Multiple Linear Regression

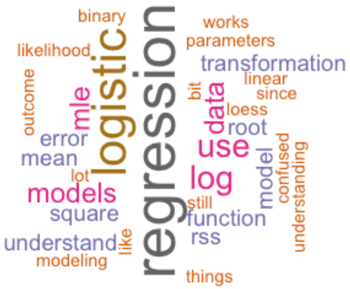
DATA 606 - Statistics & Probability for Data Analytics

Jason Bryer, Ph.D., Angela Lui, Ph.D., and George Hagstrom, Ph.D.

November 20, 2024

One Minute Paper Results

What was the most important thing you learned during this class?



What important question remains unanswered for you?

Weight of Books

```
allbacks <- read.csv('../course_data/allbacks.csv')
head(allbacks)
```

```
##   X volume area weight cover
## 1 1    885  382    800    hb
## 2 2   1016  468    950    hb
## 3 3   1125  387   1050    hb
## 4 4    239  371    350    hb
## 5 5    701  371    750    hb
## 6 6    641  367    600    hb
```

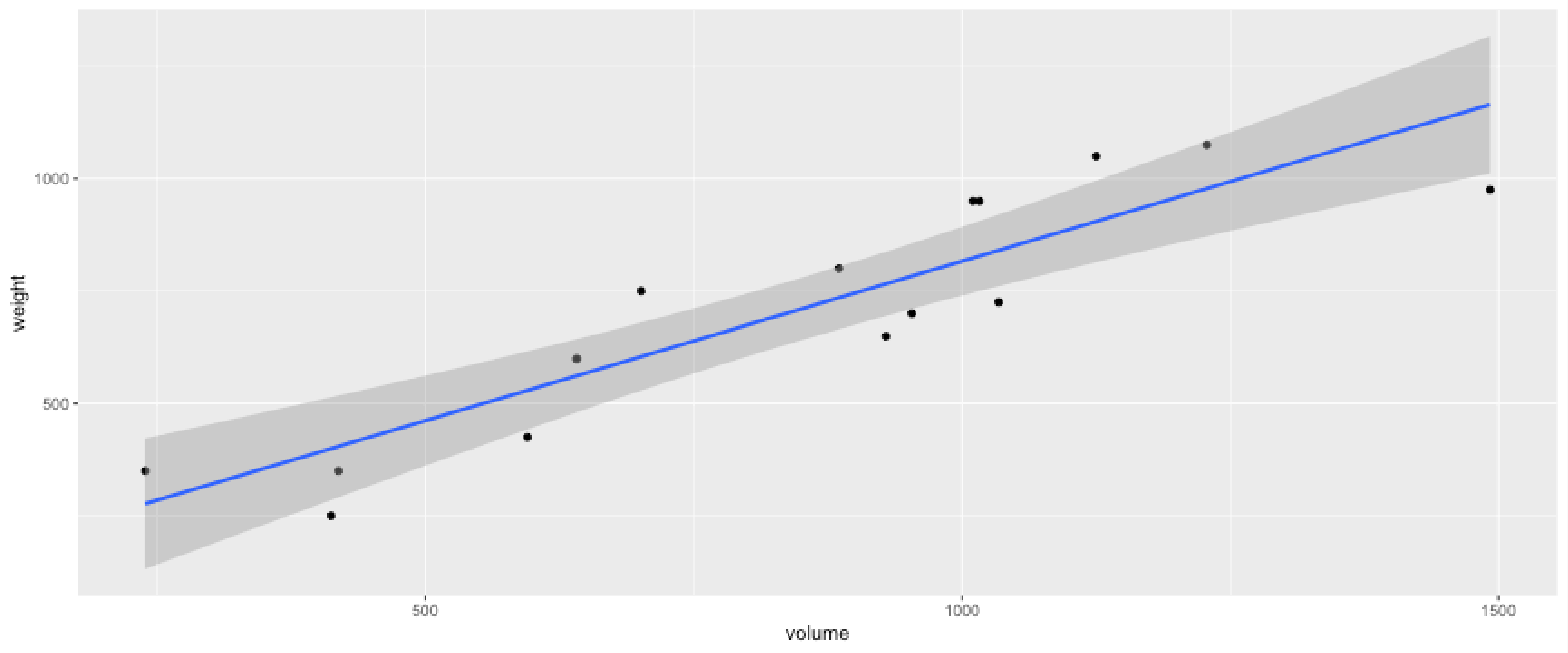
From: Maindonald, J.H. & Braun, W.J. (2007). *Data Analysis and Graphics Using R, 2nd ed.*

Weights of Books (cont)

```
lm.out <- lm(weight ~ volume, data=allbacks)
```

$$\hat{weight} = 108 + 0.71 volume$$

$$R^2 = 80\%$$



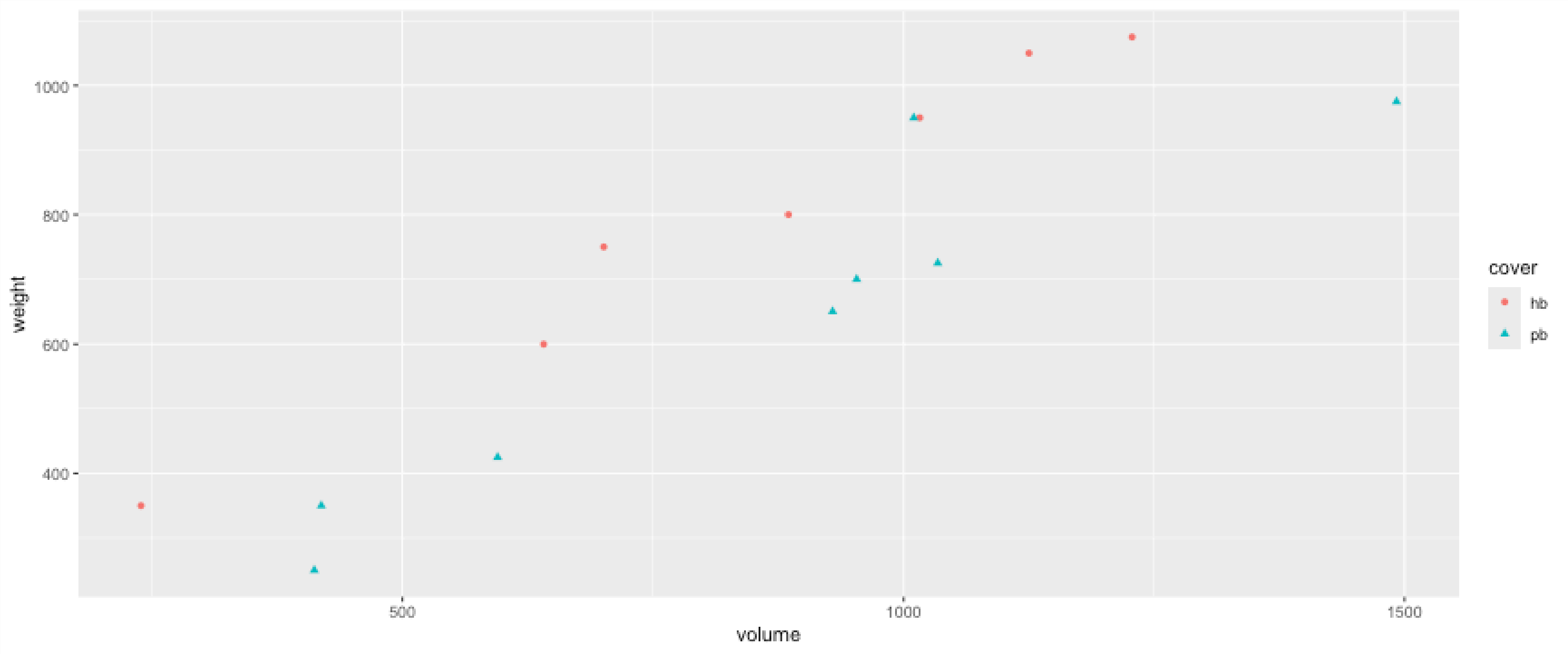
Modeling weights of books using volume

```
summary(lm.out)
```

```
## 
## Call:
## lm(formula = weight ~ volume, data = allbacks)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -189.97 -109.86   38.08  109.73  145.57 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 107.67931   88.37758   1.218    0.245    
## volume        0.70864    0.09746   7.271 6.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 123.9 on 13 degrees of freedom
## Multiple R-squared:  0.8026,	Adjusted R-squared:  0.7875 
## F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

Weights of hardcover and paperback books

- Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



- Paperbacks generally weigh less than hardcover books after controlling for book's volume.

Modeling using volume and cover type

