Regression: From Fitting to Predicting

Gov 51 Section — Week 6

Scott Cunningham
Harvard University

March 4, 2026

Today's Plan

Part 1: Fitting Lines (~35 min)

- Load new survey data
- Bivariate regression: `lm()`, coefficients
- Calculate $\hat{Y}$ by hand
- Multivariate regression
- R^2 as prediction quality

Part 2: Curves & Prediction (~35 min)

- Polynomials: let the line curve
- Interactions: when the effect depends
- Binary × binary: four group means
- Train/test split in R
- RMSE: honest evaluation

Today is **hands-on R throughout**. Open RStudio and follow along.

Part 1: Fitting Lines Through Data

The Civic Knowledge Survey

Variable	Type	Description
`knowledge`	Outcome (0–100)	Civic knowledge test score
`age`	Continuous (18–85)	Age in years
`news_hours`	Continuous (0–7)	Daily news consumption (hours)
`female`	Binary (0/1)	Gender indicator
`college`	Binary (0/1)	Has college degree
`income`	Continuous (20–150)	Household income ($K)

$n = 2{,}500$ respondents

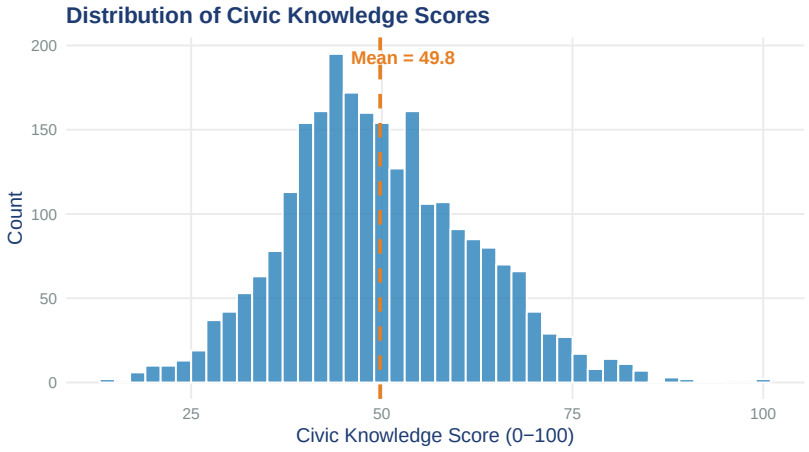
Different data from lecture — same techniques.

Load the data and explore

```
library(tidyverse)
d <- read_csv("civic_knowledge.csv")

nrow(d)
## [1] 2500

mean(d$knowledge)
## [1] 49.78

sd(d$knowledge)
## [1] 12.39
```

Run this code now. Does your mean match?

What Does Civic Knowledge Look Like?



News Consumption Predicts Civic Knowledge

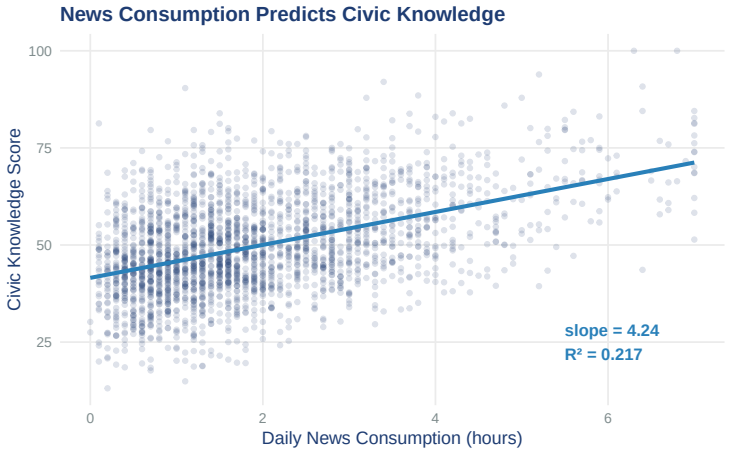
```
fit1 <- lm(knowledge ~ news_hours, data = d)
summary(fit1)
```

	Estimate	SE	t
Intercept	41.54	0.38	108.7
news_hours	4.24	0.16	26.3

$R^2 = 0.217$

$$\hat{Y}_i = 41.54 + 4.24 \cdot \text{news\_hours}_i$$

Each Additional Hour of News Raises Predicted Knowledge by 4.2 Points



Turn to your neighbor:

Someone watches **3 hours** of news per day.
What is their predicted civic knowledge score?
Show your calculation.

