AI Research Labs

Machine Learning Foundations · Practice test 3 · 45 questions

Linear Regression

How to use this test. Answer all 45 questions without notes, then mark yourself with the answer key at the back. Each question has exactly one best answer. Score 75% or more and you've mastered the material; below that, reread the lessons for the questions you missed.

3.1 Regression Fundamentals

Q1 apply Which of these is a regression problem?

A. Predicting whether an email is spam or not
B. Predicting a patient's blood type from tests
C. Grouping customers into similar segments
D. Predicting tomorrow's electricity demand in MWh

Q2 concept In the model $Y = f(X) + \varepsilon$, what does ε represent?

A. The uncertainty in the estimated slope coefficient
B. The residuals the fitted model makes on the training set
C. Random noise: everything the predictors don't explain
D. The intercept, the value of Y when every X is 0

Q3 calculate A house sold for £250k. The model's fitted price was £235k. What is the residual, and what does its sign mean?

A. +£15k: the model under-predicted
B. +£15k: the model over-predicted
C. −£15k: the model over-predicted
D. −£15k: the model under-predicted

Q4 concept How are residuals related to the noise term ε?

A. Residuals are always zero in a correct model
B. Residuals are estimates of the unobservable errors
C. Residuals are exactly equal to the true errors
D. The errors are observed directly; residuals are only guessed

Q5 calculate A fitted model is $\hat{y} = 30 + 2.6x$, where x is floor area in m² and $\hat{y}$ is price in £k. What's the predicted price of a 100 m² house?

A. £260k
B. £32.6k
C. £290k
D. £2,630k

3.2 Simple and Multiple Linear Regression

Q6 concept Is $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ a linear regression model?

- A Only when β_2 is close to zero
- B No: the X^2 term makes it logistic
- C Yes: it is linear in the coefficients
- D No: its curve isn't a straight line

Q7 concept Which of these models is NOT a linear regression model?

- A $Y = \beta_0 + \beta_1 X + \beta_2 X^3 + \varepsilon$
- B $Y = \beta_0 + \beta_1 X_1 X_2 + \varepsilon$
- C $Y = \beta_0 e^{\beta_1 X} + \varepsilon$
- D $Y = \beta_0 + \beta_1 \log X + \varepsilon$

Q8 interpret In $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, what is the effect of a one-unit increase in X_1?

- A $\beta_1 \beta_3$
- B $\beta_1 + \beta_3 X_2$
- C β_1
- D $\beta_1 + \beta_3$

Q9 concept A multiple regression with two predictors fits what shape in (X_1, X_2, Y) space?

- A A straight line
- B A plane
- C A sphere
- D An S-shaped curve

Q10 apply A scatter plot shows a clearly curved relationship between X and Y. What's the simplest fix that stays within linear regression?

- A Add X^2 as an extra column and refit with OLS
- B Switch to logistic regression for the curve
- C Run K-means first to find the curve's shape
- D Convert Y to 0/1 and fit the data again

3.3 Ordinary Least Squares (OLS)

Q11 calculate A regression has $n = 50$ observations and $p = 4$ predictors, plus an intercept. What are the dimensions of the design matrix X?

- A 50×5
- B 5×5
- C 50×4
- D 4×50

Q12 concept When does $\hat{\beta} = (X^\top X)^{-1} X^\top y$ fail to exist?

A When $X^\top X$ is singular

B When the response has units

C When some residuals are negative

D When the model has an intercept

Q13 concept What does ordinary least squares minimise?

A The sum of the residuals

B The sum of absolute residuals

C The number of predictors used

D The sum of squared residuals

Q14 calculate For the data $(x, y) = (1, 1), (2, 3), (3, 2)$, with an intercept, what is $X^\top X$?

