(1, 1pt) The number of total degree of freedom is 1+4+1+4+90=100. There are $100/(5\times 2) = 10$ replicates for each level combinations.

(2, 2pts) The p-value stays the same because all level combinations have an equal number of replicates. Under the circumstance, using Helmert codings would bring orthogonality between the spaces spanned by the codings of Age, Breed, and Breed:Age. The change of order would not influence the p-values when orthogonality exists. Furthermore, the ANOVA table is irrelevant to the choice of codings. So, no matter what codings we use, the order of testing makes no difference on the p-values.

(3, 2pts) The residuals vs. fitted-value plot can be used to check for (i) non-constant variance, and (ii) curvature in the mean of residuals. For this case, the is no need to worry about (ii) because the averages of the residuals in each of 10 groups are zero (check the answer to problem 4). For (i), Figure 1 reveals that the variance increases with the mean response. (That the response is later transformed is also a hint that such a feature might be present).

(4, 1pt) The averages of the residuals in each groups are exact zero because the fitted model is saturated (the data has 10 distinct level combinations and the model has 10 parameters in β). The $\hat{y}$-values are the averages of y's in each groups of distinct level combinations under a saturated model.

(5, 2pts) The lack-of-fit test is an F-test that compares a simpler model (called null model ω) to the saturated model (called alternative model Ω). In this case, the saturated model is the model used to produce the ANOVA output (check the answer to problem 4). So, ω: Butterfat $\sim$ Breed and Ω: Butterfat $\sim$ Breed + Age + Breed:Age. Because this ANOVA is sequential, the F-statistic of the lack-of-fit test is

$$\frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega} = \frac{(0.0000274 + 0.0000514)/(1+4)}{0.0015580/90} = 0.9103979.$$

(6, 2pts) $-1/2$, -1, and -2 are inside the 95% confidence interval for lambda. For the variable Butterfat, 1/Butterfat (the amount of milk containing 1 unit of butter fat) is more interpretable than $1/\text{Butterfat}^{1/2}$ or $1/\text{Butterfat}^2$, so I choose the reciprocal transformation.

Because 1 does not fall in the confidence interval, a transformation of the response is necessary.

(7, 1pt) Because R_a^2 $(= 1 - (RSS/(n-p))/(TSS/(n-1)))$ is a function of R^2 $(=1 - RSS/TSS)$ and R^2 is a function of the overall F-statistic $(=((TSS - RSS)/(p-1))/(RSS/(n-p)))$, we can obtain R_a^2 from the overall F as follows:

$$R_a^2 = 1 - \left[\left(F \times \frac{p-1}{n-p} + 1\right) \times \frac{n-p}{n-1}\right]^{-1} = 1 - \left[\left(61.85 \times \frac{4}{95} + 1\right) \times \frac{95}{99}\right]^{-1} = 0.7108645$$

(8, 1pt) In the regression output, the codings for Breed is the treatment codings, and the level Ayrshire is treated as the reference level. So, -2.0592 estimates the mean difference between Canadian and Ayrshire cows.

(9, 1pt) Because I used reciprocal transformation for Butterfat, the estimated mean butterfat content (note: not the mean response of the regression fit for 1/Butterfat) for Jersey cows is

$$\frac{1}{24.7298 - 5.6092} = 5.229961\%.$$

(10, 2pts) A formal method to perform this t-test is as follows. Denote the estimates of BreedCanadian and BreedGuernsey in the regression output by $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. Then, the standard error of $\hat{\beta}_1 - \hat{\beta}_2$ is

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\hat{var}(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{\hat{var}(\hat{\beta}_1) + \hat{var}(\hat{\beta}_2) - 2 \times \hat{cov}(\hat{\beta}_1, \hat{\beta}_2)}$$

and the test statistic is $|\hat{\beta}_1 - \hat{\beta}_2|/se(\hat{\beta}_1 - \hat{\beta}_2)$. Unfortunately, although we can obtain the values of $\hat{\beta}_1 - \hat{\beta}_2$, $\hat{var}(\hat{\beta}_1)$ $(= 0.5963^2)$ and $\hat{var}(\hat{\beta}_2)$ from the regression output, we do not have $\hat{cov}(\hat{\beta}_1, \hat{\beta}_2)$ in the output.

