Continuous data – multiple regression (model building, confounding)

Simple vs. Multiple Regression

Simple Linear Regression

- One continuous outcome variable, one predictor variable
- Bivariate analysis
- $\hat{\beta}_1$ is the **unadjusted** estimate of the association between x and y

$$y = \beta_0 + \beta_1 x + e$$

Multiple Linear Regression

- One continuous outcome variable, multiple predictor variables
- Multivariable analysis
- $\hat{\beta}_1$ is the estimate of the association between x_1 and y, **adjusted** for x_2, x_3, etc.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + e$$

Multiple Regression Coefficient Interpretation

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

- β_0 is the expected value of y when $x_1 = 0$ and $x_2 = 0$
- β_1 is the expected change in y for every 1 unit increase in x_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

Use What We Know from Simple Regression

- Interpretation of slope coefficients is similar
  - Must include “holding the other predictors constant” or “adjusting/controlling for the other predictors”
- Making predictions for the outcome variable is the same
  - Just plug in each of the predictor values
- Test for association between a predictor and the outcome is similar
  - *Adjusted* p-values are in the regression table
  - Tests whether association between predictor and outcome is significant after adjusting for the other predictor variables
- Concept of fitted value ($\hat{y}$) and residual (e) is the same

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.

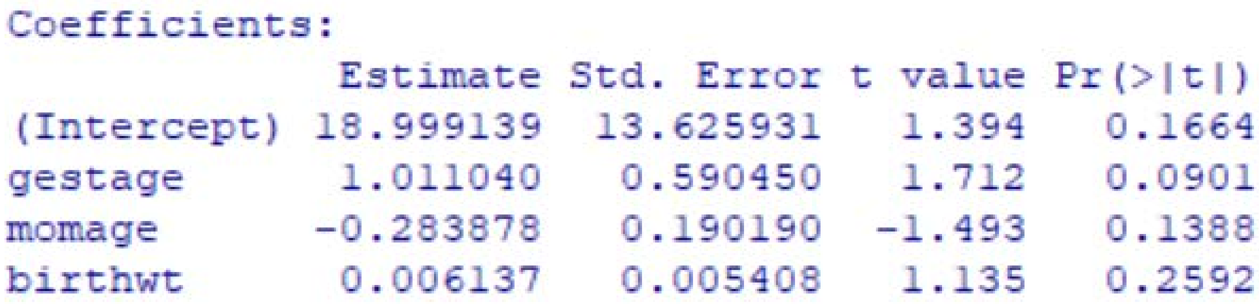
Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: SBP in LBW Infants

- Use gestational age, mothers' age, and birth weight to predict systolic blood pressure (SBP) in low birth weight (LBW) infants.
  - Interpret the coefficients in the linear regression model.
  - What is the predicted SBP for a LBW infant with gestational age of 30 weeks, mother's age of 28 years old, and birth weight of 1000 grams?

Example: SBP in LBW Infants

Summary of linear regression model for SBP:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.999139  13.625931   1.394   0.1664
gestage      1.011040   0.590450   1.712   0.0901
momage      -0.283878   0.190190  -1.493   0.1388
birthwt      0.006137   0.005408   1.135   0.2592
```

- $\hat{\beta}_0 = 19.00$
  - On average, SBP is 19.00 mmHg when gestational age is 0 weeks, mothers' age is 0 years old, and birth weight is 0 grams.
  - Doesn't make sense – that's okay.

- $\hat{\beta}_1 = 1.01$

  On average, a one week increase in gestational age is associated with a 1.01 mmHg increase in SBP, holding mothers' age and infants' birth weight constant.

- $\hat{\beta}_2 = -0.28$

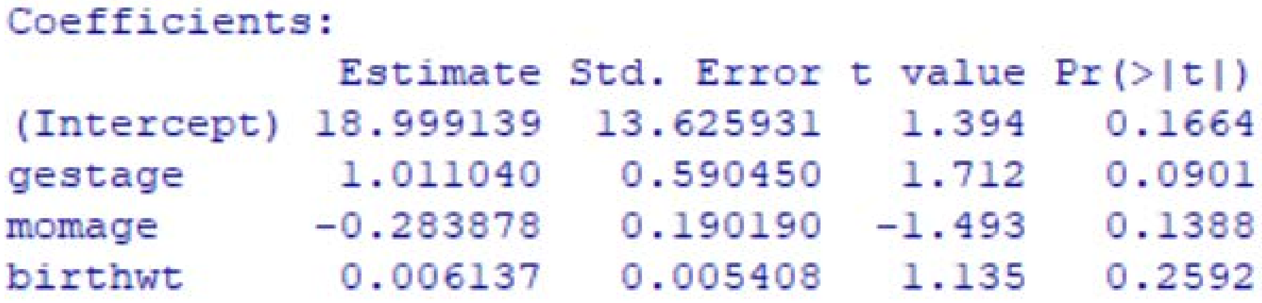
  On average, a one year increase in mothers' age is associated with a 0.28 mmHg decrease in SBP, holding gestational age and birth weight constant.

- $\hat{\beta}_3 = 0.006$

  On average, a one gram increase in birth weight is associated with a 0.006 mmHg increase in SBP, holding gestational age and mothers' age constant.

Example: SBP in LBW Infants

Summary of linear regression model for SBP:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.999139  13.625931   1.394   0.1664
gestage      1.011040   0.590450   1.712   0.0901
momage      -0.283878   0.190190  -1.493   0.1388
birthwt      0.006137   0.005408   1.135   0.2592
```

