Multiple Linear Regression - EDA and Model

Research Objective

Research Question: What determines a person's height?

Population: All BYU students.

Parameter of Interest:

- Some number measuring the "relationship" between height and various other explanatory variables such as fathers height, mother's height, etc.

Sample: A convenience sample of 1575 BYU students who are in Stat 121.

More Problem Definitions

Response Variable (y): The height of a student.

- This is a **continuous quantitative variable** meaning it can be any number (including decimals)

Explanatory Variable (x):

- Lots! The goal is to relate multiple explanatory variables to a single quantitative response variable.

Variable Encoding

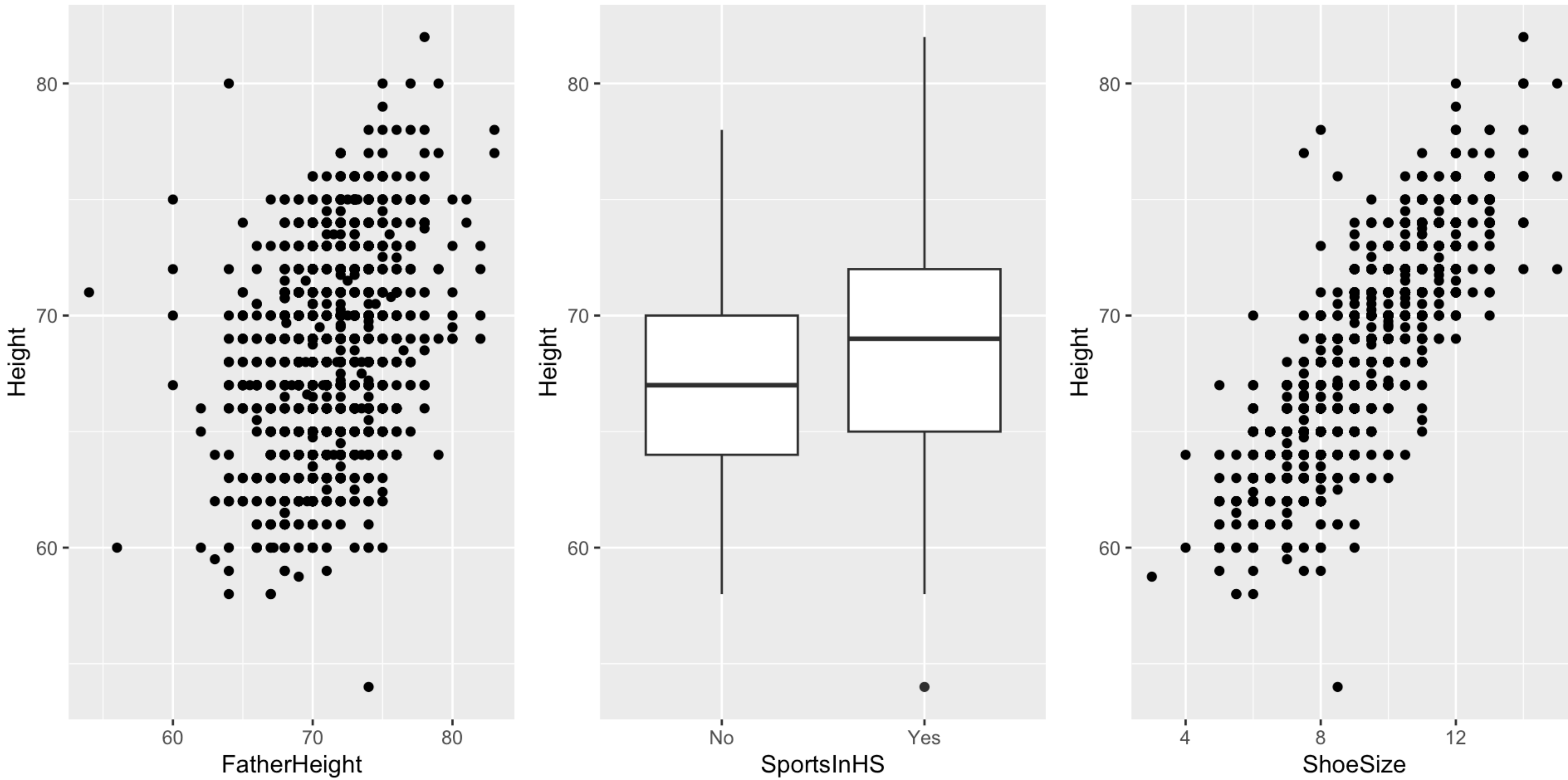
(Part of) Your Student Survey Data

Height	MotherHeight	FatherHeight	SportsInHS	Sex	ShoeSize
69	63	72	No	Female	9.5
68	71	74	Yes	Female	9.5
73	64	72	No	Male	11.5
66	64	70	Yes	Female	7.5
75	70	72	Yes	Male	12.0

