ENGR 200 – Day 1: What You Can Do With Data in Excel

Real questions engineers and athletes ask:

• Engineering: “How long can this fan run on a battery?” “What study habits predict passing Statics?”

• Athletics: “Why does my coach tell me to bench 115 lb — not 95, not 135? Where did that number come from?”

Statistics gives you a playbook to answer these: collect relevant data, summarize it, build simple models, and make predictions.

In this handout we’ll use a real dataset (SAT scores, study hours, attendance, participation, and Statics grade) to learn EXACTLY what to click in Excel — no prior knowledge required.

Class Roadmap for Today

1) Set up Excel for analysis (turn on the Data Analysis ToolPak).

2) Open the class dataset and make it easy to work with (freeze panes, format as a table).

3) Summarize variables with one-click descriptive statistics (mean, median, stdev, min, max).

4) Make quick graphs (scatterplot + trendline + R^2).

5) Run a multiple linear regression (predict Statics_Grade from SAT scores, study hours, attendance, participation).

6) Read the output (R^2, coefficients, p-values) and write a one-sentence takeaway.

7) Mini assignment: you do it once yourself with slightly different X’s.

Good Data: Getting the Right Stuff

Before you click anything, make sure the data are worth trusting.

• Representativeness: your sample should look like the population you care about (not only honors students, not only athletes, etc.).

• Clear definitions: decide exactly how each variable is measured (e.g., Study_Hours_per_week Attendance_%). Use consistent units.

• Enough observations: more rows → more reliable results. Our file has 150 students.

• No cheating with the target: don’t use information you wouldn’t have at prediction time.

• Privacy: remove names/IDs when you share results outside class.

Turn On Excel’s Data Analysis ToolPak

Windows: File → Options → Add-ins → at bottom choose “Excel Add-ins” → Go… → check “Analys
ToolPak” → OK.

Mac: Tools → Excel Add-ins… → check “Analysis ToolPak” → OK.

You will now see a new button: Data → Data Analysis. That’s where Regression lives.

Open the Dataset, Freeze the Header, Make It a Table

1) Download the Excel file to your computer and open it in Excel.

2) Freeze header row: View → Freeze Panes → Freeze Top Row.

3) Turn the range into a Table: select any cell in the data → Insert → Table → “My table has headers.”

Why? Tables auto-expand when you add rows and make formulas easier to read.

4) Optional: rename the sheet to something short like GPA_Data.

Fast Descriptive Statistics (Direct Formulas)

Pick an empty area of the sheet and type these exactly (adjust column letters to your sheet):

- Mean of Statics_Grade: =AVERAGE(H:H)
- Median: =MEDIAN(H:H)
- Standard deviation: =STDEV.S(H:H)
- Minimum and Maximum: =MIN(H:H) and =MAX(H:H)
- Correlation between SAT_Verbal and Statics_Grade: =CORREL(D:D,H:H)

Tip: Use one variable per column. Don’t mix numbers with text in the same column.

Make a Scatterplot + Trendline + R^2

Goal: visualize SAT_Verbal vs Statics_Grade.

