Regression

Matthew Sigal

December 2nd, 2015

Linear and Multiple Regression

Ordinary Least Squares Regression is a fundamental statistical method that underlies most methods of analysis. In fact, t-tests and ANOVA are special cases of regression.

The central concept is that we have **one or more variables** and we wish to see if they can be used to explain the variance of a **response variable**.

Example Dataset

gender	mathPre	readPre	group	score1	score2
Female:104	Min. :20.9	Min. :27.5	cram :105	Min. :24.9	Min. :16.4
Male :196	1st Qu.:41.2	1st Qu.:43.7	mnemonic:104	1st Qu.:43.3	1st Qu.:43.7
	Median :51.1	Median :50.3	none : 91	Median :49.3	Median :53.1
	Mean :50.0	Mean :50.3		Mean :49.5	Mean :53.8
	3rd Qu.:57.3	3rd Qu.:56.9		3rd Qu.:55.0	3rd Qu.:63.8
	Max. :77.7	Max. :84.2		Max. :77.5	Max. :95.5
	NA's :2	NA's :3			

The OLS Regression Model

Goal: In simple linear regression, we want to predict one variable (Y; the "outcome") based upon the value of another variable (X; the "predictor").

On Causality

Even if this relationship is significant, we **cannot** say that one variable *causes* the other to occur.

We usually want to believe that treatment causes the response, but unless we conduct a true experiment (with random allocation), we can't make that claim with confidence. There could be other factors or the relationship might even be in the opposite direction!

The OLS Regression Model

The simple linear regression model looks like this:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

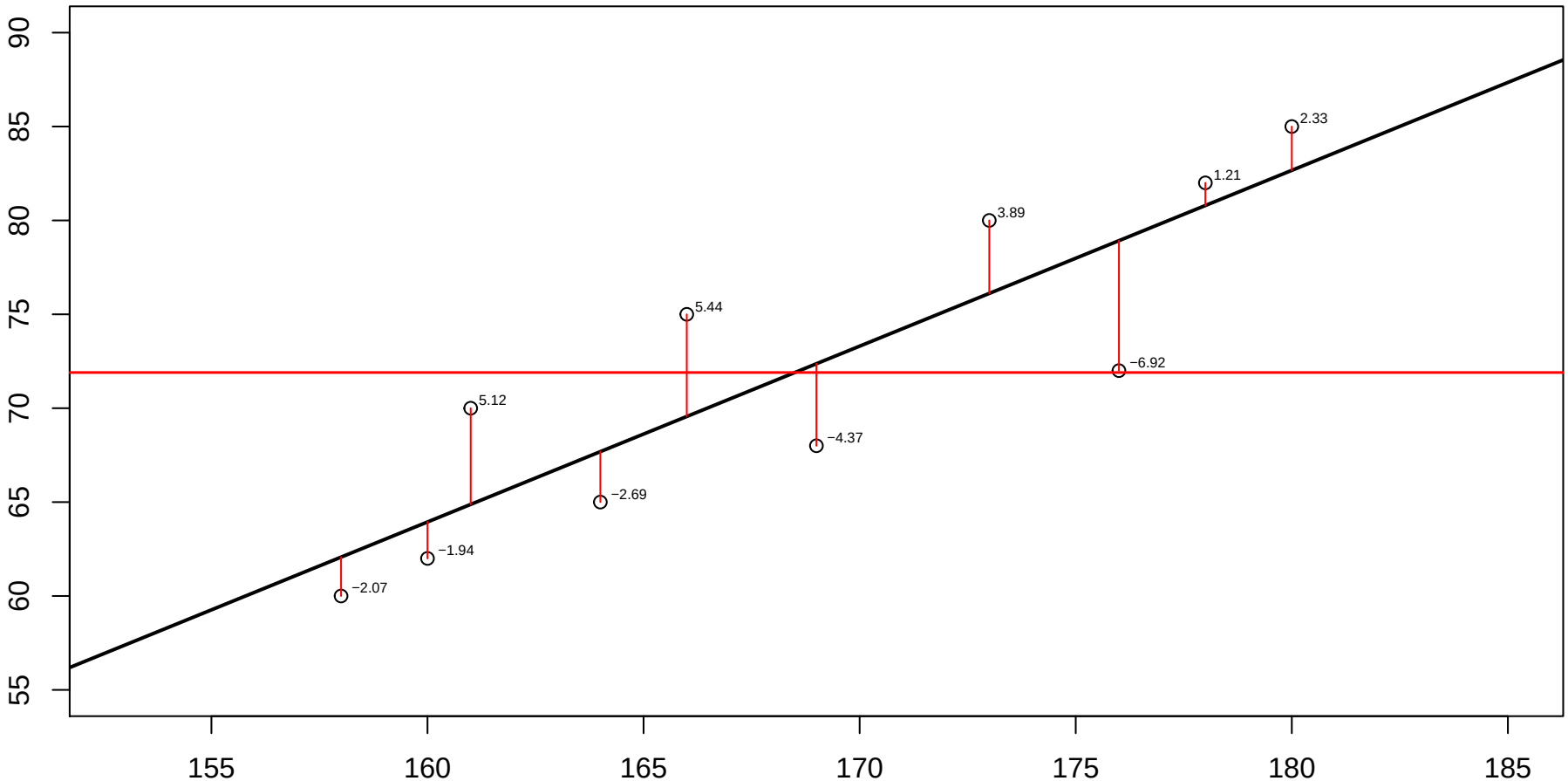
Where:

- Y_i is the value of the outcome for the i^{th} observation;
- X_i is the value of the predictor for the i^{th} observation;
- β_0 is the intercept;
- β_1 is the slope; and
- ϵ_i is an error term, based upon the difference between our model predicted Y and the actual Y for the i^{th} observation.

