# Week 9
# Multiple Linear Regression

# Residual Diagnostics

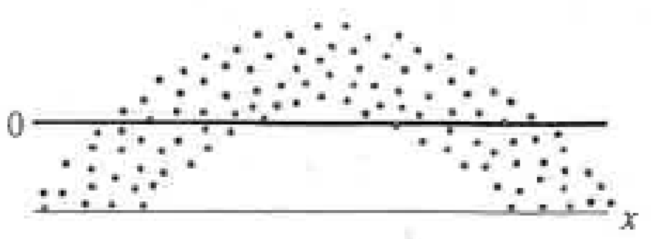



Figure 13.6: Nonlinear residual plot

- The assumption that the model is linear does not hold.  Note that the residuals are negative for low and high values of x and are positive for middle values of x.  The graph of these residuals is parabolic, not random.
- The residual plot does not have to be shaped in this manner for a nonlinear relationship to exist.  Any significant deviation from an approximately horizontal residual plot may mean that a nonlinear relationship exists between the two variables.

# Residual Diagnostics

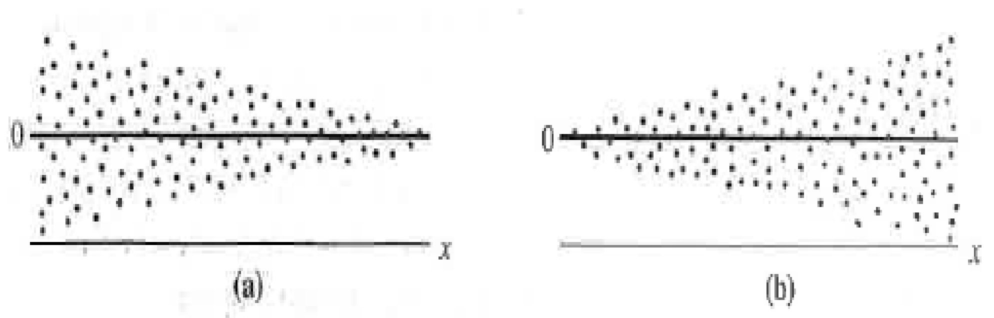



Figure 13.7: Nonconstant error variance

- The residual plots show a fan-shaped pattern, suggesting that the assumption of constant error variance (homoskedasticity) does not hold.
- Note in Figure 13.7(a) that the error variance is greater for small values of x and smaller for large values of x.  The situation is reversed in Figure 13.7 (b)

# Residual Diagnostics

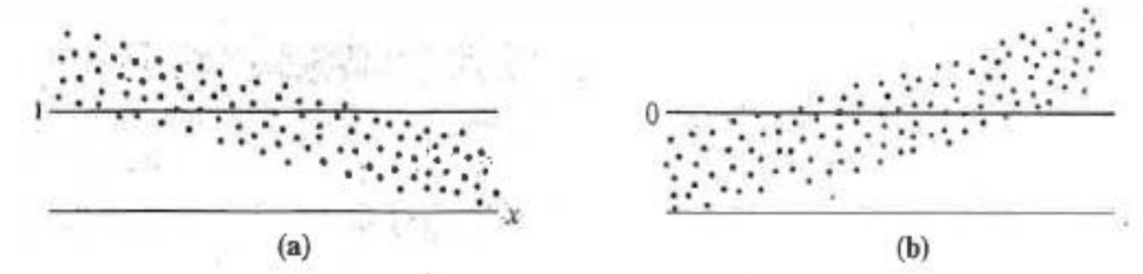



**Figure 13.8:** Graphs of non-independent error terms

- If the error terms are not independent (autocorrelation), the residual plots could look like one of the graphs in Figure 13.8.
- According to these graphs, instead of each error term being independent of the one next to it, the value of the residual is a function of the residual value next to it.
- For example, a large positive residual is next to a large positive residual and a small negative residual is next to a small negative residual.

# Residual Diagnostics

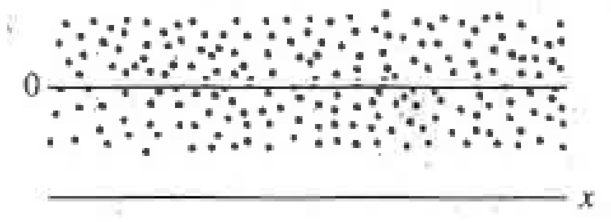



**Figure 13.9:** Healthy residual graph

The graph of the residuals from a regression analysis that meets the assumptions — a *healthy residual graph* — might look like the graph in figure 13.9. The plot has random scatter around the *x* axis; the variances of the errors are about equal for each value of *x*, and the error terms do not appear to be related to adjacent terms.