A $\begin{pmatrix} 3 & 6 \\ 6 & 12 \end{pmatrix}$

B $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$

C $\begin{pmatrix} 6 & 3 \\ 14 & 6 \end{pmatrix}$

D $\begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}$

Q15 diagnose A model includes both "height in cm" and "height in inches" as predictors. What happens when you fit it with OLS?

A Both columns get exactly the same coefficient

B It fits fine, just more slowly than usual

C $X^\top X$ is singular, so there's no unique solution

D OLS averages the two columns into one automatically

3.4 Calculating Regression Coefficients by Hand

Q16 calculate For the points $(x, y) = (0, 1), (1, 3), (2, 5), (3, 7)$, what are $\hat{\beta}_0$ and $\hat{\beta}_1$?

A $\hat{\beta}_0 = 0$, $\hat{\beta}_1 = 2.5$

B $\hat{\beta}_0 = 1$, $\hat{\beta}_1 = 2$

C $\hat{\beta}_0 = 1.5$, $\hat{\beta}_1 = 2$

D $\hat{\beta}_0 = 2$, $\hat{\beta}_1 = 1$

Q17 calculate For the points $(1, 2), (2, 1), (3, 4)$, what is the slope $\hat{\beta}_1$?

A 1.5

B 2/3

C 0.5

D 1

Q18 concept Why does the slope formula fail when every x_i has the same value?

A S_{xy} becomes negative, so the slope is invalid
B $S_{xx} = 0$, so the formula divides by zero
C The residuals no longer add up to zero
D The intercept becomes undefined first

Q19 concept In a model with an intercept, the OLS residuals always...

A have variance 1
B sum to zero
C are all positive
D equal the fitted values

Q20 interpret The fitted line for $x = 1, \ldots, 5$ and $y = 2, 4, 5, 4, 5$ is $\hat{y} = 2.2 + 0.6x$. Which point must it pass through?

A $(5, 5)$, the last observation
B $(3, 4)$, the point of means
C $(0, 0)$, the origin
D $(1, 2)$, the first observation

3.5 Interpreting Coefficients

Q21 calculate For the fitted model $\hat{y} = 5 + 2x_1 - 4x_2$, what is the prediction at $(x_1, x_2) = (3, 1)$?

A 15
B 7
C 11
D 3

Q22 interpret Predictor A is in metres with $\hat{\beta} = 40$; predictor B is in millimetres with $\hat{\beta} = 0.05$. Does A matter more?

A Yes, provided both coefficients are statistically significant
B Yes, because 40 is far bigger than 0.05
C No, because B is measured more precisely
D Can't tell from raw coefficients; standardise first

Q23 concept Why can a predictor's coefficient change sign when another predictor is added to the model?

A OLS estimates are random, so their signs often flip by chance
B Each coefficient is an effect holding the others fixed
C The intercept absorbs part of every coefficient
D Adding predictors always reverses existing signs

Q24 `interpret` An observational study finds a positive coefficient for "hours of TV per day" when predicting body weight. What can you conclude?

- A Cutting TV by an hour will reduce weight by $\hat{\beta}$
- B More TV is associated with higher predicted weight
- C Watching more TV causes people to gain weight
- D TV has no real effect, as studies are biased

Q25 `interpret` A model gives price (£k) $= -50 + 3\times$ area (m²). How should you read the intercept of −50?

- A The model must be wrong, because a house price can never be a negative number
- B Every house sells for £50k less than the model predicts
- C Every house loses £50k of value before its area is counted
- D It's the prediction at zero area, outside the data, so not meaningful

3.6 Residuals and Regression Assumptions

Q26 `interpret` A plot of residuals against fitted values shows a clear U-shape. What does it suggest?

- A The model fits the data perfectly
- B The errors are correlated over time
- C The linear form is missing curvature
- D The error variance is not constant

Q27 `interpret` The spread of the residuals triples as the fitted values increase. What is the issue, and what does it affect?

