(1) (1pt) The variables tv and doctor are skewed, but life is not.

(2) (1pt) Not necessary. A point can be an outlier without being influential if its leverage is low.

(3) (1pt) There seems to be a marked mean curvature in the residual plot, suggesting that the relationship between the response and the predictors might be non-linear (note that Model_1 has the predictors in the model in a linear way). Therefore, we may want to consider a different model, i.e., one that allows for curvature.
Note that it is not possible to determine constancy of the variance because of the bunching of the fitted values at one end (see the residual plot). Even if you could, it would be less important.

(4) (1pt) σ: square root of the error variance.

(5) (1pt) The findings given in the answer to problem (1). Notice that when a predictor is seriously skewed and there exists a strong linear relationship between the predictor and the response, the response must be skewed as well. But this is not the case here (life is not skewed). We therefore expect that a transformed predictor with skewness removed (i.e., more symmetric) can do a better job in explaining this response. Log transformation is a common tool to "cure" skewness.

(6) (1pt) Yes, because (i) with the same number of parameters, the R^2 is higher and the residual standard error is lower (a valid comparison since the responses in the two models are the same), (ii) the two effects are much significant, and (iii) the residual plot in Figure 3 represents a satisfactory pattern.

(7) (1 pt) Yes - the residual plot is satisfactory. When the mean structure of the underlying system is non-linear and complex, a linear approximation like Model_1 over a relatively wide range of predictors may be inadequate. We sometimes can find suitable *transformations of data* that permit a non-linear model to be better approximated (after transformation) by a linear one like Model_2.

(8) (1pt) Because log(population/number-of-TVs) $= -\log$(number-of-TVs/population), i.e., the new TV is a location-and-scale change of the old TV by multiplying -1, the coefficient would be $+2.9156$.

(9) (1pt) $\log(\text{population}/(2 \times \text{number-of-TVs})) = \log(\text{population}/\text{number-of-TVs}) - \log(2)$ so difference is $-\log(2) \times -2.9156 = 2.0209$ *more* years of life in A than B.

(10) (1pt) MCAR or MAR.

(11) (1pt) More negative. Errors in predictors tend to bias the size of the effect toward zero, so removing the errors would increase it.

To better explain problems (12)-(15), let y_{ij}, where $j = 1, \ldots, n_i$, $i = 1, \ldots, 12$, be the crawling age of the jth baby in the ith birth month. Let $\bar{y}_i$, $i = 1, \ldots, 12$, be the average crawling age of the babies born in the ith month. Let $N = n_1 + \cdots + n_{12}$. Note that $\bar{y}_i$'s are the response data used in the analysis, rather than the y_{ij}'s.

(12) (2pts) As we assume that $y_{i1}, y_{i2}, \ldots, y_{in_i}$ are independent and have same variance σ_i^2 (notice that the SD varies from case to case), the variance of the response $\bar{y}_i$ is σ_i^2/n_i. Weights should be proportional to $\mathsf{n}/(\mathsf{SD}^2)$.

(13) (2pts) The problem states that for each i, the $y_{i1}, y_{i2}, \ldots, y_{in_i}$ are from a skewed distribution. The response $\bar{y}_i$ is the average of $y_{i1}, y_{i2}, \ldots, y_{in_i}$, where n_i is at least 21, so that by central limit theorem, the effect of skewness in $\bar{y}_i$ is reduced and $\bar{y}_i$ is more "normal".

Simply saying the sample size is large means non-normality can be ignored (it is true that when sample size is large $\hat{\beta}$ would tend to normality even if the response is not normal) is wrong because the sample size for the regression itself is only 12.

Saying the non-normality can be ignored because the number of babies included in the study (i.e., N=431=215+216) is large is not precise enough. For example, if $n_1 = 409$, and $n_2 = \cdots = n_{12} = 2$, we still have a large $N = 431$. But, $\bar{y}_2, \ldots, \bar{y}_{12}$ would be skewed and would not be like normal.

(14) (2pts) The question asks whether $\bar{y}_i$'s would be correlated due to the correlation between y_{ij}'s. For y_{ij}'s, it is reasonable to expect that the crawling ages of babies from the *same family* are correlated, and measurements on babies from different families are not. Suppose that a family had several babies involved in the study and these babies were born in different months. It would cause $\bar{y}_i$'s to be correlated.

