The Course So Far:

- Descriptive Statistics
- Probability Theory
- Probability Distributions
- Simple Random Sampling
- Sampling Distributions
- Statistical Inference
- Estimation
- Hypothesis Testing
- Regression & Correlation
- Multiple Regression

Regression

\[ \text{MPG} = \beta_0 + \beta_1 \text{wgt} + \beta_2 \text{cyl} + \beta_3 \text{displ} + \beta_4 \text{hp} + \beta_5 \text{accel} + \epsilon \]

- Descriptive Estimation
- Horsepower
- Acceleration
- Price (?)

Number of cylinders
- Engine displacement
- Horsepower
- Acceleration
- Price (?)

```
. regress mpg cylinder
```

```
. correlate cylinder displacement horsepower acceleration
```

- `cylinder`
- `displacement`
- `horsepower`
- `acceleration`

```
. correlate cylinder displacement horsepower acceleration
```

```
. regress mpg price
```

```
. regress mpg weight
cylinder displacement horsepower acceleration
```

```
. correlate cylinder displacement horsepower acceleration
```

```
. regress mpg price
```

```
. correlate cylinder displacement horsepower acceleration
```
Fitting a multiple regression plane through a 3-dimensional scatter plot

\[ \hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

Above the plane

In the plane

<table>
<thead>
<tr>
<th>Regression</th>
<th>Simple Regression Coefficient (all significant except for price)</th>
<th>Multiple Regression Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>-0.100</td>
<td>-0.12***</td>
</tr>
<tr>
<td>Displacement</td>
<td>-0.076</td>
<td>-0.03*</td>
</tr>
<tr>
<td>Acceleration</td>
<td>-0.301</td>
<td>-0.055**</td>
</tr>
<tr>
<td>Year</td>
<td>0.43**</td>
<td></td>
</tr>
</tbody>
</table>

Multiple regression is required when relevant explanatory variables are correlated with each other.

1. Inclusion of price makes acceleration significant
2. R-square rises
3. Engine displacement’s relative effect rises
4. Price is now significant unlike the bivariate regression!

General Rule: Use multiple regression when a number of variables have an effect on a dependent variable and those variables are correlated with each other, such that leaving one of them out of the regression leads to the included variables’ coefficients reflecting some of the excluded variable’s influences.
Choose regression coefficients: \( \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \) such that the sum of squared residual errors is minimized:

\[
\text{i.e., minimize: } \sum_{i=1}^{n} e_i^2
\]

Find a [very] large number of students and assign them into homogeneous groups; that is, into groups where every student is the same as every other student with respect to GPA, high school quality, parental help, SES.

Then, in each group randomly assign one half the students to take the prep course and leave the other group without the prep course as a control group.

Test the students before and after the course to measure gains (remember the control group will, on average, see a gain in scores even without the prep course).

Do a simple regression to determine the value of \( \hat{\beta}_2 \).

By controlling the values of all other variables, we remove the correlation between the error term and prep course and the beta coefficient becomes unbiased:

\[
E[\hat{\beta}_2] = \beta_2
\]

Advantages of Multiple Regression:

- By explicitly including the other, relevant and correlated, explanatory variables, it enables us to control for them without requiring them to be exactly the same for any two people. That is, by explicitly taking account of the intercorrelations among the potential explanatory variables, multiple regression allow us to control for their influence without requiring them to take on only a restricted set of values.
- It allows for the simultaneous control of many variables even though no two people are exactly alike on all the variables.
- It allows us to generate a single estimate for the “effect” of each explanatory variable, which is analogous to the weighted average of effects in different subgroups.

Potential Problems in using Multiple Regression to “Control” other variables:

- We have to assume an explicit functional relationship among the variables -- it could be wrong.
- Unlike a randomized experiment, we have to be able to measure the explanatory variable to be able to include it in the statistical analysis. In addition, the variable must be measured well to avoid “measurement error.”
- We have to include all the relevant and correlated explanatory variables in order to avoid omitted variable bias. Randomization controls for all characteristics of the experimental subjects, regardless of whether those characteristics can be measured.

Problems with social experiments:

- Failure to randomize. It’s often difficult, or even impossible, to randomly assign observations into two groups. Difficult because our data are not often experimental, but are essentially made available to us without any possibility of arranging them into any sort of experimental design.
- Failure to follow the “treatment” protocol. Even if we can randomly assign students to the prep course -- or not -- we cannot make the students study or otherwise follow the training protocol. Also, we cannot stop control group subjects from taking the course on the sly.
- Attrition. Subjects drop out of experiments, and their tendency to drop out is not random. That is, non-random attrition may lead to correlation between errors and treatment variable (known as selection bias).
- Cost and small samples. Because experiments with humans are very expensive, the samples tend to be small, reducing the precision of our statistical estimation.