```r
lm.out2 <- lm(weight ~ volume + cover, data=allbacks)
summary(lm.out2)
```

```
##
## Call:
## lm(formula = weight ~ volume + cover, data = allbacks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -110.10  -32.32  -16.10   28.93  210.95
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  197.96284   59.19274   3.344 0.005841 **
## volume         0.71795    0.06153  11.669  6.6e-08 ***
## coverpb     -184.04727   40.49420  -4.545 0.000672 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.2 on 12 degrees of freedom
## Multiple R-squared:  0.9275,    Adjusted R-squared:  0.9154
## F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Linear Model

$$\hat{weight} = 198 + 0.72 volume - 184 coverpb$$

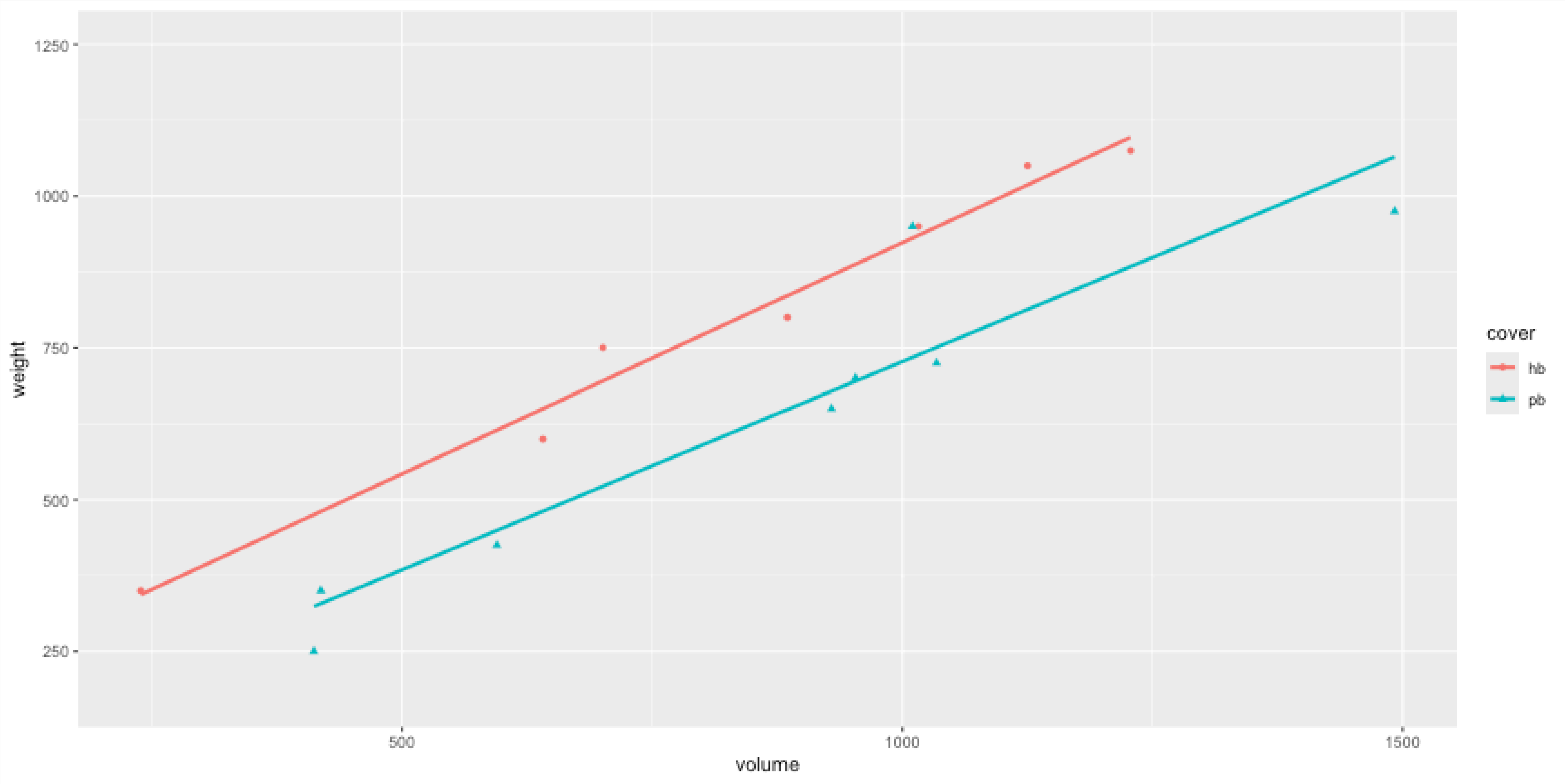
1. For **hardcover** books: plug in *0* for cover.

$$\hat{weight} = 197.96 + 0.72 volume - 184.05 \times 0 = 197.96 + 0.72 volume$$

1. For **paperback** books: put in 1 for cover.

$$\hat{weight} = 197.96 + 0.72 volume - 184.05 \times 1$$

Visualizing the linear model



Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
coverpb	-184.0473	40.4942	-4.55	0.0007

- **Slope of volume**: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- **Slope of cover**: All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- **Intercept**: Hardcover books with no volume are expected on average to weigh 198 grams.
  - Obviously, the intercept does not make sense in context. It only serves to adjust the height of the line.

