Continuous data – multiple regression (interactions)

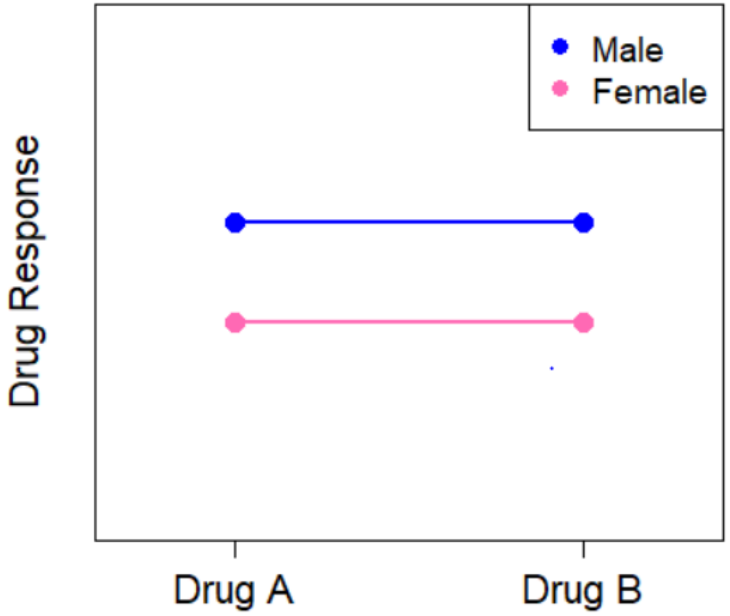
Motivating Example

- Interest lies in the average response to two different drugs: Drug A and Drug B
- We also want to know whether the response is different for men and women
- Can include indicator variables for drug (A vs. B) and sex (male vs. female) in a linear regression model with drug response as the outcome
  - We call these the **main effects** of drug and sex on the response

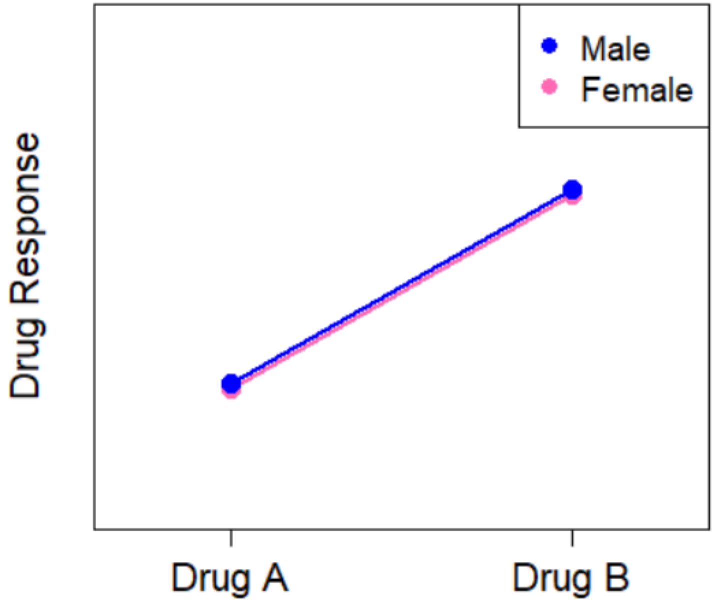
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	X	X	X	X
drug[T.B]	X	X	X	X
sex[T.female]	X	X	X	X