Now: someone who watches **0 hours**.
What is theirs?

Take 2 minutes. Use the coefficients from the previous slide.

Predicted Values Are Plug-In Arithmetic

$$\hat{Y} \quad = \quad \hat{\beta}_0 \; + \; \hat{\beta}_1 \; \cdot \; \text{value}$$

Person	Calculation	$\hat{Y}$
3 hours	$41.54 + 4.24 \times 3$	54.3
0 hours	$41.54 + 4.24 \times 0$	41.5
Difference	4.24×3	12.7

The intercept is the predicted score at news_hours $= 0$.

Adding Predictors Improves the Fit

```
fit2 <- lm(knowledge ~ news_hours + age + female + college,
           data = d)
```

	Estimate	SE	t
Intercept	43.84	0.64	68.3
news_hours	4.23	0.14	30.6
age	−0.14	0.01	−14.4
female	2.77	0.38	7.4
college	9.98	0.39	25.4

$R^2 = 0.427$

R^2: $0.217 \to 0.427$ — more than doubled

Turn to your neighbor:

Using the multivariate output:

1. Predicted knowledge for a **40-year-old college-educated woman** who watches **2 hours** of news?
2. Same person, but a **non-college man**?
3. What is the difference between the two predictions?

Take 3 minutes. Write out the full equation, then plug in.

Plug-In Calculations with Multiple Predictors

$$\hat{Y} = 43.84 + 4.23 \cdot \text{news} - 0.14 \cdot \text{age} + 2.77 \cdot \text{female} + 9.98 \cdot \text{college}$$

Person	**Calculation**	$\hat{Y}$
40, female, college, 2hrs	$43.84 + 4.23(2) - 0.14(40) + 2.77(1) + 9.98(1)$	59.3
40, male, no college, 2hrs	$43.84 + 4.23(2) - 0.14(40) + 2.77(0) + 9.98(0)$	46.5
Difference	$\hat{\beta}_{\text{female}} + \hat{\beta}_{\text{college}} = 2.77 + 9.98$	12.8

R^2 Measures How Much Variation the Model Explains

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}}$$

Model	Predictors	R^2
1	`news_hours`	0.217
2	+ `age`, `female`, `college`	0.427

R^2 always goes up when you add predictors. Is that always good?

Turn to your neighbor:

A model with **100 predictors** has $R^2 = 0.95$.

A model with **4 predictors** has $R^2 = 0.45$.

Which would you trust more for predicting a **new person's** score? Why?

Take 2 minutes. Think about what we learned in lecture.

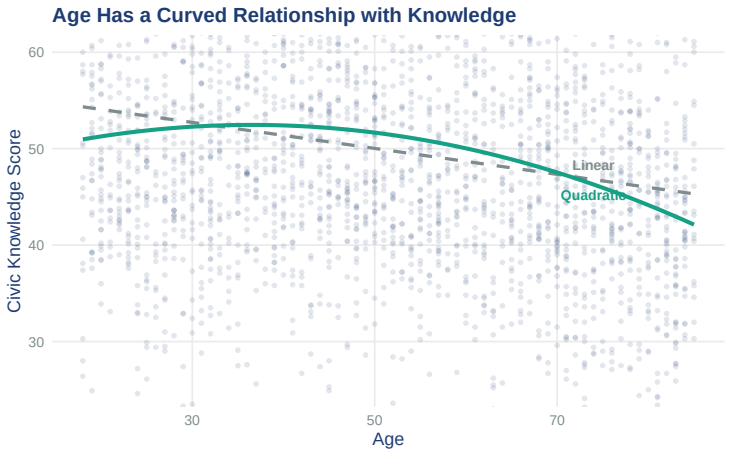
OLS finds the best line through the data.

