interactions
  - Interactions
  - SALES1 scatterplot
  - Data and model
  - Interaction model results**
  - Interaction hypothesis test
  - SALES2 data
  - SALES2 scatterplot
  - Interaction model results
  - No interaction model results
  - Interactions in practice
- 4.3 Qualitative predictors
  - 4.3.1 Qualitative predictors: 2 levels
  - 4.3.2 Qualitative predictors: >2 levels



# Interaction hypothesis test

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?

## 4.2 Interactions

Interactions

SALES1 scatterplot

Data and model

Interaction model

results

Interaction  
hypothesis test

SALES2 data

SALES2 scatterplot

Interaction model

results

No interaction

model results

Interactions in

practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

# Interaction hypothesis test

## 4.2 Interactions

### Interactions

SALES1 scatterplot

Data and model  
Interaction model  
results

### Interaction hypothesis test

SALES2 data

SALES2 scatterplot

Interaction model  
results

No interaction  
model results

Interactions in  
practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?
- NH:  $b_3 = 0$  versus AH:  $b_3 \neq 0$ .

# Interaction hypothesis test

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?
- NH:  $b_3 = 0$  versus AH:  $b_3 \neq 0$ .
- t-statistic =  $\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{-0.126 - 0}{0.039} = -3.26$ .

# Interaction hypothesis test

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?
- NH:  $b_3 = 0$  versus AH:  $b_3 \neq 0$ .
- t-statistic =  $\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{-0.126 - 0}{0.039} = -3.26$ .
- Significance level = 5%.

# Interaction hypothesis test

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?
- NH:  $b_3 = 0$  versus AH:  $b_3 \neq 0$ .
- t-statistic =  $\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{-0.126 - 0}{0.039} = -3.26$ .
- Significance level = 5%.
- Critical value,  $TINV(0.05, 8)$ , is 2.31.
- p-value,  $TDIST(-3.26, 8, 2)$ , is 0.012.

# Interaction hypothesis test

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model

#### results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model

#### results

#### No interaction

#### model results

#### Interactions in

#### practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?
- NH:  $b_3 = 0$  versus AH:  $b_3 \neq 0$ .
- t-statistic =  $\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{-0.126 - 0}{0.039} = -3.26$ .
- Significance level = 5%.
- Critical value,  $TINV(0.05, 8)$ , is 2.31.
- p-value,  $TDIST(-3.26, 8, 2)$ , is 0.012.
- Since absolute t-statistic (3.26) > critical value (2.31) and p-value (0.012) < signif. level (0.05), reject NH in favor of AH.

# Interaction hypothesis test

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction

#### model results

#### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?
- NH:  $b_3 = 0$  versus AH:  $b_3 \neq 0$ .
- t-statistic =  $\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{-0.126 - 0}{0.039} = -3.26$ .
- Significance level = 5%.
- Critical value,  $TINV(0.05, 8)$ , is 2.31.
- p-value,  $TDIST(-3.26, 8, 2)$ , is 0.012.
- Since absolute t-statistic (3.26) > critical value (2.31) and p-value (0.012) < signif. level (0.05), reject NH in favor of AH.
- Sample data favor  $b_3 \neq 0$  (at a 5% signif. level).

# Interaction hypothesis test

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model

#### results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model

#### results

#### No interaction

#### model results

#### Interactions in

#### practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

- Is there sufficient evidence of interaction (can we rule out the possibility of parallel population lines)?
- NH:  $b_3 = 0$  versus AH:  $b_3 \neq 0$ .
- t-statistic =  $\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{-0.126 - 0}{0.039} = -3.26$ .
- Significance level = 5%.
- Critical value,  $TINV(0.05, 8)$ , is 2.31.
- p-value,  $TDIST(-3.26, 8, 2)$ , is 0.012.
- Since absolute t-statistic (3.26) > critical value (2.31) and p-value (0.012) < signif. level (0.05), reject NH in favor of AH.
- Sample data favor  $b_3 \neq 0$  (at a 5% signif. level).
- The association between  $X_1$  and  $Y$  does appear to depend on the value of  $X_2$ .



# SALES2 data

$Y$ (sales)	3.8	7.8	7.9	6.5	10.6	13.3	14.7	16.1	18.7	18.8	22.9	24.2
$X_1$ (advert)	3.5	5.5	7.0	1.0	3.0	6.5	2.0	4.0	6.0	1.0	4.0	7.0
$X_2$ (stores)	1	1	1	2	2	2	3	3	3	4	4	4

- **SALES2** data:  $Y$  = annual sales (in \$m),  
 $X_1$  = annual spending on advertising (in \$m),  
 $X_2$  = number of retail stores.
- Does the relationship between sales and spending on advertising vary according to the number of stores operated?

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

Data and model  
Interaction model  
results

Interaction  
hypothesis test

### SALES2 data

#### SALES2 scatterplot

Interaction model  
results

No interaction  
model results  
Interactions in  
practice

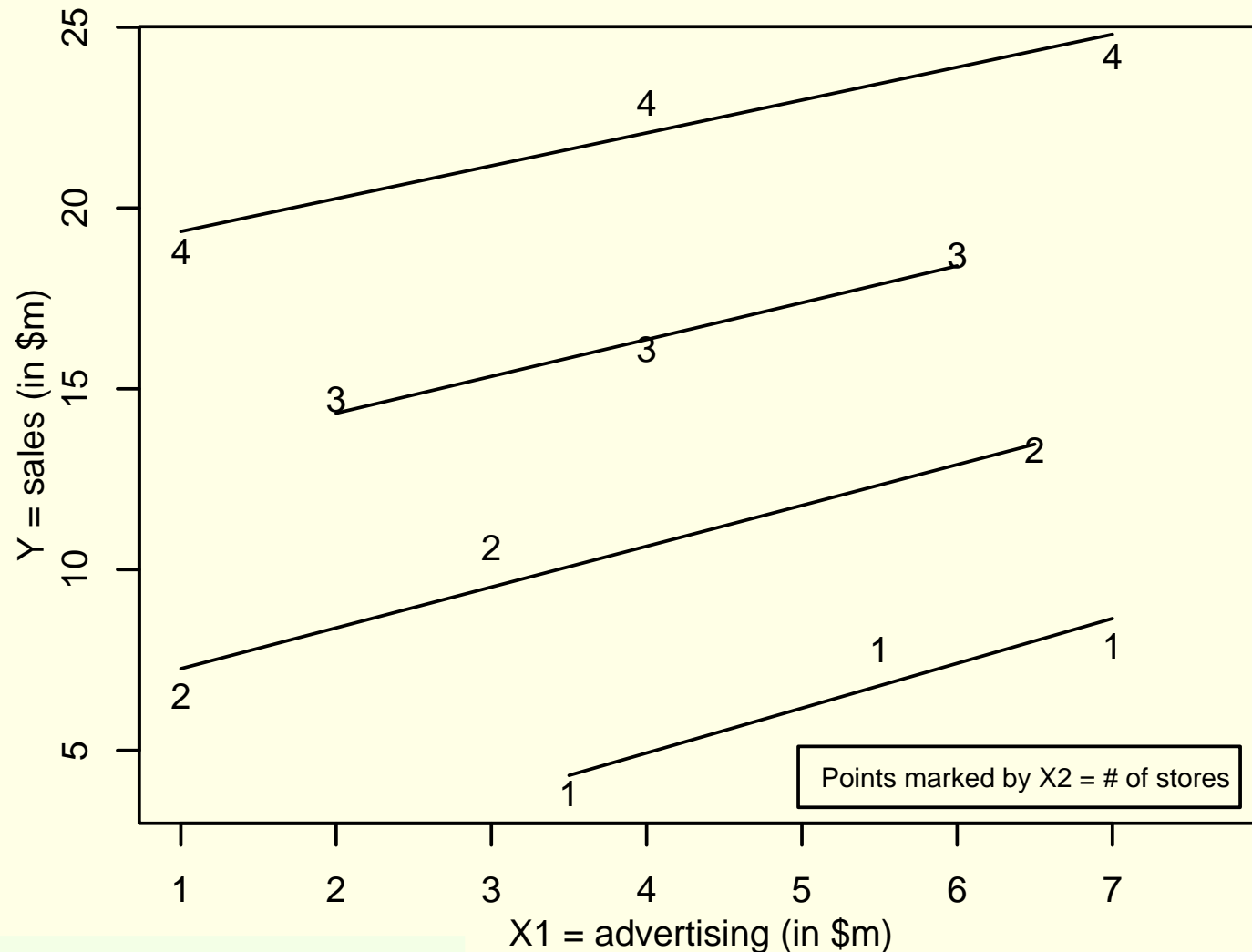
## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

