Curtin University, Semester 2, 2026

ECOM5005 Tutorial 10 Solution

Simple Linear Regression

Learning Outcomes

1. Conduct a simple regression and interpret the meaning of the regression coefficients.
2. Use regression analysis to predict the value of a dependent variable.
3. Make inferences about the slope.
4. Make inferences about the correlation coefficient.

PART 1 – Detailed solutions

Question 1. High blood pressure

Does the reported association imply that watching more television raises blood pressure?

Detailed solution

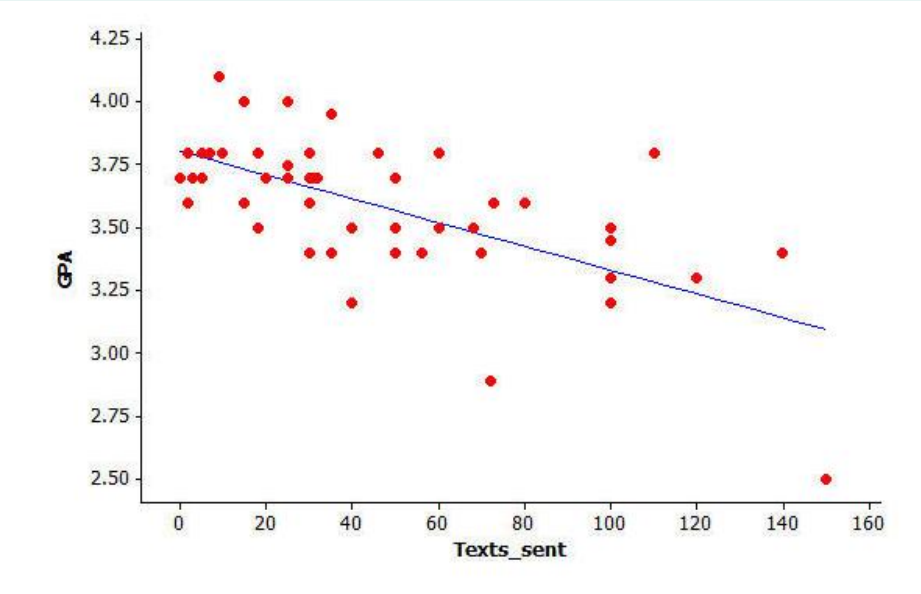
- The study described is observational: students were not randomly assigned to different amounts of television viewing.
- A statistically significant association does not establish causation. Students who watch more television may differ in physical activity, diet, sleep, stress, socioeconomic background or other factors that are also related to blood pressure.
- A randomized experiment would provide a stronger basis for causal inference because random assignment helps balance confounding variables across groups.

Answer

The study supports an association between television viewing and blood pressure, not a claim that television viewing causes higher blood pressure.

Question 2. Testing and Grades

Regression equation: GPA-hat = 3.8 - 0.005(Texts sent). Interpret -0.005 and 3.8.



Detailed solution

- Slope b1 = -0.005: for each additional text message sent, the predicted GPA decreases by 0.005 points, on average, according to the fitted linear relationship.
- A more interpretable scaling is 100 additional texts: 100 x 0.005 = 0.50. Thus, 100 more texts are associated with about a 0.50-point lower predicted GPA.
- Intercept b0 = 3.8: the model predicts an average GPA of 3.8 when Texts sent = 0. Because zero lies near the observed range in the scatter plot, this intercept has a plausible contextual meaning.
- The negative slope describes association, not causation. The model does not show that sending texts causes GPA to fall.

Answer

Slope: -0.005 GPA points per additional text. Intercept: predicted GPA = 3.8 when no texts are sent.

Question 3. Used Subaru Foresters

Least-squares equation: predicted price = 24,247.56 - 1482.06 age.

Detailed solution

- (a) X = age of the Forester, measured in years. Y = listed price, measured in dollars.
- (b) The slope is -1482.06 dollars per year.
- (c) For each additional year of vehicle age, the predicted listed price decreases by about $1,482.06, on average, over the age range represented by the data.
- (d) The intercept is $24,247.56. It is the model-predicted listed price at age 0 years. Whether this is practically useful depends on whether newly registered vehicles are reasonably represented by the observed data; regression intercepts should not be over-interpreted when X = 0 is outside the relevant data range.