Want predicted SBP for a LBW infant with gestational age of 30 weeks, mother’s age of 28 years, and birth weight of 1000 grams.

$$\widehat{\mathrm{SBP}} = 19.00 + 1.01(\mathrm{gestage}) - 0.28(\mathrm{momage}) + 0.006(\mathrm{birthwt})$$

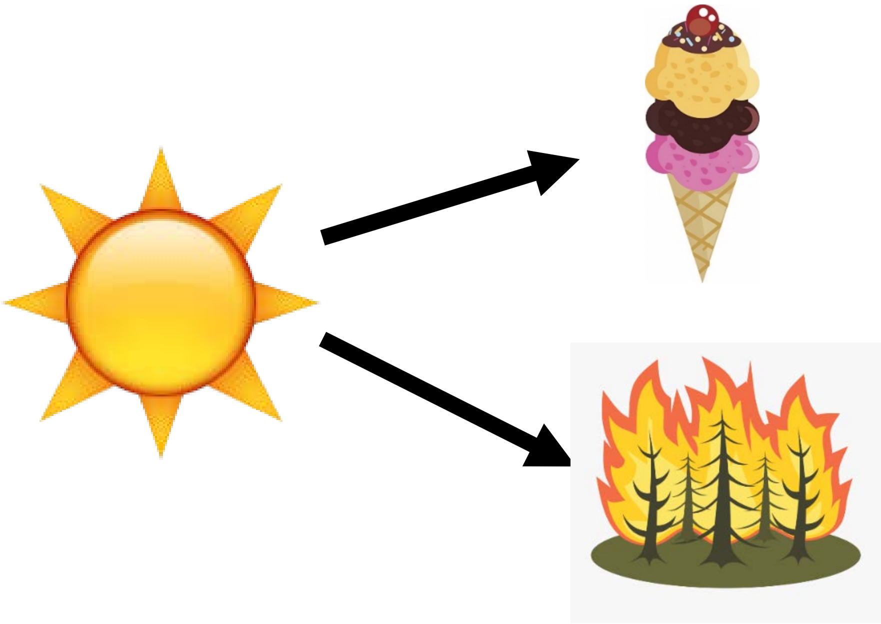
$$\widehat{\mathrm{SBP}} = 19.00 + 1.01(30) - 0.28(28) + 0.006(1000)$$

$$\widehat{\mathrm{SBP}} = 47.52\ \mathrm{mmHg}$$

Why Do Multiple Regression?

- Creates a more holistic representation of the relationships between the variables
- Makes better predictions for y
- Can adjust for confounding

Confounding Adjustment



- Unadjusted model:

$$\boxed{\text{forest fire frequency} = \beta_0 + \beta_1(\text{ice cream sales}) + e}$$

$\hat{\beta}_1$ is large and test of $H_0\colon \beta_1 = 0$ is significant

- Adjusted model:

$$\boxed{\begin{aligned}\text{forest fire frequency} = \beta_0 + \beta_1(\text{ice cream sales}) &+ \\ \beta_2(\text{temperature}) &+ e\end{aligned}}$$

$\hat{\beta}_1$ is small and test of $H_0\colon \beta_1 = 0$ is not significant

$\hat{\beta}_2$ is large and test of $H_0\colon \beta_2 = 0$ is significant

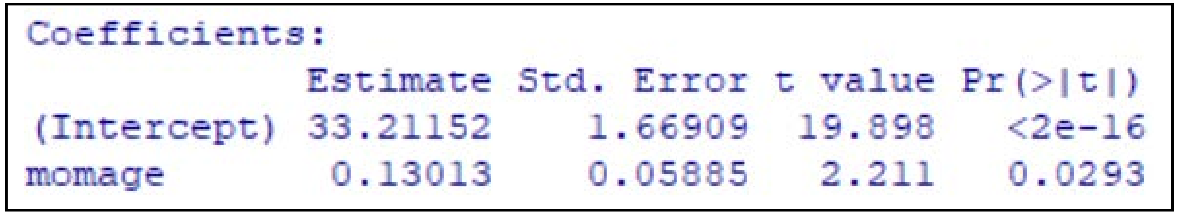
Example: Length of LBW Infants

- First, explore the unadjusted association between length of the baby and mothers’ age. Then, assess whether this association changes when gestational age is added to the model.

