

# Lesson 5

## Multiple Linear Regression

BIS 505b

Yale University  
Department of Biostatistics

Pagano Chapter 19

Date Modified: 3/21/2021

# Goals for this Lesson

## Addressing a Research Question

- ① How to describe the **linear relationship between continuous variables** when there is more than one explanatory variable of interest
- ② How to **include binary and categorical predictors**
- ③ How to **explore interactions between predictors** in their effect on the response
- ④ How to **evaluate the fit** of a multiple linear regression model
- ⑤ How to **select a final model**

# Contents

- 1 Multiple Linear Regression
  - Motivation
  - The Model
- 2 Inference
  - Confidence Interval and Hypothesis Test for  $\beta_j$
  - Overall  $F$ -Test
  - Partial  $F$ -Test
- 3 Regressors
  - Indicator Variables
  - Categorical Variables with More than Two Categories
  - Interaction Terms
- 4 Model Selection
  - Checking Assumptions
  - Model Selection

# Progress this Unit

- 1 Multiple Linear Regression
  - Motivation
  - The Model
- 2 Inference
  - Confidence Interval and Hypothesis Test for  $\beta_j$
  - Overall  $F$ -Test
  - Partial  $F$ -Test
- 3 Regressors
  - Indicator Variables
  - Categorical Variables with More than Two Categories
  - Interaction Terms
- 4 Model Selection
  - Checking Assumptions
  - Model Selection

# Multiple Linear Regression

- In Lesson 4, we used **simple linear regression** to explore the nature of the relationship between two continuous variables
- If knowing the value of a **single** explanatory variable improves our ability to predict a continuous response, we might suspect that information about **additional explanatory variables** could also be used to our advantage in **multiple linear regression**

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

# Multiple Predictors

- Purpose of multiple linear regression:
  1. **Simultaneously consider the relationships** of multiple variables with a continuous outcome
  2. Investigate relationship between explanatory (exposure) variable and a continuous outcome while **controlling for confounding**
- The effect of each independent variable is **adjusted for** all the other independent variables in the model

## Least Squares Fitted Regression Equation

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

# Statistical Regression Model

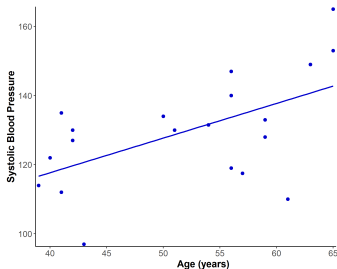
## Multiple Linear Regression Model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

- $y$  is the dependent variable
- $x_1, x_2, \dots, x_k$  are the  $k$  independent variables
- $\alpha$ : Mean value of the response when all  $k$  explanatory variables = 0
- $\beta_1, \dots, \beta_k$ : Slope parameters
  - $\beta_j$ : Change in mean response that corresponds to a one-unit increase in  $x_j$ , given that all other explanatory variables remain constant (*partial regression coefficients*)
- $\epsilon$  is the random error  $\sim N(0, \sigma_{y|x_1, \dots, x_k})$

# Multiple Linear Regression: Example

- **Example:** We previously found a significant positive linear relationship between systolic blood pressure and **age** using a sample of the Framingham data
- Given that **age** ( $x_1$ ) has already been accounted for, does the individual's systolic BP ( $y$ ) also depend on his **heart rate** ( $x_2$ )?



**Table:** Least squares regression results

Parameter	Estimate	SE	$t$	$p$ -value
Intercept	77.55	18.41	4.21	0.0005
Age	1.003	0.35	2.87	0.0101

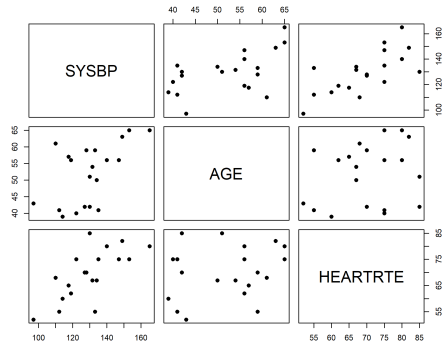
$$R^2 = 0.3144$$



# Data Structure: Example

$y_i$ (SYSBP)	$x_{1i}$ (AGE)	$x_{2i}$ (HR)
114	39	60
122	40	75
135	41	75
112	41	55
130	42	85
127	42	70
97	43	52
134	50	67
130	51	85
131.5	54	67
140	56	75
147	56	90
119	56	62
117.5	57	65
128	59	70
133	59	55
110	61	68
149	63	82
165	65	80
153	65	75

Figure: Scatterplot matrix



## Examining Bivariate Relationships: Example

- Given two regressors, can examine the linear associations of both variables with the response using a correlation matrix
- Scatterplots provide a simple description of the bivariate relationship
- However, considering bivariate relationships separately does not provide a complete picture
- Ignores the potential relationship between age and heart rate and how both variables together have a relationship with systolic BP

### R Code, Correlation Matrix

```
> vars <- c("SYSBP", "AGE", "HEARTRTE")  
> cor(fhssrs[,vars], method = "pearson", use = "pairwise.complete.obs")
```

	SYSBP	AGE	HEARTRTE
SYSBP	1.0000000	0.5607182	0.6595666
AGE	0.5607182	1.0000000	0.2246027
HEARTRTE	0.6595666	0.2246027	1.0000000

# Least Squares Regression Equation: Example

**Table:** Least squares regression results

Parameter	Estimate	SE	<i>t</i>	<i>p</i> -value
Intercept	25.116	20.246	1.24	0.232
Age	0.777	0.276	2.81	0.012
Heart Rate	0.915	0.252	3.64	0.002

$$\hat{y} = 25.116 + 0.777x_1 + 0.915x_2$$

$$\widehat{\text{SYSBP}} = 25.116 + 0.777 \text{ Age} + 0.915 \text{ HR}$$

- **Interpretation:**

- $a$ : Mean value of systolic BP when age = 0 and heart rate = 0 (?)
- $b_1$ : Holding HR constant, a one-year increase in age is associated with a 0.777 average increase in systolic BP
- $b_2$ : Holding age constant, a one-BPM increase in HR is associated with a 0.915 average increase in systolic BP

# MLR: Example

## R Code, MLR

```
> mod.mlr <- lm(SYSBP ~ AGE + HEARTRTE, data = fhssrs)
```

```
> summary(mod.mlr)
```

Call:

```
lm(formula = SYSBP ~ AGE + HEARTRTE, data = fhssrs)
```

Residuals:

Min	1Q	Median	3Q	Max
-24.727	-6.515	1.519	8.742	16.187

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	25.1165	20.2462	1.241	0.23162
AGE	0.7771	0.2765	2.811	0.01203
HEARTRTE	0.9148	0.2516	3.636	0.00204

Residual standard error: 10.63 on 17 degrees of freedom

Multiple R-squared: 0.6143, Adjusted R-squared: 0.5689

F-statistic: 13.54 on 2 and 17 DF, p-value: 0.0003042

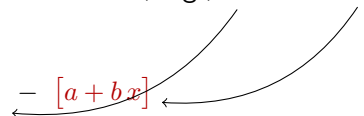
- $\hat{y} = 25.116 + 0.777 x_1 + 0.915 x_2$
- $s_{y|x_1, x_2} = 10.63$
- $n - p = 17$

