Dummy Variables for Multiple Categories

- A categorical variable can have more than two categories.
- Use multiple dummy variables to capture all categories, one for each category.
- Given the intercept term, we exclude one of the dummy variables from the regression.
  - The excluded variable represents the reference category.
  - Including all dummy variables creates a dummy variables trap (perfect multicollinearity; more later).
- Example: mode of transportation with three categories
  - Public transportation, driving alone, or car pooling
  - Use any two dummy variables; for example:
    - $d_1 = 1$ for public transportation and 0 otherwise
    - $d_2 = 1$ for driving alone and 0 otherwise
    - $d_1 = d_2 = 0$ indicates car pooling

Dummy Variables for Multiple Categories

- Data: ***Retail***

Year	Quarter	Sales	GNP
2007	1	921266	14301.854
2007	2	1013371	14512.945
⋮	⋮	⋮	⋮
2016	4	1299699	19134.463

a. Estimate a linear regression model with Sales as the dependent variable and GNP along with the relevant dummy variables, to capture seasonal variations, as explanatory variables.

b. Interpret the slope coefficient for quarter 1.

c. What are the predicted sales in quarter 2 if GNP is $18,000 (in billions)?

Dummy Variables for Multiple Categories

- We will first use Excel and then R.
- Given four quarters, we need to create only three dummy variables, using the omitted quarter as reference. All interpretations of the results are made in relation to the reference quarter.
- For illustration, we will use $d_1 = 1$ for Quarter 1 and 0 otherwise. We similarly define d_2 and d_3, using Quarter 4 as the reference quarter.
- In Excel, open the *Retail* data file. Enter "d1" in a blank column and enter the formula =IF(B2=1,1,0). Copy and paste this formula to the rest of the column. Repeat the analysis for quarter 2 and 3. A portion of the relevant data is shown below.

C	D	E	F	G
Sales	GNP	d1	d2	d3
921266	14301.854	1	0	0
1013371	14512.945	0	1	0
1000151	14717.814	0	0	1
1060394	14880.255	0	0	0
950268	14848.718	1	0	0
1028016	14997.477	0	1	0
999824	15045.188	0	0	1
957207	14671.011	0	0	0

Dummy Variables for Multiple Categories

a. $y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_3 + \varepsilon$, where y and x are sales and GNP, d_1 is a quarter 1 dummy, d_2 is a quarter 2 dummy, and d_3 is a quarter 3 dummy. (Reference Quarter: 4th)

	Coefficients	Standard Error	*t* Stat	*p*-value
Intercept	47095.6859	53963.3350	0.873	0.3888
GNP	65.0548	3.2151	20.234	6.74E-21
d1	−108765.2580	13638.1967	−7.975	2.21E-09
d2	−30486.2947	13593.5983	−2.243	0.0314
d3	−48805 0461	13570.2660	−3.596	0.0009

b. All else equal, retail sales in quarter 1 are expected to be approximately $108,765 million less than sales in quarter 4. Other dummy variables are interpreted similarly.

c. With $\text{GNP} = 18{,}000$, $d_2 = 1$, $d_3 = 0$, and $d_4 = 0$, $\hat{y} = 47095.6895 + 65.0548(18000) - 30486.2947 = 1187595$. Retail sales are predicted to be approximately $1,187,595 (in millions) in the second quarter when the GNP is $18,000 (in billions).

Dummy Variables for Multiple Categories

We will now replicate the results with R.

- # Import Retail data
- myData$d1 <- ifelse(myData$Quarter == 1, 1, 0)
- # Create d2, d3, d4 similarly

- # Using Quarter 4 as reference
- Model1 <- lm(Sales ~ GNP + d1 + d2 + d3, data = myData)
- summary(Model1)
- predict(Model1, data.frame(GNP=18000,d1=0,d2=1,d3=0))

Dummy Variables for Multiple Categories

Practice Problem: Reconsider ***Retail*** data to estimate a linear regression model with Sales as the dependent variable and GNP along with the relevant dummy variables, to capture seasonal variations, as explanatory variables.

a. Re-specify and interpret the model as $y = \beta_0 + \beta_1 x + \beta_2 d_2 + \beta_3 d_3 + \beta_4 d_4 + \varepsilon$ (Reference Quarter: 1st)

b. Interpret the slope coefficient for quarter 4.

c. What are the predicted sales in quarter 2 if GNP is $18,000 (in billions)?

d. Does using a different quarter for reference impact the inference?