The OLS Regression Model

In Least Squares Regression the estimates of the slope and the intercept are those that minimize the **sum of the squares of the residual** (ϵ_i) terms.

In other words, we want to find the line that minimizes the distances between it and our Y scores.



The OLS Regression Model

A line is made up of two properties:

- a slope (β_1); and,
- an intercept (β_0):

To find the slope: $\beta_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$ or $\frac{SS_{xy}}{SS_x}$.

To find the intercept: $\beta_0 = \bar{y} - b\bar{x}$

The OLS Regression Model

In words: the **slope** looks at the consistency in the relationship between X and Y, and will tell us: if we change X by 1 unit, how much does Y change?

If the slope is 0, our best guess for Y at every level of X is $\bar{Y}$!

This is intricately related to **correlation**. In fact, the slope can also be calculated as: $\beta_1 = r\left(\frac{S_y}{S_x}\right)$

The OLS Model

Overall, OLS posits that we can take any response variable, and look at its:

- **total variation**: $\Sigma (y_i - \bar{y})^2$, which can then be partitioned into:
  - **predictable variation**, $\Sigma (\hat{y}_i - \bar{y})^2$ (SS regression); and,
  - **unexplained variation** or error, $\Sigma (y_i - \hat{y}_i)^2$ (SS residual)

In an example, if every point fell on the regression line, which would be larger: SS regression or SS residual?

Effect Size

The ratio of SS regression to SS total ($\frac{SS_{REG}}{SS_{TOTAL}}$) yields R^2, or the proportion of Y's variance that is explained by X.

