SDS358: Applied Regression Analysis

Day 11: Multiple Regression

Dr. Michael J. Mahometa

Just a reminder:

- **Unit 1 Exam** on the 1st (*NEXT MONDAY*)

Agenda for Today:

- Multiple Regression
  - Expanding our Simple Linear Regression
  - What stays the same?
  - What changes?
- Interpretation
- Slopes and CI
- Standardized Betas

In our story so far…

- Simple Linear Regression extends Pearson.
  - It tells us about the slope (impact) of a *single* predictor in the model
  - But what if we added another predictor (or *more*) to the model?

Two predictors

- Let's take a look at some data.

```
exams <- read_csv("../data/CourseExams.csv")
vars <- c("ExamTotal", "Absences", "HoursStudied")
cor(select(exams, one_of(vars)))
```

```
##              ExamTotal   Absences HoursStudied
## ExamTotal    1.0000000 -0.4495161    0.5861130
## Absences    -0.4495161  1.0000000   -0.5558213
## HoursStudied 0.5861130 -0.5558213    1.0000000
```


Two Predictors

- Let's take a look at some data.

```
library(psych)
corr.test(select(exams, one_of(vars)))$p
```

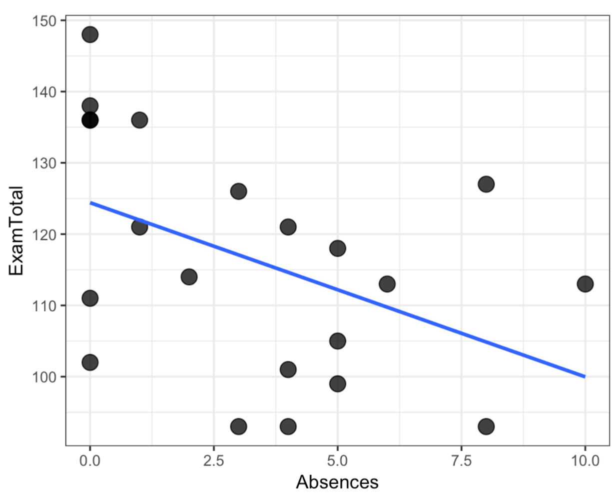
```
##                ExamTotal    Absences HoursStudied
## ExamTotal    0.000000000 0.040913219   0.01570517
## Absences     0.040913219 0.000000000   0.01778421
## HoursStudied 0.005235056 0.008892105   0.00000000
```

A Simple Question:

Primary Resarch Question:
Do Hours Studies and Number of Absences significantly predict Total Exam Scores? If so, what are their individual effects towards the prediction of Total Exam Scores?

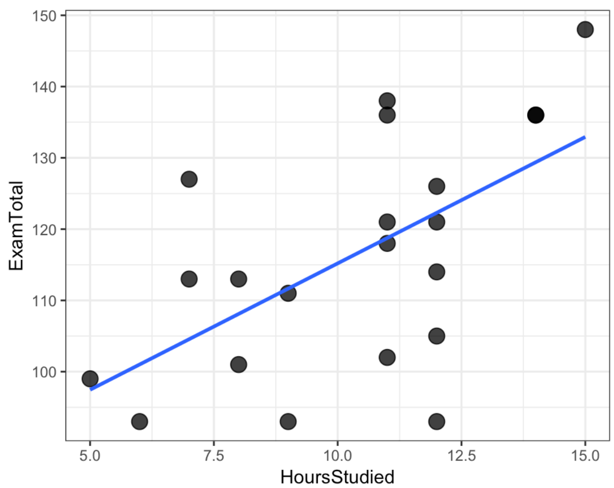
Two Simple Models

```
## (Intercept)    Absences
##  124.409302   -2.443411
```



Two Simple Models

```
##  (Intercept) HoursStudied
##    79.727739     3.547085
```



Each gives us a model….

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = b_0 + b_2 z$$

Each gives us a model….

