Practice Quiz 2: Regression

MSE 125 — Lectures 4–5

Use this practice quiz to prepare for Quiz 2 (Wednesday, April 15). The real quiz will have 2 questions in 10 minutes, closed-book. This practice set has 10 questions covering Lectures 4–5: regression, span, orthogonality, the meaning of regression coefficients, polynomial overfitting, dummy variables, log transforms, and regression to the mean.

Every concept tested on the real quiz appears somewhere on this practice set, with a different scenario.

Question 1. A regression of home prices on square footage and home type gives:

Predictor	Coefficient
Intercept	\$425,000
sqft	\$210
condo	−\$85,000
apartment	−\$120,000

`condo` and `apartment` are 0/1 indicator columns; **single-family house** is the reference level.

(a) Interpret the coefficient on `condo` in one sentence.

(b) What does the intercept represent? Is it directly meaningful?

(c) Using this model, compute the predicted price of a **1,000 sqft condo**. Show your arithmetic.

Solution

(a) Holding square footage constant, a condo is associated with a price about \$85,000 lower than a single-family house with the same square footage.

(b) The intercept is the predicted price of a single-family house (the reference level) with 0 square feet. It is the baseline that the dummy coefficients adjust from — not directly meaningful (no real house has 0 sqft) but a necessary mathematical anchor.

(c) $\widehat{\text{price}} = 425{,}000 + 210(1{,}000) - 85{,}000 = 425{,}000 + 210{,}000 - 85{,}000 = \$550{,}000$. (Condo dummy = 1, apartment dummy = 0.)

Question 2. A regression of college GPA on hours studied per week gives a coefficient of +0.12/hour. Adding SAT score as a second feature changes the hours-studied coefficient to +0.04/hour. Explain the change in two sentences. Use the word *confounding*.

Solution

Hours studied and SAT scores are positively correlated (students with higher SAT scores tend to study more, or have habits that look like both). In the simple regression, the hours-studied coefficient was confounded — it credited study time with GPA gains that were partly associated with SAT score. Adding SAT to the model splits the credit, so the hours-studied coefficient now measures the association with GPA *holding SAT constant*, which is smaller.

Question 3. A student fits polynomials of increasing degree to a small dataset and reports the R^2 values:

Degree	1	2	3	4	5	6
R^2	0.42	0.56	0.58	0.59	0.59	0.60

A classmate argues that degree 6 is best because R^2 is highest. Do you agree? Why or why not?

Solution

Not really. R^2 mechanically increases (or stays the same) every time you add a feature, even features that are pure noise — adding columns can never make the fit worse. The big jump is from degree 1 to degree 2 (+0.14). After that, each extra degree adds 0.01 or less, which is not strong evidence of real structure being captured — it could easily be the model fitting random fluctuations. Degree 2 (or maybe 3) is a more honest choice.

Question 4. A school district finds that the elementary school with the highest test scores last year dropped to 3rd place this year, and the school that ranked 20th last year jumped to 5th. The local newspaper runs a headline: *"What is the top school doing wrong, and what is the bottom school doing right?"* What is the most likely statistical explanation for both changes? Answer in two or three sentences.

Solution

Regression to the mean. Test scores fluctuate year to year due to both stable school quality and random factors (which kids took the test, test-day conditions, particular questions). The school that ranked #1 was probably good *and* lucky; without the luck, it drops. The school that ranked #20 was probably mediocre *and* unlucky; without the bad luck, it rises. Both changes are predicted by statistics alone — no change in school quality is required. The newspaper headline is almost certainly chasing noise.

Question 5. After fitting a simple regression $\hat{y} = \beta_0 + \beta_1 x_1$ on real data, you have a residual vector $\epsilon = y - \hat{y}$. By the orthogonality condition, $x_1 \cdot \epsilon = 0$.

A colleague proposes adding $x_2 = 2x_1$ as a new feature.