Conducting Linear Regression in R

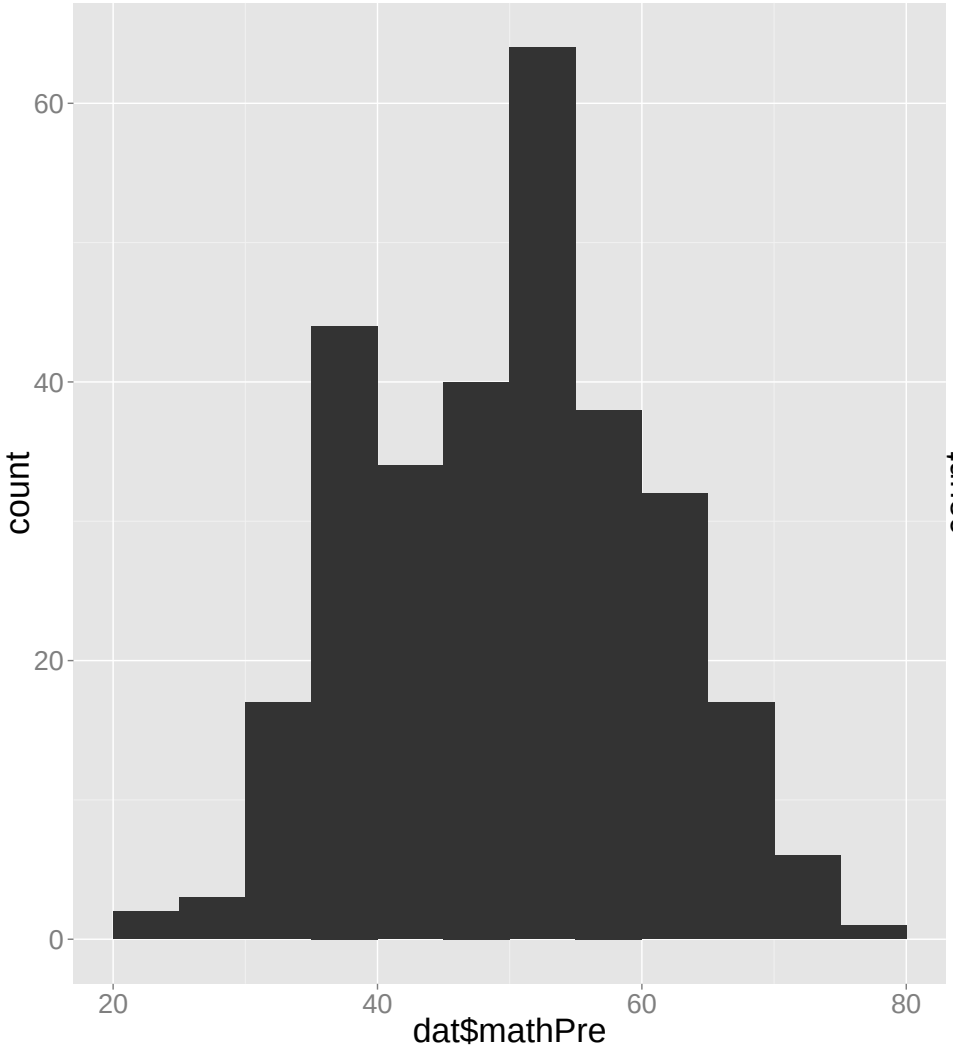
Linear Regression Example

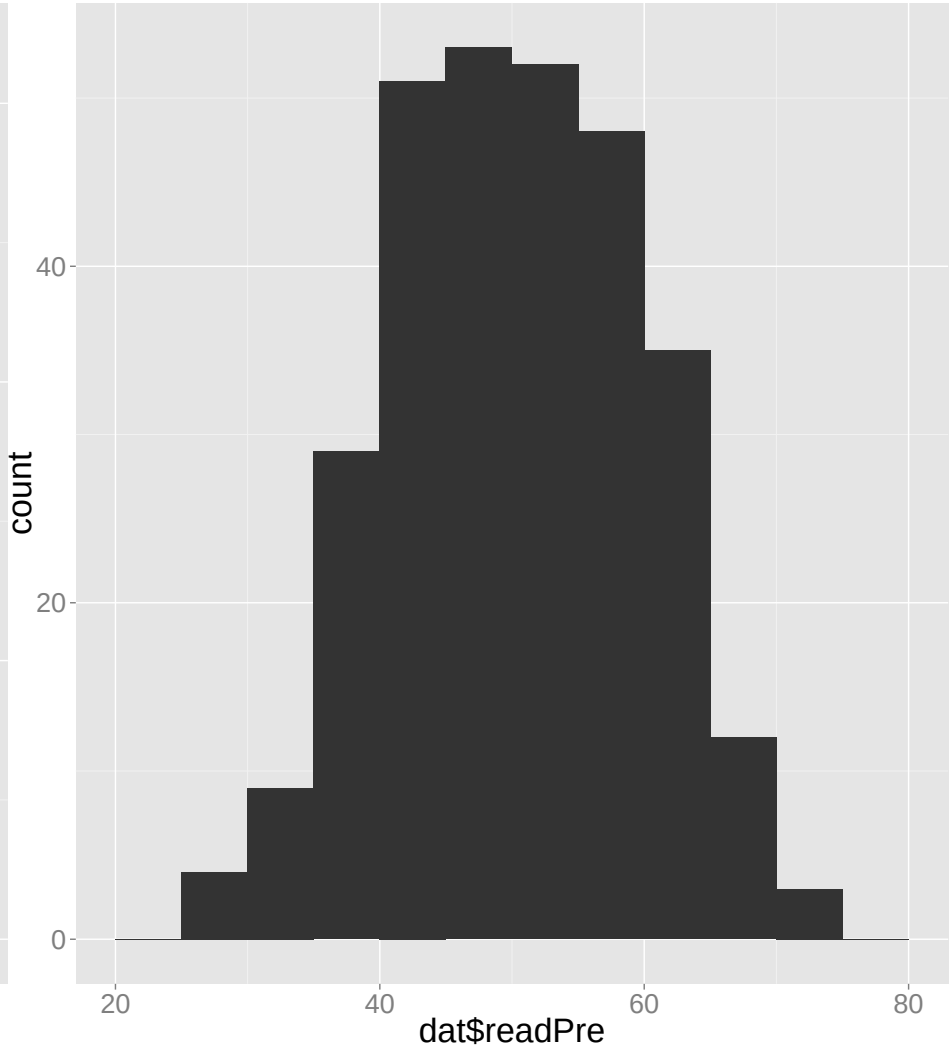
In `dat`, we have a few continuous variables. We might ask can reading scores (`readPre`) be used to predict math scores (`mathPre`)?

In this research hypothesis, we have declared: **two** variables of interest, with primary emphasis being placed on **math**, which is being predicted by **reading** . As such: `mathPre` is our "response variable" Y, and `readPre` is our "predictor variable", X.

Since we are just playing with data, we could have chosen other variables (can `score2` be predicted by `score1`?). In practical research, you should design your study or survey with full knowledge of what is going to be your response and what is going to be your predictors.

Visualization





Visualization

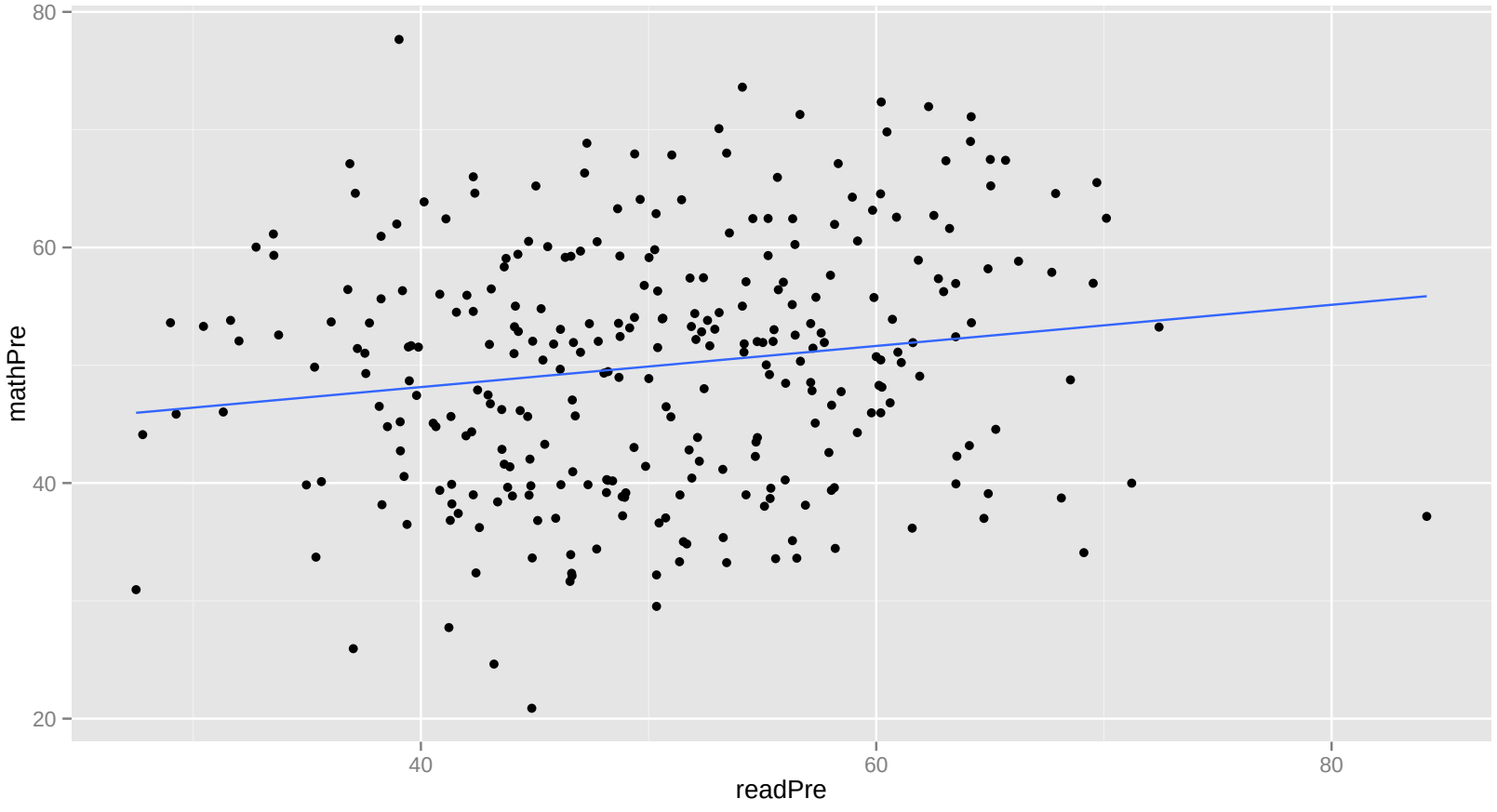
While both of these variables are relatively normal (which is good), we should note that in regression it is **not** the individual variables that need to be normally distributed, but the **residuals** (ϵ_i), which should be plotted and investigated.