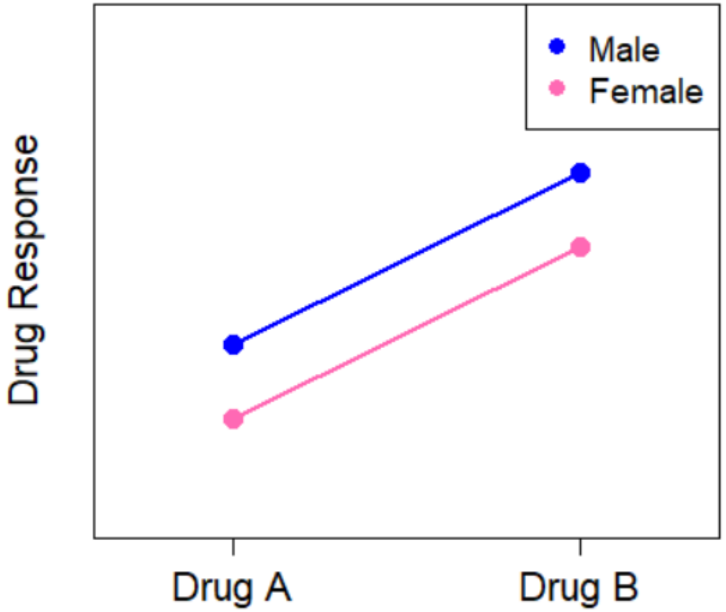
Motivating Example



Main effect of drug? No
Main effect of sex? Yes

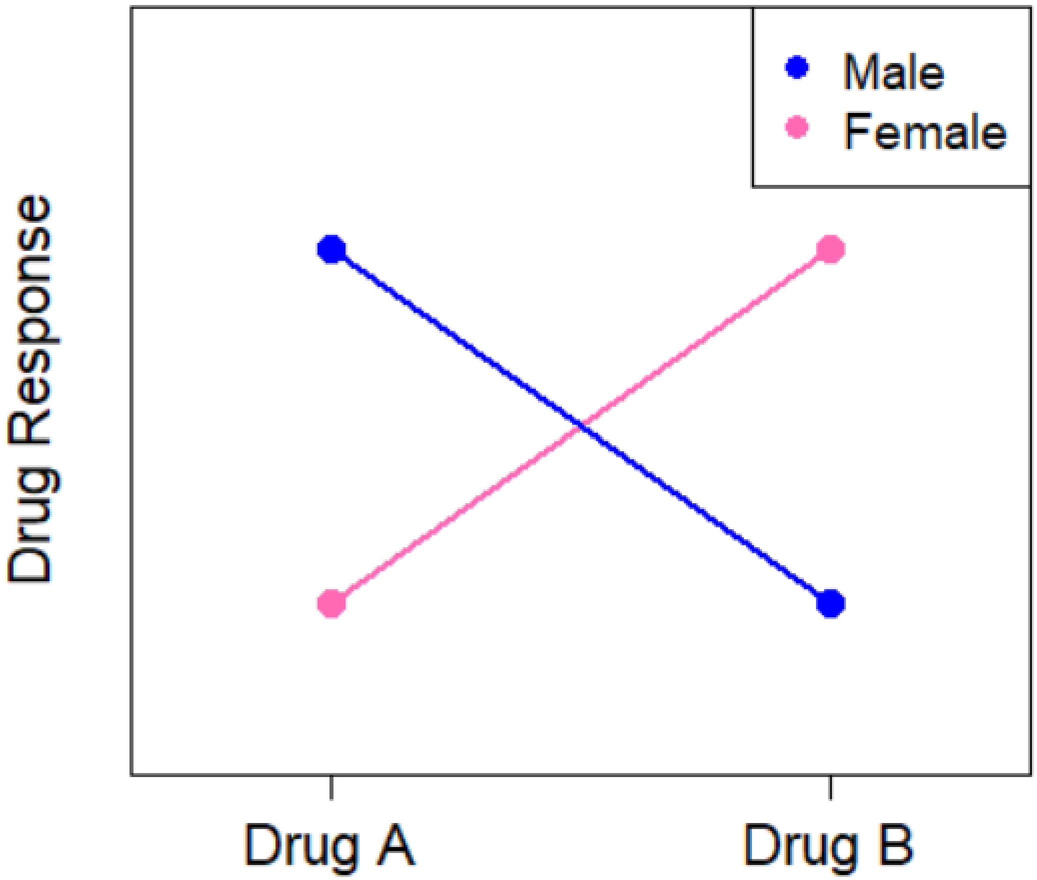


Main effect of drug? Yes
Main effect of sex? No



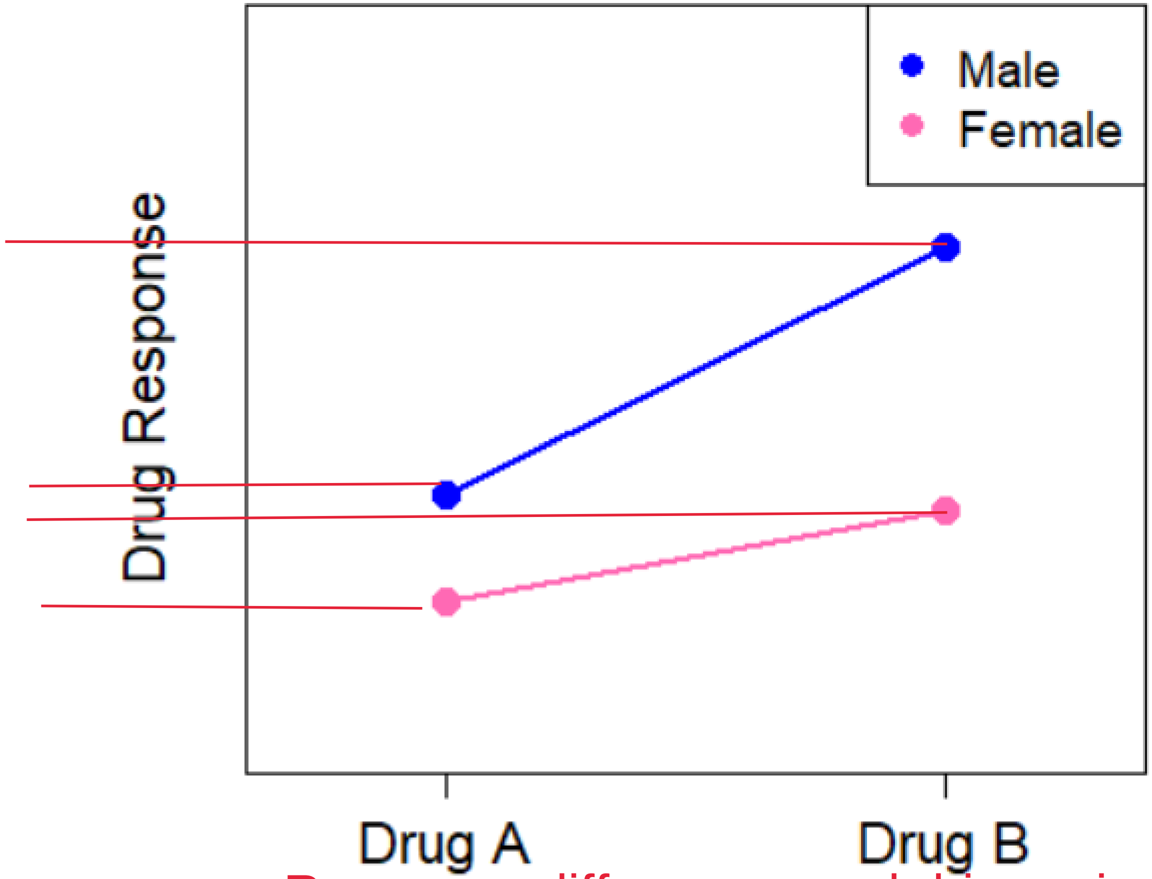
Main effect of drug? Yes
Main effect of sex? Yes

