SUMMARY · AT EXAM LEVEL

Module 5

Linear regression

Everything you must be able to define, explain, apply and assess for the open questions. With the models as figures, the comparisons as tables and the pitfalls marked.

WHAT IS INSIDE

01	Ordinary least squares and residuals	2
02	Reading coefficients: intercept, slope and ceteris paribus	2
03	Standard error, t-value, p-value and confidence interval	3
04	R-squared and adjusted R-squared	3
05	F-test and partial F-test	3
06	Dummy variables and the reference category	3
07	Omitted variable bias	4
08	Multicollinearity and VIF	4
09	Standardized coefficients	4
10	Non-linear effects: square and log	4
11	Interaction term	5
12	Mediation	5
13	Prediction and extrapolation	5
14	Outliers and the sample	5
15	Question 7.2: structure of the answer (250 words)	5
→	How to write the answer	6
→	Checklist: can I explain this?	6
→	Test yourself	7

EXAM LIKELIHOOD

Interpreting regression coefficients (ceteris paribus, units) CERTAIN	97%
R-squared and goodness of fit CERTAIN	95%
Dummy variables and reference category CERTAIN	92%
Ordinary least squares (OLS) CERTAIN	92%
Omitted variable bias CERTAIN	92%
Error term and OLS assumptions CERTAIN	88%
Comparing standardized coefficients CERTAIN	88%

Estimate per topic, based on the sample exam and the emphasis in the material.

KEY CONCEPTS

ordinary least squares · residual · error term · intercept · slope · ceteris paribus · R-squared · adjusted R-squared · standard error · t-value · p-value · confidence interval · dummy variable · reference category

ON THE EXAM

Block 5

17.5 points · 180 min in total

MC	7 multiple-choice questions	**2.5 points per question**
7	Block 7 (25 points) consists of three open questions about R output.	**200 / 250 / 150 words**

Module 5 yields 7 multiple-choice questions (17.5 points) and the open question 7.2 (9 points, at most 250 words). The formula sheet only contains R-squared, adjusted R-squared and the t-test; you must know all other formulas. The examples below come from real R output; simulated cases are marked as such.

01 Ordinary least squares and residuals



FIGURE 5.1 **Assumptions of linear regression.** Check them on the residuals.

OLS chooses the line with the smallest sum of squared **residuals** (SSR). A **residual** is observed y minus predicted y. The **error term** belongs to the population model and contains everything that affects y and is not in the model. With OLS and an intercept the sum of the residuals is always zero, which is why you square them.

> **EXAM TIP**
> Exam tip: "OLS minimises the sum of the absolute residuals" is wrong; it is the squared residuals.

02 Reading coefficients: intercept, slope and ceteris paribus

The **intercept** is the predicted y when all x are zero (with dummies: the prediction for the reference category). The **slope** b1 is the change in y on average for a rise of x by 1 unit. In a multiple regression you add **ceteris paribus**: the other predictors held constant.

Example (simulated data, 90 stores, revenue in thousand euros): **omzet = 43.27 + 2.66 reclame + 6.31 personeel – 7.08 parkeer**. Per extra 1000 euros of advertising, revenue rises on average by 2.66 thousand euros, ceteris paribus. Prediction for reclame 10, personeel 8 and parkeer 1: 43.27 + 26.63 + 50.47 – 7.08 = 113.28 (rounded).

Regression on observational data shows association, not causality. Phrase it as "is associated with".

03 Standard error, t-value, p-value and confidence interval

The **t-value** is (b – beta0)/SE; under H0: beta = 0 it is b/SE. Reclame: 2.6625/0.6187 = 4.30, p = 4.4e-05, so significant. Parkeer: -7.0823/4.8682 = -1.46, p = 0.149, so not significant at 5%.

The **p-value** is the probability of an estimate this extreme if H0 is true. It is not the probability that H0 is true. The **confidence interval** is b +/- roughly 2 x SE. If it contains zero, not significant at 5%. Statistically significant is not the same as practically important: also look at the size.

04 R-squared and adjusted R-squared

R-squared = SSE/SST = 1 – SSR/SST, with SST = SSE + SSR. Retail model: R-squared 0.6644 means 66.4% of the variation in revenue is explained. Adjusted R-squared = 1 – (1 – R2)(n – 1)/(n – k – 1) = 1 – 0.3356 × 89/86 = 0.6527.