Question 4. Digital advertising and weekly online sales

n = 24, predicted sales ($000) = 58.4 + 4.72(ad spend $000), R-squared = 0.64, SE(b1) = 0.7547; observed ad-spend range = 5 to 25 ($000).

Detailed solution

- (a) X is digital advertising spend in $000s; Y is weekly online sales in $000s. The slope 4.72 means an additional $1,000 of advertising is associated with $4,720 higher predicted weekly sales, on average. The intercept 58.4 predicts $58,400 of sales at zero advertising; since zero is outside the observed 5-25 range, this intercept should be interpreted cautiously.
- (b) At X = 18: predicted sales = 58.4 + 4.72(18) = 143.36 ($000) = $143,360. Residual = actual - predicted = 151 - 143.36 = 7.64 ($000). The positive residual means actual sales were $7,640 above the model prediction.
- (c) H0: beta1 = 0 versus H1: beta1 != 0. Test statistic t = 4.72 / 0.7547 = 6.254 with df = n - 2 = 22. The two-tailed p-value is about 0.0000027, far below 0.05, so reject H0. There is strong evidence of a linear association between advertising spend and weekly sales.
- (d) In simple linear regression, r has the same sign as the slope and |r| = sqrt(R-squared). Therefore r = +sqrt(0.64) = +0.80, indicating a strong positive linear association.
- (e) $50,000 is far outside the observed $5,000-$25,000 range. Using the line there would be extrapolation, where the linear relationship may not continue. The model should not be used for that prediction without additional data.

Answer summary

At $18,000 advertising: predicted sales = $143,360 and residual = +$7,640. t = 6.254, df = 22, p < 0.001; r = +0.80. Do not extrapolate to $50,000.

Question 5. Emergency-department staffing and patient waiting time

n = 32, predicted wait = 96.4 - 3.85(nurses), SE(b1) = 0.95, R-squared = 0.354; staffing range = 7-15 nurses.

Detailed solution

- (a) The slope -3.85 means each additional nurse is associated with a 3.85-minute decrease in predicted average waiting time. The intercept 96.4 predicts the waiting time at zero nurses; because zero is far outside the observed staffing range and operationally unrealistic, the intercept has little practical meaning.
- (b) With 12 nurses: predicted wait = 96.4 - 3.85(12) = 50.2 minutes. Residual = 61 - 50.2 = +10.8 minutes, so the actual average wait was 10.8 minutes longer than predicted.
- (c) H0: beta1 = 0. t = -3.85 / 0.95 = -4.053 with df = 30. The two-tailed p-value is about 0.00033, so reject H0 at 5%. There is evidence of a negative linear association between nurse staffing and waiting time.
- R-squared = 0.354 means about 35.4% of the sample variation in average waiting time is explained by the linear relationship with the number of nurses. Most variation is still attributable to other factors and random variation.
- (d) A fan-shaped residual plot is evidence against equal variance (homoscedasticity). The unusually large residual may be an outlier or influential case. The analyst should also check whether the fan pattern is masking nonlinearity.
- (e) These are observational shift data. Higher staffing may occur on predictably busier evenings, and patient mix, arrivals, bed availability and case severity may confound the relationship. Regression alone does not establish that adding nurses causes the observed waiting-time reduction.

Answer summary

Predicted wait at 12 nurses = 50.2 minutes; residual = +10.8 minutes. The slope is statistically significant, but equal variance is questionable and causal claims are not justified.

Question 6. Streaming promotions, subscriptions and autocorrelation

n = 36 months, predicted subscriptions (000s) = 12.6 + 1.85(promo spend $000), R-squared = 0.49, Durbin-Watson D = 0.72.