# SALES2 scatterplot

Each line represents the association between sales and advertising for a fixed number of stores.



Could the lines be exactly parallel in the population?

# Interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.995 <sup>a</sup>	0.989	0.985	0.8112

<sup>a</sup> Predictors: (Intercept), X1, X2, X1X2.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
1	(Intercept)	-6.182	1.540	-4.014	0.004
	X1	1.349	0.297	4.549	0.002
	X2	6.156	0.519	11.864	0.000
	X1X2	-0.110	0.102	-1.080	0.312

<sup>a</sup> Response variable: Y.

- Since p-value (0.312) > signif. level (0.05), cannot reject NH ( $b_3 = 0$ ) in favor of AH ( $b_3 \neq 0$ ).

### 4.2 Interactions

#### Interactions

#### SALES1 scatterplot

#### Data and model Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

### 4.3 Qualitative predictors

#### 4.3.1 Qualitative predictors: 2 levels

#### 4.3.2 Qualitative predictors: >2 levels

# Interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.995 <sup>a</sup>	0.989	0.985	0.8112

<sup>a</sup> Predictors: (Intercept), X1, X2, X1X2.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
1	(Intercept)	-6.182	1.540	-4.014	0.004
	X1	1.349	0.297	4.549	0.002
	X2	6.156	0.519	11.864	0.000
	X1X2	-0.110	0.102	-1.080	0.312

<sup>a</sup> Response variable: Y.

- Since p-value (0.312) > signif. level (0.05), cannot reject NH ( $b_3 = 0$ ) in favor of AH ( $b_3 \neq 0$ ).
- The association between  $X_1$  and  $Y$  does not appear to depend on the value of  $X_2$ .

### 4.2 Interactions

#### Interactions

#### SALES1 scatterplot

#### Data and model Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

### 4.3 Qualitative predictors

#### 4.3.1 Qualitative predictors: 2 levels

#### 4.3.2 Qualitative predictors: >2 levels

# No interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.994 <sup>a</sup>	0.987	0.985	0.8186

<sup>a</sup> Predictors: (Intercept), X1, X2.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	-4.769	0.820	-5.818	0.000
	X1	1.053	0.114	9.221	0.000
	X2	5.645	0.215	26.242	0.000

<sup>a</sup> Response variable: Y.

- $\hat{Y} = -4.77 + 1.05X_1 + 5.65X_2.$

### 4.2 Interactions

#### Interactions

#### SALES1 scatterplot

#### Data and model Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

### 4.3 Qualitative predictors

#### 4.3.1 Qualitative predictors: 2 levels

#### 4.3.2 Qualitative predictors: >2 levels

# No interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.994 <sup>a</sup>	0.987	0.985	0.8186

<sup>a</sup> Predictors: (Intercept), X1, X2.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	-4.769	0.820	-5.818	0.000
	X1	1.053	0.114	9.221	0.000
	X2	5.645	0.215	26.242	0.000

<sup>a</sup> Response variable: Y.

- $\hat{Y} = -4.77 + 1.05X_1 + 5.65X_2$ .
- When  $X_2 = 1$ ,  $\hat{Y} = 0.88 + 1.05X_1$ .

### 4.2 Interactions

#### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model results

#### Interaction

#### hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

### 4.3 Qualitative predictors

#### 4.3.1 Qualitative predictors: 2 levels

#### 4.3.2 Qualitative predictors: >2 levels

# No interaction model results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.994 <sup>a</sup>	0.987	0.985	0.8186

<sup>a</sup> Predictors: (Intercept), X1, X2.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	-4.769	0.820	-5.818	0.000
	X1	1.053	0.114	9.221	0.000
	X2	5.645	0.215	26.242	0.000

<sup>a</sup> Response variable: Y.

- $\hat{Y} = -4.77 + 1.05X_1 + 5.65X_2$ .
- When  $X_2 = 1$ ,  $\hat{Y} = 0.88 + 1.05X_1$ .
- When  $X_2 = 2/3/4$ ,  $\hat{Y} = ?$

### 4.2 Interactions

#### Interactions

#### SALES1 scatterplot

#### Data and model Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

#### Interactions in practice

### 4.3 Qualitative predictors

#### 4.3.1 Qualitative predictors: 2 levels

#### 4.3.2 Qualitative predictors: >2 levels

# Interactions in practice

- SALES1/SALES2 scatterplots illustrate interaction when we have just two predictor variables, one of which has just a few distinct values.
  - But, can employ interaction terms in more general models with more than two predictors.

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model results

#### Interaction

#### hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model

#### results

#### No interaction

#### model results

### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels



# Interactions in practice

- SALES1/SALES2 scatterplots illustrate interaction when we have just two predictor variables, one of which has just a few distinct values.
  - But, can employ interaction terms in more general models with more than two predictors.
- When using  $X_1X_2$ , say, in a model,  $X_1$  and  $X_2$  are often included, *regardless* of their significance—this is called preserving *hierarchy*.

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

# Interactions in practice

- SALES1/SALES2 scatterplots illustrate interaction when we have just two predictor variables, one of which has just a few distinct values.
  - But, can employ interaction terms in more general models with more than two predictors.
- When using  $X_1X_2$ , say, in a model,  $X_1$  and  $X_2$  are often included, *regardless* of their significance—this is called preserving *hierarchy*.
- Best if interactions suggested *before* looking at the data from background knowledge about the situation or from theoretical arguments.

