

Continuous data – multiple regression (indicator variables)

Types of Variables/Number of Groups

Variable(s)	Analysis
Continuous vs. Categorical (2 categories)	Two-group t-test
Continuous vs. Categorical (>2 categories)	ANOVA
Continuous vs. Continuous	Correlation/linear regression

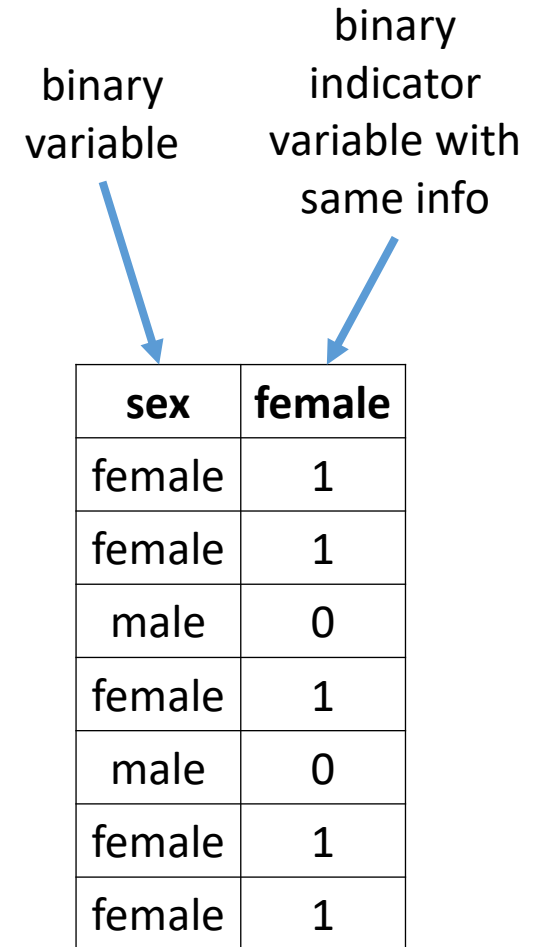
Types of Variables/Number of Groups

Variable(s)	Analysis
Continuous vs. Categorical (2 categories)	Two-group t-test or linear regression
Continuous vs. Categorical (>2 categories)	ANOVA or linear regression
Continuous vs. Continuous	Correlation or linear regression

Linear regression requires a continuous outcome variable,
but the predictor variable(s) can actually be any type!

Indicator Variables

- When a predictor variable is binary, there are only two possible values for x
- Can code observations as “0” and “1” instead of category names
 - Example: Sex (male/female) can be coded as 0/1 or 1/0
 - Example: Survival status (alive/dead) can be coded as 0/1 or 1/0
 - Example: (Yes/no) can be coded as 0/1 or 1/0 (usually 1/0)
- Consider 0 as the baseline/reference category, and the 1’s are the “indicators” of the other category
- Data dictionary tells you the meaning of “0” and “1”



The diagram illustrates the conversion of a categorical variable into a binary indicator variable. Two blue arrows point from the labels 'binary variable' and 'binary indicator variable with same info' to the respective columns of the table below.

sex	female
female	1
female	1
male	0
female	1
male	0
female	1
female	1

Indicator Variables in Regression

- When coded as 0/1, binary indicator variables can be included as predictors in a regression model
- Treated as any other numerical variable

Example: Suppose x is a binary indicator variable for sex (1=Female, 0=Male). The y variable is age at death. Calculate life expectancy for men and women.

$$\hat{y} = 76.1 + 5.0x$$

Life expectancy for men ($x = 0$):

Life expectancy for women ($x = 1$):

Coefficient Interpretation:

1 Binary Predictor

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\left\{ \begin{array}{l} y \text{ continuous} \\ x \text{ binary} \end{array} \right\}$$