- The graph of the residuals from a regression analysis that meets the assumptions – a healthy residual graph – might look like the graph in Figure 13.9.
- The plot has random scatter around the x axis; the variances of the errors are about equal for each value of x, and the error terms do not appear to be related to adjacent terms.

# Multiple Regression Model

- General form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- *k* independent variables
- $x_1, x_2, \ldots, x_k$ may be functions of variables
  - e.g. $x_2 = (x_1)^2$

# First–Order Multiple Regression Model

Relationship between 1 dependent and 2 or more independent variables is a linear function

Population Y-intercept

Population slopes

Random error

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Dependent (response) variable

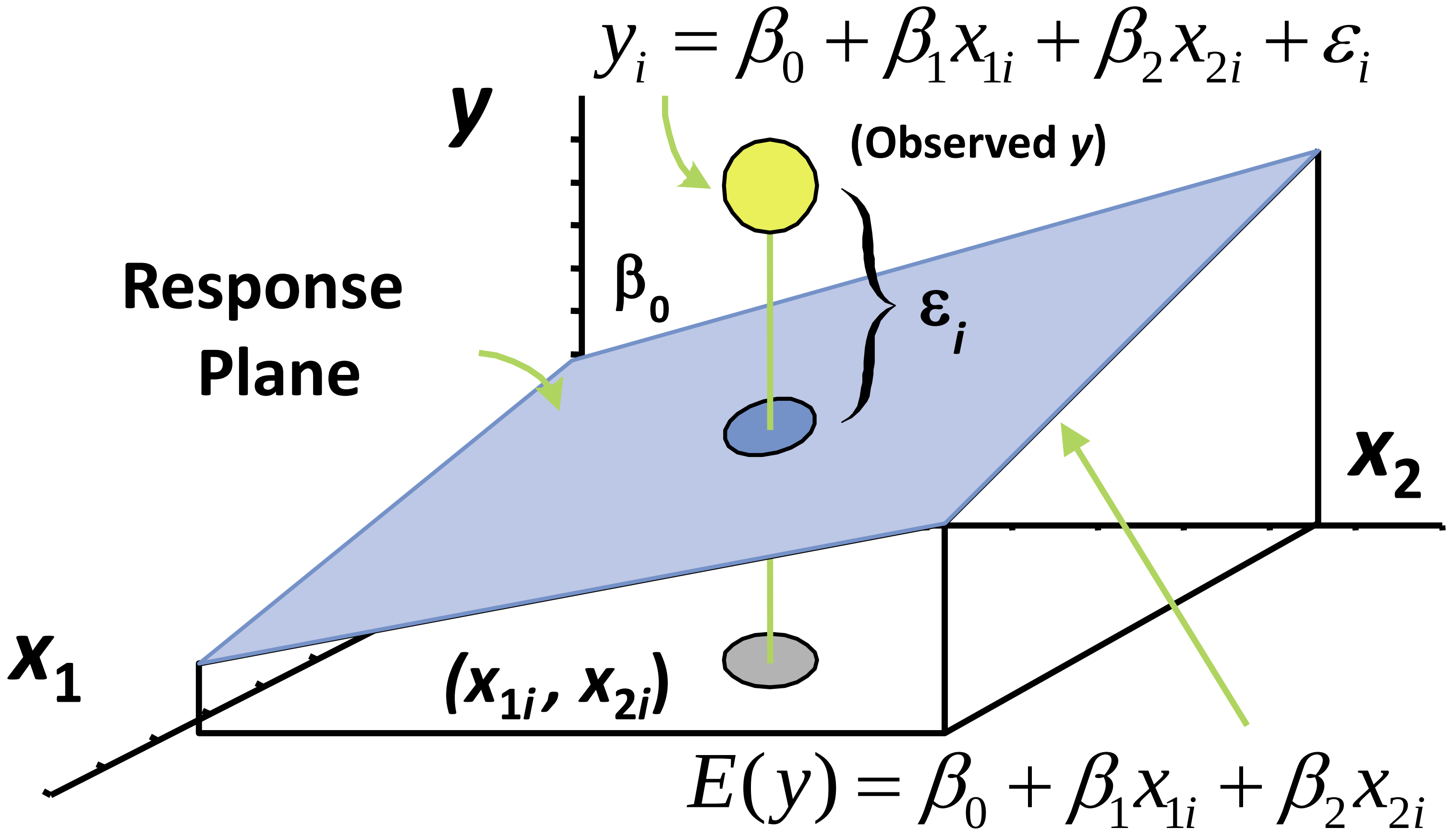
Independent (explanatory) variables

# First-order Model with 2 Independent Variables

- Relationship between 1 dependent and 2 independent variables is a linear function
- Model: $E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- Assumes no interaction between $X_1$ and $X_2$; i.e., the effect of $X_1$ on $E(Y|X_1, X_2)$ is the same regardless of $X_2$ values