Model Selection

Example: Recall the introductory case and consider three models. Which should we choose?

$$\text{Model 1: Earnings} = \beta_0 + \beta_1 \text{Cost} + \varepsilon$$
$$\text{Model 2: Earnings} = \beta_0 + \beta_1 \text{Cost} + \beta_2 \text{Grad} + \beta_3 \text{Debt} + \varepsilon$$
$$\text{Model 3: Earnings} = \beta_0 + \beta_1 \text{Cost} + \beta_2 \text{Grad} + \beta_3 \text{Debt} + \beta_4 \text{City} + \varepsilon$$

Several "goodness-of-fit" measures summarize how well the sample regression equation fits the data.

- The standard error of the estimate, s_e
- The coefficient of determination, R^2
- The adjusted coefficient of determination, adjusted R^2

Model Selection

- Recall that a residual $e_i = y_i - \hat{y}_i$.

- The sample variance, s_e^2, is the average of the squared residuals. The standard deviation (standard error of the estimate) is computed as

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

  - $SSE = \sum e_i^2$ is the error sum of squares
  - k denotes the number of explanatory variables
  - n is the sample size

- For a given sample size, increasing the number of explanatory variables reduces the numerator and denominator.

  - The net effect allows us to determine if the added explanatory variables improve the fit.
  - When comparing models, the model with the smaller s_e is preferred.

Model Selection

- R^2 denotes the sample variation in the dependent variable that is explained by the estimated regression equation.

- It is the ratio of the explained variation of the dependent variable to its total variation.

  - $SST = SSR + SSE$ (total variation = explained + unexplained variation)
  - The total variation in y is the total sum of squares, $SST = \sum(y_i - \bar{y})^2$
  - $SSR = \sum(\hat{y}_i - \bar{y})^2, SSE = \sum(y_i - \hat{y}_i)^2$

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

  - Falls between 0 and 1
  - The closer to 1, the better the fit

- Alternatively, $R^2 = \left(r_{y,\hat{y}}\right)^2$, where $r_{y,\hat{y}}$ is the sample correlation coefficient between $y \; and \; \hat{y}$ (more later).

Model Selection

- We cannot use R^2 for model comparison when the competing models with unequal number of explanatory variables.
  - R^2 never decreases as we add more variables (OLS minimizes SSE; same as maximize R^2)
  - May include variables with no economic or intuitive foundation

- Adjusted R^2 explicitly accounts for the sample size n and the number of predictor variables k.

$$= 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

  - Imposes a penalty for any additional predictors
  - The higher the adjusted R^2, the better the model

- When comparing models with the same response, the model with the higher adjusted R^2 is preferred.

Model Selection

- Recall the introductory case and consider three models.

$$\text{Model 1: Earnings} = \beta_0 + \beta_1 \text{Cost} + \varepsilon$$
$$\text{Model 2: Earnings} = \beta_0 + \beta_1 \text{Cost} + \beta_2 \text{Grad} + \beta_3 \text{Debt} + \varepsilon$$
$$\text{Model 3: Earnings} = \beta_0 + \beta_1 \text{Cost} + \beta_2 \text{Grad} + \beta_3 \text{Debt} + \beta_4 \text{City} + \varepsilon$$

a. Which of the three models is the preferred model?

b. Interpret the coefficient of determination for the preferred model.

c. What percentage of the sample variation in annual post-college earnings is unexplained by the preferred model?

Model Selection

	Model 1	Model 2	Model 3
Standard error of the estimate s_e	6,271.4407	5,751.8065	5,645.8306
Coefficient of determination R^2	0.2767	0.4023	0.4292
Adjusted R^2	0.2703	0.3862	0.4087

- Model 3 has the lowest standard error of the estimate and the highest adjusted R^2 and is therefore the preferred model for making predictions.

- We cannot rely on R^2 to compare the models because they are all based on different number of predictor variables.

- Model 3 explains 42.92% of the sample variation in the earnings; it does not explain 57.08% of the sample variation in earnings.

Model Selection

	Model 1	Model 2	Model 3
Intercept	28,375.4051* (0.000)	11819.4747 (0.129)	10,004.9665 (0.193)
Cost	0.7169* (0.000)	0.5050* (0.000)	0.4349* (0.000)
Grad	NA	192.6664* (0.007)	178.0989* (0.011)
Debt	NA	104.6573 (0.378)	141.4783 (0.230)
City	NA	NA	2,526.7888* (0.024)
se	6,271.4407	5,751.8065	5,645.8306
R^2	0.2767	0.4023	0.4292
Adjusted R^2	0.2703	0.3862	0.4087
F-test (p-value)	43.608 (0.000)	25.124 (0.000)	20.868(0.000)

Note: Parameter estimates are in the top half of the table with the *p*-values in parentheses; * represents significance at the 5% level. NA denotes not applicable. The lower part of the table contains goodness-of-fit measures.