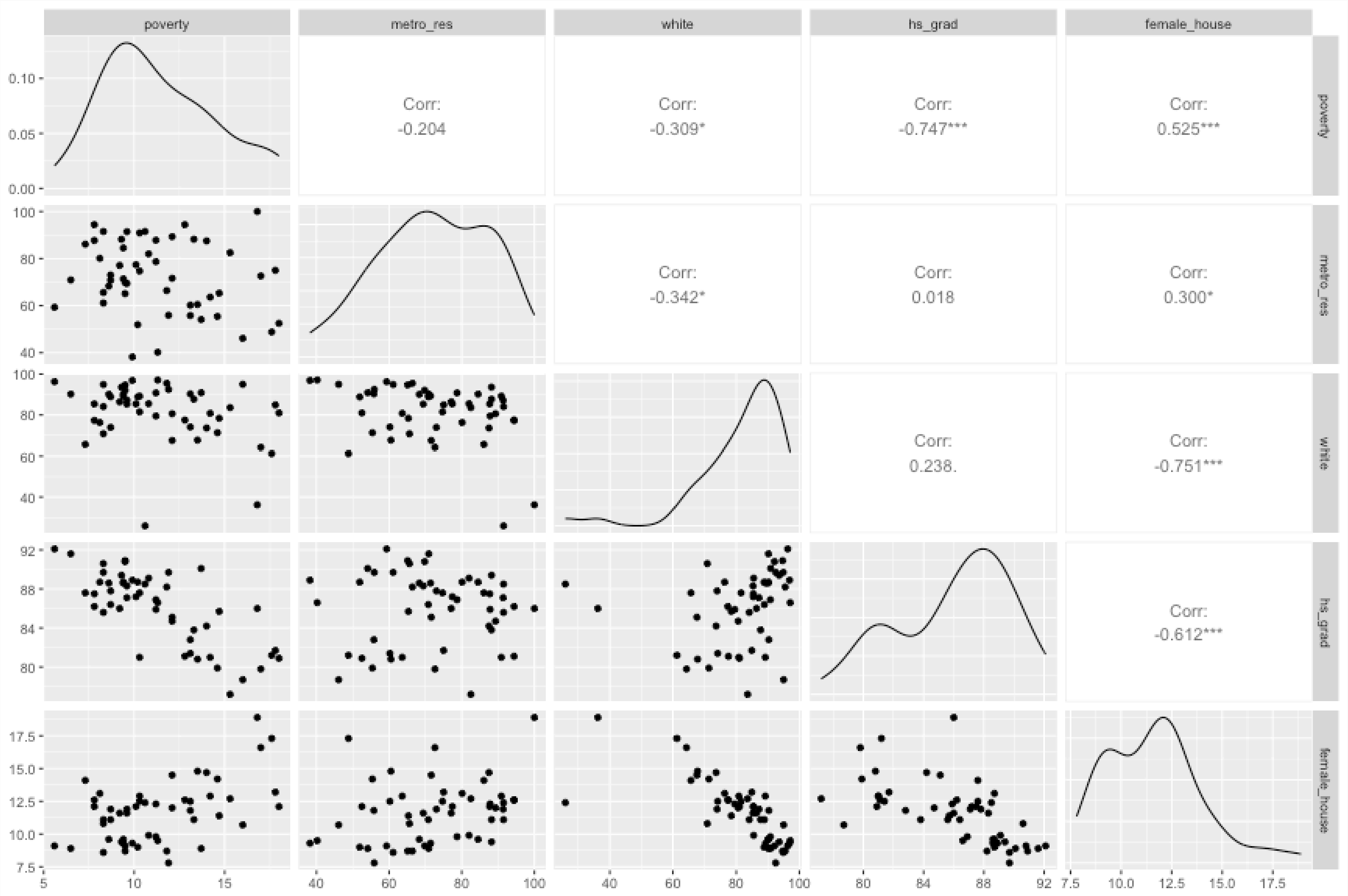
Modeling Poverty

```r
poverty <- read.table("../course_data/poverty.txt", h = T, sep = "\t")
names(poverty) <- c("state", "metro_res", "white", "hs_grad", "poverty", "female_house")
poverty <- poverty[,c(1,5,2,3,4,6)]
head(poverty)
```

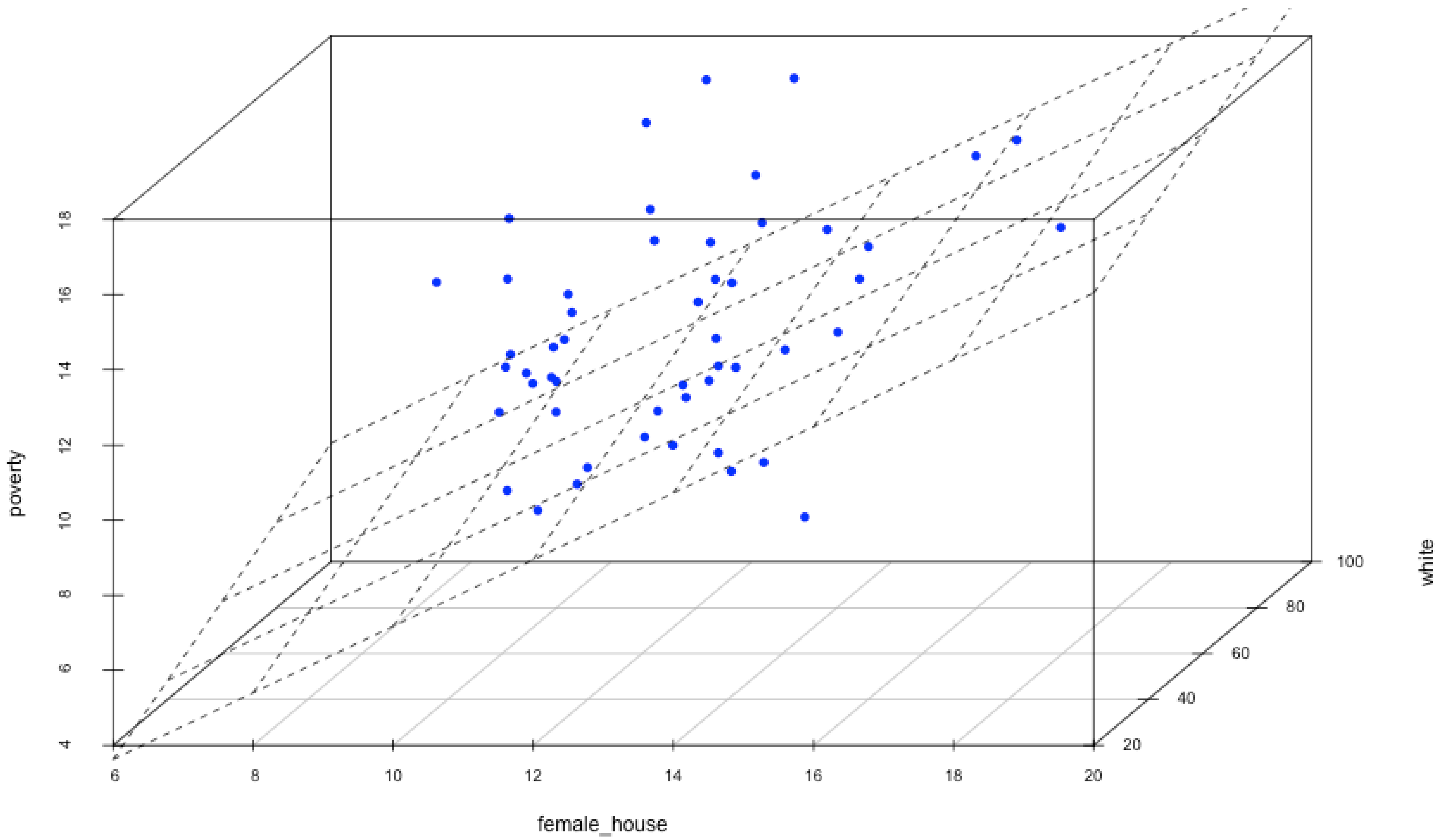
```
##        state poverty metro_res white hs_grad female_house
## 1    Alabama    14.6      55.4  71.3    79.9         14.2
## 2     Alaska     8.3      65.6  70.8    90.6         10.8
## 3    Arizona    13.3      88.2  87.7    83.8         11.1
## 4   Arkansas    18.0      52.5  81.0    80.9         12.1
## 5 California    12.8      94.4  77.5    81.1         12.6
## 6   Colorado     9.4      84.5  90.2    88.7          9.6
```