# Population Multiple Regression Model

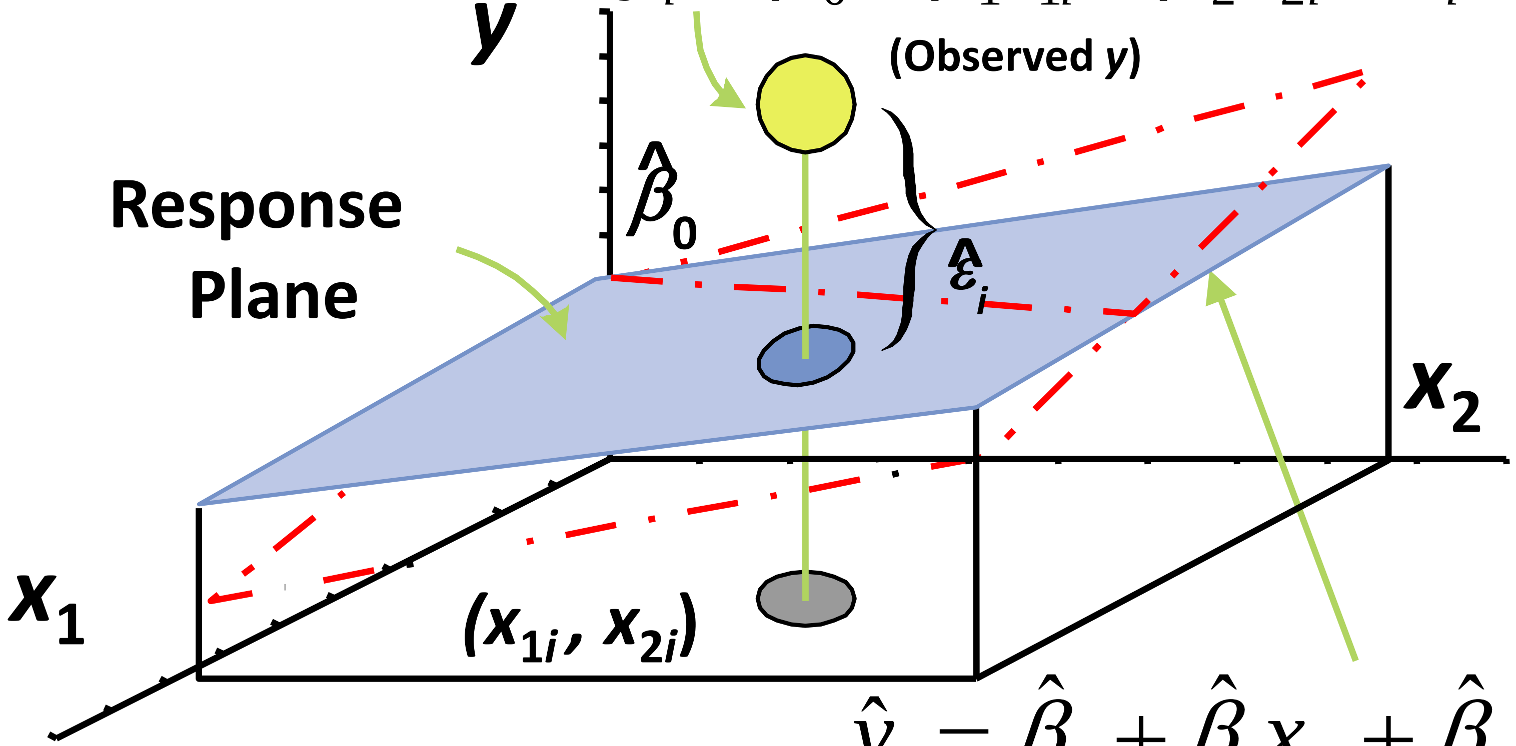
**Bivariate model:**



# Sample Multiple Regression Model

## Bivariate model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \varepsilon_i$$



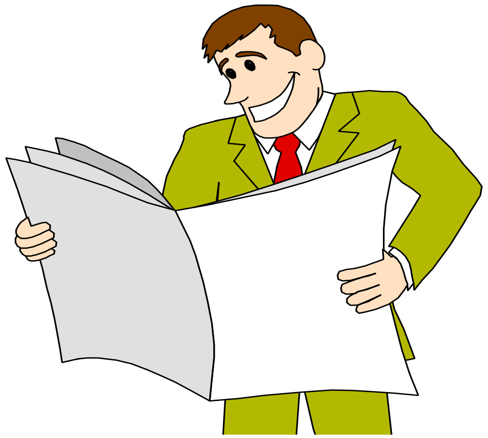

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

# Interpretation of Estimated Coefficients

- Y-intercept ($\hat{\beta}_0$)
  - Average value of Y when $X_k = 0$
- Slope ($\hat{\beta}_k$)
  - Estimated Y changes by $\hat{\beta}_k$ for each 1 unit increase in $X_k$ on average, holding all other independent variables constant.
  - If $\hat{\beta}_1 = 2$, then sales (Y) is expected to increase by 2 on average for each 1 unit increase in advertising ($X_1$) given the number of sales rep's ($X_2$).