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

#### Data and model

#### Interaction model results

#### Interaction hypothesis test

#### SALES2 data

#### SALES2 scatterplot

#### Interaction model results

#### No interaction model results

### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

# Interactions in practice

- SALES1/SALES2 scatterplots illustrate interaction when we have just two predictor variables, one of which has just a few distinct values.
  - But, can employ interaction terms in more general models with more than two predictors.
- When using  $X_1X_2$ , say, in a model,  $X_1$  and  $X_2$  are often included, *regardless* of their significance—this is called preserving *hierarchy*.
- Best if interactions suggested *before* looking at the data from background knowledge about the situation or from theoretical arguments.
- Can also combine concepts of interaction and transformations to produce increasingly sophisticated models capable of capturing quite complex relationships between variables.

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

Data and model  
Interaction model  
results

Interaction  
hypothesis test

#### SALES2 data

#### SALES2 scatterplot

Interaction model  
results

No interaction  
model results

### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

# Interactions in practice

- SALES1/SALES2 scatterplots illustrate interaction when we have just two predictor variables, one of which has just a few distinct values.
  - But, can employ interaction terms in more general models with more than two predictors.
- When using  $X_1X_2$ , say, in a model,  $X_1$  and  $X_2$  are often included, *regardless* of their significance—this is called preserving *hierarchy*.
- Best if interactions suggested *before* looking at the data from background knowledge about the situation or from theoretical arguments.
- Can also combine concepts of interaction and transformations to produce increasingly sophisticated models capable of capturing quite complex relationships between variables.
- Dangers: overcomplicating things unnecessarily, overfitting sample data (poor generalizability).

## 4.2 Interactions

### Interactions

#### SALES1 scatterplot

Data and model  
Interaction model  
results

Interaction  
hypothesis test

#### SALES2 data

#### SALES2 scatterplot

Interaction model  
results

No interaction  
model results

### Interactions in practice

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

4.2 Interactions

4.3 Qualitative predictors

**Hypothetical salaries**

Indicator variable

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Is there any suggestion of gender discrimination here?

Gender	Salary (\$k)
Female	20
Female	30
Male	45
Male	60

# Hypothetical salaries

Is there any suggestion of gender discrimination here?

Gender	Salary (\$k)	Job title
Female	20	Clerical
Female	30	Clerical
Male	45	Managerial
Male	60	Managerial

4.2 Interactions

4.3 Qualitative predictors

**Hypothetical salaries**

Indicator variable

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

# Hypothetical salaries

Is there any suggestion of gender discrimination here?

Gender	Salary (\$k)	Job title
Female	20	Clerical
Female	30	Clerical
Male	45	Managerial
Male	60	Managerial
Female	35	Managerial
Female	50	Managerial
Male	25	Clerical
Male	40	Clerical

4.2 Interactions

4.3 Qualitative predictors

Hypothetical salaries

Indicator variable

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

# Hypothetical salaries

Is there any suggestion of gender discrimination here?

Gender	Salary (\$k)	Job title	Exp (years)
Female	20	Clerical	5
Female	30	Clerical	15
Male	45	Managerial	15
Male	60	Managerial	30
Female	35	Managerial	5
Female	50	Managerial	20
Male	25	Clerical	10
Male	40	Clerical	25

What is the “salary formula” at this company?

4.2 Interactions

4.3 Qualitative predictors

Hypothetical salaries

Indicator variable

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels



4.2 Interactions

4.3 Qualitative predictors

Hypothetical salaries

Indicator variable

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

- Gender and job title are *qualitative* (or *categorical*) variables.
- To use qualitative information in a regression model we need to translate the categories into numbers a computer can understand.
- For example, define an indicator (dummy) variable  $D$ :

Job title	$D$
Clerical	0
Managerial	1

- Salary formula is  $\text{salary} = 15 + \text{experience} + 15D$ .
  - Clerical worker with 5 year's experience:

4.2 Interactions

4.3 Qualitative predictors

Hypothetical salaries

Indicator variable

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

- Gender and job title are *qualitative* (or *categorical*) variables.
- To use qualitative information in a regression model we need to translate the categories into numbers a computer can understand.
- For example, define an indicator (dummy) variable  $D$ :

Job title	$D$
Clerical	0
Managerial	1

- Salary formula is  $\text{salary} = 15 + \text{experience} + 15D$ .
  - Clerical worker with 5 year's experience:  
 $\text{salary} = 15 + 5 + 15(0) = 20$ .
  - Managerial worker with 5 year's experience:

4.2 Interactions

4.3 Qualitative predictors

Hypothetical salaries

Indicator variable

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

- Gender and job title are *qualitative* (or *categorical*) variables.
- To use qualitative information in a regression model we need to translate the categories into numbers a computer can understand.
- For example, define an indicator (dummy) variable  $D$ :

Job title	$D$
Clerical	0
Managerial	1

- Salary formula is  $\text{salary} = 15 + \text{experience} + 15D$ .
  - Clerical worker with 5 year's experience:  
 $\text{salary} = 15 + 5 + 15(0) = 20$ .
  - Managerial worker with 5 year's experience:  
 $\text{salary} = 15 + 5 + 15(1) = 35$ .

