# Chapter 7: Specific Issues in Regression Models

CZ

Fall, 2014

# One Sample Problem

Consider $y_1, \ldots, y_n$ as observations taken under uniform conditions from a stable process.

$$y_i = \beta_0 + \epsilon_i$$

In this case,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ .. \\ .. \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ . \\ . \\ \epsilon_n \end{pmatrix},$$

and $\boldsymbol{\beta} = \beta_0$. Question: what is $\hat{\beta}_0$?

# Two-Sample Problem

- Consider that the first $m$ observations are taken under one set of conditions, whereas the remaining $n - m$ are taken under a different set of conditions.

$$y_i = \begin{cases} \beta_1 + \epsilon_i & \text{if} \quad i = 1, \ldots, m \\ \beta_2 + \epsilon_i & \text{if} \quad i = m+1, \ldots, n \end{cases}$$

- This can be written as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where:

$$x_{i1} = \begin{cases} 1 & \text{if} \quad i = 1, \ldots, m \\ 0 & \text{if} \quad i = m+1, \ldots, n \end{cases} \qquad x_{i2} = \begin{cases} 0 & \text{if} \quad i = 1, \ldots, m \\ 1 & \text{if} \quad i = m+1, \ldots, n \end{cases}$$

Note: both $x_{i1}$ and $x_{i2}$ are indicator variables (dummy variables).

# Two-Sample Problem

In this case,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ .. \\ .. \\ y_n \end{pmatrix}, \mathbf{X} = \left(\begin{array}{cc} 1 & 0 \\ . & . \\ . & . \\ 1 & 0 \\ \hline 0 & 1 \\ . & . \\ . & . \\ 0 & 1 \end{array}\right), \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ . \\ . \\ \epsilon_n \end{pmatrix},$$

and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. Question: what is $\hat{\boldsymbol{\beta}}$?

# An Equivalent Formulation for Two-Sample Problem

Let $\delta = \beta_2 - \beta_1$, the previous model can be written as

$$y_i = \beta_1 + \delta x_{i2} + \epsilon_i,$$

where

$$x_{i2} = \begin{cases} 0 & \text{if} \quad \text{i=1,...,m} \\ 1 & \text{if} \quad \text{i=m+1,...,n} \end{cases}$$

The design matrix changes to:

$$\mathbf{X} = \left( \begin{array}{cc} 1 & 0 \\ . & . \\ . & . \\ 1 & 0 \\ \hline 1 & 1 \\ . & . \\ . & . \\ 1 & 1 \end{array} \right)$$

The vector of parameters is $\boldsymbol{\beta} = (\beta_1, \delta)'$. Question: what is $\hat{\boldsymbol{\beta}}$?

# K-Sample Problem

Suppose we have $K$ groups of observations. We assume

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, K; j = 1, \ldots, n_i$$

where $\mu_i = E(y_{ij})$ and $n = \sum_{i=1}^{K} n_i$.

# K-Sample Problem

For illustration purpose, let $K = 3$. We create two indicator variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th observation is in group 2} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th observation is in group 3} \\ 0 & \text{otherwise} \end{cases}$$

and we can rewrite the model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

Connection to mean: $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$ and $\beta_2 = \mu_3 - \mu_1$.

# Matrix Representation

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ \dots \\ y_{1n_1} \\ \hline y_{21} \\ \dots \\ y_{2n_2} \\ \hline y_{31} \\ \dots \\ y_{3n_3} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ . & . & . \\ 1 & 0 & 0 \\ \hline 1 & 1 & 0 \\ . & . & . \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ . & . & . \\ 1 & 0 & 1 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \dots \\ \epsilon_{1n_1} \\ \hline \epsilon_{21} \\ \dots \\ \epsilon_{2n_2} \\ \hline \epsilon_{31} \\ \dots \\ \epsilon_{3n_3} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

# Least Square Estimation

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \bar{y}_3 - \bar{y}_1 \end{pmatrix}.$$

where $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$. Consequently,

$$\widehat{\mathbf{Y}} = \begin{pmatrix} \hat{y}_{11} \\ \dots \\ \hat{y}_{1n_1} \\ \hline \hat{y}_{21} \\ \dots \\ \hat{y}_{2n_2} \\ \hline \hat{y}_{31} \\ \dots \\ \hat{y}_{3n_3} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ .. \\ \bar{y}_1 \\ \hline \bar{y}_2 \\ .. \\ \bar{y}_2 \\ \hline \bar{y}_3 \\ .. \\ \bar{y}_3 \end{pmatrix}.$$

# Hypothesis Testing

To test $H_0 : \mu_1 = \mu_2 = \mu_3$ is equivalent to testing

$$H_0 : \beta_1 = \beta_2 = 0.$$

The F statistic is:

$$F = \frac{SSR/2}{SSE/(n-3)} \sim F(2, n-3)$$

under $H_0$.