From: Gelman, H. (2007). *Data Analysis using Regression and Multilevel/Hierarchial Models.* Cambridge University Press.

Modeling Poverty



3-D scatter plot

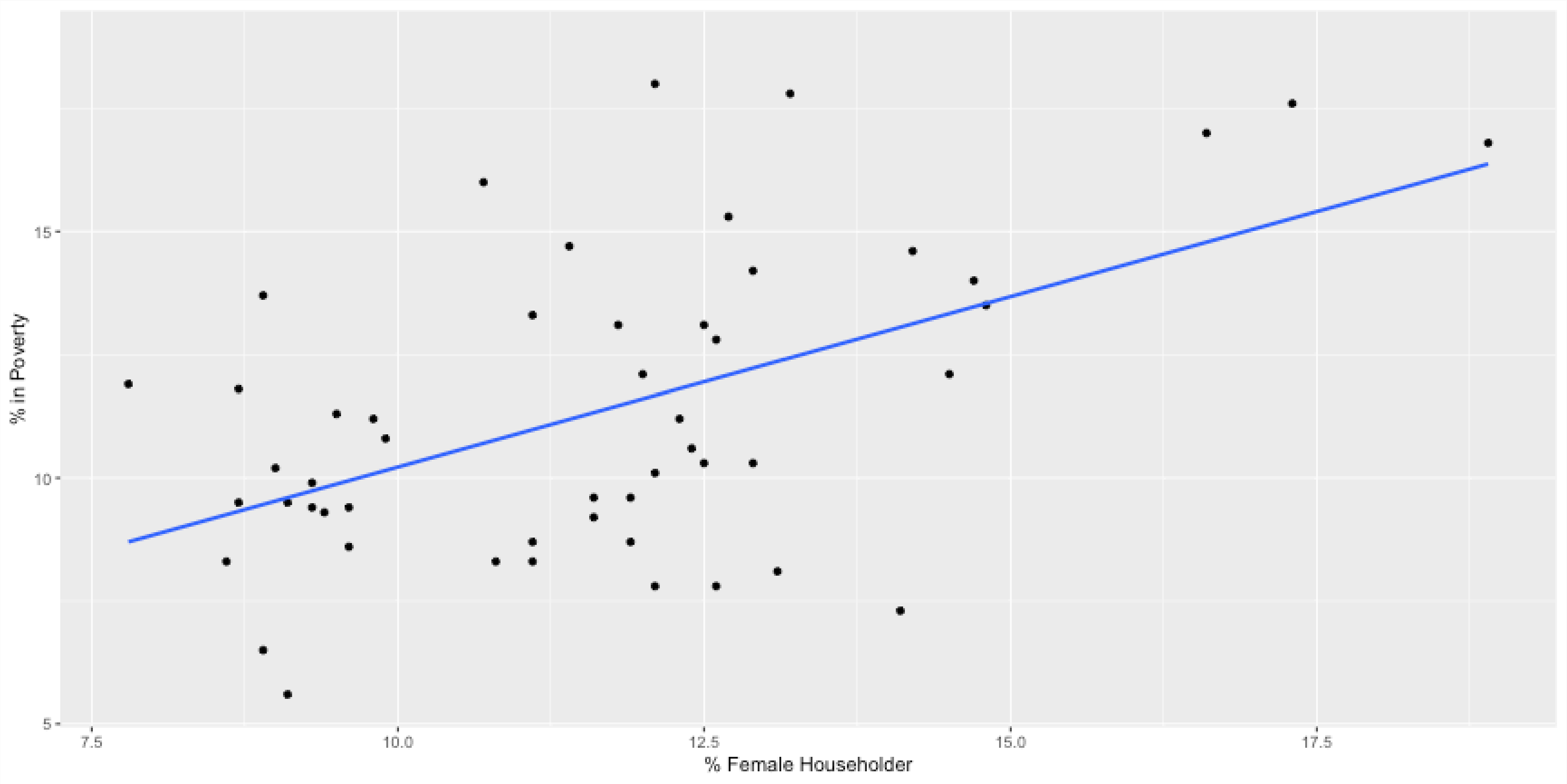


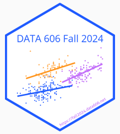
Predicting Poverty using Percent Female Householder

```r
lm.poverty <- lm(poverty ~ female_house, data=poverty)
summary(lm.poverty)
```

```
##
## Call:
## lm(formula = poverty ~ female_house, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7537 -1.8252 -0.0375  1.5565  6.3285
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.3094     1.8970   1.745   0.0873 .
## female_house   0.6911     0.1599   4.322 7.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.664 on 49 degrees of freedom
## Multiple R-squared:  0.276,	Adjusted R-squared:  0.2613
## F-statistic: 18.68 on 1 and 49 DF,  p-value: 7.534e-05
```

% Poverty by % Female Household





Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{explained \quad variability \quad in \quad y}{total \quad variability \quad in \quad y}$$

Using ANOVA we can calculate the explained variability and total variability in y.

