

BIOS663 Homework 1
Due Wednesday, Feb 6 in class

To report a test, provide H_0 , the test statistic, the degrees of freedom, the p -value, the decision (accept vs. reject H_0), and an interpretation of the decision in terms of the subject matter.

1. (a) Prove or dis-prove (with details) that

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & -6 \\ 4 & 1 & 7 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & 0 \\ 4 & 1 & -2 \end{bmatrix}$$

have linearly independent columns, respectively.

- (b) Find the eigenvalues and eigenvectors of

$$\mathbf{C} = \begin{bmatrix} 2 & 1 \\ 2 & 4 \end{bmatrix}$$

2. Suppose

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim N(0, \mathbf{\Sigma}) \text{ where } \mathbf{\Sigma} = \begin{bmatrix} 2 & 0 & 0.6 \\ 0 & 2 & 0.5 \\ 0.6 & 0.5 & 1 \end{bmatrix}$$

- (a) Derive the distribution of $3x_1 + x_2 + x_3$.
 (b) Derive the distribution of $(x_1, x_2 \mid x_3 = 3)$.
 (c) Calculate $Cov(x_1 + 2x_2, 3x_2 + x_3)$.
3. Suppose X_1, \dots, X_k are multivariate normally distributed with $X_i \sim N_n(\mu_i, \Sigma_i)$, $i = 1, \dots, k$. Further, let $Cov(X_i, X_j) = \Sigma_{ij}$ ($i \neq j$). Suppose a_1, \dots, a_k are scalars and define $Y = a_1X_1 + \dots + a_kX_k$. Find the distribution of Y .
4. *Weighted least squares* is a modification of standard regression analysis that may be used for a set of data when the assumption of variance homogeneity does not hold. (Assume the responses are independent.) If the i th response is an average of m_i equally variable observations, then $\text{Var}(y_i) = \frac{\sigma^2}{m_i}$. In this case, we have the model $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$, where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}$, and

$$\mathbf{V} = \begin{bmatrix} \frac{1}{m_1} & 0 & \dots & 0 \\ 0 & \frac{1}{m_2} & & 0 \\ \vdots & & \ddots & \\ 0 & \dots & & \frac{1}{m_n} \end{bmatrix}.$$

The fixed and known positive definite matrix $\mathbf{V}_{n \times n}$ has rank n . The weighted least squares estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_W = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

- (a) Derive the expectation of $\hat{\beta}_W$, $E[\hat{\beta}_W]$.
- (b) Derive the covariance matrix of $\hat{\beta}_W$, $\text{Cov}(\hat{\beta}_W)$.
- (c) Find the exact distribution of $\hat{\beta}_W$. If it is necessary to make any reasonable further assumptions in order to find the distribution of $\hat{\beta}_W$, provide them.
- (d) Explain why this particular choice of \mathbf{V} makes sense when our responses are averages.

Solution to HW 1

1. a. $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & -6 \\ 4 & 1 & 7 \end{bmatrix}$

solve for $c_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 4 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 8 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ 2 \\ -6 \\ 7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$

$$\Rightarrow \begin{cases} c_1 + c_2 + c_3 = 0 \\ c_1 = -2c_3 \end{cases} \Rightarrow c_2 = c_3 = -\frac{1}{2}c_1$$

and this could solve

$$\begin{cases} c_1 + 8c_2 - 6c_3 = 0 \\ 4c_1 + c_2 + 7c_3 = 0 \end{cases}$$

\therefore Not linearly independent

actually, $A \cdot \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & 0 \\ 4 & 1 & -2 \end{bmatrix}$$

solve for $c_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 4 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 8 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ 2 \\ 0 \\ -2 \end{pmatrix} = 0$

first two equations

$$\Rightarrow c_2 = c_3 = -\frac{1}{2}c_1$$

plug into the 3rd & 4th equations

$$\Rightarrow 3c_1 = 0 \Rightarrow c_1 = c_2 = c_3 = 0$$

\Rightarrow linearly independent

1. b solve

$$\left| \begin{bmatrix} 2-\lambda & 1 \\ 2 & 4-\lambda \end{bmatrix} \right| = 0$$

$$\Rightarrow \lambda = 3 \pm \sqrt{3}$$

for $\lambda_1 = 3 + \sqrt{3}$ solve v_1 s.t. $\begin{bmatrix} 2-(3+\sqrt{3}) & 1 \\ 2 & 4-(3+\sqrt{3}) \end{bmatrix} v_1 = 0$

$$\Rightarrow v_1 = c \begin{pmatrix} \sqrt{3}-1 \\ 2 \end{pmatrix}$$

similarly, for $\lambda_2 = 3 - \sqrt{3} \Rightarrow v_2 = c \begin{pmatrix} -(\sqrt{3}+1) \\ 2 \end{pmatrix}$

~~1. c. $\text{Cov}(X_1 + 2X_2, 3X_2 + X_3)$~~

2. a. linear combination of ~~joint~~ ^{multi-} normal r.v.'s is still normal; what remains to derive is the mean & variance of that normal distribution

$$E[3X_1 + X_2 + X_3]$$

$$= 3E[X_1] + E[X_2] + E[X_3] = 0$$

$$\text{Var}[3X_1 + X_2 + X_3]$$

$$= 9\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)$$

$$+ 2 \cdot 3 \cdot \text{Cov}(X_1, X_2) + 2\text{Cov}(X_2, X_3) + 2 \cdot 3 \cdot \text{Cov}(X_1, X_3)$$

$$= 25.6$$

$$\Rightarrow 3X_1 + X_2 + X_3 \sim N(0, 25.6)$$

2.b. Define $y_1 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $y_2 = (x_3)$

Then $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$

where $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mu_2 = (0)$

$\Sigma_{11} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, $\Sigma_{12} = \begin{pmatrix} 0.6 \\ 0.5 \end{pmatrix}$, $\Sigma_{21} = \Sigma_{12}^T$, $\Sigma_{22} = (1)$

Using the formula

$y_1 \mid y_2 = b \sim N \left(\mu_1 + \Sigma_{12} \cdot \Sigma_{22}^{-1} (b - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right)$

with $b = 3$

$\Rightarrow y_1 \mid y_2 = 3 \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.6 \\ 0.5 \end{pmatrix} (1)^{-1} ((3) - (0)) , \right.$

$\left. \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0.6 \\ 0.5 \end{bmatrix} [1]^{-1} [0.6 \ 0.5] \right)$

$\Rightarrow y_1 \mid y_2 = 3 \sim N \left(\begin{pmatrix} 1.8 \\ 1.5 \end{pmatrix}, \begin{pmatrix} 1.64 & -0.30 \\ -0.30 & 1.75 \end{pmatrix} \right)$

2.c. $\text{Cov} (x_1 + 2x_2, 3x_2 + x_3)$

$= \text{Cov} (x_1, 3x_2) + \text{Cov} (2x_2, 3x_2) + \text{Cov} (2x_2, x_3) + \text{Cov} (x_1, x_3)$

$= 3 \cdot 0 + 6 \cdot 2 + 2 \cdot 0.5 + 0.6$

$= 13.6$

3. Any linear combination of multi-normal r.v.'s is still normal with the mean

$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^k a_i x_i\right] \\ &= \sum_{i=1}^k a_i E[x_i] \\ &= \sum_{i=1}^k a_i \mu_i \end{aligned}$$

and variance

$$\begin{aligned} \text{Var}[Y] &= \text{Var}\left[\sum_{i=1}^k a_i x_i\right] \\ &= \sum_{i=1}^k \text{Var}[a_i x_i] + \sum_{i=1}^k \sum_{j \neq i, j=1}^k \overset{\text{Cov}}{\cancel{\text{Var}}}[a_i x_i, a_j x_j] \\ &= \sum_{i=1}^k a_i^2 \text{Var}(x_i) + 2 \sum_{i=1}^k \sum_{j=1, j \neq i}^k a_i a_j \text{Cov}(x_i, x_j) \\ &= \sum_{i=1}^k a_i^2 \Sigma_{ii} + \sum_{i=1}^k \sum_{j=1, j \neq i}^k a_i a_j \Sigma_{ij} \end{aligned}$$

$$\therefore \sum_{i=1}^k a_i x_i \sim N\left(\sum_{i=1}^k a_i \mu_i + \sum_{i=1}^k a_i^2 \Sigma_{ii} + \sum_{i=1}^k \sum_{j=1, j \neq i}^k a_i a_j \Sigma_{ij}\right)$$

4. (a)

$$\begin{aligned} E[\hat{\beta}_w] &= E[(X^T V^{-1} X)^{-1} X^T V^{-1} Y] \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} E[Y] \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} X \beta \\ &= \beta \end{aligned} \quad \begin{aligned} &\searrow \text{because } E[Y] = E[X\beta + \varepsilon] \\ &= X\beta + E(\varepsilon) \\ &= X\beta \end{aligned}$$

(b) $\text{Cov}(\hat{\beta}_w)$

$$\begin{aligned} &= \text{Cov}((X^T V^{-1} X)^{-1} X^T V^{-1} Y) \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} \text{Cov}(Y) V^{-1} X (X^T V^{-1} X)^{-1} \\ &\quad (\because \text{Cov}(Y) = \text{Var}(X\beta + \varepsilon) = \text{Var}(\varepsilon) = \sigma^2 V) \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} (\sigma^2 V) V^{-1} X (X^T V^{-1} X)^{-1} \\ &= \sigma^2 (X^T V^{-1} X)^{-1} \end{aligned}$$

(c) If ε is multi-normal, then we immediately

know that $\hat{\beta}_w \sim N(\beta, \sigma^2 (X^T V^{-1} X)^{-1})$

Otherwise, only the mean & variance

don't give enough information on the exact

distribution of $\hat{\beta}_w$

(d) The variance of the average of m_i equally variable observations could be calculated as

$$\begin{aligned}\text{Var}(y_i) &= \text{Var}\left(\sum_{i=1}^{m_i} X_i / m_i\right) \\ &= \left(\frac{1}{m_i}\right)^2 \sum_{i=1}^{m_i} \sigma^2 \\ &= \frac{\sigma^2}{m_i}\end{aligned}$$

Therefore the covariance matrix $\text{Cov}(Y) = \sigma^2 V$

where $V = \begin{bmatrix} \frac{1}{m_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{m_n} \end{bmatrix}$

In other words

{ all off-diagonal entries = 0 because we assume the observations are independent
Diagonal terms are inverse-weighted by the number of observations as more observations mean more information on that $y_i \Rightarrow$ less variable

BIOS663 Homework 2
Due Wednesday, Feb 20 in class.

1. Consider a simple linear regression $Y = X\beta + \epsilon$ with an intercept and one predictor based on a sample of size 4. Or specifically,

$$\begin{pmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{pmatrix} + \epsilon. \quad (1)$$

Calculate $(X'X)^{-1}$, $X'Y$, $\hat{\beta}$, \hat{y} and $\hat{\epsilon}$ by hand.

2. Consider the model $\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \boldsymbol{\varepsilon}$. Give the appropriate \mathbf{C} and $\boldsymbol{\theta}_0$ for testing the following hypotheses.

- (a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$
 (b) $H_0 : \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_2 + 2 \\ \beta_4 \end{pmatrix}$
 (c) $H_0 : \begin{pmatrix} \beta_1 - 2\beta_2 \\ \beta_1 + 2\beta_2 \end{pmatrix} = \begin{pmatrix} 4\beta_3 \\ -6 \end{pmatrix}$

3. Consider the model $\mathbf{y}_{5 \times 1} = \mathbf{X}_{5 \times 3}\boldsymbol{\beta}_{3 \times 1} + \boldsymbol{\varepsilon}_{5 \times 1}$, where

$$\mathbf{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \mathbf{X} = [\mathbf{J} \quad \mathbf{x}_1 \quad \mathbf{x}_2] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},$$

with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

- (a) Show as rigorously as possible whether $\theta_1 = \beta_2$ is estimable.
 (b) Show as rigorously as possible whether $\boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.
4. A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights ($y = \text{WGHT}$), average daily exercise times ($x = \text{TIME}$). One of the objectives in this study is to investigate the effect of TIME on weight loss.
- (a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete the table.

Dependent Variable: WGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184			
Error	96	1827.099916			
Corrected Total	97	4451.7701			

- (b) State the model assumptions based on which the ANOVA table was computed.
- (c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.
5. An investigator studied the ozone levels in the South Coast Air Basin of California for the years 1976-1991. He believes that the number of days the ozone levels exceeded 0.20 ppm (the response) depends on the seasonal meteorological index, which is the seasonal average temperature in degrees Celsius (the predictor). The data *hw2.dat*, is provided on Sakai.
- (a) Fit a regression model with the number of high ozone days as the response and the meteorological index as a covariate, and provide estimates of β_0 , β_1 , their standard errors, and their interpretations.
- (b) Are all of the β 's estimable? Why or why not?
- (c) Report a test of the hypothesis that the number of high ozone days is associated with the meteorological index.
- (d) Using the framework of the linear model, report an $\alpha = 0.05$ test of the hypothesis that a 1 degree increase in average temperature is associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.
- (e) Calculate the 95% confidence interval and prediction interval for the expected number of days the ozone level exceeded 0.2 ppm when the seasonal meteorological index is 16.