Explicitly taking account of the intercorrelations among the explanatory variables serves to “control” their effects, so that \( \hat{\beta}_2 \) represents the “true” (or unbiased) effect of the prep course on SAT score.
The Problem: Does performance on midterm exams predict performance on the final exam very well? Do different midterms have a different impact on final exam performance?  

The estimator of the causal effect should be unbiased and consistent. That is, a slope coefficient \( \hat{\beta} \) should be an unbiased and consistent estimator of the true population parameter \( \beta \). Hypothesis tests should have the desired significance level (the actual rejection rate of the test under the null hypothesis should equal its desired significance level).

So, for OLS regression internal validity requires that the OLS estimator is unbiased and consistent and that standard errors are computed in a way that makes confidence intervals have the desired confidence level.

**Definition:** A statistical analysis is **internally valid** if the statistical inferences about causal effects are valid for the population being studied. The analysis is **externally valid** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

Internal validity has two components:

The random errors, \( \varepsilon \), which do not depend on values of independent variables included in the model. The model is "well specified", i.e., all relevant (and correlated) explanatory variables are included in the model.

Then, it follows that:

\[
E[Y] = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k + \varepsilon
\]

\( \varepsilon \) is a random error, that, for any given set of values of independent variables \( X_1, X_2, \ldots, X_k \), is normally distributed with mean zero and variance equal to \( \sigma^2 \).

The random errors, \( \varepsilon_i \) and \( \varepsilon_j \), associated with any pair of \( y \)-values are independent of each other and independent of the values of \( x \)-variables included in the model.

The model is "well specified", i.e., all relevant (and correlated) explanatory variables are included in the model.

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon
\]

\( \beta \) is the true population parameter, \( \hat{\beta} \) is the unbiased and consistent estimator of \( \beta \).

The analysis is statistically valid if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

The analysis is **internally valid** if the OLS estimator is unbiased and consistent.

The analysis is **externally valid** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

A statistical analysis is **internally valid** if:

1. The estimator of the causal effect is unbiased and consistent.
2. Hypothesis tests have the desired significance level.
3. Confidence intervals have the desired confidence level.

The analysis is **externally valid** if:

1. The model is "well specified", i.e., all relevant (and correlated) explanatory variables are included in the model.
2. The random errors, \( \varepsilon_i \) and \( \varepsilon_j \), associated with any pair of \( y \)-values are independent of each other and independent of the values of \( x \)-variables included in the model.

The model is "well specified", i.e., all relevant (and correlated) explanatory variables are included in the model.
The probability of getting a residual this large is approximately one in ten thousand.
1. **metric (or quantitative) variable** is a variable that is measured on some well-defined scale. For variables measured on a particular scale, a change of a certain amount means the same thing no matter where one starts. e.g., growing one inch in height means the same thing whether or not one starts at 5'2" or 6'2".

2. This variable is an **ordinal variable**. Its values ascend or descend, but the distance between them is arbitrary and undefined. That is, 1<2<3<4, but we don’t know by how much.

3. **Categorical variables** are variables that do not have any order at all; that is, they are measured on **nominal scales** where each value represents a different category, but the categories themselves cannot be ranked. e.g., 1 = Male and 0 = Female for the variable "sex.”

A **dummy variable** is a categorical variable that can take on only one of two values, zero or one.

### Types of variables used in regression analysis:

**Variables**

1. `ln ln
   = Y Deep
   road
   lndroad`:

   - **Deep**:
   - **lnclor**:

   nearest salted road.

   natural logarithm of the chloride concentration (log milligrams per liter) in the well’s water — an indicator of contamination by road salt.

   shallow wells

   `ln ln
   = 1 = yes
   0 = no`

   `ln ln
   = Y Deep
   road`

   natural logarithm of the distance (log feet) between the well and the nearest salted road.

   `ln ln
   = Deep`

   `ln ln
   = Y Deep
   road`

   `ln ln
   = deeproad`

   dummy variable coded 0 for shallow wells and 1 for deeper wells drilled into bedrock.

   Study of Road Salt Contamination of New England Wells

   **Variables**

   1. `ln ln
      = ln ln
      = Y Deep
      road
      lndroad`

   inclor: natural logarithm of the chloride concentration (log milligrams per liter) in the well’s water — an indicator of contamination by road salt.

   lndroad:

   Deep:

   lnclor:

   well’s water -- an indicator of contamination by road salt.

   nearest salted road.