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = b_0 + b_2 z$$

Can we simply add these simple models together?

$$\hat{y} = b_0 + b_0 + b_1 x + b_2 z$$

That just doesn't work…

```
lm(ExamTotal ~ Absences, data=exams)$coefficients
```

```
## (Intercept)    Absences
##  124.409302   -2.443411
```

```
lm(ExamTotal ~ HoursStudied, data=exams)$coefficients
```

```
##  (Intercept) HoursStudied
##    79.727739     3.547085
```

```
lm(ExamTotal ~ Absences + HoursStudied, data=exams)$coefficients
```

```
##  (Intercept)     Absences HoursStudied
##   89.1496997   -0.9733096    2.9447681
```

Multiple (Trivariate) regression

- We can expand or underlying regression model from the population:
- From this:

$$y = \beta_0 + \beta_1 x_1 + e$$

- To this:

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

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

Multiple (Trivariate) regression

- *And* we can extend it to our sample (just like in Simple Regression)…

$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

And eventually to this:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \ldots b_k x_k$$

Amazing

```
lm(ExamTotal ~ Absences + HoursStudied, data=exams)$coefficients
```

```
##  (Intercept)     Absences HoursStudied 
##   89.1496997   -0.9733096    2.9447681
```


Not a line, but a *plane*

- Instead of a single line going through the data, the model tells us of a *plane* that is used in the prediction.

Remember our Algebra for SLR:

$$b_1 = r\frac{S_y}{S_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Simple Algebra for MLR (trivariate):

$$b_1 = \left(\frac{r_{yx} - r_{yz}r_{xz}}{1 - r_{xz}^2}\right)\left(\frac{S_y}{S_x}\right)$$

$$b_2 = \left(\frac{r_{yz} - r_{yx}r_{xz}}{1 - r_{xz}^2}\right)\left(\frac{S_y}{S_z}\right)$$

$$b_0 = \bar{y} - b_1\bar{x} - b_2\bar{z}$$

What else changes?

- Just the df of the model:

$$\hat{\sigma}^2 = \frac{\Sigma(y_i - \hat{y}_i)^2}{\text{n - k - 1}}$$

- Now, instead of k = 1, we have k > 1

What else changes?

- Just the df of the model:

$$\hat{\sigma}^2 = \frac{\Sigma(y_i - \hat{y}_i)^2}{\text{n - p}}$$

- Now, instead of p = 2, we have p > 2

What stays the same?

Assumptions:

- Independence: Error associated with each data point is independent of every other value.
- The population mean of e is 0.
  - For a given value of x, the population variance of e is: σ_e^2
  - For a given value of x, e has a normal distribution.
- Homoscedasticity

What changes?

Assumptions:

- Independence: Error associated with each data point is independent of every other value.
- The population mean of e is 0.
  - For a given value of x, the population variance of e is: σ_e^2
  - For a given value of x, e has a normal distribution.
- Homoscedasticity
- **NEW for MLR:** No *multicolinearity* of independents.

We STILL have an ANOVA model for F…

- We're still comparing Model Error to Residual Error
- STILL have Total error

$$\Sigma(y_i - \bar{y}_i)^2$$

```
exams$dev <- exams$ExamTotal - mean(exams$ExamTotal, na.rm=TRUE)
exams$dev_sq <- exams$dev^2
dev_sq <- sum(exams$dev_sq)
dev_sq
```

```
## [1] 5444.952
```


We STILL have an ANOVA model for F…

- STILL have Model error

$$\Sigma(\hat{y}_i - \bar{y}_i)^2$$

```
e_mod <- lm(ExamTotal ~ Absences + HoursStudied, exams)
exams$fit <- predictValues(e_mod)
exams$mod <- exams$fit - mean(exams$ExamTotal, na.rm=TRUE)
exams$mod_sq <- exams$mod^2
mod_sq <- sum(exams$mod_sq)
mod_sq
```

```
## [1] 1991.142
```

We STILL have an ANOVA model for F…

- STILL have Residual error

$$\Sigma(y_i - \hat{y}_i)^2$$

```
exams$res <- exams$ExamTotal - exams$fit
exams$res_sq <- exams$res^2
res_sq <- sum(exams$res_sq)
res_sq
```

```
## [1] 3453.811
```


And our ANOVA table

Source	Sums of Squares	df	Mean Squares	F-value
Regression	1991.142	?		
Error	3453.811	?		
Total	5444.952	?		

