Solution sketch for ST2304 exam - spring 2025

1. No: The independent variables are normally distributed

   No: The sample size must be greater than 100.

   Yes: The residuals are zero-mean normally distributed. This assumption is common for conducting valid hypothesis tests and constructing confidence intervals.

   Yes: The residuals have constant variance (homoscedasticity). This means the variance of the errors is the same across all levels of the independent variables. Violation of this assumption (heteroscedasticity) can lead to inefficient estimates and invalid inference.

   Yes: The independent variables are measured without error In classical linear regression, it is assumed that the independent variables (covariates) are measured accurately. Measurement error can bias coefficient estimates.

   Yes: The relationship between the independent and dependent variable is linear.

2. No: A 95% confidence interval means there is a 95% probability that the population parameter lies within the interval.

   Yes: If the confidence interval for a mean difference includes zero, there may be no significant difference between the groups.

   Yes: Increasing the sample size generally results in a narrower confidence interval.

   No: A 95% confidence interval will contain the sample mean 95% of the time.

   No: A wider confidence interval indicates more precise estimates of the population parameter.

3. Photo = 1.9853 + 0.4876 × N + 0.2387 × Light

4. The coefficient for N means that for every 1 mg/g increase in nitrogen, the photosynthetic rate increases by about 0.488 μmol CO_2 m^{-2} s^{-1}, holding sunlight constant.

   The coefficient for Light means that for every additional hour of sunlight, the photosynthetic rate increases by about 0.239μmol CO_2 m^{-2} s^{-1}, holding nitrogen constant.

5. An R-squared of 0.624 means that 62.4% of the variation in photosynthetic rate is explained by nitrogen content and sunlight exposure combined.

6. - A curved pattern in the Residuals vs Fitted plot suggests a violation of the linearity assumption.
   - In the Normal Q-Q plot, if the residuals deviate from the straight line at both ends, this suggests non-normality. While linear regression assumes normality of residuals for hypothesis testing, slight deviations are common and may not drastically impact the results. However, large deviations at the tails indicate possible outliers or skewness.
   - A point with high leverage and large residual may be an influential data point.

7. If the linearity assumption is violated a more complex model might be needed. We could transform variables, and include quadratic terms. Regarding the potential influental data point, we first need to check that they are not typos, or other errors (e.g. computers reading them in incorrectly). If it is genuine, then we can remove it and see if that makes a big difference. If it does, we have to think hard. It is possible that they really should be 'in' the data, so that we would be inflating the fit of the model by removing them.

8. Standardizing covariates allows you to directly compare the size of the coefficients, as they represent the effect of a one standard deviation change in the predictor. Standardizing the variables involves subtracting the mean and dividing by the standard deviation for each covariate. This means the covariates would have a mean of 0 and a standard deviation of 1.

9. When covariates are not standardized, the interpretation of the coefficients directly reflects the change in the outcome variable (photosynthesis rate) for a one-unit change in each predictor (e.g., an additional 1 mg/g of nitrogen or an additional hour of sunlight), based on the scale of the covariates. With standardized variables, the coefficients would indicate the change in photosynthetic rate for a one standard deviation increase in nitrogen content or sunlight exposure.

10. The maximum likelihood estimator is given by X/N = 108/120 = 0.9.

11. Under the normality assumption a 95% CI is computed by $\hat{p} \pm 1.96 * SE(\hat{p})$, where $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/N}$. This leads to CI=$0.9 \pm 1.96 \times 0.0274$, that is (0.846,0.954).

12. A log link (natrural logarithm) was used as common for Poisson regression.

13. From model2, what is the expected number of insects on a dry-habitat plant that is 10cm taller than the average (answer to one decimal place)?

    The interaction term Habitatwet:Height = -0.002655 indicates that the effect of plant height on insect counts is minimal smaller in wet habitats.

In dry habitat, the slope for height is 0.009988.

In wet habitat, it is 0.009988 - 0.002655 = 0.007333. So, for every 1 cm increase in height, dry-habitat plants attract more insects than wet-habitat ones.

For a dry habitat plant (+10 cm), the model is:

$eta = 0.968345 + 0.009988 * 10 = 1.068225$

Thus the expected count is $exp(1.068225) \approx 2.910209$ insects.

For a wet habitant plant (+10cm), the model is:

$eta = 0.968345 + 0.651597 + 0.00733 * 10 = 1.693242$

Thus the expected count is $exp(1.693242) \approx 5.437079$ insects.

14. The p-value (0.1246) from the likelihood ratio test is not significant at 5% level suggesting suggests that model1 (without interaction) fits the data s better than model2. Thus, the interaction is not important and could be removed.

15. The increasing residual deviance suggests overdispersion — variance exceeds the mean, violating Poisson assumptions.

    Solutions:

    Use a quasi-Poisson or negative binomial GLM, which allows variance ¿ mean.

    Add relevant covariates or random effects (e.g. mixed model) to account for unexplained variation.