Recall: main idea of linear regression

- Linear regression can be used to study an outcome as a linear function of a predictor.
- Example: 60 cities in the US were evaluated for numerous characteristics, including:
  - **Outcome:** the percentage of the population that had low income
  - **Predictor:** median education level

Recall: Where is our intercept?

We will introduce the idea of **centering** using the city % low income example

Need for centering

- $\beta_0$ makes no sense!
- We don’t observe any cities with median education = 0.
- We can change $X$ to fix this problem by a process called **centering**:
  1. Pick a value of $X$ (c) within the range of the data
  2. For each observation, generate $X_{\text{centered}} = X - c$
  3. Redo the regression with $X_{\text{centered}}$
Centering
We’ll use $c=12$, a high school degree

```
Centering
New equation

\[
\hat{Y}_i = \beta_0 + \beta_1 (X_{\text{Centered}_i})
\]
\[
\hat{Y}_i = \beta_0 + \beta_1 (X_i - 12)
\]
\[
\hat{Y}_i = 12.2 - 2.0(X_i - 12)
\]
- $\beta_1$ has not changed
- $\beta_0$ now corresponds to average of $y$ when $X_{\text{centered}_i}=0$ or, equivalently, $X_i=12$ (not $X_i=0$)
- Note: with $X_i=0$, we have
  \[
  \hat{Y}_i = 12.2 - 2.0(0 - 12) \\
  = 12.2 + 24 = 36.2
  \]

Centering
Interpretation

- $\beta_0$ is the mean outcome for the reference group, or the group for which $X_{\text{centered}_i}=X_i-12=0$, or when $X_i=12$.
- Here, $\beta_0$ (12.2) is the average percent of the population that is disadvantaged for cities with a median education level of 12, the equivalent of a high school degree.
- The interpretation of $\beta_1$ has not changed.

Hypothesis testing and confidence intervals of regression coefficients
Drawing conclusions about population association using our data (a sample)

- $\beta_0$: changes depending on centering of $X$, which doesn’t affect association of interest

- Real concern: is $X$ associated with $Y$?
- Assess by testing $\beta_1$:
  - Does $\beta_1=0$ in the population from which this sample was drawn?
    - Hypothesis testing
    - Confidence interval

Hypothesis testing
Null and alternate hypothesis, test statistic

- $H_0$: $\beta_1=0$
- $H_0$: $\beta_1\neq 0$
- Test statistic:
  $$t_{\text{obs}} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$
- $df = n-k-1$
  - $n$ = number of observations
  - $k$ = number of predictors ($X$’s)

Hypothesis testing
Education example

- $H_0$: $\beta_1=0$
- Test statistic: 
  $$t_{\text{obs}} = \frac{-2.0-0}{0.59} = -3.36$$
- $df = n-k-1 = 60-1-1 = 58$
  - $n$ = number of observations = 60
  - $k$ = number of predictors ($X$’s) = 1

- Calculate our p-value
  
  \[2\cdot \text{pt}(-3.36, \text{df}=58)\]
  
  \[[1] \ 0.001383108\]

- p-value=0.001

Hypothesis testing
Education example: interpretation and conclusion

- If there were no association between median education and percentage of disadvantaged citizens in the population, there would be about a 1% chance of observing data as or more extreme than ours.

- The null probability is very small, so:
  - reject the null hypothesis
  - conclude that median education level and percentage of disadvantaged citizens are associated in the population
Confidence Interval for regression coefficients

We calculate the CI using the usual formula:

$$\hat{\beta}_1 \pm t_{CR} SE(\hat{\beta}_1)$$

df of $t_{CR} = n-k-1$

For the education example, the 95% CI for $\beta_1$ is:

$$-2.0 \pm 2.021 \times 0.59$$

$$\Rightarrow (-3.2, -0.8)$$

Confidence interval

Education example interpretation and conclusion

- We are ‘95% confident’ that the true population decrease in percentage of low income citizens per additional year of median education is between **3.2 and 0.8**

- Since this interval does not contain 0, we believe percentage of low income citizens and median education are associated among cities in the United States

Dataset

- Hourly wage information from 9918 workers, along with information regarding age, gender, years of experience, etc.