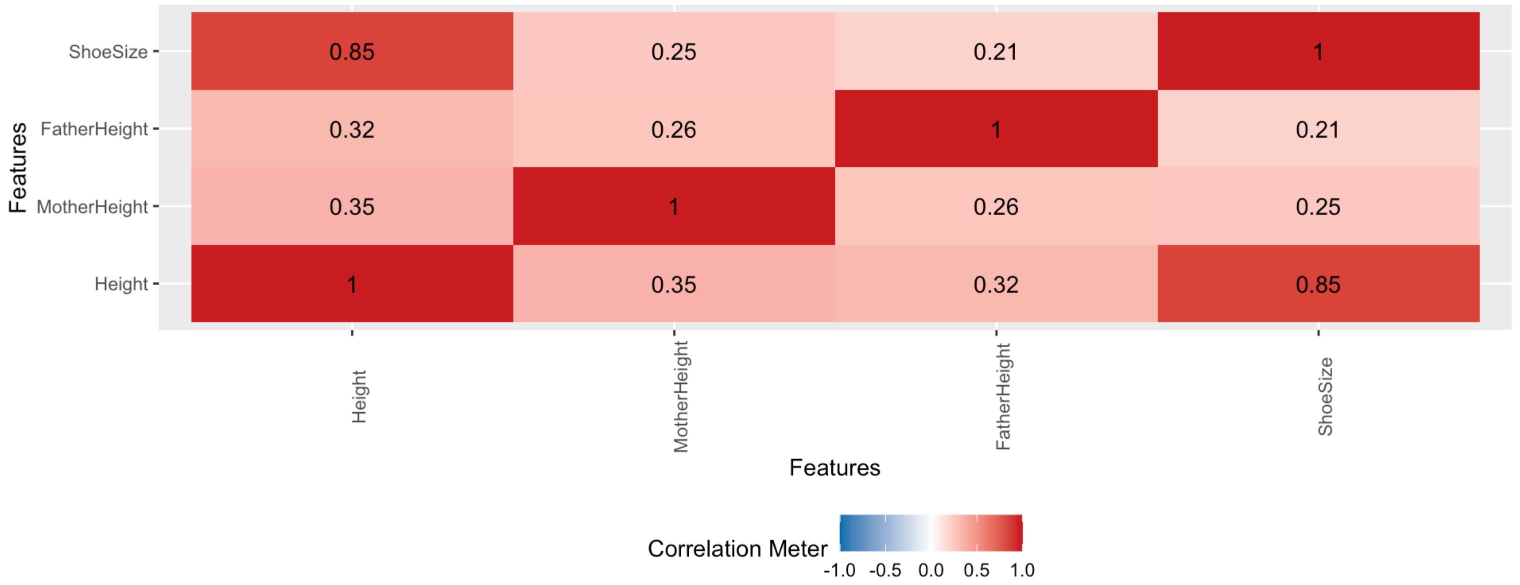
What do we do with the "Yes/No" variables?

- Encoding - the process of assigning categorical variables numerical values.
- One-hot-encoding (aka Dummy Variable encoding) - uses 1's and 0's.
- Yes=1, No=0 or Female=0, Male=1 (alphabetical)
- Much more on this in more stats classes but we'll keep in simple here.