Example: Length of LBW Infants

Unadjusted model:

- Outcome: length
- Predictors: mothers’ age



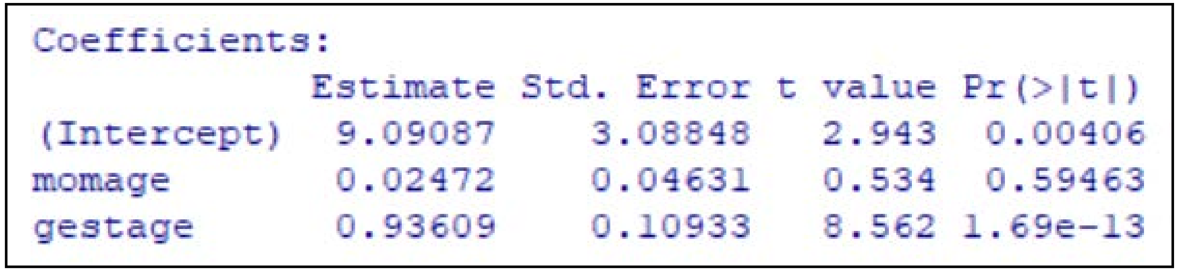
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.21152    1.66909  19.898   <2e-16
momage       0.13013    0.05885   2.211   0.0293
```

$$\widehat{\text{length}} = 33.2 + 0.13(\text{momage})$$

- On average, a one year increase in mothers’ age is associated with a 0.13 cm increase in infants’ length.
- This unadjusted effect is significant (p=0.029).

Adjusted model:

- Outcome: length
- Predictors: mothers’ age, gestational age



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.09087    3.08848   2.943  0.00406
momage       0.02472    0.04631   0.534  0.59463
gestage      0.93609    0.10933   8.562 1.69e-13
```

$$\widehat{\text{length}} = 9.09 + 0.02(\text{momage}) + 0.94(\text{gestage})$$

- On average, a one year increase in mothers’ age is associated with a 0.02 cm increase in infants’ length, holding gestational age constant.
- This adjusted effect is not significant (p=0.595).

Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between each predictor variable and the outcome variable
- **Constant variance** of the residuals
- **Normality** of the residuals
- Absence of **multicollinearity**

Absence of Multicollinearity

What it means

- Predictor variables aren't too strongly correlated
- Multicollinearity can cause:
    - Instability (small changes in the data cause large changes in coefficient estimates)
    - Large standard errors (indicates a lot of uncertainty in our coefficient estimates)

How to assess it

- Calculate correlation between each pair of predictor variables
- No set threshold for correlation that is "too high", but r between -0.5 and 0.5 should be okay

How to fix it

- Include only one of the highly correlated predictor variables in the model

Model Building

- Which predictor variables belong in the regression model? Are any transformations of the variables necessary?
- Epidemiologists, subject-matter experts, and biostatisticians are all needed to provide valuable knowledge about the relationships between the variables and verify that the model satisfies important statistical properties
- **Stepwise model selection** uses the data to decide what should go in the model
  - Not usually recommended because doesn’t incorporate subject-matter knowledge about the variables
  - Beyond the scope of this course

Comparing Models

- Oftentimes, we want to compare two or more models that contain different predictor variables
- Many statistics exist that assess the fit of the model (AIC, BIC, R^2, etc.)
- Remember: R^2 measures the proportion of variation in the outcome that is explained by the predictors
    - In general, higher R^2 means the model fits the data better
    - Problem: R^2 always increases as you put more predictors in the model, and we prefer for models to be parsimonious (as small as possible)

Adjusted R^2

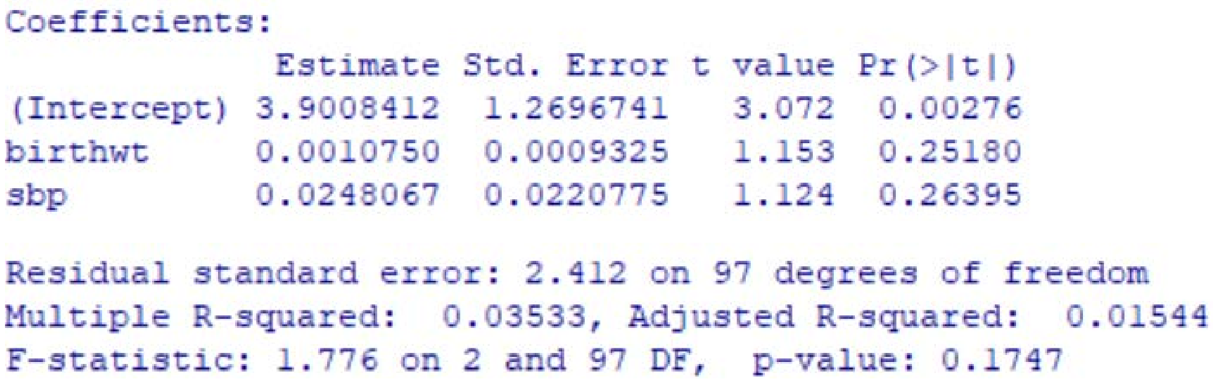
- **Adjusted $\boldsymbol{R^2}$** includes a small penalty for each added predictor
  - Only goes up if the added predictor really does contribute to predicting the outcome
  - Allows for a fair comparison of models