- A Heteroskedasticity; SEs and p-values become unreliable
- B Multicollinearity; coefficients flip their signs
- C Underfitting; the R^2 drops all the way down to zero as well
- D Leakage; the test score becomes inflated

Q28 `interpret` Residuals from a model of monthly sales, plotted in time order, show long runs of positive values followed by runs of negative ones. What's the likely issue?

- A Heteroskedasticity: the spread keeps growing
- B Perfect multicollinearity among the Xs
- C Outliers: a few of the months are extreme
- D Autocorrelation: the errors aren't independent

Q29 `diagnose` A colleague says: "My residual plot looks fine, so the model must be correct." What's the best response?

- A Residual plots only matter when the sample size is small, so it's fine here
- B Agreed: a clean residual plot confirms the model is correct
- C A clean plot is consistent with the assumptions but proves nothing
- D Only a Q–Q plot can show whether the model is correct

Q30 concept Does fitting OLS (computing $\hat{\beta}$) require normally distributed errors?

A Yes: without normal errors $\hat{\beta}$ can't be computed

B Yes: R^2 is only defined when the errors are normal

C No: the assumptions matter for SEs, p-values and CIs

D No: OLS results never rely on any assumptions at all

3.7 R^2, Adjusted R^2, RMSE and Other Metrics

Q31 calculate A model has RSS = 30 and TSS = 120. What is R^2?

A 0.25

B 0.75

C 0.80

D 4

Q32 calculate RSS = 30, TSS = 120, $n = 25$ and $p = 4$. What is the adjusted R^2?

A 0.80

B 0.70

C 0.75

D 0.65

Q33 calculate On three test cases the residuals are $2, -1, 3$. What are the MAE and the RMSE?

A MAE 2, RMSE 2.16

B MAE 2, RMSE 4.67

C MAE 2, RMSE 1.33

D MAE 1.33, RMSE 2.16

Q34 interpret A model's out-of-sample R^2 is −0.2. What does that mean?

A 20% of its predictions on new data are wrong

B It predicts worse than the training mean on new data

C The predictors are negatively correlated with Y

D There must be an error, since R^2 can't be negative

Q35 interpret A column of pure random noise is added to an OLS model. What typically happens to R^2 and adjusted R^2?

A R^2 falls slightly; adjusted R^2 tends to rise

B Neither changes, since it's unrelated to Y

C Both fall, since the column is pure noise

D R^2 rises slightly; adjusted R^2 tends to fall

3.8 Standard Errors, Confidence Intervals and p-values

Q36 calculate A coefficient is $\hat{\beta}_1 = 1.5$ with SE $= 0.5$. What are the t-statistic and an approximate 95% CI (use $t \approx 2$)?

- A $t = 3$, CI $(0.5,\ 2.5)$
- B $t = 3$, CI $(1,\ 2)$
- C $t = 0.75$, CI $(0.5,\ 2.5)$
- D $t = 0.33$, CI $(1,\ 2)$

Q37 calculate For $x = 1, 2, 3, 4$ and $y = 1, 3, 2, 4$: $\hat{\beta}_1 = 0.8$, RSS = 1.8 and $S_{xx} = 5$. What is $\mathrm{SE}(\hat{\beta}_1)$?

- A 0.18
- B 0.60
- C 0.424
- D 0.36

Q38 interpret What does a p-value of 0.03 for a coefficient mean?

- A If the true coefficient were 0, results this extreme would occur 3% of the time
- B The coefficient is large enough to matter in practice
- C There is a 3% chance that the true coefficient is exactly zero
- D There is a 97% chance that the fitted regression model is the correct one for the data

Q39 interpret A coefficient has $\hat{\beta} = 0.9$ with a 95% CI of $(-0.2,\ 2.0)$. What can you conclude?

- A The predictor has no effect, since the interval includes zero
- B Not significant at 5%; the true effect could be negative, zero or quite large
- C The effect is significant, since the estimate is positive and larger than the interval's lower end
- D The true coefficient is 0.9, with 95% certainty

Q40 calculate A regression has $n = 40$ observations and $p = 3$ predictors plus an intercept. How many degrees of freedom do the coefficient t-tests use?

