Agenda

1. Finish notes from Monday
2. Multiple regression

Multiple Regression Multiple regression is a natural extension of simple linear regression.

- SLR: one response variable, one explanatory variable

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon, \text{ where } \epsilon \sim N(0, \sigma_\epsilon)$$

- MLR: one response variable, *more than one* explanatory variable

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p + \epsilon, \text{ where } \epsilon \sim N(0, \sigma_\epsilon)$$

- Estimated coefficients (e.g. $\hat{\beta}_i$'s) now are interpreted in relation to (or "conditional on") the other variables
- β_i reflects the *predicted* change in Y associated with a one unit increase in X_i, conditional upon the rest of the X_i's.
- R^2 has the same interpretation (proportion of variability explained by the model)

Multiple Regression with a Categorical Variable Consider the case where X_1 is quantitative, but X_2 is an *indicator* variable that can only be 0 or 1 (e.g. $coffeeTemp$). Then,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot X_2$$

So then,

$$\begin{aligned}
\text{For hot coffee,} \qquad \hat{Y}|_{X_1, X_2=0} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 \\
\text{For cold coffee,} \qquad \hat{Y}|_{X_1, X_2=1} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot 1 \\
&= \left(\hat{\beta}_0 + \hat{\beta}_2\right) + \hat{\beta}_1 \cdot X_1
\end{aligned}$$

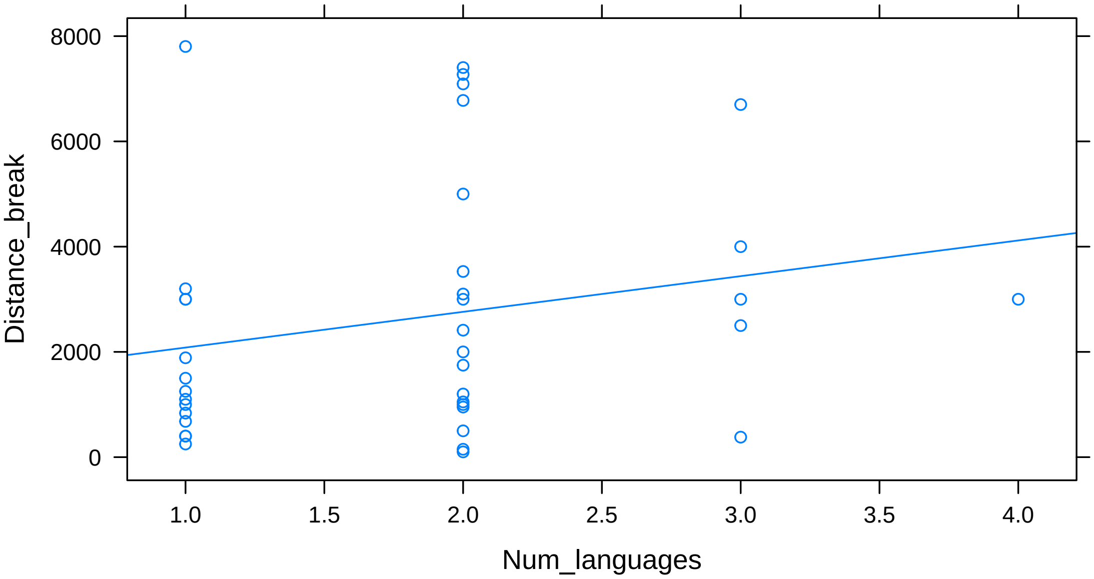
This is called a *parallel slopes* model. [Why?]

Example: Class data Let's think back to the data we collected at the beginning of class again. We're going to try to predict how far from Smith you each were. We could start with a simple linear regression model,

```
m1 <- lm(Distance_break~Num_languages, data=us)
coef(m1)

##   (Intercept) Num_languages
##     1404.6036      678.2473

plotModel(m1)
```



1. Write out the equation of the line

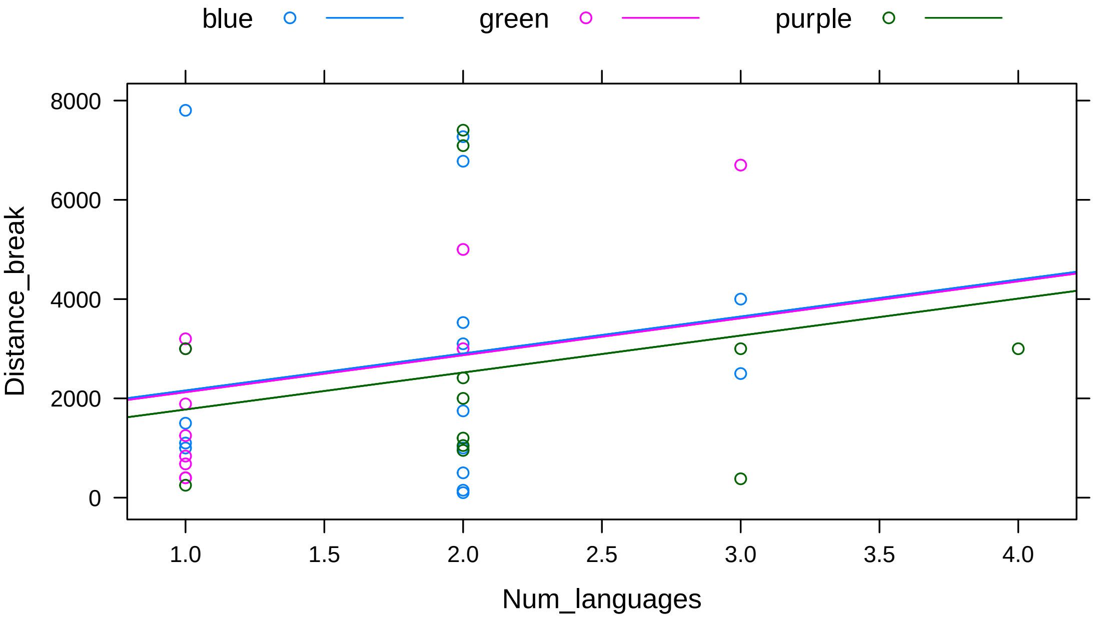
2. Interpret the slope and intercept of the line

But, maybe we want to include some additional information. We can add variables into our linear model:

```
m2 <- lm(Distance_break~Num_languages+Sheet_color, data=us)
summary(m2)

##
## Call:
## lm(formula = Distance_break ~ Num_languages + Sheet_color, data = us)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2885.6 -1464.4  -767.5   812.0  5643.1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1415.17    1201.08   1.178    0.247
## Num_languages       744.70     555.15   1.341    0.189
## Sheet_colorgreen    -34.75     974.41  -0.036    0.972
## Sheet_colorpurple  -383.68     925.93  -0.414    0.681
##
## Residual standard error: 2352 on 34 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.05534,Adjusted R-squared:  -0.02801
## F-statistic: 0.664 on 3 and 34 DF,  p-value: 0.58

plotModel(m2)
```



1. Write out the equation of the line

2. Interpret all the coefficeints in the model

3. Find the predicted value for yourself

4. Find your residual– is it negative or positive?