$\hat{Y}$ = plug in your values and calculate.

R^2 measures in-sample fit — not prediction quality.

Part 2: Curves, Interactions, and Honest Evaluation

Age Has a Curved Relationship with Civic Knowledge



Add Age-Squared to Let the Line Curve

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{age}_i + \hat{\beta}_2 \cdot \text{age}_i^2$$

```
fit3 <- lm(knowledge ~ age + I(age^2), data = d)
```

	Estimate	**SE**	t
Intercept	46.62	1.79	26.0
`age`	0.321	0.075	4.3
`I(age^2)`	−0.0044	0.0007	−6.1

Still OLS — `I(age^2)` is just another column in the data.

Turn to your neighbor:

Using $\hat{Y} = 46.62 + 0.321 \cdot \text{age} - 0.0044 \cdot \text{age}^2$:

1. Predicted knowledge at age 25?
2. Predicted knowledge at age 50?
3. Predicted knowledge at age 75?
4. At what age is predicted knowledge highest?
   (Hint: peak $= -\hat{\beta}_1 \,/\, (2\hat{\beta}_2)$)

Take 3 minutes. This is the trickiest calculation today.

Polynomial Predictions Are Still Plug-In Arithmetic

$$\hat{Y} = 46.62 + 0.321 \cdot \text{age} - 0.0044 \cdot \text{age}^2$$

Age	Calculation	$\hat{Y}$
25	$46.62 + 0.321(25) - 0.0044(625)$	51.9
50	$46.62 + 0.321(50) - 0.0044(2500)$	51.7
75	$46.62 + 0.321(75) - 0.0044(5625)$	45.9

$$\text{Peak age} = -0.321 \,/\, (2 \times -0.0044) = 36.5 \text{ years}$$

Interactions: When the Effect Depends on Who You Are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{college} + \hat{\beta}_2 \cdot \text{female} + \hat{\beta}_3 \cdot (\text{college} \times \text{female})$$

```
fit4 <- lm(knowledge ~ college * female, data = d)
```

	Estimate	SE	t
Intercept	46.46	0.40	116.1
`college`	5.76	0.67	8.6
`female`	−0.64	0.55	−1.2
`college:female`	9.04	0.93	9.7

$\hat{\beta}_3 = 9.04$: the college effect is 9 points *larger* for women

Binary × Binary: Four Predicted Group Means

$$\hat{Y} = 46.46 + 5.76 \cdot \text{college} - 0.64 \cdot \text{female} + 9.04 \cdot (\text{college} \times \text{female})$$

	No College	**College**
Male	$\hat{\beta}_0 = 46.5$	$\hat{\beta}_0 + \hat{\beta}_1 = 52.2$
Female	$\hat{\beta}_0 + \hat{\beta}_2 = 45.8$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 60.6$

- College effect for men: $\hat{\beta}_1 = 5.8$
- College effect for women: $\hat{\beta}_1 + \hat{\beta}_3 = 14.8$
- Difference of differences: $\hat{\beta}_3 = 9.0$

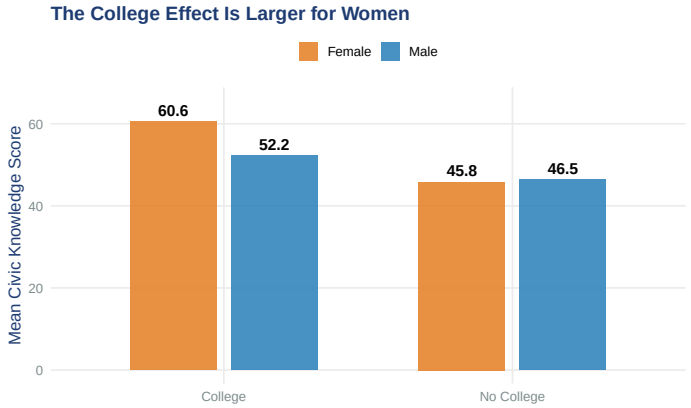
Turn to your neighbor:

Looking at the 2×2 table:

1. What is the college effect for men? For women?
2. Why are these different?
3. Which coefficient captures the difference?