It is common that non-normality in X or Y can cause problems in our residuals, but not necessarily. We will explore how to determine whether or not the residuals are appropriately distributed later on.

The Bivariate Relationship

In the most basic form, regression is a method for fitting a line to a cloud of data. This is easy to visualize with two variables, and is done using the familiar scatterplot:

The Bivariate Relationship



Notice how the line is not flat: as reading scores increase, so do math scores - but there is a lot of noise!

Fit the Linear Regression

While the visualization is useful, it does not address everything we need:

- What are the values for the intercept and the slope?
- Is this apparently positive relationship significant?
  - In other words: is the slope greater than 0?
- What proportion of variance in math scores is accounted for by our reading scores?
  The linear regression will address each of these.

Fit the Linear Regression

Table 2

	Dependent variable:
	mathPre
readPre	0.175*** (0.064)
Constant	41.200*** (3.300)
Observations	295
R^2	0.024
Adjusted R^2	0.021
Residual Std. Error	10.400 (df = 293)
F Statistic	7.350*** (df = 1; 293)
Note:	*p<0.1; **p<0.05; ***p<0.01

Output: The Intercept

$$Y_i = \boldsymbol{\beta_0} + \beta_1 X_i + \epsilon_i$$

The **estimate** for the Intercept (β_0) is 41.16, which is the math score we would predict for someone who received a 0 on their reading test.

The **standard error** of this estimate is 3.29.

If we divide the estimate by the standard error, it gives the **t-value** (12.48), which is used to compute a **p-value** ($p < .001$).

Output: The Intercept

The **p-value** is the probability of obtaining a t-statistic as large or larger if the true value of the intercept is *zero*.

In this case, the p-value is very very small, much smaller than the traditionally used $\alpha = 0.05$. So, we reject the null hypothesis that the intercept, β_0, is zero.

Interpretation

In psychology, having a significant intercept isn't usually all that interesting. We often use measures that aren't even meaningful for a score of zero. For example, if we were trying to predict income by IQ, having a zero or non-zero intercept doesn't mean too much since IQ isn't even defined for a score that low.

Output: Slopes!

$$Y_i = \beta_0 + \boldsymbol{\beta_1} X_i + \epsilon_i$$

Next we will look at the line that starts with `readPre`. This is the slope for `readPre` (β_1). For every unit increase in reading score, we predict that the respective math score will increase by 0.175 points.

The **standard error** of the estimate for the slope is .064 and the **t-value** is 2.711, which is also significant ($p = 0.007$).

So, we reject the null hypothesis that the slope is zero ($H_0 : \beta_1 = 0$).

Output: Model Summaries

Two other important values to note are:

1. Multiple *R*-squared: This tells us the proportion of variance in the outcome (`mathPre`) that is explained (or "accounted for") by our predictor (`readPre`). In this case, only 2.45% of the variability in math scores is accounted for by the reading scores.

Output: Model Summaries

2. The F-statistic: This is a test for the overall model, and for our model it is significant, with $F(1,293) = 7.35$, p $= 0.007$.

In this example, we only have one predictor. In fact, the F-statistic is simply the square of the t-statistic for the one predictor.

However, if there were more predictors, then the F test tells us that, taken together, all of the predictors explain a significant proportion of the variability in the outcome measure.

