PS 211: Introduction to Experimental Design

Fall 2025 · Section C1

Lecture 17: Regression

Updates & Reminders

- Homework 4 has been posted and is due **Dec. 2.**
    - Remember, only your top 3 homework grades count.
- Data Write-up grades will be posted by the end of the day tomorrow.
    - We also directly commented on your submitted PDFs.
    - Good job! 🎉
    - Please come to office hours if you have questions.
    - If you have feedback about how the assignment can be improved, please let me know.
        - Next year I will ask better questions on the pre-course survey.
- **No class or office hours next week!**

Part 1: What is Regression?

What is Regression?

Regression builds directly on correlation — today we move from *describing* relationships to *predicting* outcomes.

- While correlation *describes* a relationship, regression quantifies how much change to expect in one variable given a change in another.
- This lets us use one variable to **predict** another.
- **Correlation answers:**
  "Are X and Y related?"
- **Regression answers:**
  "How much does Y change when X changes?"

Regression vs. Correlation

- Imagine we have data about ice cream sales per month (in dollars) and the number of shark attacks per month.
- **Correlation** can tell us that ice cream sales and shark attacks are related (r = .7).
- But what if we want to **predict** shark attacks based on ice cream sales? For example, if monthly ice cream sales are $5000, how many shark attacks should we expect?
- **Regression** allows us to make this prediction by quantifying the relationship between ice cream sales and shark attacks.

Regression Does Not Tell Us About Causation!

- Even though regression predicts, it **cannot** establish causality.
- Why?
  - Predictors are **not manipulated**.
  - Confounds still exist.
  - Observational data is **not** experimental data.

Key idea:
Regression tells us how we should expect a variable to change *if a predictor changes*, not *why* the change occurs.

Regression Variables

In regression, we use **new terms** instead of IV and DV:

- **Predictor variable** (X):
  The variable we use to make our prediction

- **Outcome variable** (Y):
  The variable we want to predict

Previously, we have used “independent” and “dependent” variable labels. But we have used these terms to describe **experiments** in which our independent variable was manipulated (e.g., condition assignment).

In regression, we are often working with **observational data** where no variables are manipulated. Thus, we use the terms **predictor** and **outcome** to avoid implying causality.

What is Regression Used For?

Regression is used in almost every field:

- Content-recommendation algorithms
- Dating apps
- Economic forecasting
- Public health forecasting
- Risk assessment tools (e.g., recidivism)
- Investment behavior
- Etc.

Regression in Psychology & Neuroscience

Part 2: Simple Linear Regression

Simple Linear Regression

Simple Linear Regression (SLR):

- Used to **predict an outcome** (Y) based on **one predictor** (X).
- We find the **line of best fit** through the data points.

$$\hat{Y} = bX + a$$

- **b** = slope
- **a** = intercept
- **Ŷ** = predicted outcome value

Understanding the Slope and Intercept

- The **slope (b)** tells us: How much Y is expected to change when X increases by 1 unit.
- If **b > 0** → higher X predicts higher Y (positive relationship)
- If **b < 0** → higher X predicts lower Y (negative relationship)
- If **b = 0** → X does not predict Y
- The **intercept (a)** tells us: The predicted value of Y when X = 0.
- Sometimes X = 0 is not meaningful (e.g., GPA = 0, height = 0) --> remember interval scales!
  - In those cases, the **intercept is still needed mathematically**, but we interpret it cautiously.

What Does the Regression Line Represent?

- The regression line is the **prediction rule** for Y based on X.
- It represents our **best guess** for the average value of Y at each level of X.
- Example:
  - A therapist wants to predict their clients’ depression scores based on stressful life events.
  - The therapist counts the number of stressful life events for every client in the past 2 years and their current depression scores.
  - The therapist finds that the best fitting regression line has a slope of 2 and an intercept of 10.
  - This means that for each additional stressful life event, the therapist predicts the depression score to increase by 2 points, starting from a baseline of 10 when there are no stressful life events.

Finding the Regression Line: Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is a common method used to find the **best-fitting line** in simple regression.