- 1) Consider a simple linear regression $Y = X\beta + \epsilon$ with an intercept and one predictor based on a sample of size 4. Or specifically,

$$\begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} + \epsilon. \quad (1)$$

Calculate $(X'X)^{-1}$, $X'Y$, $\hat{\beta}$, \hat{y} , and $\hat{\epsilon}$ by hand.

$$Y = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1+1+1+1 & 1+1+0.5+2 \\ 1+1+0.5+2 & 1+1+0.25+4 \end{bmatrix} = \begin{bmatrix} 4 & 4.5 \\ 4.5 & 6.25 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{|4(6.25) - 4.5(4.5)|} \begin{bmatrix} 6.25 & -4.5 \\ -4.5 & 4 \end{bmatrix} = \frac{1}{4.75} \begin{bmatrix} 6.25 & -4.5 \\ -4.5 & 4 \end{bmatrix} = \begin{bmatrix} 25/19 & -18/19 \\ -18/19 & 16/19 \end{bmatrix} \\ \approx \begin{bmatrix} 1.3158 & -0.9474 \\ -0.9474 & 0.8421 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 2 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} = \begin{bmatrix} 0.5 - 0.5 + 0.3 + 1.2 \\ 0.5 - 0.5 + 0.5 * 0.3 + 2 * 1.2 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 2.55 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 25/19 & -18/19 \\ -18/19 & 16/19 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2.55 \end{bmatrix} = \begin{bmatrix} -42/95 \\ 69/95 \end{bmatrix} \approx \begin{bmatrix} -0.4421 \\ 0.7263 \end{bmatrix}$$

$$\hat{y} = X\hat{\beta} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -42/95 \\ 69/95 \end{bmatrix} = \begin{bmatrix} 27/95 \\ 27/95 \\ -3/38 \\ 96/95 \end{bmatrix} \approx \begin{bmatrix} 0.2842 \\ 0.2842 \\ -0.0789 \\ 1.0105 \end{bmatrix}$$

$$\hat{\epsilon} = (y - \hat{y}) = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} - \begin{bmatrix} 27/95 \\ 27/95 \\ -3/38 \\ 96/95 \end{bmatrix} = \begin{bmatrix} 41/190 \\ -149/190 \\ 36/95 \\ 18/95 \end{bmatrix} \approx \begin{bmatrix} 0.2158 \\ -0.7842 \\ 0.3789 \\ 0.1895 \end{bmatrix}$$

- 2) Consider the model $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \beta_4 \mathbf{x}_4 + \epsilon$. Give the appropriate \mathbf{C} and $\boldsymbol{\theta}_0$ for testing the following hypotheses.

$$(a) H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 \equiv \begin{matrix} \beta_1 - \beta_4 = 0 \\ \beta_2 - \beta_4 = 0 \\ \beta_3 - \beta_4 = 0 \end{matrix}$$

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0 \quad \Rightarrow \quad \begin{bmatrix} 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}_{3 \times 5} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}_{5 \times 1} = \begin{bmatrix} \beta_1 - \beta_4 \\ \beta_2 - \beta_4 \\ \beta_3 - \beta_4 \end{bmatrix}_{3 \times 1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3 \times 1}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}_{3 \times 5} \quad \boldsymbol{\theta}_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3 \times 1}$$

$$(b) H_0 : \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_2 + 2 \\ \beta_4 \end{pmatrix}$$

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0 \quad \Rightarrow \quad \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}_{2 \times 5} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}_{5 \times 1} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_3 - \beta_4 \end{bmatrix}_{2 \times 1} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}_{2 \times 1}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}_{2 \times 5} \quad \boldsymbol{\theta}_0 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}_{2 \times 1}$$

$$(c) H_0 : \begin{pmatrix} \beta_1 - 2\beta_2 \\ \beta_1 + 2\beta_2 \end{pmatrix} = \begin{pmatrix} 4\beta_3 \\ -6 \end{pmatrix}$$

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0 \quad \Rightarrow \quad \begin{bmatrix} 0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0 \end{bmatrix}_{2 \times 5} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}_{5 \times 1} = \begin{bmatrix} \beta_1 - 2\beta_2 - 4\beta_3 \\ \beta_1 + 2\beta_2 \end{bmatrix}_{2 \times 1} = \begin{bmatrix} 0 \\ -6 \end{bmatrix}_{2 \times 1}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0 \end{bmatrix}_{2 \times 5} \quad \boldsymbol{\theta}_0 = \begin{bmatrix} 0 \\ -6 \end{bmatrix}_{2 \times 1}$$

3) Consider the model $\mathbf{y}_{5 \times 1} = \mathbf{X}_{5 \times 3} \boldsymbol{\beta}_{3 \times 1} + \boldsymbol{\varepsilon}_{5 \times 1}$, where

$$\mathbf{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \mathbf{X} = [\mathbf{J} \quad \mathbf{x}_1 \quad \mathbf{x}_2] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \text{ with } \boldsymbol{\varepsilon} \sim N_5(\mathbf{0}, \sigma^2 \mathbf{I}).$$

(a) Show as rigorously as possible whether $\theta_1 = \beta_2$ is estimable.

\mathbf{X} is not full rank. $r(\mathbf{X}) = 2 < 3$

$$\mathbf{x}_2 = \mathbf{J} - \mathbf{x}_1$$

$\theta_1 = \beta_2$ is estimable if there exists a \mathbf{T} matrix for $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta} = \mathbf{T}\mathbf{E}(\mathbf{Y})$ such that $\mathbf{C} = \mathbf{T}\mathbf{X}$.

$$\theta_1 = \beta_2 \quad \Rightarrow \quad \boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta} \equiv \theta_1 = [0 \quad 0 \quad 1]\boldsymbol{\beta} = \beta_2$$

$$\Rightarrow \quad \mathbf{C} = \mathbf{T}\mathbf{X} \equiv [0 \quad 0 \quad 1] = [t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5] \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$\Rightarrow \quad [0 \quad 0 \quad 1] = [\sum_{i=1}^5 t_i \quad t_1 + t_2 + t_5 \quad -t_3 - t_4]$$

$$\Rightarrow \quad \begin{array}{l} t_1 + t_2 + t_3 + t_4 + t_5 = 0 \\ t_1 + t_2 + t_5 = 0 \\ -t_3 - t_4 = 1 \end{array} \Rightarrow \begin{array}{l} t_3 + t_4 = 0 \\ -t_3 - t_4 = 1 \equiv t_3 + t_4 = -1 \end{array}$$

Since $t_3 + t_4 = -1 \neq 0$, there is no \mathbf{T} that can satisfy the equation $\mathbf{C} = \mathbf{T}\mathbf{X}$.

$\therefore \theta_1 = \beta_2$ is not estimable

3) Consider the model $\mathbf{y}_{5 \times 1} = \mathbf{X}_{5 \times 3} \boldsymbol{\beta}_{3 \times 1} + \boldsymbol{\varepsilon}_{5 \times 1}$, where

$$\mathbf{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \mathbf{X} = [\mathbf{J} \quad \mathbf{x}_1 \quad \mathbf{x}_2] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \text{ with } \boldsymbol{\varepsilon} \sim N_5(\mathbf{0}, \sigma^2 \mathbf{I}).$$

(b) Show as rigorously as possible whether $\boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.

For $\boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ to be testable, it must also be estimable with either \mathbf{C} or \mathbf{M} being full rank.

$$\begin{aligned} \boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix} &\Rightarrow \boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta} \equiv \boldsymbol{\theta}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{bmatrix} \\ \Rightarrow \mathbf{C} = \mathbf{T}\mathbf{X} &\equiv \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix} \\ \Rightarrow \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} &= \begin{bmatrix} \sum_{i=1}^5 t_{1i} & t_{11} + t_{12} + t_{15} & -t_{13} - t_{14} \\ \sum_{i=1}^5 t_{2i} & t_{21} + t_{22} + t_{25} & -t_{23} - t_{24} \end{bmatrix} \end{aligned}$$

These equations are satisfied by $\mathbf{T} = \begin{bmatrix} 1/3 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$. This is one of many solutions for \mathbf{T} .

$$\therefore \boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix} \text{ is estimable}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix} \Rightarrow r(\mathbf{C}) = 2 \quad \therefore \mathbf{C} \text{ is full rank}$$

Since $\boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is estimable and \mathbf{C} is full rank, $\boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.

- 4) A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights ($y = \text{WGHT}$), average daily exercise time ($x = \text{TIME}$). One of the objectives in this study is to investigate the effect of TIME on weight loss.

- (a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete this table.

Dependent Variable: WGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184	$2624.670184/1 = 2624.670184$	137.906162	$\sim F(1,96)$ 2.86E-20
Error	96	1827.099916	$1827.099916/96 = 19.032291$		
Corrected Total	97	4451.7701			

- (b) State the model assumptions based on which the ANOVA table was computed.

HILE-Gauss:

1. Homogeneity Assumption: We assume each row of ε has same variance σ^2 .
2. Independence Assumption: We assume each row of ε is statistically independent of every other row.
3. Linearity Assumption: We assume the expected value of the response are linear functions of the parameter. $E(y) = X\beta$.
4. Existence Assumption: We observe values of random variables with finite variance. $H_0: \sigma_{model}^2 = \sigma_{error}^2$
5. The error term follows a Gaussian distribution. $\varepsilon_i \sim N(0, \sigma^2)$

- (c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.

Yes, daily exercise time is a significant predictor for predicting weight loss.

The test statistic is 137.9062.

The F-test of $\beta_{\text{WGHT}} = 0$ for $y = X\beta + \varepsilon$ generates a p-value < 0.0001 .

We reject the null hypothesis that $\beta_{\text{WGHT}} = 0$ (the average daily exercise time is not significant).

Therefore, daily exercise time appears to be a significant predictor for predicting weight loss.

- 5) An investigator studied the ozone levels in the South Coast Air Basin of California for the years 1976-1991. He believes that the number of days the ozone levels exceeded 0.20 ppm (the response) depends on the seasonal meteorological index, which is the seasonal average temperature in degrees Celsius (the predictor). The data *hw2.dat*, is provided on Sakai.

- (a) Fit a regression model with the number of high ozone days as the response and the meteorological index as a covariate and provide estimates of β_0, β_1 , their standard errors, and their interpretations.

Table 5.1: Regression Model for Number of High Ozone Days

Variable	Parameter Estimate	Standard Error	Interpretation
β_0	-192.98	163.503	The number of high ozone days with a meteorological index temperature of 0.
β_1	15.30	9.421	The change in the number of days with high ozone with every increase in the meteorological index temperature by 1 degree Celsius.

- (b) Are all of the β 's estimable? Why or why not?

Yes. X is full rank ($r(X) = 2 = p = r$). Therefore, all the β 's are estimable.

- (c) Report a test of the hypothesis that the number of high ozone days is associated with the meteorological index.

Hypothesis: $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

Test Statistic: $t_{obs} = \frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}} = \frac{15.30 - 0}{9.421} = 1.62 \sim t_{14}$

Degrees of Freedom: $df = 16 - 2 = 14$

P-value: $\Pr(|t| > 1.62) = 2 * (1 - \Pr(t \leq 1.62)) = 0.1267$

Decision: We fail to reject the null hypothesis.

Interpretation: There is insufficient evidence to suggest that there is an association between the number of high ozone days and meteorological index.

- (d) Using the framework of the linear model, report an $\alpha = 0.05$ test of the hypothesis that a 1 degree increase in average temperature is associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.

Hypothesis: $H_0: \beta_1 = 12$ vs. $H_1: \beta_1 \neq 12$

Test Statistic: $t_{obs} = \frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}} = \frac{15.30 - 12}{9.421} = 0.35 \sim t_{14}$ or
 $F = 0.12 \sim F(1, 14)$

Degrees of Freedom: $df = 16 - 2 = 14$

Critical Region: $C_\alpha = \{t: |t| > t_{v, 1-\frac{\alpha}{2}}\} \rightarrow C_{0.05} = \{t: |t| > 2.1448\}$

P-value: $\Pr(|t| > 0.35) = 2 * (1 - \Pr(t \leq 0.35)) = 0.7316$ or
 $\Pr(F_{1,14} > 0.12) = 1 - \Pr(F_{1,14} \leq 0.12) = 0.7316$

Decision: We fail to reject the null hypothesis.

Interpretation: There is insufficient evidence to suggest that a 1 degree increase in average temperature is not associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.

- (e) Calculate the 95% confidence interval and prediction interval for the expected number of days the ozone level exceeded 0.2 ppm when the seasonal meteorological index is 16.

95% Confidence Interval: (21.7579, 81.7581)

When the seasonal meteorological index is 16, there is a 95% confidence that the average number of days the ozone level exceeded 0.2 ppm is between 21.76 and 81.76 days.

95% Prediction Interval: (-7.4416, 110.9576)

Based on the observed data, there is a 95% chance that a seasonal meteorological index of 16 will result in between 0 and 110.96 days that the ozone level exceeded 0.2ppm.

BIOS663 Homework 3

Due noon on Tuesday, March 5 to my mailbox.

1. A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights ($y = \text{WGHT}$), average daily exercise times ($x_1 = \text{TIME}$) and average daily running mileages ($x_2 = \text{RUN}$). One of the objectives in this study is to investigate the effect of x_1 and x_2 on weight loss.

- (a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete the table.

Dependent Variable: WGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184			
Error	96	1827.099916			
Corrected Total	97	4451.7701			

- (b) State the model assumptions based on which the ANOVA table was computed.
- (c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.
- (d) To further explore the unexplained variation in the data, and in order to improve the predictive power of the model, the average daily running mileage (RUN) is also considered. The result is summarized in the following table. Does the analysis suggest that neither variable is significant? Why and why not?

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TIME	1	88.173	88.173	3.10	0.08
RUN	1	70.339	70.339	2.47	0.12

2. A data set was collected by the Environmental Protection Agency (EPA) at the Health Effects Research Laboratory at UNC: Chapel Hill. One hundred seventy-two young adult males received a battery of pulmonary function tests. (The data are described in more detail in Muller and Fetterman on page 536).

For this homework, fit a model with average forced vital capacity (FVC) (in ml) as the outcome and height, weight, body mass index ($BMI = \frac{weight(kg)}{(height(m))^2}$), age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity as predictors. For the purpose of added-in-order tests, assume that this order is the preferred order for testing. The data are available on the course website in FILEN.DAT with associated SAS file hw3.SAS.

To report a test, provide H_0 , the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter.

- (a) Use Proc GLM to produce a table like Table 4.8.1 in Muller and Fetterman (pg56) (details about the table can be found in Lecture7.pdf, pg29) with the following predictors: height, weight and age. The table should contain six df values, six SS values, four MS values, three F values, and three p-values.
- (b) Report the test of whether the group of predictors (height, weight, body mass index, age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity) is important.
- (c) Report the corrected R^2 for these data.
- (d) Give the two models being compared in testing the following hypotheses for these data, and report each test.
 - i. H_1 : The entire group of predictors provides no useful information about FVC.
 - ii. H_2 : Height provides no information about FVC, not adjusting for effects of other predictors (i.e., in a simple regression model).
 - iii. H_3 : Adjusting for weight and BMI, height does not provide any additional information about FVC.
 - iv. H_4 : After adjusting for weight, BMI, age, elevation, speed, temperature, barometric pressure, and humidity, height does not provide any additional information about FVC.
 - v. H_5 : The group of body size variables (height, weight, BMI) provides no additional information about FVC compared to a model for only the mean level of FVC.
 - vi. H_6 : The group of body size variables (height, weight, BMI) provides no additional information about FVC after adjusting for age, elevation, speed, temperature, barometric pressure, and humidity.
- (e) Report a test of the hypothesis that humidity has no affect on FVC after adjusting for all the other variables in the model.
- (f) Describe the relationship between the body size variables and FVC in these data.
- (g) Based on the original model, which characteristics are associated with the best (largest) FVC?

3. For the same data in Q2, consider the following model

$$\begin{aligned} FVC_i = & \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \beta_3 BMI_i + \beta_4 AREA_i \\ & + \beta_5 AGE_i + \beta_6 AVTREL_i + \beta_7 AVTRSP_i + \beta_8 AVTREL_i AVTRSP_i \\ & + \beta_9 TEMP_i + \beta_{10} BARM_i + \beta_{11} HUM_i + \varepsilon_i, \end{aligned}$$

$$i = 1, \dots, n.$$

- (a) Compute the following correlations, giving the interpretation of each, between FVC and age, and report tests of the hypotheses that each correlation equals zero.
 - i. the correlation between age and FVC, controlling both for all the other variables in the model
 - ii. the correlation between age and FVC, controlling only age for all the other variables in the model
 - iii. the simple correlation between age and FVC (not controlling for any other variables)
- (b) Provide and interpret the following diagnostics (include subject ID when appropriate) for the regression model.
 - i. Largest 5 studentized residuals (in absolute value)
 - ii. Results of a test of the Gaussian distribution for the studentized residuals
 - iii. Histogram of the studentized residuals
 - iv. Plot of studentized residuals versus predicted values

1.