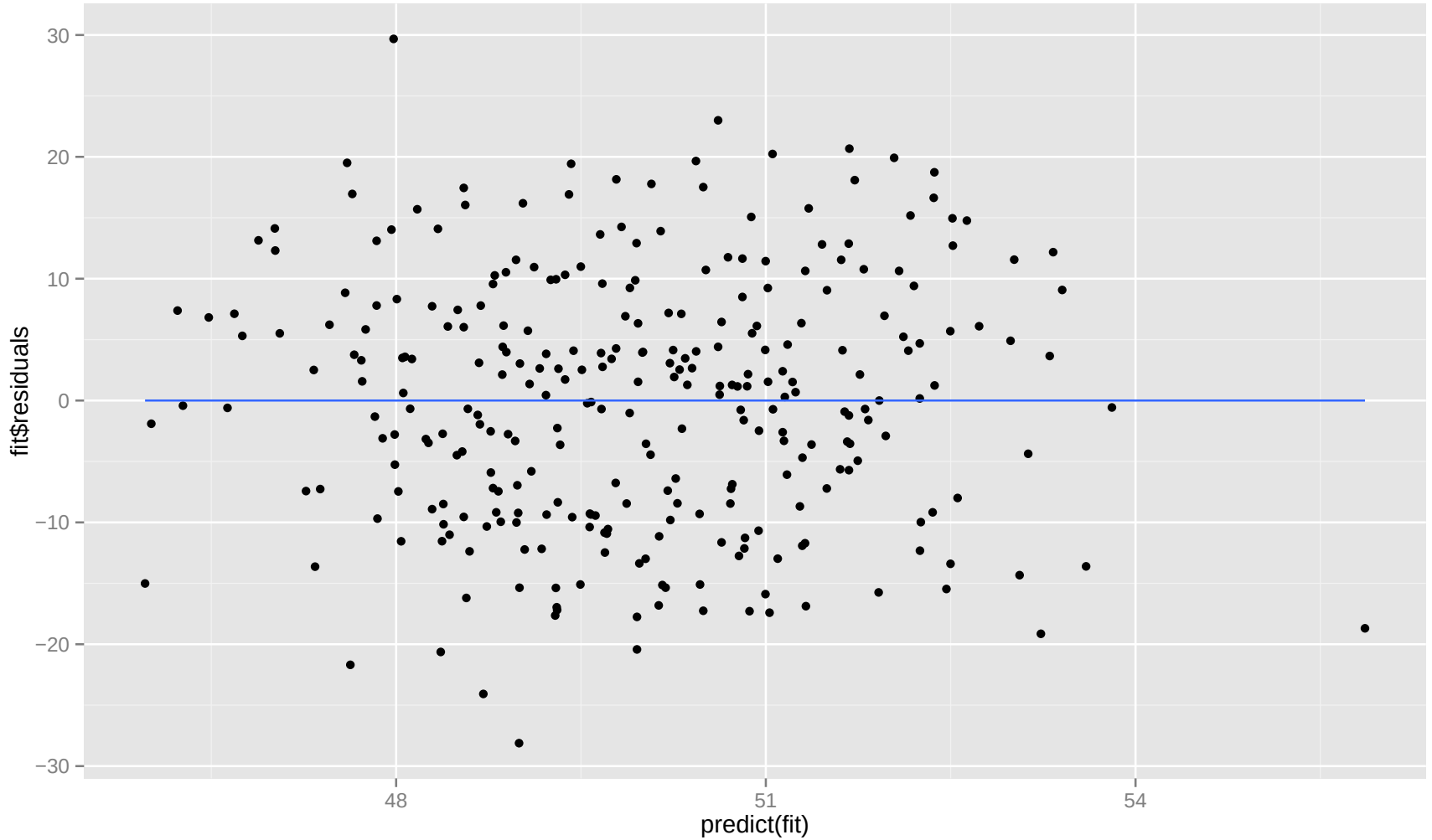
Check Assumptions

The first assumption of this particular linear regression is that the relationship between the two variables can be approximated by a straight line (linearity).

So, looking at the bivariate scatterplot is actually very important in assessing whether this assumption was met. Curvature in the scatterplot would be indicative of a problem! In our case this looked pretty good.

Check Assumptions

Most important: "residuals vs fitted values" plot. You are hoping to see residuals spread evenly above and below zero across the range of fitted values. Any increasing or decreasing funnel is a sign of a violation of the **homoscedasticity**, while any curvature (a U-shape) is a sign of a violation of **linearity**.



Outliers, Leverage, and Influential Cases

Standardized residuals are divided by an estimate of their standard deviation, which puts them onto familiar standard scale with a mean of 0, and standard deviation of 1.

Observations with a **standardized residual** $>$ **3.29** are a cause for concern, this is a very large residual!

Cases with a **standardized residual** $>$ **2** are a bit of a red flag.

Leverage

Points that are extreme in `X` (our predictor) can exert a lot of pressure on the regression line. That is to say, **high leverage points can pull the regression line towards it**, while points that are close to the mean of `X` will have less of an effect on where the regression line goes.