It finds the line that **minimizes the total squared prediction error**:

$$\mathrm{SSE} = \sum (Y - \hat{Y})^2$$

where:

- **SSE** = Sum of Squared Errors
- **Y** = Actual outcome values
- **Ŷ** = Predicted outcome values from the regression line

Finding the Regression Line: Ordinary Least Squares (OLS)

Finding the Regression Line: Ordinary Least Squares (OLS)

$$\mathrm{SSE} = \sum (Y - \hat{Y})^2$$

Why squared errors?

- Prevents positive and negative errors from canceling.
- Penalizes **large errors** more than small ones.
- Makes math tractable and guarantees a unique best line.

OLS: Practice

Which of the following describes (in words) the best-fitting regression line found using OLS?

A. The line that has an equal number of points above and below it.

B. The line that minimizes the sum of squared deviations between actual and predicted Y values.

C. The line that best predicts the **mean** of Y.

D. The line with the **largest** slope.

Answer: B.
OLS finds the line with the **smallest sum of squared errors (SSE)**, which is the sum of squared deviations between actual and predicted Y values.

Error in Prediction

- Regression predictions are **not perfect**.
- There will always be some error in our predictions.
- These errors can come from:
  - **Unexplained variance** in the data (i.e., factors not included in the model)
  - **Random noise** in measurements
  - **Individual differences** among subjects

Error in Prediction

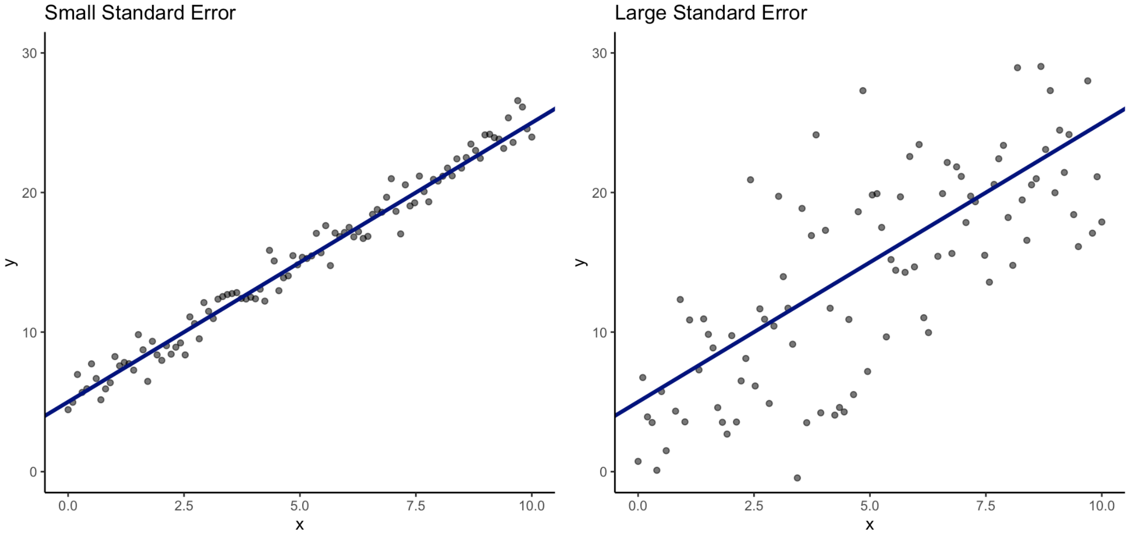
- The amount of error around the line of best fit can be quantified.
- The standard error of the estimate describes how far away (on average) the data points are from the line of best fit.
- Computed as:

$$SE_{estimate} = \sqrt{\frac{SSE}{df}}$$

Where:

- **SSE** = Sum of Squared Errors (from OLS)
- **df** = degrees of freedom
- For simple regression, **df = n - 2** (n = number of data points)

Error in Prediction



Question: How will increasing the number of data points (n) affect the standard error of the estimate?

Answer: Increasing the number of data points (n) will generally **decrease** the standard error of the estimate. This is because a larger sample size provides more information about the population, leading to more accurate estimates of the regression parameters.

Proportionate Reduction in Error (PRE)

- PRE quantifies how much more accurate our predictions become when using regression versus simply using the **mean** of Y.