   `ln ln
   = Deep`

   `ln ln
   = Distance`

   `ln ln
   = 1 = yes
   0 = no`

   Interaction of Deep & Distance

   `ln ln
   = Y Deep
   Road
   lndroad`

   Log of Distance

   `ln ln
   = Y Deep
   Road`

   Log of Chloride Concentration

   `ln ln
   = ln ln
   = β + β • Deep + β • ln droad + ε`

   `ln ln
   = Deep`

   `ln ln
   = Y Deep
   Road`

   `ln ln
   = deeproad`

   `ln ln
   = β + β • deep + β • ln droad + β • deeproad + ε`

   `ln ln
   = β + β • Deep + β • ln droad + β • deep × ln droad`

   `ln ln
   = β + β • deep`

   `ln ln
   = β + β • deep`
### Chloride Concentrations and Distance*

<table>
<thead>
<tr>
<th>Distance (in feet)</th>
<th>Shallow Wells</th>
<th>Deep Wells</th>
</tr>
</thead>
<tbody>
<tr>
<td>(20 feet)</td>
<td>2.414 feet</td>
<td>2.414 feet</td>
</tr>
<tr>
<td>(10 feet)</td>
<td>2.414 feet</td>
<td>2.414 feet</td>
</tr>
</tbody>
</table>

*Closest well is 20 feet, furthest is 2,641 feet

---

### Chloride Concentrations and Distance*

<table>
<thead>
<tr>
<th>Natural Log of Chloride Concentration (mg per liter)</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000</td>
<td>2.059</td>
<td>4.01</td>
<td>6.96</td>
</tr>
<tr>
<td>2.0000</td>
<td>6.96</td>
<td>10.01</td>
<td>13.97</td>
</tr>
</tbody>
</table>

---

### Chloride Concentrations and Distance*

<table>
<thead>
<tr>
<th>Natural Log of Distance (in feet)</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000</td>
<td>4.689</td>
<td>6.15</td>
<td>7.64</td>
</tr>
<tr>
<td>2.0000</td>
<td>7.64</td>
<td>9.12</td>
<td>10.60</td>
</tr>
</tbody>
</table>

---

### Chloride Concentrations and Distance*

<table>
<thead>
<tr>
<th>Weight (lbs)</th>
<th>Price (dollars)</th>
<th>Year</th>
<th>Horsepower</th>
<th>Cylinder</th>
<th>Acceleration</th>
<th>MPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,500</td>
<td>15,000</td>
<td>1998</td>
<td>250</td>
<td>6</td>
<td>12.0</td>
<td>22</td>
</tr>
</tbody>
</table>

---

### Handout on using dummy variables to test interaction effects in regression analysis

#### No Interaction

| Coef. | Std. Err. | t     | P>|t| |
|-------|-----------|-------|-----|
| weight | 0.0167 | 0.0000745 | 3.83 | 0.000335 |
| price  | -0.285  | 0.0000745  | -3.83 | 0.000335 |
| year   | 0.285   | 0.0000745  | 3.83 | 0.000335 |

#### With Interaction

| Coef. | Std. Err. | t     | P>|t| |
|-------|-----------|-------|-----|
| weight | 0.0167 | 0.0000745 | 3.83 | 0.000335 |
| price  | -0.285  | 0.0000745  | -3.83 | 0.000335 |
| year   | 0.285   | 0.0000745  | 3.83 | 0.000335 |

---

### The marginal impact of a one pound change in a car’s weight depends on the horsepower of the engine in the car.

\[
\frac{\delta MPG}{\delta wgt} = \beta_w + \beta_{hp} \times \text{horsepower}.
\]
### Car Mileage: No Interaction Effect

| Variable | Coefficient | Std. Error | t-Statistic | P>|t| | 95% Confidence Interval |
|----------|-------------|------------|-------------|-----|--------------------------|
| cons     | 42.22624    | 22.0676    | 1.91        | 0.058 | -1.399893 - 85.85238     |
| wthp     | 0.0000745   | 0.0000212  | 3.51        | 0.001 | 0.0000325 - 0.0001165   |
| price    | 0.000649    | 0.0001878  | 3.45        | 0.001 | 0.0002776 - 0.0010203   |
| year     | 0.3410433   | 0.2613443  | 1.30        | 0.0194| -0.1756165 - 0.8577031  |
| horsepow | -0.2849999  | 0.0729368  | -3.91       | 0.000 | -0.429191 - 0.1408088   |
| cylinder | 0.0609603   | 0.603188   | 0.10        | 0.12 | -1.131501 - 1.253422    |
| weight   | -0.0167031  | 0.0021642  | -7.72       | 0.000 | -0.0209816 - 0.0124246  |