Sum of Squares

```r
anova.poverty <- anova(lm.poverty)
print(xtable::xtable(anova.poverty, digits = 2), type='html')
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1.00	132.57	132.57	18.68	0.00
Residuals	49.00	347.68	7.10		

Sum of squares of *y*: $SS_{Total} = \sum (y - \bar{y})^2 = 480.25 \rightarrow$ **total variability**

Sum of squares of residuals: $SS_{Error} = \sum e_i^2 = 347.68 \rightarrow$ **unexplained variability**

Sum of squares of *x*: $SS_{Model} = SS_{Total} - SS_{Error} = 132.57 \rightarrow$ **explained variability**

$$R^2 = \frac{explained \quad variability \quad in \quad y}{total \quad variability \quad in \quad y} = \frac{132.57}{480.25} = 0.28$$

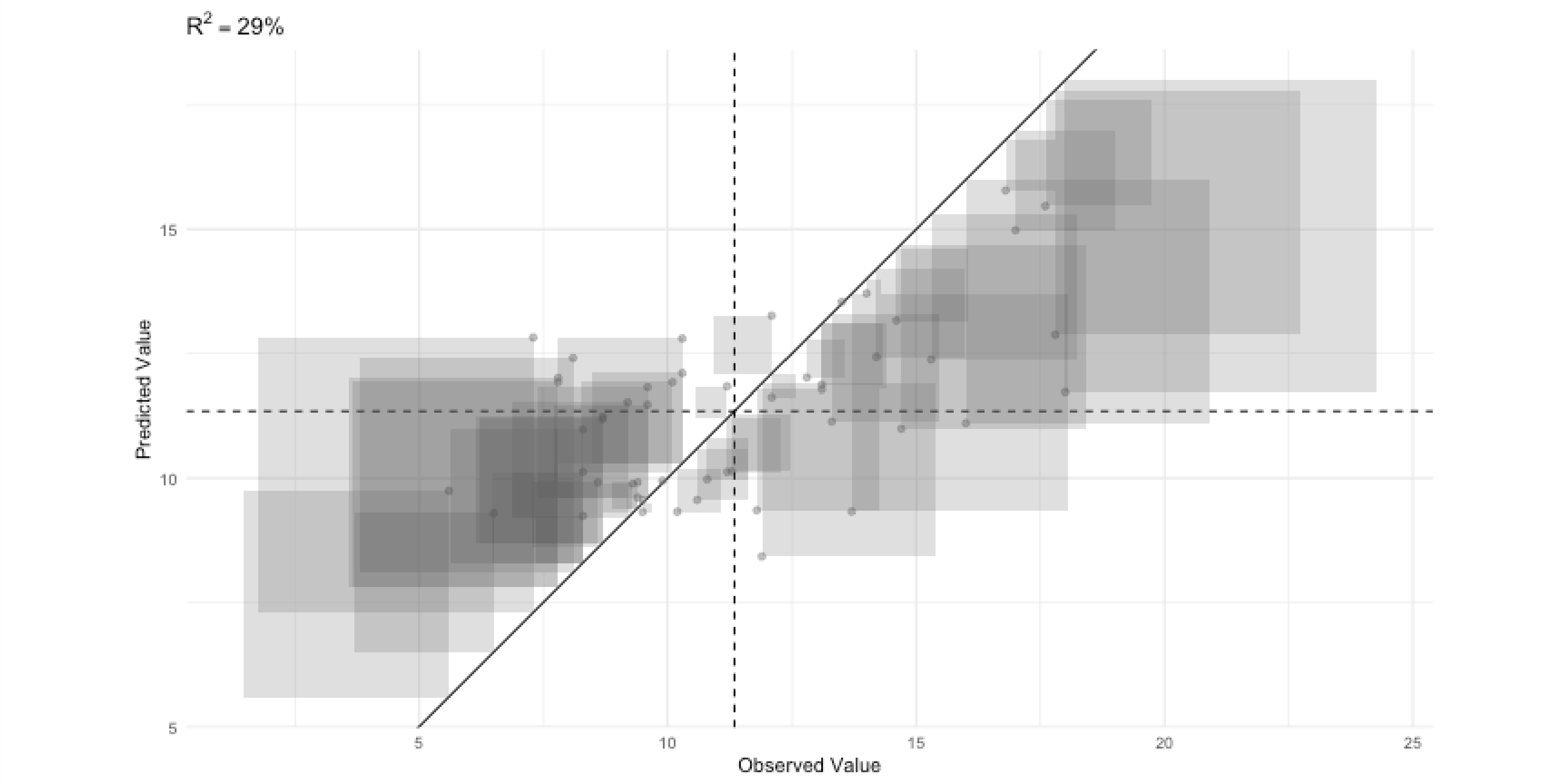
Why bother?

- For single-predictor linear regression, having three ways to calculate the same value may seem like overkill.

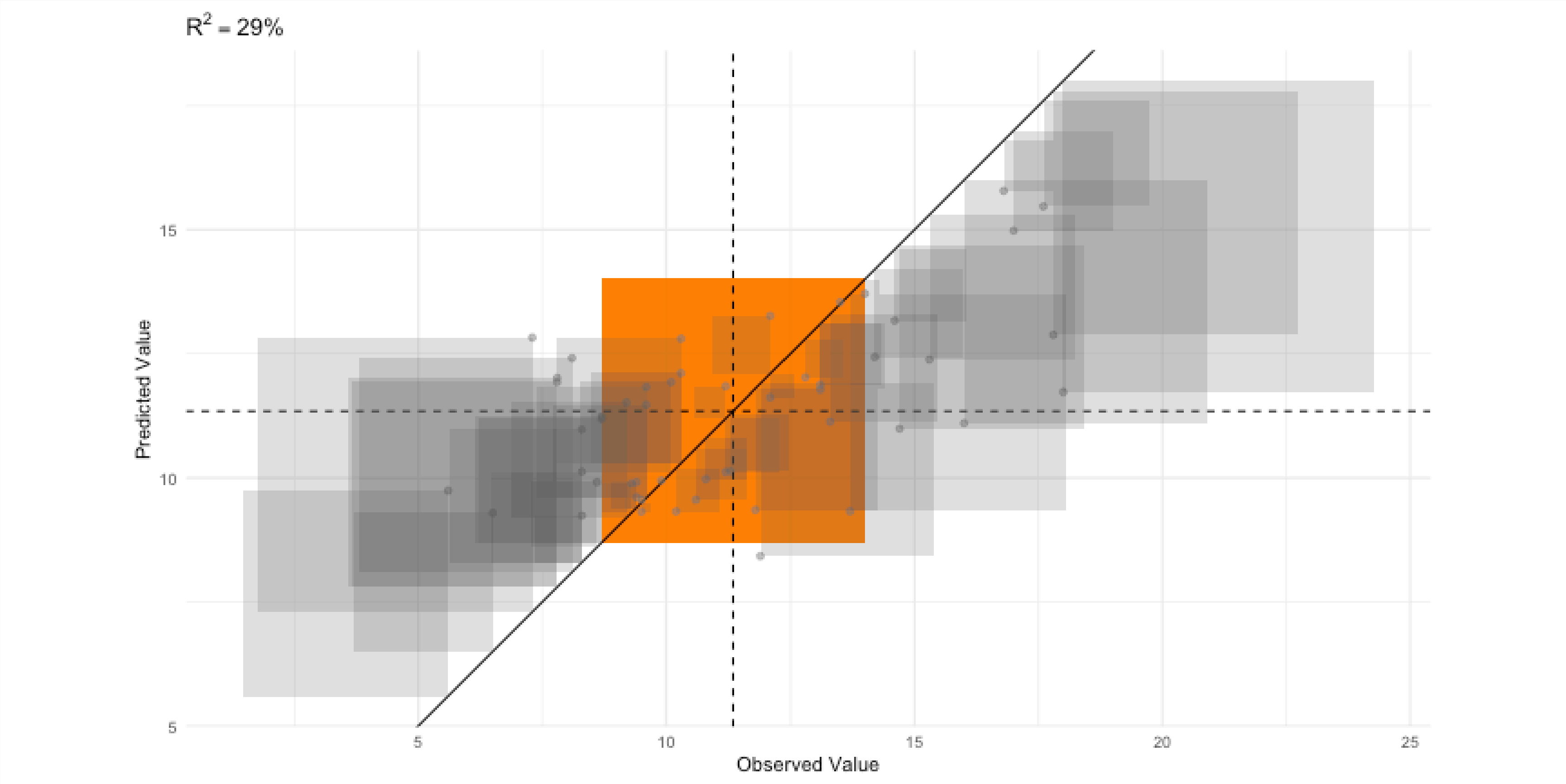
- However, in multiple linear regression, we can't calculate R^2 as the square of the correlation between x and y because we have multiple xs.

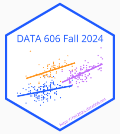