# Interpretation of Coefficients: SLR vs. MLR

- The estimated coefficient of **age** in MLR is slightly different than it was when age was the only explanatory variable in the model
  - SLR:  $\hat{y} = 77.550 + 1.003 \text{ Age}$
  - MLR:  $\hat{y} = 25.116 + 0.777 \text{ Age} + 0.915 \text{ HR}$
- The interpretation of a coefficient depends on which additional explanatory variables are included in the model
- Coefficients in a multiple regression account for the effects of the *other* variables in the model
- Interpretation of the effect of one covariate is stated in terms of holding all other covariates **constant**

## Expected Impact on $y$ of $c$ -unit Increase in $x$

- Determine the expected impact on  $y$  for a increase in  $x$  that is greater than 1-unit by subtracting the fitted equations ( $\hat{y}$ ) under the two conditions, e.g.,  $x + c$  vs.  $x$ 
  - $\hat{y} = a + bx$


$$\begin{aligned}\hat{y}_1 - \hat{y}_0 &= [a + b(x + c)] - [a + bx] \\ &= a + bx + cb - a - bx \\ &= cb\end{aligned}$$

- The expected change in  $y$  that corresponds to a 10-unit increase in  $x$  is equal to  $10 \times \text{slope}$ ,  $10b$

## Expected Impact on $y$ of $c$ -unit Change in $x$ : Example

$$\hat{y} = 77.5 + 1.003 \text{ Age}$$

- When age =  $x + 10$ , the fitted equation is:  $\hat{y}_1 = 77.5 + 1.003(x + 10)$
- When age =  $x$ , the fitted equation is:  $\hat{y}_0 = 77.5 + 1.003x$

$$\begin{aligned}\hat{y}_1 - \hat{y}_0 &= [77.5 + 1.003(x + 10)] - [77.5 + 1.003x] \\ &= 77.5 + 1.003x + 10(1.003) - 77.5 - 1.003x \\ &= 10(1.003) = 10.03\end{aligned}$$

- The expected change in systolic BP that corresponds to a 10-unit change in age is equal to  $10 \times \text{slope}$ , 10.03



# Variance about Regression, $s_{y|x_1, \dots, x_k}^2$

- $\sigma_{y|x_1, \dots, x_k}^2$  is estimated by  $s_{y|x_1, \dots, x_k}^2$
- As before, estimated using residuals,  $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$

	Simple linear regression	Multiple linear regression
Fitted equation	$\hat{y} = a + b x$	$\hat{y} = a + b_1 x_1 + \dots + b_k x_k$
Explanatory variables ( $k$ )	1	$k$
Coefficients estimated ( $p$ )	2	$p = k + 1$
$s_{y x}^2$	$\frac{SSE}{n - 2}$	$\frac{SSE}{n - p}$

- Divide  $SSE$  by  $n - p$  since  $p$  parameters are estimated in  $\hat{y}$  (intercept +  $k$  explanatory variables)



# Variance about Regression: Example

## R Code, ANOVA Table

```
# ANOVA table of fitted model
```

```
> anova(mod.mlr)
```

Analysis of Variance Table

Response: SYSBP

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
AGE	1	1565.0	1565.01	13.857	0.001693
HEARTRTE	1	1492.8	1492.75	13.217	0.002045
Residuals	17	1919.9	112.94		

$$\begin{aligned} \bullet s_{y|x_1, x_2}^2 &= \frac{SSE}{n-3} \\ &= \frac{1919.9}{17} \\ &= 112.94 \quad \text{Mean Sq Residuals} \end{aligned}$$

$$\begin{aligned} \bullet s_{y|x_1, x_2} &= \sqrt{\frac{SSE}{n-3}} \\ &= \sqrt{112.94} \\ &= 10.63 \quad \text{Residual SE} \end{aligned}$$

# Summary: Least Squares Estimation

## Population

$$\mu_{y|x_1, x_2} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Unknown  
parameters:

Regression  
coefficients

$$\text{Var}(Y|x_1, x_2) = \sigma_{y|x_1, x_2}^2$$

Variance about  
regression

## Estimates from Sample

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

$$\hat{\text{Var}}(Y|x_1, x_2) = s_{y|x_1, x_2}^2 = \frac{SSE}{n - 3} = MSE$$

- Least squares estimators  $a$ ,  $b_1$ ,  $b_2$  are chosen to minimize  $SSE$ , sum of squared residuals (residuals  $e_i = y_i - \hat{y}_i$ )

$n - p$  (# regression coefficients)

## Progress this Unit

- 1 Multiple Linear Regression
  - Motivation
  - The Model
- 2 Inference
  - Confidence Interval and Hypothesis Test for  $\beta_j$
  - Overall  $F$ -Test
  - Partial  $F$ -Test
- 3 Regressors
  - Indicator Variables
  - Categorical Variables with More than Two Categories
  - Interaction Terms
- 4 Model Selection
  - Checking Assumptions
  - Model Selection

# Tests of Interest

- There are three main tests of inference:
  - 1 Individual  $t$ -tests for specific predictors
    - $H_0 : \beta_j = 0$
  - 2 Overall  $F$ -test for testing if *any* of the predictors is useful
    - $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$
  - 3 Partial  $F$ -test for testing if *group/subset* of covariates is useful
    - For example,  $H_0 : \beta_3 = \beta_4 = 0$

## Inference on the Parameters, $\beta_j$

- Once again, we can use estimated **standard errors** of the regression coefficients to:
  1. Construct confidence intervals for  $\beta_j$
  2. Perform hypothesis tests for  $\beta_j$
- Inference for  $\beta_j$  is based on the  $t$ -distribution:

$$T = \frac{b_j - \beta_j}{s_{b_j}} \sim t_{n-p}$$

- Where  $p$  = Number of regression coefficients estimated =  $k + 1$
- Recall,  $\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$

Confidence Interval for  $\beta_j$ : ExampleConfidence Interval for  $\beta_j$ :

Parameter	100(1 - $\alpha$ )% CI
$\beta_j$	$b_j \pm t_{n-p, 1-\frac{\alpha}{2}} s b_j$

## R Code, CI for Parameters

```
# 95% CIs for regression parameters
> confint(mod.mlr)
```

	2.5 %	97.5 %
(Intercept)	-17.5993455	67.832251
AGE	0.1938239	1.360444
HEARTRTE	0.3839153	1.445667

- 95% CI for  $\beta_1$  (Age): (0.19, 1.36)
- 95% CI for  $\beta_2$  (HRte): (0.38, 1.45)

## R Code, CI for Parameters

```
> summary(mod.mlr)
```

Call:

```
lm(formula = SYSBP ~ AGE + HEARTRTE, data = fhssrs)
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	25.1165	20.2462	1.241	0.23162
AGE	0.7771	0.2765	2.811	0.01203
HEARTRTE	0.9148	0.2516	3.636	0.00204