Detailed solution

- (a) The slope 1.85 means an additional $1,000 of promotional spending is associated with 1.85 thousand additional predicted subscriptions. At X = 20, predicted subscriptions = 12.6 + 1.85(20) = 49.6 thousand.
- (b) Because the slope is positive and R-squared = 0.49, r = +sqrt(0.49) = +0.70, indicating a moderately strong positive sample correlation.
- (c) The Durbin-Watson statistic ranges from 0 to 4 and is close to 2 when first-order autocorrelation is absent. D = 0.72 is well below 2, consistent with positive autocorrelation. The long same-sign runs in the residual plot reinforce this concern.
- (d) Positive autocorrelation violates the independence-of-errors component of LINE. When errors are dependent, the usual OLS standard errors, t tests, p-values and confidence intervals can be misleading, even if the fitted line itself looks reasonable.
- (e) A sensible next step is to re-specify the model to account for time structure, for example by including time trend/seasonality variables or using an appropriate time-series regression method, then recheck the residuals before relying on inference.

Answer summary

Prediction at $20,000 promo spend = 49.6 thousand subscriptions; r = +0.70; D = 0.72 suggests positive autocorrelation and a violation of independence.

PART 2 – Data exercise: EPA fuel economy

Dataset context

The workbook contains 234 observations from 38 popular vehicle models in model years 1999 and 2008. Highway MPG and engine displacement are used for this exercise. The dataset is derived from U.S. EPA FuelEconomy.gov records and distributed as the mpg teaching dataset.

Step-by-step Excel instructions

1. Open “EPA_Fuel_Economy_Data.xlsx” and select the Regression_Data sheet. The explanatory variable is displ_L in column E and the response variable is hwy_MPG in column F.

2. If Data Analysis is not visible on the Data tab, enable the Analysis ToolPak. Windows: File > Options > Add-ins > Manage Excel Add-ins > Go > tick Analysis ToolPak. Mac: Tools > Excel Add-ins > tick Analysis ToolPak.

3. Create the scatter plot. Select E1:F235, choose Insert > Scatter > Scatter with only markers. Put engine displacement on the horizontal axis and highway MPG on the vertical axis. Add a Linear Trendline and tick Display Equation on chart and Display R-squared value on chart.

4. Calculate the correlation coefficient in a blank cell using =CORREL(E2:E235,F2:F235). Because the slope is negative, the correlation should also be negative.

5. Run the regression. Choose Data > Data Analysis > Regression. Set Input Y Range to F1:F235 and Input X Range to E1:E235. Tick Labels. Choose a new worksheet ply or an output range. Tick Residuals, Residual Plots, Line Fit Plots and Normal Probability Plots. Keep the 95% confidence level and click OK.

6. In the Regression Statistics block, record Multiple R, R Square, Adjusted R Square, Standard Error and Observations. Remember: Excel reports Multiple R as a non-negative magnitude; use the sign of the slope (or CORREL) when discussing the correlation coefficient.

7. In the ANOVA table, record Regression SS, Residual SS, Total SS, the F statistic and Significance F. In simple regression, the overall F test and the two-sided slope t test reach the same significance conclusion.

8. In the Coefficients table, record the Intercept and displ_L coefficient, their standard errors, t statistics, p-values and 95% confidence limits.

9. For a 3.0-litre engine, compute predicted MPG with =intercept+slope*3. For the first row (Audi A4, 1.8 L, observed highway MPG = 29), calculate predicted MPG = intercept+slope*1.8 and residual = observed-predicted.

10. Inspect the residual plot for curvature, changing spread and unusual observations. Inspect the normal probability plot and residual histogram for approximate normality. Do not assess independence only from the residual-vs-X plot; independence is mainly a design/time-order issue.