# General Form of Hypothesis Testing

For a $K$-Sample Problem, the F statistic is:

$$F = \frac{SSR/(K-1)}{SSE/(n-K)} \sim F(K-1, n-K)$$

under $H_0$. In addition,

- $SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$ where $\bar{y} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}$
- $SSE = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$
- $SSR = SST - SSE = \sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2$

# Example

Suppose we are concerned with the effects of $K$ catalysts on the yield of a chemical process. Consider $K = 4$ groups with equal number of observations ($n_1 = n_2 = n_3 = n_4 = 5$) in each group. The data are:

| Catalyst | Observations | | | | | $\bar{y}_i$ |
|---|---|---|---|---|---|---|
| 1 | 91.5 | 92.1 | 93.9 | 91.0 | 94.5 | 92.60 |
| 2 | 94.1 | 91.7 | 93.5 | 89.9 | 92.0 | 92.24 |
| 3 | 84.4 | 85.7 | 86.5 | 88.5 | 87.4 | 86.50 |
| 4 | 86.0 | 87.3 | 85.5 | 84.8 | 83.2 | 85.36 |

# Example

We fit the following regression model to the data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

where

- $x_{i1} = 1$ if $i$th observation is from the 2nd catalyst group and 0 if not.
- $x_{i2} = 1$ if $i$th observation is from the 3rd catalyst group and 0 if not.
- $x_{i3} = 1$ if $i$th observation is from the 4th catalyst group and 0 if not.

Question: What is $\widehat{\beta}$?

# Example

To test equality of groups means ($\beta_1 = \beta_2 = \beta_3 = 0$), we create the following ANOVA table:

| Source | DF | Sum of Squares | MS | F |
|---|---|---|---|---|
| Regression | 3 | $\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2 = 214.17$ | 71.39 | 29.12 |
| Error | 16 | $\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = 39.22$ | 2.45 | |
| Total | 19 | $\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = 253.40$ | | |

Since p-value$= P(F(3,16) > 29.12) < 0.05$, we reject $H_0$ and conclude at least one of the means is different from the others.

# Regression Model with Both Continuous Variables and Indicator Variables: Example 1

$$\begin{aligned} y &= \text{muscle mass} \\ x_1 &= \text{age} \\ x_2 &= \text{gender} \end{aligned}$$

We could code the gender variable as $x_2 = 1$ if female and $x_2 = 0$ if male. Consider the multiple regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

We can view the model above as two individual sub-models:

$$\begin{aligned} \text{Model for male } (x_{i2} = 0): \quad E(y_i) &= \beta_0 + \beta_1 x_{i1} \\ \text{Model for female } (x_{i2} = 1): \quad E(y_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 \end{aligned}$$

**What is the interpretation of $\beta_2$?**

$\beta_2$ is the difference of average muscle mass between females and males of the same age.

# Example 1

- These two models could be graphed as parallel lines (**why?**) in the plane with $x_1$ as the horizontal axis and $y$ as the vertical axis.

- What is the difference between fitting one joint model and two separate models?
- One reason for combining genders with an indicator variable in one model is that we usually get a smaller $MSE$, so that prediction and confidence intervals for $y$ or $E(y)$ are narrower.