Motivating Example



- Difference in average response on Drug A and average response on Drug B: None
- Difference in average response of males and average response of females: None
- Does this mean that drug and sex have no impact on response? No!
  - Strong interaction between drug and sex

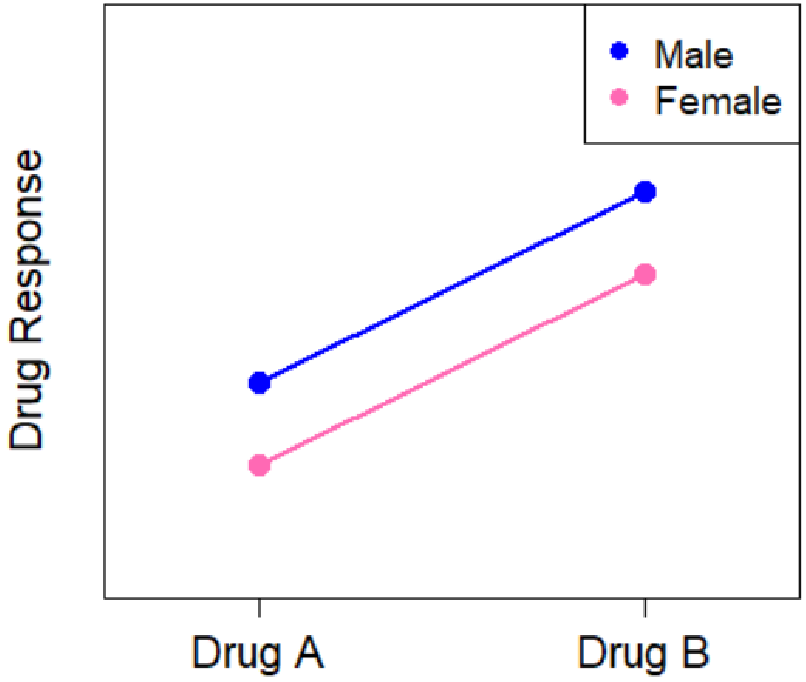
Motivating Example



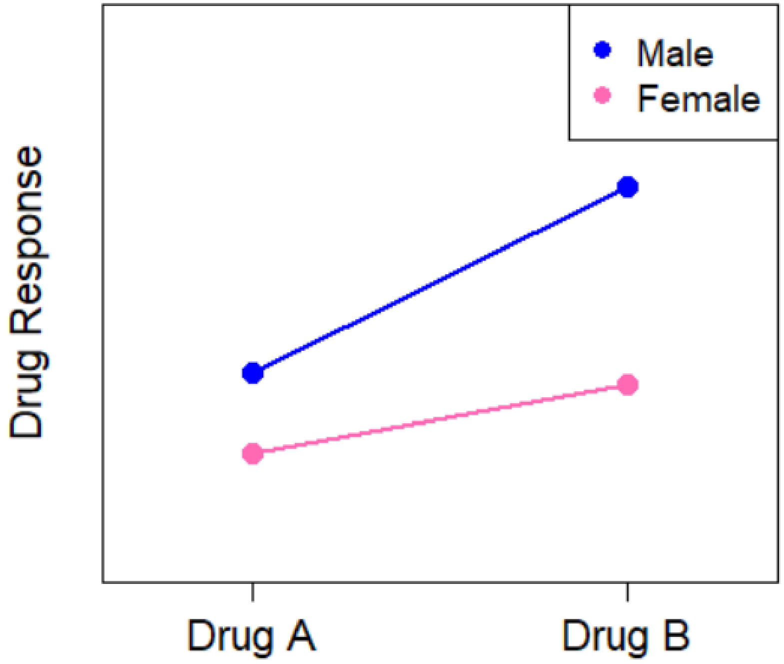
- If the association between one predictor and the outcome varies within levels of another predictor, we say that there is an **interaction** between the two predictor variables

Response difference much bigger in males vs females...there is an interaction in drug response due to sex

Motivating Example



Main effect of drug? yes
Main effect of sex? yes
Drug-sex interaction? no



Main effect of drug? yes
Main effect of sex? yes
Drug-sex interaction? yes

Regression with an Interaction Effect

- To incorporate an interaction effect in a regression model, multiply the two predictors by each other and include that as another predictor:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Shows up in regression output as $x_1{:}x_2$