- It is calculated as:

$$PRE = \frac{SSE_{mean} - SSE_{regression}}{SSE_{mean}}$$

This is the same as our earlier definition of R^2: Variance explained by the regression model!

$$PRE = R^2$$

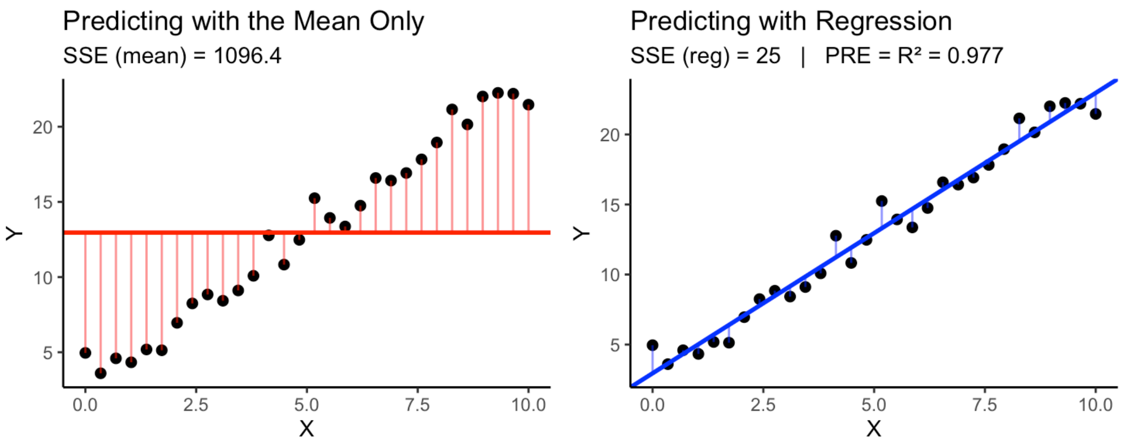
This is also called the **coefficient of determination**.

Proportionate Reduction in Error (PRE)

Interpreting R^2:

- **$R^2 = 0$** → regression does not improve prediction
- **$R^2 = 1$** → perfect prediction
- **$R^2 = .40$** → 40% of variance explained

Proportionate Reduction in Error (PRE)



Predicting with the Mean Only

SSE (mean) = 1096.4

Predicting with Regression

SSE (reg) = 25 | PRE = R² = 0.977

Proportionate Reduction in Error (PRE)

$$PRE = \frac{SSE_{mean} - SSE_{regression}}{SSE_{mean}}$$

This is very similar to the formula for R² or η^2 in ANOVA:

$$R^2 = \frac{SS_{between}}{SS_{total}}$$

Our numerators reflect the amount of variance explained by the model, and our denominators reflect the total variance in the outcome variable.

- Thus, R² in regression and η^2 in ANOVA both quantify the **proportion of variance explained** by the model.
- They are measures of **effect size**.

R^2 and Significance Testing

- **R^2** tells us how much variance in Y is explained by the regression model.
  → It is an **effect size**.
- We can also ask: **Is this R^2 likely to be > 0 in the population, or could it be due to chance?**
  - In simple regression, this is equivalent to testing whether the **slope = 0** (or $r = 0$).
  - Like with an ANOVA, we can use an **F-test** to determine whether the overall regression model explains a significant amount of variance in Y.
  - We use an ***F*-test** for the overall model.
  - We use a *t* test for the slope.
- Remember, our test statistic values depend on both effect size and sample size.
  - With a large enough sample, even a small R^2 can be statistically significant.

Unstandardized Coefficients

- Regression coefficients can be reported in **unstandardized** or **standardized** form.
- Unstandardized coefficients are reported in the **original units** of the variables.

Unstandardized slope (b)

- Interpreted in **raw units**
- Example: “For each additional stressful event, depression scores increase by 3.2 points.”

Intercept (a)

- Predicted Y when X = 0 (may or may not be meaningful)

Standardized Coefficients

- Standardized coefficients are reported in **standard deviation units**.

- Variables are standardized (z-scored) before running the regression.

Standardized slope (β)

- Interpreted in **standard deviation units**

- Example: “For each additional standard deviation increase in stressful events, depression scores increase by 0.45 standard deviations.”