```
# CI for beta1 (AGE)
```

```
> 0.7771 - qt(.975, df = 17)*0.2765
```

```
[1] 0.193736
```

```
> 0.7771 + qt(.975, df = 17)*0.2765
```

```
[1] 1.360464
```

# Hypothesis Test for $\beta_j$

## Hypothesis Test for $\beta_j$ :

- $H_0 : \beta_j = 0$
- $H_1 : \beta_j \neq 0$

### Test Statistic for $H_0 : \beta_j = 0$

$$t = \frac{b_j}{s_{b_j}}$$

- Test statistics are compared to a  $t$ -distribution with  $n - p$   $df$
- Under  $H_0 : \beta_j = 0$ , there is no linear relationship between the predictor variable  $x_j$  and the response variable after controlling for the effects of all other predictor variables in the model

# Hypothesis Test for $\beta_j$ : Example

Table: Least squares regression results

Parameter	Estimate	SE	$t$	$p$ -value
Intercept	25.116	20.246	1.24	0.232
Age	0.777	0.276	2.81	0.012
Heart Rate	0.915	0.252	3.64	0.002

- **Example:** Is there evidence of a linear relationship between systolic blood pressure and age given HR is included in the model?

- **Step 1:** State the hypotheses

- $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$

- **Step 2:** Specify significance level

- $\alpha = 0.05$

- **Step 3:** Compute the appropriate test statistic

- Age:  $t = \frac{b_1}{s_{b_1}} = \frac{0.777}{0.276} = 2.81$

$$T \sim t_{n-p} = t_{20-3}$$



## Hypothesis Test for $\beta_j$ : Example

- **Step 4:** Generate the decision rule

$$t_{.975} = \text{qt}(1 - .05/2, \text{df} = 20 - 3)$$

- Reject  $H_0$  if  $|t| \geq t_{n-3, 1-\frac{\alpha}{2}} = t_{17, .975} = t^* = 2.110$

- **Step 5:** Draw a conclusion about  $H_0$

$$\text{pval} = 2 * (1 - \text{pt}(2.81, \text{df} = 20 - 3))$$

- $t = 2.81$

$$\bullet p = 2 \times P(T \geq 2.81) = 0.012$$

- $|t| \geq t^* \rightarrow \text{Reject } H_0$

$$\bullet p \leq 0.05 \rightarrow \text{Reject } H_0$$

- **Conclusion:** The data provide evidence of a linear association between age and systolic blood pressure ( $b_1 = 0.777, p = 0.012$ ) after heart rate is controlled for/included in the model

## Hypothesis Test for $\beta_j$ : Example

**Table:** Least squares regression results

Parameter	Estimate	SE	$t$	$p$ -value
Intercept	25.116	20.246	1.24	0.232
Age	0.777	0.276	2.81	0.012
Heart Rate	0.915	0.252	3.64	0.002

- Table of least squares regression results includes the estimated slopes, standard errors,  $t$ -statistics, and  $p$ -values for each explanatory variable
- Looking at the relationship between *heart rate* and systolic blood pressure when age is included in the model, we see a significant positive linear association ( $b_2 = 0.915, p = 0.002$ )
- **Interpretation:** On average, systolic blood pressure increases as either age or heart rate increases

## ANOVA Table for Multiple Linear Regression

Table: ANOVA Table for Linear Regression with  $k$  Explanatory Variables

Source of Variation	Sum of Squares ( $SS$ )	Degrees of Freedom ( $df$ )	Mean Squares ( $MS$ )	F
Model	$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$k^*$	$MSM = \frac{SSM}{df_1}$	$\frac{MSM}{MSE}$
Error	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - p^\dagger$	$MSE = \frac{SSE}{df_2}$	
Total	$SST = \sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$		

\* $df_1$  : Number of explanatory variables,  $k = p - 1$

$^\dagger df_2$  : Sample size – number of estimated parameters =  $n - p$

- In multiple regression, can again decompose the total sum of squares  $SST$  into  $SSM$  and  $SSE$

- $s_y^2 = \frac{SST}{n - 1}$

## Overall $F$ -Test

- Overall  $F$ -Test (a.k.a., “Global  $F$ -test” or “Omnibus  $F$ -test”) tests whether the explanatory variables collectively have an effect on the response variable
  - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
  - $H_1$ : At least one  $\beta_j \neq 0$  for  $j = 1, \dots, k$
  - For MLR,  $F \sim F(df_1 = k, df_2 = n - p)$  under  $H_0$
  - Reject  $H_0$  if  $F \geq F_{1-\alpha}(df_1, df_2)$
- A significant  $F$ -test ( $p\text{-value} \leq 0.05$ ) does not necessarily mean the model fits the data well; means at least one of the  $\beta$ s is non-zero

### ANOVA Overall $F$ -Test Statistic

$$F = \frac{MSM}{MSE}$$

## Overall $F$ -Test

- Compare a **full model** (all predictors included:  $x_1, x_2, x_3, \dots, x_k$ ) vs. **reduced model** (model under  $H_0$  i.e., no predictors included)
  - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ 
    - This model is known as the “null model” and includes no predictors  $x$
  - $H_1$ : At least one  $\beta_j \neq 0$  for  $j = 1, \dots, k$
- **Full model:**  $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$
- **Reduced model:**  $y = \alpha + \epsilon$
- Compare “**nested models**”: Reduced model is a subset of the full model

Overall  $F$ -Test: Example

## R Code, Overall F-Test

```
> summary(mod.mlr)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.1165    20.2462   1.241  0.23162
AGE           0.7771     0.2765   2.811  0.01203
HEARTRTE      0.9148     0.2516   3.636  0.00204
Residual standard error: 10.63 on 17 degrees of freedom
Multiple R-squared:  0.6143, Adjusted R-squared:  0.5689
F-statistic: 13.54 on 2 and 17 DF,  p-value: 0.0003042

# Null Model (intercept only)
> mod.null <- lm(SYSBP ~ 1, data = fhssrs)
# Full Model
> mod.mlr <- lm(SYSBP ~ AGE + HEARTRTE, data = fhssrs)
# Testing H0: beta1 = beta2 = 0
> anova(mod.null, mod.mlr)

Analysis of Variance Table
Model 1: SYSBP ~ 1
Model 2: SYSBP ~ AGE + HEARTRTE
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      19 4977.7
2      17 1919.9  2      3057.8 13.537 0.0003042
```

	$SS$	$df$	$MS$	$F$
Model	3057.8	2	$\frac{3057.8}{2} = 1528.88$	13.54
Error	1919.9	17	$\frac{1919.9}{17} = 112.94$	
Total	4977.7	19	$F \sim F(2, 17)$	

- $H_0 : \beta_1 = \beta_2 = 0$  vs.  $H_1 : \text{At least one } \beta_j \neq 0$

- $F = \frac{MSM}{MSE} = \frac{1528.88}{112.94} = 13.54 > 3.59$

- $p = P(F \geq 13.54) = 0.0003 < 0.05$

- **Conclusion:** Reject  $H_0$ . Evidence to conclude there is a linear relationship between the response and at least one explanatory variable.

$$3.59 = f_{95} = \text{qf}(.95, df1=2, df2=17)$$

$$0.0003 = pval = 1 - \text{pf}(13.54, df1=2, df2=17)$$

# $R^2$ and $R_a^2$

- **Coefficient of Determination ( $R^2$ )**: Proportion of the total variation in the response variable that is explained by the model (i.e., explanatory variables); including more predictors in model increases  $R^2$
- **Adjusted  $R^2$  ( $R_a^2$ )**: Compensates for added complexity of a larger model by accounting for the number of predictors in the model (not directly interpreted as the proportion of the variability in  $y$  that is explained by the regression model)

## Coefficient of Determination

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

## Adjusted $R^2$

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

Model	$R^2$	$R_a^2$
Age	$0.3144 = \frac{1565.0}{4977.7} = 1 - \frac{3412.7}{4977.7}$	$0.2763 = 1 - \frac{3412.7/18}{4977.7/19}$
Age, Heart Rate	$0.6143 = \frac{3057.8}{4977.7} = 1 - \frac{1919.9}{4977.7}$	$0.5689 = 1 - \frac{1919.9/17}{4977.7/19}$



## Partial $F$ -Test

- **Partial  $F$ -Test** is used to test if a subset of covariates (e.g.,  $x_3, x_4$ ) is useful
- Compare a **full model** (all predictors included:  $x_1, x_2, x_3, x_4$ ) vs. **reduced model** (model under  $H_0$  i.e.,  $x_3$  and  $x_4$  not included)
  - $H_0: \beta_3 = \beta_4 = 0$ 
    - The reduced model,  $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ , is sufficient
  - $H_1: \beta_3$  and  $\beta_4$  not both 0
- **Full model:**  $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
- **Reduced model:**  $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$



# Partial $F$ -Test

## ANOVA Partial $F$ -Test Statistic

$$F_0 = \frac{\frac{SSM(F) - SSM(R)}{\text{Number parameters tested under } H_0}}{\frac{SSE(F)}{df_2(F)}}$$

- Full model  $MSE$  is in the denominator
- $F_0 \sim F(df_1 = \text{Number of parameters tested under } H_0, df_2 = n - p)$  under  $H_0$
- Reject  $H_0$  if  $F_0 \geq F_{1-\alpha}(df_1, df_2)$

# Progress this Unit

- 1 Multiple Linear Regression
  - Motivation
  - The Model
- 2 Inference
  - Confidence Interval and Hypothesis Test for  $\beta_j$
  - Overall  $F$ -Test
  - Partial  $F$ -Test
- 3 Regressors
  - Indicator Variables
  - Categorical Variables with More than Two Categories
  - Interaction Terms
- 4 Model Selection
  - Checking Assumptions
  - Model Selection

# Binary Predictor

- So far, have only considered **quantitative predictor variables**
- Can include **categorical predictors** by constructing artificial variables known as **dummy variables** or **indicator variables**
- R automatically creates dummy variables for categorical variables when using **factor** variables (be sure to create *un-ordered* factors)
- For example, consider the **binary predictor**, Sex

$$z_1 = \begin{cases} 1 & \text{Male} \\ 0 & \text{Female} \end{cases}$$

## R Code, Summarizing Binary Predictor