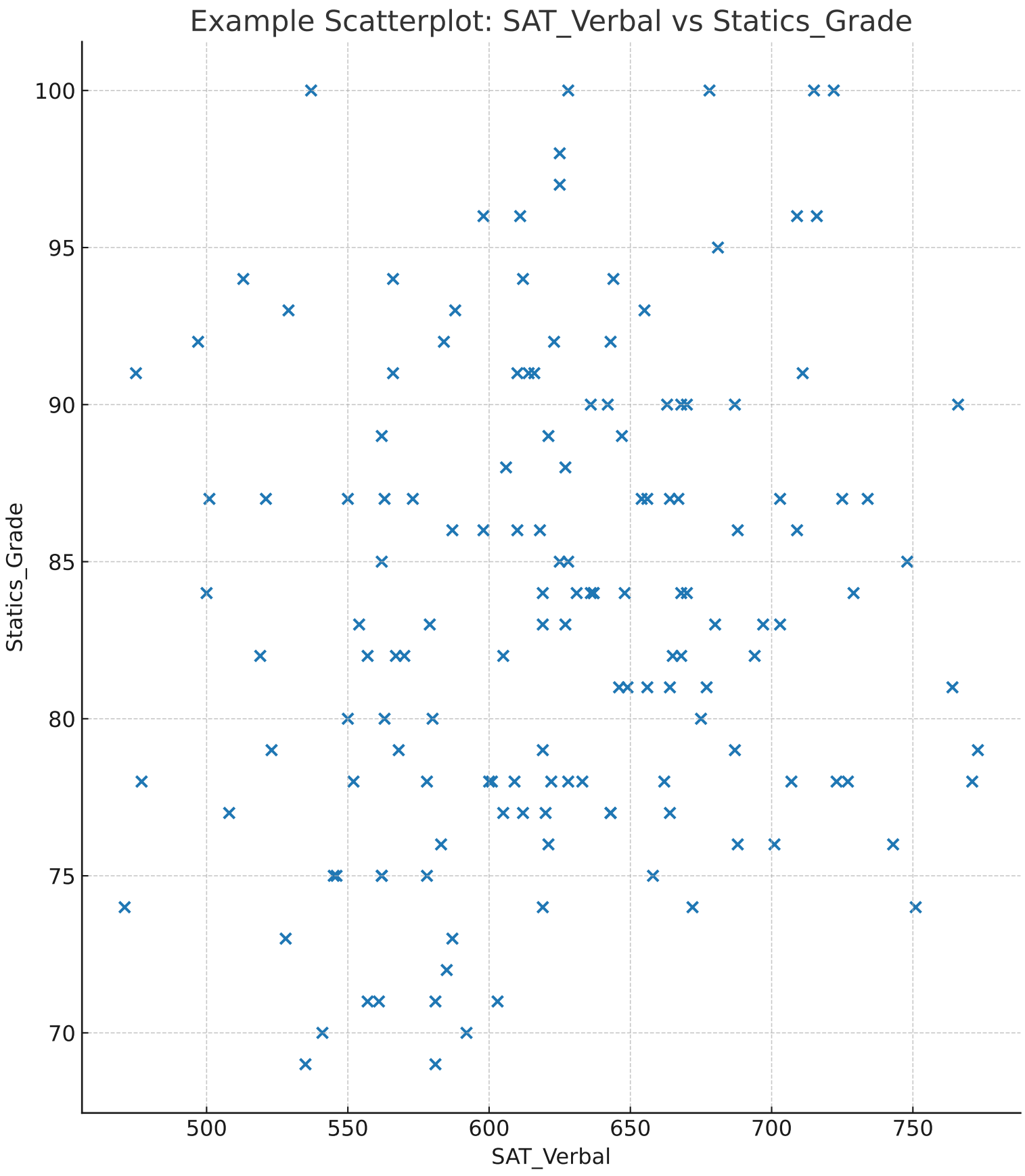
1) Select the two columns D (SAT_Verbal) and H (Statics_Grade) including headers.

2) Insert → Scatter → first option (markers only).

3) Click a point → right-click → Add Trendline → Linear → check “Display Equation on chart” and “Display R-squared value on chart”.

4) Read the equation ($y = a + bx$). If b is positive, higher SAT_Verbal tends to higher Statics_Grade.

5) R^2 shows how much of the variation in grades is explained by the line (0 to 1).


Example Scatterplot: SAT_Verbal vs Statics_Grade
Statics_Grade
100
95
90
85
80
75
70
500
550
600
650
700
750
SAT_Verbal

Run a Multiple Regression in Excel (ToolPak)

Goal: Predict Statics_Grade (Y) from SAT_Math, SAT_Verbal, Study_Hours_per_week, Attendanc
Class_Participation (X’s).

1) Data → Data Analysis → Regression → OK.

2) Input Y Range: select the Statics_Grade column including header.

3) Input X Range: select the five predictor columns including headers (they must be side-by-side).

4) Check “Labels.” Choose an Output Range (e.g., a blank area of the sheet).

5) Click OK. Excel prints Regression Statistics, ANOVA, and Coefficients tables.

What to read first:

• R Square: how much variation is explained (0–1).

• P-values: < 0.05 suggests the coefficient is statistically different from 0.

• Coefficient sign: + means higher X → higher Y; − means higher X → lower Y.

Reading the Output — What to Say in Plain English

Example with this class dataset (computed here for illustration — your Excel output will look similar):

• $R^2 \approx 0.051$ (about 5.1% of variation explained).