Twins are correlated. But twins are born in the same month so they do not cause correlation between $\bar{y}_i$'s.

Notice that we cannot determine whether $\bar{y}_i$'s are correlated by using SD. The SD are values that appear in the diagonal part of the covariance matrix of $\bar{y}_i$'s, but information about correlation is in the off-diagonal part of the covariance matrix.

Although data are collected over time, there is no good reason to expect correlation between measurements from different families who come in succession. Time serial correlation is only plausible when the same process is being measured over time.

(15) (2pts) According to the statement of the problem, the gender of babies is an important predictor. On the other hand, no association between gender and birth month indicates that the (sample) correlation between gender and temperature is very low. When an important predictor is not included in the model, its effect, denoted by β_2, is divided into two parts, i.e., $(X_1^T X_1)^{-1} X_1^T X_2 \beta_2$ and $I - (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$ as shown in LNp.7-2 (for the current case, X_1 is the model matrix of the model $\mathsf{CrawlingAge} \sim \mathsf{temperature}$). For gender, the first part is very small (close to zero), but the second part would result in a larger estimate of σ^2. When gender is included in the model, the standard error for the temperature effect would tend to be reduced because of a smaller $\hat{\sigma}^2$ obtained in the new model. Because of the lack of association between gender and temperature, we would not expect any significant change in the size of the temperature effect estimate. The t-test for temperature would become more significant.

(16) (2pts) Yes. Figure 4 suggests that the response might have a heavy right tail (although Figure 4 only presents the shape of the marginal distribution of the response, rather than the conditional distribution of the response given the predictors). This is further confirmed in Figure 5 because the residuals are heavily skewed to the right. When the (conditional) distribution of response has a heavy tail, ordinary least square estimator would often be significantly biased by outliers. But a robust estimator, which is designed to resist the influence of outliers, can substantially reduce the bias.

(17) (2pts) (i) In the histogram (Figure 4), the response are "bunched" near zero, but, in markedly decreasing numbers, large responses do occur. (ii) The residual plot (Figure 5) shows a pattern of $\hat{\sigma} \propto \hat{y}^2$, especially on those positive residuals. Both suggest an inverse transformation.
FYI. The logarithm of incomes in general are roughly normal distributed. But the Box-Cox method did not identify logarithm as an adequate transformation for this data. The data came from billionaires so they represent the *extreme* right tail of the usual income distribution. The tail of a normal (say, the distribution of $Y|Y > 2$ where $Y \sim N(0,1)$) is skewed, which explains why we need a *stronger* transform than the log transformation for income. Log is not a good transformation for this data judging also by the pattern in the residual plot not being $\hat{\sigma} \propto \hat{y}$. Therefore, the result here (i.e., inverse transformation) is *consistent* with incomes in general being lognormal.

(18) (1pt) Yes. For example, if $R^2 = 0$, the adjusted R^2 will be negative. If R^2 is close to zero and the number of parameters in β is large, the adjusted R^2 would very likely be negative.

(19) (1pt) $\mathsf{wealth}^{-1} = 1.0453 - 0.0077 \times \mathsf{age}$, where $1.0453 = 0.4105 + 0.6348$ and $-0.0077 = 0.0015 - 0.0092$.

(20) (1pt) There is no reason to expect Cook's distances to follow a half-normal distribution (they can follow an F-distribution when some assumptions hold) so we do not care that the points follow a curve.
Some might say there might be some influential points - maybe so, although probably not, but that was not the question asked.

(21) (2pts) Let ω and Ω represent Model_5 and Model_4 respectively. Then, $RSS_\omega = 14.3391$, $RSS_\Omega = 0.2568^2 \times 215 = 14.1784$, $df_\omega = 221$, $df_\Omega = 215$. So the F-statistic is

$$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega} = \frac{(14.3391 - 14.1784)/6}{14.1784/215} = 0.4061.$$