- β_0 is the expected value of y when $x = 0$
- $\beta_0 + \beta_1$ is the expected value of y when $x = 1$
- Thus, β_1 is the expected difference in y between subjects with $x = 0$ and $x = 1$

Coefficient Interpretation:

2 Binary Predictors

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\left\{ \begin{array}{l} y \text{ continuous} \\ x_1 \text{ binary} \\ x_2 \text{ binary} \end{array} \right\}$$

- β_0 is the expected value of y when $x_1 = 0$ and $x_2 = 0$
- β_1 is the expected difference in y between subjects with $x_1 = 0$ and $x_1 = 1$, holding x_2 constant
- β_2 is the expected difference in y between subjects with $x_2 = 0$ and $x_2 = 1$, holding x_1 constant

Coefficient Interpretation:

1 Binary & 1 Continuous Predictor

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\left\{ \begin{array}{l} y \text{ continuous} \\ x_1 \text{ binary} \\ x_2 \text{ continuous} \end{array} \right\}$$

- β_0 is the expected value of y when $x_1 = 0$ and $x_2 = 0$
- β_1 is the expected difference in y between subjects with $x_1 = 0$ and $x_1 = 1$, holding x_2 constant
- β_2 is the expected change in y for every 1 unit increase in x_2 , holding x_1 constant (or, within levels of x_1)

Low Birth Weight Data

- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

Variable	Description
sex	Sex of the baby (Male, Female)
gestage	Gestational age at time of birth (weeks)
length	Length of the baby (cm)
birthwt	Birth weight of the baby (g)
headcirc	Baby's head circumference (cm)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.

Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: Apgar Score/Hemorrhage

- Fit a linear regression model to predict the Apgar score for low birth weight infants with and without germinal matrix hemorrhage.

y = outcome variable (continuous) = apgar score for low birth weight babies
x = predictor variable (binary) = germinal matrix hemorrhage

Example: Apgar Score/Hemorrhage

Summary of linear regression model for Apgar score:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.5059	0.2564	25.373	<2e-16
hemorrhage[T.Yes]	-1.7059	0.6620	-2.577	0.0115

Residual standard error: 2.364 on 98 degrees of freedom
Multiple R-squared: 0.06345, Adjusted R-squared: 0.05389
F-statistic: 6.639 on 1 and 98 DF, p-value: 0.01147

Estimated Apgar score for infant without germinal matrix hemorrhage ($x = 0$):

Estimated Apgar score for infant with germinal matrix hemorrhage ($x = 1$):

$$\hat{y} = 6.51 - 1.71x$$

6.51 is expected apgar score without germinal matrix hemorrhage

-1.71 is expected difference in apgar score with gmh

Example: Apgar Score/Hemorrhage

- Add birth weight to the model.
- Does germinal matrix hemorrhage have a significant association with Apgar score after adjusting for birth weight?
- Does including birth weight in the model improve model fit?

y = apgar score of lbw babies

x1 = germinal matrix hemorrhage

x2 = birth weight

Example: Apgar Score/Hemorrhage

Summary of linear regression model for Apgar score:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.1163051	1.0002428	5.115	0.00000158
hemorrhage[T.Yes]	-1.6631592	0.6591426	-2.523	0.0133
birthwt	0.0012587	0.0008761	1.437	0.1540

Residual standard error: 2.351 on 97 degrees of freedom

Multiple R-squared: 0.08297, Adjusted R-squared: 0.06406

F-statistic: 4.388 on 2 and 97 DF, p-value: 0.01499

p -value for hemorrhage = 0.013

Since the p -value is less than 0.05, we reject the null hypothesis and conclude that there is sufficient evidence to suggest that there is an association between germinal matrix hemorrhage and Apgar score after adjusting for birth weight.

Adjusted- R^2 from model with birth weight = 0.064

Adjusted- R^2 from model without birth weight = 0.054

Since the adjusted- R^2 is higher in the model with birth weight, we prefer this model.

