

# Module 1: Multiple Linear Regression Review

BIOS 526

Instructor: Benjamin Risk

## Reading

- Review matrix algebra. See notes on github, [M0\\_MatrixReview\\_bios526.pdf](#). For a more advanced reference, see [The Matrix Cookbook](#).
- Review notes from Applied Linear Regression (e.g., BIOS 509).
- A detailed reference: Sheather, Simon J. *A Modern Approach to Regression with R*. Springer, 2009.

## Concepts

- Linear regression model in matrix notation.
- Inference for regression coefficient estimates, expected values, and predictions.
- Dummy variables.
- Effect modification and confounding.

## Acknowledgments

- Lecture notes build upon materials from Prof. Howard Chang.

## Motivating Example

What maternal traits are associated with a child's cognitive test score at age 3?

- score: cognitive test score at age 3.
- age: maternal age at delivery.
- edu: maternal education: (1) less than high school, (2) high school, (3) some college, (4) college and above.

```
> dat = read.csv ("testscore.csv")
> str(dat)
'data.frame': 400 obs. of  3 variables:
 $ score : int  120 89 78 42 115 97 94 68 103 94 ...
 $ edu: int   2  1 2  1 4  1 1 2  3 3 ...
 $ age: int   21 17 19 20 26 20 20 24 19 24 ...

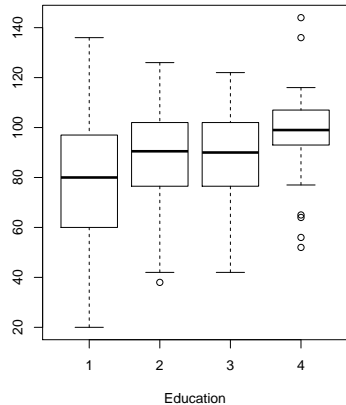
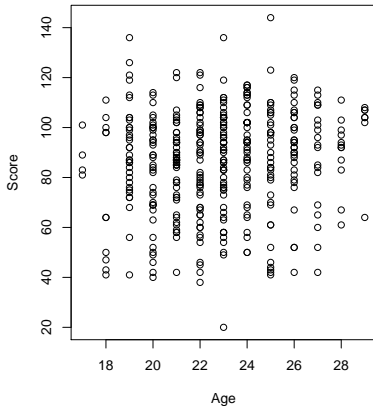
> summary (dat$score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 20.00   74.00   90.00   86.93  102.00  144.00

> table (dat$edu)
 1    2    3    4
85 212  76  27

> summary (dat$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 17.00   21.00   23.00   22.79   25.00   29.00
```

# Exploratory Plots

```
> par (mfrow =c(1,2))  
> plot (score~age, data = dat, xlab="Age",ylab="Score")  
> boxplot (score~edu, xlab = "Education", data = dat)
```



## Simple Linear Regression

Let  $y_i$  denote the test score for child  $i$ ,  $Age_i$  denote the corresponding maternal age at delivery, and  $\epsilon_i$  denote the error term. We will consider the following linear model:

$$y_i = \beta_0 + \beta_1 Age_i + \epsilon_i, \quad i = 1, 2, \dots, 400. \quad (1)$$

- $y_i$  is a linear function of  $Age_i$ .
- $\beta_0$  = the test score for a child born of a mother at age zero. (Not meaningful directly!)
- $\beta_1$  = increase in test score associated with one year increase in maternal age.
- $\mathbb{E} \epsilon_i = 0$ .
- To start, we do not make assumptions regarding the  $\epsilon_i$ .

Clearly, we cannot find  $\beta_0$  and  $\beta_1$  such that model (1) fits all of our observations perfectly.

## Least Squares Estimate

We can write model (1) in matrix form by stacking observations by row:

$$\begin{aligned}\mathbf{Y} &= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \beta_1 \text{Age}_1 \\ \beta_0 + \beta_1 \text{Age}_2 \\ \vdots \\ \beta_0 + \beta_1 \text{Age}_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\ &= \begin{bmatrix} 1 & \text{Age}_1 \\ 1 & \text{Age}_2 \\ \vdots & \\ 1 & \text{Age}_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\ &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.\end{aligned}$$

## Least Squares Estimate, cont.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{Y}$  is an  $n \times 1$  response (dependent variable) vector.
- $\mathbf{X}$  is an  $n \times p$  (design) matrix where  $p$  is the number of covariates (i.e., predictors, i.e., independent variables).
- $\boldsymbol{\beta}$  is a  $p \times 1$  vector of regression coefficients.
- $\boldsymbol{\epsilon}$  is an  $n \times 1$  vector of errors.

We can view the expected value of  $\mathbf{Y}$  as a **linear combination** of the two columns of  $\mathbf{X}$ .

$$\mathbb{E} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$$

## Least Squares Solution

One approach is to minimize a loss function. A popular loss function is the sum of squared differences between the observed  $\mathbf{Y}$  and  $\mathbf{X}\boldsymbol{\theta}$  for some vector  $\boldsymbol{\theta} \in \mathbb{R}^2$ .