EDA Tool #1 - Plots



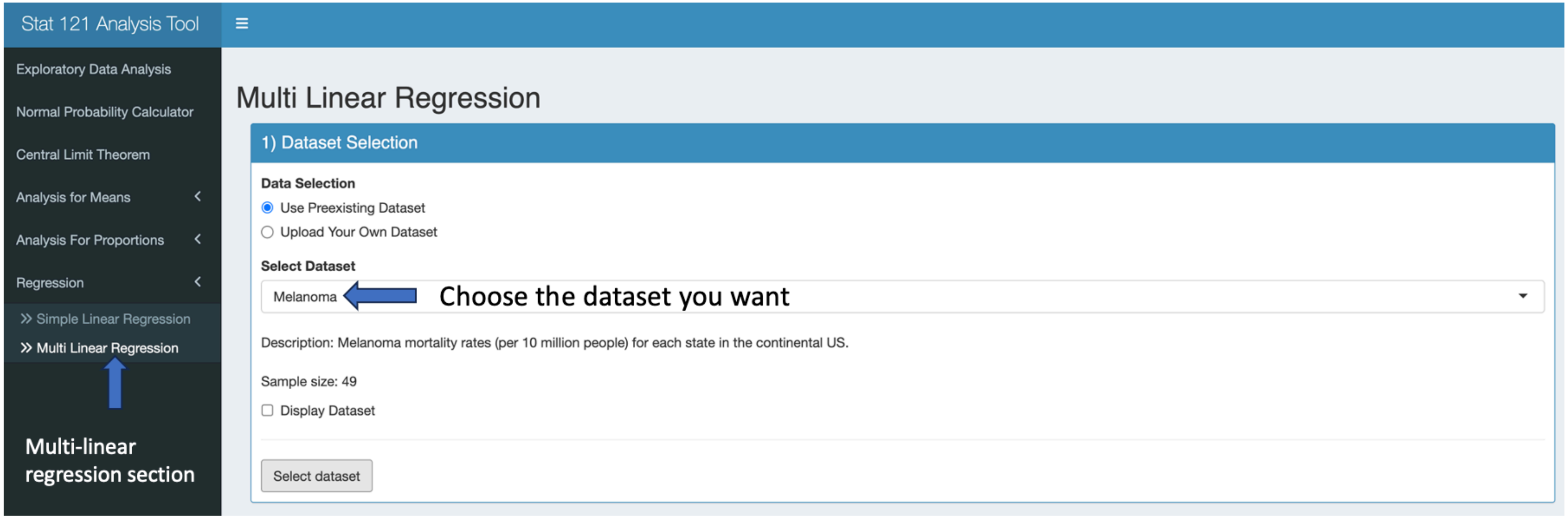
EDA Tool #2 - Correlations



<u>Reminder on Properties of Correlation (r):</u>

- $-1 < r < 1$
- Only appropriate for LINEAR relationships
- NOT impacted by scale of data (scale invariant).
- Highly impacted by outliers
- $\text{Cor}(X, Y) = \text{Cor}(Y, X)$

Using the Analysis Tool



Using the Analysis Tool

2) Select Variables

Please select up to 30 explanatory variables to use. Each explanatory variable should "explain" what happens to the response variable. NOTE: Only numeric variables will be given as options

Select Response Variable:

Mort

Set the response variable

Select Explanatory Variable(s):

Lat Ocean Long

Choose any explanatory variables – please READ questions carefully about what explanatory variables to use

Show 5 **entries**

Search:

	Lat	Ocean	Long
1	33	1	87
2	34.5	0	112
3	35	0	92.5
4	37.5	1	119.5
5	39	0	105.5

Showing 1 to 5 of 49 entries

Previous 1 2 3 4 5 … 10 Next

Proceed to EDA

Using the Analysis Tool

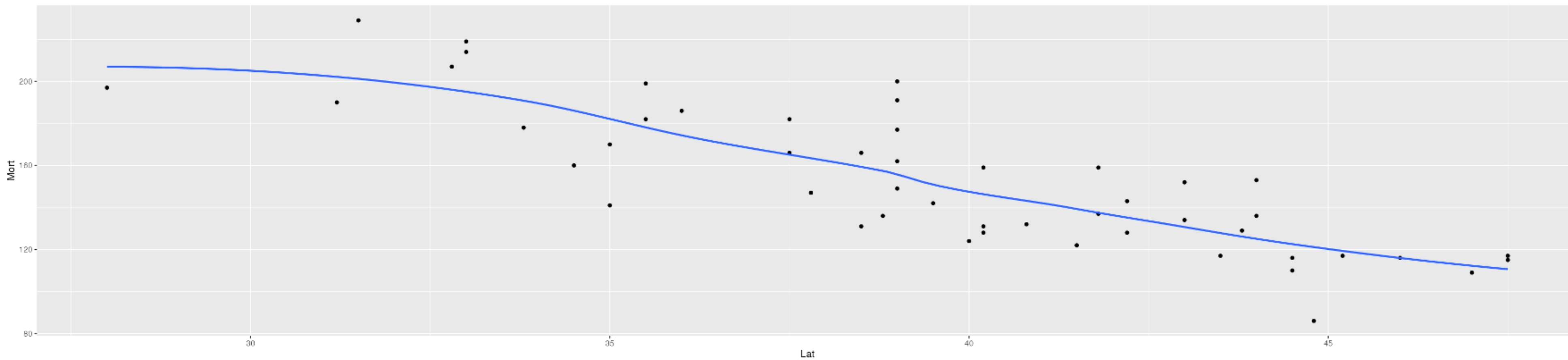
Because there are many variables, we have to explore them one at a time

3) Exploratory Data Analysis

Choose a variable to explore:

Lat — Choose the explanatory variable to explore

Create Scatterplot



Which numerical summary do you want to calculate?