- We’ll focus on predicting hourly wage with available information.

  **Outcome**: hourly wage

  **Multiple predictors**: age, gender, years experience, etc...
Regression: Hourly wage vs. years of experience

Simple linear regression since only one covariate (years of experience)

How do we estimate the coefficients? Use least squares

- For each person, their actual hourly wage \( Y_i \) and predicted hourly wage \( \hat{Y}_i \) are known
  \[
  \hat{\epsilon}_i = (Y_i - \hat{Y}_i) = (Y_i - (\beta_0 + \beta_1 X_i)) \quad \text{is the residual or error}
  \]
- The coefficient estimates are found by minimizing the sum of the squared error
  \[
  \min \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2
  \]
- The coefficients are the “least squares” estimates

Notes on regression analysis

- \( \hat{Y}_i = \beta_0 + \beta_1 X_i \) for any known point on the line
- \( \bar{Y} = \beta_0 + \beta_1 \bar{X} \) is always true
- Recall the regression line equation
  \[
  Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i
  \]

Model 1: years of experience

- Model 1: Predict income by years of experience
  \[
  \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad \Rightarrow \quad \hat{Y}_i = 8.38 + 0.04 X_i
  \]
- \( \hat{\beta}_0 = 8.38 \) so the average hourly wage for someone with no experience at all is about $8.40
- \( \hat{\beta}_1 = 0.04 \) so for every additional year of experience, the predicted hourly wage increases about 4 cents
  - For 10 years of additional experience, the predicted hourly wage increases about 40 cents
Should we center X?

- 0 years of experience is within the range of the data
- The average hourly wage corresponding to 0 years of experience makes sense
- No need to center X

Model 2: **for no experience**

years of experience, gender (0=man, 1=woman)

\[
\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Experience}_i + \hat{\beta}_2 \text{Gender}_i
\]

\[\Rightarrow \hat{Y}_i = 9.27 + 0.04(\text{Experience}_i) - 2.20(\text{Gender}_i)\]

- For a man with no experience:
  \[\hat{Y}_i = 9.27 + 0.04(0) - 2.20(0) = $9.27\]
- For a woman with no experience:
  \[\hat{Y}_i = 9.27 + 0.04(0) - 2.20(1) = $7.07\]

Model 2: **for 10 years experience**

years of experience, gender (0=man, 1=woman)

\[
\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Experience}_i + \hat{\beta}_2 \text{Gender}_i
\]

\[\Rightarrow \hat{Y}_i = 9.27 + 0.04(\text{Experience}_i) - 2.20(\text{Gender}_i)\]

- For a man with 10 years of experience:
  \[\hat{Y}_i = 9.27 + 0.04(10) - 2.20(0) = $9.67\]
  \[= \hat{\beta}_0 + \hat{\beta}_1(10)\]
- For a woman with 10 years of experience:
  \[\hat{Y}_i = 9.27 + 0.04(10) - 2.20(1) = $7.47\]
  \[= \hat{\beta}_0 + \hat{\beta}_1(10) + \hat{\beta}_2(1)\]
Model 2: **for males**

years of experience, gender (0=man, 1=woman)

\[
\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{(Experience)} + \hat{\beta}_2 \text{(Gender)}
\]

\[\Rightarrow \hat{Y}_i = 9.27 + 0.04(\text{Experience}) - 2.20(\text{Gender})\]

- For a man with no experience:
  \[\hat{Y}_i = 9.27 + 0.04(0) - 2.20(0) = \$9.27\]

- For a man with 10 years of experience:
  \[\hat{Y}_i = 9.27 + 0.04(10) - 2.20(0) = \$9.67\]

  \[= \hat{\beta}_0 + \hat{\beta}_1(10)\]

---

Model 2: **for females**

years of experience, gender (0=man, 1=woman)

\[
\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{(Experience)} + \hat{\beta}_2 \text{(Gender)}
\]

\[\Rightarrow \hat{Y}_i = 9.27 + 0.04(\text{Experience}) - 2.20(\text{Gender})\]

- For a woman with no experience:
  \[\hat{Y}_i = 9.27 + 0.04(0) - 2.20(1) = \$7.07\]

  \[= \hat{\beta}_0 + \hat{\beta}_2\]

- For a woman with 10 years of experience:
  \[\hat{Y}_i = 9.27 + 0.04(10) - 2.20(1) = \$7.47\]

  \[= \hat{\beta}_0 + \hat{\beta}_1(10) + \hat{\beta}_2\]

---

**Model 2 Interpretation**

\[\hat{Y}_i = 9.27 + 0.04(\text{Experience}) - 2.20(\text{Gender})\]

- \(\hat{\beta}_0 = 9.27\) : the average hourly wage **for a man with no experience at all** is about \$9.30

- \(\hat{\beta}_1 = 0.04\) : for every additional year of experience, the predicted hourly wage increases about 4 cents **for both men and women**.