(22) (1pt) They use different null and alternative models in testing $\beta_{\mathsf{regionM}} = 0$.
For the test with p-value=0.94065, the null model is

$$\omega : \mathsf{wealth}^{-1} \sim \mathsf{age},$$

and the alternative model is

$$\Omega : \mathsf{wealth}^{-1} \sim \mathsf{age} + \mathsf{regionM}.$$

For the test with p-value=0.01337, the null model is

$$\omega : \mathsf{wealth}^{-1} \sim \mathsf{age} + \mathsf{age{:}regionM},$$

and the alternative model is

$$\Omega : \mathsf{wealth}^{-1} \sim \mathsf{age} + \mathsf{regionM} + \mathsf{age{:}regionM}.$$

(23) (2pts) Containing regionM or not is the only distinction between the two simplified models. In other words, we need to decide whether regionM should be included into model when age:regionM is already in model. We should choose $\mathsf{wealth}^{-1} \sim$ regionM+age:regionM, the one based on regression output, for at least two reasons. First, regression output tested regionM when age:regionM is already in model as shown in the answer to problem (22) but ANOVA did not. On the other hand, the insignificance of regionM in ANOVA is not reliable because

it is probably due to the bias in the coefficient estimate of regionM, which occurs when some important effects (i.e., age:regionM in the case) are not included in the model (i.e., the Ω: $\mathsf{wealth}^{-1} \sim \mathsf{age} + \mathsf{regionM}$ in the answer to problem(22)).
Notice that if we change the order of effects entering ANOVA model to be

$$\text{anova}(\mathsf{wealth}^{-1} \sim \mathsf{age} + \mathsf{age{:}regionM} + \mathsf{regionM}) \ ,$$

where regionM is placed as the last one, then the test for regionM in this ANOVA is exactly the one in the regression output. Usually, we would expect an effect to become less significant when its order in ANOVA is moved backward. But this did not happen on regionM in this case. The reason is as stated above.

(24) (1pt) The average response (i.e., wealth^{-1}) of non-Middle-East billionaires.

(25) (1pt) $(5.18e-04)/\sqrt{7.1348} = 0.0001939273$.

(26) (1pt) All the predictors were measured in the same unit: mg/liter (in contrast to the Longley dataset demonstrated in the Lab, where the variables were had different units and so rescaling would be more suitable). Here, it is most reasonable to consider principal components on the unscaled data to maintain the comparable units.

(27) (2pts) Because of using covariance matrix in constructing PCs, the sum of the sample variances is the sum of eigenvalues:

$$(5.04e+06) + (6.51e+05) + (1.79e+04) + (2.17e+03) + (1.84e+01) = 5711088.$$

The predictor with the smallest sample variance is x4 because (i) $\mathsf{x4} \approx$ PC5 (which can be found from the coefficients of the eigenvector corresponding to PC5) and (ii) PC5 has the smallest eigenvalue.

(28) (1pt) Because the model spaces (i.e., the space spanned by the columns of model matrix) of Model_6 and Model_7 are identical, we can use the information offered in "Sum Sq" column of the ANOVA table under Model_7 to calculate the R^2 of Model_6. The answer is

$$\frac{0.02+0.07+0.03+0.00+3.98}{0.02+0.07+0.03+0.00+3.98+0.96} = 0.8102767 \approx 0.811.$$

(29) (2pts) The best sub-model is y $\sim$ PC1 + PC4. When considering only those sub-models with two (or an equal number of) PCs, we know that maximizing adjusted-R^2 is equivalent to maximizing R^2. Because of the orthogonality existing between PCs, the R^2 of a two-PCs sub-model would be proportional to the sum of the two values corresponding to the two PCs in the "Sum Sq" column of ANOVA. Because 3.98 and 0.07 are the first two largest values (or using their corresponding F-values or p-values to choose) in the column, we get the answer.
Note that if the predictors are not orthogonal, the above method won't work any more for the job of identifying the sub-model that maximizes R^2.

(30) (1pt) The model y$\sim$z would have a good fit because (i) z is almost a linear transformation of PC1 (check the coefficients of the eigenvector corresponding to PC1) and (ii) PC1 explains a much larger variation of y than the other PCs as shown in the ANOVA table.
Actually, we can roughly obtain the R^2 of the model y$\sim$z by:

$$3.98/(0.02+0.07+0.03+0.00+3.98+0.96) = 0.7865613,$$

which is very close to 0.8111. It is therefore a simple and easy-to-interpret model with a good fit.