# Recap: Linear Regression

DS701 Session 12 — Wed Oct 14, 2026

# Today’s plan

# Knowledge-Check Review

# KC 1: The objective and the equations

*In one or two sentences: what does the least-squares line minimize, and what are the* ***normal equations*** *that its coefficients* $\hat{\beta}$ *satisfy? (Write the equations; no derivation needed.)*

# KC 2: Why $R^2$ only ever goes up

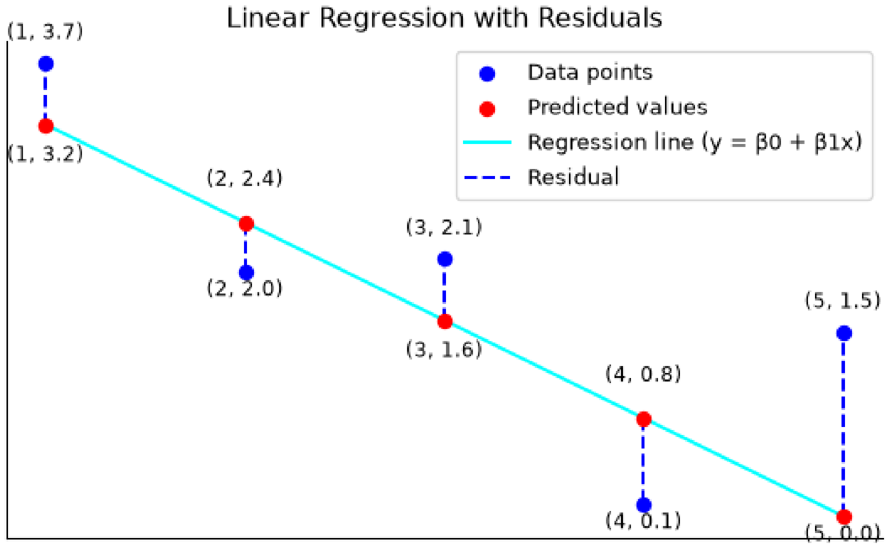
*$R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$ is "the fraction of the variance of $y$ explained by the model." Explain why $R^2$ can never go **down** when you add another feature (column) to $X$ and refit — and why that makes a high in-sample $R^2$ weak evidence on its own that the model is good.*

# KC 3: “Linear” means linear in $\beta$

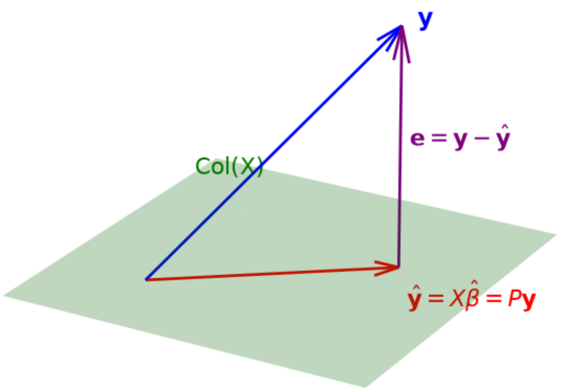
*A colleague says “linear regression can only fit straight lines, so it is useless for my data, which clearly curves.” Using the lecture’s fits of a parabola and of $y = \beta_0 + \beta_1 \log x$, explain what “linear” in* linear model* actually refers to, and what changes (and what does not) in the least-squares procedure when you fit a curve.*

# Highlights

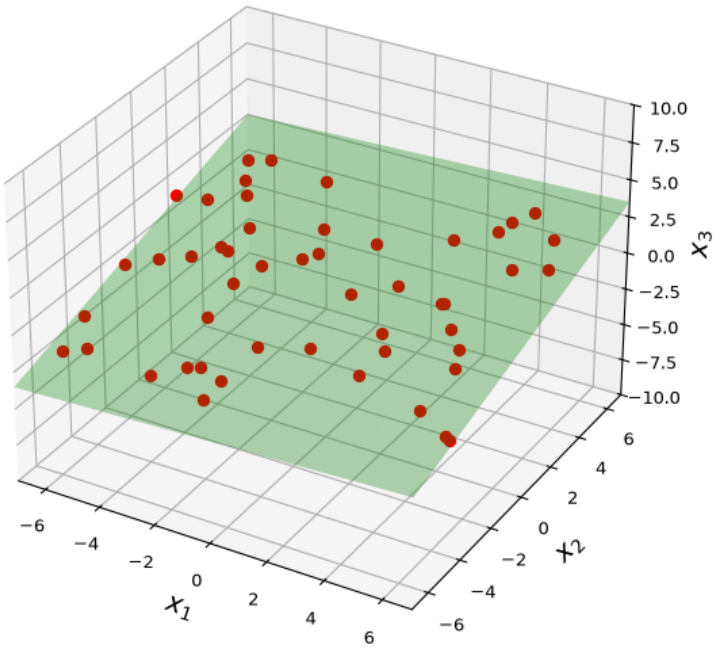
# The model, the residual, the objective



# Least squares is a projection



# Same equations, any design matrix




Once $X$ is built, the procedure never changes:

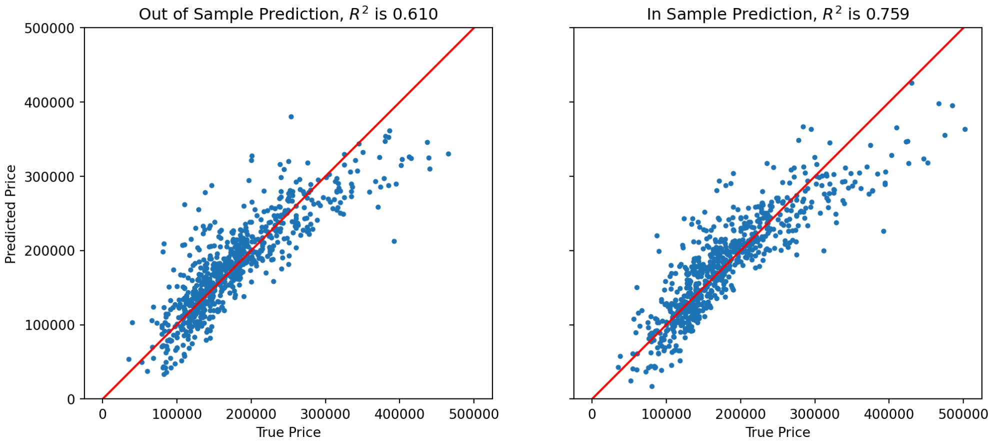
| **model** | **columns of $X$** |
|---|---|
| line | $\mathbf{1},\ x$ |
| parabola | $\mathbf{1},\ x,\ x^2$ |
| log | $\mathbf{1},\ \log x$ |
| plane / multiple regression | $\mathbf{1},\ u,\ v,\ \ldots$ |
| trend surface | $\mathbf{1},\ u,\ v,\ u^2,\ uv,\ v^2$ |

# Reading the answer: coefficients

From the Ames model (seven features, sale price in dollars):

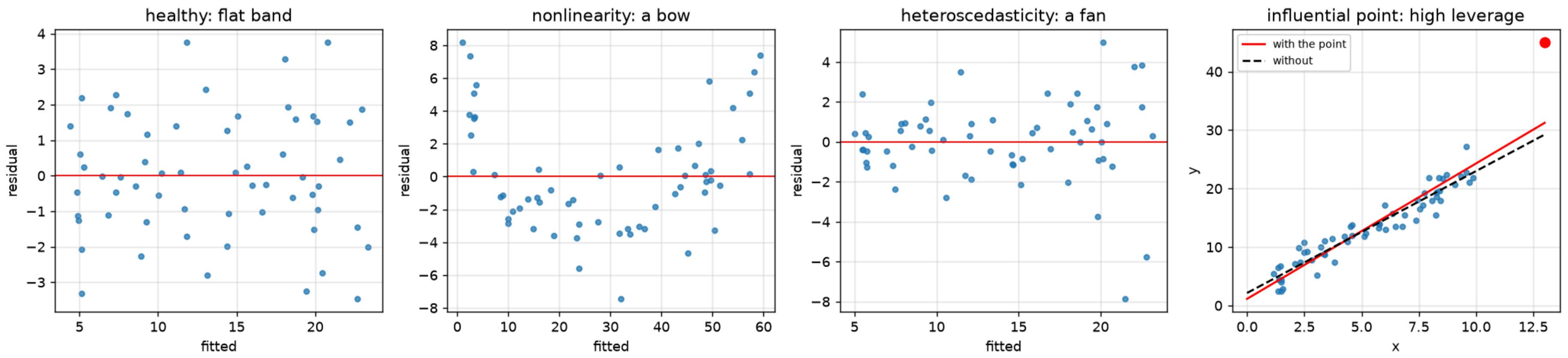
| **feature** | $\hat{\beta}$ | **reads as** |
|---|---|---|
| gross living area | \$48 / sq ft | one more square foot, **other features held fixed** |
| full bath | \$14,900 | one more full bath, others fixed |
| garage area | \$99 / sq ft | |
| lot area | \$0.23 / sq ft | **CI includes 0** — not distinguishable from no effect |

# Reading the answer: $R^2$ — and its limits




$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \in [0, 1]$$

# The diagnostic habit: plot the residuals



# In-Class Activity

# Activity: least squares by hand, then read the residuals

Goal: on a real dataset (200 diabetes patients — your section's own subset), build $X$ with an intercept, solve the normal equations, verify against `lstsq` and scikit-learn, compute $R^2$; form the projection matrix and check the residual is orthogonal to the columns; add a correlated feature and watch the coefficients move. Then the stretch: a `diagnostics` function, a mystery dataset with a hidden problem, and the fix.

- Work in groups of 2–3. Open **your section's** notebook.
- Parts marked **(autograded)** are submitted to Gradescope; the rest is participation.
- Staff will circulate — be ready to explain any part of your work.
- Dependencies: `numpy`, `pandas`, `matplotlib`, `scipy`, `scikit-learn`.

### Section A1

CO Open in Colab | Open on GitHub

### Section B1

CO Open in Colab | Open on GitHub

# Going deeper

Full lecture notes: Linear Regression

- Least-Squares Lines — residuals and the objective
- A least-squares problem — $X\beta = \mathbf{y}$, the normal equations, when $X^\top X$ is invertible
- Least-Squares Fitting of Other Models — parabola and log fits; linear in $\beta$
- Multiple Regression — the least-squares plane
- Measuring the fit: $R^2$ and TSS = RSS + ESS — the orthogonality behind the identity
- Aside: Summary Statistics — reading a statsmodels table
- House Prices in Ames, Iowa — coefficients in dollars, in- vs out-of-sample $R^2$
- **Next lecture:** when $y$ is a category (logistic regression), and when $\beta$ overfits (regularization)