(a)

Dependent Variable: WGHT					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184	2624.670184	137.91	<0.001
Error	96	1827.099916	19.032291		
Corrected total	97	4451.7701			

(b)

The model assumptions behind the model $WGHT = \beta_0 + \beta_1 \times TIME + \varepsilon$ are:

- (1) existence assumption: assume ε_i has finite first and second moment. In other words, we observe values of random variables with finite variance.
- (2) linearity assumption: we assume that the expected values of the weight (WGHT) are linear functions of the average daily exercise times (TIME).
- (3) independent assumption: we assume that each element of ε is statistically independent of every other.
- (4) homogeneity assumption: we assume that each element of ε has the same variance σ^2 .
- (5) Gaussian error assumption: we assume that $\varepsilon_i \sim N(0, \sigma^2)$. Note that the normality assumption is needed for the F test, but not needed for the estimates.

(c)

The average daily exercise time is a significant predictor for predicting weight loss. In this specific hypothesis testing, the null hypothesis is that the average daily exercise time is not a significant predictor for predicting weight loss ($H_0: \beta_1 = 0$); and the alternative hypothesis is that the average daily exercise time is a significant predictor for predicting weight loss ($H_A: \beta_1 \neq 0$). According to the ANOVA table, the test statistic $F_{obs} = 137.91$, which follows F distribution with degree of freedom of 1 and 96. The corresponding p-value is less than 0.05. Therefore, we reject the null hypothesis. The result can be interpreted as the following: assuming that the average daily exercise time is not a significant predictor for predicting weight loss, then the probability of

observing the data that are as extreme as ours or more extreme is less than 0.05, which is too small for us to believe that the null hypothesis is true.

(d)

The analysis does not suggest that neither variable is significant, essentially because they are correlated covariates so that the addition of one additional covariate does not provide additional significant information, given the other covariate is in the model. More specifically, the output is for the type III test, which refers to the statistical significance of the added-last test, given that all the other variables are in the model. In other words, the results are interpreted as: given the RUN is in the model, then the p-value for the test of the regression coefficient of TIME after being added to the model is 0.08; and given the TIME is in the model, then the p-value for the test of the regression coefficient of RUN after being added to the model is 0.12.

However, if we add either TIME or RUN to the intercept only model of WGHT, it is still possible that they are significant variables.

2

a.

Source	Df	SS	MS	Fobs	p
Intercept	1.0	4839362527.0	4839362527.0	11694.6	<.0001
Model (Un.)	4.0	4885896692.4	1221474173.1	2951.8	<.0001
Model (Cor.)	3.0	46534165.4	15511388.5	37.5	<.0001
Error (Res.)	166.0	68692765.6	413811.8		
Total (Un.)	170.0	4954589458.0	29144643.9		
Total (Cor.)	169.0	115226931.0	681816.2		

- b. $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (This set of predictors does not contribute to explaining any of the variability in FVC.) We reject the null hypothesis that these predictors, as a set,

do not significantly predict FVC, as the set of predictors significantly predict FVC, $F(9,160) = 13.90$, $p < 0.0001$.

$$c. R^2_C = \frac{CSS(\text{Regression})}{CSS(\text{Total})} = \frac{50570942}{115226931} = 0.4389$$

d. For all of the following calculations, the 2 individuals with any missing data were deleted before running the model.

i. Comparing intercept-only model to full model

Full Model: FVC_i

$$= \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$$

Intercept - Only Model: $FVC_i = \beta_0 + \varepsilon_i$

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

$$F_{obs} = \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} = \frac{\frac{115226931 - 64655989}{169 - 160}}{64655989/160} = \frac{50570942/9}{404100} = 13.90$$

We reject the null hypothesis that these predictors, as a set, do not significantly predict FVC, as the set of predictors significantly predict FVC, $F(9,160) = 13.90$, $p < 0.0001$.

ii. Comparing intercept-only model to model with height as predictor

Larger Model: $FVC_i = \beta_0 + \beta_1(\text{height}) + \varepsilon_i$

Intercept - Only Model: $FVC_i = \beta_0 + \varepsilon_i$

$$H_0: \beta_1 = 0 \text{ in } FVC_i = \beta_0 + \beta_1(\text{height}) + \varepsilon_i$$

$$F_{obs} = \frac{\frac{SSE(\text{smaller}) - SSE(\text{larger})}{dfE(\text{smaller}) - dfE(\text{larger})}}{SSE(\text{full})/dfE(\text{full})} = \frac{\frac{115226931 - 76194753}{169 - 168}}{64655989/160} = \frac{39032178}{404100} = 96.59$$

Height provides information about FVC in a simple regression, $F(1,160) = 96.59$, $p < 0.001$. We reject the null hypothesis that height provides no information in predicting FVC, height is significantly related to FVC.

iii. Comparing model with weight and BMI as predictors to model with height, weight, and BMI as predictors

Larger Model: $FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \varepsilon_i$

Smaller Model: $FVC_i = \beta_0 + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \varepsilon_i$

$$H_0: \beta_1 = 0 \text{ in } FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \varepsilon_i$$

$$F_{obs} = \frac{\frac{SSE(\text{smaller}) - SSE(\text{larger})}{dfE(\text{smaller}) - dfE(\text{larger})}}{SSE(\text{full})/dfE(\text{full})} = \frac{\frac{70003745 - 69983719}{167 - 166}}{64655989/160} = \frac{20026}{404100} = 0.0496$$

After adjusting for weight and BMI, height provides no additional information in predicting FVC, $F(1,160) = 0.0496$, $p = 0.82$. We fail to reject the null hypothesis that height provides information about FVC after controlling for weight and BMI.

iv. Comparing full model to full model minus height

Full Model: FVC_i

$$= \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$$

Reduced Model: FVC_i

$$= \beta_0 + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$$

$H_0: \beta_1 = 0$ in

$FVC_i =$

$$\beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$$

$$F_{obs} = \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} = \frac{64719380 - 64655989}{161 - 160} = \frac{63391}{404100} = 0.1569$$

After adjusting for all the other predictors in the model, height provides no additional information in predicting FVC, $F(1,160)=0.1569$, $p=0.69$. We fail to reject the null hypothesis that height provides information about FVC after controlling for the other variables in the model.

- v. Comparing intercept-only model to model with model with height, weight, and BMI as predictors

$$\text{Larger Model: } FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \varepsilon_i$$

$$\text{Intercept - Only Model: } FVC_i = \beta_0 + \varepsilon_i$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ in $FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(\text{smaller}) - SSE(\text{larger})}{dfE(\text{smaller}) - dfE(\text{larger})}}{SSE(\text{full})/dfE(\text{full})} = \frac{115226931 - 69983719}{169 - 166} = \frac{15081070}{404100} = 37.32$$

Height, weight, and BMI together provide additional information about FVC compared to a model for only the mean level of FVC, $F(3,160)=37.32$, $p<0.001$. We reject the null hypothesis that the body size variables provide no more information than the mean-only model. As a set, these three variables provide significant information about FVC.

- vi. Comparing model with age, elevation, speed, temp, barometric pressure, and humidity with model with age, elevation, speed, temp, barometric pressure, humidity, height, weight, and BMI

Full Model: FVC_i

$$= \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$$

Reduced Model: FVC_i

$$= \beta_0 + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ in $FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{99234383 - 64655989}{163 - 160} = \frac{11526131}{404100} = 28.523 \quad \checkmark$$

As a set, height, weight, and BMI together provide additional information about FVC after controlling for all of the other variables in the model, $F(3,160)=28.523$, $p<0.001$. We reject the null hypothesis that the body size variables provide no more information after controlling for all of the other variables in the model. As a set, these three variables provide significant information about FVC after controlling for the other variables.

- e. Full Model: $FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$

Reduced Model: $FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \varepsilon_i$

$H_0: \beta_9 = 0$ in

$FVC_i = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{weight}) + \beta_3(\text{BMI}) + \beta_4(\text{age}) + \beta_5(\text{avg treadmill elevation}) + \beta_6(\text{avg treadmill speed}) + \beta_7(\text{temp}) + \beta_8(\text{barom}) + \beta_9(\text{humid}) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{64768642 - 64655989}{161 - 160} = \frac{112653}{404100} = 0.2788, p = 0.5982 \quad \checkmark \quad \checkmark$$

We fail to reject the null hypothesis that humidity has no effect on FVC after controlling for the other variables in the model, $F(1,160)=0.2788$, $p=0.5982$. After controlling for the other variables in the model, humidity does not significantly relate to FVC.

- f. The body size variables together are significantly related to FVC, $F(3,160)=37.32$, $p<0.001$. They are significantly related even after controlling for all other variables in the model, $F(3,160)=28.523$, $p<0.001$. However, the variables are correlated with each other, as height is significantly related to FVC, $F(1,160)=96.59$, $p<0.001$, but not after controlling for weight and BMI, $F(1,160)=0.0496$, $p=0.82$. \checkmark
- g. Examining the added-in-order test, we see that height and weight are both highly related to FVC, but controlling for the other variables, neither height nor weight are significantly related (collinearity with BMI). The added-in-order also reveals that average treadmill elevation and average treadmill speed are significantly related to FVC as well. Again, controlling for the other variables in the model, neither variable individually is significantly related to FVC. Age is significantly related to FVC when controlling for the other variables, but is not significantly related to FVC when controlling for only height, weight, and BMI. \checkmark

Examining parameter estimates from the added-in-order tests, we find that increasing height as well as increasing weight is associated with an increased FVC. Controlling for height, weight, BMI, and age, increasing average treadmill elevation as well as increasing treadmill speed are associated with increased FVC. While age is not significantly related to FVC controlling for height, weight, and BMI, an increased age is associated with increased FVC controlling for all other variables in the model.

7. This problem involves the FEV data described above. We will consider the model $FVC_i = \beta_0 + \beta_1 X_H + \beta_2 X_W + \beta_3 X_{BMI} + \beta_4 X_{Area} + \beta_5 X_{Age} + \beta_6 X_{avtrei} + \beta_7 X_{avtrsp} + \beta_8 X_{avtrei \cdot avtrsp} + \beta_9 X_{temp} + \beta_{10} X_{barm} + \beta_{11} X_{hum} + \epsilon$.

a. Compute the following correlations, giving the interpretation of each, between FVC and age, and report tests of the hypotheses that each correlation equals zero.

i. The correlation between age and FVC, controlling both for all the other variables in the model

• **Hypotheses:**

- $H_0: \rho_{(age, FVC | other\ variables)} = 0$
- $H_A: \rho_{(age, FVC | other\ variables)} \neq 0$

• **Test Statistic:** $F_{obs} = \frac{\frac{[SSE(no\ age) - SSE(full)]}{[df_e(no\ age) - df_e(full)]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[64828716.44 - 62761458.29]}{[159 - 158]}}{\frac{62761458.29}{158}} = 5.20426$ ✓

• **Degrees of Freedom:** $df_e(no\ age) = 159, df_e(full) = 158$

• **P-value:** $\Pr(F_{obs} > F_{(1, 158)}) = 0.02387$ ✓

• **Decision:** Reject the null hypothesis

• **Interpretation:** There is a nonzero correlation between age and FVC after adjusting both for the other variables in the model.

• **Correlation:**

- According to SAS, $\rho_{(age, FVC | other\ variables)} = 0.17857$ ✓
- This suggests that, after controlling both variables for all of the other variables, with one increase in standard deviation of age, FVC is expected to increase by 0.17857 standard deviations.

ii. The correlation between age and FVC, controlling only age for all the other variables in the model

• **Hypotheses:**

- $H_0: \rho_{FVC(age | other\ variables)} = 0$
- $H_A: \rho_{FVC(age | other\ variables)} \neq 0$

• **Test Statistic:**

- First, used SAS to model obtain studentized residuals of age=(other variables)
- Second, used SAS to model avfvc=(studentized residuals)

• $F_{obs} = \frac{\frac{[SSE(\beta_0) - SSE(full)]}{[df_e(\beta_0) - df_e(full)]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[CSS(model)]}{[169 - 168]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[2067258.2]}{[169 - 168]}}{\frac{113159672.8}{168}} = 3.07$

• **Degrees of Freedom:** $df_e(\beta_0) = 169, df_e(full) = 168$

• **P-value:** $\Pr(F_{obs} > F_{(1, 168)}) = 0.0816$ ✓

• **Decision:** Fail to reject the null hypothesis

• **Interpretation:** After controlling the other variables on age, there isn't enough evidence to suggest a significant correlation between age (adjusted) and FVC.

• **Correlation:** $r_{FVC(age | other\ variables)} = r(FVC_i, \hat{\epsilon}_{age}) = 0.13394$. If this correlation was statistically significant, it would suggest that with one increase in standard deviation of age (adjusted), FVC (unadjusted) is expected to increase by 0.13570 standard deviations.

iii. The simple correlation between age and FVC (not controlling for any other variables)

• **Hypotheses:** $H_0: \rho_{FVC, age} = 0$ vs. $H_A: \rho_{FVC, age} \neq 0$

• **Test Statistic:**

○ $F_{obs} = \frac{\frac{[SSE(\beta_0) - SSE(\beta_{0, age})]}{[df_e(\beta_0) - df_e(\beta_{0, age})]}}{\frac{SSE(\beta_{0, age})}{df_e(\beta_{0, age})}} = \frac{\frac{[115226930.97 - 112547661.64]}{[169 - 168]}}{\frac{112547661.64}{168}} = 3.99935$ ✓

• **Degrees of Freedom:** $df_e(\beta_0) = 169, df_e(full) = 168$

• **P-value:** $\Pr(F_{obs} > F_{(1, 168)}) = 0.047129$

• **Decision:** Reject the null hypothesis

• **Interpretation:** There is enough evidence to suggest a simple correlation between age and FVC.

• **Correlation:** $r_{FVC, age} = 0.15249$. This suggests that with one increase in standard deviation of age (unadjusted), FVC (unadjusted) is expected to increase by 0.15249 standard deviations.

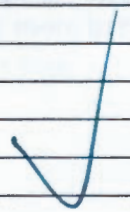
b. Provide and interpret the following diagnostics (include subject ID when appropriate) for the regression model.

(b)

i.

the largest 5 studentized residuals in absolute values are:

Subject ID	Studentized residuals in absolute value
60	3.882
185	2.904
49	2.903
181	2.660
99	2.267



ii.

To test whether the studentized residuals are normal, we can use one of:

- 1) Shapiro-Wilk Test;
- 2) Kolmogorov-Smirnov Test;
- 3) Cramer-von Mises Test;
- 4) Anderson-Darling Test.