- A 37
- B 40
- C 36
- D 39

3.9 Multicollinearity and VIF

Q41 calculate Regressed on the other predictors, predictor X_j has $R_j^2 = 0.95$. What is its VIF?

- A 95
- B 1.05
- C 0.95
- D 20

Q42 `calculate` A predictor's VIF is 16. By roughly how much is its coefficient's standard error inflated?

- A About 1.6 times
- B About 8 times
- C About 16 times
- D About 4 times

Q43 `interpret` Two sensors measure almost the same pollutant ($r = 0.98$). In a regression with both, one has a large p-value, though on its own it predicts the response well. Why?

- A The sensor is physically unrelated to the response being predicted
- B p-values are unreliable whenever the sample size is very large, as it is here
- C Given the other sensor, it adds little information, so its SE is inflated
- D The model has overfitted, so every p-value is meaningless

Q44 `concept` Why can a model with severe multicollinearity still predict well on new data from the same source?

- A The combined effect is well determined; only its split is unstable
- B Multicollinearity only affects the intercept, not the slopes
- C VIF only changes R^2, never the fitted coefficients
- D Correlated predictors always make predictions more accurate on similar data

Q45 `apply` Which of these is NOT a sensible way to deal with strongly correlated predictors?

- A Add a second copy of one of the correlated predictors
- B Fit ridge regression instead of plain OLS
- C Collect data where the predictors vary independently
- D Combine the correlated predictors into a single one

Answer key

Each answer comes with the reasoning, including why the most tempting wrong option is wrong.

Q1 **D** · lesson 3.1

Demand in MWh is a quantity, so it's regression. Spam and blood type are categories (classification), and grouping without labels is clustering.

Q2 **C** · lesson 3.1

ε is the unobservable noise in the true relationship. Residuals $y_i - \hat{y}_i$ are our *estimates* of it from a fitted model, which is a common mix-up.

Q3 **A** · lesson 3.1

Residual = observed − fitted = 250 − 235 = +15. A positive residual means the actual value is above the prediction, so the model predicted too low.

Q4 **B** · lesson 3.1

We never see the true errors, because we don't know the true f. Residuals use the fitted model instead, so they approximate the errors. Even a correct model has non-zero residuals because of the noise.

Q5 **C** · lesson 3.1

$30 + 2.6 \times 100 = 290$. £260k forgets the intercept; £2,630k multiplies the intercept too; £32.6k drops the factor of 100.

Q6 **C** · lesson 3.2

"Linear" refers to the unknown coefficients. X^2 is just another column computed from the data, so OLS fits it exactly as before, even though the fitted curve bends.

Q7 **C** · lesson 3.2

In $\beta_0 e^{\beta_1 X}$ the coefficient β_1 sits inside an exponential, so the model isn't linear in its parameters. The others only transform the inputs ($\log X$, $X_1 X_2$, X^3), which is allowed.

Q8 **B** · lesson 3.2

With an interaction, the slope for X_1 depends on the value of X_2: differentiate (or compare X_1 and $X_1 + 1$) to get $\beta_1 + \beta_3 X_2$. Plain β_1 is the effect only when $X_2 = 0$.

Q9 **B** · lesson 3.2

$\beta_0 + \beta_1 X_1 + \beta_2 X_2$ is a flat plane over the (X_1, X_2) floor. With more predictors it's a hyperplane, which can't be drawn but works the same way.

Q10 **A** · lesson 3.2

Adding a squared term lets OLS fit a parabola while staying linear in the coefficients. Logistic regression is for binary outcomes, and K-means doesn't fit relationships at all.

Q11 **A** · lesson 3.3

One row per observation and one column per coefficient: four predictors plus the column of ones for the intercept. 5×5 is the size of $X^\top X$.