5 Number Summary — Choose any numerical summary to look at

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.00   36.00   39.50   39.53   43.00   47.50
```

Proceed to Checking Assumptions

Multiple Linear Regression Model

In specifying a model for the *population* relationship between height and all the explanatory variables, we want to,

1. include all the variables at the same time,
2. keep it linear in all the variables at the same time,
3. account for the fact that the data is not a perfect relationship.

Multiple Linear Regression Model

An observation in our dataset

Residual (how far the observation is away from the line)

$$Y_i = \overbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_P X_{Pi}}^{\text{Follows a linear relationship with the explanatory variables}} + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma)$$

And common standard deviation ”sigma”

“come from”

Residuals

with mean 0

A normal distribution

where:

- X_{1i} is the i^{th} observation the of 1st explanatory variable
    - E.g X_{13} the mother’s height for the 3rd observation in our dataset.
- P = total number of explanatory variables you have

Multiple Linear Regression Model

$$\text{Height}_i = \beta_0 + \beta_1 \text{MH}_i + \beta_2 \text{FH}_i + \beta_3 \text{Sports}_i + \beta_4 \text{Sex}_i + \beta_5 \text{Shoe}_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

How do we interpret $\beta_0, \beta_1, \ldots, \beta_5$ (these are called slopes” or “effects”)?

- β_1 (MotherHeight): Holding everything else constant (or all else being equal), as the height of the mother goes up by 1, we expect height to go up by β_1 on average.

- β_2 (FatherHeight): Holding everything else constant (or all else being equal), as the height of the father goes up by 1, we expect height to go up by β_2 on average.

- β_3 (Sports): Holding everything else constant (or all else being equal), student’s who play sports in high school are expected to be β_3 inches taller than those who didn’t.

Multiple Linear Regression Model

$$\text{Height}_i = \beta_0 + \beta_1 \text{MH}_i + \beta_2 \text{FH}_i + \beta_3 \text{Sports}_i + \beta_4 \text{Sex}_i + \beta_5 \text{Shoe}_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

How do we interpret $\beta_0, \beta_1, \ldots, \beta_5$?

- β_5 (shoe size): Holding everything else constant (or all else being equal), as shoe size increases by 1, students get β_5 inches taller on average.

- β_0: Female student's whose parents are 0 inches tall, did not play sports in HS and wear a 0 shoe size, we expect their height to be β_0 on average.

Multiple Linear Regression Model

$$\text{Height}_i = \beta_0 + \beta_1 \text{MH}_i + \beta_2 \text{FH}_i + \beta_3 \text{Sports}_i + \beta_4 \text{Sex}_i + \beta_5 \text{Shoe}_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

INTERCHANGEABLE TERMINOLOGY WARNING! The $\beta_1, \ldots, \beta_5$ can be called any of the following:

- Slopes: “what is the slope for shoe size on height?”
- Effects: “what is the effect of shoe size on height?”
- Coefficients: “What is the coefficient for shoe size in the regression model?”

