Exam 2 Solutions - STAT 324

## Question 4

### Question:

Based on the best model output and supporting diagnostics (e.g., VIF values), is multicollinearity a problem in the model? Explain your reasoning clearly.

**Total Points:** 15

### Solution:

Based on the provided model output and VIF values, multicollinearity is **not considered a significant problem** in this model.

Here’s the reasoning:

1. **Examination of VIF Values:** The Variance Inflation Factors (VIFs) for the predictor variables are as follows:
   * Agriculture: 2.147
   * Education: 1.816
   * Catholic: 1.299
   * Infant.Mortality: 1.108
2. **Interpretation of VIF Values:** VIF values quantify how much the variance of the estimate of a regression coefficient is increased due to multicollinearity. A commonly accepted guideline is that VIF values above 5 or 10 indicate problematic levels of multicollinearity.
3. **Conclusion:** All the calculated VIF values in this model are well below the common thresholds of 5 or 10. The highest VIF is approximately 2.15 for Agriculture, which is far from suggesting a significant issue with multicollinearity. Therefore, the predictor variables in this model do not appear to be excessively correlated with each other to a degree that would severely impact the reliability or interpretation of the coefficient estimates.

**In summary, the low VIF values indicate that multicollinearity is not a significant concern for the best\_model\_swiss.**

### Rubric:

| Criteria | Points | Descriptors |
| --- | --- | --- |
| **Identification of VIF Values** | 3 | Correctly identifies and lists the VIF values for each predictor variable from the output. |
| **Understanding of Multicollinearity & VIF** | 5 | Explains what multicollinearity is and how VIF values are used to detect it. Mentions common VIF thresholds (e.g., 5 or 10). |
| **Analysis of VIF Values** | 4 | Compares the observed VIF values to the accepted thresholds. Correctly states whether the values indicate a multicollinearity problem. |
| **Clear Explanation and Conclusion** | 3 | Provides a clear and concise explanation of the reasoning. Draws a correct conclusion about the presence or absence of problematic multicollinearity in the model. |
| **Overall Clarity and Organization** |  | (Implicitly assessed throughout) Response is well-organized and easy to understand. |

## Question 8

### Question:

Plant scientists are investigating how the rate of photosynthesis (Y, measured in mols) in a newly developed rice variety responds to different light intensities (X, measured in lux). The experiments cover a wide range of light intensities, from very low light (e.g., 50 lux) up to moderately high levels (e.g., 15,000 lux), as rice photosynthesis tends to saturate. They found that the model fits the observed data extremely well across this broad spectrum of light. This model effectively captures the initial steep rise in photosynthesis at low light levels and the subsequent diminishing returns as light intensity increases further, before other factors become limiting.

Suppose we want to predict the photosynthetic rate for this rice variety at a light intensity of 7,500 lux. Show your work and explained your answer while mention any assumptions you are making.

A fake R output for this model might look like:

# Model: lm(formula = photosynthetic\_rate ~ log10(light\_intensity\_lux), data = rice\_photosynthesis\_study)  
# Coefficients:  
# (Intercept) log10(light\_intensity\_lux)  
# -8.500 7.200  
# ---  
# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
# ---  
# Residual standard error: 1.85 on 78 degrees of freedom  
# Multiple R-squared: 0.844, Adjusted R-squared: 0.842  
# F-statistic: 423.2 on 1 and 78 DF, p-value: < 2.2e-16

### Solution:

The model relates photosynthetic rate () to the base-10 logarithm of light intensity (). Based on the R output, the estimated linear model is:

where: - is the predicted photosynthetic rate - is the light intensity in lux - is the intercept coefficient - is the coefficient for

From the provided R output: - Estimated Intercept (): -8.500 - Estimated coefficient for (): 7.200

We want to predict the photosynthetic rate when the light intensity () is 7,500 lux.

First, calculate the base-10 logarithm of 7,500:

Now, substitute this value and the estimated coefficients into the model equation:

Therefore, the predicted photosynthetic rate at a light intensity of 7,500 lux is approximately 19.400 mols.

**Explanation:** The model uses the logarithm of light intensity because the relationship between light and photosynthesis is often non-linear, exhibiting diminishing returns at higher intensities. Taking the log can linearize this relationship. By plugging the log of the target light intensity (7,500 lux) into the estimated linear equation, we obtain the predicted photosynthetic rate based on the relationship learned from the data.

**Assumption:** A key assumption being made is that the linear relationship between the photosynthetic rate and the log of light intensity, as modeled, holds true and is generalizable to a light intensity of 7,500 lux, which falls within or is close to the range of light intensities the model was trained on (50 to 15,000 lux). We also assume that other factors not included in this simple model (like CO2 concentration, temperature, etc.) are not limiting or are held constant in a way consistent with the data used to build the model.

### Rubric:

| Criteria | Points | Descriptors |
| --- | --- | --- |
| **Identification of Model Structure** | 2 | Correctly identifies the model as a linear relationship between photosynthetic rate and the log10 of light intensity (). |
| **Extraction of Coefficients** | 2 | Correctly extracts both the intercept and the log10 coefficient values from the provided R output. |
| **Calculation of (7500)** | 3 | Correctly calculates or approximates the base-10 logarithm of 7,500 lux. |
| **Setting up and Performing Prediction Calculation** | 5 | Correctly substitutes the calculated log value and the coefficients into the model equation and performs the calculation to find the predicted photosynthetic rate. |
| **Showing Work** | 1 | Clearly shows the steps involved in the calculation process. |
| **Explanation of Prediction** | 1 | Provides a clear explanation of how the prediction is made and what the result represents. |
| **Stating Relevant Assumption(s)** | 1 | Mentions at least one relevant assumption made when performing the prediction (e.g., model validity within the range, other factors constant). |

## Question 9

### Question:

Based on the previous model, the slope coefficient for log(light\_intensity\_lux) is 7.2. What does this coefficient tell you about how photosynthetic rate changes with increasing light intensity? Explain using interpretation of a log transformation in context.