Q12 **A** · lesson 3.3

The formula needs the inverse of $X^\top X$. It doesn't exist with perfect multicollinearity or with more coefficients than observations. Units, intercepts and residual signs are irrelevant.

Q13 **D** · lesson 3.3

OLS minimises $\sum(y_i - \hat{y}_i)^2$. The plain sum of residuals is always zero with an intercept, so it can't pick a line; absolute residuals define a different method.

Q14 **D** · lesson 3.3

Entries are $n = 3$, $\sum x = 6$ and $\sum x^2 = 1 + 4 + 9 = 14$. (With $X^\top y = (6, 13)^\top$ the solution is $\hat{\beta} = (1, 0.5)$.) 12 would be $(\sum x)^2/3$, not $\sum x^2$.

Q15 **C** · lesson 3.3

One column is an exact multiple of the other (perfect multicollinearity), so infinitely many coefficient pairs give the same fit. Software either fails or silently drops one column.

Q16 **B** · lesson 3.4

$\bar{x} = 1.5$, $\bar{y} = 4$, $S_{xy} = 10$, $S_{xx} = 5$, so $\hat{\beta}_1 = 2$ and $\hat{\beta}_0 = 4 - 2(1.5) = 1$. The points lie exactly on $y = 1 + 2x$.

Q17 **D** · lesson 3.4

$\bar{x} = 2$, $\bar{y} = 7/3$. $S_{xy} = (-1)(-1/3) + 0 + (1)(5/3) = 2$ and $S_{xx} = 2$, so the slope is $2/2 = 1$. A common slip is to divide by n in only one of the two sums; both S_{xy} and S_{xx} are plain sums.

Q18 **B** · lesson 3.4

With no variation in x there's no information about how y changes with x, and the denominator $\sum(x_i - \bar{x})^2$ is zero. With little variation, the slope is computable but very unstable.

Q19 **B** · lesson 3.4

This follows from the normal equation for β_0. It's a handy check on hand calculations: in dataset B the residuals $(-0.8, 0.6, 1.0, -0.6, -0.2)$ sum to 0.

Q20 **B** · lesson 3.4

An OLS line with an intercept always passes through $(\bar{x}, \bar{y}) = (3, 4)$: check $2.2 + 0.6 \times 3 = 4$. It need not pass through any data point or the origin.

Q21 **B** · lesson 3.5

$5 + 2(3) - 4(1) = 7$. 15 ignores the minus sign; 11 forgets the x_2 term; 3 drops the intercept.

Q22 **D** · lesson 3.5

Coefficient size depends on units: 0.05 per mm is 50 per metre, larger than A's 40. Standardise predictors (per-SD effects) before comparing, and consider their uncertainty too.

Q23 **B** · lesson 3.5

A coefficient answers "what happens to Y when this predictor changes and the other included ones don't?" Change which predictors are included and you change the question, so the answer (even its sign) can change.

Q24 **B** · lesson 3.5

Observational regression shows association, holding the included predictors fixed. Confounders (activity, diet) could drive both. Causal claims, like "cutting TV reduces weight", need experiments or strong assumptions.

Q25 **D** · lesson 3.5

The intercept is the predicted value when all predictors are zero. A 0 m² house doesn't exist, so −50 is just where the line crosses the axis. The model can still predict well in the range of real houses.

Q26 **C** · lesson 3.6

A systematic curve means the model consistently over- then under-predicts: structure is left in the residuals. Adding X^2 or other terms may help. Non-constant variance shows as a fan, not a U.

Q27 **A** · lesson 3.6

A fan shape means the noise variance isn't constant. The usual standard-error formula assumes it is, so p-values and confidence intervals can't be trusted. Transforming Y (e.g. log) or weighted least squares can help.

Q28 **D** · lesson 3.6

Runs of same-sign residuals over time mean one month's error predicts the next. OLS assumes independent errors, so its standard errors are off. Time-series methods are designed for this.