Final numerical results

Regression statistic	Value
Observations	234
Multiple R (magnitude)	0.7660
Correlation r	-0.7660
R Square	0.5868
Adjusted R Square	0.5850
Standard Error of estimate	3.8360 MPG

Fitted regression equation:

predicted highway MPG = 35.6977 - 3.5306(engine displacement in litres)

ANOVA source	df	SS	MS	F	Significance F
Regression	1	4,847.833	4,847.833	329.453	2.04E-46
Residual	232	3,413.829	14.715		
Total	233	8,261.662			

Term	Coefficient	Std. Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	35.6977	0.7204	49.5548	2.12E-125	34.2784	37.1169
Engine displacement	-3.5306	0.1945	-18.1508	2.04E-46	-3.9138	-3.1473

Interpretation of the slope: for each additional 1 litre of engine displacement, the model predicts about 3.53 fewer highway MPG, on average, across these observations.

Interpretation of R-squared: approximately 58.68% of the sample variation in highway MPG is explained by the linear relationship with engine displacement. The remaining 41.32% is not explained by this one-predictor linear model.

Inference for the slope: t = -18.1508 with df = 232 and p ≈ 2.04 x 10^-46. Reject H0: beta1 = 0. There is extremely strong evidence of a linear association between engine displacement and highway MPG.

Prediction for 3.0 L: 35.6977 - 3.5306(3) = 25.1059 MPG, so approximately 25.11 MPG.

First observation residual: predicted MPG for 1.8 L = 29.3426; residual = 29 - 29.3426 = -0.3426 MPG. The observed value is about 0.34 MPG below the fitted line.

Final figures

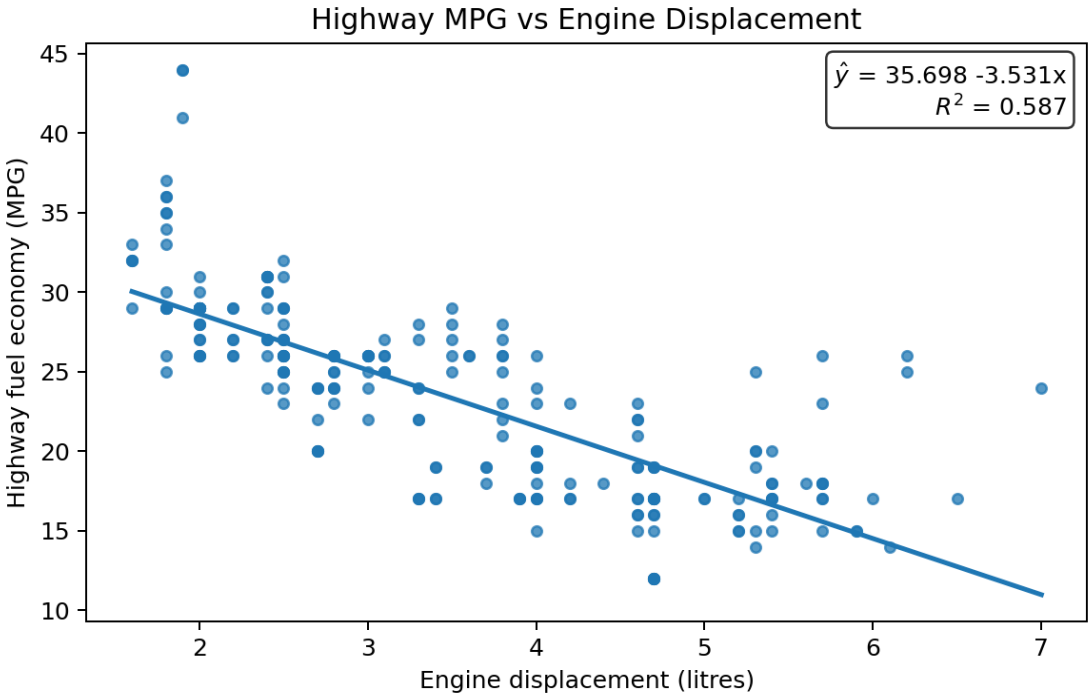


Figure 1: Highway MPG versus engine displacement with the fitted regression line

The plot shows a clear negative association: vehicles with larger engines tend to have lower highway MPG. The relationship is reasonably strong but not perfectly linear.