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 \text{Age}_i)^2 \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathbf{Y}^T \mathbf{Y} - 2\mathbf{Y}^T \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\theta}\end{aligned}$$

Differentiate the above with respect to  $\boldsymbol{\theta}$  and set to zero, we get

$$-2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X}\boldsymbol{\theta} = 0$$

If  $\mathbf{X}$  is full-rank, solving this gives us the **ordinary least squares estimate**:

$$\boxed{\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}}.$$



## Properties of LS Estimate

$$\hat{\beta}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- For OLS, we have a nice closed-form solution. Other loss functions or models often require iterative optimization routines.
- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$  is the fitted value.
- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  is the **hat matrix**. It shows that  $\hat{\mathbf{Y}}$  can be expressed as a **linear combination** of  $\mathbf{Y}$ .
- $\hat{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}$  is the residual.

Geometric interpretation of the least squares estimate:

1.  $\hat{\mathbf{Y}}$  and  $\hat{\epsilon}$  are orthogonal.
2.  $\mathbf{H}$  is a projection matrix of  $\mathbf{Y}$  on the column space of  $\mathbf{X}$ . Also note that  $\mathbf{H}\mathbf{H} = \mathbf{H}$  (definition of idempotent = projection matrix)
3.  $\text{tr}(\mathbf{H}) = p = \text{number of covariates (including intercept)}$ .

# Least squares and hyperplanes

The least squares solution minimizes the sum of squared distances between  $\mathbf{Y}$  and  $\hat{\mathbf{Y}}$ .

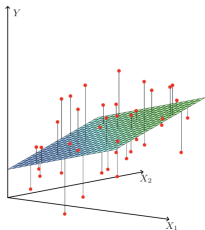


FIGURE 3.1. Linear least squares fitting with  $X \in \mathbb{R}^2$ . We seek the linear function of  $X$  that minimizes the sum of squared residuals from  $Y$ .

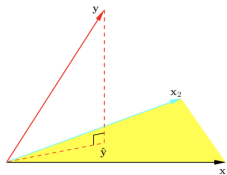
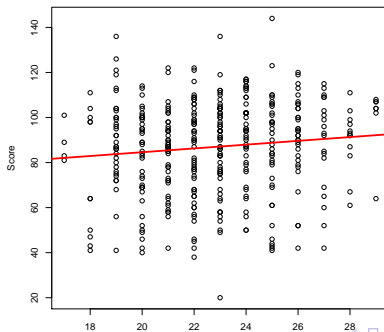


FIGURE 3.2. The  $N$ -dimensional geometry of least squares regression with two predictors. The outcome vector  $\mathbf{y}$  is orthogonally projected onto the hyperplane spanned by the input vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . The projection  $\hat{\mathbf{y}}$  represents the vector of the least squares predictions

**Figure:** The view of OLS in  $\mathbb{R}^2$  (left) and the view in  $\mathbb{R}^n$  (right). Note the picture on the right is an abstract view of  $n$ -dimensional vectors, which humans generally aren't able to visualize. Taken from Hastie et al ESL.

# OLS calculations in R

```
> X = cbind (1, dat$age)
> Y = dat$score
> beta = solve ( t(X) %*% X ) %*% t(X) %*% Y
> beta
      [,1]
[1,] 67.7826813
[2,] 0.8402729
> plot (score~age, data = dat, xlab="Age",ylab="Score")
> abline (beta, lwd = 3, col = 2)
```



# Properties of the Estimator

The estimator

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

has the following properties.

1.  $\hat{\beta}$  is an unbiased estimator of  $\beta$ :  $\mathbb{E} \hat{\beta} = \beta$ .
2.  $\hat{\beta}$  is a consistent estimator of  $\beta$  under very general assumptions. Let  $\hat{\beta}_n$  be the OLS estimator from  $n$  observations. Then  $\lim_{n \rightarrow \infty} P(|\hat{\beta}_n - \beta| > \epsilon) = 0$ . (Details omitted.)
3.  $\hat{\mathbf{Y}}$  is an unbiased estimator of the mean trend  $\mathbf{X}\beta$ :  $\mathbb{E} \hat{\mathbf{Y}} = \mathbf{X}\beta$ .
4. Supposing  $\epsilon_i \stackrel{iid}{\sim} (0, \sigma^2)$ ,  $\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$ .
5.  $\text{Cov}(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ .

# Statistical Linear Regression Model

So far, we have not discussed the distribution of the errors in Model (1). For **inference**, let's now assume

$$y_i = \beta_0 + \beta_1 \text{Age}_i + \epsilon_i, \quad i = 1, 2, \dots, 400. \quad (2)$$

$$\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

This implies the following model:

- $y_i$  is normally distributed:

$$y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 \text{Age}_i, \sigma^2).$$

- i.e., the observed  $y_i$  is normally distributed around the linear trend  $\beta_0 + \beta_1 \text{Age}_i$ .

# Statistical Linear Regression Model

- Note: In our data application predicting child's test score from maternal age at delivery and mother's education, this model is wrong. Why?
- Can a wrong model be useful?

# Inference for Regression Coefficients

Model (2) also implies that the joint distribution of  $\mathbf{Y}$  is

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where  $\mathbf{I}$  is an  $n \times n$  identity matrix.

But we don't know  $\sigma^2$ .