# Example 2

$$\begin{aligned} y &= \text{sale price of home} \\ x_1 &= \text{square foot area of home} \\ x_2 &= \text{whether home has air conditioning or not} \end{aligned}$$

To put the air conditioning variable into a model, we use the indicator variable $x_2 = 1$ if the home has air conditioning and $x_2 = 0$ if the home does not have air conditioning. Thus, we have the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

# Example 2

- Interpretation of $\beta_2$:
  $\beta_2$ measures the difference in average sale prices of homes that have air conditioning versus homes that do not, given that the homes have the same square foot area.
- The model can be considered as two separate models:

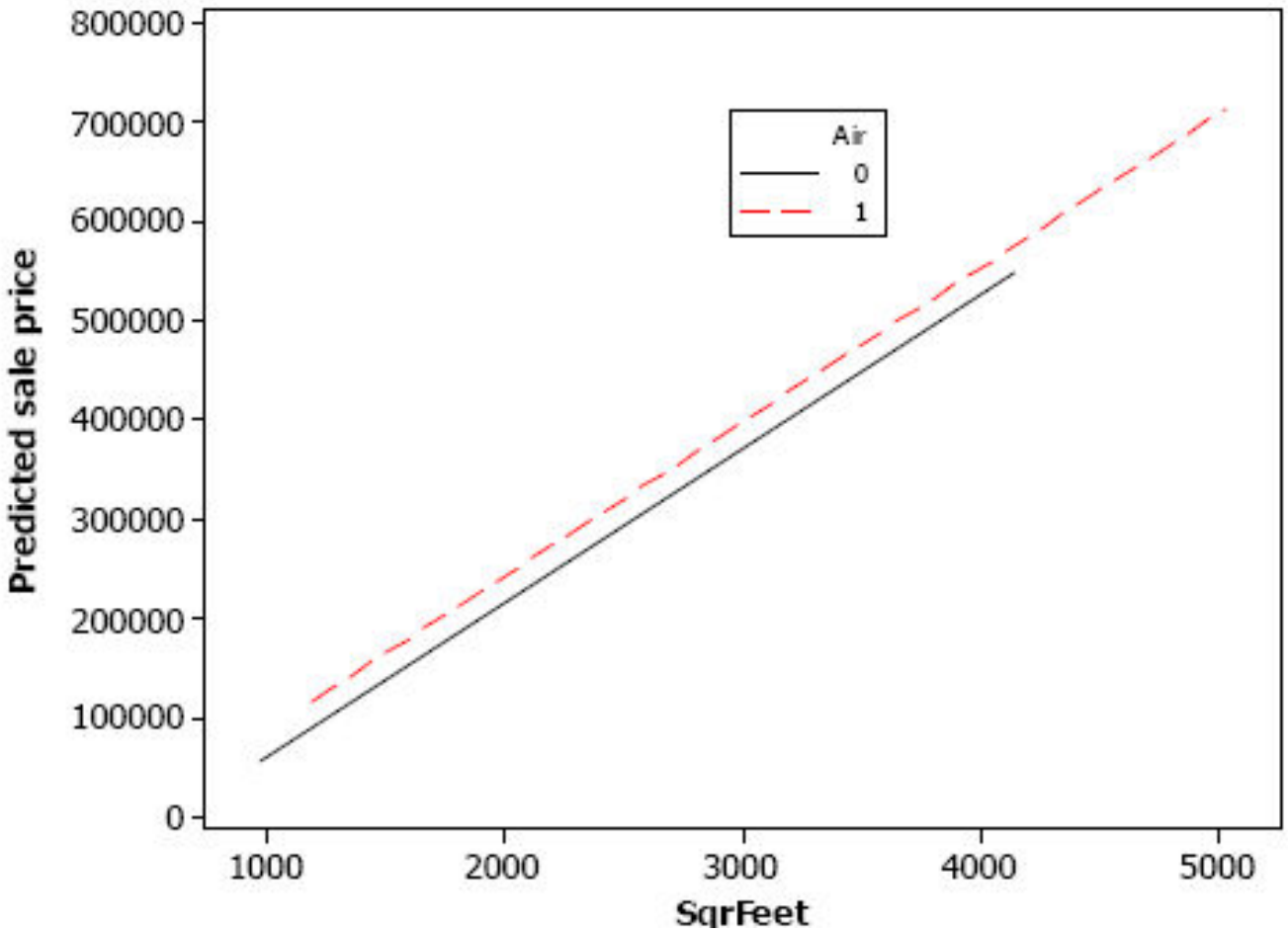
$$\begin{aligned} \text{Model for no air conditioning } E(y_i) &= \beta_0 + \beta_1 x_{i1} \\ \text{Model for having air conditioning } E(y_i) &= (\beta_0 + \beta_2) + \beta_1 x_{i1} \end{aligned}$$

- Again, we would have two parallel lines separated by distance $\beta_2$ for all $x_1$.