- And next we'll learn another measure of explained variability, *adjusted* R^2, that requires the use of the third approach, ratio of explained and unexplained variability.

R^2: Error variance

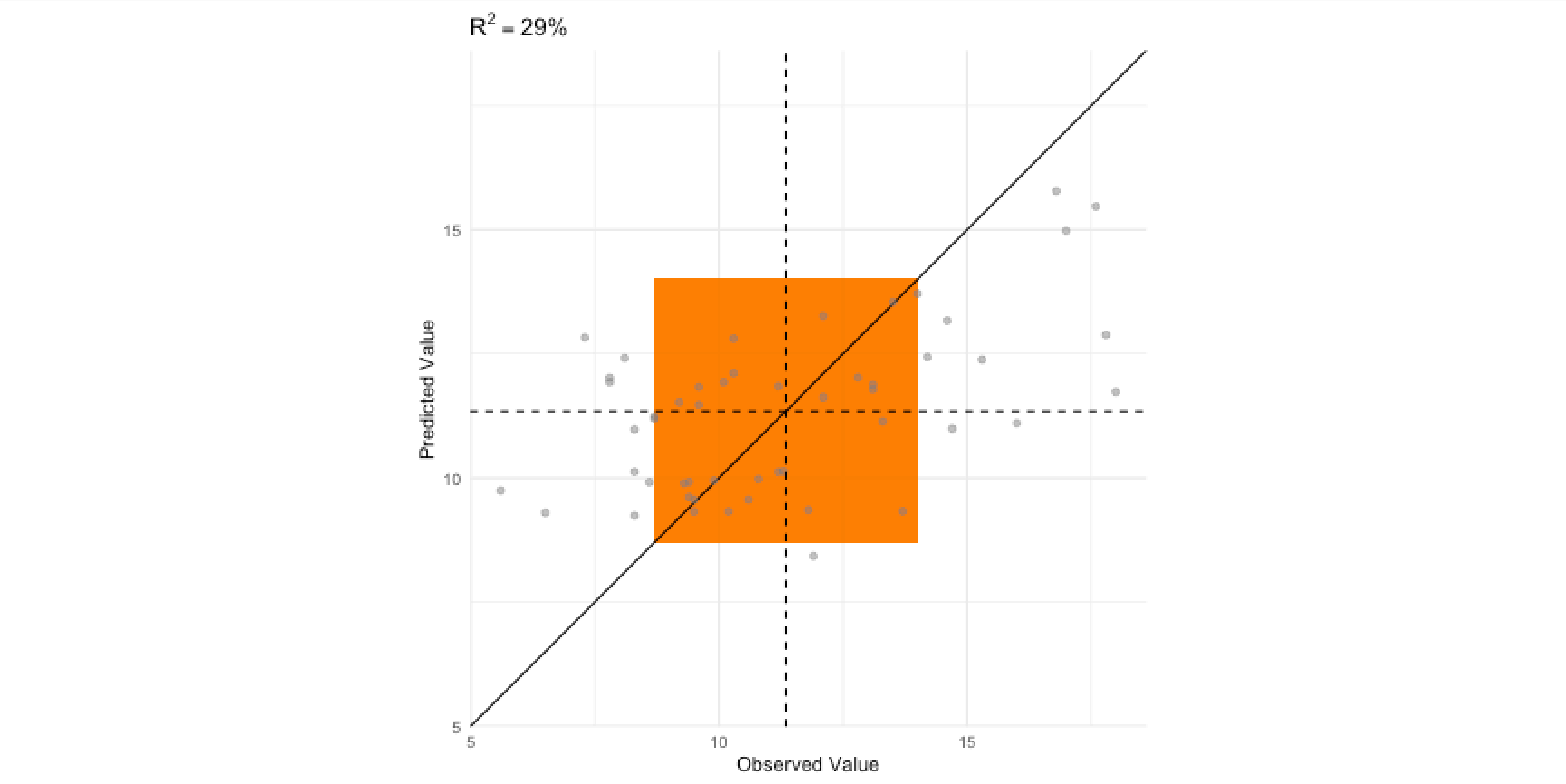


R^2: Error variance (cont.)

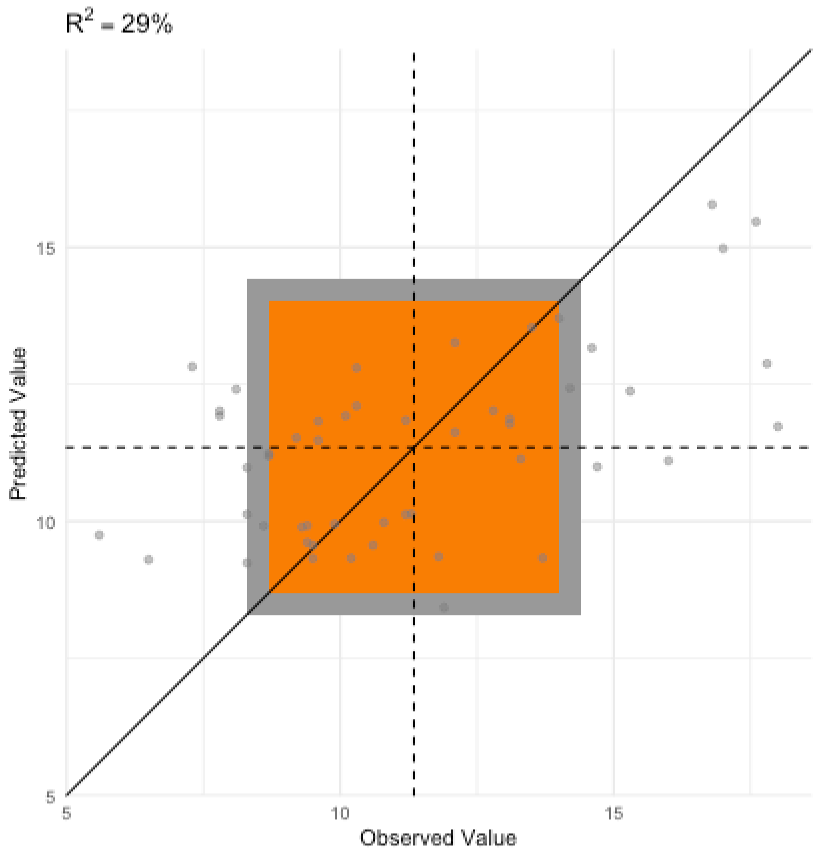




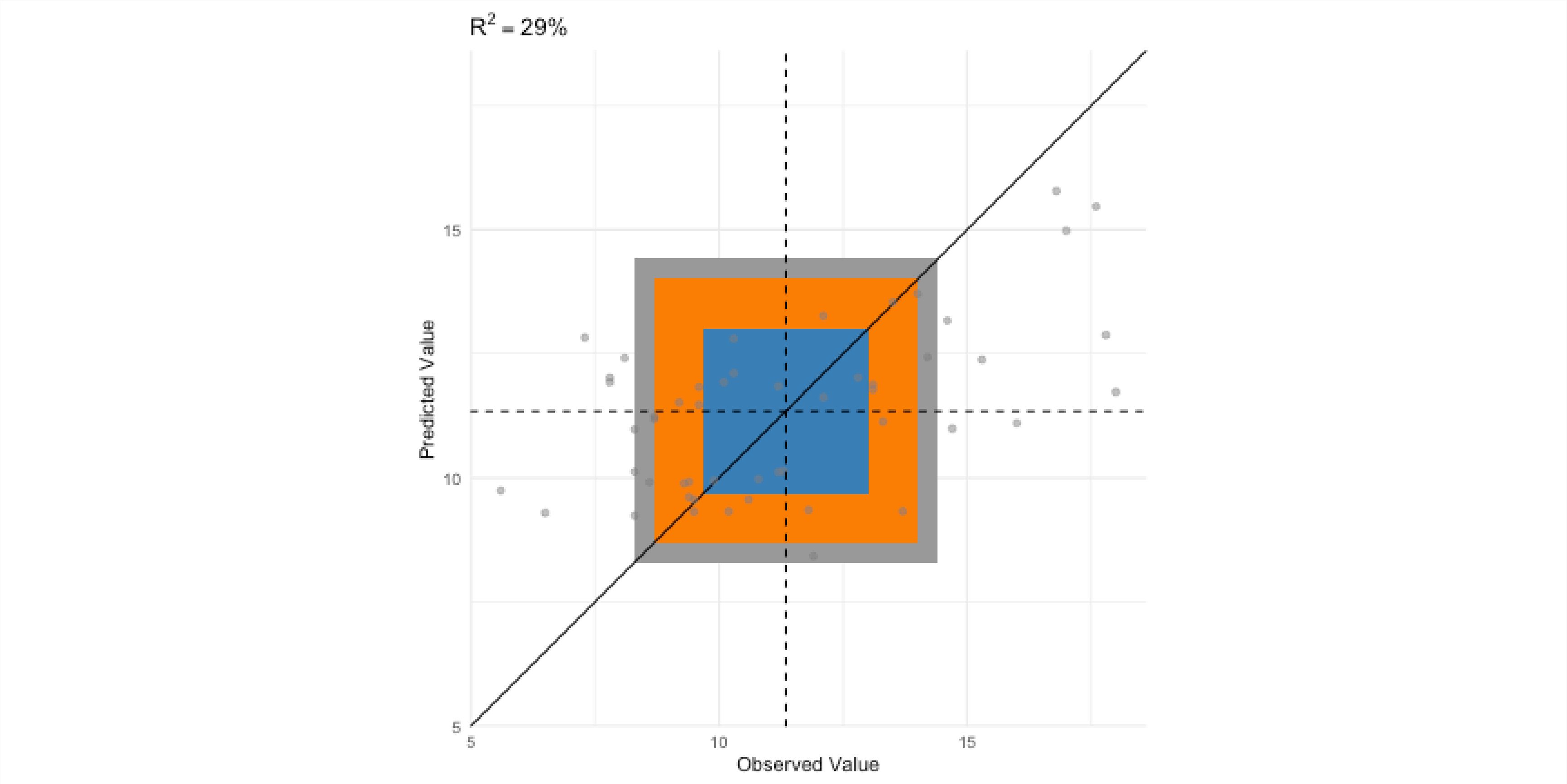
R^2: Error variance (cont.)



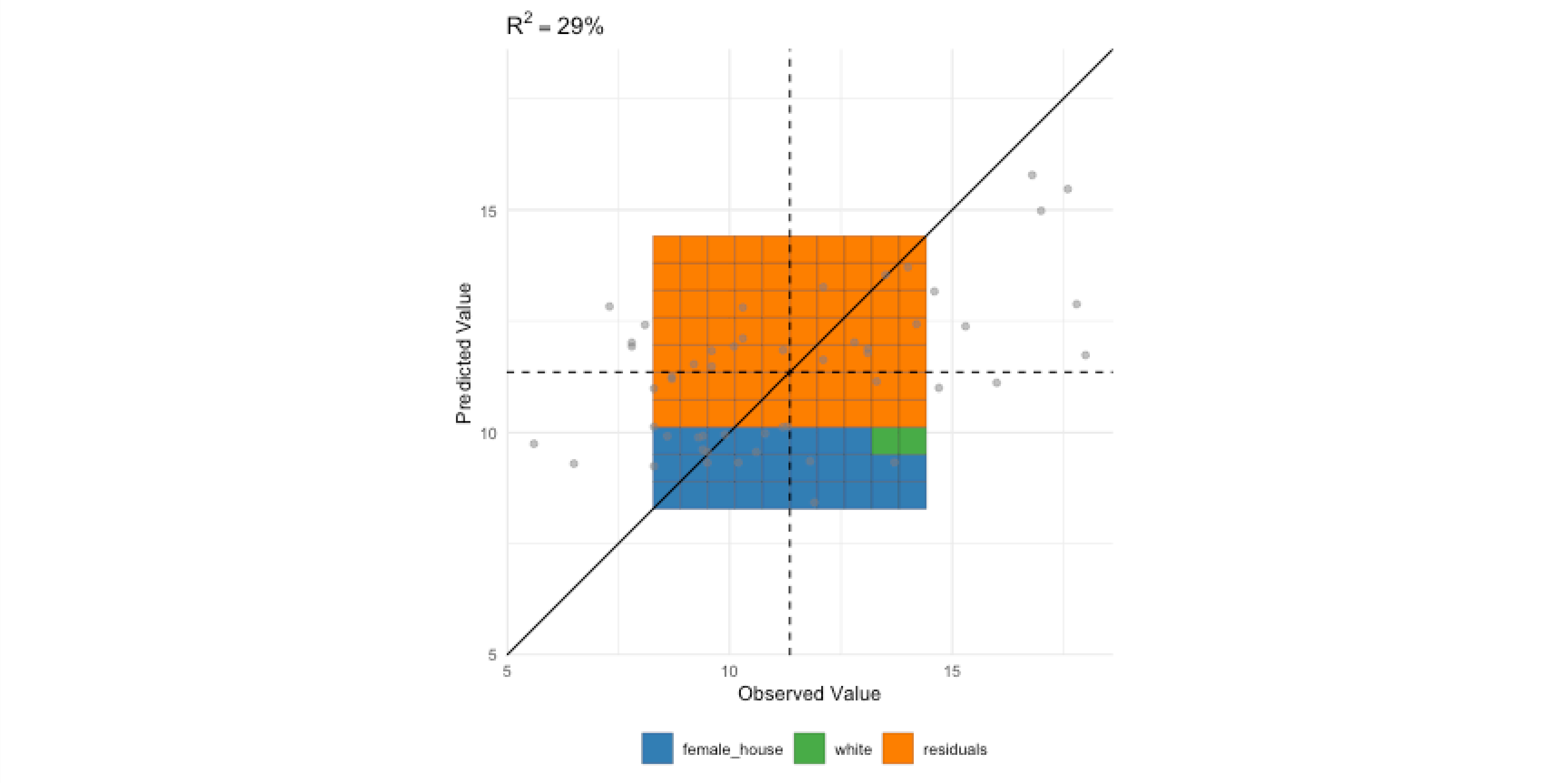
R^2: Total (grey) and error (orange) variance



R^2: Total (grey), error (orange), and regression (blue)



R^2: All variances