# $1^{st}$ Order Model Example

You work in advertising for the New York Times. You want to find the effect of **ad size** (sq. in.) and newspaper **circulation** (000) on the number of **ad responses** (00). Estimate the unknown parameters.



You've collected the following data:

| $(y)$ Resp | $(x_1)$ Size | $(x_2)$ Circ |
|---|---|---|
| 1 | 1 | 2 |
| 4 | 8 | 8 |
| 1 | 3 | 1 |
| 3 | 5 | 7 |
| 2 | 6 | 4 |
| 4 | 10 | 6 |

# Parameter Estimation R Output

```
> y = c(1,4,1,3,2,4)
> x1 = c(1,8,3,5,6,10)
> x2 = c(2,8,1,7,4,6)
>
> reg1 <- lm(y ~ x1+x2)
> summary(reg1)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
       1        2        3        4        5        6 
 0.17012  0.05272  0.04077 -0.05202 -0.41547  0.20387 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.06397    0.25986   0.246   0.8214  
x1           0.20492    0.05882   3.484   0.0399 *
x2           0.28049    0.06860   4.089   0.0264 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2888 on 3 degrees of freedom
Multiple R-squared:  0.9737,    Adjusted R-squared:  0.9561 
F-statistic: 55.44 on 2 and 3 DF,  p-value: 0.004276
```

The fitted multiple regression is

$\hat{Y} = 0.06397 + 0.20492X_1 + 0.28049X_2$

# Interpretation of Coefficients Solution

- The fitted multiple regression is

  $\hat{Y} = 0.06397 + 0.20492X_1 + 0.28049X_2$

- $\hat{\beta}_1 = 0.20492$

  Number of responses to ad is expected to increase by 0.2049* 100 = 20.49, on average, for each 1 sq. in. increase in ad size, holding circulation constant.

- $\hat{\beta}_2 = 0.28049$

  Number of responses to ad is expected to increase by 0.2805* 100 = 28.05, on average, for each 1000 unit increase in circulation, holding ad size constant

# Estimation of $\sigma^2$

For a model with *k* independent variables,

$$\hat{\sigma}^2 = \frac{\text{RSS}}{\text{n}-\text{p}} = \text{MSE where RSS} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$\hat{\sigma} = \text{SER} = \sqrt{\text{MSE}} = \sqrt{\frac{\text{RSS}}{\text{n}-\text{p}}}$$

In the previous example, RSS = 0.2503, $\hat{\sigma}^2$= MSE = 0.0834, $\hat{\sigma}$ = SER = 0.2888

```
> anova(reg1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x1         1 7.8551  7.8551  94.162 0.002324 **
x2         1 1.3946  1.3946  16.718 0.026442 *
Residuals  3 0.2503  0.0834
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Testing Overall Significance

- Shows if there is a linear relationship between **all** X variables **together** and Y
- Hypotheses

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

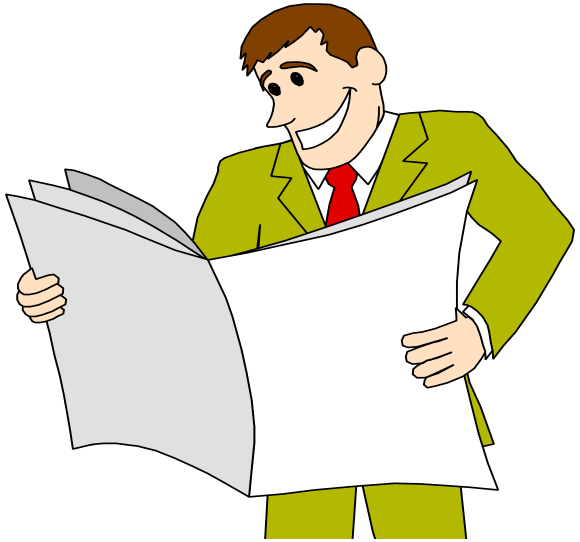
  $H_1$: At least one $\beta_j \neq 0$
- Test Statistic

  $F_{stat} = \frac{\text{RegMS}}{\text{MSE}} \sim F_{p-1,n-p}$ under $H_0$

# Testing Overall Significance Example

You work in advertising for the New York Times. You want to find the effect of **ad size** (sq. in.), $x_1$, and newspaper **circulation** (000), $x_2$, on the number of **ad responses** (00), $y$.

Conduct the overall F–test of model usefulness. Use $\alpha = 0.05$.



# Testing Overall Significance Example

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

$H_1$: At least one $\beta_j \neq 0$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06397    0.25986   0.246   0.8214
x1           0.20492    0.05882   3.484   0.0399 *
x2           0.28049    0.06860   4.089   0.0264 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2888 on 3 degrees of freedom
Multiple R-squared:  0.9737,    Adjusted R-squared:  0.9561
F-statistic: 55.44 on 2 and 3 DF,  p-value: 0.004276
```