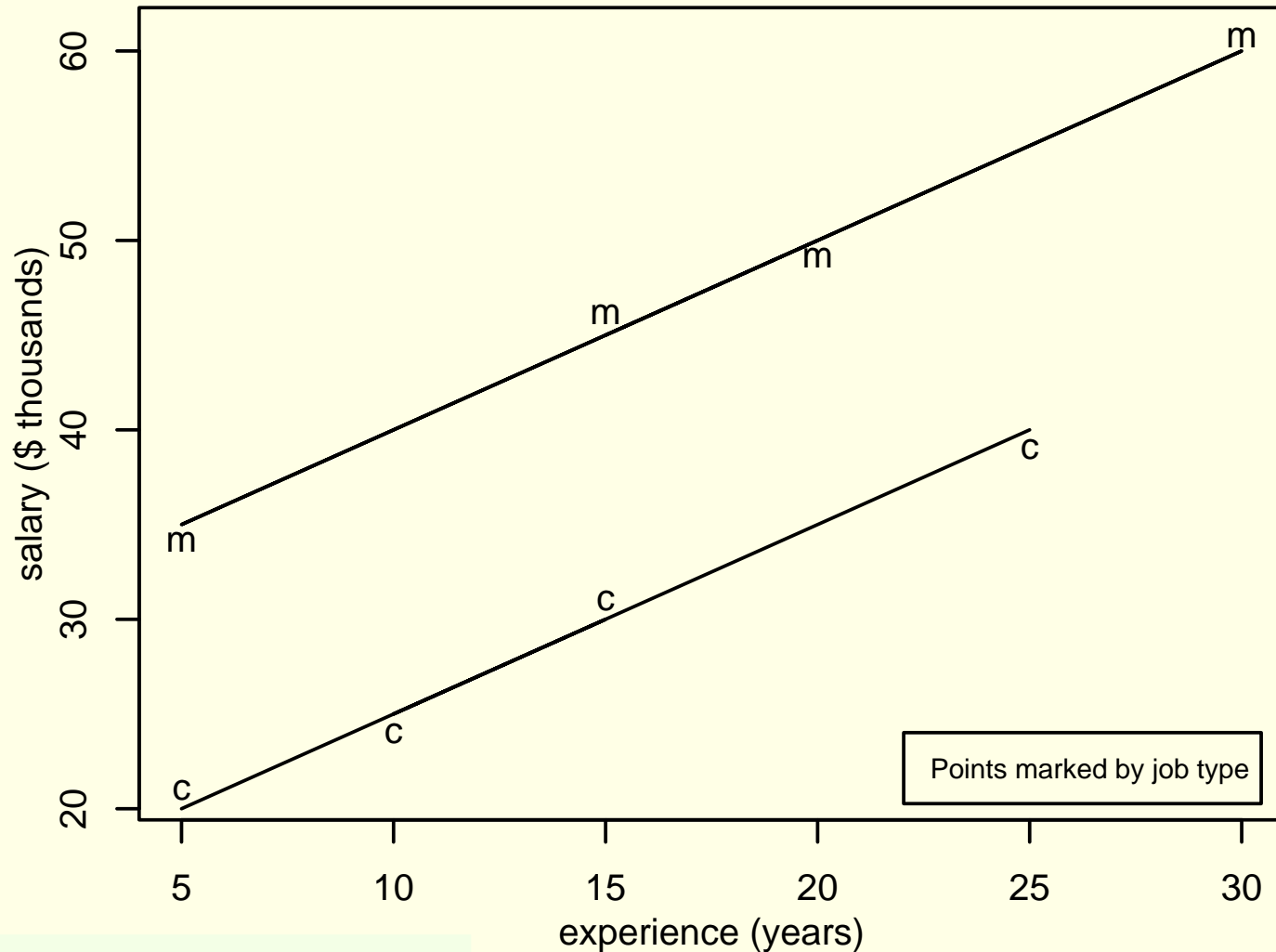
# Qualitative predictors: 2 levels

- Use one indicator variable to code information in a qualitative predictor with two levels (categories).
- One category takes the value 0 (called the reference category):
  - doesn't matter which, but often the most common category or the most meaningful "baseline" category.
- The other category takes the value 1.

Salary (\$k)	Job title	$D$	Exp (years)
21.0	Clerical	0	5
31.0	Clerical	0	15
46.1	Managerial	1	15
60.7	Managerial	1	30
34.1	Managerial	1	5
49.1	Managerial	1	20
24.0	Clerical	0	10
39.0	Clerical	0	25

# Salary/experience by job scatterplot

Each line represents the association between salary and experience for managerial/clerical workers.



How much more are managers paid (for same exp.)?

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial  
Model results with  $D^*=1$  for clerical  
Example you can do in your head

Salary/experience by gender scatterplot

Model results  
Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Model results with D=1 for managerial

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.997 <sup>a</sup>	0.994	0.992	1.2100

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	15.000	0.935	16.035	0.000
	experience	1.000	0.052	19.272	0.000
	D	15.000	0.877	17.096	0.000

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 15.00 + 1.00 \text{ experience} + 15.00D.$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with D=1 for managerial

Model results with D\*=1 for clerical  
Example you can do in your head

Salary/experience by gender scatterplot

Model results  
Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction  
Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Model results with $D=1$ for managerial

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.997 <sup>a</sup>	0.994	0.992	1.2100

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
1	(Intercept)	15.000	0.935	16.035	0.000
	experience	1.000	0.052	19.272	0.000
	D	15.000	0.877	17.096	0.000

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 15.00 + 1.00 \text{ experience} + 15.00D.$
- When  $D=0$  (clerical),  
 $\hat{Y} = 15.00 + 1.00 \text{ experience}.$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical  
 Example you can do in your head

Salary/experience by gender scatterplot

Model results  
 Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Model results with $D=1$ for managerial

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.997 <sup>a</sup>	0.994	0.992	1.2100

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	15.000	0.935	16.035	0.000
	experience	1.000	0.052	19.272	0.000
	D	15.000	0.877	17.096	0.000

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 15.00 + 1.00 \text{ experience} + 15.00D.$
- When  $D=0$  (clerical),  
 $\hat{Y} = 15.00 + 1.00 \text{ experience}.$
- When  $D=1$  (managerial),  $\hat{Y} = ?$



# Model results with $D^*=1$ for clerical

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.997 <sup>a</sup>	0.994	0.992	1.2100

<sup>a</sup> Predictors: (Intercept), experience, Dstar.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	30.000	1.091	27.495	0.000
	experience	1.000	0.052	19.272	0.000
	Dstar	-15.000	0.877	-17.096	0.000

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 30.00 + 1.00 \text{ experience} - 15.00 D^*.$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Model results with $D^*=1$ for clerical

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.997 <sup>a</sup>	0.994	0.992	1.2100

<sup>a</sup> Predictors: (Intercept), experience, Dstar.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	30.000	1.091	27.495	0.000
	experience	1.000	0.052	19.272	0.000
	Dstar	-15.000	0.877	-17.096	0.000

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 30.00 + 1.00 \text{ experience} - 15.00 D^*$ .
- When  $D^* = 0$  (managerial),  
 $\hat{Y} = 30.00 + 1.00 \text{ experience}$ .

# Model results with $D^*=1$ for clerical

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.997 <sup>a</sup>	0.994	0.992	1.2100

<sup>a</sup> Predictors: (Intercept), experience, Dstar.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	30.000	1.091	27.495	0.000
	experience	1.000	0.052	19.272	0.000
	Dstar	-15.000	0.877	-17.096	0.000

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 30.00 + 1.00 \text{ experience} - 15.00D^*$ .
- When  $D^* = 0$  (managerial),  
 $\hat{Y} = 30.00 + 1.00 \text{ experience}$ .
- When  $D^* = 1$  (clerical),  $\hat{Y} = ?$

# Example you can do in your head

- For workers with the same amount of experience, who is getting paid more, males or females?
- How much more are they being paid, on average?

Salary (\$k)	Exp (years)	Gender	<i>D</i>
25.5	1	Female	0
26.5	1	Female	0
26.0	2	Male	1
27.0	2	Female	0
27.0	3	Male	1
28.0	3	Female	0
28.0	4	Male	1
29.0	4	Female	0
28.5	5	Male	1
29.5	5	Male	1