```
> table(fhssrs$SEX_factor)
Female    Male
      7      13

> prop.table(table(fhssrs$SEX_factor))
Female    Male
 0.35    0.65
```

## Binary Predictor in SLR

- **Question:** Is there a relationship between sex and systolic BP?
  - How would you answer this question without using a regression model?

$$z_1 = \begin{cases} 1 & \text{Male} \\ 0 & \text{Female} \end{cases}$$

- Choose one category as the default (*reference*) category (=0, Females)
- A simple model of  $y$ , systolic blood pressure, with 1 dummy variable:

$$\mu_{y|z} = \alpha + \beta_1 z_1$$

# Binary Predictor in SLR

$$\mu_{y|z} = \alpha + \beta_1 z_1$$

- Regression line for **males** ( $z_1 = 1$ ):

$$\mu_{y|z_1=1} = \alpha + \beta_1(1) = \alpha + \beta_1$$

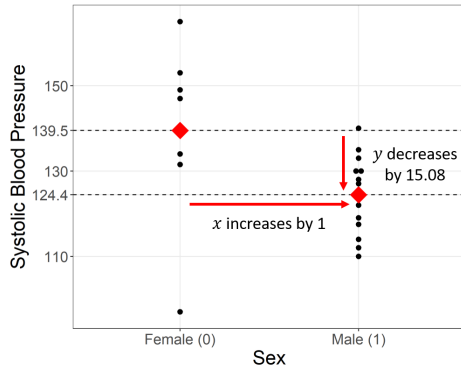
- $\alpha + \beta_1$ : Mean systolic BP in males

- Regression line for **females** ( $z_1 = 0$ ):

$$\mu_{y|z_1=0} = \alpha + \beta_1(0) = \alpha$$

- $\alpha$ : Mean systolic BP in females
- $\beta_1$ : Difference in mean systolic BP between males and females (*reference*)
  - For a one-unit increase in  $z_1$ , there is a  $\beta_1$  change in the mean of systolic BP

# Binary Predictor in SLR: Example



- Level 1 (males) have average systolic BP that is 15.08 mmHg **less** than level 0 (females)
- Slope  $b_1 = -15.08$

**Table:** Summary statistics of systolic blood pressure in the sample by sex

Sex	$n$	Mean	Difference
Male (1)	13	124.42	<b>-15.08</b>
Female (0)	7	139.50	

# Binary Predictor in SLR: Example

## R Code, Binary Predictor

```
> mod.sex <- lm(SYSBP ~ SEX_factor, data = fhssrs)
> summary(mod.sex)

Call:
lm(formula = SYSBP ~ SEX_factor, data = fhssrs)

Residuals:
    Min       1Q   Median       3Q      Max
-42.500  -7.192   3.077   8.808  25.500

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    139.500     5.594   24.936 2.07e-15
SEX_factorMale  -15.077     6.939   -2.173  0.0434

Residual standard error: 14.8 on 18 degrees of freedom
Multiple R-squared:  0.2078, Adjusted R-squared:  0.1638
F-statistic: 4.721 on 1 and 18 DF,  p-value: 0.04339
```

$$\hat{y} = 139.5 - 15.08 \text{ Sex}$$

- Test of  $H_0 : \beta_1 = 0$ : Reject  $H_0$ ,  $p = 0.0434$
- We have evidence of a significant difference in systolic blood pressure in males vs. females. On average, males have a systolic BP that is 15.08 mmHg *less* than females (*reference*).
- Equivalent to 2-sample pooled  $t$ -test



## Binary Predictor in MLR

$$\mu_{y|\mathbf{x}} = \alpha + \beta_1 x_1 + \beta_2 z_1$$

- Model with 1 continuous covariate, heart rate ( $x_1$ ), and 1 dummy variable, sex ( $z_1$ )
- Regression line for **males** ( $z_1 = 1$ ):

$$\mu_{y|x_1, z_1=1} = \alpha + \beta_1 x_1 + \beta_2(1) = (\alpha + \beta_2) + \beta_1 x_1$$

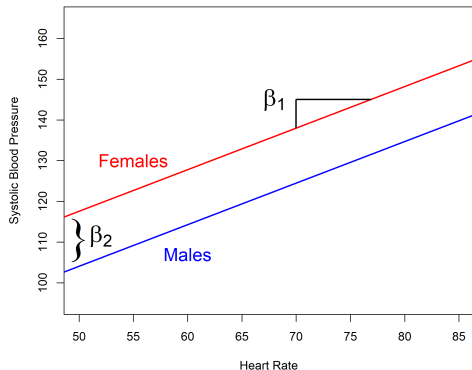
- Regression line for **females** ( $z_1 = 0$ ):

$$\mu_{y|x_1, z_1=0} = \alpha + \beta_1 x_1 + \beta_2(0) = \alpha + \beta_1 x_1$$

- Same slope ( $\beta_1$ ); male intercept shifted by  $\beta_2$  (intercept for males:  $\alpha + \beta_2$ )
  - $\beta_2$ : Slope for the binary dummy variable ( $z_1$ ) is the average difference in systolic BP for the males compared to females (reference), adjusted for heart rate
  - $\alpha$ : Mean systolic BP in females with a heart rate = 0



## Binary Predictor in MLR: Example



$$\mu_{y|x_1, x_2} = \alpha + \beta_1 \text{HR} + \beta_2 \text{Sex}$$

- Test if the slope of the line(s) = 0 (linear association between *HR* and systolic BP, adjusting for sex)
  - $H_0 : \beta_1 = 0$
- Test if there are two separate parallel lines for the two sexes, or if one is sufficient to describe the data (difference in sexes, adjusting for HR)
  - $H_0 : \beta_2 = 0$

# Binary Predictor in MLR: Example

## R Code, Binary Predictor

```
> mod2 <- lm(SYBP ~ HEARTRTE + SEX_factor, data = fhssrs)
> summary(mod2)
```

Call:

```
lm(formula = SYBP ~ HEARTRTE + SEX_factor, data = fhssrs)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.9044	-4.8031	0.8015	3.9004	23.5380

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	66.6744	18.2550	3.652	0.001971
HEARTRTE	1.0237	0.2501	4.093	0.000758
SEX_factorMale	-13.5133	5.0816	-2.659	0.016519

Residual standard error: 10.81 on 17 degrees of freedom

Multiple R-squared: 0.601, Adjusted R-squared: 0.5541

F-statistic: 12.8 on 2 and 17 DF, p-value: 0.0004057

$$\hat{y} = 66.67 + 1.02 \text{ HR} - 13.51 \text{ Sex}$$

- Step 1:  $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$
- Step 2:  $\alpha = 0.05$
- Step 3: HR:  $t = \frac{b_1}{s_{b_1}} = \frac{1.02}{0.25} = 4.09$
- Step 4: Reject  $H_0$  if  $|t| \geq t_{n-3, 1-\frac{\alpha}{2}}$   
 $t_{17, .975} = t^* = 2.110$
- Step 5:  $|t| \geq t^* \rightarrow \text{Reject } H_0$   
 $p = 0.0008 \leq 0.05 \rightarrow \text{Reject } H_0$
- Conclusion: There is a linear association between *HR* and *SBP* after controlling for sex. After adjusting for sex, a one-unit increase in *HR* increases *SBP* by 1.02 mmHg, on average.

# Binary Predictor in MLR: Example

## R Code, Binary Predictor

```
> mod2 <- lm(SYBP ~ HEARTRTE + SEX_factor, data = fhssrs)
> summary(mod2)
```

Call:

```
lm(formula = SYBP ~ HEARTRTE + SEX_factor, data = fhssrs)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.9044	-4.8031	0.8015	3.9004	23.5380

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	66.6744	18.2550	3.652	0.001971
HEARTRTE	1.0237	0.2501	4.093	0.000758
SEX_factorMale	-13.5133	5.0816	-2.659	0.016519

Residual standard error: 10.81 on 17 degrees of freedom

Multiple R-squared: 0.601, Adjusted R-squared: 0.5541

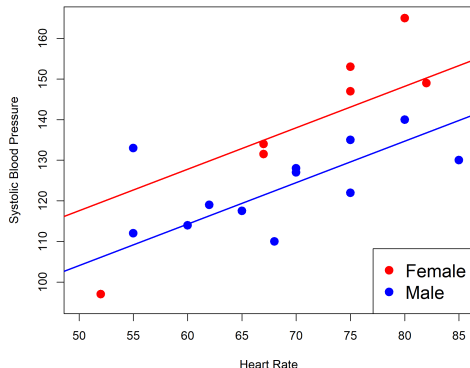
F-statistic: 12.8 on 2 and 17 DF, p-value: 0.0004057

$$\hat{y} = 66.67 + 1.02 \text{ HR} - 13.51 \text{ Sex}$$

- Step 1:  $H_0 : \beta_2 = 0$  vs.  $H_1 : \beta_2 \neq 0$
- Step 2:  $\alpha = 0.05$
- Step 3: Sex:  $t = \frac{-13.51}{5.08} = -2.66$
- Step 4: Reject  $H_0$  if  $|t| \geq t^* = 2.110$
- Step 5:  $|t| \geq t^* \rightarrow \text{Reject } H_0$   
 $p = 0.017 \leq 0.05 \rightarrow \text{Reject } H_0$
- Conclusion: There is a significant difference in average SBP in M and F after adjusting for HR. On average, M tend to have lower systolic blood pressure by 13.51 mmHg compared to F, after accounting for HR.

# Binary Predictor in MLR: Example

$$\hat{y} = 66.67 + 1.02 \text{ HR} - 13.51 \text{ Sex}$$



- **Females:** ( $z_1 = 0$ )

$$\begin{aligned}\hat{y} &= 66.67 + 1.02 \text{ HR} - 13.51(0) \\ &= 66.67 + 1.02 \text{ HR}\end{aligned}$$

- **Males:** ( $z_1 = 1$ )

$$\begin{aligned}\hat{y} &= 66.67 + 1.02 \text{ HR} - 13.51(1) \\ &= (66.67 - 13.51) + 1.02 \text{ HR} \\ &= 53.16 + 1.02 \text{ HR}\end{aligned}$$

- When MLR involves binary predictors, the slopes associated with the dummy variables provide **adjusted mean differences**, adjusting for all the other variables in the model

## Categorical Predictor with More than 2 Levels

- A **categorical variable with more than two categories**, or levels, can be represented as a set of indicator (dummy) variables
- If the categorical variable has  $C$  levels, then  $C - 1$  dummy variables are needed to represent the categorical variable in the model

Category	$z_1$	$z_2$
Young age group ( $< 45$ )	0	0
Middle age group ( $45 - 56$ )	1	0
Old age group ( $\geq 57$ )	0	1

- For a given individual, if know his values of  $z_1$  and  $z_2$ , know what age group he is in
- The youngest age group is the **reference group**

# Categorical Predictor in MLR

$$\mu_{y|\mathbf{x}} = \alpha + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2$$

- Model with 1 continuous covariate, heart rate ( $x_1$ ), and categorical variable, age group ( $z_1, z_2$ )
- Regression line for **young age group** ( $z_1 = 0, z_2 = 0$ ):  $\mu_{y|\mathbf{x}} = \alpha + \beta_1 x_1$
- Regression line for **middle age group** ( $z_1 = 1, z_2 = 0$ ):  $\mu_{y|\mathbf{x}} = (\alpha + \beta_2) + \beta_1 x_1$
- Regression line for **old age group** ( $z_1 = 0, z_2 = 1$ ):  $\mu_{y|\mathbf{x}} = (\alpha + \beta_3) + \beta_1 x_1$ 
  - $\beta_2$ : Slope for the middle age group dummy variable ( $z_1$ ) is the average difference in systolic BP for the middle age group compared to the youngest age group, adjusted for heart rate
  - $\beta_3$ : Slope for the old age group dummy variable ( $z_2$ ) is the average difference in systolic BP for the old age group compared to the youngest age group, adjusted for heart rate
  - $\alpha$ : Mean systolic BP in young age group when heart rate = 0

# Categorical Predictor in MLR: Example

## R Code, Categorical Predictor

```
> fhssrs$AGEGRP_factor <- factor(fhssrs$AGEGRP,  
  levels = 1:3, labels = c("<45", "45-56", "57+"))  
> levels(fhssrs$AGEGRP_factor)  
[1] "<45"    "45-56"  "57+"  
  
> mod3 <- lm(SYSBP ~ HEARTRTE + AGEGRP_factor, data = fhssrs)  
> summary(mod3)  
...  
Coefficients:  
                Estimate Std. Error t value Pr(>|t|)  
(Intercept)      53.9555    19.0212   2.837  0.01191  
HEARTRTE          0.9731     0.2745   3.545  0.00269  
AGEGRP_factor45-56  8.9146     6.6130   1.348  0.19642  
AGEGRP_factor57+  13.7312     6.2668   2.191  0.04359  
  