```
VisualStats::r_squared_shiny()
```

Predicting poverty using % female household & %

```r
lm.poverty2 <- lm(poverty ~ female_house + white, data=poverty)
print(xtable::xtable(lm.poverty2), type='html')
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.5789	5.7849	-0.45	0.6577
female_house	0.8869	0.2419	3.67	0.0006
white	0.0442	0.0410	1.08	0.2868

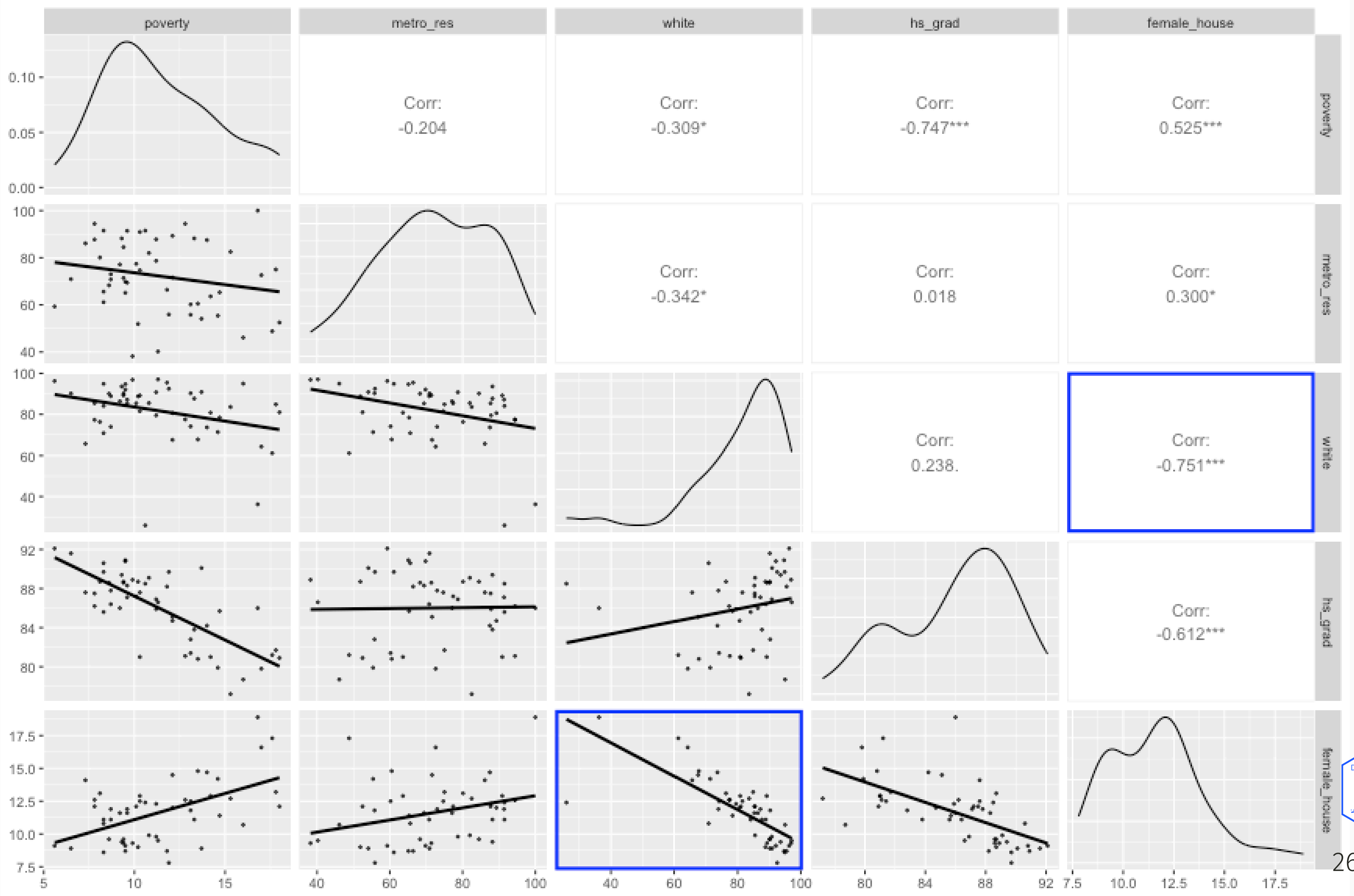
```r
anova.poverty2 <- anova(lm.poverty2)
print(xtable::xtable(anova.poverty2, digits = 3), type='html')
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1.000	132.568	132.568	18.745	0.000
white	1.000	8.207	8.207	1.160	0.287
Residuals	48.000	339.472	7.072		

$$R^2 = \frac{explained \quad variability \quad in \quad y}{total \quad variability \quad in \quad y} = \frac{132.57 + 8.21}{480.25} = 0.29$$

Unique information

Does adding the variable `white` to the model add valuable information that wasn't provided by `female_house`?



Collinearity between explanatory variables

poverty vs % female head of household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.3094	1.8970	1.74	0.0873
female_house	0.6911	0.1599	4.32	0.0001

poverty vs % female head of household and % female household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.5789	5.7849	-0.45	0.6577
female_house	0.8869	0.2419	3.67	0.0006
white	0.0442	0.0410	1.08	0.2868

Note the difference in the estimate for `female_house`.

Collinearity between explanatory variables

- Two predictor variables are said to be collinear when they are correlated, and this collinearity complicates model estimation.
  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors

R^2 vs. adjusted R^2

Model	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{error}}{SS_{total}} \times \frac{n-1}{n-p-1} \right)$$

where *n* is the number of cases and *p* is the number of predictors (explanatory variables) in the model.

- Because *p* is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

One Minute Paper

Complete the one minute paper:

https://forms.gle/ESBAdHRhzT65fW6c6

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?