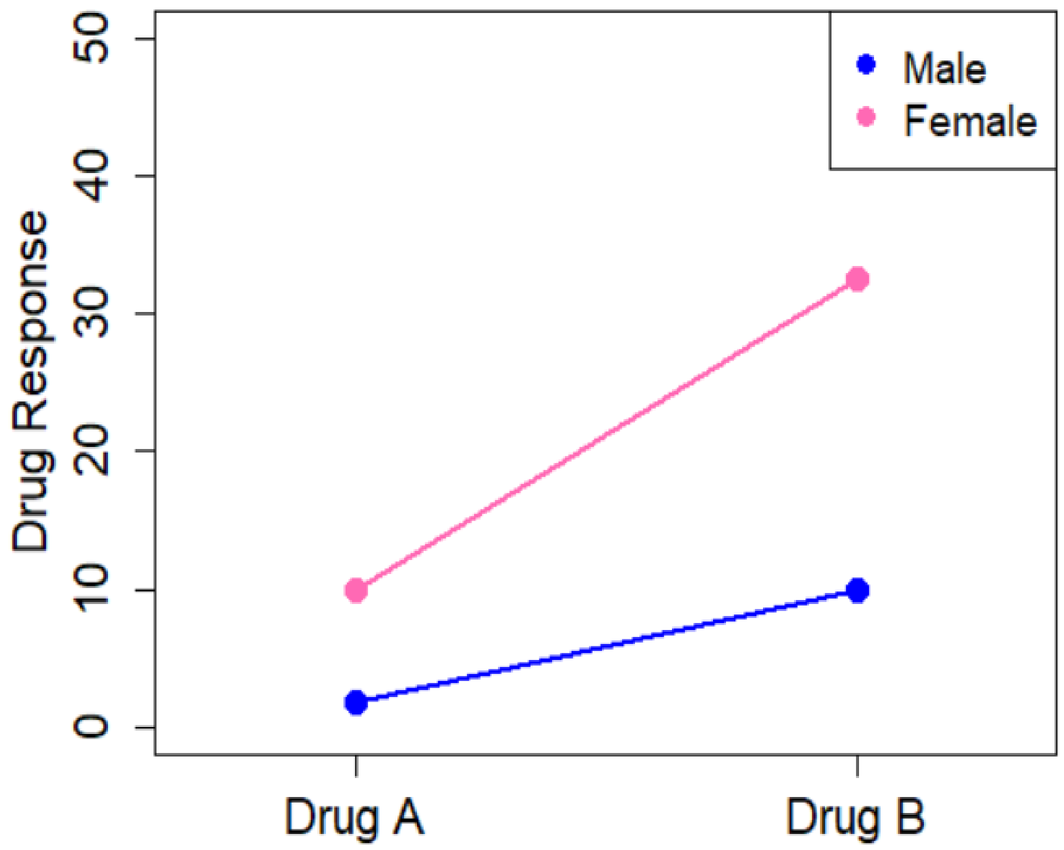
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	X	X	X	X
drug[T.B]	X	X	X	X
sex[T.female]	X	X	X	X
drug[T.B]:sex[T.female]	X	X	X	X

Example: Drug/Sex Interaction

- Using the regression output below, what is the predicted response for each sex on each drug?
- Is there evidence of a drug/sex interaction?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.785      2.078   0.859  0.39258
drug           8.109      2.939   2.759  0.00694
sex            9.450      2.939   3.216  0.00177
drug:sex      13.204      4.156   3.177  0.00200
```



Example: Drug/Sex Interaction

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.785      2.078   0.859  0.39258
drug           8.109      2.939   2.759  0.00694
sex            9.450      2.939   3.216  0.00177
drug:sex      13.204      4.156   3.177  0.00200
```

Drug is coded as 0=Drug A, 1=Drug B
Sex is coded as 0=Male, 1=Female

y = drug response

yhat = 1.79+8.11(drug)+9.45(sex)+13.2(drug*sex interaction)

- Predicted response for males on Drug A:

  =1.79

- Predicted response for males on Drug B:

  yhat = 1.79+8.11(1)+9.45(0)+13.2(1)(0)
  =9.90

- Predicted response for females on Drug A:

  =11.25

- Predicted response for females on Drug B:

  =32.55

Example: Drug/Sex Interaction

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.785      2.078   0.859  0.39258
drug           8.109      2.939   2.759  0.00694
sex            9.450      2.939   3.216  0.00177
drug:sex      13.204      4.156   3.177  0.00200
```

Drug is coded as 0=Drug A, 1=Drug B
Sex is coded as 0=Male, 1=Female

$$\hat{y} = 1.79 + 8.11(drug) + 9.45(sex) + 13.20(drug * sex)$$

$H_0: \beta_{\text{int}} = 0 \qquad H_A: \beta_{\text{int}} \neq 0$

p-value = 0.002

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 interaction between drug and sex on response.

Coefficient Interpretation: 2 Binary Predictors & Interaction

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

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

$\hat{y}$ for each combination of x_1 and x_2:

	$x_1 = 0$	$x_1 = 1$
$x_2 = 0$	β_0	$\beta_0 + \beta_1$
$x_2 = 1$	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

- 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 $x_1 = 1$, when $x_2 = 0$
- β_2 is the expected difference in y between subjects with $x_2 = 0$ and $x_2 = 1$, when $x_1 = 0$
- β_3 is the expected additional difference (on top of β_1) in y between subjects with $x_1 = 0$ and $x_1 = 1$, when $x_2 = 1$

  OR

  β_3 is the expected additional difference (on top of β_2) in y between subjects with $x_2 = 0$ and $x_2 = 1$, when $x_1 = 1$

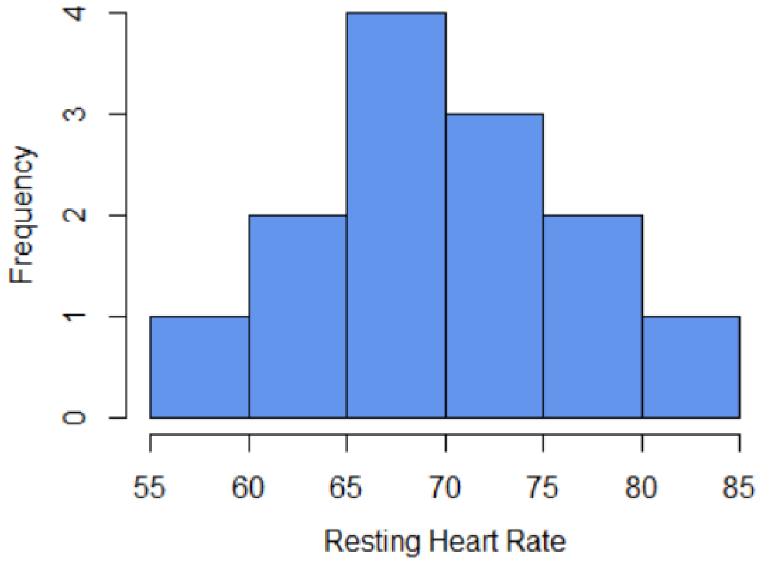
Coefficient Interpretation: 1 Binary, 1 Continuous Predictor, & Interaction

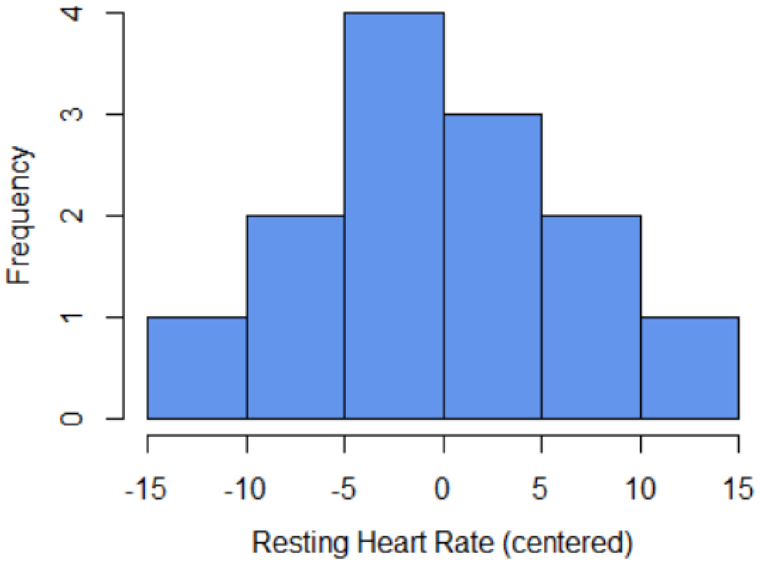
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \qquad \left\{ \begin{array}{l} y \text{ continuous} \\ x_1 \text{ binary} \\ x_2 \text{ continuous} \end{array} \right\}$$