Residual standard error: 11.6 on 16 degrees of freedom  
Multiple R-squared:  0.5673, Adjusted R-squared:  0.4862  
F-statistic: 6.993 on 3 and 16 DF,  p-value: 0.003215
```

$$\hat{y} = 54.0 + 0.97 \text{ HR} \\ + 8.9 \text{ Age}_{45-56} + 13.7 \text{ Age}_{57+}$$

- $b_1 = 0.97$ : 1-unit increase in HR associated with 0.973-unit average increase in systolic BP, controlling for age group
  - Test of  $H_0 : \beta_1 = 0$ : Reject  $H_0$ ,  $p = 0.003$
  - Evidence of a significant association between HR and systolic BP, controlling for age group

# Categorical Predictor in MLR: Example

**Table:** Least squares regression results

Parameter	Estimate	SE	$t$	$p$ -value
Intercept	53.956	19.021	2.837	0.012
Heart Rate	0.973	0.275	3.545	0.003
Age 45-56	8.915	6.613	1.348	0.196
Age 57+	13.731	6.267	2.191	0.044

## R Code, Unadjusted Mean SYSBP

```
> aggregate(x = list(meansysbp = fhssrs$SYSBP),  
            by = list(agegrp = fhssrs$AGEGRP_factor),  
            FUN = mean, na.rm = TRUE)  
agegrp meansysbp  
1    <45    119.5714  
2   45-56    133.5833  
3    57+    136.5000
```

- $b_2 = 8.915$ : Heart rate-adjusted difference in mean systolic BP in 45-56 vs. <45 (*ref*)
  - Test of  $H_0 : \beta_2 = 0$ : Fail to reject  $H_0$ ,  $p = 0.196$
  - No evidence of a significant difference in systolic BP in these two groups, controlling for HR
- $b_3 = 13.731$ : Heart rate-adjusted difference in mean systolic BP in 57+ vs. <45 (*ref*)
  - Test of  $H_0 : \beta_3 = 0$ : Reject  $H_0$ ,  $p = 0.044$
  - Evidence of a significant difference in systolic BP in these two groups, controlling for HR



# Categorical Predictor in MLR: Example

## R Code, Changing Reference Category to 57+

```
> levels(fhssrs$AGEGRP_factor)
[1] "<45"    "45-56"  "57+"
# Change reference category to 57+ (use factor label in ref=)
> fhssrs$AGEGRpv2_factor <- relevel(fhssrs$AGEGRP_factor,
                                     ref = "57+")
> levels(fhssrs$AGEGRpv2_factor)
[1] "57+"    "<45"    "45-56"
> mod4 <- lm(SYSBP ~ HEARTRTE + AGEGRpv2_factor, data=fhssrs)
> summary(mod4)
...
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      67.6866    19.8999   3.401  0.00365
HEARTRTE          0.9731     0.2745   3.545  0.00269
AGEGRpv2_factor<45 -13.7312     6.2668  -2.191  0.04359
AGEGRpv2_factor45-56 -4.8166     6.4770  -0.744  0.46787
```

$$\hat{y} = 67.7 + 0.97 \text{ HR} \\ - 13.7 \text{ Age}_{<45} - 4.8 \text{ Age}_{45-56}$$

- $a = 67.7$ : mean systolic BP in old age group when HR = 0
- $b_1 = 0.97$ : 1-unit increase in HR associated with 0.97-unit increase in mean systolic BP, controlling for age group
- $b_2 = -13.7$ : Heart rate-adjusted difference in mean systolic BP in <45 vs. 57+ (ref)
- $b_3 = -4.8$ : Heart rate-adjusted difference in mean systolic BP in 45-56 vs. 57+ (ref)



## Partial $F$ -Test: Example

$$\mu_{y|\mathbf{x}} = \alpha + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2$$

- Would like to test if the **age group** variable is an important predictor in this model
- Since age group consists of two dummy variables,  $z_1$  and  $z_2$ , will perform a **Partial  $F$ -Test** to test the significance of a **group of parameters** (i.e.,  $\beta_2$  and  $\beta_3$ )
- Compare **full model** (all predictors included:  $x_1, z_1, z_2$ ) to **reduced model** (model under  $H_0$ : age group dummy variables not included)
  - $H_0$ :  $\beta_2 = \beta_3 = 0$
  - $H_1$ :  $\beta_2$  and  $\beta_3$  are not both 0
- **Full model:**  $y = \alpha + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + \epsilon$
- **Reduced model:**  $y = \alpha + \beta_1 x_1 + \epsilon$

# Partial $F$ -Test: Example

Table: Full Model: ANOVA Table

Source	$SS$	$df$
Model	2824.0	3
Error	2153.7	16

Table: Reduced Model: ANOVA Table

Source	$SS$	$df$
Model	2165.4	1
Error	2812.3	18

- **Example:** Is age group an important predictor in this model?
- **Step 1:** State the hypotheses
  - $H_0: \beta_2 = \beta_3 = 0$  vs.  
 $H_1: \beta_2$  and  $\beta_3$  not both 0
- **Step 2:** Specify significance level
  - $\alpha = 0.05$
- **Step 3:** Compute the appropriate test statistic

$$F_0 \sim F(2, 16)$$

$$\bullet F_0 = \frac{\frac{SSM(F) - SSM(R)}{\text{Number parameters tested}}}{\frac{SSE(F)}{df_2(F)}} = \frac{\frac{2824 - 2165.4}{2}}{\frac{2153.7}{16}} = \frac{329.3}{124.6} = 2.45$$

## Partial $F$ -Test: Example

- **Step 4:** Generate the decision rule `f95 = qf(.95, df1 = 2, df2 = 16)`
  - Reject  $H_0$  if  $F_0 \geq F_{1-\alpha}(2, 16) = F_{.95}(2, 16) = F^* = 3.634$
- **Step 5:** Draw a conclusion about  $H_0$  `pval = 1 - pf(2.45, df1 = 2, df2 = 16)`
  - $F_0 = 2.45$
  - $p = P(F \geq 2.45) = 0.12$
  - $F_0$  not  $\geq 3.634 \rightarrow$  Fail to reject  $H_0$
  - $p$  not  $\leq 0.05 \rightarrow$  Fail to reject  $H_0$
- **Conclusion:** Fail to reject the null hypothesis that  $\beta_2 = \beta_3 = 0$ . Thus, we cannot conclude age category is contributing significantly to this model containing heart rate.

# Partial $F$ -Test: Example

## R Code, Partial F-Test

```
# Full Model
> mod.full <- lm(SYSBP ~ HEARTRTE + AGEGRP_factor,
  data = fhssrs)
> # anova(mod.full)      # to get SSM(F), SSE(F) and df2
# Reduced Model
> mod.reduced <- lm(SYSBP ~ HEARTRTE,
  data = fhssrs)
> # anova(mod.reduced) # to get SSM(R)
# Partial F-test, H0: beta2 = beta3 = 0
> anova(mod.reduced, mod.full)
Analysis of Variance Table