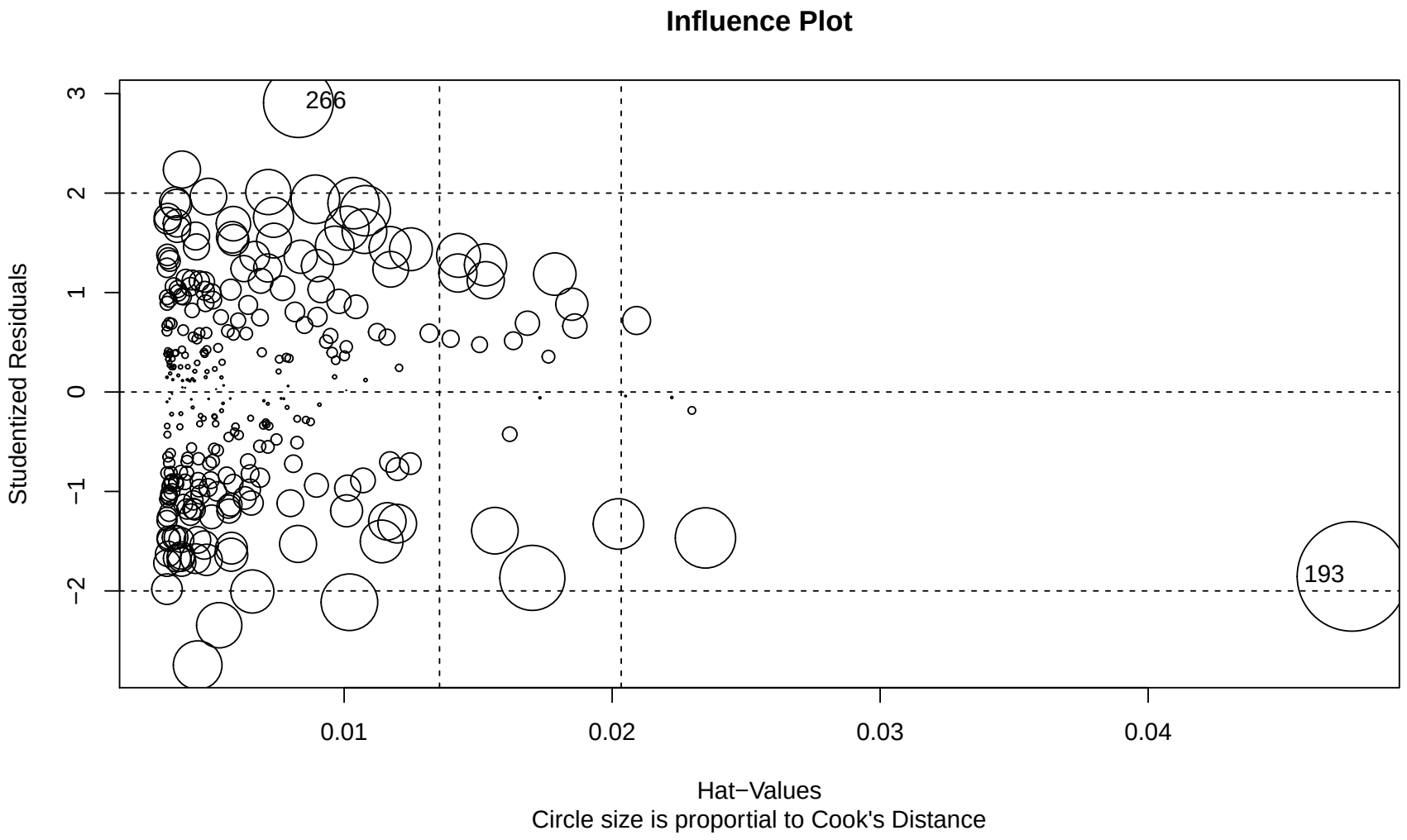
High leverage don't necessarily mean that a data point influences the regression estimates (slope and intercept). A point can be extreme in `X` but fit very well with the pattern of rest of the data!

Influential Cases

When a data point has high leverage (extreme in $\mathtt{X}$) *and* is an outlier in terms of $\mathtt{Y}$ it can greatly change the values of the parameter estimates (slope and intercept) that we would get if it weren't in the sample. It is then considered an **influential case**.

There are a few measures for assessing influence. **Cook's distance** is a popular one, and considers the effect that a single case has on the model as a whole, with values > 1 often flagged as causes for concern.

Influential Cases: Plotting



Summary

So, we have multiple ways to describe datapoints:

- Weird on one variable, ignoring the others:
  - **univariate outlier**
- Weird for the regression model:
  - **regression outlier** (large residual)
- Weird on predictor variable:
  - **high leverage**
- Weird on predictor *and* weird on response:
  - **influential!**

Reporting

This study investigated the relationship between reading and math scores. It was hypothesized that higher reading scores would be associated with higher math scores. The overall model was significant, $F(1,293) = 7.35$, p=0.007, and a 1 unit increase in reading was associated with 0.17 (SE = 0.064) unit increase in math score. While there were no substantial deviations from linearity or influential cases, this model only accounts for a small proportion of the variability in math scores, $R^2 = 0.0245$.

Multiple Regression Introduction

When we have two or more predictors, then we are conducting a "multiple regression". You can actually use multiple regression to answer a lot of different types of questions!

The most straightforward is to try and ascertain the best possible prediction for our outcome (Y), based upon a set of predictor variables.