The null hypothesis is that the studentized residuals follow normal distribution, and the followings are the corresponding test statistics and the p-values:

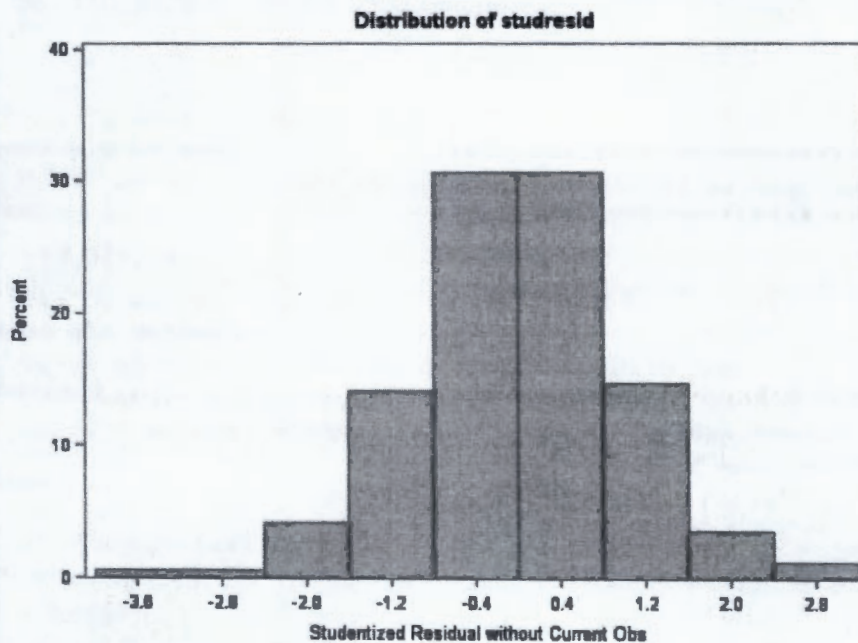
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.989513	Pr < W	0.2420
Kolmogorov-Smirnov	D	0.044551	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.039587	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.308589	Pr > A-Sq	>0.2500

All the tests show p-values that are greater than 0.05, so that we fail to reject the null hypothesis. The result can be interpreted as the following: assuming that the studentized residuals are normally distributed, then the probability of observing the data that are as extreme as ours or more extreme is larger than 0.05, which is not too small for us to question the null hypothesis.

Note that the type of test should be decided before conducting the test to avoid "p shopping", although the test statistics and p-values are all reported in the table above.

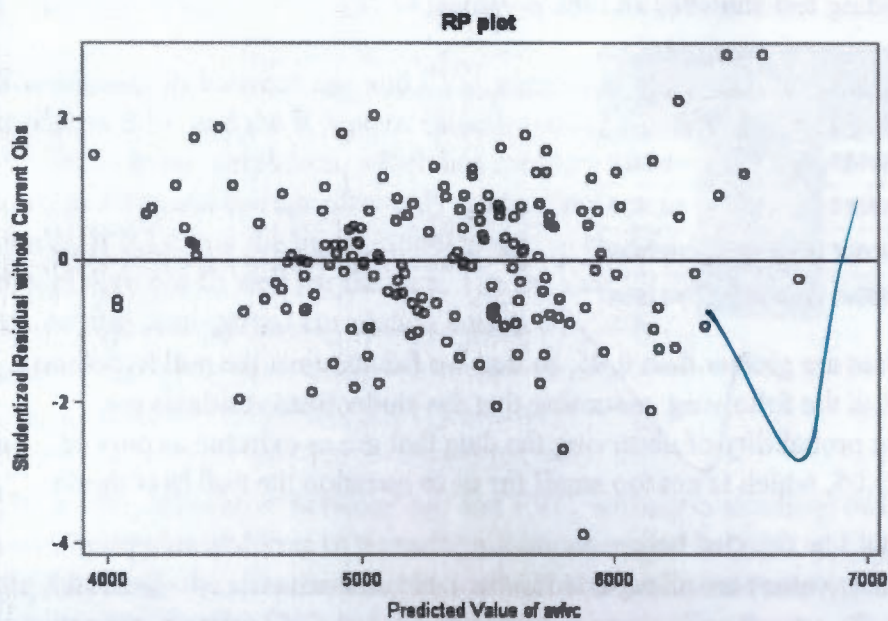
iii.

The histogram of the studentized residuals is plotted as the following:



The histogram can be interpreted as the following: the shape of the studentized residuals appears to be normal in this histogram, which coincide with our hypothesis testing in part ii.

iv. the RP plot of studentized residuals is plotted as the following:



The RP plot can be interpreted as the following: from the RP plot we can see that the linearity assumption, the homogeneity of variance assumption, and Gauss error assumption holds, because we cannot see any nonlinear pattern or heterogeneity of variance of the residuals. Also the studentized residuals appear to be normal as well.

BIOS663 Homework 4
Due Monday, April 8 in class.

1. The following questions are on the data and model described in Q3 of HW3:
 - Examine the tolerances and variance inflation factors from this model. Do you think any collinearity is present based on the tolerance and VIF? Why or why not?
 - Conduct an eigenanalysis of the scaled SSCP and correlation matrices, presenting a table formatted like Table 8.6.2 in Muller and Fetterman.
 - (a) Does there appear to be any collinearity between the intercept and the covariates? Why or why not? If so, list the variables?
 - (b) Does there appear to be any collinearity among the covariates? Why or why not? If so, list the variables?
2. Find the Box-Cox transformation of the simulated data (BoxCox.dat) and compare the residual plots of the raw and transformed data.
3. Investigators are interested in the effect of dermal nicotine exposure in a population of Latino tobacco workers in North Carolina. (Nicotine can be absorbed from tobacco leaves through the skin and can cause nicotine poisoning, which is characterized by nausea, vomiting, headache, and dizziness.) Data were collected on tobacco work tasks and risk factors for exposure to nicotine during a summer tobacco work season. Nicotine exposure was measured by levels of cotinine, a nicotine metabolite, contained in saliva. Other covariates of interest include age, body mass index, education, work conditions (working in wet conditions is believed to increase nicotine absorption), type of tobacco work (“priming” refers to picking or harvesting the tobacco and is expected to result in highest nicotine exposures, “barning” refers to putting the harvested tobacco into a barn for curing, “topping” refers to breaking the flower off the top of the plant, and “other” refers to farm work that does not involve tobacco contact, such as driving a truck), and smoking (smokers would also have nicotine exposure through cigarettes, and it is not known whether exposure to tobacco leaves would increase cotinine levels to a similar extent in both smokers and non-smokers).

The variables are available in the file tobacco.dat and listed in the following order.

- COTININE: salivary cotinine concentration (in ng/mL)
- AGE: age (in years)
- BMI: body mass index (in kg/m²)
- EDUC: years of education
- WET: takes value 1 if work conditions on day of measurement were wet and takes value 0 otherwise

- TASK: takes value 1 for priming, 2 for barning, 3 for topping, and 4 for other work not involving tobacco contact
- LNNSMOKE: natural logarithm of $(1 + \text{number of cigarettes smoked per day})$

To report a test, provide H_0 , the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter.

- One-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task as the only predictor.
 - Report a test of whether all cell means are equal.
 - If your overall test of the task effect was significant, examine all pairwise comparisons using the Scheffe correction. Summarize your findings in a table including columns for the estimated mean difference, degrees of freedom, F statistic, p -value, and a Scheffe confidence interval for the mean difference. Explain your findings in language the investigator can understand.
 - Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding, and give the interpretations of parameters in both coding schemes. In addition, provide the \mathbf{C} and $\boldsymbol{\theta}_0$ matrices used to test the hypothesis that average cotinine levels for workers involved in priming are greater than the average cotinine levels for all other workers.
- Two-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task and wet as predictors.
 - Fit the two-way ANOVA model with full interaction, and interpret all parameter estimates in your model, clearly stating which coding scheme you used. Discuss the validity of the HILE Gauss assumptions for this model.
 - *Based on this model*, create a table of the estimated mean log cotinine levels, associated standard errors, and how each estimated mean is obtained from the model parameters (e.g., $\hat{\beta}_0 + \hat{\beta}_1$) for each task-wet combination.
- The Full Model in Every Cell: For these questions, use the log of salivary cotinine as the response and task, and lnnsnsmoke as predictors.
 - Fit the full model in every cell. Provide and interpret estimates of all parameters for this model.
 - Report an appropriate test of whether task is related to cotinine levels. If this test is significant, report step-down tests to determine exactly where differences lie. For all tests reported, be sure to state H_0 clearly and give explicit justification for which tests were used and why.

BIOS 663 Homework 4

4/8/2019

Problem 1

The following questions are on the data and model described in Q3 of HW3.

part i:

Examine the tolerances and variance inflation factors from this model. Do you think any collinearity is present based on tolerance and VIF? Why or why not? Define tolerance as follows:

$$T_j = 1 - R_j^2$$

where $R_j^2 = R^2(X_j, \{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_{p-1}\})$ is the squared multiple correlation. Tolerances close to 1 are good, where tolerances close to 0 show worse collinearity.

Define the variance inflation factor as follows:

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{T_j}$$

A VIF close to 1 is good, where a VIF implies worse collinearity as it approaches infinity. The R-Code below calculates the Tolerance and VIF values for the model.

```
fit = lm(AVFVC ~ HEIGHT + WEIGHT + BMI + AREA + AGE + AVTREL + AVTRSP + AVTREL*AVTRSP + TEMP + BAROM + HUMID, data = dat2)
VIF = vif(fit)
Tol = 1/VIF
df = (rbind(VIF, Tol))
df %>% knitr::kable(align = c("c", "c"))
```

	HEIGHT	WEIGHT	BMI	AREA	AGE	AVTREL	AVTRSP	TEMP	BAROM	HUMID	AVTREL:AVTRSP
VIF	458.0476405	703.4462861	177.4503720	1364.8975242	1.0833448	580.389689	78.3930237	29.0137857	1.0590953	29.1669187	795.4489506
Tol	0.0021832	0.0014216	0.0056354	0.0007327	0.9230672	0.001723	0.0127562	0.0344664	0.9442021	0.0342854	0.0012572

Based on these values, it appears that there is a lot of collinearity present. This is because very few VIF values are close to 1, and many of the tolerance values are close to 0.

part ii:

Conduct an eigenanalysis of the scaled SSCP and correlation matrices, presenting a table formatted like Table 8.6.2 in Muller and Fetterman.

The Scaled SSCP Matrix can be defined as follows:

$$SSCP_s = D_s^{-0.5}(X'X)D_s^{-0.5}$$

where X is the design matrix includign the intercept, and D_s is a diagonal matrix with elements extracted from the diagonal of $X'X$.

and the correlation matrix as follows:

$$R = D_c^{-0.5}CD_c^{-0.5}$$

where C is the covariance matrix of the centered design matrix excluding the intercept, and D_c is a diagonal matrix of the extracted diagonal values of C .

The following R code calculates these two matrices for the model above, and performs an eigenanalysis on them both.


```
cov_int <- dat2 %>% mutate(INT = 1, AVTRELTRSP = AVTREL*AVTRSP) %>% select(INT, HEIGHT, WEIGHT, BMI, AREA, AGE, AVTREL, AVTRSP, AVTRELTRSP, T
EMP, BAROM, HUMID) %>% as.matrix()
xtx = t(cov_int)%*%cov_int
Ds_half <- diag(diag(xtx)^-0.5)
sscp <- Ds_half %*% xtx %*% Ds_half
eig_sscp <- eigen(sscp)$values
PCs_sscp <- prcomp(sscp)[2]
CI_sscp <- sqrt(eig_sscp[1])/eig_sscp)

covariates <- dat2 %>% mutate(AVTRELTRSP = AVTREL*AVTRSP) %>% select(HEIGHT, WEIGHT, BMI, AREA, AGE, AVTREL, AVTRSP, AVTRELTRSP, TEMP, BAROM,
HUMID)
cov_center <- apply(covariates, 2, function(y) y - mean(y))
C <- (t(cov_center)%*%cov_center)/dim(cov_center)[1]
Dc_half <- diag(diag(C)^-0.5)
R <- Dc_half %*% C %*% Dc_half
eig_corr <- eigen(R)$values
CI_corr <- sqrt(eig_corr[1])/eig_corr)
PCs_corr <- prcomp(R)[2]

df <- data.frame("Eigenvalue" = c("Correlation Matrix", eig_corr), "Condition Index" = c("Correlation Matrix", CI_corr))
df2 <- data.frame("Eigenvalue" = c("Scaled SSCP", eig_sscp), "Condition Index" = c("Scaled SSCP", CI_sscp))

df %>% knitr::kable(align = c("c", "c"))
```

Eigenvalue	Condition.Index
Correlation Matrix	Correlation Matrix
3.00984215183995	1
2.44782688677429	1.10887223583127
2.02476480717731	1.21922699035498
1.11320126901436	1.64431498130338
1.01325012874562	1.72350888345776
0.809279431894358	1.92851316812643
0.561095110548582	2.31608032840747
0.0177075849003461	13.0374361244415
0.00187373597726271	40.0790724581539
0.000705172710081123	65.3317229298841
0.000453720417828643	81.4474887486485

```
df2 %>% knitr::kable(align = c("c", "c"))
```

Eigenvalue	Condition.Index
Scaled SSCP	Scaled SSCP
11.9049479582918	1
0.0360382910658672	18.1753028593325
0.0293708868488598	20.132848271369
0.0161788245373072	27.1262817394066
0.00669911241875652	42.1555847347362
0.0049048528907778	49.2663919299516
0.0015549706468733	87.4989120284385
0.000251548277639727	217.546988935216
3.7467685203547e-05	563.683495875739
9.67075357565232e-06	1109.51605795048
4.86755151319751e-06	1563.89817359537

(a):

Does there appear to be any collinearity between the intercept and the covariates? Why or why not? If so, list the variables.

Since the eigenvalues from the Scaled SSCP (which includes the intercept) show several eigenvalues near 0 and condition indices above 30 (namely the last 8), we know that there does appear to be collinearity issues. To identify which covariates this collinearity is between, we take a look at the last few PCs below.

```
PCs_sscp$rotation[,9:12]
```

```
##          PC9          PC10          PC11          PC12
## [1,] -0.362557672  0.2106087110  0.6871965147  0.1473741435
## [2,] -0.302287651 -0.3282233486 -0.4589345265  0.4834046812
## [3,]  0.228929531  0.4172069832  0.1173983196  0.3239142829
## [4,] -0.186467777 -0.2773877633 -0.1708233087 -0.0167347071
## [5,] -0.097183258 -0.3259677478  0.1402711189 -0.7151909522
## [6,]  0.008240427  0.0007674032  0.0006411418 -0.0004944857
## [7,] -0.093561530  0.3902109785 -0.2888650047 -0.2038318770
## [8,] -0.094260468  0.4147408015 -0.2908961834 -0.2059574809
## [9,]  0.100295075 -0.3935164095  0.2925559731  0.2064610370
## [10,] -0.025805362  0.0012700485 -0.0127654815 -0.0270116541
## [11,]  0.806990526 -0.1072759346 -0.0237080809 -0.0096174923
## [12,]  0.029751201 -0.0035722176  0.0110939585  0.0220843112
```

From the PCA analysis, we can see that the covariates with the largest departures from 0 in the last four PCs are covariates 1 (i.e. the intercept), 2 (height), 3 (weight), 5 (area), 7 (avtrel), 8 (avtrsp), and 9 (avtrel*avtrsp).