Intercept (β_0)

- Predicted Y when X is at its mean (i.e., when standardized X = 0)

Unstandardized vs. Standardized Coefficients

Unstandardized Coefficients

Advantages

- Easy to interpret in real-world terms
- Useful for practical applications

Disadvantages

- Hard to compare across different scales

Standardized Coefficients

Advantages

- Easy to compare across different predictors
- Useful for understanding relative importance

Disadvantages

- Harder to interpret in real-world terms

Example: Unstandardized vs. Standardized Coefficients

Imagine you are predicting house prices based on square footage and number of bedrooms.

- **Unstandardized coefficients:**
  - Slope (b) for square footage = 150 (each additional square foot increases price by \$150)
  - Slope (b) for number of bedrooms = 20,000 (each additional bedroom increases price by \$20,000)

Which predictor has a larger impact on price?

Answer: It depends on the scale of the predictors. Square footage is measured in hundreds or thousands, while number of bedrooms is a small integer. Thus, we cannot directly compare their unstandardized coefficients.

Example: Unstandardized vs. Standardized Coefficients

Imagine you are predicting house prices based on square footage and number of bedrooms.

- **Standardized coefficients:**
    - Slope (β) for square footage = 0.6
    - Slope (β) for number of bedrooms = 0.4

Now, which predictor has a larger impact on price?

Answer: Square footage has a larger impact on price, as indicated by its higher standardized coefficient (β = 0.6 vs. β = 0.4).

Part 3: Multiple Regression: Testing multiple predictors

Multiple Regression

Multiple regression predicts an outcome using **2+ predictors**:

$$\hat{Y} = b_1 X_1 + b_2 X_2 + \cdots + a$$

Examples:

- Predict GPA from hours studied + attendance
- Predict depression from stress + sleep + social support

Key idea:

Each slope (**b$_1$**, **b$_2$**, ...) represents the relationship between that predictor and Y **holding all other predictors constant**.

This makes regression different from simple correlations.

Multiple Predictors

Imagine predicting Y from **two** predictors: X_1 and X_2.

Each predictor explains a **different slice of variance**.

Why use multiple predictors?

- Increases accuracy
- Reduces omitted-variable bias
- Helps control for confounds

Multiple Predictors

You want to understand how stress and sleep relate to depression.

- You first run a simple linear regression predicting depression from sleep alone. You find that sleep significantly predicts depression (b = -2, p < .01).

- However, you then run a multiple regression including both stress and sleep as predictors.
    - Here, you find that stress significantly predicts depression (b = 3, p < .001), but sleep no longer significantly predicts depression (b = -0.5, p = .15).

What do these two findings suggest?

A. Sleep is not related to depression.

B. Stress predicts sleep.

C. Stress explains some of the same variance in depression as sleep.

D. Stress and sleep are unrelated.

Answer: C. If including stress in the model reduces the effect of sleep on depression, it suggests that stress and sleep share some variance in predicting depression.

Interaction Terms in Regression

- An **interaction** occurs when the effect of one predictor **depends on** the level of another predictor.
- In other words:
  - The relationship between X_1 and Y changes depending on X_2.
- This is the same as in ANOVA, where the effect of one IV depends on the level of another IV.

Example: The effect of stress on depression may depend on social support.

- High social support may buffer the negative effect of stress on depression.

*To test for interactions in regression, we include an **interaction term** in the model:*

$$\hat{Y} = b_1 X_1 + b_2 X_2 + b_3 (X_1 \times X_2) + a$$

Part 4: Regression in R

Simple Linear Regression in R

```r
# Fit regression model
model <- lm(depression ~ stress, data = df)
summary(model)
```

Simple Linear Regression in R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7714     1.1665   6.662  0.00117 **
stress        2.8571     0.2878   9.928  0.00020 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.897 on 4 degrees of freedom
Multiple R-squared:  0.961,
Adjusted R-squared:  0.951
F-statistic: 98.57 on 1 and 4 DF,  p-value: 0.000195
```

- The slope for stress is 2.8571, meaning that for each additional unit increase in stress, depression scores are predicted to increase by approximately 2.86 points.
- The R^2 value is 0.961, indicating that approximately 96.1% of the variance in depression scores is explained by stress.

Multiple Regression in R