Example: Apgar Score in LBW Infants

- Researchers want to build a model for predicting Apgar score. Since the Apgar score measures the infants’ clinical status at birth, they know that birth weight and systolic blood pressure (SBP) are important determinants. However, they’re unsure whether or not gestational age should be included as a predictor in the model. To answer this question, fit the following two linear regression models and assess which model has better fit.
  - Model 1: apgar ~ birthwt + SBP
  - Model 2: apgar ~ birthwt + SBP + gestage

Example: Apgar Score in LBW Infants

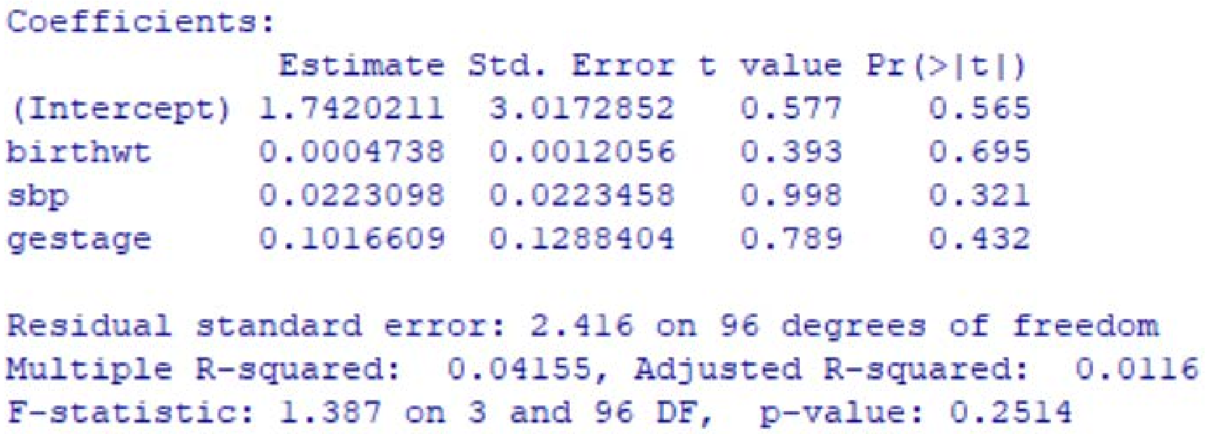
Summary of Model 1 (outcome = Apgar score, predictors = birth weight, SBP)



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.9008412  1.2696741   3.072  0.00276
birthwt     0.0010750  0.0009325   1.153  0.25180
sbp         0.0248067  0.0220775   1.124  0.26395

Residual standard error: 2.412 on 97 degrees of freedom
Multiple R-squared:  0.03533, Adjusted R-squared:  0.01544
F-statistic: 1.776 on 2 and 97 DF,  p-value: 0.1747
```

Summary of Model 2 (outcome = Apgar score, predictors = birth weight, SBP, gestational age)



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.7420211  3.0172852   0.577    0.565
birthwt     0.0004738  0.0012056   0.393    0.695
sbp         0.0223098  0.0223458   0.998    0.321
gestage     0.1016609  0.1288404   0.789    0.432

Residual standard error: 2.416 on 96 degrees of freedom
Multiple R-squared:  0.04155, Adjusted R-squared:  0.0116
F-statistic: 1.387 on 3 and 96 DF,  p-value: 0.2514
```

- Model 1 adjusted-$R^2 = 0.0154$, Model 2 adjusted-$R^2 = 0.0116$
- We prefer Model 1 since the adjusted-R^2 value is larger (gestational age does not contribute enough to explaining variation in Apgar score beyond birth weight and SBP, and thus should not be added to the model)

Model Building Summary

- Building a regression model is just as much an art as it is a science
- Remember: Aim for a model that is parsimonious (as simple and interpretable as possible) while still being valid
- Checking assumptions can be really hard, especially when there are many predictors
- Collaborate with a biostatistician!

Important Points

- Interpretation of coefficients in multiple linear regression models
- Concept of adjusted and unadjusted effects
- Multicollinearity: what is it and why does it matter
- Assessing model fit/comparing models