Categorical Variables with >2 Categories

- When there are more than two categories, we can include these categorical variables as predictors in regression by using multiple indicator variables
- Multiple indicator variables
 - One category is always the baseline/reference category
 - Need an indicator variable (0/1) for each of the other categories
 - Examples: 3-category variable needs 2 indicator variables
4-category variable needs 3 indicator variables
5-category variable needs 4 indicator variables ...

Example: Gestational Age Categories

- *gestage3* is a 3-category variable specifying whether an infant's gestational age is Preterm (32-36 weeks), Very preterm (28-31 weeks), or Extremely preterm (<28 weeks). We can code this information by using two indicator variables.
 - Infants with "Preterm" gestational age have `verypre=0` and `extpre=0`
 - Infants with "Very preterm" gestational age have `verypre=1` and `extpre=0`
 - Infants with "Extremely preterm" gestational age have `verypre=0` and `extpre=1`

gestage3	verypre	extpre
Preterm	0	0
Preterm	0	0
Extremely preterm	0	1
Very preterm	1	0
Preterm	0	0
Very preterm	1	0

Example: Gestational Age Categories

- To look at the association between *gestage3* and some outcome variable, we should include *verypre* and *extpre* as predictors in the linear regression model.
- Example: What is the predicted Apgar score for infants in each of the three gestational age categories?

gestage3	verypre	extpre
Preterm	0	0
Preterm	0	0
Extremely preterm	0	1
Very preterm	1	0
Preterm	0	0
Very preterm	1	0

Example: Gestational Age Categories

Summary of linear regression model for Apgar score:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.0000	0.6260	11.182	<2e-16
verypre	-0.7091	0.7062	-1.004	0.318
extpre	-1.2000	0.7667	-1.565	0.121

Residual standard error: 2.424 on 97 degrees of freedom

Multiple R-squared: 0.02498, Adjusted R-squared: 0.004872

F-statistic: 1.242 on 2 and 97 DF, p-value: 0.2933

$\hat{y} = \hat{\text{apgar}}$

$\hat{y} = 7.00 - 0.71(\text{verypre}) - 1.2(\text{extremepre})$

Estimated Apgar score for preterm infant
(verypre = 0, extpre = 0):

$\hat{\text{apgar}} = 7.00$ points

Estimated Apgar score for very preterm
infant (verypre = 1, extpre = 0):

$\hat{\text{apgar}} = 7.00 - 0.71(1) - 1.20(0)$
 $= 6.29$ points

Estimated Apgar score for extremely preterm
infant (verypre = 0, extpre = 1):

$\hat{\text{apgar}} = 7.00 - 0.71(0) - 1.20(1)$
 $= 5.80$ points

Coefficient Interpretation: One 3-Category Predictor (2 Indicators)

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\left\{ \begin{array}{l} y \text{ continuous} \\ x_1 \text{ binary indicator for 2}^{\text{nd}} \text{ category} \\ x_2 \text{ binary indicator for 3}^{\text{rd}} \text{ category} \\ 1^{\text{st}} \text{ category is the reference} \end{array} \right\}$$

- β_0 is the expected value of y for subjects in the first category
- β_1 is the expected difference in y between subjects in the first category and subjects in the second category
- β_2 is the expected difference in y between subjects in the first category and subjects in the third category

(If there are any other predictors in the model, just add “holding the other predictor variables constant” to the interpretation.)

Categorical Variables with >2 Categories

The linearity assumption - we assume that the difference between preterm and very preterm vs very preterm and extreme preterm has to be the same...linearity forces difference in outcome to be same between each category

- Why not just code the categories as 0/1/2 and include a single predictor in the model?
- Example: Suppose that x was coded as 0 if the infant is “Preterm”, 1 if the infant is “Very preterm”, and 2 if the infant is “Extremely preterm”.