**Total Points:** 15

### Solution:

The model is given by: Photosynthetic Rate

From the previous question, the estimated slope coefficient () for is 7.200.

The slope coefficient in a regression model with a log-transformed predictor tells us about the change in the response variable associated with a *unit increase in the log-transformed predictor*. In the case of a transformation, a one-unit increase in corresponds to a tenfold increase in the original variable . This is because if , then .

Therefore, the interpretation of the slope coefficient 7.200 in this context is:

For every tenfold increase in light intensity (measured in lux), the photosynthetic rate (measured in mols) is estimated to increase by 7.200 units, holding other factors constant (although no other factors are in this simple model).

**Explanation using context:** The log transformation is used here likely because the relationship between light intensity and photosynthetic rate is not linear in its original scale; it typically shows a rapid increase at low light levels and then plateaus or saturates at higher levels. The log transformation helps to linearize this relationship. The coefficient of 7.200 signifies the average change in photosynthetic rate for a proportional increase in light intensity (specifically, a 10-fold increase). This means that increasing light intensity from, say, 100 lux to 1000 lux is associated with an estimated 7.200 mols increase in photosynthetic rate, just as increasing it from 500 lux to 5000 lux is associated with a similar estimated increase of 7.200 mols. This reflects the diminishing *absolute* return as light intensity increases, although the return per tenfold increase remains constant in this log-linear model.

### Rubric:

| Criteria | Points | Descriptors |
| --- | --- | --- |
| **Identification of Slope Coefficient** | 2 | Correctly identifies the slope coefficient value for as 7.2 or 7.200. |
| **Understanding of Interpretation** | 5 | Explains the general principle of interpreting a slope when the predictor is transformed – relates a unit change in the log scale to a tenfold change in the original scale. |
| **Contextual Interpretation (Photosynthetic Rate & Light Intensity)** | 4 | Applies the interpretation specifically to the variables in the problem (photosynthetic rate and light intensity). |
| **Using the Specific Coefficient Value** | 3 | Includes the specific value of the coefficient (7.200) in the interpretation statement. |
| **Explanation of Log Transformation Role** | 1 | Briefly explains why a log transformation might be used in this context (e.g., to handle non-linearity or diminishing returns) or clarifies the meaning of the coefficient in the context of the non-linear relationship. |

## Question 10

### Question

A market analyst is studying the relationship between the number of unique monthly visitors to e-commerce websites (X) and their total monthly sales revenue (Y, in US dollars). The dataset includes a diverse range of websites, from small niche stores with a few thousand visitors (e.g., 1,000) to larger platforms attracting up to 1,000,000 or more visitors per month. The analyst found that a log-log model that provides a very strong and interpretable fit. This model suggests that a percentage change in website traffic tends to correspond to a relatively consistent percentage change in sales revenue across websites of different sizes.

Suppose we want to predict the monthly sales revenue for a website that receives 90,000 unique monthly visitors. Show your work and explained your answer while mention any assumptions you are making.

A fake R output for this model might look like:

# Model: lm(formula = log10(monthly\_sales\_usd) ~ log10(unique\_monthly\_visitors), data = ecommerce\_traffic\_sales)  
# Coefficients:  
# (Intercept) log10(unique\_monthly\_visitors)  
# 0.1500 1.1200  
# ---  
# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
# ---  
# Residual standard error: 0.18 on 148 degrees of freedom  
# Multiple R-squared: 0.950, Adjusted R-squared: 0.950  
# F-statistic: 2844 on 1 and 148 DF, p-value: < 2.2e-16

**Total Points:** 15

### Solution:

The model is , where is sales revenue and is unique visitors. From the R output, the estimated coefficients are and .

We need to predict for .

1. Calculate :
2. Calculate the predicted :
3. Exponentiate to find the predicted :

The predicted monthly sales revenue for a website with 90,000 unique monthly visitors is approximately $500,000. The log-log model indicates a proportional relationship; here, a 1% increase in visitors is associated with an estimated 1.12% increase in sales.

**Assumption:** The model is valid for predicting at 90,000 visitors, and other factors influencing sales are consistent with the model’s data.

### Rubric:

| Criteria | Points | Descriptors |
| --- | --- | --- |
| **Identify Model & Coefficients** | 4 | Correctly identifies the log-log model structure () and extracts the intercept (0.1500) and slope (1.1200) coefficients. |
| **Calculate (90000)** | 3 | Correctly calculates or approximates . |
| **Predict (Y)** | 4 | Correctly sets up and calculates the predicted value of using the extracted coefficients and the calculated log visitor count. |
| **Exponentiate to find Predicted Y** | 3 | Correctly performs the base-10 exponentiation to convert the predicted back to the original sales revenue scale (). |
| **Show Work & State Assumption** | 1 | Clearly shows the main calculation steps and states a relevant assumption for the prediction. |