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

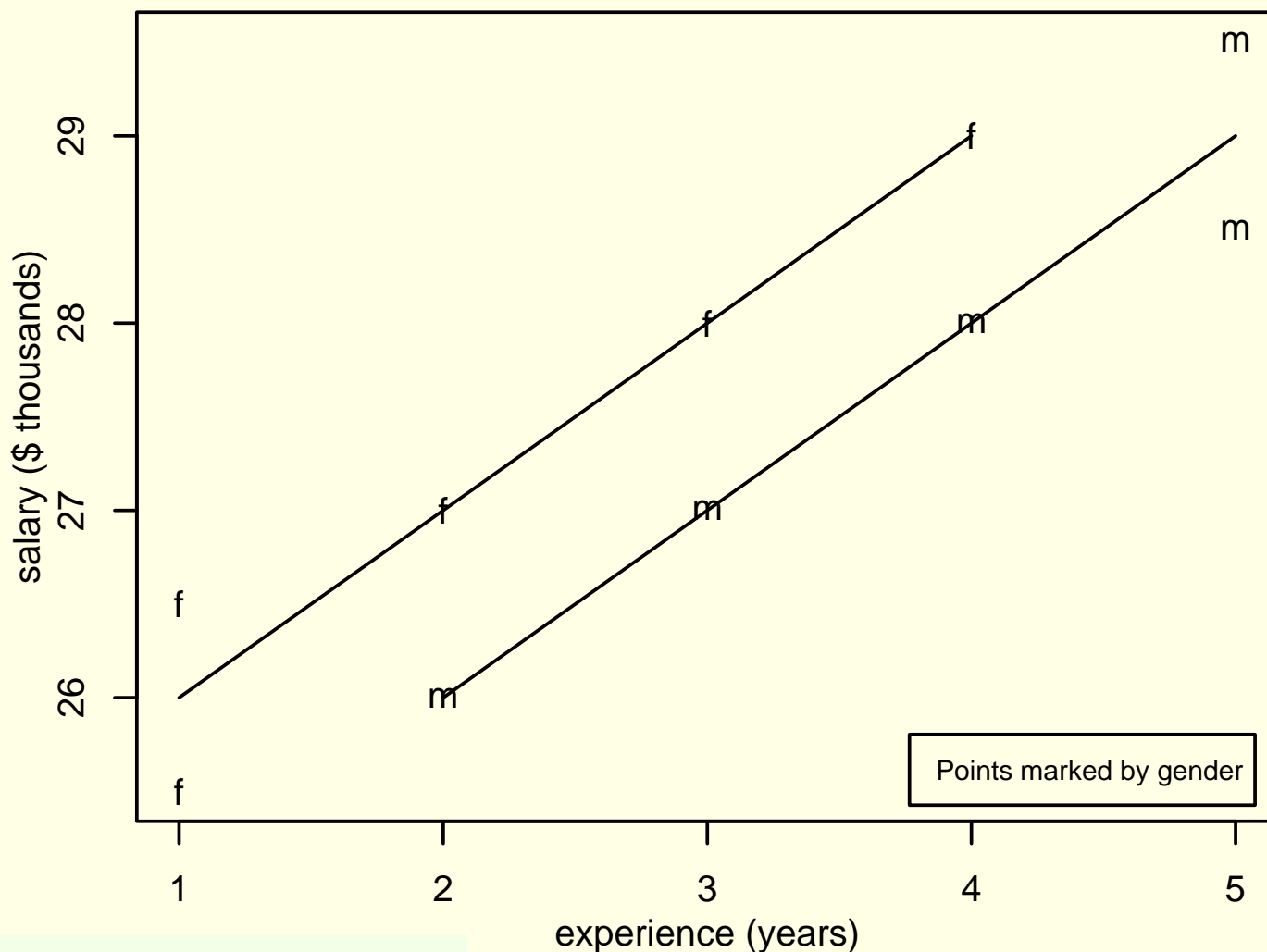
Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Salary/experience by gender scatterplot

Each line represents the association between salary and experience for females/males.



How much more are females paid (for same exp.)?

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.967 <sup>a</sup>	0.935	0.917	0.3780

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
1	(Intercept)	25.000	0.282	88.715	0.000
	experience	1.000	0.102	9.757	0.000
	D	-1.000	0.290	-3.450	0.011

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.00 + 1.00 \text{ experience} - 1.00D.$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with D=1 for managerial

Model results with D\*=1 for clerical

Example you can do in your head

Salary/experience by gender scatterplot

**Model results**

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.967 <sup>a</sup>	0.935	0.917	0.3780

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
1	(Intercept)	25.000	0.282	88.715	0.000
	experience	1.000	0.102	9.757	0.000
	D	-1.000	0.290	-3.450	0.011

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.00 + 1.00 \text{ experience} - 1.00D.$
- When  $D=0$  (female),  
 $\hat{Y} = 25.00 + 1.00 \text{ experience}.$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.967 <sup>a</sup>	0.935	0.917	0.3780

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
1	(Intercept)	25.000	0.282	88.715	0.000
	experience	1.000	0.102	9.757	0.000
	D	-1.000	0.290	-3.450	0.011

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.00 + 1.00 \text{ experience} - 1.00D.$
- When  $D=0$  (female),  
 $\hat{Y} = 25.00 + 1.00 \text{ experience}.$
- When  $D=1$  (male),  $\hat{Y} = ?$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels



# Similar example requiring software

- For workers with the same amount of experience, who is getting paid more, males or females?
- How much more are they being paid, on average?

Salary (\$k)	Exp (years)	Gender	<i>D</i>
25.455	1.0	Female	0
25.956	0.5	Female	0
25.923	2.0	Male	1
26.434	1.5	Female	0
27.023	3.0	Male	1
27.535	2.5	Female	0
28.123	4.0	Male	1
28.621	3.5	Female	0
28.617	5.0	Male	1
28.819	4.5	Male	1

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

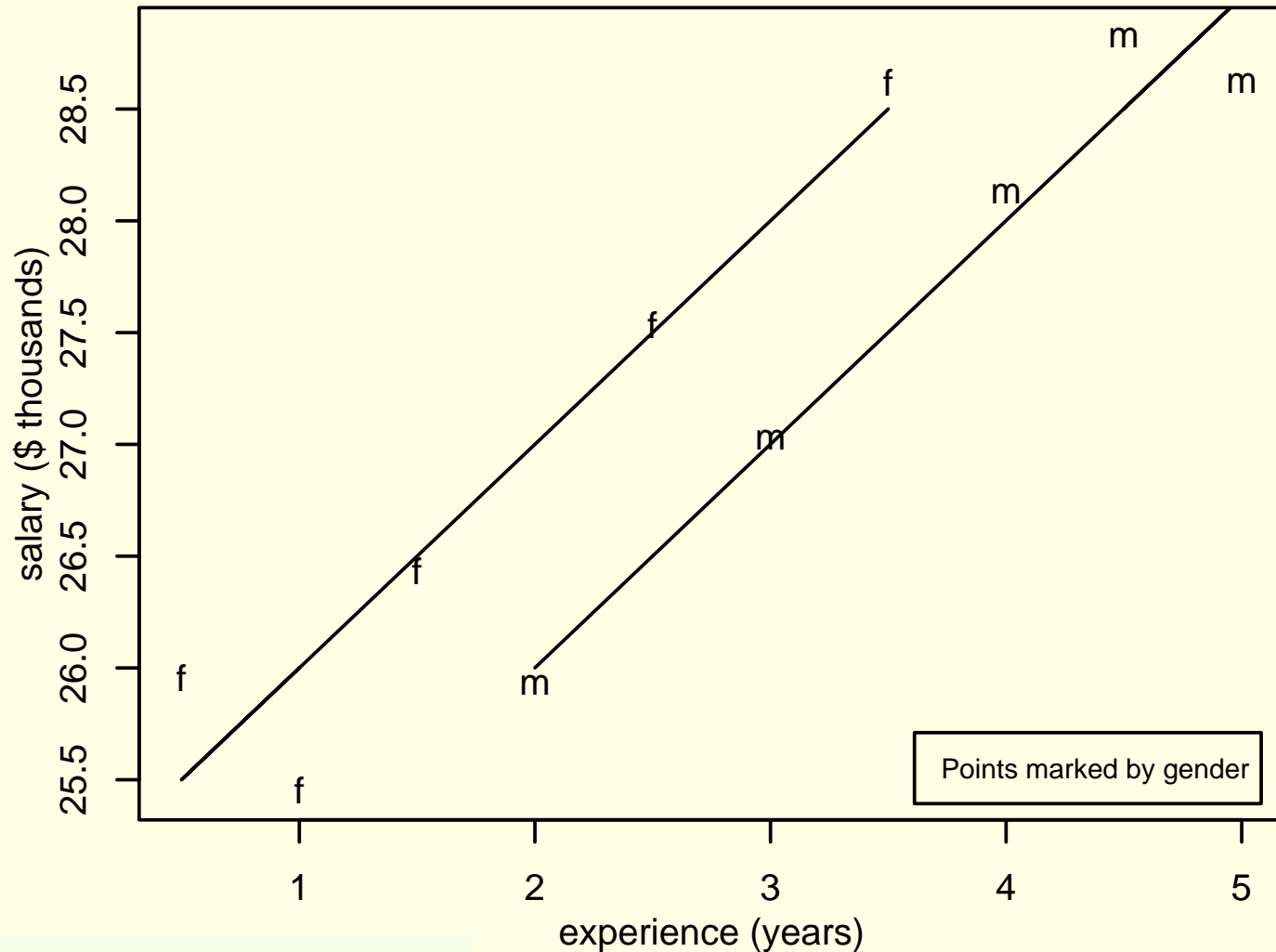
Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Salary/experience by gender scatterplot