## lm() and Residual Error

We often estimate  $\sigma^2$  with an unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2 .$$

```
> fit = lm (score~age, data = dat)
```

```
> summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	67.7827	8.6880	7.802	5.42e-14	***
age	0.8403	0.3786	2.219	0.027	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.34 on 398 degrees of freedom

Multiple R-squared: 0.01223, Adjusted R-squared: 0.009743

F-statistic: 4.926 on 1 and 398 DF, p-value: 0.02702



# Confidence Interval for the Estimate of the Mean

## Method 1 (manual-ish):

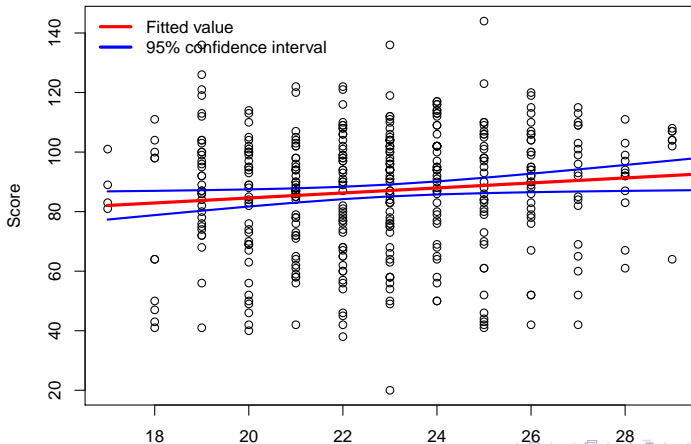
```
> invXtX = solve(t(X)%*%X)
> beta = invXtX %*% t(X) %*% Y
> sigmahat = sum((Y - X%*%beta)^2) / (length(Y)-2)
> V = sigmahat*invXtX
> X = cbind (1, 17:30) #Design matrix with age=17, 18, ..., 30
> Est = X %*% beta
> SE = sqrt( diag(X%*%V%*%t(X)) )
> Upper95 = Est + 1.96*SE
> Lower95 = Est - 1.96*SE
> cbind(Est, Lower95, Upper95)[1:3,]
      [,1]      [,2]      [,3]
[1,] 82.06732 77.33092 86.80372
[2,] 82.90759 78.83235 86.98283
[3,] 83.74787 80.30069 87.19504
```

## Method 2:

```
> confInt = predict (fit, newdata = data.frame(age=17:30), interval="conf")
> confInt[1:3,]
      fit      lwr      upr
1 82.06732 77.31656 86.81808
2 82.90759 78.82000 86.99519
3 83.74787 80.29024 87.20549
```

## Confidence Interval for the Estimate of the Mean, ii

```
> plot (score~age, data = dat, xlab="Age",ylab="Score")  
> lines (Est~c(17:30), col = 2, lwd =3)  
> lines (Upper95~c(17:30), col = 2, lwd = 3, lty = 3)  
> lines (Lower95~c(17:30), col = 2, lwd = 3, lty = 3)  
> legend ("topleft", legend = c("Fitted value", "95% confidence interval"),  
lty = c(1,3), bty="n", lwd=3)
```



# Prediction Interval for New Observation

Let  $y_i$  be a **new observation** with covariate value  $age_i$ . How do we predict its uncertainty?

We want to capture the uncertainty in our estimate of the expected value,  $\hat{y}_i$ , plus the uncertainty due to measurement error  $\epsilon_i$ . Recall

$$y_i = \beta_0 + \beta_1 age_i + \epsilon_i$$

for  $\epsilon_i \sim N(0, \sigma^2)$ .

If we knew  $\epsilon_i$ , a point estimator would be

$$\tilde{y}_i = \hat{\beta}_0 + \hat{\beta}_1 age_i + \epsilon_i .$$

Then

$$\text{Var}(\tilde{y}_i) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 age_i) + \text{Var}(\epsilon_i)$$

## Prediction Interval for New Observation, ii

We have

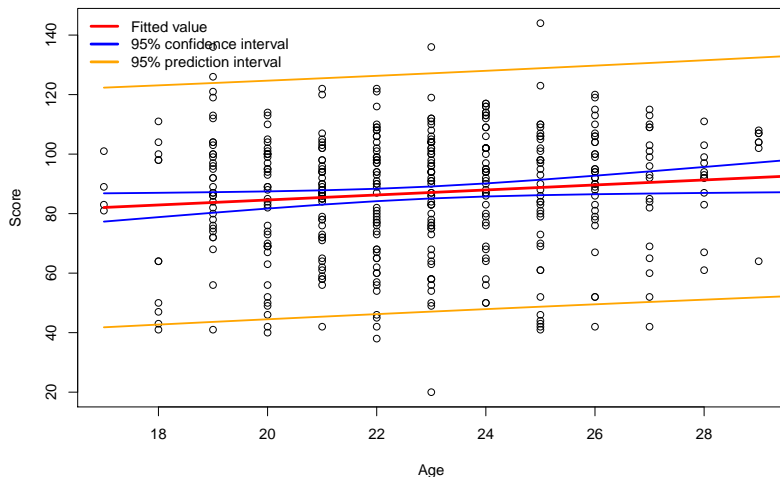
$$\text{Var}(\tilde{y}_i) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 \text{age}_i) + \text{Var}(\epsilon_i)$$

So we plug in our estimator to get the predictive variance:

$$\tilde{\sigma}^2 = \text{Var}(\tilde{y}_i) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 \text{age}_i) + \hat{\sigma}^2$$

and the approximate prediction interval is  $\tilde{y}_i \pm 1.96\tilde{\sigma}$ .

# Prediction Interval for New Observation, iii

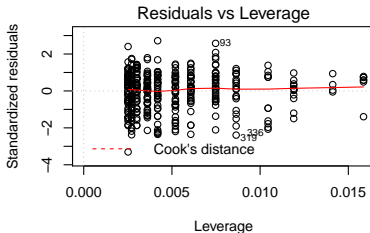
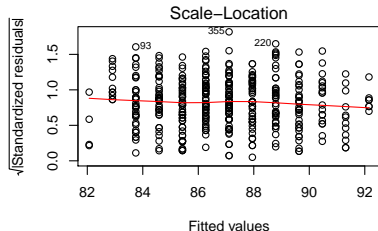
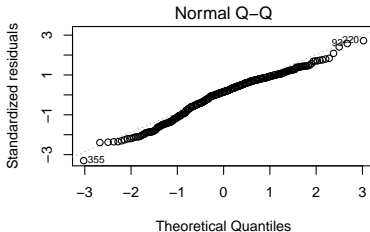
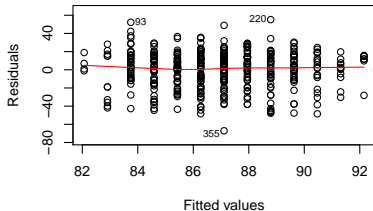


# Assumptions of Linear Regression

1. Residuals are independent.
2. Linearity.
3. Residuals are normal.
4. Homoscedasticity (variance of residuals is constant).

# Model Diagnostics