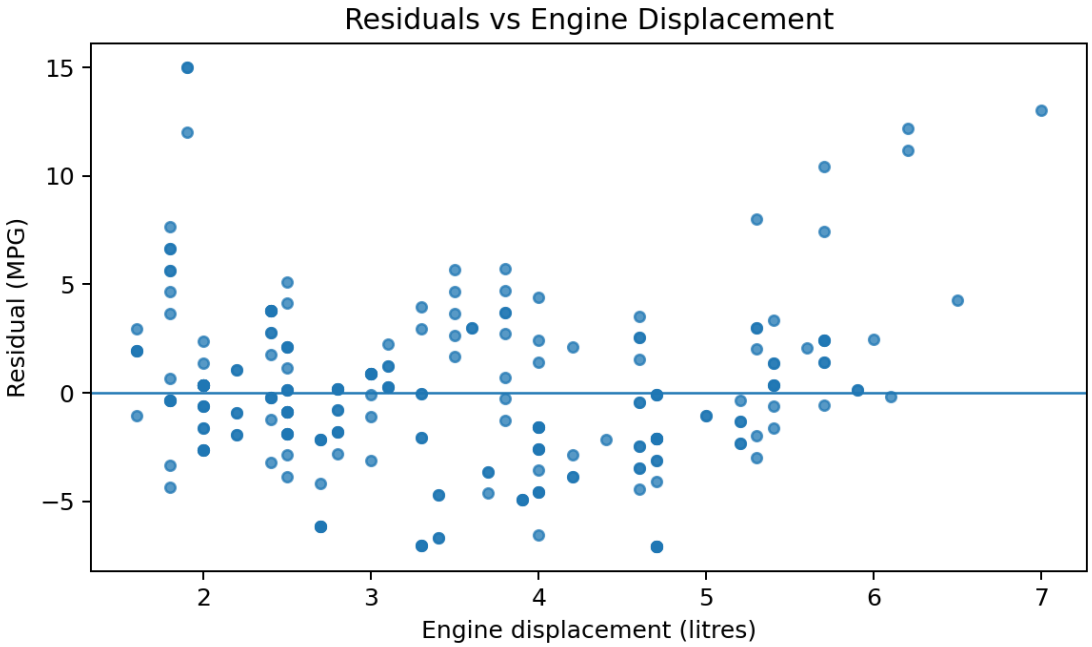


Figure 2: Residuals versus engine displacement

The residuals are centred around zero, but there is visible structure and changing spread across engine sizes. This suggests the simple linear form and equal-variance assumption are only approximate.