And our ANOVA table

Source	Sums of Squares	df	Mean Squares	F-value
Regression	1991.142	k		
Error	3453.811	n-k-1		
Total	5444.952	n-1		

- What is the F-value for the Overall Model?

And our ANOVA table

Source	Sums of Squares	df	Mean Squares	F-value
Regression	1991.142	2	995.57	5.19
Error	3453.811	18	191.88	
Total	5444.952	20		

And our ANOVA table

From R

```
e_mod <- lm(ExamTotal ~ Absences + HoursStudied, data=exams)
simpleAnova(e_mod)
```

```
## Analysis of Variance Table
##
## Response: ExamTotal
##            Df Sum Sq Mean Sq F value  Pr(>F)
## Predictors  2 1991.1  995.57  5.1886 0.01662 *
## Residuals  18 3453.8  191.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


And our ANOVA table

From R

```
e_mod <- lm(ExamTotal ~ Absences + HoursStudied, data=exams)
summary(e_mod)
```

```
##
## Call:
## lm(formula = ExamTotal ~ Absences + HoursStudied, data = exams)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.5937  -8.5403   0.4312   9.0768  25.0234
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   89.1497    16.9747   5.252  5.4e-05 ***
## Absences      -0.9733     1.2275  -0.793    0.438
## HoursStudied   2.9448     1.3666   2.155    0.045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.85 on 18 degrees of freedom
## Multiple R-squared:  0.3657, Adjusted R-squared:  0.2952
## F-statistic: 5.189 on 2 and 18 DF,  p-value: 0.01662
```

Our Proportion of Variance Accounted For

$$R^2 = \frac{SSModel}{SSTotal}$$

```
mod_sq / dev_sq
```

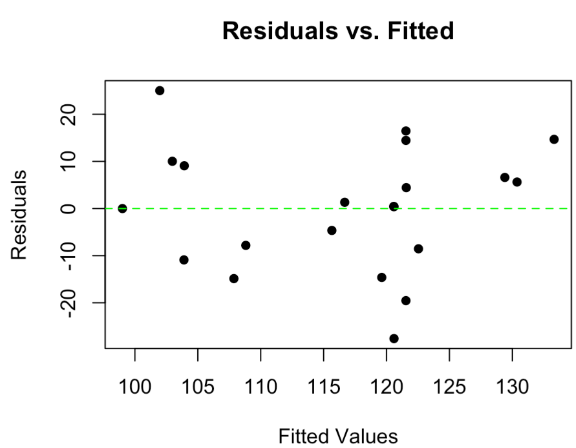
```
## [1] 0.3656858
```

- And how do we interpret this *new* R^2? (Notice it's also *not* r^2.)

Homoscedasticity

- We can *still* look for Homoscedasticity:

```
residFitted(e_mod)
```

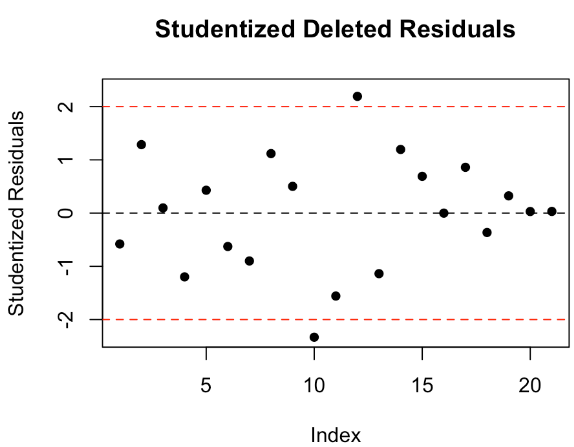


Outlier Checks

- Outliers are based on two things:
  - Residuals
  - Leverage
- Outlier checks *will still work* for MLR

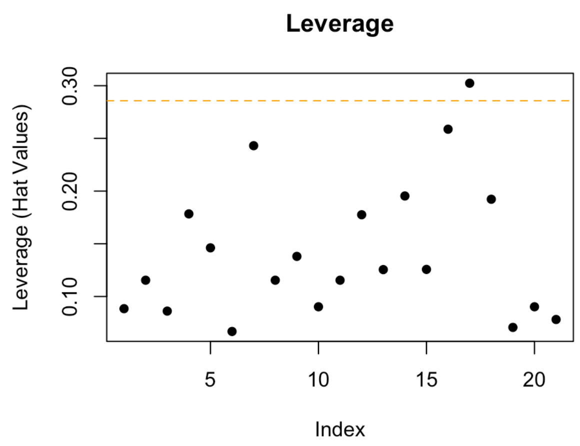
Studentized Deleted Residuals