# Example 2

Fit the model:

```
> model=lm(SalePrice~SqrFeet+Air,data=mydata)
> summary(model)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -95886.952  12341.639  -7.769 4.26e-14 ***
SqrFeet       155.323      4.974  31.226  < 2e-16 ***
Air         26630.747   9440.052   2.821  0.00497 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 77770 on 518 degrees of
freedom
Multiple R-squared: 0.6818,  Adjusted R-squared: 0.6806
F-statistic:   555 on 2 and 518 DF, p-value: < 2.2e-16
```


800000
700000
600000
500000
400000
300000
200000
100000
0
Predicted sale price
1000
2000
3000
4000
5000
SqrFeet
Air
0
1

# Interaction Model

- In these models, the effects of the predictors on the response are not affected by other predictors. These models are called **additive model**. That is, there is NO **interaction effect** among predictors.
- The **interaction model** allows the effect of one predictor on the response to depend on the values of other predictors.

# Interaction Model: Back to Example 2

Suppose we suspect that the effect of air conditioning (yes or no) depends upon the size of the home. In other words, suppose that there is interaction between the two predictors. To put an interaction term in the model, we multiply the variables involved. The model here becomes

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon$$

# Interaction Model: Back to Example 2

The two models become:

- For homes with no air conditioning, $x_{i2} = 0$, so

$$E(y_i) = \beta_0 + \beta_1 x_{i1}$$

  intercept $= \beta_0$ and slope $= \beta_1$.

- For homes with air conditioning, $x_{i2} = 1$, so

$$E(y_i) = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_{i1}$$

  intercept $= \beta_0 + \beta_2$ and slope $= \beta_1 + \beta_3$.

- $\beta_2$ is the difference in intercepts between two models;
  $\beta_3$ is difference in slopes between two models.
  Two lines are not parallel any more.

# Interaction Model: Back to Example 2

**Note:** the interaction term may or may not be significant. To test it,

$$H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0.$$

Apply t-test. The interaction term may be dropped if we fail to reject $H_0$.

# Interaction Model: Back to Example 2

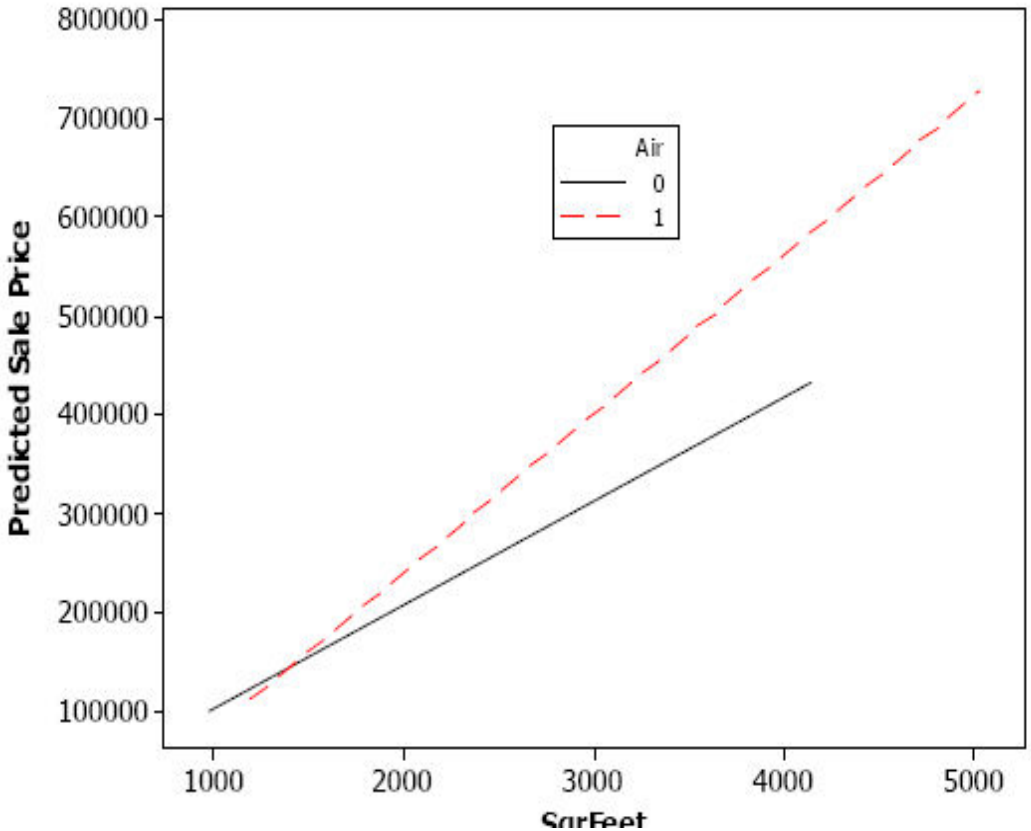
```
>model=lm(SalePrice~SqrFeet+Air+SqrFeet*Air,data=mydata
)
>summary(model)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3217.55   30085.04  -0.107 0.914871
SqrFeet        104.90      15.75   6.661 6.96e-11 ***
Air         -78867.83   32663.33  -2.415 0.016100 *
SqrFeet:Air     55.89      16.58   3.371 0.000805 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 77010 on 517 degrees of
freedom
Multiple R-squared: 0.6887,  Adjusted R-squared: 0.6869
F-statistic: 381.2 on 3 and 517 DF,  p-value: < 2.2e-16
```