Decision rule based on p-value: reject $H_0$ if p-value < $\alpha$

Decision: reject $H_0$ because p-value = 0.004276 < 0.05

Conclusion: There is sufficient evidence to show that the model is useful. In other words, at least one of the independent variables is contributing significant information for the prediction of number of ad. responses

# $R^2$ in Multiple Regression

$S_y^2 = \frac{\sum(y - \bar{y})^2}{n-1} = \frac{TSS}{n-1}$. Therefore, TSS = $(n - 1)S_y^2$

TSS = RegSS + RSS

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

- When another regressor is added to the regression,
  - TSS remains unchanged; it is a pure feature of the data
  - RSS cannot increase: the larger model cannot fit the data worse
  - Note that RSS remains unchanged if the additional regressor has a coefficient of zero (i.e., contributes nothing to the model)

- As a result, $R^2$ cannot decrease as we include more regressors, even if the extra regressors are irrelevant !

# Adjusted $R^2$

- Adjusted $R^2$ is an improved measure over $R^2$
- It adjusts for the number of regressors, k, in the model

  Adjusted $R^2 = 1 - \frac{\frac{RSS}{n-p}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-p}\left(\frac{RSS}{TSS}\right) = 1 - \frac{n-1}{n-p}(1 - R^2)$

  $\lim_{n\to\infty}$ Adjusted $R^2 = \lim_{n\to\infty}\left[1 - \frac{\frac{n-1}{n}}{\frac{n-p}{n}}(1 - R^2)\right] = \lim_{n\to\infty}\left[1 - \frac{1-\frac{1}{n}}{1-\frac{p}{n}}(1 - R^2)\right] = R^2$

  Adjusted $R^2 \leq R^2$
- When another regressor is added to the regression,
  - TSS remains unchanged; RSS cannot increase
  - n – p decreases by 1
- Adjusted $R^2$ may increase or decrease
  - If the additional regressor does not provide much explanatory power to the model, RSS will change little. Adjusted $R^2$ will decrease!
  - Adjusted $R^2$ increases if the additional regressor is important in explaining Y.
  - $R^2$ never decreases when an additional regressor is included.

# Categorical Variable

- A *categorical variable* is a variable that can take on a fixed number of possible values
  - Example 1: X has 2 possible values – 0 or 1
  - Example 2: X has 3 possible values – 1, 2, or 3
  - Example 3: X has 3 possible values – “A”, “B”, or “C”
- A categorical variable assigns one value to each category in data.
  - Example 1: Season. X = 1 if winter, X = 2 if spring, X = 3 if summer, X = 4 if autumn
  - Example 2: Education. X = 1 if high school dropout, X = 2 if high school graduate, X = 3 if others
  - Example 3: Grade. X = HD if score ≥ 90, X = P if 50 ≤ score < 90, X = F if score < 50.
- A categorical variable must be transformed into dummy variables before regression can be used.
  - Choose a base category.
  - Define a separate dummy variable for each category other than the base category.
  - Perfect multicollinearity arises if a dummy variable for the base category is also defined.

# Example

A researcher is interested in estimating the effect of geographical location on house price (Y, expressed in thousand dollars). He has access to a house location variable (X), which is defined as follows: 1 if *East*, 2 if *South*, 3 if *West*, 4 if *North*. The first five observations of the data set is given below:

| Observation | Y | X | East dummy | South dummy | West dummy | North dummy |
|---|---|---|---|---|---|---|
| 1 | 787 | 1 | 1 | 0 | 0 | 0 |
| 2 | 274 | 1 | 1 | 0 | 0 | 0 |
| 3 | 115 | 2 | 0 | 1 | 0 | 0 |
| 4 | 313 | 3 | 0 | 0 | 1 | 0 |
| 5 | 611 | 4 | 0 | 0 | 0 | 1 |

a. What type of variable is X called?
b. Using *East* as the base category, write down an appropriate multiple regression model for the researcher. Briefly explain how the coefficients in the model should be interpreted.