Q29 **C** · lesson 3.6

Diagnostics can reveal problems; they can't certify a model. Missing variables or poor out-of-sample performance can hide behind a tidy plot. Check held-out error as well.

Q30 **C** · lesson 3.6

Minimising RSS is pure algebra and works on any data. The assumptions (mean zero, constant variance, independence, roughly normal) are what justify the uncertainty measures built on top of $\hat{\beta}$.

Q31 **B** · lesson 3.7

$R^2 = 1 - 30/120 = 0.75$. 0.25 is RSS/TSS itself (the share of error left over); 4 is TSS/RSS.

Q32 **B** · lesson 3.7

$1 - \frac{30/20}{120/24} = 1 - \frac{1.5}{5} = 0.70$. Residual df is $n - p - 1 = 20$; total df is $n - 1 = 24$. 0.75 is the unadjusted R^2.

Q33 **A** · lesson 3.7

MAE $= (2 + 1 + 3)/3 = 2$. MSE $= (4 + 1 + 9)/3 = 4.67$, so RMSE $= \sqrt{4.67} = 2.16$. 1.33 is the signed average error, and 4.67 is the MSE, not its root.

Q34 **B** · lesson 3.7

Out-of-sample R^2 compares the model with predicting $\bar{y}_{\text{train}}$ for everyone. Negative means the model's squared error is larger than that baseline's: worse than useless on new data. Only in-sample R^2 (with an intercept) is bounded below by 0.

Q35 **D** · lesson 3.7

In-sample R^2 can never decrease when a predictor is added: OLS finds some small chance fit. Adjusted R^2 charges for the extra parameter, so a useless predictor usually lowers it.

Q36 **A** · lesson 3.8

$t = 1.5/0.5 = 3$. CI $= 1.5 \pm 2 \times 0.5 = (0.5, 2.5)$, which excludes 0, so it's significant at 5%. $(1, 2)$ uses one SE instead of two.

Q37 **C** · lesson 3.8

$\hat{\sigma}^2 = \text{RSS}/(n - p - 1) = 1.8/2 = 0.9$, so $\text{SE} = \sqrt{0.9/5} = \sqrt{0.18} = 0.424$. 0.18 forgets the square root; 0.60 and 0.36 use RSS instead of $\hat{\sigma}^2$.

Q38 **A** · lesson 3.8

A p-value is calculated *assuming* H_0 (and the model's assumptions). It is not the probability that H_0 is true, and it says nothing about effect size.

Q39 **B** · lesson 3.8

The interval includes 0, so we fail to reject H_0. But it's wide: the data are consistent with a small negative effect or a sizeable positive one. "Not significant" isn't "no effect".

Q40 **C** · lesson 3.8

Residual df $= n - p - 1 = 40 - 3 - 1 = 36$: one is lost for each estimated coefficient, including the intercept. 37 forgets the intercept; 39 is $n - 1$.

Q41 **D** · lesson 3.9

$\text{VIF} = 1/(1 - 0.95) = 20$: its coefficient's variance is 20 times what it would be with uncorrelated predictors. 1.05 is $1/0.95$.

Q42 **D** · lesson 3.9

VIF inflates the **variance**; the standard error is its square root, so it's inflated by $\sqrt{16} = 4$.

Q43 **C** · lesson 3.9

With near-duplicate predictors the model can't tell which deserves the credit, so each coefficient is poorly determined. Insignificance here means "not needed alongside the other", not "irrelevant".

Q44 **A** · lesson 3.9

Like two people carrying one sofa, the total contribution is clear but how it's divided between them isn't. Predictions depend on the total, interpretation on the split.

Q45 **A** · lesson 3.9

Duplicating information pushes $X^\top X$ even closer to singular. Combining, regularising (ridge shares the credit) or getting more varied data all reduce the problem.