(b):

Does there appear to be any collinearity among the covariates? Why or why not? If so, list the variables.

Since the eigenvalues from the Correlation Matrix (which does NOT includes the intercept) show several eigenvalues near 0 and condition indices above 30 (namely the last three), we know that there does appear to be collinearity issues. To identify which covariates this collinearity is between, we take a look at the last few PCs below.

```
PCs_corr$rotation[,10:11]
```

```
##          PC10          PC11
## [1,]  0.108339574 -0.507595144
## [2,]  0.544305474 -0.014314017
## [3,] -0.130059319 -0.310181321
## [4,] -0.544577040  0.582884584
## [5,] -0.004893314 -0.001127961
## [6,] -0.390534996 -0.350817108
## [7,] -0.138224165 -0.124569696
## [8,]  0.454476015  0.409062361
## [9,] -0.012113720  0.009805947
## [10,] -0.002270905 -0.002276515
## [11,]  0.012911758 -0.012268259
```

From the PCA analysis, we can see that the covariates with the largest departures from 0 in the last four PCs are covariates 1 (height), 3 (BMI), 4 (area), 6 (avtrel), and 8 (avtrel*avtrsp).

Problem 2

Find the Box-Cox Transformation of the simulated data (BoxCox.dat) and compare the residual plots of the raw and transformed data.

The Box-Cox Transformations are a family of transformations of the response variables defined as:

$$Y_i(\pi) = \left\{ \frac{Y_i^\pi - 1}{\pi Y^{*(\pi-1)}} \quad \pi \neq 0 \quad \& \quad Y^* \ln(Y_i) \quad \pi = 0 \right\}$$

where $Y^* = (\prod_{i=1}^N Y_i)^{1/N}$. This corresponds to a transformation that is y^π for $\pi \neq 0$ and $\log(y)$ otherwise. The transformation above puts the SSE of these on the same scale for the purpose of comparison and choosing the best π . We try the values of π between -1 and 1 incremented every 0.25 to compare the likelihoods and find the best transformation.

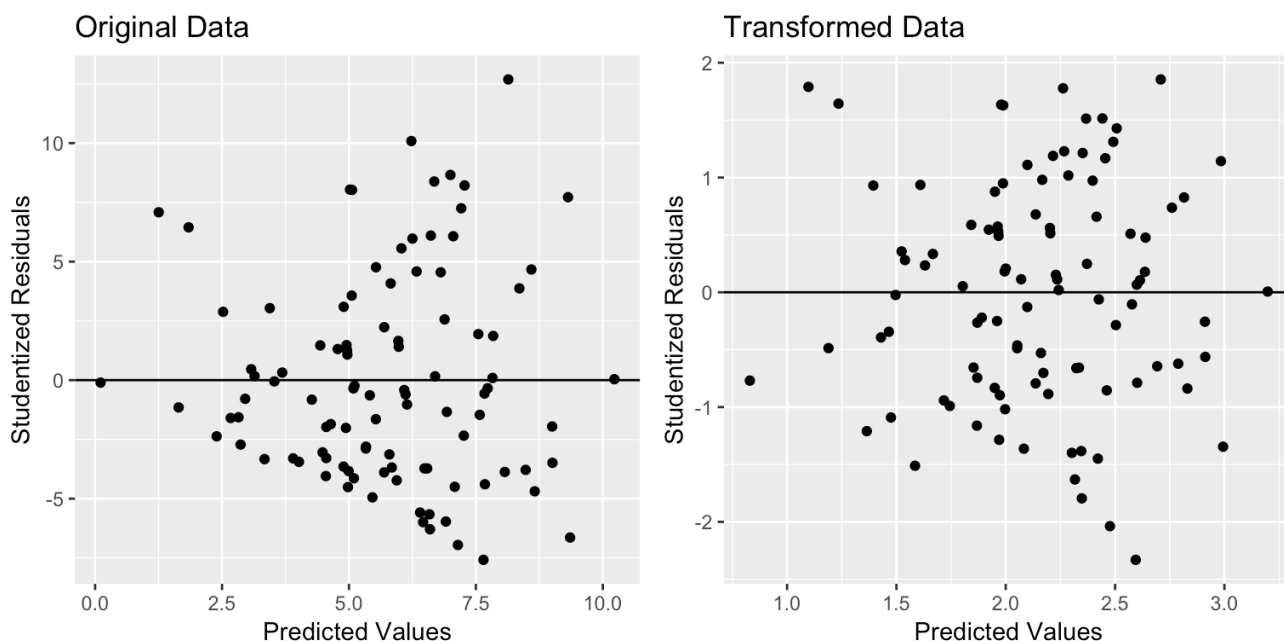
```
fit_bc <- lm(bc$V2 ~ bc$V1)
# Transformation Analysis
cols <- MASS::boxcox(fit_bc, seq(-1,1,1/4), plotit = FALSE)$x
like <- MASS::boxcox(fit_bc, seq(-1,1,1/4), plotit = FALSE)$y %>% as.matrix() %>% t()
knitr::kable(like, col.names = cols)
```

-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
-685.0735	-549.0958	-425.2876	-325.0825	-264.4143	-241.4822	-239.6977	-248.8666	-264.7483

From the values above, we can see that the choice of π with the smallest likelihood is $\pi = 0.5$. Below, we transform the data using this result, and compare the residual plots of the two datasets.

```
lambda = 0.5
bc$V3 <- (bc$V2^lambda)
fit_bc_2 <- lm(bc$V3 ~ bc$V1)

plot1 <- ggplot() + geom_point(aes(fit_bc$fitted.values, fit_bc$residuals)) + geom_hline(aes(yintercept = 0)) + labs(x = "Predicted Values",
y = "Studentized Residuals", title = "Original Data")
plot2 <- ggplot() + geom_point(aes(fit_bc_2$fitted.values, fit_bc_2$residuals)) + geom_hline(aes(yintercept = 0)) + labs(x = "Predicted Value
s", y = "Studentized Residuals", title = "Transformed Data")
cowplot::plot_grid(plot1, plot2, nrow = 1)
```



From the plots above, it is clear that the transformation of the data yields better assumption validations than the original data. In particular, the graph on the left of the original data seems to fan out (i.e. more extreme residuals) as the predicted values are increased. The residuals are more randomly distributed around the x-axis in the transformed data.

Problem 3

part i: One-Way ANOVA:

For these questions, use the log of salivary cotinine as the response and task as the only predictor.

```
tobacco1 <- tobacco %>% mutate(LOGCOT = log(COTININE),
                                TASK1 = case_when(TASK == 1 ~ 1, TASK != 1 ~ 0),
                                TASK2 = case_when(TASK == 2 ~ 1, TASK != 2 ~ 0),
                                TASK3 = case_when(TASK == 3 ~ 1, TASK != 3 ~ 0),
                                TASK4 = case_when(TASK == 4 ~ 1, TASK != 4 ~ 0))
```

(a):

Report a test of whether all cell means are equal.

Consider the following model using the cell mean coding scheme:

$$y = \beta_1 I_{T1} + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4}$$

where y is the log cotinine, and I_{Ti} is the indicator function associated with the i th task. In order to test whether all cell means are equal, we want to test the following set of hypotheses:

$$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4$$

which is equivalent to:

$$H_0 = \beta_1 - \beta_2 = 0, \beta_1 - \beta_3 = 0, \beta_1 - \beta_4 = 0$$

In order to test these hypotheses, we can use the overall F test, where:

$$F = (SSH/a)/\hat{\sigma}^2 \sim F_{G-1, n-G}$$

where $SSH = (\hat{\theta} - \theta_0)'M^{-1}(\hat{\theta} - \theta_0)$, $G = 4$, $n = 694$, and $\hat{\sigma}^2 = \text{mse}$. It should also be noted that $M = C(X'X)^{-1}C'$.

```
X = tobacco1 %>% select(TASK1, TASK2, TASK3, TASK4) %>% as.matrix()
fit = lm(LOGCOT ~ -1 + TASK1 + TASK2 + TASK3 + TASK4, data = tobacco1)
thetahat = c((fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[2,1],
             (fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[3,1],
             (fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[4,1])
mse = sum(fit$residuals^2)/(694-4)
C = matrix(c(1, 1, 1, -1, 0, 0, 0, -1, 0, 0, 0, -1), nrow = 3)
M = C %*% solve(t(X) %*% X) %*% t(C)
ssh = t(thetahat) %*% solve(M) %*% thetahat
f_obs = (ssh/3)/mse
p = 1-pf(f_obs, 4-1, 694-4)
## CAN ALSO USE linearHypothesis(fit, C)
```

From the code above, the test statistic $F = 116.2032527$ and the p-value is approximately 0. This means that we can reject the null hypothesis that all four cell means are equal. In other words, there is evidence to reject the fact that the four types of tobacco work have the same mean log cotinine level.

(b):

If your overall test of the task effect was significant, examine all pairwise comparisons using the Scheffe correction. Summarize your findings in a table including columns for the estimated mean difference, degrees of freedom, F statistic, p-value, and a Scheffe confidence interval for the mean difference. Explain your findings in language the investigator can understand.

Since the overall test of the task effect in (a) was significant, we will go forward with the pairwise comparisons. Since TASK has four levels, there will be $4 * (4 - 1)/2 = 6$ pairwise comparisons. Scheffe's correction provides a general technique for account for the fact that we are performing 6 tests.

To find the F statistic, we can take the square of the t statistic as follows:

$$F = t^2 = \left(\frac{\hat{\beta}_i - \hat{\beta}_j}{\sqrt{MSE(1/n_i + 1/n_j)}} \right)^2 \sim F_{G-1, n-G}$$

where $\hat{\beta}_i$ and n_i are the mean log cotinine level and sample size for the i th task level. MSE is the mean squared error as calculated in the previous test. The critical region for this F test can be calculate by multiplying the F statistic by $G - 1 = 4 - 1 = 3$ to account for multiplicity in testing.

```
scheffe <- ScheffeTest(aov(LOGCOT ~ factor(TASK), data = tobacco1))
f1 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[2,1]) / sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK2))))^2
f2 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[3,1]) / sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK3))))^2
f3 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[4,1]) / sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK4))))^2
f4 <- ((summary(fit)$coefficients[2,1] - summary(fit)$coefficients[3,1]) / sqrt(mse*(1/sum(tobacco1$TASK2) + 1/sum(tobacco1$TASK3))))^2
f5 <- ((summary(fit)$coefficients[2,1] - summary(fit)$coefficients[4,1]) / sqrt(mse*(1/sum(tobacco1$TASK2) + 1/sum(tobacco1$TASK4))))^2
f6 <- ((summary(fit)$coefficients[3,1] - summary(fit)$coefficients[4,1]) / sqrt(mse*(1/sum(tobacco1$TASK3) + 1/sum(tobacco1$TASK4))))^2
df <- data.frame(Diff = scheffe$`factor(TASK)`[,1], DF = "(3,690)", f = c(f1, f2, f3, f4, f5, f6), pval = scheffe$`factor(TASK)`[,4], CI_L =
scheffe$`factor(TASK)`[,2], CI_U = scheffe$`factor(TASK)`[,3])
df %>% knitr::kable(align = c("c", "c"))
```

	Diff	DF	f	pval	CI_L	CI_U
2-1	-0.9207815	(3,690)	19.57598	0.0002350	-1.503991	-0.3375720
3-1	-1.6738481	(3,690)	131.87148	0.0000000	-2.082328	-1.2653684
4-1	-2.6992523	(3,690)	332.68364	0.0000000	-3.113975	-2.2845298
3-2	-0.7530666	(3,690)	12.98968	0.0049075	-1.338617	-0.1675167
4-2	-1.7784708	(3,690)	71.37797	0.0000000	-2.368393	-1.1885490
4-3	-1.0254042	(3,690)	47.25875	0.0000000	-1.443412	-0.6073970

From the table above, it appears that all pairwise null hypotheses can be rejected. This means there is evidence to suggest that every mean log cotinine level for a certain task level is different than the mean log cotinine level for any other task level.

(c):

Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding, and give the interpretations of parameters in both coding schemes. In addition, provide the C and θ_0 matrices used to test the hypothesis that average cotinine levels for workers involved in priming are greater than the average cotinine levels for all other workers.

For the cell mean coding, we use the model proposed in (a).

```
summary(fit)$coefficients
```

```
##           Estimate Std. Error  t value      Pr(>|t|)
## TASK1  4.508557    0.1022201  44.10636 5.762099e-203
## TASK2  3.587775    0.1812765  19.79172  2.168761e-69
## TASK3  2.834709    0.1039098  27.28047  9.869409e-112
## TASK4  1.809304    0.1070123  16.90745  6.478023e-54
```

The parameter estimates and standard errors are given in the code summary above. The interpretations are as follows: β_1 is the mean log cotinine level for priming, β_2 is the mean log cotinine level for barning, β_3 is the mean log cotinine level for topping, and β_4 is the mean log cotinine level for work not involving tobacco contact.

For the reference cell coding, we consider the following model:

$$y = \beta_1 + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4}$$

where y is the log cotinine, and I_{Ti} is the indicator function associated with the i th task. It should be noted that TASK1 is the reference.

```
fit_ref = lm(LOGCOT ~ TASK2 + TASK3 + TASK4, data = tobacco1)
summary(fit_ref)$coefficients
```

```
##           Estimate Std. Error  t value      Pr(>|t|)
## (Intercept)  4.5085566    0.1022201  44.106361 5.762099e-203
## TASK2       -0.9207815    0.2081109  -4.424476  1.123234e-05
## TASK3       -1.6738481    0.1457607 -11.483531  4.699458e-28
## TASK4       -2.6992523    0.1479884 -18.239617  5.849323e-61
```

Again, the parameter estimates and standard errors are given in the code summary above. The interpretations are different than for the cell mean coding and are as follows: β_1 is the intercept which again is the mean log cotinine level for priming, which is the reference level. β_2 is the difference between the mean log cotinine level for barning and the mean log cotinine level for priming. Similarly, β_3 is now the difference between the mean log cotinine level for topping and the mean log cotinine level for priming, and β_4 is the difference between the mean log cotinine level for work not involving tobacco contact and the mean log cotinine level for priming.