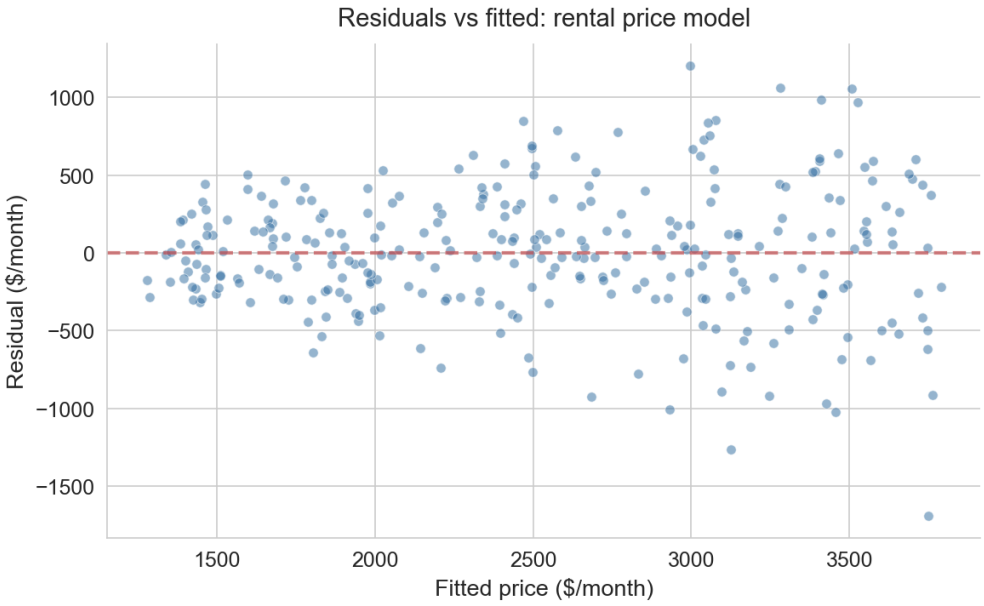
(a) Without computing anything new, what is the value of $x_2 \cdot \epsilon$? Show your reasoning in one line.

(b) Will adding x_2 to the model improve R^2? Connect your answer to the concept of linear dependence.

Solution

(a) $x_2 \cdot \epsilon = (2x_1) \cdot \epsilon = 2\,(x_1 \cdot \epsilon) = 2 \cdot 0 = 0$. Inner products are linear, so scaling a vector scales the inner product by the same factor.
(b) No. $x_2 = 2x_1$ is a scalar multiple of x_1, so $\{\mathbf{1}, x_1, x_2\}$ are linearly dependent — x_2 adds no new direction to the span. The set of reachable predictions does not grow, so $\hat{y}$ cannot move and R^2 is unchanged. The fact that $x_2 \cdot \epsilon = 0$ is automatic, not coincidence: any vector already in the span of the existing features is orthogonal to the residual.

Question 6. The plot below shows residuals vs. fitted values for a regression of monthly rental prices on apartment square footage.



(a) What pattern do you see in the residuals?

(b) What does the pattern tell you about the model's errors, and what is one fix you could try?

Solution

(a) The residuals fan out — they are tightly clustered near zero for low fitted prices and spread out widely (both above and below zero) for higher fitted prices.

(b) This pattern is **heteroscedasticity**: the error variance is not constant across fitted values. The model is more wrong, in absolute dollar terms, for expensive apartments than for cheap ones. A common fix is to log-transform the response (use log(price) instead of price), which compresses the right tail and stabilizes the variance.

Question 7. A data scientist one-hot encodes the categorical variable `body_style` (with three levels: sedan, SUV, truck) as **three** columns: `is_sedan`, `is_SUV`, `is_truck`. He includes all three in the regression along with the intercept. The regression software returns an error or produces nonsensical coefficients. Why?

Solution

For every car, exactly one of `is_sedan`, `is_SUV`, `is_truck` equals 1, so the three columns sum to the all-ones column — the same column the intercept already contributes. The four columns $\{\mathbf{1}, \text{is\_sedan}, \text{is\_SUV}, \text{is\_truck}\}$ are **linearly dependent**: one is a linear combination of the others, e.g., $\text{is\_truck} = \mathbf{1} - \text{is\_sedan} - \text{is\_SUV}$. The matrix X^TX becomes singular and the normal equations have no unique solution. The fix: drop one dummy (the *reference level*), leaving three independent columns.

Question 8. A study of post-graduation earnings regresses log(income) on years of post-secondary education and reports a coefficient of **0.08**. How should this coefficient be interpreted? Give two equivalent interpretations: one in terms of percentage change, and one in terms of multiplication.

Solution

Percentage: each additional year of education is associated with roughly an 8% higher income.
Multiplicative: each additional year of education multiplies the predicted income by $e^{0.08} \approx 1.083$.
Both say the same thing on the original scale: a year of education corresponds to a fixed *fraction* of existing income, so the dollar increase is larger for someone earning \$100,000 than for someone earning \$30,000.