Each line represents the association between salary and experience for females/males.



How much more are females paid (for same exp.)?

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 <sup>a</sup>	0.945	0.929	0.3370

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	25.000	0.233	107.136	0.000
	experience	1.000	0.099	10.107	0.000
	D	-1.000	0.284	-3.517	0.010

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.00 + 1.00 \text{ experience} - 1.00D.$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with D=1 for managerial

Model results with D\*=1 for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 <sup>a</sup>	0.945	0.929	0.3370

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	25.000	0.233	107.136	0.000
	experience	1.000	0.099	10.107	0.000
	D	-1.000	0.284	-3.517	0.010

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.00 + 1.00 \text{ experience} - 1.00D.$
- When  $D=0$  (female),  
 $\hat{Y} = 25.00 + 1.00 \text{ experience}.$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 <sup>a</sup>	0.945	0.929	0.3370

<sup>a</sup> Predictors: (Intercept), experience, D.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	25.000	0.233	107.136	0.000
	experience	1.000	0.099	10.107	0.000
	D	-1.000	0.284	-3.517	0.010

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.00 + 1.00 \text{ experience} - 1.00D.$
- When  $D=0$  (female),  
 $\hat{Y} = 25.00 + 1.00 \text{ experience}.$
- When  $D=1$  (male),  $\hat{Y} = ?$

- For workers with the same amount of experience, who is getting paid more, males or females?
- Does it depend on the actual amount of experience?

Salary (\$k)	Exp (years)	Gender	<i>D</i>
25.5	0.5	Female	0
26.0	1.0	Female	0
25.9	2.0	Male	1
26.5	2.0	Female	0
26.9	2.5	Male	1
27.5	3.0	Female	0
27.9	3.5	Male	1
28.5	4.5	Female	0
28.4	3.5	Male	1
28.9	4.0	Male	1

## 4.2 Interactions

### 4.3 Qualitative predictors

#### 4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with  $D=1$  for managerial

Model results with  $D^*=1$  for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

### Interaction

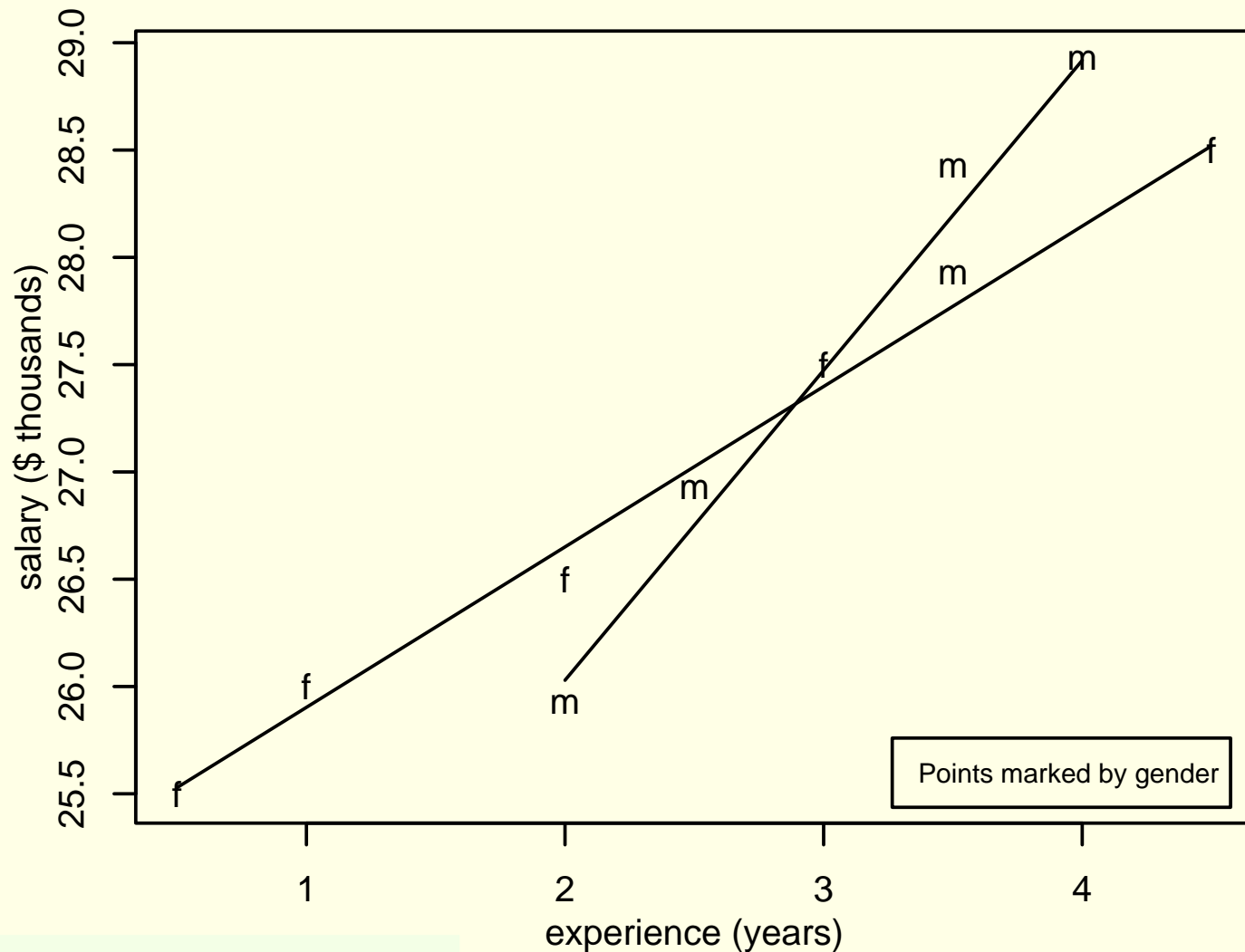
Salary/experience by gender scatterplot

Interaction results

#### 4.3.2 Qualitative predictors: >2 levels

# Salary/experience by gender scatterplot

Each line represents the association between salary and experience for females/males.