The TASK value corresponding with priming is 1, and so we want to test that the mean cotinine level for TASK1 is greater than the mean cotinine level for all others. We can test the following hypothesis using the cell mean coding:

$$H_0 = \beta_1 = (\beta_2 + \beta_3 + \beta_4)/3 \quad \text{vs.} \quad H_A = \beta_1 > (\beta_2 + \beta_3 + \beta_4)/3$$

This null hypothesis corresponds to:

$$H_0 = \beta_1 - \frac{1}{3}\beta_2 - \frac{1}{3}\beta_3 - \frac{1}{3}\beta_4 = 0$$

so $\theta_0 = [0]$ and $C = [1 \quad -1/3 \quad -1/3 \quad -1/3]$

part ii: Two-Way ANOVA:

For these questions, use the log of salivary cotinine as the response and task and wet as predictors.

```
tobacco$WET <- tobacco$WET %>% as.factor()
tobacco$TASK <- tobacco$TASK %>% as.factor()
tobacco2 = tobacco %>% mutate(LOGCOT = log(COTININE),
                              WET0TASK1 = case_when(WET == 0 & TASK == 1 ~ 1,
                                                         WET != 0 | TASK != 1 ~ 0),
                              WET1TASK1 = case_when(WET == 1 & TASK == 1 ~ 1,
                                                         WET != 1 | TASK != 1 ~ 0),
                              WET0TASK2 = case_when(WET == 0 & TASK == 2 ~ 1,
                                                         WET != 0 | TASK != 2 ~ 0),
                              WET1TASK2 = case_when(WET == 1 & TASK == 2 ~ 1,
                                                         WET != 1 | TASK != 2 ~ 0),
                              WET0TASK3 = case_when(WET == 0 & TASK == 3 ~ 1,
                                                         WET != 0 | TASK != 3 ~ 0),
                              WET1TASK3 = case_when(WET == 1 & TASK == 3 ~ 1,
                                                         WET != 1 | TASK != 3 ~ 0),
                              WET0TASK4 = case_when(WET == 0 & TASK == 4 ~ 1,
                                                         WET != 0 | TASK != 4 ~ 0),
                              WET1TASK4 = case_when(WET == 1 & TASK == 4 ~ 1,
                                                         WET != 1 | TASK != 4 ~ 0))
```

(a):

Fit the two-way ANOVA model with full interaction, and interpret all parameter estimates in your model, clearly stating which coding scheme you used. Discuss the validity of the HILE Gauss assumptions for this model.

Consider the following model using the cell mean coding scheme:

$$y = \beta_1 I_{W0,T1} + \beta_2 I_{W1,T1} + \beta_3 I_{W0,T2} + \beta_4 I_{W1,T2} + \beta_5 I_{W0,T3} + \beta_6 I_{W1,T3} + \beta_7 I_{W0,T4} + \beta_8 I_{W1,T4}$$

where y is the log cotinine, and $I_{Ti,Wj}$ is the indicator function associated with the i th task and j th wet categories.

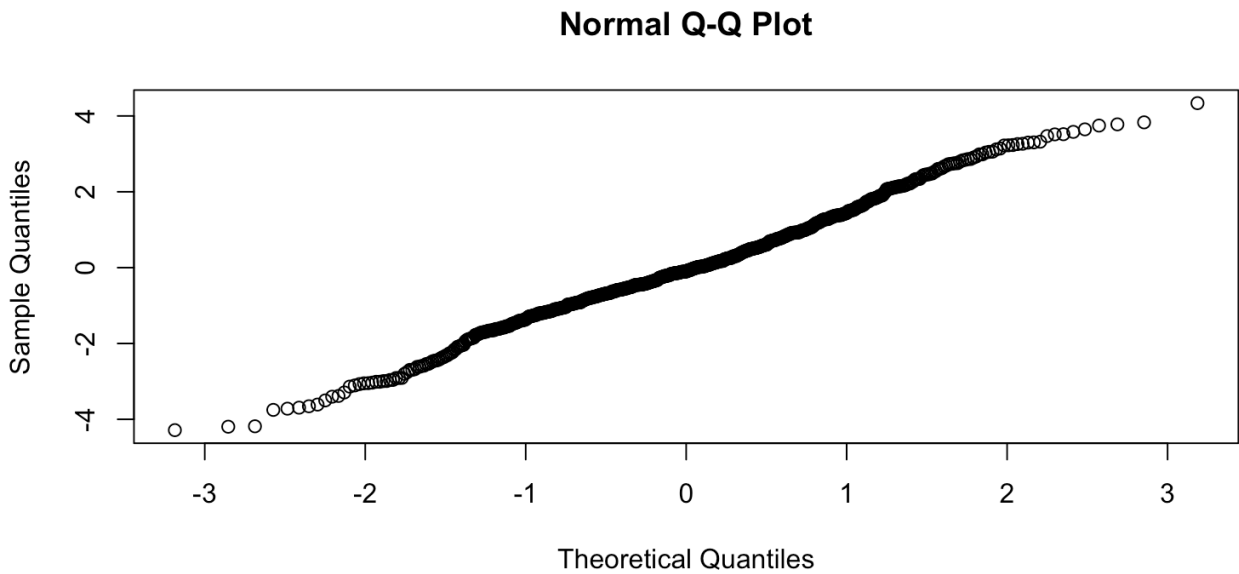
```
fit <- lm(LOGCOT ~ -1 + WET0TASK1 + WET1TASK1 + WET0TASK2 + WET1TASK2 + WET0TASK3 + WET1TASK3 + WET0TASK4 + WET1TASK4, data = tobacco2)
summary(fit)$coefficients
```

##		Estimate	Std. Error	t value	Pr(> t)
##	WET0TASK1	4.269337	0.1854712	23.018868	2.548339e-87
##	WET1TASK1	4.613116	0.1226196	37.621351	6.459772e-169
##	WET0TASK2	3.542748	0.2271549	15.596174	3.665240e-47
##	WET1TASK2	3.667024	0.3013550	12.168449	5.528611e-31
##	WET0TASK3	2.688185	0.2152536	12.488456	2.142659e-32
##	WET1TASK3	2.879303	0.1187505	24.246652	2.808906e-94
##	WET0TASK4	1.808889	0.1238562	14.604754	2.911782e-42
##	WET1TASK4	1.810534	0.2130902	8.496563	1.211849e-16

The estimates for each of the parameters are shown in the summary statistics above. Since the cell mean coding is used, the interpretations are clear; each parameter represents the mean log cotinine level for the combination of WET and TASK levels listed. For example, β_1 is estimated by 4.269337 noted by the WET0TASK1 indicator variable.

In terms of HILE Gauss assumptions for this model, the only assumptions that are generally checked for ANOVA are H, I, and Gauss. The independence assumption is dependent on the design and the sampling scheme, and from the description of the design, I do not see any issues that would question the validity of the independence assumption. In terms of the homogeneity and gaussian errors assumptions, we can perform tests as done below to check these assumptions. We also note that the design is unbalanced (i.e. the sample size per cell ranges from 25 to 161), which is something to consider when using this model.

```
qqnorm(fit$residuals)
```



The linearity of the QQ-Plot above verifies the gaussian errors assumption.

```
leveneTest(LOGCOT ~ TASK*WET, data = tobacco2, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value    Pr(>F)
## group  7 18.635 < 2.2e-16 ***
##      686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the Levene Test above, the p-value is $< 2.2e-16$, which means we reject the hypothesis that the homogeneity of variance assumption is satisfied. We should proceed with caution when using this model.

(b):

Based on this model, create a table of the estimated mean log cotinine levels, associated standard errors, and how each estimated mean is obtained from the model parameters (e.g., $\hat{\beta}_0 + \hat{\beta}_1$) for each task-wet combination.

```
beta <- intToUtf8(946)
params <- summary(fit)$coefficients[1:8,1:2] %>% as.data.frame()
params <- data.frame(params, "Parameter" = paste0(beta, c(1:8)))
params %>% knitr::kable(align = c("c", "c"))
```

	Estimate	Std..Error	Parameter
WET0TASK1	4.269337	0.1854712	β_1
WET1TASK1	4.613116	0.1226196	β_2
WET0TASK2	3.542748	0.2271549	β_3
WET1TASK2	3.667024	0.3013550	β_4
WET0TASK3	2.688185	0.2152536	β_5
WET1TASK3	2.879303	0.1187505	β_6
WET0TASK4	1.808889	0.1238562	β_7
WET1TASK4	1.810534	0.2130902	β_8

The estimates for mean log cotinine levels, their standard errors, and their relationship to the parameters are summarized in the table above. With cell mean coding, the parameter interpretations are clear; they each simply represent the mean log cotinine level for one task-wet combination.

part iii: The Full Model in Every Cell:

For these questions, use the log of salivary cotinine as the response and task and Innsmoke as predictors.

```
tobacco3 <- tobacco %>% mutate(LOGCOT = log(COTININE),
                                TASK1 = case_when(TASK == 1 ~ 1, TASK != 1 ~ 0),
                                TASK2 = case_when(TASK == 2 ~ 1, TASK != 2 ~ 0),
                                TASK3 = case_when(TASK == 3 ~ 1, TASK != 3 ~ 0),
                                TASK4 = case_when(TASK == 4 ~ 1, TASK != 4 ~ 0))
```

(a):

Fit the full model in every cell. Provide and interpret estimates of all parameters for this model.

Consider the following model using the reference cell coding scheme:

$$y = \beta_1 + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4} + \beta_5 X + \beta_6 I_{T2} X + \beta_7 I_{T3} X + \beta_8 I_{T4} X$$

where y is the log cotinine, X is the Innsmoke variable, and I_{Ti} is the indicator function associated with the i th task level. Note that TASK1 is the reference.

```
fit = lm(LOGCOT ~ factor(TASK) + factor(TASK)*LNNSMOKE, data = tobacco3)
summary(fit)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	4.3344142	0.09319895	46.507113	2.802358e-214
## factor(TASK)2	-1.2210290	0.19086147	-6.397462	2.924652e-10
## factor(TASK)3	-2.3221216	0.13561455	-17.122953	5.432182e-55
## factor(TASK)4	-3.4207011	0.13356892	-25.610007	4.927483e-102
## LNNSMOKE	0.2945994	0.08869134	3.321625	9.424000e-04
## factor(TASK)2:LNNSMOKE	0.4274696	0.16993370	2.515508	1.211344e-02
## factor(TASK)3:LNNSMOKE	0.9359153	0.12592833	7.432126	3.191248e-13
## factor(TASK)4:LNNSMOKE	1.4843094	0.13531844	10.969010	6.613806e-26

The estimates for each parameter can be seen by the summary above. The intercept, β_1 , is the mean log cotinine level for priming when Innsmoke is 0. The factor(TASK)2 estimate, for β_2 , is the difference between the mean log cotinine level for barning and for priming when Innsmoke is 0. Similarly, The factor(TASK)2 estimate, for β_3 , is the difference between the mean log cotinine level for topping and for priming when Innsmoke is 0 and the factor(TASK)4 estimate, for β_4 , is the difference between the mean log cotinine level for work not involving tobacco and for priming when Innsmoke is 0. The LNNSMOKE estimate, for β_5 , is the

mean increase log cotinine level for a one unit increase in $\ln(\text{smoke})$ (the natural log of 1 + number of cigarettes smoked a day) in those whose task is priming. Similarly, $\beta_6, \beta_7, \beta_8$ are the mean increase log cotinine level for a one unit increase in $\ln(\text{smoke})$ for those whose task is burning, topping, and no tobacco involvement, respectively.

(b):

Report an appropriate test of whether task is related to cotinine levels. If this test is significant, report step-down tests to determine exactly where differences lie. For all tests reported, be sure to state H_0 clearly and give explicit justification for which tests were used and why.

In order to test whether task is related to cotinine levels, we want to test the following hypotheses:

$$H_0 = 0 = \beta_2 = \beta_3 = \beta_4 \quad \& \quad 0 = \beta_6 = \beta_7 = \beta_8$$

which is equivalent to:

$$H_0 = \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_6 = 0, \beta_7 = 0, \beta_8 = 0,$$

These will test whether the intercepts for each task and the slopes for each task are equivalent to each other. In order to test these hypotheses, we can use the overall F test, where:

$$F = (SSH/a)/\hat{\sigma}^2 \sim F_{G-2, n-G}$$

where $SSH = (\hat{\theta} - \theta_0)'M^{-1}(\hat{\theta} - \theta_0)$, $G = 8$, $n = 694$, and $a^2 = \hat{\sigma}^2$. It should also be noted that $M = C(X'X)^{-1}C'$.

```
X = tobacco3 %>% mutate(INT = 1, LNNTASK2 = LNNSMOKE*TASK2, LNNTASK3 = LNNSMOKE*TASK3, LNNTASK4 = LNNSMOKE*TASK4) %>% select(INT, TASK2, TASK3, TASK4, LNNSMOKE, LNNTASK2, LNNTASK3, LNNTASK4) %>% as.matrix()
C = matrix(c(0, 0, 0, 0, 0, 0,
             1, 0, 0, 0, 0, 0,
             0, 1, 0, 0, 0, 0,
             0, 0, 1, 0, 0, 0,
             0, 0, 0, 0, 0, 0,
             0, 0, 0, 1, 0, 0,
             0, 0, 0, 0, 1, 0,
             0, 0, 0, 0, 0, 1), nrow = 6)

linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2 = 0
## factor(TASK)3 = 0
## factor(TASK)4 = 0
## factor(TASK)2:LNNSMOKE = 0
## factor(TASK)3:LNNSMOKE = 0
## factor(TASK)4:LNNSMOKE = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##      Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      692 1806.03
## 2      686  883.86   6    922.17 119.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 119.29 and the p-value is $< 2.2e-16$, so we can reject the null hypothesis. This means there is evidence to reject the fact that the slopes for all four task levels are equal and the intercepts for all four task levels are equal.

Next, we perform a step down test to test whether the intercepts are equal. The null hypothesis for this test will be the first set of hypotheses previously listed above.

```
C = matrix(c(0, 0, 0,
             1, 0, 0,
             0, 1, 0,
             0, 0, 1,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0), nrow = 3)

linearHypothesis(fit, C)
```



```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2 = 0
## factor(TASK)3 = 0
## factor(TASK)4 = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      689 1785.88
## 2      686  883.86   3    902.01 233.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 233.36 and the p-value is $< 2.2e-16$, so we can reject the null hypothesis. This means there is evidence to reject the fact the the intercepts for all four task levels are equal.

We will now follow the same process with the slopes for each task level. The null hypothesis for this test will be the second set of hypotheses previously listed above at the first test for this problem part.

```
C = matrix(c(0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             1, 0, 0,
             0, 1, 0,
             0, 0, 1), nrow = 3)

linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2:LNNSMOKE = 0
## factor(TASK)3:LNNSMOKE = 0
## factor(TASK)4:LNNSMOKE = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      689 1053.51
## 2      686  883.86   3    169.64 43.889 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 43.889 and the p-value is $< 2.2e-16$, so we can reject the null hypothesis. This means there is evidence to reject the fact the the slopes for all four task levels are equal.

Next, we would want to step down even further to test pairwise comparisons of the intercepts and of the slopes. For the intercepts, to test whether the intercepts for TASK1 and TASK2 are equivalent, we would want to test $H_0 : \beta_1 = \beta_1 + \beta_2 \rightarrow \beta_2 = 0$. To test the equivalence of TASK1 and TASK3, as well as TASK1 and TASK4, we can follow the same process. For TASK2 and TASK3, we would want to test $H_0 : \beta_2 = \beta_3 \rightarrow \beta_2 - \beta_3 = 0$. Similar hypotheses would be tested for TASK2 and TASK4, as well as TASK3 and TASK4. We will perform a normal F test, but will look to reject the null as p-values smaller than $\alpha = 0.05/6 = .008$ using the Bonferroni correction, since we are running 6 tests. All F-statistics have degrees of freedom (1, 686), since we are performing one test with a size $n - G$ fitted model.