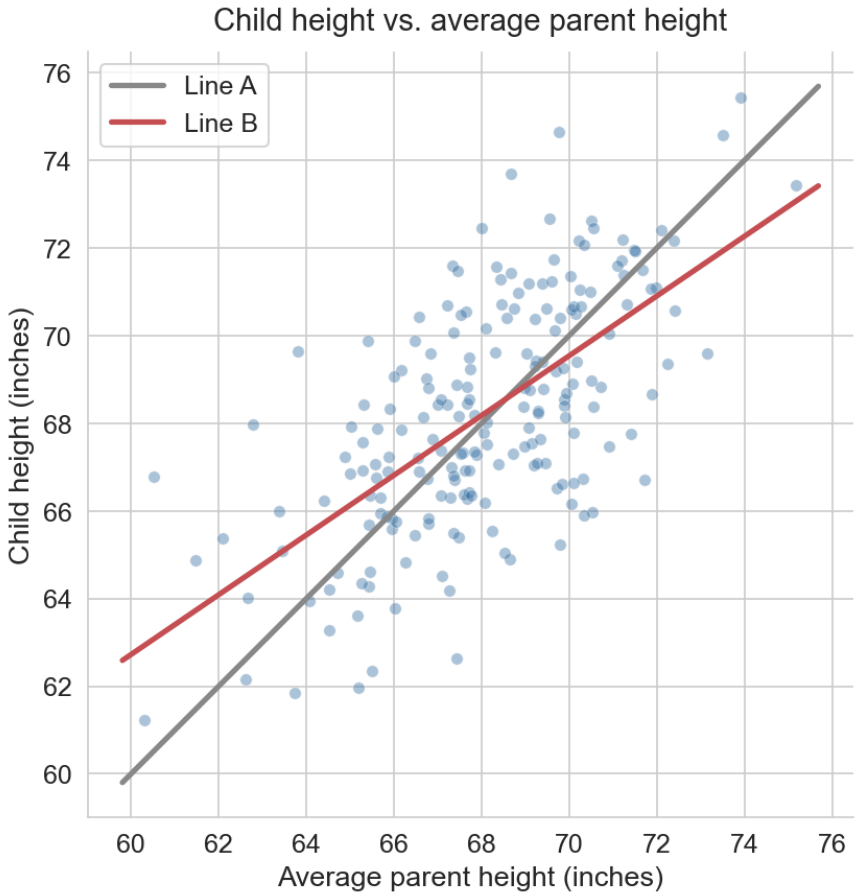
Question 9. A bicycle shop regresses used-bike resale prices on frame height and gets a coefficient of \$80/inch. Adding frame weight to the model changes the height coefficient to \$15/inch (and weight gets a coefficient of \$5/lb). Are the two models inconsistent? Use the word *multicollinearity* in your answer.

Solution

Not inconsistent — they answer different questions. Frame height and frame weight are highly correlated (taller frames weigh more), which is **multicollinearity**. The simple-regression height coefficient (\$80/inch) measures the *total* association between height and price, including

the effect of the extra weight that comes with taller frames. The multiple-regression coefficient ($15/inch) measures the association *holding weight constant*. Multicollinearity makes individual coefficients unstable — small data changes can swing them substantially — even though the model's predictions may stay similar. Decide which coefficient to report based on the question you are answering.

Question 10. The plot below shows child height vs. average parent height for a sample of families. Two lines are overlaid: **Line A** (gray) and **Line B** (red).



(a) Which line is the regression line of child on parent, and which is $y = x$ (child has the same height as the parents)?

(b) What does the relative slope of the two lines reveal about heights across generations?

Solution

(a) Line B (red, flatter) is the regression line. Line A is $y = x$. The regression line always has slope $r \cdot \mathrm{SD}_y/\mathrm{SD}_x$, and since $|r| < 1$, the slope is shallower than the 45-degree line. **(b)** The flatter regression line shows **regression to the mean**: children of tall parents tend to be tall, but less tall than their parents on average; children of short parents are short but

less short. This is purely a statistical consequence of imperfect parent-child correlation, not a biological mechanism.

Question 11. A regression has been fit by OLS and includes an intercept, so the all-ones column $\mathbf{1}$ is a column of X. Let $\epsilon = y - \hat{y}$ be the residual.

(a) Use the orthogonality condition $X^\top \epsilon = 0$ to prove that $\sum_i \epsilon_i = 0$ — the residuals always sum to zero when the model has an intercept. (One line of algebra.)

(b) A student fits a regression with an intercept, then plots residuals vs. fitted values and eyeballs the cloud as centered around $+1.2$ instead of 0. Based on part (a), what must have gone wrong?

Solution

(a) Since $\mathbf{1}$ is a column of X, the orthogonality condition $X^\top \epsilon = 0$ implies $\mathbf{1}^\top \epsilon = 0$. But $\mathbf{1}^\top \epsilon = \sum_i \epsilon_i$. So the sum of residuals is exactly zero — and therefore the mean residual is zero too.

(b) Something is wrong with the setup. If the model was truly fit by OLS *with an intercept*, the residuals cannot have a nonzero mean — that would violate $\mathbf{1}^\top \epsilon = 0$. Possible culprits: the student forgot to include an intercept column, computed residuals against a different $\hat{y}$ (e.g., predictions on a held-out set, not the training set), or dropped observations after fitting. The nonzero mean is a red flag that the "residuals" aren't really the OLS residuals they think they are.