A trick to resolve this problem is to change the reference level of treatment coding from Ayrshire to Canadian. After doing so, the coefficient of the term BreedGuernsey would be the difference of mean response between Canadian and Guernsey cows. The estimate of this coefficient is $(-4.3417) - (-2.0592) = -2.2825$ with a std. Error being 0.5963. The std. Error does not change because each level combinations were performed an equal number of times so that no matter what level is treated as the reference level, it has no impact on the value of the std. Error. The t-value is $-2.2825/0.5963 = -3.827771$, which clearly indicates a significant difference.

(11, 2pts) The standard error is inversely proportional to the sqrt(sample size) so going from 10 to 4 observations would increase the standard error by a factor of sqrt(10/4)=1.581139. The new standard error for the effect would be around $0.5963 \times 1.581139 = 0.9428331$ and the corresponding t-statistic would be $-4.3417/0.9428331 = -4.604951$ which is still clearly significant. So reducing the sample size to 4 replicates would still allow enough power to detect an effect of the current amount.

(12, 1pt) Five observations were collected repeatedly on each cow. They are repeated measures (check lecture note p.6-1 for its definition). We might suspect the observations from same cow are correlated.

(13, 1.5pts) Robust estimators are only worth considering when the errors are not normal. Without normality (check the statistical modeling in LNp.10-1), the F-statistic can be calculated but not easily evaluated for significance because its null distribution is not an F-distribution any more.

Furthermore, the geometric property under OLS that $\hat{y}$ is orthogonal to $\hat{\epsilon}$ does not hold for the robust estimator $\hat{\beta}$ because $\hat{y} = X\hat{\beta}$ is not the orthogonal projection of y onto the space spanned by the columns of X. Notice that this property is required in the calculation of R^2 and in proving that the null distribution of an F-test is an F-distribution.

FYI, the t-values in the output do not follow a t-distribution under the null because of the lack of normality. However, their asymptotic distributions still follow a t-distribution.

(14, 1pt) No correlation assumption because the data were observed over time. We might suspect that their errors have temporal correlation.

(15, 1pt) It means gas usage changes with temperature at a different rate before and after the installation of insulation (the slopes of the two groups are different).

(16, 1pt) At a temperature of zero degrees, the gas usage was 2.2632 units higher before compared to after the installation of insulation. Note that we do need to say zero degrees because the difference between the two groups varies according to temperature because of the significant interaction Insulate:Temp.

(17, 1pt) Suppose that the fitted lines are $a_0+a_1\times Temp$ and $b_0+b_1\times Temp$ for Before and After groups respectively. The saved gas consumption is $(a_0-a_1)+(b_0-b_1)\times Temp$. So, the gas consumption is $2.26321 + (-0.14361) \times 4 = 1.68877$ units higher usage before compared to after.

(18, 2pts) The Cook statistics are used to detect influential points so we would look for values far greater than the rest. If such a point or points are identified, we should exclude them and see the effect on the fit.

No practical consequence. We do not need the Cook statistics to follow a half-normal distribution (i.e., do not care whether the points follow a straight line) - we only care about the (few) points that are unusually large based on the trend observed in the bulk of all the points.

(19, 2pts) Under-estimate for this case. It is because the mean outside temperature were generally higher in the before group than the after group (i.e., there exists a negative collinearity between Temp and Insulate). So, the unadjusted mean usage difference between the two groups underestimates the true difference because gas usage is less if the outdoor temperature is higher (i.e., the estimated coefficient of Temp is negative). Thus, including temperature removes a bias - it may also increase the precision of the comparison but the question is about the underestimation.