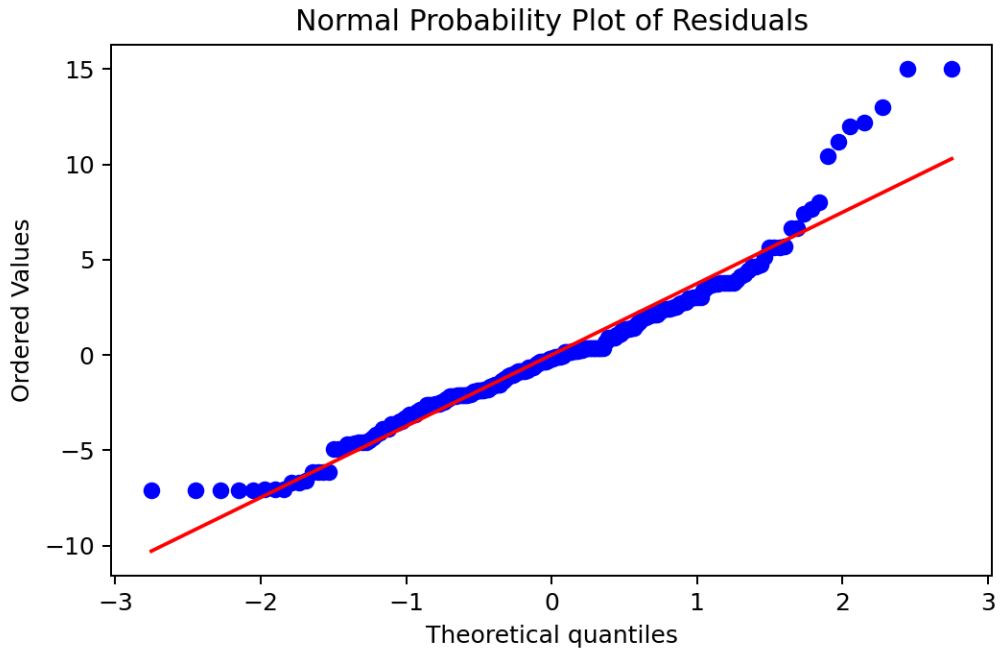


Figure 3: Normal probability plot of residuals

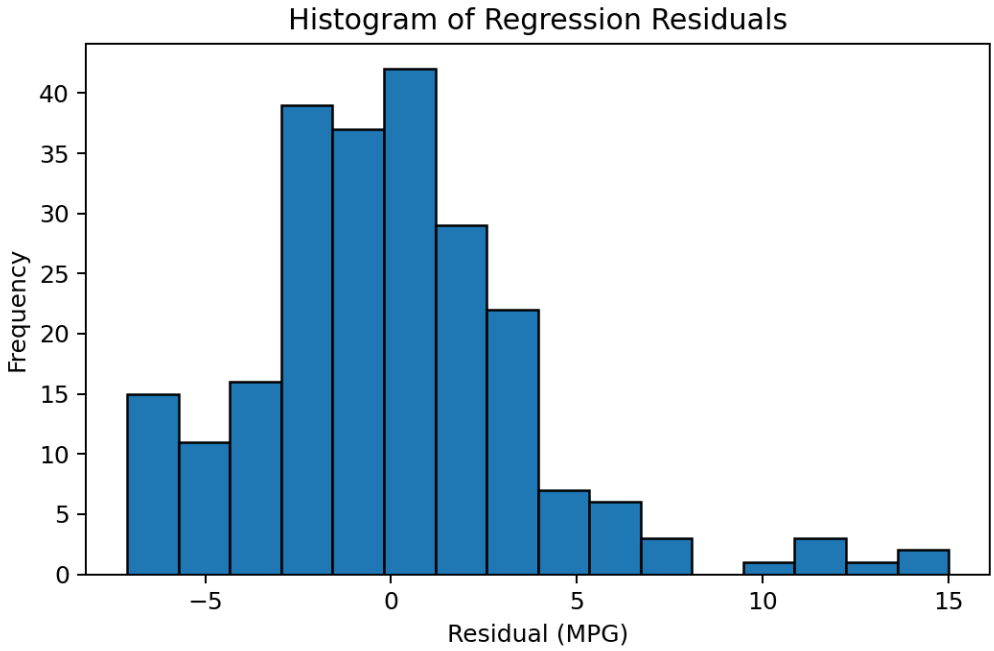


Figure 4: Histogram of residuals

The residual distribution is not perfectly normal; the normal probability plot departs from a straight line in the tails and the histogram is asymmetric. With n = 234, slope inference is still numerically very strong, but students should report the diagnostic limitations rather than treating the LINE assumptions as exact.

Part 2 conclusion

Engine displacement is strongly and negatively associated with highway fuel economy in this sample. The fitted model predicts about 3.53 fewer highway MPG for each additional litre of displacement and explains about 58.7% of the observed variation in highway MPG.

This is an association, not a causal estimate. Engine displacement is related to vehicle class, weight, drivetrain, cylinders, transmission and many other engineering choices. A one-predictor regression cannot isolate a causal effect.

Use the model as a useful descriptive relationship within the observed 1.6-7.0 litre range, while acknowledging residual-pattern and omitted-variable limitations.

PART 3 – Pearson MyLab

1. Please ensure that you have completed Quiz 5 will be due on this Friday (23-Oct-2026) at 5:pm AWST.

Note: If you are having issues with obtaining access to MyLab, please re-read the instructions (available in the welcome topic). If you are still unsuccessful, please send an email to the teaching team explaining the issue. It is important that we complete MyLab tasks on a weekly basis since the final e-test in this unit will be completed using MyLab.