```
Call:
lm(formula = depression ~ stress + sleep + social_support, data = df)
Residuals:
       1        2        3        4        5        6
-1.20000 -0.80000  0.40000  1.60000  0.20000 -0.20000
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.1234     1.2345   4.150  0.00567 **
stress           2.4567     0.3456   7.112  0.00045 ***
sleep           -0.9876     0.4567  -2.163  0.07543 .
social_support    1.2345     0.5678   2.173  0.06543 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.567 on 3 degrees of freedom
Multiple R-squared:  0.978,
Adjusted R-squared:  0.962
F-statistic: 57.89 on 3 and 3 DF,  p-value: 0.000123
```

Which predictor has the largest impact on depression?

Recap: Regression

- Regression predicts an outcome variable (Y) based on one or more predictor variables (X).
- We find the best-fitting line using **Ordinary Least Squares (OLS)**, which minimizes the sum of squared errors between our predictions and actual values.
- Simple linear regression uses one predictor, while multiple regression uses two or more predictors.
- Regression coefficients can be unstandardized (b) or standardized (β).
- R^2 quantifies the proportion of variance in Y explained by the model.
- We test the significance of individual predictors with *t*-tests and the overall model with an *F*-test.

Regression: Practice Questions

A researcher uses hours studied (X) to predict final exam score (Y). The regression slope is b = 4.2, p < .01.

What does this result mean in practical terms?

A. Students who study score exactly 4.2 points higher on their exam relative to students who do not study.

B. For each extra hour studied, exam scores increase on average by 4.2 points.

C. For each extra hour studied, exam scores increase by 4.2%.

D. Students who study more tend to perform worse.

Answer: B.
The slope indicates that for each additional hour studied, the predicted exam score increases by 4.2 points on average.

Regression: Practice Questions

The best-fitting regression line predicting stress level (Y) from hours of sleep (X) is:

$$\hat{Y} = 22 - 1.5X$$

What does the intercept (22) represent?

A. Predicted stress level when sleep = 0 hours

B. Average stress level in the sample

C. Minimum stress score possible

Answer: A.
The intercept represents the predicted stress level when the predictor (hours of sleep) is equal to 0.

How should we interpret this intercept in practical terms?

Since 0 hours of sleep is not a realistic scenario, the intercept may not have a meaningful interpretation in this context. It is primarily a mathematical necessity for the regression equation.

Regression: Practice Questions

A study predicts job performance from years of experience. The model yields $R^2 = .32$.

Which is the most accurate interpretation of this R^2 value?

A. Experience causes 32% better job performance.

B. Experience explains 32% of the variance in performance.

C. 32% of employees are high performers.

D. Performance can be predicted with 32% accuracy.

Answer: B.
R^2 indicates that 32% of the variance in job performance is explained by years of experience

Regression: Practice Questions

A study predicts college GPA using SAT score (X_1) and high school GPA (X_2).

When SAT score is used alone, it significantly predicts college GPA ($b = 0.03$, $p < .01$). When both SAT score and high school GPA are included as predictors in the model, the slope for SAT becomes non-significant.

What is the best interpretation of this result?

A. SAT does not predict college GPA at all.

B. SAT only predicts GPA for high-performing students.

C. SAT and high school GPA explain overlapping variance.

D. High school GPA causes SAT performance.

Answer: C.

Including high school GPA in the model accounts for some of the same variance in college GPA that SAT score explains, leading to a non-significant slope for SAT.

Regression: Practice Questions

A regression predicts happiness from daily exercise duration. The slope is $b = 1.4, p = .28.$

Which conclusion is correct?

A. Exercise has no effect on happiness.

B. We failed to find evidence that exercise predicts happiness.

C. Exercise decreases happiness.

D. The effect size is zero.

Answer: B.
A non-significant p-value indicates that we did not find evidence that exercise duration predicts happiness in this sample. It does not prove that there is no effect in the population.

That’s all for today!

Enjoy Thanksgiving! 🦃