- \(\hat{\beta}_2 = -2.20\) : the expected hourly wage is \$2.20 lower for women than it is for men **at any experience level**.

---

**Confounding by gender? Model 1 vs. Model 2**

- **Model 1**: \(\hat{Y}_i = 8.38 + 0.04(\text{Experience})\)

- **Model 2**: \(\hat{Y}_i = 9.27 + 0.04(\text{Experience}) - 2.20(\text{Gender})\)

95% CI for \(\hat{\beta}_1\) in Model 1: (0.001, 0.07)

and \(\hat{\beta}_1\) from Model 2 is within this CI

Gender is not a **confounder**

The association between salary and experience does not change when we control for gender.
Confounding: the epidemiologic definition

C is a confounder of the relation between X and Y if:

- Outcome Y
- Predictor X
- Confounder C

Confounding: example

Smoking is a confounder of the relation between coffee consumption (X) and lung cancer (Y) since:

- Lung Cancer Y
- Coffee Consumption X
- Smoking C

Model 3: years of experience, age

- Let’s try a model that includes age instead of gender:
  \[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Experience}_i + \hat{\beta}_2 \text{Age}_i - 40 \]
- The relationship is harder to graph with two continuous predictors, since now the regression is in a 3-dimensional space
  - We will learn AV plots in a few days...
- Notice that age is centered at 40 years. Age ranged between 18 and 64 in this dataset

Model 3: no experience years of experience, age

- \[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Experience}_i + \hat{\beta}_2 (\text{Age}_i - 40) \]
  \[ \Rightarrow 26.5 - 0.82(\text{Experience}_i) + 0.92(\text{Age}_i - 40) \]
- For a 40-year-old with no experience:
  \[ \hat{Y}_i = 26.5 - 0.82(0) + 0.92(40 - 40) = 26.50 \]
  \[ = \hat{\beta}_0 \]
- For a 41-year-old with no experience:
  \[ \hat{Y}_i = 26.5 - 0.82(0) + 0.92(41 - 40) = 27.42 \]
  \[ = \hat{\beta}_0 + \hat{\beta}_2 \]
Model 3: 10 years experience
years of experience, age

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Experience}_i) + \hat{\beta}_2(Age_i - 40)$$

$$\Rightarrow 26.5 - 0.82(\text{Experience}_i) + 0.92(Age_i - 40)$$

- For a 40-year-old with 10 years of experience:
  $$\hat{Y}_i = 26.5 - 0.82(10) + 0.92(40 - 40) = $18.30$$
  $$= \hat{\beta}_0 + \hat{\beta}_1 \times 10$$

- For a 41-year-old with 10 years of experience:
  $$\hat{Y}_i = 26.5 - 0.82(10) + 0.92(41 - 40) = $19.22$$
  $$= \hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 1$$

Model 3: 40 years of age
years of experience, age

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Experience}_i) + \hat{\beta}_2(Age_i - 40)$$

$$\Rightarrow 26.5 - 0.82(\text{Experience}_i) + 0.92(Age_i - 40)$$

- For a 40-year-old with no experience:
  $$\hat{Y}_i = 26.5 - 0.82(0) + 0.92(40 - 40) = $26.50$$
  $$= \hat{\beta}_0$$

- For a 40-year-old with 10 years of experience:
  $$\hat{Y}_i = 26.5 - 0.82(10) + 0.92(40 - 40) = $18.30$$
  $$= \hat{\beta}_0 + \hat{\beta}_1 \times 10$$

Model 3: 41 years of age
years of experience, age

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Experience}_i) + \hat{\beta}_2(Age_i - 40)$$

$$\Rightarrow 26.5 - 0.82(\text{Experience}_i) + 0.92(Age_i - 40)$$