Assumptions of the MLR Model

Easy way to remember what we are assuming about the population in a multiple linear regression model:

- L - Linear relationship between y and all the quantitative x's simultaneously
- I - Independence (one obs. doesn't impact the other)
- N - Normal residuals (distance from "line" is normal)
- E - Equal spread of residuals around the "line"

More on why these assumptions are important and how to check these in the next subunit.

Parameter Estimation

Parameters we want to estimate: β_0 & $\beta_1, \ldots, \beta_P$ (which defines the line) and σ (so we know how spread out things are)

Goal: Find the predictions that goes "closest" to the data points.

Parameter Estimation

What do we mean by "line closest to points"? We want to find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_P$ so that:

$$\begin{aligned}\sum_{i=1}^{n}(\text{Obs}_i - \text{Pred}_i)^2 &= \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_P X_{Pi}))^2 \\ &= \sum_{i=1}^{n}(\text{residual}_i)^2\end{aligned}$$

is as small as possible. This is called the least squares regression line.

A few notes:

1. We "square" distances to account for "above" and "below" the line distances.
2. We sum squared residuals because we look at all the data.
3. We use "hats" to denote estimates from sample (for example, $\hat{\beta}_1$ is our estimate of β_1)
4. We include all the explanatory variables simultaneously.

Parameter Estimation

How do we find $\hat{\beta}_0, \ldots, \hat{\beta}_P$ that minimizes

$$\begin{aligned}\sum_{i=1}^{n}(\mathrm{Obs}_i - \mathrm{Line}_i)^2 &= \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_P X_{Pi}))^2 \\ &= \sum_{i=1}^{n}(\mathrm{residual}_i)^2?\end{aligned}$$

1. Guess and check
2. Use calculus

- In both cases, we'll let the computer do the hard work for us.

The Fitted MLR Model

Fitted MLR Model Output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.9398704	1.3214456	18.116426	0.0000000
MotherHeight	0.2294728	0.0166812	13.756401	0.0000000
FatherHeight	0.2445963	0.0155991	15.680122	0.0000000
SportsInHSYes	0.1241309	0.1183613	1.048746	0.2944567
SexMale	3.0174074	0.1361838	22.156871	0.0000000
ShoeSize	1.1225896	0.0375218	29.918346	0.0000000

Fitted Regression Line Equation:

$$\hat{y} = 23.94 + 0.23 \times \text{MotherHeight}_i + 0.24 \times \text{FatherHeight}_i + 0.12 \times \text{Sports}_i + \\ 3.02 \times \text{Sex}_i + 1.12 \times \text{ShoeSize}_i$$

The Fitted MLR Model

How do we interpret $\hat{\beta}_0 = 23.94$?

- β_0: For female children with 0 inch tall parents who do not play sports in HS and wear a 0 shoe size, we expect their height to be 27.28in on average.

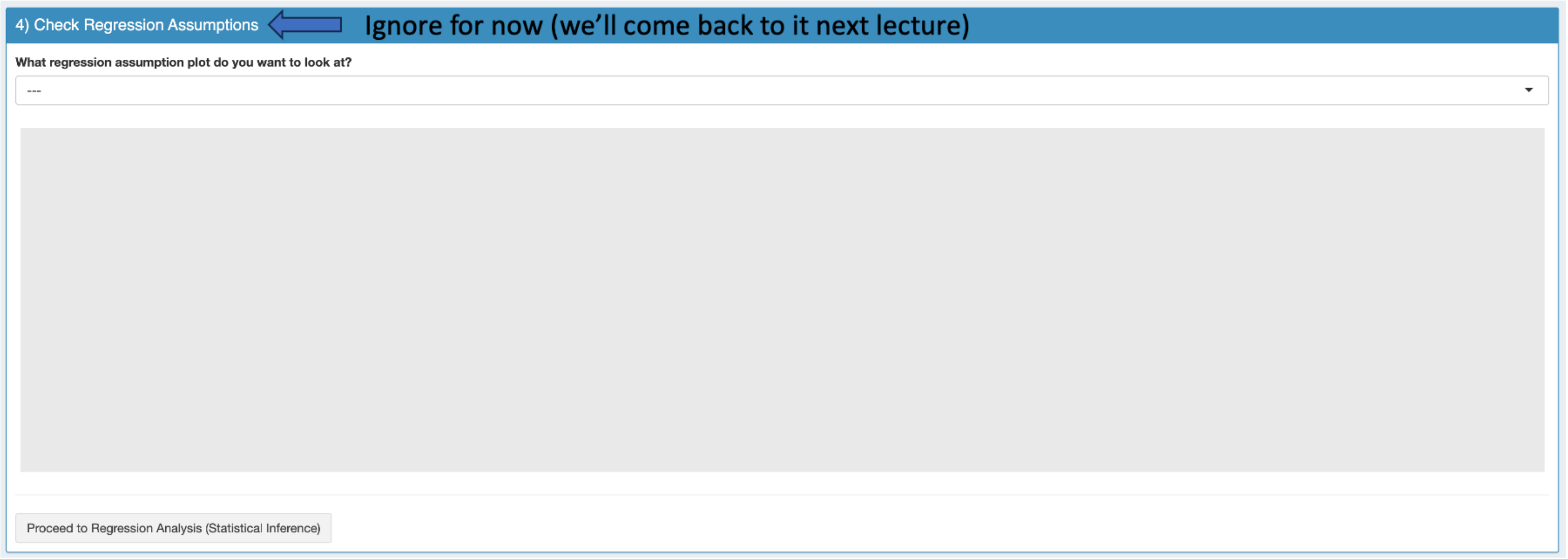
How do we interpret $\hat{\beta}_3 = 0.124$ (sports)?

- All else being equal, students who play sports in high school are 0.124 inches taller, on average.

How do we interpret $\hat{\beta}_5 = 1.123$ (shoe size)?

- Holding everything else constant (or all else being equal), as the shoe size goes up by 1, we expect height to go up by 1.123 on average.

Using the Analysis Tool



Using the Analysis Tool

5) Regression Analysis