```
> par (mfrow = c(2,2))
> plot (fit)  Residuals vs Fitted
```



# Model Diagnostics

## Residual vs Fitted

- $\hat{\epsilon}_i$  should be independent of  $\hat{y}_i$  (no patterns):
- Linearity: Red line should be flat.
- Variance constant (homoscedasticity)

## Normal Q-Q Plot

- Standardized residual  $\hat{\epsilon}_i / (\hat{\sigma} \sqrt{1 - h_{ii}})$  should be standard normal.
- We expect the points to follow a straight line. Check non-normality, particularly skewed tails.

## Scale-Location

- Similar to residual-vs-fitted, but use  $\hat{\epsilon}_i / (\hat{\sigma} \sqrt{1 - h_{ii}})$ . Diagnose heteroscedasticity (e.g., red line increasing).

## Residual vs Leverage

- Leverage  $h_{ii}$  is how far away  $\mathbf{x}_i$  is from other  $\mathbf{x}_{i'}$ .  $h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$ .
- high leverage and outlier = problem.
- Cook's Distance: measures how much model changes when remove  $i$ th point;  $> 1$  is problematic. (Note: when this occurs, a contour line appears in plot.)



# Model Coefficient Interpretations

```
> fit = lm (score~age, data = dat)
```

```
> summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	67.7827	8.6880	7.802	5.42e-14	***
age	0.8403	0.3786	2.219	0.027	*

Residual standard error: 20.34 on 398 degrees of freedom

Multiple R-squared: 0.01223, Adjusted R-squared: 0.009743

F-statistic: 4.926 on 1 and 398 DF, p-value: 0.02702

- A one year increase in mother's age at delivery was associated with a 0.84 ( $CI_{95\%} 0.84 \pm 2 \cdot 0.38$ ) increase in the child's average test score, where test score was measured at age 3. This association was statistically significant at a type I error rate of 0.05.
- There was considerable heterogeneity in test scores ( $\hat{\sigma} = 20.34$ ). Therefore the regression model does not predict individual test scores well ( $R^2 = 0.012$ ).