- For a 41-year-old with no experience:
  $$\hat{Y}_i = 26.5 - 0.82(0) + 0.92(41 - 40) = $27.42$$
  $$= \hat{\beta}_0 + \hat{\beta}_2$$

- For a 41-year-old with 10 years of experience:
  $$\hat{Y}_i = 26.5 - 0.82(10) + 0.92(41 - 40) = $19.22$$
  $$= \hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 1$$

Model 3: Interpretation

- $\hat{\beta}_0 = 26.5$: the average hourly wage for a 40-year-old with no experience at all is about $26.50
- $\hat{\beta}_1 = -0.82$: for every additional year of experience, the predicted hourly wage decreases about 82 cents for two people of the same age (or “adjusting for age”)
- $\hat{\beta}_2 = 0.92$: for every additional year of age, the expected hourly wage increases about 92 cents for two people with the same amount of experience (or “adjusting for experience”)

Model 1 vs. Model 3
What happens when we include age?

- Model 1:
  \[ \hat{Y}_i = 8.38 + 0.04(\text{Experience}_i) \]

- Model 3:
  \[ \hat{Y}_i = 26.5 - 0.82(\text{Experience}_i) + 0.92(\text{Age}_i - 40) \]

- 95% CI for \( \beta_1 \) in Model 1: (0.001, 0.07)
  and \( \hat{\beta}_1 \) from Model 3 is outside this CI

- **Age is a confounder**!
  When we adjust for age, the apparent effect of experience on wage changes

The Coefficient of Determination, \( R^2 \)

- \( R^2 \) measures the ability to predict \( Y \) using \( X \)
- Variability explained by \( X \) is \( SSM = \sum (\hat{y}_i - \bar{y})^2 \)
- Total variability is \( SST = \sum (y_i - \bar{y})^2 \)
- \( R^2 \) is defined as
  \[ R^2 = \frac{SSM}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \]
- Measures the proportion of total variability explained by the model
- \( R^2 \) is the square of \( r \), “Pearson’s correlation coefficient”
  - Recall: \( r \) is a rough way of evaluating the association between two continuous variables

Using \( R^2 \)

as a model selection criteria

- The coefficient of determination, \( R^2 \) evaluates the entire model
- \( R^2 \) shows the **proportion of the total variation in \( Y \) that has been predicted by this model**
  - Model 1: 0.0076; 0.8% of variation explained
  - Model 2: 0.05; 5% of variation explained
  - Model 3: 0.20; 20% of variation explained
- You want a model with large \( R^2 \)

Adjusted \( R^2 \)

another model selection criteria

- In both models 2 and 3, the new predictor added a great deal to the model
  - \( R^2 \) increased a lot
  - More importantly, both new predictors were statistically significant
- \( R^2 \) increases **any time you include a new variable!**
- The adjusted \( R^2 \) is adjusted for the number of \( X \)'s in the model, so it only increases when helpful predictors are added
- Balance the need for
  - Simple model
  - Good predictive ability
Summary of Lecture 9

- Centering
- Hypothesis testing and confidence intervals
- Regression by least squares
- Interpreting regression coefficients
- Adding a 2nd predictor to a model
  - Binary X added: 2 parallel lines
  - Continuous X added: 3-dimensional graph
    - for both, new interpretation reflecting new model
- Is the new X a confounder?
  - Compare $\beta_1$ across models