# Interaction Model: Back to Example 2

- The interaction effect is significant, which indicates that the effect of SqrFeet on SalePrice depends on whether a home has air conditioning.
- Rule of hierarchy: If the interaction term is significant, we should keep all the first-order terms no matter whether they are significant or not, for the sake of interpretation.

# Interaction Model: Back to Example 2

The graph of the relationship between sale price and square foot area for homes with air conditioning and homes without air conditioning.



# Interaction Model: Back to Example 2

Consider the interaction model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon$$

We aim to test whether air conditioning has an effect at all. This means we are testing:

$$H_0 : \beta_2 = \beta_3 = 0 \quad H_1 : \beta_2 \neq 0 \quad \text{or } \beta_3 \neq 0.$$

Apply general linear F-test (additional sum of squares) to compare the full and restricted model.

# Interaction Model: Back to Example 2

```
> anova(full)
Analysis of Variance Table

Response: SalePrice
             Df     Sum Sq    Mean Sq   F value    Pr(>F)
SqrFeet       1 6.6664e+12 6.6664e+12 1124.1585 < 2.2e-16
Air           1 4.8138e+10 4.8138e+10    8.1175 0.0045589
SqrFeet:Air   1 6.7382e+10 6.7382e+10   11.3626 0.0008054
Residuals   517 3.0659e+12 5.9302e+09

> anova(restricted)
Analysis of Variance Table

Response: SalePrice
            Df     Sum Sq    Mean Sq F value    Pr(>F)
SqrFeet      1 6.6664e+12 6.6664e+12  1087.5 < 2.2e-16 ***
Residuals  519 3.1814e+12 6.1299e+09
```

# Interaction Model: Back to Example 2

- $SSE(\text{Restricted}) - SSE(\text{full}) = 1.155 \times 10^{11}$,
  $df_{SSE}(\text{Restricted}) - df_{SSE}(\text{Full}) = 2$, $MSE(\text{full}) = 5.9302 \times 10^{9}$.
- $F_0 = 1.155 \times 10^{11}/2/(5.9302 \times 10^{9}) = 9.74$.
- Since $F_0 > F_{0.05}(2, 517) = 3.01$, we reject $H_0$ and conclude that whether the house has air conditioning or not has a significant effect on sale price.

# Categorical Variable with More than 2 Levels: Example

- Data set "Senic": response is Infection risk in hospital. Predictors are Stay (average length of stay), and Region$= 1, 2, 3$ or 4, indicating which one of four regions of U.S. the hospital is in.

# Categorical Variable with More than 2 Levels: Example

- We do not want to use the original Region values 1, 2, 3, 4 in the model because this implies a numerical relationship among the regions
- Region is a categorical variable, so instead of using Region as a single quantitative predictor, we create 3 indicator variables to account for the 4 levels of Region.

# More than 2 Levels: Model Construction

The indicator variables describing the four regions are:

1. $I_1 = 1$ if hospital is in region 1, and 0 otherwise.
2. $I_2 = 1$ if hospital is in region 2, and 0 otherwise.
3. $I_3 = 1$ if hospital is in region 3, and 0 otherwise.
4. $I_1 = I_2 = I_3 = 0$ if hospital is in region 4, and not all 0 otherwise.

Here, Region 4 is the reference level (group).