# Example – cont'd

a. X is called a categorical variable.

b. The model is $Y = \beta_0 + \beta_1 South + \beta_2 West + \beta_3 North + \varepsilon$

$\beta_0$: average house price in the East (or the base group).

$\beta_1$: differential house price between South and East.

$\beta_2$: differential house price between West and East.

$\beta_3$: differential house price between North and East.

# Perfect Multicollinearity among Dummy Variables

| Location | $D_{East}$ | $D_{South}$ | $D_{West}$ | $D_{North}$ |
|---|---|---|---|---|
| East | 1 | 0 | 0 | 0 |
| South | 0 | 1 | 0 | 0 |
| West | 0 | 0 | 1 | 0 |
| North | 0 | 0 | 0 | 1 |

- Perfect multicollinearity arises if a dummy variable for the base category, $D_{East}$, is also defined.
- It is easy to verify that $D_{East} = 1 – D_{south} – D_{West} – D_{north}$

# Dummy Variables in Multiple Regression

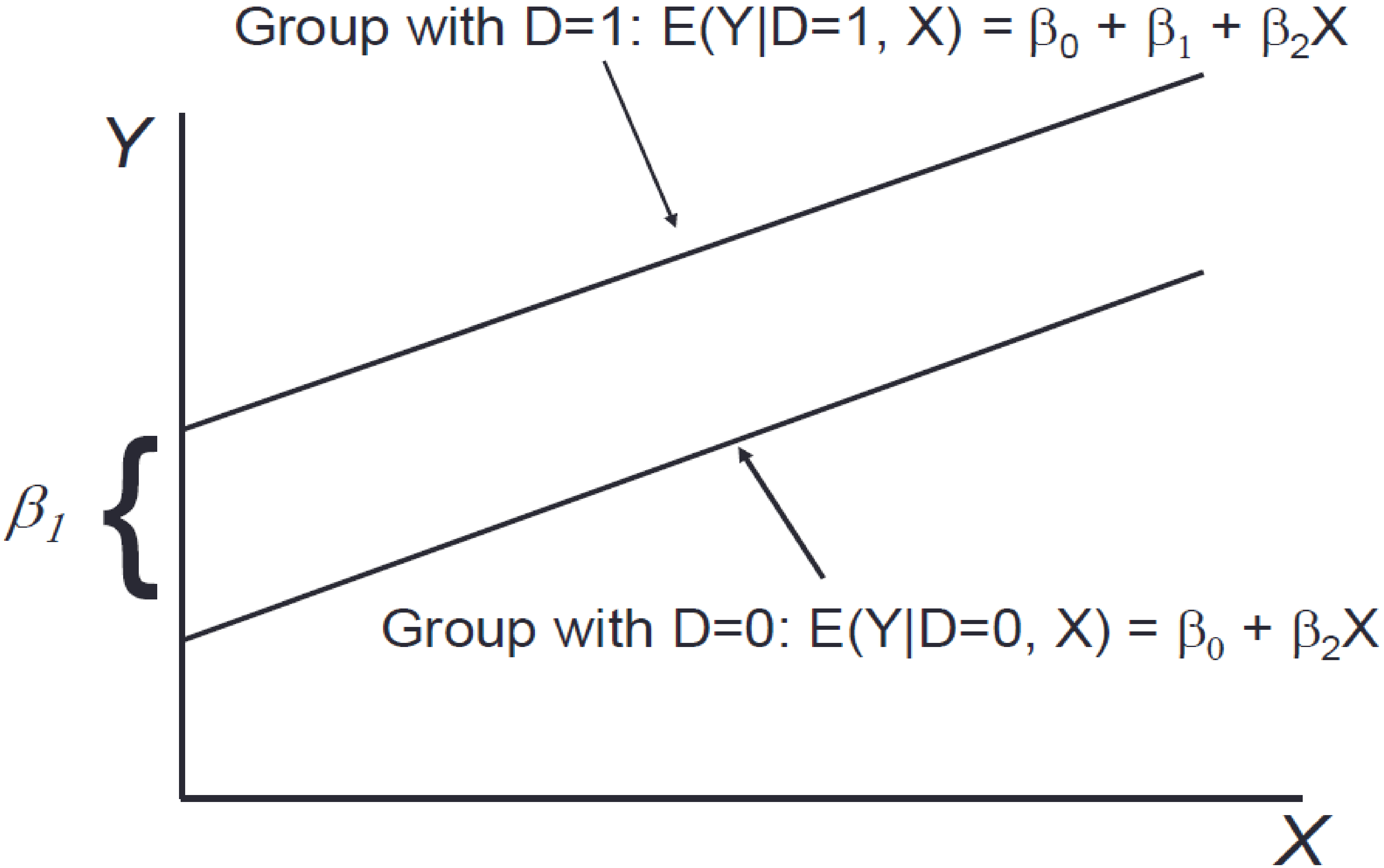
- Consider a model with one continuous regressor (X) and one dummy regressor (D)