• Coefficients (intercept and slopes):

- Intercept: 68.9827

- SAT_Math: -0.0111

- SAT_Verbal: 0.0160

- Study_Hours_per_week: -0.2226

- Attendance_%: 0.1311

- Class_Participation: 0.3829

Interpretation example: holding other variables constant, a 10-point increase in SAT_Verbal is associated with about 0.16 points change in Statics_Grade (per this simple model).

Caution: Correlation is not causation. Use this as a prediction/exploration tool, not proof of cause.

Do-It-Now Mini Assignment (10–15 minutes)

1) Make a scatterplot of HS_GPA vs Statics_Grade with a linear trendline and R^2.

2) Run a regression with Y = Statics_Grade and X’s = HS_GPA, Attendance_%.

3) Write one sentence: which predictor seems most useful and why? Use sign and p-value.

4) Bonus: replace HS_GPA with SAT_Verbal — does R^2 go up or down?

Troubleshooting (Read Me If Something Breaks)

• “Regression” not in Data → Data Analysis: turn on the Analysis ToolPak (see earlier page).

• Ranges don’t match error: make sure Y Range has the same number of rows as X Range.

• P-values are blank: you may have selected non-numeric cells; clean your columns.

• Trendline options missing: click the dots on the chart first, then right-click → Add Trendline.

• Weird results: check that there are no text strings in numeric columns; remove missing values.

Glossary (One-Line Each)

- Variable: a column of numbers you measure (e.g., SAT_Math).
- Dependent variable (Y): what you want to predict (e.g., Statics_Grade).
- Independent variable (X): what you use to predict Y.
- Mean/Median: average/middle value.
- Standard deviation: how spread out the numbers are.
- Scatterplot: a picture of how two variables move together.
- Regression: a math way to draw the best line/plane through data to predict Y from X's.
- R^2: 0–1 measure of how well the model fits (higher is better).
- p-value: the chance you'd see a coefficient this far from zero if the true effect were zero.

Data Dictionary (What Each Column Is)

Student_ID — int64

HS_GPA — float64

SAT_Math — int64

SAT_Verbal — int64

Study_Hours_per_week — float64

Attendance_% — float64

Class_Participation — float64

Statics_Grade — int64

ENGR 200 - Day 1 Excel Starter Guide

Prepared by Instructor

In this first class, we explored how engineers, athletes, and other professionals can use data to answer practical questions. For example, determining how long a fan can run, or why a coach sets a specific weight target for training. We saw that statistics can help find answers — but only if the data is representative and collected correctly.

We then used Excel to load the ENGR 200 Term Project dataset, enabled the Data Analysis ToolPak, and learned how to run a regression analysis to explore relationships between variables. Along the way, we discussed how to interpret the output and connect it back to real-world questions.

Term Project Connection

This dataset is not just for practice — it's the same one you will use for your ENGR 200 Term Project. Don't worry if you've never done this before. The steps you just learned are the foundation. Once you work through the full process, you'll understand both the how and the why behind the results.

Think of this as learning to drive in an empty parking lot before heading onto the highway — the project is the highway, but you've already learned how to steer, brake, and accelerate.

By the end of this, you'll not only complete the term project successfully, but you'll also be able to look at real-world problems and say: "I can find the answer with data — and I know exactly how to do it."

ENGR 200 – Day 1 Excel Regression (Detailed Guide)

In this exercise, we will learn how to run a regression analysis in Excel using the ENGR200 dataset. We will focus on two predictors (Attendance and Class Participation) and the outcome (Statics Grade). The purpose is to give you step-by-step practice with Excel's Data Analysis ToolPak and help you learn how to interpret the results in plain English.

Steps to Run Regression in Excel

1. Open the dataset (ENGR200_Term_Project_Data.xlsx) in Excel. 2. Go to File → Options → Add-ins → Excel Add-ins → Go → check Analysis ToolPak → OK. You should now see 'Data Analysis' on the Data tab. 3. Select Data → Data Analysis → Regression → OK. 4. Input Y Range: select the Statics_Grade column (include header). 5. Input X Range: select the Attendance_% and Class_Participation columns (side by side, include headers). 6. Check the 'Labels' box since we included headers. 7. Choose an output range on the same sheet (blank area). Click OK.