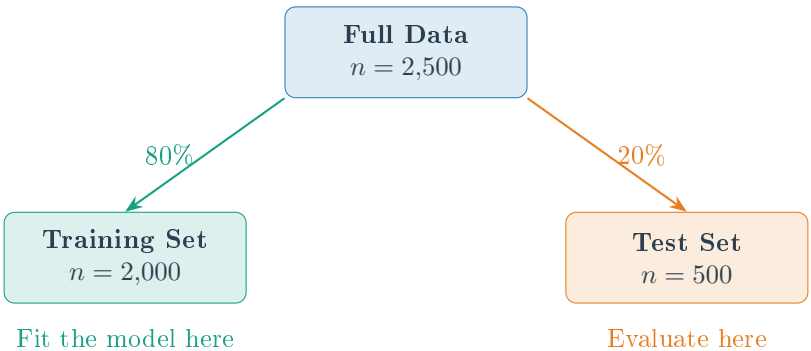
Take 2 minutes. Match each cell to its formula.

The Interaction in a Picture



$\hat{\beta}_3 = 9.04$ — the extra college boost for women

The Train/Test Split Tests Whether Your Model Generalizes



Never let the model see the test data during fitting.

Four Lines of R Do the Train/Test Split

```r
# 1. Split 80/20
set.seed(51)
train_idx <- sample(1:nrow(d), size = 0.8 * nrow(d))
train <- d[train_idx, ];  test <- d[-train_idx, ]

# 2. Fit on training data only
fit <- lm(knowledge ~ news_hours + age + female
          + college, data = train)

# 3. Predict on test data
y_hat <- predict(fit, newdata = test)

# 4. Test RMSE
sqrt(mean((test$knowledge - y_hat)^2))   ## 9.67
```

Turn to your neighbor:

Run the train/test code yourself. Then:

1. Fit the **bivariate** model (`news_hours` only) on the training data. What is its test RMSE?
2. Fit the **full** model with `age`, `I(age^2)`, `college*female`, `income`. What is its test RMSE?
3. Which model predicts better on new data?

Take 4 minutes. Run the code — compare the numbers.

RMSE Tells You How Far Off Your Predictions Are

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

	Model	Train RMSE	Test RMSE
1	`news_hours`	10.91	11.14
2	+ `age`, `female`, `college`	9.30	9.67
3	+ $\texttt{age}^2$	9.19	9.54
4	+ `college`×`female`	8.89	9.34
5	+ `income` (full)	7.80	**8.07**

Lower = better. Train RMSE always goes down. Test RMSE is what matters.

Turn to your neighbor:

Why does Train RMSE always go *down*
as we add predictors,
but Test RMSE doesn't always?

What would happen if we added
200 random noise variables?

Take 2 minutes. Think back to the lecture on overfitting.

Predicted values are plug-in arithmetic.

Interactions let effects vary across groups.

Test RMSE — not R^2 — measures prediction quality.

Both Halves Build the Same Skill: Plug In and Calculate

- **Part 1:** $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots$
- **Part 2:** Add X^2, add $X_1 \times X_2$ — still plug-in, just more terms
- **Key insight:** The model is always a sum of coefficients $\times$ values

Exam 1 is **March 12**. Practice plug-in calculations!

Before Next Section

1. **Problem Set 2** due Thursday, March 5
2. Practice: fit `lm()` with interactions and polynomials on any dataset
3. Review: can you calculate $\hat{Y}$ by hand from any regression output?

Questions?