  $Y = \beta_0 + \beta_1 D + \beta_2 X + \varepsilon$

- $\beta_1$ is the effect of D on Y, keeping X constant
- $\beta_1$ also represents an *intercept shift* of the population regression line (PRL) between the two groups
  - For the group with D=1, $E(Y|D=1, X) = (\beta_0 + \beta_1) + \beta_2 X$
  - For the group with D=0, $E(Y|D=0, X) = \beta_0 + \beta_2 X$

# Graphical Depiction of the Model

$Y = \beta_0 + \beta_1 D + \beta_2 X + u$



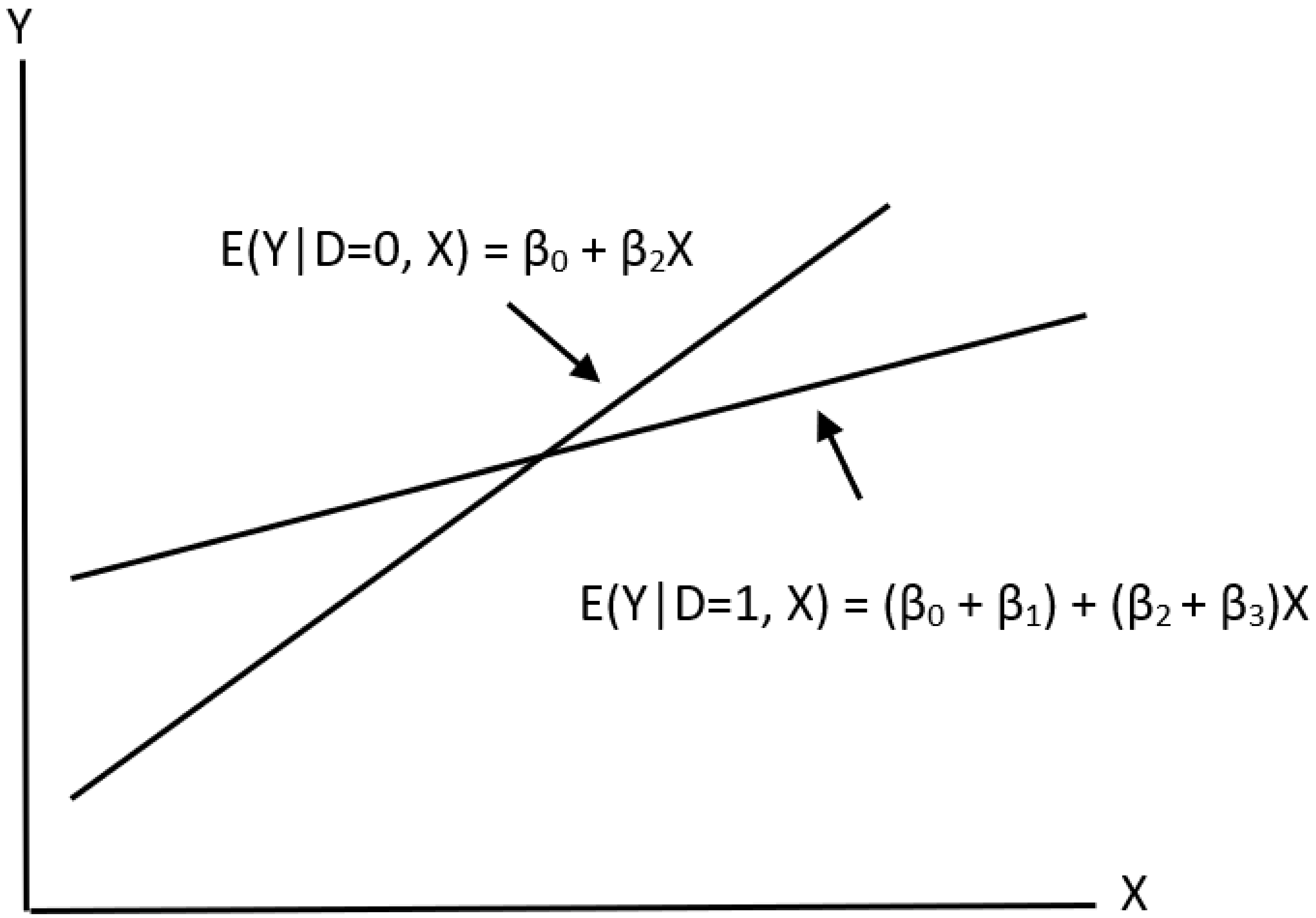
# Dummy Variables in Multiple Regression

- Consider a model interacting a dummy variable, D, with a continuous variable , X.