```

### 1 and 2
C = matrix(c(0, 1, 0, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f12 <- test$F[2]
p12 <- test$`Pr(>F)`[2]

### 1 and 3
C = matrix(c(0, 0, 1, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f13 <- test$F[2]
p13 <- test$`Pr(>F)`[2]

### 1 and 4
C = matrix(c(0, 0, 0, 1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f14 <- test$F[2]
p14 <- test$`Pr(>F)`[2]

### 2 and 3
C = matrix(c(0, 1, -1, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f23 <- test$F[2]
p23 <- test$`Pr(>F)`[2]

### 2 and 4
C = matrix(c(0, 1, 0, -1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f24 <- test$F[2]
p24 <- test$`Pr(>F)`[2]

### 3 and 4
C = matrix(c(0, 0, 1, -1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f34 <- test$F[2]
p34 <- test$`Pr(>F)`[2]

df <- data.frame(Test = c("TASK1:TASK2", "TASK1:TASK3", "TASK1:TASK4", "TASK2:TASK3", "TASK2:TASK4", "TASK3:TASK4"), Fvalue = c(f12, f13, f14, f23, f24, f34), Pvalue = c(p12, p13, p14, p23, p24, p34))
df %>% knitr::kable(align = c("c", "c"), digits = c(10, 10, 20))

```

Test	Fvalue	Pvalue
TASK1:TASK2	40.92752	2.924652e-10
TASK1:TASK3	293.19551	0.000000e+00
TASK1:TASK4	655.87245	0.000000e+00
TASK2:TASK3	32.37628	1.882119e-08
TASK2:TASK4	131.13804	0.000000e+00
TASK3:TASK4	63.99177	5.325440e-15

From the summary table above, we can reject the null hypothesis in every case. That is, there is evidence to support the fact that all combinations of TASK intercepts are significantly different than one another. We will perform an analogous analysis for the pairwise testing of the equivalence of slopes of the TASK levels.

```

### 1 and 2
C = matrix(c(0, 0, 0, 0, 0, 1, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f12 <- test$F[2]
p12 <- test$`Pr(>F)`[2]

### 1 and 3
C = matrix(c(0, 0, 0, 0, 0, 0, 1, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f13 <- test$F[2]
p13 <- test$`Pr(>F)`[2]

### 1 and 4
C = matrix(c(0, 0, 0, 0, 0, 0, 0, 1), nrow = 1)
test<-linearHypothesis(fit, C)
f14 <- test$F[2]
p14 <- test$`Pr(>F)`[2]

### 2 and 3
C = matrix(c(0, 0, 0, 0, 0, 1, -1, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f23 <- test$F[2]
p23 <- test$`Pr(>F)`[2]

### 2 and 4
C = matrix(c(0, 0, 0, 0, 0, 1, 0, -1), nrow = 1)
test<-linearHypothesis(fit, C)
f24 <- test$F[2]
p24 <- test$`Pr(>F)`[2]

### 3 and 4
C = matrix(c(0, 0, 0, 0, 0, 0, 1, -1), nrow = 1)
test<-linearHypothesis(fit, C)
f34 <- test$F[2]
p34 <- test$`Pr(>F)`[2]

df <- data.frame(Test = c("TASK1:TASK2", "TASK1:TASK3", "TASK1:TASK4", "TASK2:TASK3", "TASK2:TASK4", "TASK3:TASK4"), Fvalue = c(f12, f13, f14
, f23, f24, f34), Pvalue = c(p12, p13, p14, p23, p24, p34))
df %>% knitr::kable(align = c("c", "c"), digits = c(10, 10, 20))

```

Test	Fvalue	Pvalue
TASK1:TASK2	6.327780	1.211344e-02
TASK1:TASK3	55.236500	3.191248e-13
TASK1:TASK4	120.319184	0.000000e+00
TASK2:TASK3	8.913427	2.931579e-03
TASK2:TASK4	35.506803	4.063003e-09
TASK3:TASK4	16.311803	5.979385e-05

From the summary table above, we can reject the null hypothesis in almost every case. That is, there is evidence to support the fact that all combinations of TASK slopes (except TASK1:TASK2) are significantly different than one another. the p-value for TASK1:TASK2 is greater than the Bonferroni corrected $\alpha = 0.008$ so there is not evidence to reject the fact that the mean increase in log cotinine levels when Innsmoke is increased by one is different between priming and barning tasks.

```

1  ods pdf file = "/home/sedwards/UNC/B663_H4.pdf";
2  ods graphics on;
3
4  LIBNAME home "/home/sedwards/UNC/663/";
5  %LET filepath = /home/sedwards/UNC/663/;
6
7  /**** Question #1 ****
8  172 young adult males received a battery of pulmonary function tests.
9
10 Fit a model with average forced vital capacity (FVC) (in ml) as the outcome
11 with height, weight, body mass index, age, average treadmill elivation,
12     average treadmill speed, temperature, barometric pressure, and humidity
13     as predictors.
14 Consider the model -
15     FVC = B0 + B1(height) + B2(weight) + B3(BMI) + B4(area) + B5(age) +
16         B6(avtrel) + B7(avtrsp)
17         + B8(avtrel)(avtrsp) + B9(temp) + B10(barm) + B11(hum) + e
18
19 ***/
20 TITLE "QUESTION 1";
21 /* READ IN DATA & SET UP VARIABLES */
22 DATA filen;
23     INFILE "&filepath.\FILEN.DAT";
24     INPUT subject year cohort date days timsess height weight age area temp
25         barm
26         hum avtrel avtrsp avfvc;
27     LABEL
28     subject="subject id"
29     year="year of study"
30     cohort="Ozone Dosage Level Group"
31     date="Date of Study, Julian Date"
32     days="# Days After 12-31-79"
33     timsess="Time of Session"
34     height="Height (cm)"
35     weight="Weight (kg)"
36     age="Age (years)"
37     area="Body Surface Area (M**2)"
38     temp="Air Temperature (deg C)"
39     barm="Barometric Pressure (mmHg)"
40     hum="Relative Humidity %"
41     avtrel="Average Treadmill Elevation (deg)"
42     avtrsp="Average Speed of Treadmill (mph)"
43     avfvc="Average Forced Vital Capacity (mL)";
44 RUN;
45
46 DATA filen;
47     SET filen;
48
49     int=avtrel*avtrsp;
50     bmi=10000*weight/(height*height);
51     tim2=timsess*timsess;
52 RUN;
53
54 PROC FREQ DATA=filen;
55     TABLES height weight BMI area avtrel avtrsp int temp barm hum age / LIST
56     MISSING;
57 RUN;
58
59 PROC FREQ DATA=filen;
60     WHERE weight=.;
61     TABLES weight*bmi / LIST MISSING;
62 RUN;
63
64 /***

```

```

63 (a) Examine the tolerances and variance inflation factors (VIF) from this
    model.
64 Do you think any collinearity is present based on the tolerance and VIF?
    Why or why not?
65 ***/
66 PROC REG DATA=filen;
67     MODEL avfvc = height weight BMI area age avtrel avtrsp int temp barm hum
        / TOL VIF;
68     TITLE "Q1-A";
69 QUIT;
70
71 /***
72 (b) Conduct an eigenanalysis of the scaled SSCP and correlation matrices,
    presenting a table formatted
73     like Table 8.6.2.
74
75     i) Does there appear to be collinearity with the intercept? Why or why not?
76         If so which variables are suspect?
77
78     ii) Does there appear to be any other collinearity? Why or why not?
79         If so which variables are suspect?
80 ***/
81 DATA B;
82     SET filen;
83     inter = 1;
84 RUN;
85 ods output Eigenvalues=EIa;
86 PROC PRINCOMP DATA=B NOINT;          /* SCALED SSCP MATRIX */
87     VAR inter height weight BMI area age avtrel avtrsp int temp barm hum;
88     TITLE "Q1-B";
89 RUN;
90
91 DATA EIa;
92     IF _N_ = 1 THEN DO; SET EIa(RENAME=(eigenvalue=e1)); WHERE number = 1; END;
93     SET EIa;
94     CondIndex = sqrt(e1/eigenvalue);
95 RUN;
96 PROC PRINT DATA=EIa;    RUN;
97
98 ods output Eigenvalues=EIb;
99 PROC PRINCOMP DATA=B;          /* CORRELATION MATRIX */
100     VAR inter height weight BMI area age avtrel avtrsp int temp barm hum;
101 RUN;
102 DATA EIb;
103     IF _N_ = 1 THEN DO; SET EIb(RENAME=(eigenvalue=e1)); WHERE number = 1; END;
104     SET EIb;
105     CondIndex = sqrt(e1/eigenvalue);
106 RUN;
107 PROC PRINT DATA=EIb; RUN;
108
109
110 /***** Question #2 *****/
111 Find the Box-Cox transformation of the simulated data (boxcox.dat) and
112     compare the residual plots of the raw and transformed data.
113 ***/
114 TITLE "QUESTION 2";
115 /* READ IN DATA & SET UP VARIABLES */
116 DATA filen;
117     INFILE "&filepath.BoxCox.dat";
118     INPUT x y;
119 RUN;
120
121 /* BEST LAMBDA = 0.5 */
122 PROC TRANSREG DATA=filen SS2 DETAIL;
123     TITLE "DEFAULTS";
124     MODEL BOXCOX(y) = IDENTITY(x);

```

```

125 RUN;
126
127 PROC IML;
128 USE filen;
129 READ ALL VAR "y" INTO y;
130 lny=log(y);
131 n=nrow(y);
132 one=j(n,1,1);
133 avglog=lny`*one/n;
134 PRINT avglog;
135 geomean = exp(avglog);
136 PRINT geomean;
137 geomeany=geomean#one;
138 create gmean var {geomeany};
139 append from geomeany;
140 close gmean;
141 QUIT;
142 RUN;
143
144 DATA p1;
145     MERGE gmean filen;
146
147     y_5 = ((y**0.5)-1)/(0.5*(geomeany**(0.5-1)));
148 RUN;
149
150 ODS GRAPHICS ON;
151 PROC REG DATA=p1 PLOTS=ALL;
152     MODEL y = x;
153     TITLE "1-RAW";
154 QUIT;
155
156 PROC REG DATA=p1 PLOTS=ALL;
157     MODEL y_5 = x;
158     TITLE "1-TRANSFORMED";
159 QUIT;
160 ODS GRAPHICS OFF;
161
162
163 /***** Question #3 *****/
164 Effect of dermal nicotine exposure in a population of Latino tobacco workers
165 in North Carolina.
166 ***/
167 TITLE "QUESTION 3";
168
169 /* READ IN DATA & SET UP VARIABLES */
170 DATA filen;
171     INFILE "&filepath.\tobacco.dat";
172     INPUT cotinine age bmi educ wet task lnnsnsmoke;
173
174     ID = _N_;
175     lnCot = LOG(cotinine);
176
177     IF task=1 THEN t1=1; ELSE t1=0;
178     IF task=2 THEN t2=1; ELSE t2=0;
179     IF task=3 THEN t3=1; ELSE t3=0;
180     IF task=4 THEN t4=1; ELSE t4=0;
181
182     wet_0 = 1 - wet;
183     wet_1 = wet;
184
185     w0t1 = wet_0 AND t1;
186     w0t2 = wet_0 AND t2;
187     w0t3 = wet_0 AND t3;
188     w0t4 = wet_0 AND t4;
189     w1t1 = wet_1 AND t1;
190     w1t2 = wet_1 AND t2;

```



```

254                                one 0 t1 0 t2 0 t3 1;
255        CONTRAST "T1 > AVG(T234)" one 0 t1 1 t2 -0.333333 t3 -0.333333;
256
257        TITLE "REF CELL - 2";
258    QUIT;
259
260    PROC GLM DATA=filen;
261        CLASS task;
262        MODEL lnCot = task / NOINT;
263        LSMEANS task / PDIF F ADJUST=SCHEFFE;
264        MEANS task / SCHEFFE;
265
266        TITLE "SCHEFFE";
267    QUIT;
268
269
270    /***
271    B - TWO-WAY ANOVA - LOG(cotinine) = task + wet
272
273    lnCot = mu + alpha i + beta j + gamma ij
274
275    i=2      j=4      ij=8
276
277    1 + 2 + 4 + 8 = 15
278
279    let wet=1 and task=4 be the reference
280    i=1
281    j=3      1 + 1 + 3 + 3 = 8
282
283    A(mu) + B(wet_0) + C(t1) + D(t2) + E(t3) + BC(w0t1) + BD (w0t2) + BE (w0t3)
284
285    now 4 parameters describe the dose effect
286
287    Fit model with full interaction.
288        - Interpret all parameter estimates.
289        - Clearly state coding scheme used.
290        - Discuss HILE-G assumptions.
291    Create a table of estimated mean log(cotinine) levels, standard errors, how
    each estimated mean is obtained from the model parameters
292    ***/
293    PROC FREQ DATA=filen;
294        TABLES task*wet / MISSING NOROW NOCOL NOPERCENT;
295        TITLE "Balanced Cells?";
296    RUN;
297
298    PROC GLM DATA=filen;
299        MODEL lnCot = w0t1 w1t1 w0t2 w1t2 w0t3 w1t3 w0t4 w1t4 / NOINT SOLUTION;
300
301        ESTIMATE "GRAND MEAN" w0t1 1 w1t1 1 w0t2 1 w1t2 1 w0t3 1 w1t3 1 w0t4 1
    w1t4 1 / DIVISOR=8;
302        ESTIMATE "MARG MEAN: WET 0" w0t1 1 w1t1 0 w0t2 1 w1t2 0 w0t3 1 w1t3 0
    w0t4 1 w1t4 0 / DIVISOR=4;
303        ESTIMATE "MARG MEAN: WET 1" w0t1 0 w1t1 1 w0t2 0 w1t2 1 w0t3 0 w1t3 1
    w0t4 0 w1t4 1 / DIVISOR=4;
304        ESTIMATE "MARG MEAN: T1" w0t1 1 w1t1 1 w0t2 0 w1t2 0 w0t3 0 w1t3 0 w0t4 0
    w1t4 0 / DIVISOR=2;
305        ESTIMATE "MARG MEAN: T2" w0t1 0 w1t1 0 w0t2 1 w1t2 1 w0t3 0 w1t3 0 w0t4 0
    w1t4 0 / DIVISOR=2;
306        ESTIMATE "MARG MEAN: T3" w0t1 0 w1t1 0 w0t2 0 w1t2 0 w0t3 1 w1t3 1 w0t4 0
    w1t4 0 / DIVISOR=2;
307        ESTIMATE "MARG MEAN: T4" w0t1 0 w1t1 0 w0t2 0 w1t2 0 w0t3 0 w1t3 0 w0t4 1
    w1t4 1 / DIVISOR=2;
308
309        CONTRAST "MAIN EFFECT WET" w0t1 1 w1t1 -1 w0t2 1 w1t2 -1 w0t3 1 w1t3 -1
    w0t4 1 w1t4 -1;
310

```