Model 1: SYSBP ~ HEARTRTE
Model 2: SYSBP ~ HEARTRTE + AGEGRP_factor
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1      18 2812.3
2       16 2153.7  2     658.56 2.4462 0.1183
```

$$\begin{aligned} F_0 &= \frac{\frac{SSM(F) - SSM(R)}{\text{Number parameters tested}}}{\frac{SSE(F)}{df_2(F)}} \\ &= \frac{2824 - 2165.4}{\frac{2}{2153.7}} \\ &= \frac{658.6/2}{2153.7/16} = 2.45 \end{aligned}$$

# Interaction

- In the previous example with sex and heart rate, we assumed the same effect of heart rate (slope) for both groups, but allowed different intercepts
- In some situations, an **interaction** might exist
- An **interaction** between two variables,  $x_1$  and  $x_2$ , exists when one explanatory variable ( $x_1$ ) has a different effect on the predicted response depending on the value of a second explanatory variable ( $x_2$ )
  - For example, the effect of heart rate on systolic blood pressure might differ by the sex of the individual (different slope of heart rate for males and females)

# Interaction

- To allow for an effect of this type, we create an **interaction term**
- An interaction term is generated by multiplying together two different explanatory variables  $x_1$  and  $x_2$  to create a third variable  $x_1 x_2$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

Main effects                      Interaction term

## Interaction: Example

$$\mu_{y|\mathbf{x}} = \alpha + \beta_1 x_1 + \beta_2 z_1 + \beta_3 x_1 z_1$$

- A model with heart rate ( $x_1$ ), sex ( $z_1$ ) and the interaction of heart rate and sex ( $x_1 z_1$ )
- Regression line for **males**:

$$\mu_{y|x_1, z_1=1} = \alpha + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) = (\alpha + \beta_2) + (\beta_1 + \beta_3) x_1$$

- Regression line for **females**:

$$\mu_{y|x_1, z_1=0} = \alpha + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) = \alpha + \beta_1 x_1$$

- $\beta_2$ : Difference in the intercept between males and females
- $\beta_3$ : Difference in the slope of heart rate between males and females
- The interaction term allows for a different slope for M and F
- To test if the lines are parallel, test if the interaction term is significant  $H_0: \beta_3 = 0$



# Interaction: Example

## R Code, Interaction Model

```
# Interaction model
> mod.intx <- lm(SYSBP ~ HEARTRTE + SEX_factor + HEARTRTE*SEX_factor, data = fhssrs)
# Printing fitted model
> summary(mod.intx)
...
Coefficients:

```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-5.7122	22.7673	-0.251	0.80509
HEARTRTE	2.0411	0.3172	6.434	0.00000826
SEX_factorMale	94.6501	27.7405	3.412	0.00357
HEARTRTE:SEX_factorMale	-1.5314	0.3892	-3.935	0.00118

```
Residual standard error: 7.942 on 16 degrees of freedom
Multiple R-squared: 0.7972, Adjusted R-squared: 0.7592
F-statistic: 20.97 on 3 and 16 DF, p-value: 0.000008637
```

# Interaction: Example

$$\hat{y} = -5.71 + 2.04 \text{ HR} + 94.65 \text{ Sex} - 1.53 \text{ HR} \times \text{Sex}$$

**Table:** Least squares regression results

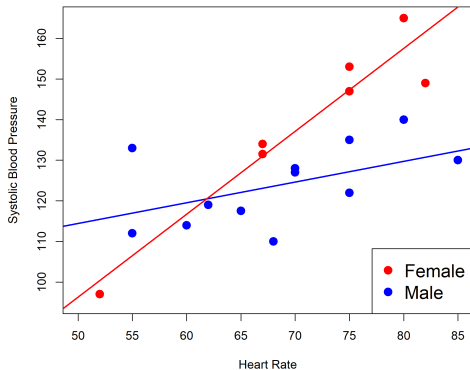
Parameter	Estimate	SE	<i>t</i>	<i>p</i> -value
Intercept	-5.712	22.767	-0.251	0.805
Heart Rate	2.041	0.317	6.434	<.0001
Sex (M vs. F)	94.650	27.741	3.412	0.004
Heart Rate $\times$ Sex	-1.531	0.389	-3.935	0.001

- $b_3 = -1.531$ : Difference in effect of heart rate (slope of heart rate) in males vs. females (*reference*)
  - Test of  $H_0: \beta_3 = 0$ : Reject  $H_0$ ,  $p = 0.0012$
  - Evidence of a significant difference in effect of heart rate on systolic blood pressure in males vs. females. The slope in males is 1.53-units *lower* than the slope in females (*reference*).



# Interaction: Example

$$\hat{y} = -5.71 + 2.04 \text{ HR} + 94.65 \text{ Sex} - 1.53 \text{ HR} \times \text{Sex}$$



- **Females:** [Slope =  $b_1$ ]

$$\begin{aligned}\hat{y} &= -5.71 + 2.04 \text{ HR} + 94.65(0) - 1.53 \text{ HR}(0) \\ &= -5.71 + 2.04 \text{ HR}\end{aligned}$$

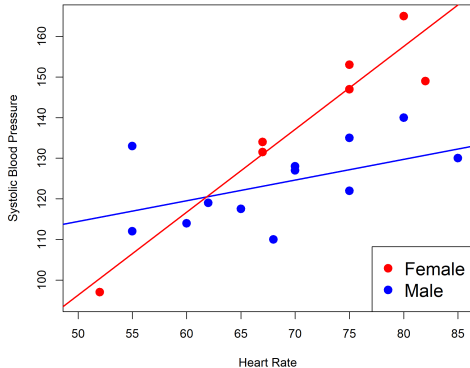
- **Males:** [Slope =  $b_1 + b_3$ ]

$$\begin{aligned}\hat{y} &= -5.71 + 2.04 \text{ HR} + 94.65(1) - 1.53 \text{ HR}(1) \\ &= (-5.71 + 94.65) + (2.04 - 1.53) \text{ HR} \\ &= 88.94 + 0.51 \text{ HR}\end{aligned}$$

- $b_3 = -1.53 = 0.51 - 2.04$  (slope in males - slope in females): Since  $b_3 < 0$ , slope in males is flatter than slope in females (*ref*)

# Interaction: Example

$$\hat{y} = -5.71 + 2.04 \text{ HR} + 94.65 \text{ Sex} - 1.53 \text{ HR} \times \text{Sex}$$



Females:  $\hat{y} = -5.71 + 2.04 \text{ HR}$

Males:  $\hat{y} = 88.94 + 0.51 \text{ HR}$

- Significant interaction: The effect of heart rate on the response is different for different values of sex
- The effect on systolic BP of a 1-unit increase in heart rate is much greater for females than males (steeper slope)
  - A one-unit increase in heart rate increases systolic BP by  $b_1 = 2.04$  mmHg, on average, for females and  $b_1 + b_3 = 0.51$  mmHg, on average, for males

# Testing Linear Combination of Coefficients: Example

## R Code, Estimating Slope in Males: $\beta_1 + \beta_3$

```
# Interaction model
> mod.intx <- lm(SYSBP ~ HEARTRTE + SEX_factor + HEARTRTE*SEX_factor, data = fhssrs)
> library(multcomp)      # Load required package
> names(coef(mod.intx))  # 4 coefficients estimated: a, b1, b2, b3
[1] "(Intercept)"      "HEARTRTE"      "SEX_factorMale"      "HEARTRTE:SEX_factorMale"
# Vector that specifies linear combination of coefficients interested in
> K <- rbind(c(0, 1, 0, 1))  # 1 = coefficients "on" when estimating slope in Males
> rownames(K) <- "b1+b3 (slope in Males)"
> summary(glht(mod.intx, linfct = K))
```

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t )
b1+b3 (slope in Males) == 0	0.5097	0.2255	2.261	0.0381