(20, 1pt) Examining whether Insulate is significant in the models, "|Residuals| $\sim$ Insulate" or "|Residuals| $\sim$ Insulate + Temp + Insulate:Temp", where $|\cdot|$ stands for taking absolute value, would do it. Bartlett's test or Levene's test is on the right track but you can only apply it on the model "Residuals $\sim$ Insulate", not "Residuals $\sim$ Insulate + Temp + Insulate:Temp" because the former has replicates but the latter does not (or too few replicates, check Figure 3). Actually, none of these tests are significant (might be due to insufficient sample size) although Figure 5 shows a pattern of non-constant variance. Notice that applying Bartlett's test or Levene's test on the model "Gas $\sim$ Insulate" is not suitable for examing whether the error variances in Figure 5 are identical. Can you see why?

(21, 1pt) The following are acceptable answers:

- MAR, assuming that missing probabilities are identical within each group, so that although the missing probabilities vary with the variable Insulate, it is still an informative missing when Insulate is included in the model.
- MNAR, assuming that the missing probabilities for the cases in a group (Before or After) vary and depend on some variable(s) that were not observed.

(22, 1pt) The predictors are likely to be collinear.

(23, 1pt) No. Every values in Distance is an average of 10 records, which makes the weight matrix in WLS proportional to the identity matrix. So, WLS and OLS generate same estimates in this case.

(24, 2pts) Yes, at least it can be used to determine a model with similar performance as the final fitted model of backward selection (BS). Normally, just the ANOVA table would be insufficient. But, it so happens that the predictors have been ordered so that the p-values increase down the table. This means that the nested series of models represented in the table probably correspond exactly to those chosen in the backward elimination process. This claim is based on the following property: if we use the reverse order of removal in a BS to perform a sequential ANOVA analysis (first removing in BS, last adding in ANOVA), then (i) the resulting ANOVA table would generally have such an increasing pattern on p-values, and (ii) the null (ω) and alternative (Ω) models of the test for a predictor in the ANOVA are identical to the ones of the test in BS that determines the removal of this predictor. Furthermore, the differences between any pairs of adjacent p-values are quite large. Therefore, we are confident (but, of course, not 100% sure) that the final fitted model is "Distance $\sim$ RStr + OStr".

(25, 1pt) An average index of (standardized) leg strength and flexibility because all five coefficients for PC1 are about the same size.

(26, 1.5pts) Because the X^TX matrix of the standardized predictors is proportional to the correlation matrix of these predictors, the two matrices have the same condition number. The trace of the correlation matrix, which is also the sum of its eigenvalues, is the number of predictors, 5. So, the largest eigenvalue is $5 - (0.7599 + 0.5217 + 0.1183 + 0.0797) = 3.5204$ and the condition number is sqrt$(3.5204/0.0797) = 6.6461$.

(27, 1.5pts) The confidence region is an ellipse with the major and minor axes parallel to the vertical and horizontal axes. The axis corresponding to PC1 is the minor axis and the one corresponding to PC2 is the major one because PC1 explains more variation in X than PC2.

(28, 1.5pts) LFlex. The property that λ=0 corresponds to ordinary least squares (OLS) makes it possible to identify this predictor. Although the predictors in the ridge regression had been standardized (i.e., different from what used in the regression and ANOVA outputs), note that standardizing a predictor would not change the signs of its OLS estimates. In Figure 6, the bottom line converges to zero the fastest and its OLS estimate (at λ=0) is negative. The only negative estimate in the regression output is LFlex.

(29, 1pt) The ridge estimators typically have lower variances (i.e., an advantage) than the OLS estimators, but the former have more bias (a disadvantage) than the latter. When the predictors have strong collinearity as in this case, the mean square error (=variance + bias^2) is typically lower for ridge estimators.