# Hypothesis testing

```
> fit = lm (score~age, data = dat)
```

```
> summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	67.7827	8.6880	7.802	5.42e-14	***
age	0.8403	0.3786	2.219	0.027	*

Residual standard error: 20.34 on 398 degrees of freedom

Multiple R-squared: 0.01223, Adjusted R-squared: 0.009743

F-statistic: 4.926 on 1 and 398 DF, p-value: 0.02702

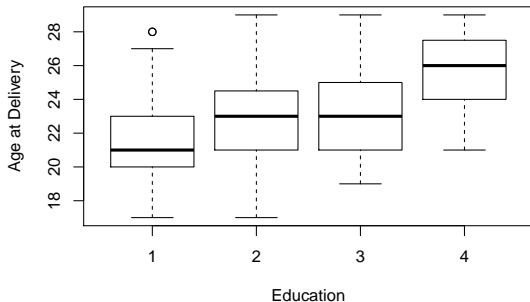
- (Intercept):  $H_0 : \beta_0 = 0$  when age=0. In words, test scores are equal to zero for a mother at age=0.  $H_A : \beta_0 \neq 0$ .  
Conclusion:  $p < 0.05$  so we reject the null hypothesis at  $\alpha = 0.05$  and conclude the intercept differs from zero.
- age:  $H_0 : \beta_1 = 0$ . In words, the slope of age is equal to zero.  
Equivalently, there is no linear effect of age.  $H_A : \beta_1 \neq 0$ .  
Conclusion:  $p < 0.05$ , so we reject the null hypothesis and conclude there is a significant linear effect of age.

# Quiz 1

Class activity: Break-out session and quiz 1.

## Confounding, i

We found that older mothers had children on average with higher test scores. However, higher maternal education was associated with higher age as shown in the boxplots below.



How do we estimate the age association accounting for the effect of education?

## Confounding, ii

- Confounding causes spurious correlation.
- This is also called “omitted variables bias.”
- We can use multiple regression to account for possible confounding factors.
- Other words meaning the same thing: “adjusting for” mother’s education, “accounting for” mother’s education.
- “controlling for” mother’s education.
- Note: Causal inference provides a formal treatment for confounding; EPI 760 / BIOS 761. For  $X, Y, Z$ , if  $Z \rightarrow X$  and  $Z \rightarrow Y$ , then  $Z$  is a confounding variable for  $X \rightarrow Y$ .
- This course: we model  $E[Y|X, Z]$ , such that the effect of  $X$  is adjusted for  $Z$ , whereas  $E[Y|X]$  is unadjusted.
- In this class, if a coefficient for a predictor of interest changes greatly when accounting for a new predictor, then this may be due to confounding.

# Assumptions of Linear Regression (plus omitted variables assumption)

1. Residuals are independent.
2. Linearity.
3. Residuals are normal.
4. Homoscedasticity (variance of residuals is constant).
5. All variables that are not included in the model have coefficient equal to zero (omitted variables bias = 0).

## Categorical Variables

First, examine the effects of education on scores.

edu is coded as 1, 2, 3, 4 – don't want to code as a continuous variable.

**Approach 1:** an indicator (dummy) for each group

$$X_{1i} = 1_{\{edu_i=1\}}, X_{2i} = 1_{\{edu_i=2\}}, X_{3i} = 1_{\{edu_i=3\}}, X_{4i} = 1_{\{edu_i=4\}}$$

If we have  $edu_i = [1, 1, 2, 2, 3, 3, 4, 4]$ , the design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

What does each element of the corresponding  $\beta$  vector mean?

## Average Score by Maternal Education, i

```
> E1 = as.numeric(dat$edu == 1)
> E2 = as.numeric(dat$edu == 2)
> E3 = as.numeric(dat$edu == 3)
> E4 = as.numeric(dat$edu == 4)

> fit = lm (dat$score ~ E1 + E2 + E3 + E4 - 1) #use "-1" to remove intercept
> summary (fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
E1	78.447	2.159	36.33	<2e-16 ***
E2	88.703	1.367	64.88	<2e-16 ***
E3	87.789	2.284	38.44	<2e-16 ***
E4	97.333	3.831	25.41	<2e-16 ***

Residual standard error: 19.91 on 396 degrees of freedom  
Multiple R-squared: 0.9508, Adjusted R-squared: 0.9503  
F-statistic: 1913 on 4 and 396 DF, p-value: < 2.2e-16

This is equivalent to calculating the mean IQ for each education group:

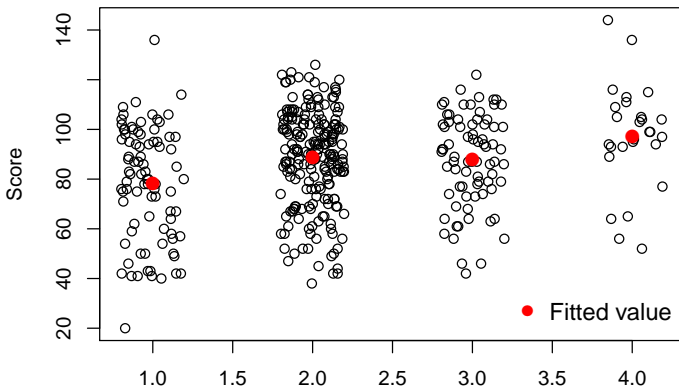
```
> tapply(dat$score, INDEX = dat$edu, FUN = mean)
      1      2      3      4
78.44706 88.70283 87.78947 97.33333
```



## Average Score by Maternal Education, ii

What is the estimated difference in average score between mothers without high school and mothers with college and above?

```
> plot (score~jitter(edu), data = dat, xlab="Education Group",ylab="Score")  
> lines (coef(fit)~c(1,2,3,4), col = 2, cex=1.5, pch=16, type = "p")  
> legend ("bottomright", legend=c("Fitted value"),pch=16,col=2,cex=1.2,bty="n")
```



## Categorical Variables, Approach 2

**Approach 2:** an intercept + three indicators for group 2, 3, and 4.

$$X_{1i} = 1, X_{2i} = 1_{\{edu_i=2\}}, X_{3i} = 1_{\{edu_i=3\}}, X_{4i} = 1_{\{edu_i=4\}}$$

Similarly, if we have  $edu_i = [1, 1, 2, 2, 3, 3, 4, 4]$ , the design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

What does each element of the corresponding  $\beta$  vector mean?