- General Rule: When a categorical predictor variable has $k$ categories, we only need to use $k - 1$ indicator variables.
- For the overall fit of the model, it doesn't matter which set of $k - 1$ indicators we use.
- However, the estimated regression coefficients and t-tests depend on the choice of the reference group.

# More than 2 Levels: Model Construction

Fit the additive model:

$$E(y) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 \text{Stay}$$

1. For hospitals in Region 1: $E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_1$
2. For hospitals in Region 2: $E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_2$
3. For hospitals in Region 3: $E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_3$
4. For hospitals in Region 4: $E(y) = \beta_0 + \beta_4 \text{Stay}$

Each equation has the same slope $\beta_4$, but their intercepts differ. Therefore, **the four regression lines are parallel**.

# R

- In R, if we fit

  ```
  >model=lm(InfctRsk~factor(Region)+Stay,data=senic)
  ```

  R creates 3 indictor variables for "Region". By default, Region $= 1$ is the reference group.

- However, we need Region $= 4$ to be the reference group

  ```
  >Region_factor=relevel(factor(Region), "4")
  >model=lm(InfctRsk~Region_factor+Stay,data=senic)
  ```

# More than 2 Levels: R Output

R output:

```
lm(formula = InfctRsk ~ Region_factor + Stay,
data = senic)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.33906    0.70876  -0.478  0.63336
Region_factor1 -1.01158    0.39297  -2.574  0.01143
Region_factor2 -0.90069    0.35352  -2.548  0.01227
Region_factor3 -1.08114    0.33401  -3.237  0.00161
Stay            0.58177    0.08079   7.201 9.05e-11
```

# More than 2 Levels: Test Region Effect

- **But how do we address the question of whether Region is a significant predictor of InfctRsk?**
- Each t-test for the indicators tells part of the story, e.g., the $p$-value 0.011 is testing $H_0 : \beta_1 = 0$. This null hypothesis says that there is no difference of InfctRisk between regions 1 and 4, if the value of Stay is held constant. Similarly, the $p$-values for $I_2$ and $I_3$ correspond to comparisons between regions 2 and 4 and between 3 and 4, respectively.
- Therefore, to test the whole effect of Region, the null hypothesis should be $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

$\Rightarrow$ General Linear F-test.

# More than 2 Levels: Test Region Effect

General Linear F-test for the Region Effect:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

- Full model:

$$E(y) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 \text{Stay}$$

```
> anova(lm(InfctRsk~Region_factor+Stay,data=senic))
Analysis of Variance Table
Response: InfctRsk
               Df  Sum Sq Mean Sq F value    Pr(>F)
Region_factor   3  10.929   3.643  3.1363   0.02855
Stay            1  60.221  60.221 51.8479 9.051e-11
Residuals     106 123.119   1.161
```

# More than 2 Levels: Test Region Effect

General Linear F-test for the Region Effect:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

- Restricted model:

$$E(y) = \beta_0 + \beta_4 \text{Stay}$$

```
> anova(lm(InfctRsk~Stay,data=senic))
Analysis of Variance Table
Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  58.652  58.652  47.141 4.233e-10
Residuals 109 135.616   1.244
```

# More than 2 Levels: Test Region Effect

General Linear F-test for the Region Effect:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$\begin{aligned} F &= \frac{SSE(Restricted) - SSE(Full)}{[df_{SSE}(Restricted) - df_{SSE}(Full)] \cdot MSE(full)} \\ &= \frac{135.62 - 123.12}{3 \times 1.161} = 3.59 \end{aligned}$$

- Critical value approach: $F_{0.05}(3, 106) = 2.69 < 3.59$
- p-value approach: the p-value is $P(F(3, 106) > 3.59) = 0.016 < 0.05$.
- Therefore, we reject the null hypothesis at 0.05% level and conclude that there is some region effect on risk of infection.

# More than 2 Levels: Test Region Effect

What if we want to test whether Region 1 and 2 differ significantly?
Recall:

- For hospitals in region 1, $E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_1$.
- For hospitals in region 2, $E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_2$.
- For hospitals in region 3, $E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_3$.
- For hospitals in region 4, $E(y) = \beta_0 + \beta_4 \text{Stay}$.

Then the reasonable null hypothesis is $H_0 : \beta_1 = \beta_2$.

$\Rightarrow$ General Linear F-test.