- 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 $x_1 = 1$, when $x_2 = 0$
- β_2 is the expected change in y for every 1 unit increase in x_2, when $x_1 = 0$
- β_3 is the expected additional change in y (on top of β_2) for every 1 unit increase in x_2, when $x_1 = 1$

Variable Centering

- Interpretability of coefficients in a model with an interaction is easier when the continuous variable(s) are **centered**
- To center a variable, subtract the mean of the variable from each observation
  - Now, observations represent number of units above or below the mean





hr	hr_ctr
57	-12.0769
61	-8.07692
62	-7.07692
66	-3.07692
66	-3.07692
68	-1.07692
69	-0.07692
70	0.923077
71	1.923077
74	4.923077
76	6.923077
76	6.923077
82	12.92308

mean hr = 69.07692

Variable Centering

- Interpretability of coefficients in a model with an interaction is easier when the continuous variable(s) are **centered**
- To center a variable, subtract the mean of the variable from each observation
  - Now, observations represent number of units above or below the mean
  - $x = 0$ in interpretations is now meaningful: it represents the average value

Coefficient Interpretation: 1 Binary, 1 Continuous Predictor, & Interaction

Centered

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$$\left\{ \begin{array}{l} y \text{ continuous} \\ x_1 \text{ binary} \\ x_2 \text{ continuous} \\ \quad \text{(and centered!)} \end{array} \right\}$$

- β_0 is the expected value of y for a subject with $x_1 = 0$ and the average value of x_2
- β_1 is the expected difference in y between subjects with $x_1 = 0$ and $x_1 = 1$, when they have the average value of x_2
- β_2 is the expected change in y for every 1 unit increase in x_2, when $x_1 = 0$
- β_3 is the expected additional change in y (on top of β_2) for every 1 unit increase in x_2, when $x_1 = 1$