### Car Mileage: With Interaction Effect

| Variable | Coefficient | Std. Error | t-Statistic | P>|t| | 95% Confidence Interval |
|----------|-------------|------------|-------------|-----|--------------------------|
| cons     | 42.22624    | 22.0676    | 1.91        | 0.058 | -1.399893 - 85.85238     |
| wthp     | 0.0000745   | 0.0000212  | 3.51        | 0.001 | 0.0000325 - 0.0001165   |
| price    | 0.000649    | 0.0001878  | 3.45        | 0.001 | 0.0002776 - 0.0010203   |
| year     | 0.2809669   | 0.2709474  | 1.04        | 0.302 | -0.2546449 - 0.8165786  |
| accel    | 0.431885    | 0.193399   | 2.23        | 0.027 | 0.0495716 - 0.8141984   |
| horsepow | -0.0554725  | 0.166672   | -1.66       | 0.10  | -0.3822891 - 0.1729181  |
| cylinder | 0.0609603   | 0.603188   | 0.10        | 0.12 | -1.131501 - 1.253422    |
| weight   | -0.0123731  | 0.0018463  | -6.70       | 0.000 | -0.0160229 - 0.0087234  |

### Handout on using metric variables in testing for interaction effects in multiple regression

The Question:
When comparing the total variance explained by the "base model" containing only the three significant variables to the variance explained by the "expanded model" which adds the insignificant variables, is the addition to explained variation large enough to say that this extra set of variables together has significantly raised total explained variance?

\[ H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \]

\[ H_a : \text{at least one of these } \beta \text{ not equal to zero} \]

\[ F_{n,k} = \frac{(RSS(K-H) - RSS(K))}{H} / \frac{RSS(K)}{n-K} \]

Residual Sum of Squares (RSS) is the sum of squared deviations of the residuals around their mean value of zero:

\[ RSS(K) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} \beta_k x_{ik})^2 \]

Remember, it's RSS that least squares regression seeks to minimize.

### Automobile Mileage Data: Residual Plots

quietly regress mpg weight cylinder displac e horsepow accel year
predict mpghat
generate resid = mpg - mpghat
graph twoway scatter resid mpghat, yline(0) ti("Plot of Mileage Residuals vs. Predicted Values")

plot of mileage residuals vs. predicted values

grid lines are 5, 10, 25, 50, 75, 90, and 95 percentiles

### Handout on using metric variables in testing for interaction effects in multiple regression

\[ H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \]

\[ H_a : \text{at least one of these } \beta \text{ not equal to zero} \]

\[ F_{n,k} = \frac{(RSS(K-H) - RSS(K))}{H} / \frac{RSS(K)}{n-K} \]

Residual Sum of Squares (RSS) is the sum of squared deviations of the residuals around their mean value of zero:

\[ RSS(K) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} \beta_k x_{ik})^2 \]

Remember, it's RSS that least squares regression seeks to minimize.

quietly regress mpg weight cylinder displace horsepow accel year

\[ H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \]

\[ H_a : \text{at least one of these } \beta \text{ not equal to zero} \]

\[ F_{n,k} = \frac{(RSS(K-H) - RSS(K))}{H} / \frac{RSS(K)}{n-K} \]

Residual Sum of Squares (RSS) is the sum of squared deviations of the residuals around their mean value of zero:

\[ RSS(K) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} \beta_k x_{ik})^2 \]

Remember, it's RSS that least squares regression seeks to minimize.

quietly regress mpg weight cylinder displac e horsepow accel year

\[ H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \]

\[ H_a : \text{at least one of these } \beta \text{ not equal to zero} \]

\[ F_{n,k} = \frac{(RSS(K-H) - RSS(K))}{H} / \frac{RSS(K)}{n-K} \]

Residual Sum of Squares (RSS) is the sum of squared deviations of the residuals around their mean value of zero:

\[ RSS(K) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} \beta_k x_{ik})^2 \]

Remember, it's RSS that least squares regression seeks to minimize.
Symmetry Plot for Regression Residuals

Residual Plots for mileage data regression

Nonnormality of the regression error term:
- We lose the justification for applying the t and the F distributions, especially with small samples.
- Since Ordinary Least Squares tends to be sensitive to outliers, heavy-tailed distributions can cause great sample-to-sample variation. I.e., our results from one random sample out of a population might not look very much like results from other samples.