```
> confint(glht(mod.intx, linfct = K))
```

Linear Hypotheses:

	Estimate	lwr	upr
b1+b3 (slope in Males) == 0	0.50973	0.03179	0.98767

- Test of  $H_0: \beta_1 + \beta_3 = 0$ : Reject  $H_0$ ,  $p = 0.0381$ ; Estimate  $b_1 + b_3 = 0.51$  [95% CI: (0.03, 0.99)]

# Interaction

- If we did not reject  $H_0$ , would not have sufficient evidence to say that the relationship between systolic BP and HR differs depending on the sex of the individual
- If an interaction term between two variables is included in the model, then the “main effects” for those variables should also be included
  - If  $x_1 x_2$  is in the model, then  $x_1$  and  $x_2$  should also be included (*hierarchical principle*)
- If there **is not** a significant interaction, remove the interaction term (unnecessary term), re-run the model that includes only the main effects
- Look at the interaction significance first and proceed from there

# Progress this Unit

- 1 Multiple Linear Regression
  - Motivation
  - The Model
- 2 Inference
  - Confidence Interval and Hypothesis Test for  $\beta_j$
  - Overall  $F$ -Test
  - Partial  $F$ -Test
- 3 Regressors
  - Indicator Variables
  - Categorical Variables with More than Two Categories
  - Interaction Terms
- 4 Model Selection
  - Checking Assumptions
  - Model Selection

# Diagnostics

- Have seen that there are certain assumptions made when we model the data
- These assumptions allow us to calculate test statistics and know the distribution of the test statistic under a null hypothesis
- We will review some commonly used model diagnostics that can identify problems with the fitted model



# Residuals

- The use of a residual plot is analogous to that in the simple linear regression model
  - **Residual**  $e_i = y_i - \hat{y}_i$
  - **Standardized residual**: Divide residual  $e_i$  by the standard deviation of the residual. Quantifies size of residual in standard deviation units: can be easily used to identify outliers.
  - **Studentized residual (a.k.a. jackknife residuals)**: Refit model after removing  $i$ th observation (using  $n - 1$  observations), compare  $y_i$  to fitted  $y$  from model with  $i$ th observation deleted (deleted residual). Standardizing deleted residuals gives studentized residuals.
- Studentized residuals that are large relative to  $N(0, 1)$  (e.g.,  $> 3$  or  $4$ ) may be considered outliers in the sense that  $y_i$  far from  $\hat{y}_i$  under the linear regression model

# Residual Plots

- **Residual Plots**

- Plot of residuals, standardized residuals, or studentized residuals vs.  $\hat{y}$
- Plot of residuals  $e_i$  vs. covariates  $x_1, x_2, \dots, x_k$
- Normal Q-Q plot or histogram of residuals
- Help identify:
  - Outliers
  - Non-normal error distributions
  - Non-constant variance
  - Non-linearity in individual variables

## R Code, Residuals

```
> mod.mlr <- lm(SYSBP ~ AGE + HEARTRTE,
                 data = fhssrs)
> dat <- fhssrs[, c("RANDID", "SYSBP", "AGE",
                   "HEARTRTE")]

# Extract predicted values from mod.mlr
dat$predicted <- predict(mod.mlr)

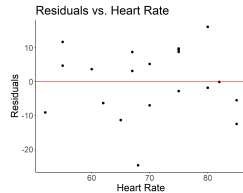
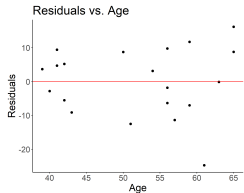
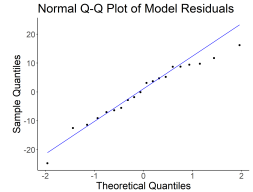
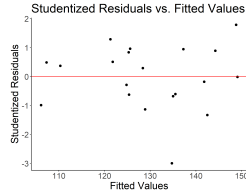
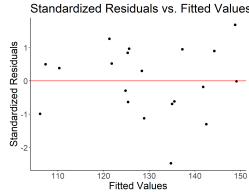
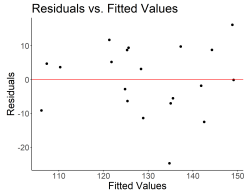
# Extract residuals from mod.mlr
dat$residuals <- resid(mod.mlr)

# Extract standardized residuals from mod.mlr
dat$stdres <- rstandard(mod.mlr)

# Extract studentized residuals from mod.mlr
library(MASS)
dat$studres <- studres(mod.mlr)
```

# Residual Plots

$$\hat{y} = 25.116 + 0.777 \text{ Age} + 0.915 \text{ HR}$$



- No indication of non-linearity, non-constant variance, and non-normality
- Possible outlier with studentized residual  $\approx -3$

# Multicollinearity

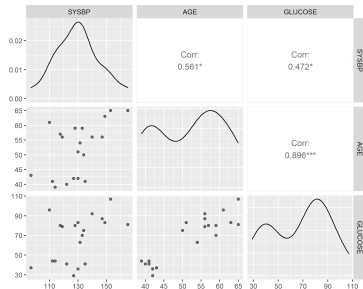
- Another important issue to consider in MLR is the problem of **collinearity** (a.k.a., **multicollinearity**)
- Multicollinearity occurs when two or more explanatory variables are correlated to the extent that they convey essentially the same information about the variation in  $y$
- Some of the  $x$  variables may be redundant in predicting  $y$

# Multicollinearity

- Some **symptoms of multicollinearity**:
  - Instability of the estimated coefficients and their standard errors. Standard errors are inflated, implying there is a large amount of sampling variability in the estimated coefficients.
  - Regression coefficients change greatly when predictors are included/excluded from the model
  - Significant overall  $F$ -test but no significant  $t$ -tests for the slopes ( $\beta$ s)
  - Regression coefficients may not “make sense”, i.e., don’t match scatterplot and/or intuition
- Note that multicollinearity is not a violation of model assumptions, but should be investigated when performing MLR

# Scatterplot and Correlation Matrix: Example

**Figure:** Scatterplot matrix



- Pairwise correlations can be used to check for “pairwise” collinearity
- Useful but note that pairwise correlations do not show more complicated linear dependence between the independent variables

**Table:** Correlation matrix

	SYSBP	AGE	GLUCOSE
SYSBP	1	$r=0.56$ $p=0.01$	$r=0.47$ $p=0.04$
AGE		1	$r=0.90$ $p < 0.001$
GLUCOSE			1

**Table:** Simple linear regression models of systolic BP

	Parameter	SE	$p$
Age	1.003	0.35	0.01

	Parameter	SE	$p$
Glucose	0.326	0.14	0.04

# Multicollinearity: Example

Table: ANOVA Table

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Model	1594.7	2	797.35	4.01	0.037
Error	3383	17	199		
Total	4977.6	19			

Table: Least squares regression results

Parameter	Estimate	SE	<i>t</i>	<i>p</i>
Intercept	76.06	19.71	3.86	0.001
Age	1.16	0.70	1.67	0.11
Glucose	-0.09	0.35	-0.27	0.79

- Overall  $F$ -test is significant, but the tests of individual  $\beta$  are not
- Choose to include either age or glucose in the model
- Association of glucose and systolic blood pressure now negative in MLR

# Model Selection

- In general, if presented with a number of potential explanatory variables, must decide which to include in a regression model
- All else being equal, the simpler model is often easier to interpret and work with
- To study the full effect of each explanatory variable on the response, it would be necessary to perform a separate regression analysis for each possible combination of explanatory variables
- While thorough, the “all possible models” approach can be time-consuming



## Automatic Selection Procedures: Presented in Lab

- Automated stepwise approaches have been developed to choose the “best-fitting” model
- Procedures add and/or subtract variables one at a time according to prespecified inclusion/exclusion criteria
- Can be useful when you have a large number of potential predictors
  - 1 Forward selection
  - 2 Backward elimination
  - 3 Forward stepwise selection
- There are selection methods based on  $p$ -values of tests of the parameters,  $R^2$  and other model diagnostics

## Caution Using Automatic Selection Procedures

- Note that different selection procedures could result in different final models
- Automated procedures cannot assess a good functional form for a predictor and do not think about which interactions might be important
- Although automated variable selection procedures are tempting, it is important to spend time thinking about the research question, predictors of interest, possible confounders, functional form of covariates, and plausible interactions