Coefficient Interpretation: 2 Continuous Predictors & Interaction Effect

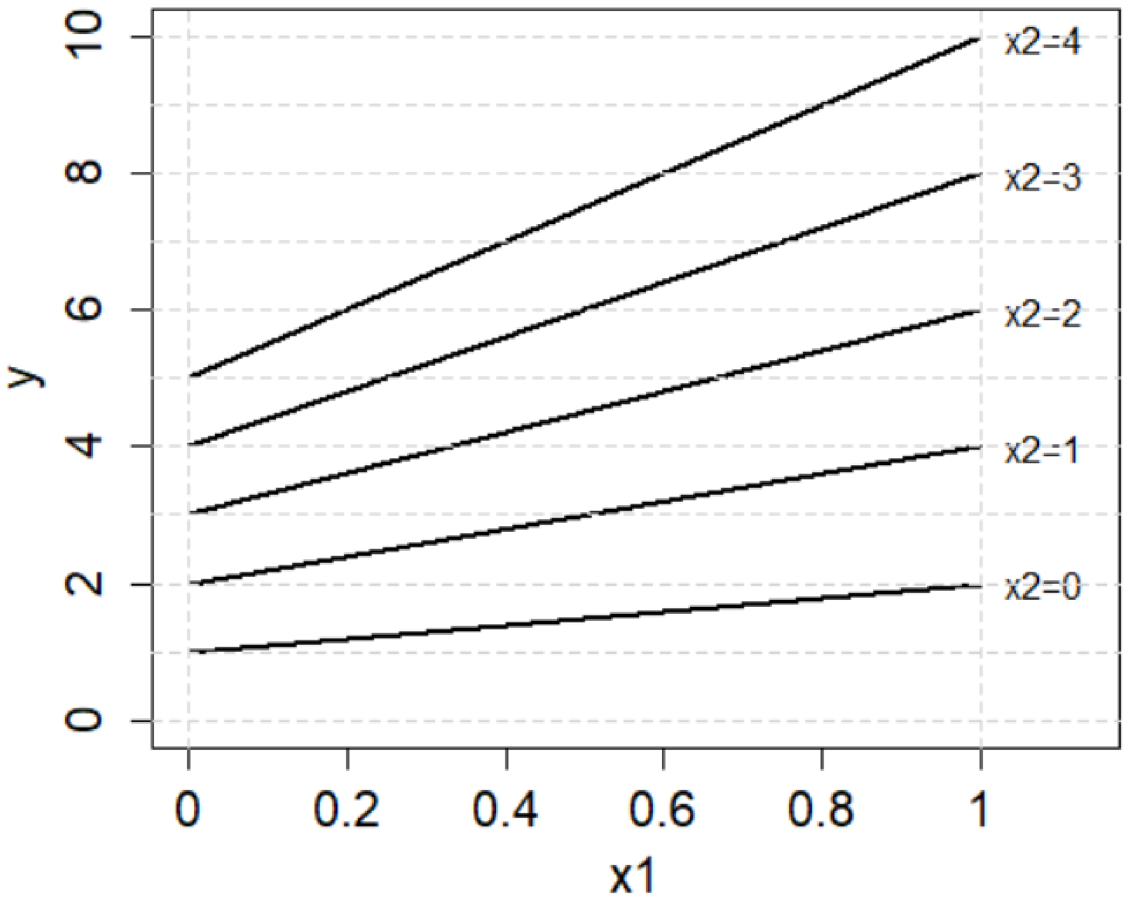
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$$\left\{ \begin{matrix} y \text{ continuous} \\ x_1 \text{ continuous} \\ x_2 \text{ continuous} \end{matrix} \right\}$$

- 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, when $x_2 = 0$
- β_2 is the expected change in y for every 1 unit increase in x_2, when $x_1 = 0$
- β_3 doesn't have a very intuitive interpretation. Think of it as the added effect (in addition to β_1 and β_2) when x_1 and x_2 both increase by 1 unit

It's a good idea to center x_1 and x_2 so that the value of 0 refers to the average value

Coefficient Interpretation: 2 Continuous Predictors & Interaction Effect



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.

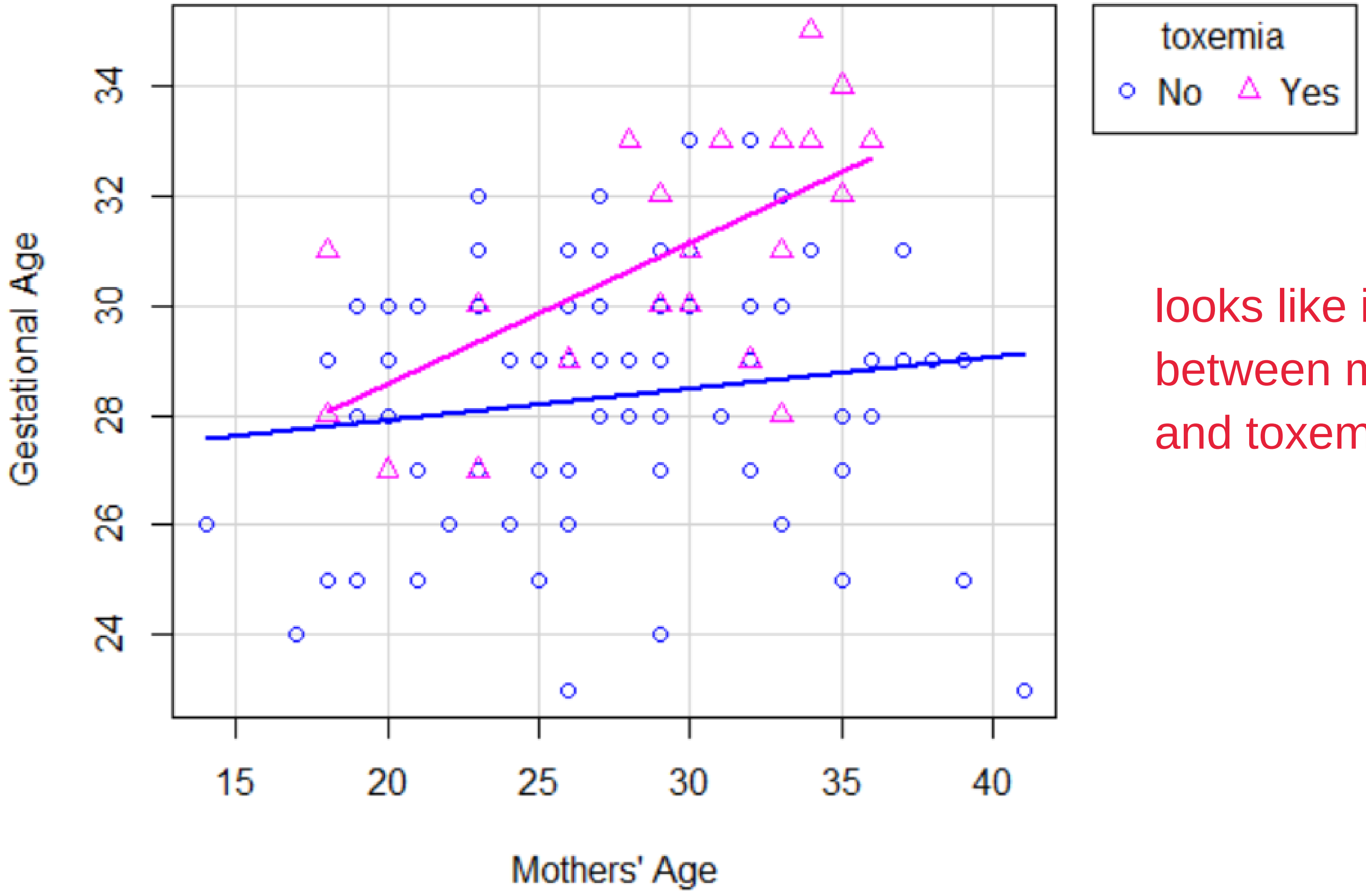
Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: Gestational Age

- Fit a linear regression model to predict the gestational age for low birth weight infants using mothers’ age, toxemia diagnosis, and their interaction as predictors.
- Write down the regression equation and interpret the coefficients.

y - gestational age
x1=mother's age (continuous, centered)
x2=toxemia

Example: Gestational Age



looks like interaction between mother's age and toxemia diagnosis

Example: Gestational Age

Summary of linear regression model for gestational age:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             28.37457    0.25103 113.033  < 2e-16
momage_ctr               0.05750    0.04166   1.380 0.170736
toxemia[T.Yes]           2.19180    0.55924   3.919 0.000167
momage_ctr:toxemia[T.Yes] 0.19932   0.09747   2.045 0.043598
```