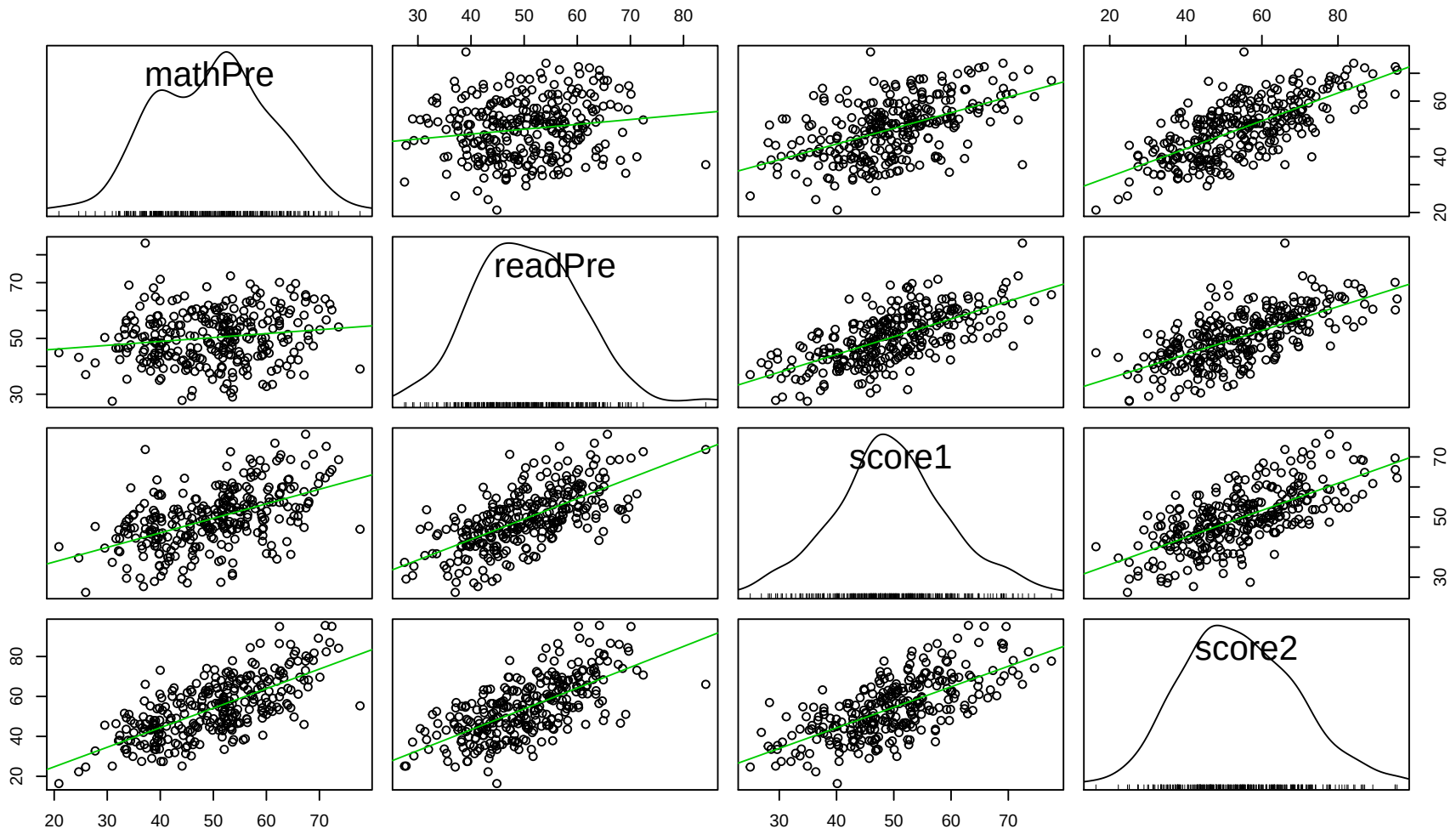
Multiple Regression Introduction

In Psychology, three common applications are:

1. Controlling for variables we know to be important, is our target predictor significant?
    - including another predictor "controls" for the remaining predictors
    - does `mathPre` predict `score2`, controlling for `readPre`?
2. Comparing the relative "importance" of predictors - which is most important in predicting `score2`, `mathPre` or `readPre`?
3. Describing inter-relationships of predictors with respect to the outcome.
    - We can incorporate *interactions*!
    - Perhaps the relationship between `score2` and `readPre` is different across the different `groups`?

Multiple Regression Introduction

Bottom line, this is a huge topic that we could spend weeks on, and we are only going to scratch the surface today.



Multiple Regression Example

Table 3:mathPre as Predictor

	Dependent variable:
	score2
mathPre	0.968*** (0.059)
Constant	5.520* (3.000)
Observations	298
R^2	0.479
Adjusted R^2	0.477
Residual Std. Error	10.600 (df = 296)
F Statistic	272.000*** (df = 1; 296)
Note:	*p<0.1; **p<0.05; ***p<0.01

The effect of `mathPre` is significant, $F(1,296) = 272$, $p < .0001$, $R^2 = 0.479$.

Adding `readPre`

Table 4:mathPre and readPre

	Dependent variable:
	score2
mathPre	0.853^{***} (0.036)
readPre	0.894^{***} (0.040)
Constant	-33.500^{***} (2.520)
Observations	295
R^2	0.810
Adjusted R^2	0.808
Residual Std. Error	6.420 (df = 292)
F Statistic	621.000^{***} (df = 2; 292)
Note:	*p<0.1; **p<0.05; ***p<0.01

Adding `readPre` yields a sig. model, $F(2,292) = 621$, $p < .0001$, and R^2 increased to 0.81!

Adding Factors!

But perhaps this relationship is different across levels of our predictors. This is easiest to see with a categorical variable. Let's investigate how `readPre` relates to `score2`, given `group`.

Main Effects Model

The main effects model says that the relationship between `readPre` and `score2` is the same for each `group`, but the different `groups` might have different relationships with `score2`.

This is done via **contrast coding**, which takes one level of `group` as our reference group.

Main Effects Model

Table 5

	Dependent variable:
	score2
readPre	1.030^{***} (0.068)
groupmnemonic	-3.380^{**} (1.520)
groupnone	-3.750^{**} (1.580)
Constant	4.580 (3.610)
Observations	297
R^2	0.448
Adjusted R^2	0.443
Residual Std. Error	11.000 (df = 293)
F Statistic	79.400^{***} (df = 3; 293)
Note:	*p<0.1; **p<0.05; ***p<0.01

This says that, controlling for `readPre` and compared to the cram group, both mnemonic and none had significantly different `score2`s.

Interactions!

But maybe `readPre` and `group` *interact* - meaning they may have different relationships with `score2`.