```
studResidPlot(e_mod)
```



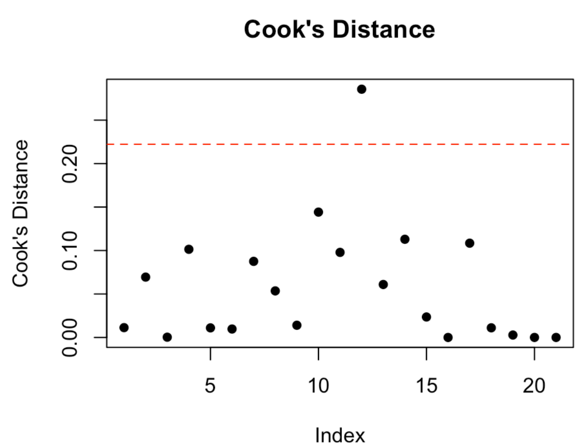
Leverage

```
levPlot(e_mod)
```



Cook's Distance

```
cooksPlot(e_mod)
```



Now, something ACTUALLY changes

- First, our interpretations
- The intercept
  - Now, it's the "constant coefficient"
  - When all coefficient values are zero
- The slope(s) (or regression coefficients)
  - Now, they are the "partial regression coefficients"
  - b_k is the effect on y after being adjusted by the other predictor(s)
  - OR, b_k is the effect on y while holding the other predictor(s) constant (fixed)

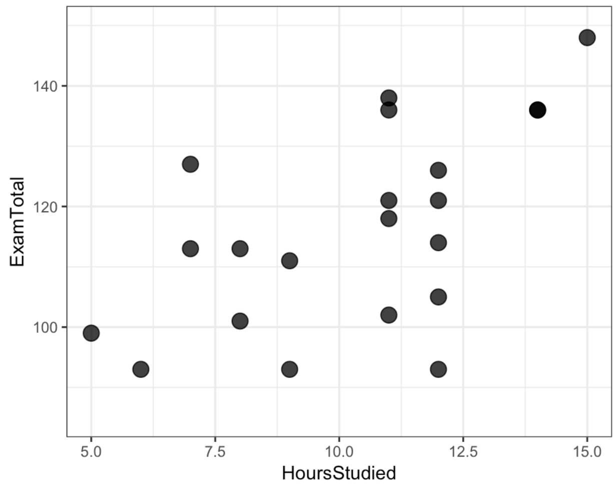
Let's see this in *Action*:

- Let's look at Hours Studied predicting Exam Total.