Understanding the Output

Excel produces three main sections in the regression output: 1. Regression Statistics - R Square: how much of the grade variation is explained. In our example, $R^2 \approx 0.012$, meaning only about 1% is explained by Attendance and Participation. - Multiple R: correlation between actual and predicted grades. Low in this case. - Standard Error: typical prediction error, about 7.4 points. 2. ANOVA Table - Significance F ≈ 0.42. Since this is much greater than 0.05, the model is not statistically significant. 3. Coefficients Table - Intercept ≈ 70. This is the baseline grade if Attendance and Participation were 0. - Attendance ≈ +0.13 (p = 0.27). A 1% increase in attendance predicts a 0.13 point increase in grade, but the p-value shows it is not significant. - Class Participation ≈ +0.24 (p = 0.49). A 1-unit increase in participation predicts a 0.24 point increase in grade, also not significant.

Variable	Coefficient	p-value	Interpretation
Intercept	70.08	<0.001	Baseline grade when Attendance and Participation = 0
Attendance_%	+0.13	0.27	Slight positive effect, not significant
Class Participation	+0.24	0.49	Slight positive effect, not significant

Plain English Takeaway:
In this dataset, attendance and participation alone do not strongly predict Statics grades. The regression explains very little variation (about 1%), and neither coefficient is statistically significant. This shows why engineers use multiple predictors (GPA, SAT, study hours, etc.) for stronger models. Still, this practice run teaches you the mechanics of Excel regression and how to read the output.

ENGR 200 – Worked Example: Two-Predictor Regression in Excel

Example variables: HS_GPA and SAT_Verbal → predict Statics_Grade. This example shows how to run a two-predictor regression in Excel's Data Analysis ToolPak and how to read the output. We use the class dataset (150 students). The steps and interpretation below are written so a first-time Excel user can follow.

A) Step-by-Step: Run the Model in Excel

• Open the dataset in Excel. Make sure each variable is a single clean numeric column.
• Turn on the ToolPak: File → Options → Add-ins → Excel Add-ins → Go → check Analysis ToolPak → OK (Mac: Tools → Excel Add-ins).
• Go to Data → Data Analysis → Regression → OK.
• Input Y Range: select the Statics_Grade column (include the header).
• Input X Range: select HS_GPA and SAT_Verbal (place them side-by-side, include headers).
• Check Labels. Choose an Output Range on the same sheet (blank area). Click OK.

B) What You Should See (Numbers Rounded)

Regression Statistics
R Square ≈ 0.0218 (≈2.2%); Std. Error ≈ 7.40; N = 150
ANOVA
F(2,147) ≈ 1.64; Significance F ≈ 0.197 (model NOT significant)
Coefficients
Intercept ≈ 70.43; HS_GPA ≈ +1.05 (p ≈ 0.52); SAT_Verbal ≈ +0.016 (p ≈ 0.088)

C) How to Interpret (Plain English)

• Overall model: F-test p ≈ 0.197 > 0.05 → the model is NOT statistically significant. Together, HS_GPA and SAT_Verbal do not explain much of the grade variation.
• Fit: R^2 ≈ 0.022 → about 2% of the variation in Statics_Grade is explained (very weak). Most variation is due to other factors not included here.
• HS_GPA: coefficient ≈ +1.05, p ≈ 0.52 → no reliable evidence of an effect (not significant).
• SAT_Verbal: coefficient ≈ +0.016, p ≈ 0.088 → suggests a small positive effect, but it is only marginal and not below the usual 0.05 cutoff.
• One-sentence takeaway: This two-predictor model is weak and not significant; use stronger predictors (e.g., study hours, attendance, or combine more variables) for better performance.

D) If You Want to Pick ANY Two Predictors

Choose your Y (what you want to predict) and two X's (predictors) in adjacent columns. Data → Data Analysis → Regression. Set Input Y Range to the Y column; Input X Range to the two X columns; check Labels. Click OK. Read three things first:
• F and Significance F (ANOVA): if p < 0.05, the model is significant.
• R^2: how much variation is explained (higher is better).
• Coefficients & p-values: signs (+/–) tell direction; p < 0.05 suggests a reliable effect.
Write your conclusion: "Using X1 and X2 to predict Y, the overall model is/ isn't significant (F, p). R^2 is ___. X1 has a (positive/negative) effect of __ (p = __). X2 has __ (p = __)."