## Difference in Average Score w.r.t. Edu = 1

```
> fit = lm (dat$score ~ E2 + E3 + E4 )
```

```
> summary (fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	78.447	2.159	36.330	< 2e-16	***
E2	10.256	2.556	4.013	7.18e-05	***
E3	9.342	3.143	2.973	0.00313	**
E4	18.886	4.398	4.294	2.21e-05	***

## R can automatically create dummy variables with "factor"

```
> fit = lm (dat$score ~ factor(dat$edu))
```

```
> summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	78.447	2.159	36.330	< 2e-16	***
factor(dat\$edu)2	10.256	2.556	4.013	7.18e-05	***
factor(dat\$edu)3	9.342	3.143	2.973	0.00313	**
factor(dat\$edu)4	18.886	4.398	4.294	2.21e-05	***

Residual standard error: 19.91 on 396 degrees of freedom

Multiple R-squared: 0.05856, Adjusted R-squared: 0.05142

F-statistic: 8.21 on 3 and 396 DF, p-value: 2.59e-05

# Adjusting for Confounding

Now consider the following model:

$$y_i = \beta_0 + \beta_1 E2_i + \beta_2 E3_i + \beta_3 E4_i + \beta_4 age_i + \epsilon_i. \quad (3)$$

- Each education group has their own intercept.
- The effect (slope) of maternal age is constant across groups (parallel lines).
- $\beta_4$  is the association between score and age, **accounting for different averages in score** in different education groups.

# Activity and Quiz ii

- Breakout session with activity and quiz 2

## Adjusting for confounding, ii

```
> fit = lm (score~factor(edu) + age, data = dat)
```

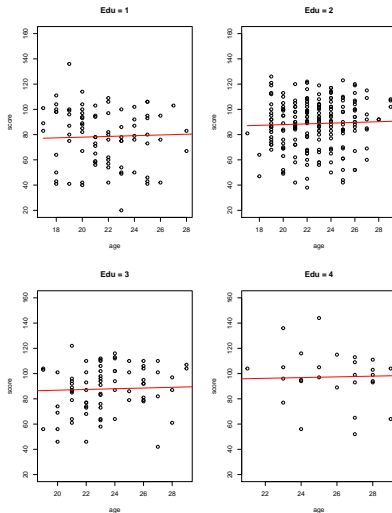
```
> summary (fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	72.2360	8.8700	8.144	5.07e-15	***
factor(edu)2	9.9365	2.5953	3.829	0.000150	***
factor(edu)3	8.8416	3.2203	2.746	0.006316	**
factor(edu)4	17.6809	4.7065	3.757	0.000198	***
age	0.2877	0.3985	0.722	0.470736	

Age effect is no longer statistically significant!

# Adjusting for Confounding



# Variance Inflation Factors

One must always be on the look out for issues with **multicollinearity**.

The general issue is that when two variables are highly correlated, it is hard to disentangle their effects.

Mathematically, the standard errors are inflated. Suppose we have a design matrix  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ , and we want to calculate the variance inflation factor for  $x_1$ . We regress  $\mathbf{x}_1 \in \mathbb{R}^n$  against  $[\mathbf{x}_2, \dots, \mathbf{x}_p]$ .

Let  $R_1^2$  be the associated R-squared. Then  $VIF_1 = \frac{1}{1-R_1^2}$ .

It can be shown  $\text{Var } \hat{\beta}_j = VIF_j \frac{\sigma^2}{(n-1)S_{x_j}^2}$ .



# VIFs

- There are different rules of thumb:  $VIFs > 10$  or 5 or 3 are cause for concern, but this is only a rough guide!
- Sheather (2008) uses 5.
- For large sample sizes, we can tolerate more multicollinearity.
- The GVIF is a generalization of VIF for factors. For  $DF=1$ ,  $GVIF^{1/2/DF}$  is the square root of the usual VIF, so one approach is to square it in order to apply the rules of thumb.

# VIF

```
> library(car)
> vif(fit)

          GVIF Df GVIF^(1/(2*Df))
FactorEdu 1.15514 3      1.024328
age       1.15514 1      1.074774
>
> temp = vif(fit)
>
> temp[,3]^2
FactorEdu      age
 1.049248  1.155140
>
> temp[,3]^2<5
FactorEdu      age
      TRUE      TRUE
```

## Interaction Effects: Effect Modification

Model (3) assumes an identical effect of age across all education groups. We can relax this by including **interaction** terms.

$$y_i = \beta_0 + \beta_1 E2_i + \beta_2 E3_i + \beta_3 E4_i + \beta_4 age_i + \beta_5 E2_i age_i + \beta_6 E3_i age_i + \beta_7 E4_i age_i + \epsilon_i. \quad (4)$$

For each edu group:

$$y_i = \begin{cases} \beta_0 + \beta_4 age_i & Edu = 1 \\ \beta_0 + \beta_1 + (\beta_4 + \beta_5) age_i & Edu = 2 \\ \beta_0 + \beta_2 + (\beta_4 + \beta_6) age_i & Edu = 3 \\ \beta_0 + \beta_3 + (\beta_4 + \beta_7) age_i & Edu = 4 \end{cases}$$

By examining whether  $\beta_5, \beta_6$ , and  $\beta_7$  are zero, we can determine if the effect of maternal age is **modified** by education.

# Interaction Effects: Effect Modification

```
> fit = lm (score~FactorEdu*age, data = dat)
> summary(fit)
```

Call:

```
lm(formula = score ~ FactorEdu * age, data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-56.70	-11.80	2.07	14.58	54.34

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	105.2202	17.6127	5.974	5.2e-09	***
FactorEdu2	-33.0929	21.5732	-1.534	0.1258	
FactorEdu3	-53.4970	27.9460	-1.914	0.0563	.
FactorEdu4	36.4537	49.5065	0.736	0.4620	
age	-1.2402	0.8097	-1.532	0.1264	
FactorEdu2:age	1.9704	0.9764	2.018	0.0443	*
FactorEdu3:age	2.7862	1.2293	2.266	0.0240	*
FactorEdu4:age	-0.4799	1.9635	-0.244	0.8070	

## Effect Modification, continued

```
> vif(fit)
```

	GVIF	Df	$GVIF^{1/(2*Df)}$
FactorEdu	9.093097e+05	3	9.842800
age	4.821946e+00	1	2.195893
FactorEdu:age	1.039839e+06	3	10.065322

Note how the variance is inflated.

This is common with interaction variables since by construction there is dependence between an interaction variable and the main effects. It is often an issue we have to live with.

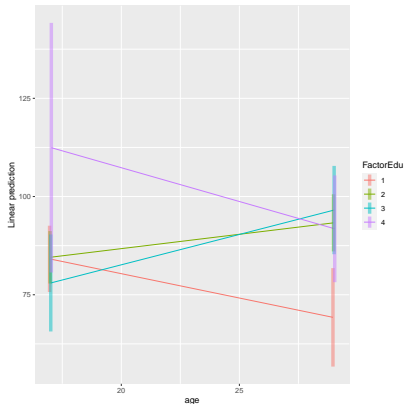
In general, it means we need larger sample sizes to examine interactions.

# Interaction Plots

It is important to visualize your data and model results.

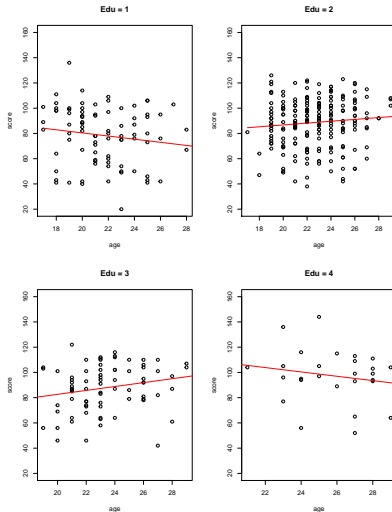
`emmeans::emmip` provides a quick plot:

```
emmip(fit, FactorEdu ~ age, cov.reduce = range, CIs = TRUE)
```



# Effect Modification

We can also examine the effects by “manually” creating these plots; see R code:



## F-test of interaction

To test whether the interaction was significant, look at whether the inclusion of the interaction significantly improved model fit.

$H_0$ : The interaction between education and age does not improve model fit.

$H_A$ : The interaction improves model fit.

```
> fit_nointer = lm(score~FactorEdu+age,data=dat)
> anova(fit_nointer,fit)
Analysis of Variance Table
```

Model 1: score ~ FactorEdu + age

Model 2: score ~ FactorEdu \* age

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	395	156733				
2	392	153857	3	2876.5	2.4429	0.06376 .

Interestingly, it is not significant. Overall, there is only limited statistical evidence of an interaction.



## Linear Combination of Coefficients

Let's take a closer look at the individual slopes.

We wish to estimate the slope of *age<sub>i</sub>* among *Edu<sub>i</sub>* = 3. The point estimate is  $\hat{\beta}_4 + \hat{\beta}_6$ . Also,

$$\text{Var}(\hat{\beta}_4 + \hat{\beta}_6) = \text{Var}(\hat{\beta}_4) + \text{Var}(\hat{\beta}_6) + 2\text{Cov}(\hat{\beta}_4, \hat{\beta}_6) .$$

```
> fit = lm (score~factor(edu)*age, data = dat)
> Est = coef(fit)[5] + coef(fit)[7]
> SE = sqrt ( vcov(fit)[5,5] + vcov(fit)[7,7] + 2*vcov(fit)[5,7])
> Est
      age
1.545989
> SE
[1] 0.9249421
```

So a 95% confidence interval for  $\hat{\beta}_4 + \hat{\beta}_6$  is

$$1.54 \pm 1.96 \times 0.92 = (-0.27, 3.36) .$$

and thus is not statistically different from 0 at  $\alpha = 0.05$ .