Model Selection

Example. First subset the **House_Price** data by the college town of Ames (Iowa State University) or the college town of Lincoln (University of Nebraska). Also, only include Single Family houses. You should obtain 612 observations between these two towns, of which 209 are in Ames and 403 are in Lincoln.

1. Find the averages for the sale amount, number of bedrooms, number of bathrooms, square footage, and lot size in both campus towns.

2. Estimate the following two models to predict the price of a house.

   Model 1: $Sale\_amount = \beta_0 + \beta_1 Beds + \beta_2 Baths + \beta_3 Sft\_home + \beta_4 Sft\_lot + \beta_5 Ames + \varepsilon$,
   where Ames is equal to 1 if house is located in Ames, 0 otherwise.

   Model 2: $Sale\_amount = \beta_0 + \beta_1 Beds + \beta_2 Baths + \beta_3 Sft_{home} + \beta_4 Sft_{lot} + \beta_5 Ames + \beta_6 Newer + \varepsilon$,
   where Newer is equal to 1 if the house was built in 2000 or later, 0 otherwise.

3. Which of the above models is better for making predictions?

Confidence and Prediction Intervals

$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$

We can estimate the above model as $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{ki}$

Let $x_1^0, x_2^0, \ldots, x_K^0$ denote specific values of the explanatory variables. At these specific values, $\hat{y}_i^0 = b_1 x_{1i}^0 + b_2 x_{2i}^0 + \cdots + b_K x_{ki}^0$.

- $\hat{y}_i^0$ is the point estimate for both y_i and $E(y_i)$.
- Point estimators are subject to sampling variations. In other words, the estimates will change if we use a different sample to estimate the regression model.
- We use the point estimate along with the margin of error to construct two types of interval estimates.

Confidence Interval: Interval estimate for $E(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$

Prediction Interval: Interval estimate for $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$

The prediction interval is always wider that the confidence interval because it also accounts for the variability due to the random error term ε_i.

Confidence and Prediction Intervals

For specific values $x_1^0, x_2^0, \ldots, x_K^0$, the $100(1-\alpha)\%$ **confidence interval** for the expected value of y_i is:

$$\hat{y}_i^0 \pm t_{\alpha/2,df} se(\hat{y}_i^0)$$

- $\hat{y}_i^0 = b_1 x_{1i}^0 + b_2 x_{2i}^0 + \cdots + b_k x_{ki}^0$; $se(\hat{y}_i^0)$ is the standard error of $\hat{y}_i^0$
- $t_{\alpha/2,df}$ is the value associated with the probability $\alpha/2$ in the upper tail of the distribution with $df = n - k - 1$.

For specific values $x_1^0, x_2^0, \ldots, x_k^0$, the $100(1-\alpha)\%$ **prediction interval** for an individual value of y_i is:

$$\hat{y}_i^0 \pm t_{\alpha/2,df} \sqrt{(se(\hat{y}_i^0))^2 + s_e^2}$$

- $\hat{y}_i^0 = b_1 x_{1i}^0 + b_2 x_{2i}^0 + \cdots + b_K x_{ki}^0$; $se(\hat{y}_i^0)$ is the standard error of $\hat{y}_i^0$
- $t_{\alpha/2,df}$ is defined as above $df = n - k - 1$
- s_e is the standard error of the estimate

Confidence and Prediction Intervals

Let us revisit the ***College*** data example.

Dependent Variable: Annual post-college earnings (Earnings in $)

Explanatory Variables:

The average annual cost (Cost in $)
The graduation rate (Grad in %)
The percentage of students paying down debt (Debt in %)
If a college is located in a city (City equals 1 if a city location, 0 otherwise)

Question: Estimate the model to predict Earnings if Cost = $25,000, Grad = 60, Debt = 80, and City = 1.

$\widehat{\text{Earnings}} = 10{,}004.97 + 0.434 \times 25{,}000 + 178.10 \times 60 + 141.48 \times 80 + 2{,}526.79 \times 1 = 45{,}408.7991$ or about $45,409

$45,409 is the point estimate.

Confidence and Prediction Intervals

Using Excel for Confidence and Prediction Intervals

Confidence Interval: $\hat{y}_i^0 \pm t_{\alpha/2,df} se(\hat{y}_i^0)$

To derive $\hat{y}_i^0$ and $se(\hat{y}_i^0)$, run an auxiliary regression of y on all x_j^* variables where $x_j^* = x_j - x_j^0$.