Estimated Apgar score for preterm infant ($x = 0$):

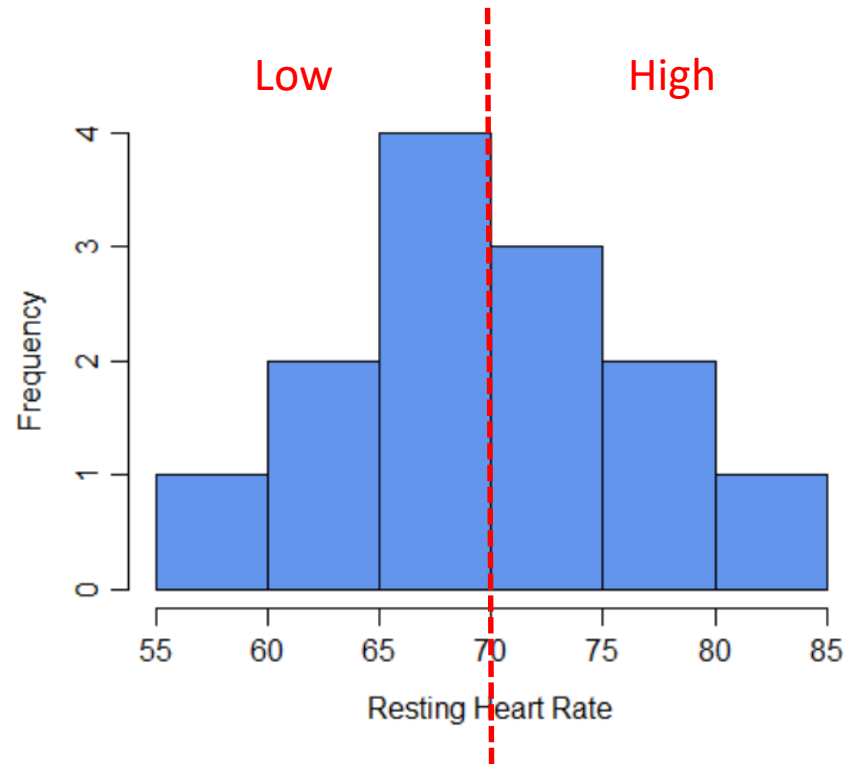
$$\hat{y} = 6.92 - 0.58x$$

Estimated Apgar score for very preterm infant ($x = 1$):

Estimated Apgar score for extremely preterm infant ($x = 2$):

Categorizing Continuous Variables

- Common in medical research to split a continuous predictor variable into a small number of categories
- Try to avoid this!



hr	hr_cat
57	Low
61	Low
62	Low
66	Low
66	Low
68	Low
69	Low
70	High
71	High
74	High
76	High
76	High
82	High

Categorizing Continuous Variables

- Disadvantages:

1. Loss of power to detect association between x and y
2. Loss of precision in estimating/predicting y
3. Different cutpoints in different studies make it difficult to compare results
4. Think of two observations very close to each other but on different sides of the cutpoint... Dichotomization assumes these subjects are more similar to the others in their group than to each other.
5. Can lead to misunderstanding that your risk of something doesn't change until you hit a certain point
 - Example: relationship between BMI and risk of heart disease
6. Especially problematic when the data are used to determine the cutpoint
 - Example: split at median into high/low groups

Categorizing Continuous Variables

- When it can be okay:
 - Splits represent meaningful cutpoints
 - Example: retirement planning study, dichotomize age at 55 years old since that's when eligibility for social programs kicks in
 - Relationship between continuous x and y is not linear
 - We could perform a transformation of x or y , but if we want to preserve interpretability of the model, it could be preferable to categorize x instead.

Disadvantages of categorizing continuous variables outweigh the advantages 99%* of the time.

Important Points

- Concept of an indicator variable
- Interpretation of linear regression models with binary predictors
- Interpretation of linear regression models with categorical predictors (>2 categories)
- Pros/cons of categorizing a continuous variable