Do M/F salary differences depend on experience?

# Interaction results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.992 <sup>a</sup>	0.984	0.976	0.1872

<sup>a</sup> Predictors: (Intercept), experience, D, D\_exp.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	25.155	0.153	164.143	0.000
	experience	0.748	0.058	12.814	0.000
	D	-2.033	0.394	-5.159	0.002
	D_exp	0.697	0.128	5.444	0.002

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.155 + 0.748 \text{ experience} - 2.033D + 0.697D_{\text{exp}}$ .

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with D=1 for managerial

Model results with D\*=1 for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels



# Interaction results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.992 <sup>a</sup>	0.984	0.976	0.1872

<sup>a</sup> Predictors: (Intercept), experience, D, D\_exp.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	25.155	0.153	164.143	0.000
	experience	0.748	0.058	12.814	0.000
	D	-2.033	0.394	-5.159	0.002
	D_exp	0.697	0.128	5.444	0.002

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.155 + 0.748 \text{ experience} - 2.033D + 0.697D_{\text{exp}}$ .
- When  $D=0$  (female),  $\hat{Y} = 25.155 + 0.748 \text{ exp}$ .

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with D=1 for managerial

Model results with D\*=1 for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Interaction results

## Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.992 <sup>a</sup>	0.984	0.976	0.1872

<sup>a</sup> Predictors: (Intercept), experience, D, D\_exp.

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	25.155	0.153	164.143	0.000
	experience	0.748	0.058	12.814	0.000
	D	-2.033	0.394	-5.159	0.002
	D_exp	0.697	0.128	5.444	0.002

<sup>a</sup> Response variable: salary.

- $\hat{Y} = 25.155 + 0.748 \text{ experience} - 2.033D + 0.697D_{\text{exp}}$ .
- When  $D = 0$  (female),  $\hat{Y} = 25.155 + 0.748 \text{ exp}$ .
- When  $D = 1$  (male),  $\hat{Y} = ?$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

Qualitative predictors: 2 levels

Salary/experience by job scatterplot

Model results with D=1 for managerial

Model results with D\*=1 for clerical

Example you can do in your head

Salary/experience by gender scatterplot

Model results

Similar example requiring software

Salary/experience by gender scatterplot

Model results

Interaction

Salary/experience by gender scatterplot

Interaction results

4.3.2 Qualitative predictors: >2 levels

# Qualitative predictors: $>2$ levels

- Use two indicator variables to code information in a qualitative predictor with three levels (categories).
- One category takes the value 0 for both indicator variables (called the reference category):
  - doesn't matter which, but often the most common category or the most meaningful "baseline" category.
- Each of the other categories takes the value 1 for one of the indicator variables and 0 for the other.
- For example, define  $D_1$  and  $D_2$  to model difference between three states:

State	$D_1$	$D_2$
Oregon	0	0
California	1	0
Washington	0	1

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors:  $>2$  levels

Qualitative predictors:  $>2$  levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for  $D_2$  and  $D_1$  cost

Reduced model results

Qualitative predictors in practice

# Revenue/cost data

Revenue (\$m)	Cost (\$m)	State	$D_1$	$D_2$
25.5	0.10	CA	1	0
26.0	0.05	WA	0	1
25.9	0.20	CA	1	0
26.5	0.15	OR	0	0
27.3	0.30	OR	0	0
27.5	0.25	WA	0	1
27.9	0.40	OR	0	0
28.5	0.35	WA	0	1
28.4	0.50	OR	0	0
28.9	0.45	WA	0	1
26.6	0.35	CA	1	0
27.4	0.45	CA	1	0

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

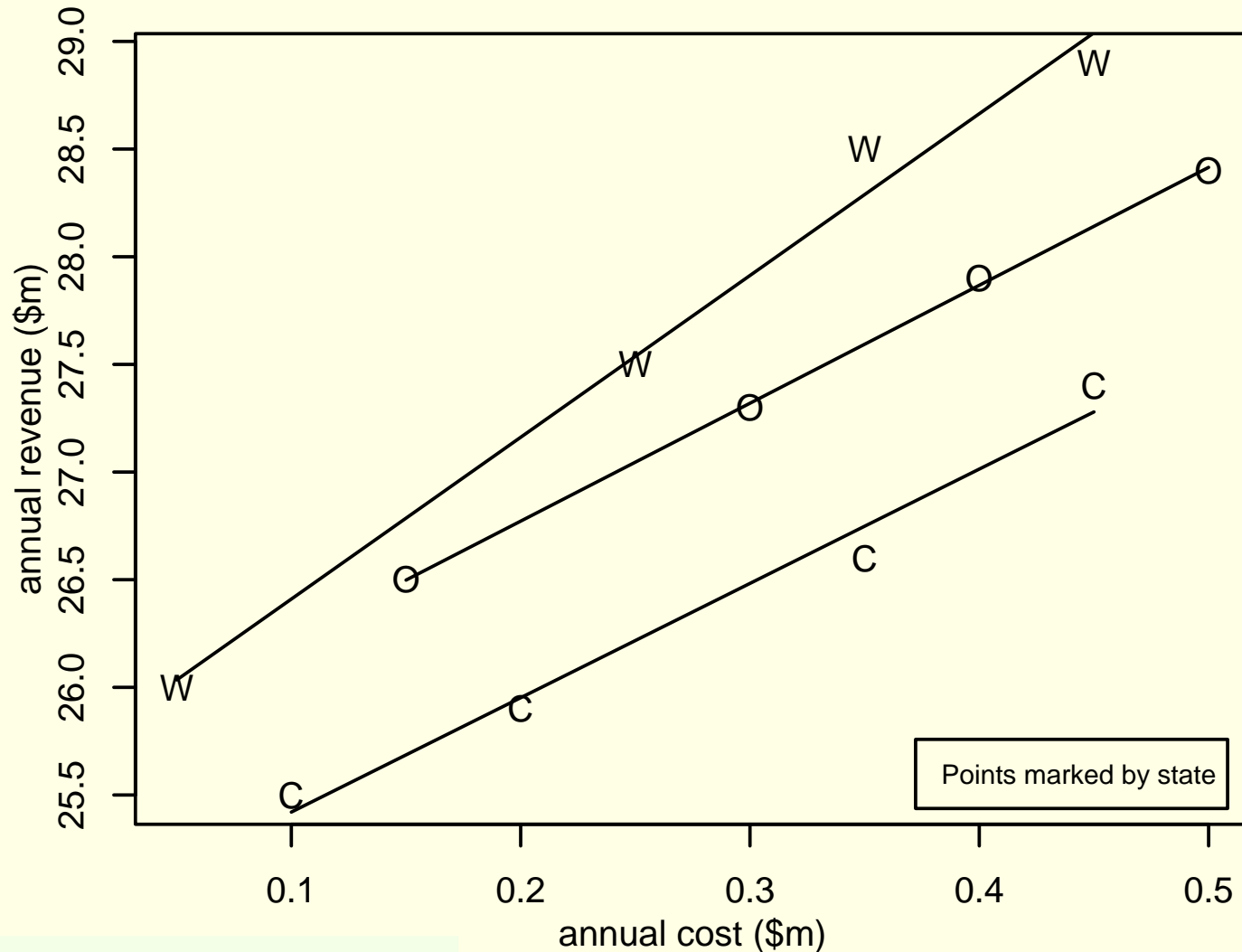
Nested F-test for  $D_2$  and  $D_1$  cost

Reduced model results

Qualitative predictors in practice

# Revenue/cost by state scatterplot

Each line represents the association between revenue and cost for different states.