```
Regression Analysis of: Mort (Y) explained by Lat, Ocean, Long (X's)
Coefficient Table:
```

Confidence Level for Slope and Intercept:

0.5 — 0.95 — 0.99

0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 0.99

Show 5 **entries**

	Test	F-statistic	p.value
1	F-test for all slopes are equal to zero	50.826	0

Showing 1 to 1 of 1 entries — Previous 1 Next

There is a lot going on here but for this subunit, all you need are the slopes (we'll get to the rest later)

Show 5 **entries**

	Estimate	Std. Error	t value	p value	CI Lower Bound	CI Upper Bound
(Intercept)	349.2369	27.0596	12.9062	0	294.7361	403.7377
Lat	-5.495	0.5289	-10.3898	0	-6.5602	-4.4297
Ocean	21.7976	5.2263	4.1707	0.0001	11.2712	32.324
Long	0.1219	0.1732	0.7037	0.4852	-0.227	0.4708

Showing 1 to 4 of 4 entries — Previous 1 Next

```
R-squared:  0.7721
sigma-hat:  16.4806
```

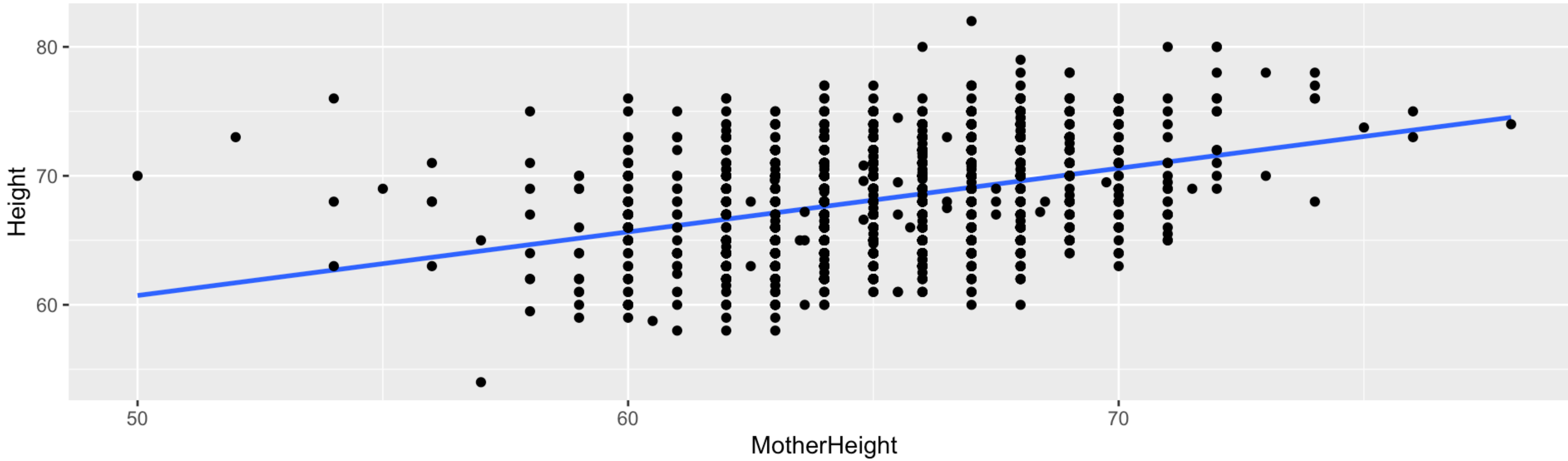
Proceed to Predictions

Fitted regression equation:

$$\hat{y} = 349.2369 - 5.495 \times \text{Lat} + 21.7976 \times \text{Ocean} + 0.1219 \times \text{Long}$$

Visualizing the Fitted MLR Model

When we only had 1 explanatory variable, we could visualize the fitted model:

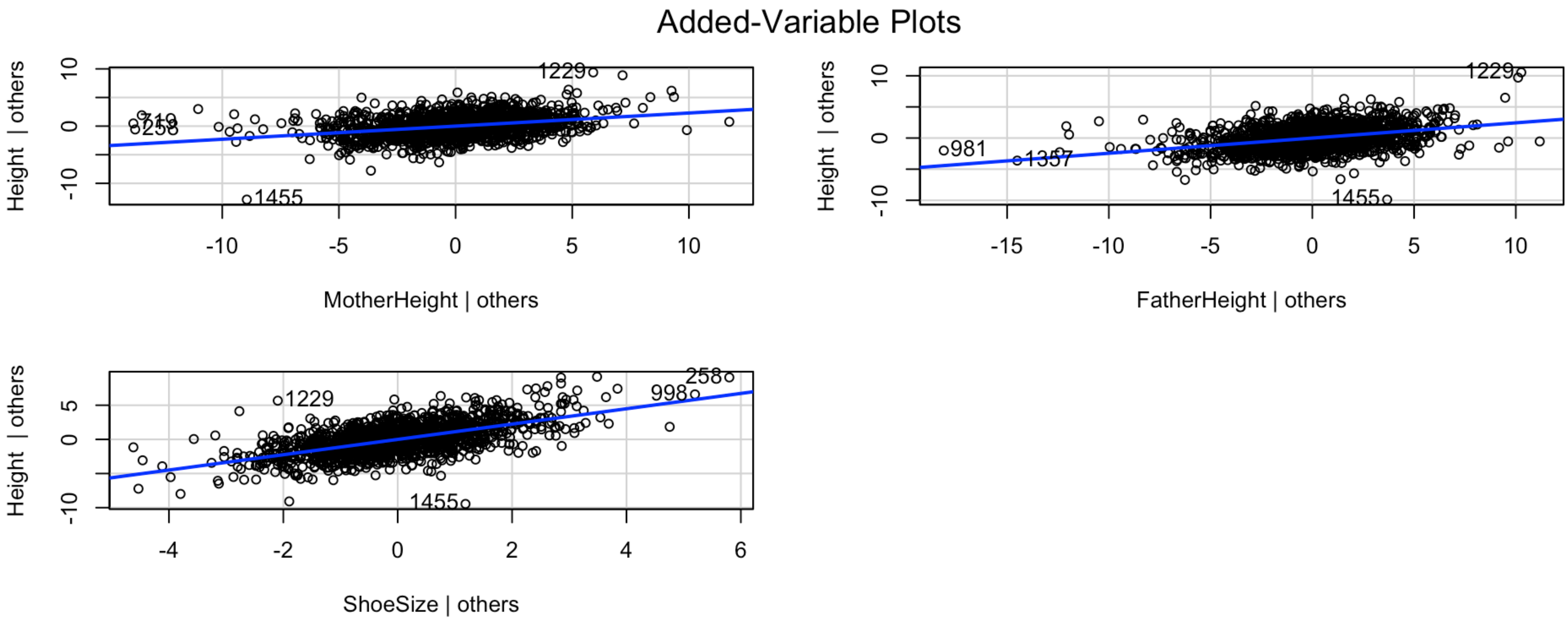


But we can't do that here because we have multiple explanatory variables that all work together.

Visualizing the Fitted MLR Model

Added variable plots (also known as partial regression plots):

- Intuition: Make a scatterplot of one x vs y AFTER "adjusting" for the other x's (math detail beyond this course so we'll just let the computer do it for us).



Parameter Estimation

An estimate of σ is more complicated to explain (take more stats courses), so for purposes of this class, the computer estimates it for us.

- $\hat{\sigma} = 1.809$

How do we interpret $\hat{\sigma}$?

- On average, the actual heights are about 1.809 far away from the estimated heights.
- Is this “better” or “worse” than if we just included mother’s height?
- $\hat{\sigma}$ = 3.896 if we only use mother height.
- It’s hard to tell just from $\hat{\sigma}$ how good a model is. A better measure is R^2.

Assessing Model Fit

Mathematical formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_P X_{Pi}))^2}{\sum_{i=1}^{n}(Y_i - \bar{y})^2} = 0.811$$

Intuition:

- Formal interpretation: The percent of variability in Y that is explained by all X's simultaneously.
- R^2 is between 0 and 1 with 1 meaning the explanatory variables perfectly explain the response.
- R^2 is a percentage grade on how well all the X's are doing in telling us about Y.
- For our study, 81.1% of the variation in student's height can be explained by mother's height, father's height, if you played sports in HS, biological sex and shoe size.

Using the Analysis Tool

5) Regression Analysis