R-squared rises or stays equal with every extra predictor; **adjusted R-squared** can fall and is the measure to compare models with a different number of predictors. It does not decide whether a predictor is significant: the p-value does. A high R-squared says nothing about causality or bias.

05 F-test and partial F-test

The **F-test** at the bottom of **summary(lm)** tests H0: all slopes are zero at the same time. F = (R2/k)/((1 – R2)/(n – k – 1)) = (0.6644/3)/(0.3356/86) = 56.75. A significant F says the model as a whole explains something, not that every predictor is significant.

The **partial F-test** (**anova(small, large)**) compares two nested models. Long Island housing data: the large model with four extra style dummies (both models have **taxes** and **rooms**) lowers the residual sum of squares from 793,934 to 607,120; F = ((793,934 – 607,120)/4)/(607,120/83) = 6.38, p = 0.00016. The four dummies add something together. R-squared rises from 0.0785 to 0.2953, adjusted R-squared from 0.0573 to 0.2444.

06 Dummy variables and the reference category

A category with k levels becomes k – 1 **dummy variables**. The remaining category is the **reference category**; the intercept belongs to that group and each dummy coefficient is the difference with that group, ceteris paribus. Pitfall: k dummies at once gives perfect multicollinearity (the dummy trap).

Example (houses on Long Island, 90 houses): **valuem2 = 208.11 + 0.029 taxes – 157.98 f.regionSuffolk**; R-squared 0.427. A house in Suffolk is worth on average 157.98 dollars per m2 less than in Nassau at equal taxes. If you change the reference

to Suffolk, the dummy gets +157.98 and the intercept changes; R-squared and predictions stay equal.

07 Omitted variable bias

If a variable is missing that is related to both y and an included x, that coefficient is distorted (**omitted variable bias**): it also picks up the effect of the omitted variable. You recognise it because the coefficient changes strongly once you add the variable. The reverse is not true: a variable that is unrelated to x distorts nothing.

08 Multicollinearity and VIF

Multicollinearity is strong association between predictors. Consequence: larger standard errors and unstable, often non-significant coefficients; the estimate stays unbiased. The **VIF** of variable j is 1/(1 – R2_j), with R2_j from the regression of x_j on the other predictors. At R2_j = 0.8 the VIF is 5. Housing data: rooms 1.713, bedrooms 1.532, bath 1.229: no problem. Solution: drop or combine one of the variables.

09 Standardized coefficients

A **standardized coefficient** expresses the effect in standard deviations: 1 sd more in x goes with b sd more in y. You use it to compare the relative strength of predictors with different units. The p-values do not change.

10 Non-linear effects: square and log

A **quadratic term**: y = b0 + b1 x + b2 x^2 gives a marginal effect b1 + 2 b2 x and a minimum or maximum at -b1/(2 b2) (a minimum if b2 > 0). Housing data: lotsize -0.6455 and lotsize^2 0.0001853; minimum 0.6455/(2 × 0.0001853) = 1742 within a range of 381 to 2044; R-squared 0.7204. Outside that range the curve does not hold: **extrapolation** is unreliable.

A **log transformation**: in log(y) = b0 + b1 x, b1 is roughly a relative change of 100 x b1 percent per unit of x (exactly: e^b1 – 1); in log-log b1 is an elasticity (percent per percent). Example (simulated data, 240 alumni): in **log(salaris)** the coefficient for vrouw is -0.056; that is roughly 5.6% lower salary, ceteris paribus (p = 0.081, not significant at 5%).

11 Interaction term

With **y ~ x1 * x2** the effect of x1 equals b1 + b3 ×2. The main coefficient b1 is the effect at x2 = 0. With a dummy and a number each group has its own slope. Example (simulated data, 150 deliveries): the interaction afstand x regen is 0.774 extra minutes per kilometre when it rains. Housing data **rooms * f.region**: interaction 6.92 with p = 0.584, so no evidence for a difference in slope.

12 Mediation

Mediation means x works on y through an intermediate variable m. You estimate three regressions: y on x (total effect), m on x and y on x and m (direct effect). If the direct effect of x becomes zero once m is in the model, there is full mediation. Moderation is different: there a variable changes the strength of the effect (interaction term).