## Effect of centering with interactions

Now let's fit the linear model with age centered.

```
> dat$ageC = scale(dat$age,center=TRUE,scale=FALSE)
> fit_inter_ageC = lm(score~FactorEdu*ageC,data=dat)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	76.9567	2.3589	32.624	< 2e-16	***
FactorEdu2	11.8133	2.7237	4.337	1.84e-05	***
FactorEdu3	9.9996	3.3132	3.018	0.002710	**
FactorEdu4	25.5160	6.9760	3.658	0.000289	***
ageC	-1.2402	0.8097	-1.532	0.126440	
FactorEdu2:ageC	1.9704	0.9764	2.018	0.044263	*
FactorEdu3:ageC	2.7862	1.2293	2.266	0.023969	*
FactorEdu4:ageC	-0.4799	1.9635	-0.244	0.807027	

Residual standard error: 19.81 on 392 degrees of freedom  
Multiple R-squared: 0.07705, Adjusted R-squared: 0.06057  
F-statistic: 4.675 on 7 and 392 DF, p-value: 4.756e-05

## Effect of centering, cont.

What changed?

Why?

## Effect Modification

We now have the following model:

$$y_i = \beta_0^* + \beta_1^* E2_i + \beta_2^* E3_i + \beta_3^* E4_i + \beta_4 (age_i - a\bar{age}.) + \\ \beta_5 E2_i (age_i - a\bar{age}.) + \beta_6 E3_i (age_i - a\bar{age}.) + \beta_7 E4_i (age_i - a\bar{age}.) + \epsilon_i.$$

Compare this to the previous model:

$$y_i = \beta_0 + \beta_1 E2_i + \beta_2 E3_i + \beta_3 E4_i + \beta_4 age_i + \\ \beta_5 E2_i age_i + \beta_6 E3_i age_i + \beta_7 E4_i age_i + \epsilon_i.$$

Then  $\beta_1 = \beta_1^* - \beta_5 a\bar{age}.$

In our estimates,  $-33.09 = 11.81 - 1.97 * 22.79$

## More on centering

Compare the VIF before and after centering age:

```
> vif(fit) #not centered
```

	GVIF	Df	$GVIF^{(1/(2*Df))}$
FactorEdu	9.093097e+05	3	9.842800
age	4.821946e+00	1	2.195893
FactorEdu:age	1.039839e+06	3	10.065322

```
> vif(fit_inter_ageC) #centering tends to improve VIFs
```

	GVIF	Df	$GVIF^{(1/(2*Df))}$
FactorEdu	3.411158	3	1.226922
ageC	4.821946	1	2.195893
FactorEdu:ageC	12.285804	3	1.519033

## F-tests and location invariance

We can test the overall effect of education using F-tests. Invariant to centering.

```
> fit_ageC = lm(score~ageC,data=dat)
> anova(fit_ageC,fit_inter_ageC)
Analysis of Variance Table

Model 1: score ~ ageC
Model 2: score ~ FactorEdu * ageC
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     398 164663
2     392 153857   6     10807 4.5889 0.0001617 ***
---
```

```
>
> fit_age = lm(score~age,data=dat)
> anova(fit_age,fit)
Analysis of Variance Table

Model 1: score ~ age
Model 2: score ~ FactorEdu * age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     398 164663
2     392 153857   6     10807 4.5889 0.0001617 ***
---
```

## Interactions: compare slopes

We can test the difference between each pair of slopes.

First, let's ignore corrections for multiple comparisons.

```
> library(emmeans)
> emtrends(fit, pairwise~FactorEdu, var="age", adjust='none')
```

```
$emtrends
```

FactorEdu	age.trend	SE	df	lower.CL	upper.CL
1	-1.24	0.810	392	-2.832	0.352
2	0.73	0.546	392	-0.342	1.803
3	1.55	0.925	392	-0.272	3.364
4	-1.72	1.789	392	-5.237	1.797

Confidence level used: 0.95

```
$contrasts
```

contrast	estimate	SE	df	t.ratio	p.value
1 - 2	-1.970	0.976	392	-2.018	0.0443
1 - 3	-2.786	1.229	392	-2.266	0.0240
1 - 4	0.480	1.964	392	0.244	0.8070
2 - 3	-0.816	1.074	392	-0.760	0.4479
2 - 4	2.450	1.870	392	1.310	0.1909
3 - 4	3.266	2.014	392	1.622	0.1056

## Interactions: compare slopes

Now let's use a Bonferroni correction for the pairwise comparisons: multiply p-values by number of tests. This will be discussed later in the course.

```
> emtrends(fit, pairwise~FactorEdu, var="age", adjust='bonferroni')
```

```
$emtrends
```

FactorEdu	age.trend	SE	df	lower.CL	upper.CL
1	-1.24	0.810	392	-2.832	0.352
2	0.73	0.546	392	-0.342	1.803
3	1.55	0.925	392	-0.272	3.364
4	-1.72	1.789	392	-5.237	1.797

Confidence level used: 0.95

```
$contrasts
```

contrast	estimate	SE	df	t.ratio	p.value
1 - 2	-1.970	0.976	392	-2.018	0.2656
1 - 3	-2.786	1.229	392	-2.266	0.1438
1 - 4	0.480	1.964	392	0.244	1.0000
2 - 3	-0.816	1.074	392	-0.760	1.0000
2 - 4	2.450	1.870	392	1.310	1.0000
3 - 4	3.266	2.014	392	1.622	0.6338



# Interactions, Factor Levels, and Centering

*These are important*

- Centering a covariate affects the location of the main effects of the terms it interacts with.
- Main effects of Education now describe education at the *average* age instead of Age=0.
- F-test of overall effect of Education is invariant to location of Age.
- Additionally, the reference levels of a factor will impact the p-values in `summary()`.
- Always write out the model you are fitting.
- Include main effects whenever interactions are included.
- Examine overall effects using `anova(modelreduced,modelfull)`.