  $Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 D{*}X + \varepsilon$

- For the group with D = 1, $E(Y|D=1, X) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X$
- For the group with D = 0, $E(Y|D=0, X) = \beta_0 + \beta_2 X$

# Dummy Variables in Multiple Regression



# Relationship between pooled-variance t test and OLS:

The daily catch of 2 fishing boats was recorded on a random basis. The results for 2 independent random samples are given in the accompanying table.

| **Boat 1** | **120** | **136** | **107** | **109** | **129** | **117** | **125** | **110** | **124** |
|---|---|---|---|---|---|---|---|---|---|
| Boat 2 | 131 | 144 | 116 | 111 | 103 | 122 | 141 | 139 | 130 |
| | 133 | 132 | 135 | 148 | | | | | |

Is there a statistically significant difference in mean daily catch between the two fishing boats?

## Pooled variance t test
$(\sigma_1^2 = \sigma_2^2 = \sigma^2)$

$H_0$: $\mu_1 - \mu_2 = 0$
$H_1$: $\mu_1 - \mu_2 \neq 0$

```
> boat1=c(120,136,107,109,129,117,125,110,124)
> boat2=c(131,144,116,111,103,122,141,139,130,133,132,135,148)
> boat=c(rep(1,9),rep(0,13))
> boat
 [1] 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
>
> pooled_var <- function(x, y, integer = FALSE) {
+   n1 <- length(x)
+   n2 <- length(y)
+   return(((n1 - 1) * var(x) + (n2 - 1) * var(y)) / (n1 + n2 - 2))
+ }
>
> pooled_var(boat1, boat2, integer = FALSE)
[1] 144.2538
>
> t.test(boat1, boat2, alternative = "two.sided", var.equal = TRUE)

        Two Sample t-test

data:  boat1 and boat2
t = -1.9102, df = 20, p-value = 0.07055
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.8126941   0.9152582
sample estimates:
mean of x mean of y
 119.6667  129.6154
```

However, the separate variance t test (non-equal variances) is NOT equivalent to OLS with het-robust (OR constant) variance.

## Simple linear regression with a binary predictor

Daily catch = $\beta_0 + \beta_1$Boat + $\varepsilon$

where Boat = 1 if Boat 1; 0 otherwise

```
> catch=c(120,136,107,109,129,117,125,110,124,131,144,116,111,103,122,141,139,130,133
,132,135,148)
>
> reg2 <- lm(catch ~ boat)
> summary(reg2)

Call:
lm(formula = catch ~ boat)

Residuals:
    Min      1Q  Median      3Q     Max
-26.615  -9.154   1.885   8.346  18.385

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  129.615      3.331   38.91   <2e-16 ***
boat          -9.949      5.208   -1.91   0.0705 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.01 on 20 degrees of freedom
Multiple R-squared:  0.1543,    Adjusted R-squared:  0.112
F-statistic: 3.649 on 1 and 20 DF,  p-value: 0.07055
```

```
> anova(reg2)
Analysis of Variance Table

Response: catch
          Df  Sum Sq Mean Sq F value  Pr(>F)
boat       1  526.38  526.38   3.649 0.07055
Residuals 20 2885.08  144.25
```

# Multicollinearity

- $X_2 = 2X_1$ is a linear combination of $X_1$.
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 (2X_1) + \varepsilon$
- Regression cannot be run because of perfect multicollinearity. In this setup, there is only one source of data variation in the independent variable, but there are 2 slope parameters to estimate.
- $\mathrm{Var}(\hat{\beta}_j) = \frac{\mathrm{SER}^2}{(n-1)S_{xj}^2}\mathrm{VIF}_j$ where $\mathrm{VIF}_j = \frac{1}{1-R_j^2}$ is the variance inflation factor
- $\mathrm{Var}(\hat{\beta}_j)$ is inflated by a factor of $\mathrm{VIF}_j$ if $X_j$ is correlated with the other independent variables.
- Rule of thumb:
  - VIF = 1, there is no multicollinearity among independent variables
  - VIF > 1, the independent variables may be moderately correlated.
  - $5 \le \mathrm{VIF} \le 10$, it indicates high correlation that may be problematic.
  - If $R_j^2 = 0.8$, $\mathrm{VIF}_j = \frac{1}{1-0.8} = 5$ and if $R_j^2 = 0.9$, $\mathrm{VIF}_j = \frac{1}{1-0.9} = 10$,
  - If $\mathrm{VIF}_j > 10$, it definitely raises the concern of multicollinearity.