```
e_mod
```

```
##
## Call:
## lm(formula = ExamTotal ~ Absences + HoursStudied, data = exams)
##
## Coefficients:
##  (Intercept)      Absences  HoursStudied
##      89.1497       -0.9733        2.9448
```

Let's see this in *Action*:



Our Mission:

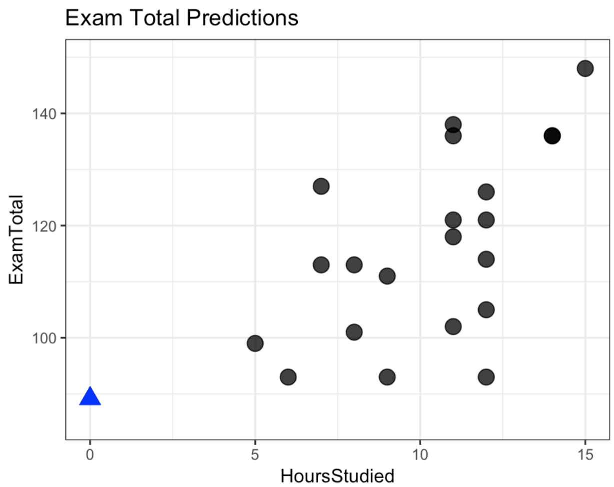
- We want to put a *predicted* line on the graph…to capture the effect of Hours Studied on the Exam Total score.
- We'll use:

$$\hat{y} = 89.15 + -0.9733b_1 + 2.9448b_2$$

$$\hat{y} = 89.15 + (-0.9733 \times \text{Absences}) + (2.9448 \times \text{Hours Studied})$$

- First: Where does the line begin?

Predicting Exam Total

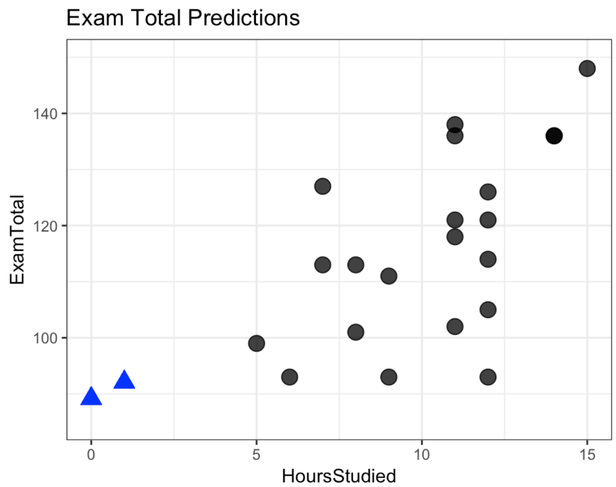


Predicting Exam Total

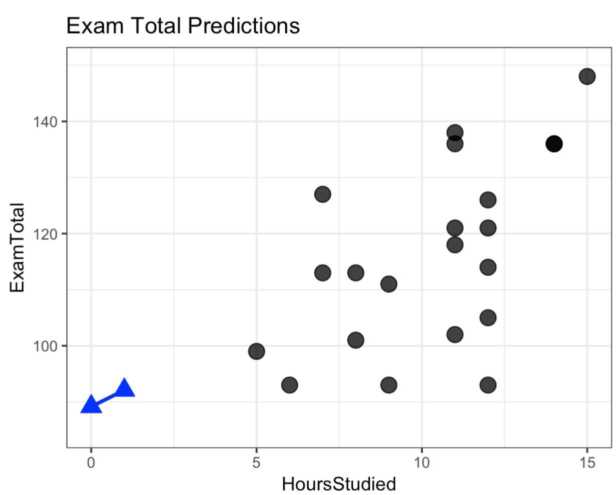
- Where does the line go to?
- Let's predict the outcome at **one** hour of Hours Studied:
- Remember: In the model, the *effect* of any predictor on the outcome is while holding all other predictors *constant*.

$$\hat{y} = 89.15 + (-0.9733 \times \text{Absences}) + (2.9448 \times \text{Hours Studied})$$

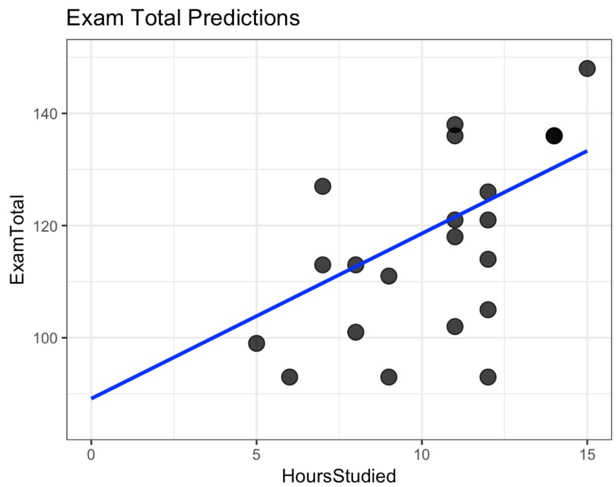
Predicting Exam Total



Predicting Exam Total



Predicting Exam Total

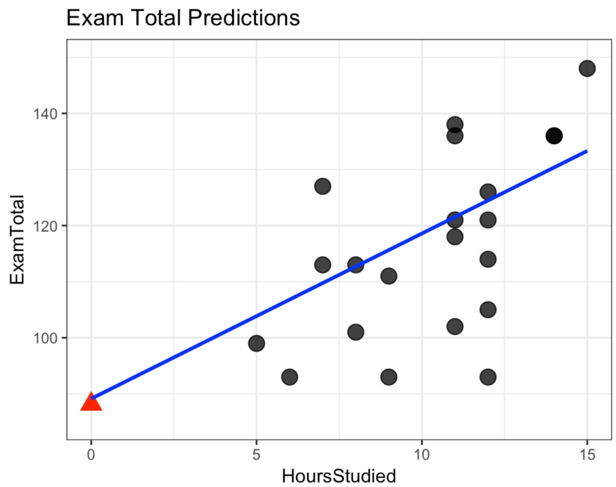


Let's Investigate

- Since we predicted the Exam Total when Absences was *constant* at 0, let's see what happens when Absences is moved to 1.
- Again, where does the line begin?
- Use:

$$\hat{y} = 89.15 + (-0.9733 \times \text{Absences}) + (2.9448 \times \text{Hours Studied})$$

Let's Investigate

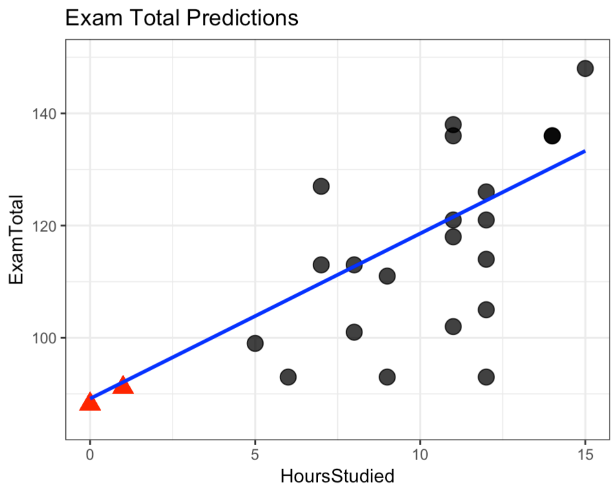


Let's Investigate

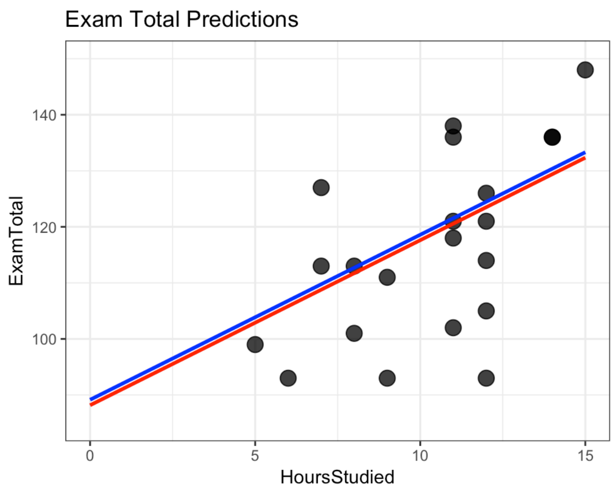
- Where does the line go to?
- Let's predict the outcome at **one* hour of Hours Studied:
- Remember: In the model, the *effect* of any predictor on the outcome is while holding all other predictors *constant*.

$$\hat{y} = 89.15 + (-0.9733 \times \text{Absences}) + (2.9448 \times \text{Hours Studied})$$

Let's Investigate



Let's Investigate



Riddle me this:

```
e_mod
```

```
##
## Call:
## lm(formula = ExamTotal ~ Absences + HoursStudied, data = exams)
##
## Coefficients:
##  (Intercept)      Absences  HoursStudied
##      89.1497       -0.9733        2.9448
```

- What's the (vertical) distance between the two lines on the previous graph?
- What would be the distance between the blue line (Absences = 0) and a new line holding Absences constant at 5?

What's the "Best Practice"?

- What do you think we *should* hold the variable(s) that we are *not* graphing (on the x-axis) constant at?
- What value universally makes sense?

What's the "Best Practice"?

```
library(emmeans)
ref_grid(e_mod)
```

```
## 'emmGrid' object with variables:
##     Absences = 3.2857
##     HoursStudied = 10.333
```

- We'll cover prediction in *more depth* (and the `emmeans()` function) on Wednesday, but to see quickly:

What's the "Best Practice"?