$$\hat{y} = 28.37 + 0.06x_1 + 2.19x_2 + 0.20x_1x_2$$

Average mothers' age = 27.73 years

- $\hat{\beta}_0 = 28.37$: The average gestational age for a low birth weight infant whose mother is 27.73 years old and does not have toxemia is 28.37 weeks.
- $\hat{\beta}_1 = 0.06$: On average, every one year increase in mothers' age is associated with a 0.06 week increase in gestational age when the mother does not have a toxemia diagnosis.
- $\hat{\beta}_2 = 2.19$: For mothers who are 27.73 years old, we expect the gestational age to increase by 2.19 weeks if the mother has a toxemia diagnosis.
- $\hat{\beta}_1 + \hat{\beta}_3 = 0.06 + 0.20 = 0.26$: On average, every one year increase in mothers' age is associated with a 0.26 week increase in gestational age when the mother has a toxemia diagnosis.

Example: Gestational Age

What happens if we don’t center mothers’ age before fitting the model?

Model 1: using centered mothers’ age:

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 28.37457    0.25103 113.033  < 2e-16
momage_ctr                   0.05750    0.04166   1.380 0.170736
toxemia[T.Yes]               2.19180    0.55924   3.919 0.000167
momage_ctr:toxemia[T.Yes]    0.19932    0.09747   2.045 0.043598
```

Model 2: using uncentered mothers’ age:

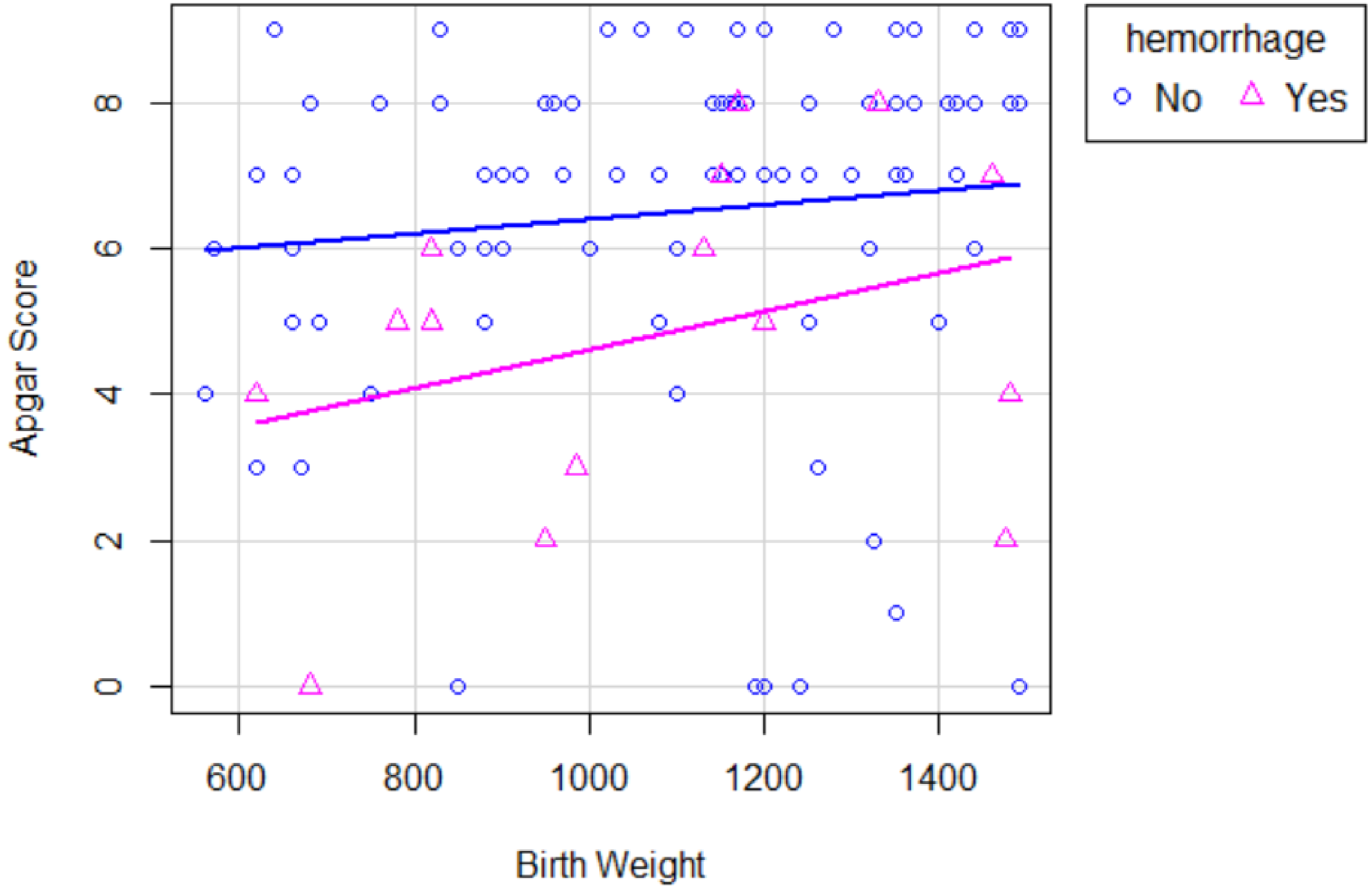
```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             26.78012    1.16786  22.931   <2e-16
momage                   0.05750    0.04166   1.380   0.1707
toxemia[T.Yes]          -3.33535    2.85514  -1.168   0.2456
momage:toxemia[T.Yes]    0.19932    0.09747   2.045   0.0436
```

- Main effect of toxemia changes (coefficient and significance)
- In Model 1, β_2 is the expected difference in gestational age for mothers with/without a toxemia diagnosis when they are 27.73 years old.
- In Model 2, β_2 is the expected difference in gestational age for mothers with/without a toxemia diagnosis when they are 0 years old.

Example: Apgar Score

- Does occurrence of germinal matrix hemorrhage modify the association between birth weight and Apgar score?

Example: Apgar Score



Example: Apgar Score

Summary of linear regression model for Apgar score:

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      6.5008735  0.2557276  25.421   <2e-16
birthwt_ctr                      0.0009838  0.0009601   1.025   0.3081
hemorrhage[T.Yes]               -1.6238370  0.6631600  -2.449   0.0162
birthwt_ctr:hemorrhage[T.Yes]    0.0016864  0.0023780   0.709   0.4799
```

$H_0: \beta_{\text{int}} = 0$ $H_A: \beta_{\text{int}} \neq 0$ p-value = 0.48

Since the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is not sufficient evidence to say that the association between birth weight and Apgar score is different for infants with and without germinal matrix hemorrhage.

Example: Apgar Score

- Since the interaction effect isn't significant, we should report results from a model without the interaction effect.

Summary of linear regression model for Apgar score (without interaction effect):

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.4994739  0.2550639  25.482   <2e-16
birthwt_ctr        0.0012587  0.0008761   1.437   0.1540
hemorrhage[T.Yes] -1.6631592  0.6591426  -2.523   0.0133
```

$$\hat{y} = 6.50 + 0.001(\text{birthwt\_ctr}) - 1.66(\text{hemorrhage})$$

Important Points

- Concept of an interaction
- What it means to center a variable and when/why it's helpful
- Interpretation of linear regression model with an interaction effect