# More than 2 Levels: With Interaction

- Until now we only focus on the additive model (no interaction terms). Maybe we need to allow unequal slopes among the 4 regions.
- To allow for interaction (or unequal slopes) in the model, we add the second-order terms $I_1 Stay$, $I_2 Stay$, and $I_3 Stay$ (Stay multiplied by each indicator) so that the population regression model is

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 \text{Stay} \\ &+ \beta_5 I_1 \text{Stay} + \beta_6 I_2 \text{Stay} + \beta_7 I_3 \text{Stay} \end{aligned}$$

# More than 2 Levels: With Interaction

This model includes four (perhaps nonparallel) lines for four regions:

- Region 1, $I_1 = 1, I_2 = 0$, and $I_3 = 0$

$$E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_1 + \beta_5 \text{Stay} = (\beta_0 + \beta_1) + (\beta_4 + \beta_5)\text{Stay}$$

- Region 2, $I_1 = 0, I_2 = 1$, and $I_3 = 0$

$$E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_2 + \beta_6 \text{Stay} = (\beta_0 + \beta_2) + (\beta_4 + \beta_6)\text{Stay}$$

- Region 3, $I_1 = 0, I_2 = 0$, and $I_3 = 1$

$$E(y) = \beta_0 + \beta_4 \text{Stay} + \beta_3 + \beta_7 \text{Stay} = (\beta_0 + \beta_3) + (\beta_4 + \beta_7)\text{Stay}$$

- Region 4, $I_1 = 0, I_2 = 0$, and $I_3 = 0$

$$E(y) = \beta_0 + \beta_4 \text{Stay}$$

# More than 2 Levels: With Interaction

- $\beta_1, \beta_2, \beta_3$ represent the differences of intercept between regions 1 and 4, 2 and 4, 3 and 4, respectively.
- $\beta_5, \beta_6, \beta_7$ represent the differences of slope between regions 1 and 4, 2 and 4, 3 and 4, respectively.
- Although these interpretations all involve comparison with region 4, we could make any region the "baseline" region by creating indicators for the other regions.

# More than 2 Levels: Test Interaction

Question: How to test the significance of interaction effect? In other words, whether or not the interaction between Region and Length of Stay significant?

$$H_0: \ \beta_5 = \beta_6 = \beta_7 = 0 \quad H_1: \ \text{at least one of them} \ \neq 0.$$

$$\Rightarrow \text{General Linear F-test.}$$

# More than 2 Levels: Test Interaction

General Linear F-test for the Interaction Effect:

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$

Full model: with all single-order terms and interactions.

```
> anova(lm(InfctRsk~Region_factor+Stay+
Region_factor*Stay,data=senic))
Analysis of Variance Table
Response: InfctRsk
                    Df  Sum Sq Mean Sq F value    Pr(>F)
Region_factor        3  10.929   3.643  3.2142   0.02601
Stay                 1  60.221  60.221 53.1341 6.559e-11
Region_factor:Stay   3   6.381   2.127  1.8765   0.13819
Residuals          103 116.738   1.133
```

# More than 2 Levels: Test Interaction

General Linear F-test for the Interaction Effect:

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$

Restricted model: with only single-order terms (Additive model).

```
> anova(lm(InfctRsk~Region_factor+Stay,data=senic))
Analysis of Variance Table
Response: InfctRsk
              Df  Sum Sq Mean Sq F value    Pr(>F)
Region_factor  3  10.929   3.643  3.1363   0.02855
Stay           1  60.221  60.221 51.8479 9.051e-11
Residuals    106 123.119   1.161
```

# More than 2 Levels: Test Interaction

General Linear F-test for the Region Effect:

$$H_0: \ \beta_5 = \beta_6 = \beta_7 = 0$$

$$\begin{aligned} F &= \frac{SSE(Restricted) - SSE(Full)}{[df_{SSE}(Restricted) - df_{SSE}(Full)] \cdot MSE(full)} \\ &= \frac{(123.119 - 116.738)}{3 \times 1.133} = 1.88 \end{aligned}$$

- Critical value approach: $F_{0.05}(3, 103) = 2.69 > 1.88$
- p-value approach: the p-value is $P(F(3, 103) > 1.88) = 0.138 < 0.05$.
- Conclusion: the interaction terms are not significant at 0.05% level, and the additive model can fit the data adequately.