Some possible solutions to the nonnormality of residuals:
1. Transform the dependent variable.
2. Transform one (or more) independent variables.
Possible transformations include: log transformations or power transformations.

Are there ways to deal with heteroscedasticity?
- Stata can provide “heteroscedasticity robust” statistics after OLS estimation (see Hamilton “Robust Regression”), or
- We can use Stata’s “weighted least squares” to produce asymptotically unbiased estimates of standard errors. (See Hamilton)

Why is this a problem?
- Our initial assumption is that the variance of epsilon, conditional on the explanatory variables, is the same for all combinations is independent of the values of any of the independent variables. If this is incorrect, then the model exhibits heteroscedasticity.
- Heteroscedasticity is present, the coefficient estimates are unbiased, but the estimated standard errors of the estimated coefficients are biased.
- Consequently, t statistics, test statistics, F statistics, and the coefficients of determination (R-squared) will be incorrect, leading to incorrect inferences about our estimates.

\[ E[\epsilon_i | X] = E[\epsilon_j | X] = 0 \text{ for all } i \]
\[ E[\epsilon_i \epsilon_j] = \begin{cases} 0 & \text{if } i = j, i, j = 1, 2, \ldots, n \\ \sigma^2 & \text{if } i \neq j, i, j = 1, 2, \ldots, n \end{cases} \]

Constant sigma implies homoscedasticity
Changing sigma implies heteroscedasticity
Influential Observations

Omitted Variable Bias -- Specification Error
The solution: Include all relevant explanatory variables. For this you need a strong theory of the causal process that you are trying to explain, or, lacking the appropriate variables, you need to run a well-controlled experiment.

Functional Form -- Misspecification

Career Statistics for 353 Major League Baseball Players: 1993 Season
Goal: Predict determinants of their 1993 Salaries

The Model:

\[ \text{lsalary} = \beta_0 + \beta_1 \text{years} + \beta_2 \text{gamesyr} + \beta_3 \text{bavg} + \beta_4 \text{hrunsyr} + \beta_5 \text{rbisy} + \epsilon \]

I'm using the logarithm of salary because of the wide variation in player salaries (Range: $109,000 to $6,329,213)

As a group the three variables are significant!

Multicollinearity refers to a situation in which there is a strong linear relationship among two or more independent variables in a multiple regression. In the more extreme cases, this linear correlation is so strong that the contribution of individual variables to the regression cannot be adequately ascertained.

\[ \text{Multicollinearity} = \frac{s^2}{(n-1)\text{var}(X_j) - R_j^2} \]

where
- \( s^2 \) is the variance of the error term in the regression.
- \( \text{var}(X_j) \) is the variance of the independent variable \( j \).
- \( R_j^2 \) is the multiple \( R^2 \) for the regression of \( X_j \) on the other covariates

As \( R_j^2 \) rises from 0 to one, the VIF approaches infinity.
```plaintext
**Dealing with Multicollinearity**

- Use Stata's post estimation command *estat vif* to see if any coefficients show signs of trouble. E.g., for our baseball salary model we have:

  ```plaintext
  . quietly regress lsalary rbisyr hrunsyr gamesyr years bavg
  . estat vif
  ```

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
<th>1/VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>rbisyr</td>
<td>17.89</td>
<td>0.055901</td>
</tr>
<tr>
<td>hrunsyr</td>
<td>7.94</td>
<td>0.125913</td>
</tr>
<tr>
<td>gamesyr</td>
<td>6.10</td>
<td>0.163982</td>
</tr>
<tr>
<td>years</td>
<td>1.47</td>
<td>0.678753</td>
</tr>
<tr>
<td>bavg</td>
<td>1.20</td>
<td>0.834258</td>
</tr>
<tr>
<td>_cons</td>
<td>6.92</td>
<td></td>
</tr>
</tbody>
</table>
  
  Mean VIF = 6.92

- Experiment with adding and deleting the suspect variables. Do standard errors or coefficient estimates change substantially?

- Employ one of the following coping strategies:
  - Keep the variables in the equation, but understand that we cannot generalize (beyond the sample) about their separate effects.
  - Drop one or more of the offending variables, since their information is mostly redundant.
  - Combine the variables, since there is evidence that they are measuring the same thing.
  - Collect more data. Multicollinearity is basically a problem of not enough information. Adding cases generally makes the coefficient estimates more precise, by [1] increasing n in the denominator of the formula, and [2] by increasing the variance of the independent variables.

- Larger n increases denominator \( \frac{1}{n-1} \text{var}(X) \) likely to rise also.