Table 6

	Dependent variable:
	score2
readPre	1.010*** (0.113)
groupmnemonic	−3.600 (8.020)
groupnone	−6.460 (9.230)
readPre:groupmnemonic	0.004 (0.156)
readPre:groupnone	0.054 (0.180)
Constant	5.310 (5.850)
Observations	297
R^2	0.448
Adjusted R^2	0.439
Residual Std. Error	11.000 (df = 291)
F Statistic	47.300*** (df = 5; 291)
Note:	*p<0.1; **p<0.05; ***p<0.01

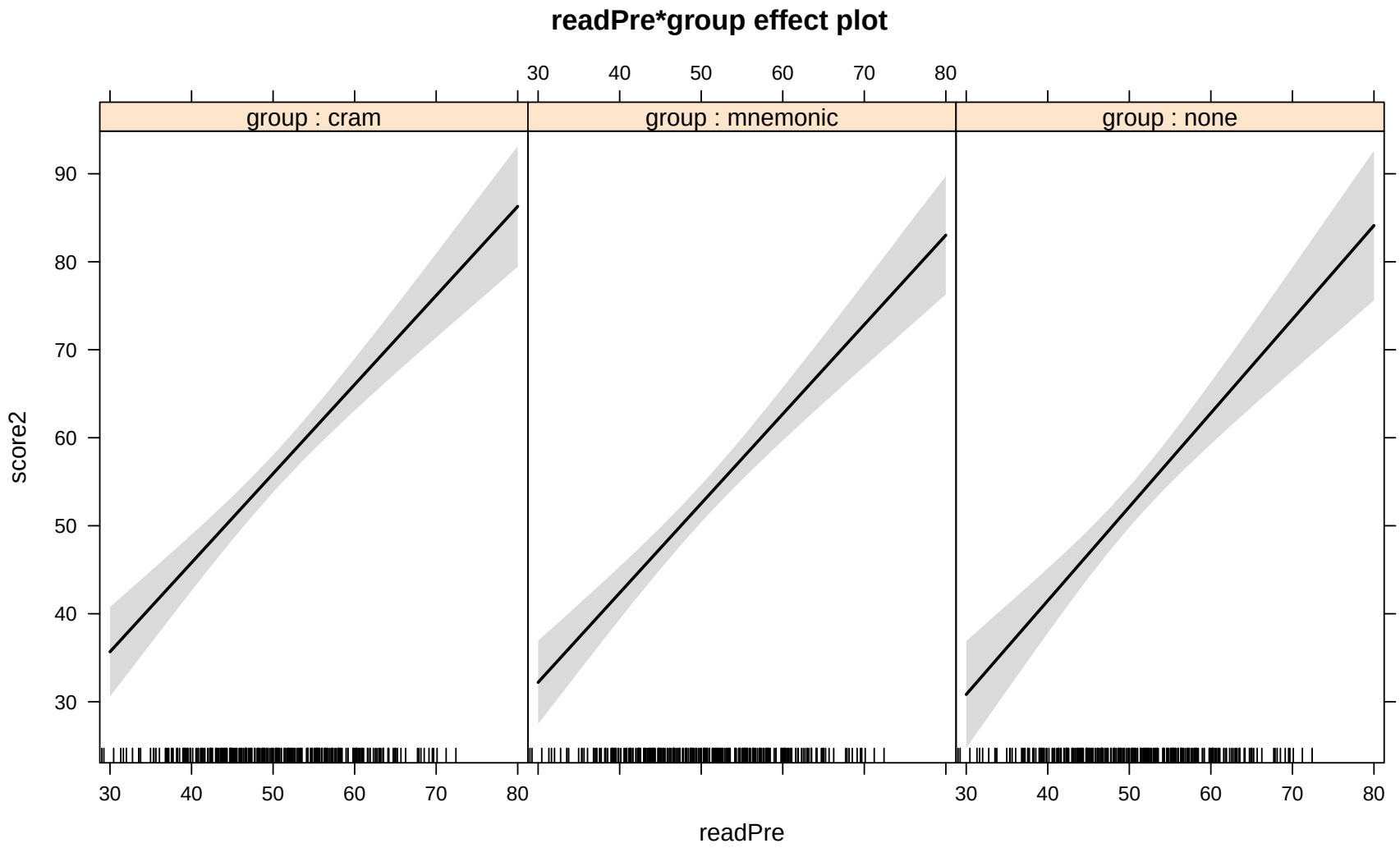
Interactions!

When incorporating interactions, we look at the p values for the crossings first. Often, if they are insignificant, we remove them from the model.

```
## Analysis of Variance Table
##
## Response: score2
##                Df Sum Sq Mean Sq F value Pr(>F)
## readPre         1  27748   27748  229.41 <2e-16 ***
## group           2    862     431    3.56   0.03 *
## readPre:group   2     13       6    0.05   0.95
## Residuals     291  35198     121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions!

To get an idea of what we are looking for, look at this plot:



Other Notables

- Diagnostics
  - Same as in linear regression, with the addition of **multicollinearity**
  - Basically means: if some of your predictor variables are too similar, bad things will happen.
  - Assessed via **variance inflation factors** (VIF) - if any are greater than 10, indicates a problem.
  - All of the regression plots we discussed before still work here!

Other Notables

- Importance of Predictors
  - With multiple significant predictors, we might want to know which explains the "most" variability
  - Standardize our predictors to put them on the same scale, rerun regression!
- Comparing Models
  - Can compare sequential models using the ANOVA framework

Final Thoughts

Multiple Regression is very cool. We can incorporate:

- multiple types of predictors from any scale of measurement,
- more than one dependent variable ("multivariate multiple regression"), different kinds of response variables (e.g., logistic regression predicts binary response)
- "block" designs, where we build models hierarchically ("controlling for this set of demographic variables, the relationship between stress and exam grade was...")
- even things we cannot directly measure (latent variables, via factor analysis/structural equation modeling)!