```
Regression Analysis of: Mort (Y) explained by Lat, Ocean, Long (X's)
Coefficient Table:
```

Confidence Level for Slope and Intercept:

0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 0.99

Show 5 entries

	Test	F-statistic	p.value
1	F-test for all slopes are equal to zero	50.826	0

Showing 1 to 1 of 1 entries

Previous 1 Next

Show 5 entries

	Estimate	Std. Error	t value	p value	CI Lower Bound	CI Upper Bound
(Intercept)	349.2369	27.0596	12.9062	0	294.7361	403.7377
Lat	-5.495	0.5289	-10.3898	0	-6.5602	-4.4297
Ocean	21.7976	5.2263	4.1707	0.0001	11.2712	32.324
Long	0.1219	0.1732	0.7037	0.4852	-0.227	0.4708

Showing 1 to 4 of 4 entries

Previous 1 Next

```
R-squared:  0.7721
sigma-hat:  16.4806
```

Proceed to Predictions

Additional MLR Practice

Measuring possum head size can be difficult. What is the relationship between possum head size and sex, age, skull width, total length and tail length? Use a multiple linear regression model (and the course app) to answer the following questions:

1. What is the estimated head size for a newborn, female possum with 0 skull width, length and tail length?
2. How much should head size go up (or down) as the possum gets 1 cm bigger?
3. How much are male head sizes bigger (or smaller) than female head sizes (on average)?
4. On average, how far away are true head sizes from estimated head sizes?
5. How well do the explanatory variables explain head size?

Additional MLR Practice

1. What is the estimated head size for a newborn, female possum with 0 skull width, length and tail length?
   - $\hat{\beta}_0 = 33.4974481$
2. How much should head size go up (or down) as the possum gets 1 cm bigger?
   - $\hat{\beta}_1 = 0.4528877$
3. How much are male head sizes bigger (or smaller) than female head sizes (on average)?
   - $\hat{\beta}_2 = 1.1695384$
4. On average, how far away are true head sizes from estimated head sizes?
   - This is the $\hat{\sigma} = 2.080432$
5. How well do the explanatory variables explain head size?
   - This is $R^2 = 0.669$

Homework Choices for Unit 7

Same as Unit 6 but we’re going to add more variables to the regression:

1. Rate my professor - what matters in determining a rate my professor score?
2. Supervisor - what makes people like their manager?
3. Body Fat - what body measurements are predictive of your BMI?
4. Basketball Salary - what skills lead to a higher salary?

Key Terminology

- EDA for MLR
- Multiple linear regression model
- R^2
- Interpretation of Coefficients
- Added-variable Plots
- Least squares estimation