The resulting estimate of the intercept and its standard error equal $\hat{y}_i^0$ and $se(\hat{y}_i^0)$, respectively.

Prediction Interval: $\hat{y}_i^0 \pm t_{\alpha/2,df} \sqrt{(se(\hat{y}_i^0))^2 + s_e^2}$

To derive $\hat{y}_i^0$, $se(\hat{y}_i^0)$, and s_e, again run the above auxiliary regression of y on all x_j^*.

The resulting estimate of the intercept and its standard error equal $\hat{y}_i^0$ and $se(\hat{y}_i^0)$, respectively and s_e is the standard error of the estimate.

Let's illustrate with Excel and R.

Confidence and Prediction Intervals

Using R for Confidence and Prediction Intervals

```
> # Import College data for confidence/prediction intervals
> Model <- lm(Earnings ~ Cost + Grad + Debt + City, data = myData)
> predict(Model, data.frame(Cost=25000,Grad=60,Debt=80,City=1), level=0.95, interval="confidence")
      fit      lwr      upr
1 45408.8 43392.99 47424.61
> predict(Model, data.frame(Cost=25000,Grad=60,Debt=80,City=1), level=0.95, interval="prediction")
      fit      lwr      upr
1 45408.8 34041.05 56776.55
```

For the specific values of the explanatory variables:

- The 95% confidence interval is [$43,393, $47,425]
- The 95% prediction interval is [$34,041, $56,777]
- As expected, the prediction interval is wider because it also accounts for the variability caused by the random error term.
- At the 90% level, the confidence and the prediction intervals are [$43,721, $47,096] and [$35,893, $54,924].
- Why are the intervals narrower at 90% confidence?
- $t_{0.025,111} = 1.981567$ and $t_{0.05,111} = 1.658697$[qt(α/2, df, lower.tail=FALSE)

Hypothesis Testing

Hypothesis testing is used to resolve conflicts between two competing hypotheses on a particular population of interest.

The null hypothesis is denoted H_0.

The alternative hypothesis is denoted H_A.

We conduct a hypothesis test to determine whether or not sample evidence contradicts H_0.

We can make one of two decisions: reject the null hypothesis or do not reject the null hypothesis.

- If sample evidence is inconsistent with the null hypothesis, we reject the null hypothesis.
- Conversely, if sample evidence is not inconsistent with the null hypothesis, then we do not reject the null hypothesis.

It is not correct to conclude that "we accept the null hypothesis" because while the sample information may not be inconsistent with the null hypothesis, it does not necessarily prove that the null hypothesis is true. (guilty or not guilty)

Hypothesis Testing

Type I Error: reject the null hypothesis when it is true.

Type II Error: do not reject the null hypothesis when it is false.

It is not always easy to determine which of the two errors has more serious consequences.

- For given evidence, there is a trade-off between these errors; by reducing the likelihood of a Type I error, we implicitly increase the likelihood of a Type II error, and vice versa.
- The only way we can reduce both errors is by collecting more evidence.

Let α and β be the probability of Type I error and Type II error, respectively.

- For a given n, we can reduce α only at the expense of a higher β and reduce β only at the expense of a higher α.
- We can lower both α and β by increasing the sample size n.
- The optimal choice of α and β depends on the relative cost of these two types of errors, and determining these costs is not always easy.

Hypothesis Testing

Example: An online retailer is deciding whether or not to build a brick-and-mortar store in a new marketplace. A market analysis determines that the venture will be profitable if average pedestrian traffic exceeds 500 people per day. The competing hypotheses are specified as follows.

H_0: Do not build brick-and-mortar store

H_A: Build brick-and-mortar store

Discuss the consequences of a Type I error and a Type II error.

- A Type I error occurs when the retailer rejects H_0, but H_0 is true; that is, the retailer builds the brick-and-mortar store, but average pedestrian traffic does not exceed 500 people per day and the venture will not be profitable.
- A Type II error occurs when the retailer does not reject H_0, but H_0 is false; that is, the retailer does not build the brick-and-mortar store, but average pedestrian traffic exceeds 500 people per day and the venture would have been profitable.
- Arguably, the consequences of a Type I error in this example are more serious than those of a Type II error.

Practice Problem: The manager of a large manufacturing firm is considering switching to new and expensive software that promises to reduce its assembly costs. Before purchasing the software, the manager wants to conduct a hypothesis test to determine if the new software does reduce its assembly costs.

a. Would the manager of the manufacturing firm be more concerned about a Type I error or a Type II error? Explain.

b. Would the software company be more concerned about a Type I error or a Type II error? Explain.