Are revenue/cost associations different across states?

# Full interaction model results

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for  $D_2$  and  $D_1$  cost

Reduced model results

Qualitative predictors in practice

- $E(\text{revenue}) = b_0 + b_1 \text{cost} + b_2 D_1 + b_3 D_2 + b_4 D_1 \text{cost} + b_5 D_2 \text{cost}.$
- Allows for different intercepts and slopes for each state. How?

# Full interaction model results

- $E(\text{revenue}) = b_0 + b_1 \text{cost} + b_2 D_1 + b_3 D_2 + b_4 D_1 \text{cost} + b_5 D_2 \text{cost}.$
- Allows for different intercepts and slopes for each state. How?

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
1	(Intercept)	25.677	0.193	133.252	0.000
	cost	5.477	0.533	10.272	0.000
	D1	-0.787	0.248	-3.168	0.019
	D2	-0.018	0.241	-0.075	0.943
	D1cost	-0.166	0.739	-0.225	0.829
	D2cost	2.038	0.708	2.877	0.028

<sup>a</sup> Response variable: rev.

- Which intercepts/slopes could be the same?

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for D<sub>2</sub> and D<sub>1</sub> cost

Reduced model results

Qualitative predictors in practice

# Nested F-test for $D_2$ and $D_1$ cost

## Model Summary

Model	R Squared	Adjusted R Squared	Regression Std. Error	Change Statistics F-stat	df1	df2	Pr(>F)
2	0.991 <sup>a</sup>	0.988	0.120				
1	0.992 <sup>b</sup>	0.985	0.138	0.052	2	6	0.950

<sup>a</sup> Predictors: (Intercept), cost, D1, D2cost.

<sup>b</sup> Predictors: (Intercept), cost, D1, D2, D1cost, D2cost.

- There is a suggestion that adding  $D_2$  and  $D_1$  cost to the model causes overfitting. Why?

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels  
Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for  $D_2$  and  $D_1$  cost

Reduced model results

Qualitative predictors in practice



# Nested F-test for $D_2$ and $D_1$ cost

## Model Summary

Model	R Squared	Adjusted R Squared	Regression Std. Error	Change Statistics F-stat	df1	df2	Pr(>F)
2	0.991 <sup>a</sup>	0.988	0.120				
1	0.992 <sup>b</sup>	0.985	0.138	0.052	2	6	0.950

<sup>a</sup> Predictors: (Intercept), cost, D1, D2cost.

<sup>b</sup> Predictors: (Intercept), cost, D1, D2, D1cost, D2cost.

- There is a suggestion that adding  $D_2$  and  $D_1$  cost to the model causes overfitting. Why?
  - Adjusted  $R^2$  is higher for the reduced model.
  - The regression standard error,  $s$ , is lower for the reduced model.
  - The nested F-stat is not significant (high p-value), so the reduced model is favored.

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for  $D_2$  and  $D_1$  cost

Reduced model results

Qualitative predictors in practice

# Reduced model results

- $E(\text{revenue}) = b_0 + b_1 \text{cost} + b_2 D_1 + b_5 D_2 \text{cost}.$

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	25.683	0.089	288.819	0.000
	cost	5.433	0.255	21.334	0.000
	D1	-0.829	0.081	-10.219	0.000
	D2cost	2.004	0.253	7.921	0.000

<sup>a</sup> Response variable: rev.

- Plug-in  $D_1 = D_2 = 0$  for OR:  
$$\widehat{\text{rev}} = 25.683 + 5.433 \text{cost} - 0.829(0) + 2.004(0)\text{cost}$$
$$= 25.683 + 5.433 \text{cost}.$$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for  $D_2$  and  $D_1 \text{cost}$

Reduced model results

Qualitative predictors in practice

# Reduced model results

- $E(\text{revenue}) = b_0 + b_1 \text{cost} + b_2 D_1 + b_5 D_2 \text{cost}.$

## Parameters<sup>a</sup>

Model		Estimate	Std. Error	t-stat	Pr(>  t )
2	(Intercept)	25.683	0.089	288.819	0.000
	cost	5.433	0.255	21.334	0.000
	D1	-0.829	0.081	-10.219	0.000
	D2cost	2.004	0.253	7.921	0.000

<sup>a</sup> Response variable: rev.

- Plug-in  $D_1 = D_2 = 0$  for OR:  
$$\widehat{\text{rev}} = 25.683 + 5.433 \text{cost} - 0.829(0) + 2.004(0)\text{cost}$$
$$= 25.683 + 5.433 \text{cost}.$$
- Plug-in  $D_1 = 1, D_2 = 0$  for CA:  $\widehat{\text{rev}} = ?$
- Plug-in  $D_1 = 0, D_2 = 1$  for WA:  $\widehat{\text{rev}} = ?$

4.2 Interactions

4.3 Qualitative predictors

4.3.1 Qualitative predictors: 2 levels

4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for  $D_2$  and  $D_1 \text{cost}$

Reduced model results

Qualitative predictors in practice

# Qualitative predictors in practice

## 4.2 Interactions

## 4.3 Qualitative predictors

### 4.3.1 Qualitative predictors: 2 levels

### 4.3.2 Qualitative predictors: >2 levels

Qualitative predictors: >2 levels

Revenue/cost data

Revenue/cost by state scatterplot

Full interaction model results

Nested F-test for  $D_2$  and  $D_1$  cost

Reduced model results

Qualitative predictors in practice

- Can employ indicator variables in more general applications with qualitative predictors that have any number of levels/categories (within reason).
- For example, consider an application with one quantitative predictor ( $X$ ) and two qualitative predictors, one with 2 levels and the other with 4.
- If association between  $X$  and  $Y$  could differ for 2 levels of first qualitative predictor, then need to create an indicator variable— $D_1$ , say—and include it in model together with  $D_1X$  interaction.
- If association between  $X$  and  $Y$  could also differ for 4 levels of second qualitative predictor, then need to create three additional indicator variables— $D_2$ ,  $D_3$ , and  $D_4$ , say—and include them together with  $D_2X$ ,  $D_3X$ , and  $D_4X$  interactions.