13 Prediction and extrapolation

You predict with **predict(model, newdata = data.frame(...))**. A **prediction interval** is wider than a confidence interval, because it also contains the spread of an individual y. Energy data: 36.961 + 0.14 × 5 = 37.66. Predicting outside the range of x (**extrapolation**) is unreliable.

14 Outliers and the sample

Two outliers can reverse a result: in the energy data the p-value is 0.216 with all observations and 0.047 without the two observations with pCool above 35 (R-squared 0.7% to 1.9%). That is a reason to look at the data, not to delete observations without reason.

15 Question 7.2: structure of the answer (250 words)

1. Coefficients: read b with units, ceteris paribus, and say what it means for the case (2 points).
2. P-values: significant at 5% yes or no, what H0 says, no probability of H0; significance says nothing about the size or importance of the effect (2 points).
3. Model fit: R-squared or adjusted R-squared as share of explained variation, and RSE/RMSE as typical prediction error; a good fit on this data does not guarantee good predictions for new data (2 points).
4. Sample: who is in it, bias, representativeness, internal and external validity (3 points).

MODEL ANSWER

How to write the answer

Four building blocks, in this order. Each one earns points; leave none out.

1. Term
Name the model or concept with the English term.

2. Definition
Give the definition or features, precisely.

3. Case fact
Link it to a concrete fact from the case.

4. Conclusion
End with a judgement, advice or consequence.

QUESTION 1 · WORD LIMIT 75 WORDS (WORD LIMIT ASSUMED)

In a regression the coefficient of a dummy variable is 4.2 with $p < 0.05$. What does that mean?

TERM The coefficient of a dummy variable compares a group with the reference group. DEFINITION It gives the expected difference in Y between group 1 and the reference group (0), other variables held constant. CASE FACT Here the difference is 4.2 units, and because $p < 0.05$ it is significantly different from zero. CONCLUSION Group 1 scores 4.2 higher on average than the reference group; this is an association, not a proven cause.

68 / 75 words

KEY CONCEPTS

Checklist: can I explain this?

Tick what you can explain from memory in one sentence, using the English term.

- ☐ ordinary least squares
- ☐ residual
- ☐ error term
- ☐ intercept
- ☐ slope
- ☐ ceteris paribus
- ☐ R-squared
- ☐ adjusted R-squared
- ☐ standard error
- ☐ t-value
- ☐ p-value
- ☐ confidence interval
- ☐ dummy variable
- ☐ reference category
- ☐ omitted variable bias
- ☐ multicollinearity
- ☐ VIF
- ☐ standardized coefficient
- ☐ interaction term
- ☐ F-test
- ☐ partial F-test
- ☐ mediation
- ☐ extrapolation

FLASHCARDS

Test yourself

Cover the right-hand column and say the answer out loud.

Question	Answer
1 What does OLS minimise?	The sum of squared residuals (SSR) around the regression line.
2 What does the intercept mean?	The predicted y when all x variables are zero (or the reference group).
3 When is a coefficient significant at the 5% level?	When the p-value is smaller than 0.05 (|t| roughly larger than 2).
4 What relation holds between SST, SSE and SSR?	SST = SSE + SSR: total variation = explained + residual.
5 How do you interpret a dummy coefficient?	The average difference in y with the reference group, ceteris paribus.
6 What happens to the coefficient of x if you omit a relevant confounder?	It is biased: it also picks up the effect of the confounder.
7 Which VIF value points to a problem?	Roughly a VIF above 5 to 10 (R2_j above 0.8 to 0.9).
8 How do you interpret b1 in log(y) = b0 + b1 x?	A rise of x by 1 unit goes with roughly 100 x b1 percent change in y.
9 What does an interaction term x1 * x2 do?	The effect of x1 on y depends on the value of x2 (and vice versa).
10 What does the F-test at the bottom of summary(lm) test?	H0: all slope coefficients are zero at the same time; the model explains nothing.
11 What is mediation?	x affects y through an intermediate variable m (x to m to y).
12 How do you predict in R for a new observation?	predict(model, newdata = data.frame(...)), with all predictors included.