```

311     CONTRAST "MAIN EFFECT TASK" w0t1 1 w1t1 1 w0t2 -1 w1t2 -1 w0t3 0 w1t3 0
      w0t4 0 w1t4 0,
312                                     w0t1 1 w1t1 1 w0t2 0 w1t2 0 w0t3 -1 w1t3 -1
      w0t4 0 w1t4 0,
313                                     w0t1 1 w1t1 1 w0t2 0 w1t2 0 w0t3 0 w1t3 0
      w0t4 -1 w1t4 -1;
314
315     CONTRAST "INTERACTION WET V TASK" w0t1 1 w1t1 -1 w0t2 -1 w1t2 1 w0t3 0
      w1t3 0 w0t4 0 w1t4 0,
316                                     w0t1 1 w1t1 -1 w0t2 0 w1t2 0 w0t3 -1
      w1t3 1 w0t4 0 w1t4 0,
317                                     w0t1 1 w1t1 -1 w0t2 0 w1t2 0 w0t3 0
      w1t3 0 w0t4 -1 w1t4 1;
318
319     TITLE "CELL MEAN";
320     QUIT;
321
322     PROC GLM DATA=filen;
323         MODEL lnCot = one wet_1 t2 t3 t4 w1t2 w1t3 w1t4 / NOINT SOLUTION;
324
325         ESTIMATE "GRAND MEAN" one 8 wet_1 4 t2 2 t3 2 t4 2 w1t2 1 w1t3 1 w1t4 1 /
      DIVISOR=8;
326         ESTIMATE "MARG MEAN: WET 0" one 4 wet_1 0 t2 1 t3 1 t4 1 w1t2 0 w1t3 0
      w1t4 0 / DIVISOR=4;
327         ESTIMATE "MARG MEAN: WET 1" one 4 wet_1 4 t2 1 t3 1 t4 1 w1t2 1 w1t3 1
      w1t4 1 / DIVISOR=4;
328         ESTIMATE "MARG MEAN: T1" one 2 wet_1 1 t2 0 t3 0 t4 0 w1t2 0 w1t3 0 w1t4
      0 / DIVISOR=2;
329         ESTIMATE "MARG MEAN: T2" one 2 wet_1 1 t2 2 t3 0 t4 0 w1t2 1 w1t3 0 w1t4
      0 / DIVISOR=2;
330         ESTIMATE "MARG MEAN: T3" one 2 wet_1 1 t2 0 t3 2 t4 0 w1t2 0 w1t3 1 w1t4
      0 / DIVISOR=2;
331         ESTIMATE "MARG MEAN: T4" one 2 wet_1 1 t2 0 t3 0 t4 2 w1t2 0 w1t3 0 w1t4
      1 / DIVISOR=2;
332
333         CONTRAST "MAIN EFFECT WET" one 0 wet_1 4 t2 0 t3 0 t4 0 w1t2 1 w1t3 1
      w1t4 1;
334
335         CONTRAST "MAIN EFFECT TASK" one 0 wet_1 0 t2 2 t3 0 t4 0 w1t2 1 w1t3 0
      w1t4 0,
336                                     one 0 wet_1 0 t2 0 t3 2 t4 0 w1t2 0 w1t3 1
      w1t4 0,
337                                     one 0 wet_1 0 t2 0 t3 0 t4 2 w1t2 0 w1t3 0
      w1t4 1;
338
339         CONTRAST "INTERACTION WET V TASK" one 0 wet_1 0 t2 0 t3 0 t4 0 w1t2 1
      w1t3 0 w1t4 0,
340                                     one 0 wet_1 0 t2 0 t3 0 t4 0 w1t2 0
      w1t3 1 w1t4 0,
341                                     one 0 wet_1 0 t2 0 t3 0 t4 0 w1t2 0
      w1t3 0 w1t4 1;
342
343     TITLE "REF CELL";
344     QUIT;
345
346     ODS GRAPHICS ON;
347     PROC REG DATA=filen PLOTS=ALL;
348         MODEL lnCot = one wet_1 t2 t3 t4 w1t2 w1t3 w1t4 / NOINT;
349         TITLE "3-B";
350         OUTPUT OUT=B PREDICTED=y_hat RSTUDENT=r_i;
351     QUIT;
352
353     PROC UNIVARIATE DATA=B PLOT NORMAL;
354         CLASS wet task;
355         VAR r_i;
356         QQPLOT r_i / NORMAL;

```

```

357     HISTOGRAM r_i / NORMAL;
358     TITLE "3-B";
359 RUN;
360
361 PROC MEANS DATA=B STD;
362     CLASS y_hat;
363     VAR r_i;
364     OUTPUT OUT=B_2 STD(r_i)=sd;
365 RUN;
366
367 PROC SGPLOT DATA=B_2;
368     SCATTER X=y_hat Y=sd;
369     label sd="Within Group Sample SDs";
370     TITLE "Within Group Residuals";
371 RUN;
372
373 ODS GRAPHICS OFF;
374
375
376 /***
377 C - FULL MODEL IN EVERY CELL - LOG(cotinine) = task + lnnsnake - categorical
    & continuous
378
379     Fit full model in every cell.
380     - Interpret all parameter estimates.
381     Report test for whether task is related to cotinine levels.
382     - If sig, report step-down tests to determine exactly where differences
        lie.
383     - State H0 & Give justification for which test is used and why.
384 ***/
385 PROC MEANS DATA=filen;
386     VAR lnnsnake;          /* mean = 0.5960 */
387 RUN;
388
389 PROC GLM DATA=filen;
390     MODEL lnCot = t1 t2 t3 t4 lnsnkt1 lnsnkt2 lnsnkt3 lnsnkt4 / NOINT SOLUTION;
391
392     ESTIMATE "ADJ CELL MEAN: T1" t1 1 t2 0 t3 0 t4 0 lnsnkt1 0.5960 lnsnkt2 0
        lnsnkt3 0 lnsnkt4 0;
393     ESTIMATE "ADJ CELL MEAN: T2" t1 0 t2 1 t3 0 t4 0 lnsnkt1 0 lnsnkt2 0.5960
        lnsnkt3 0 lnsnkt4 0;
394     ESTIMATE "ADJ CELL MEAN: T3" t1 0 t2 0 t3 1 t4 0 lnsnkt1 0 lnsnkt2 0
        lnsnkt3 0.5960 lnsnkt4 0;
395     ESTIMATE "ADJ CELL MEAN: T4" t1 0 t2 0 t3 0 t4 1 lnsnkt1 0 lnsnkt2 0
        lnsnkt3 0 lnsnkt4 0.5960;
396
397     ESTIMATE "MEAN OF ADJ CELL MEANS" t1 1 t2 1 t3 1 t4 1 lnsnkt1 0.5960
        lnsnkt2 0.5960 lnsnkt3 0.5960 lnsnkt4 0.5960 / DIVISOR=4;
398     ESTIMATE "MEAN INTERCEPT" t1 1 t2 1 t3 1 t4 1 lnsnkt1 0 lnsnkt2 0 lnsnkt3
        0 lnsnkt4 0 / DIVISOR=4;
399     ESTIMATE "MEAN SLOPE" t1 0 t2 0 t3 0 t4 0 lnsnkt1 1 lnsnkt2 1 lnsnkt3 1
        lnsnkt4 1 / DIVISOR=4;
400
401     CONTRAST "TEST OF COINCIDENCE" t1 1 t2 -1 t3 0 t4 0 lnsnkt1 0 lnsnkt2
        0 lnsnkt3 0 lnsnkt4 0,
402                                     t1 1 t2 0 t3 -1 t4 0 lnsnkt1 0 lnsnkt2
        0 lnsnkt3 0 lnsnkt4 0,
403                                     t1 1 t2 0 t3 0 t4 -1 lnsnkt1 0 lnsnkt2
        0 lnsnkt3 0 lnsnkt4 0,
404                                     t1 0 t2 0 t3 0 t4 0 lnsnkt1 1 lnsnkt2
        -1 lnsnkt3 0 lnsnkt4 0,
405                                     t1 0 t2 0 t3 0 t4 0 lnsnkt1 1 lnsnkt2
        0 lnsnkt3 -1 lnsnkt4 0,
406                                     t1 0 t2 0 t3 0 t4 0 lnsnkt1 1 lnsnkt2
        0 lnsnkt3 0 lnsnkt4 -1;
407

```

```

408 CONTRAST "STEPDOWN: EQUAL INTERCEPTS"
409          t1 1 t2 -1 t3 0 t4 0 lnsmt1 0 lnsmt2
          0 lnsmt3 0 lnsmt4 0,
410          t1 1 t2 0 t3 -1 t4 0 lnsmt1 0 lnsmt2
          0 lnsmt3 0 lnsmt4 0,
411          t1 1 t2 0 t3 0 t4 -1 lnsmt1 0 lnsmt2
          0 lnsmt3 0 lnsmt4 0;
412 CONTRAST "STEPDOWN: EQUAL SLOPES"
413          t1 0 t2 0 t3 0 t4 0 lnsmt1 1 lnsmt2
          -1 lnsmt3 0 lnsmt4 0,
414          t1 0 t2 0 t3 0 t4 0 lnsmt1 1 lnsmt2
          0 lnsmt3 -1 lnsmt4 0,
415          t1 0 t2 0 t3 0 t4 0 lnsmt1 1 lnsmt2
          0 lnsmt3 0 lnsmt4 -1;
416
417 CONTRAST "PAIRWISE INTERCEPTS T1 V T2" t1 1 t2 -1 t3 0 t4 0 lnsmt1 0
lnsmt2 0 lnsmt3 0 lnsmt4 0;
418 CONTRAST "PAIRWISE INTERCEPTS T1 V T3" t1 1 t2 0 t3 -1 t4 0 lnsmt1 0
lnsmt2 0 lnsmt3 0 lnsmt4 0;
419 CONTRAST "PAIRWISE INTERCEPTS T1 V T4" t1 1 t2 0 t3 0 t4 -1 lnsmt1 0
lnsmt2 0 lnsmt3 0 lnsmt4 0;
420 CONTRAST "PAIRWISE INTERCEPTS T2 V T3" t1 0 t2 1 t3 -1 t4 0 lnsmt1 0
lnsmt2 0 lnsmt3 0 lnsmt4 0;
421 CONTRAST "PAIRWISE INTERCEPTS T2 V T4" t1 0 t2 1 t3 0 t4 -1 lnsmt1 0
lnsmt2 0 lnsmt3 0 lnsmt4 0;
422 CONTRAST "PAIRWISE INTERCEPTS T3 V T4" t1 0 t2 0 t3 1 t4 -1 lnsmt1 0
lnsmt2 0 lnsmt3 0 lnsmt4 0;
423
424 CONTRAST "PAIRWISE SLOPES T1 V T2" t1 0 t2 0 t3 0 t4 0 lnsmt1 1
lnsmt2 -1 lnsmt3 0 lnsmt4 0;
425 CONTRAST "PAIRWISE SLOPES T1 V T3" t1 0 t2 0 t3 0 t4 0 lnsmt1 1
lnsmt2 0 lnsmt3 -1 lnsmt4 0;
426 CONTRAST "PAIRWISE SLOPES T1 V T4" t1 0 t2 0 t3 0 t4 0 lnsmt1 1
lnsmt2 0 lnsmt3 0 lnsmt4 -1;
427 CONTRAST "PAIRWISE SLOPES T2 V T3" t1 0 t2 0 t3 0 t4 0 lnsmt1 0
lnsmt2 1 lnsmt3 -1 lnsmt4 0;
428 CONTRAST "PAIRWISE SLOPES T2 V T4" t1 0 t2 0 t3 0 t4 0 lnsmt1 0
lnsmt2 1 lnsmt3 0 lnsmt4 -1;
429 CONTRAST "PAIRWISE SLOPES T3 V T4" t1 0 t2 0 t3 0 t4 0 lnsmt1 0
lnsmt2 0 lnsmt3 1 lnsmt4 -1;
430
431 CONTRAST "EQUAL ADJ CELL MEANS" t1 1 t2 -1 t3 0 t4 0 lnsmt1 0.5960
lnsmt2 -0.5960 lnsmt3 0 lnsmt4 0,
432          t1 1 t2 0 t3 -1 t4 0 lnsmt1 0.5960
          lnsmt2 0 lnsmt3 -0.5960 lnsmt4 0,
433          t1 1 t2 0 t3 0 t4 -1 lnsmt1 0
          lnsmt2 0 lnsmt3 0 lnsmt4 -0.5960;
434 CONTRAST "PAIRWISE ADJ T1 V T2" t1 1 t2 -1 t3 0 t4 0 lnsmt1 0.5960
lnsmt2 -0.5960 lnsmt3 0 lnsmt4 0;
435 CONTRAST "PAIRWISE ADJ T1 V T3" t1 1 t2 0 t3 -1 t4 0 lnsmt1 0.5960
lnsmt2 0 lnsmt3 -0.5960 lnsmt4 0;
436 CONTRAST "PAIRWISE ADJ T1 V T4" t1 1 t2 0 t3 0 t4 -1 lnsmt1 0.5960
lnsmt2 0 lnsmt3 0 lnsmt4 -0.5960;
437 CONTRAST "PAIRWISE ADJ T2 V T3" t1 0 t2 1 t3 -1 t4 0 lnsmt1 0
lnsmt2 0.5960 lnsmt3 -0.5960 lnsmt4 0;
438 CONTRAST "PAIRWISE ADJ T2 V T4" t1 0 t2 1 t3 0 t4 -1 lnsmt1 0
lnsmt2 0.5960 lnsmt3 0 lnsmt4 -0.5960;
439 CONTRAST "PAIRWISE ADJ T3 V T4" t1 0 t2 0 t3 1 t4 -1 lnsmt1 0
lnsmt2 0 lnsmt3 0.5960 lnsmt4 -0.5960;
440
441 TITLE "3-C CELL MEANS";
442 QUIT;
443
444 PROC GLM DATA=filen;
445 CLASS task;
446 MODEL lnCot = task lnnsnsmoke / NOINT;

```

```
447         MEANS task / SCHEFFE;  
448         LSMEANS task / PDIFF ADJUST=SCHEFFE;  
449         TITLE "SCHEFFE";  
450 QUIT;  
451  
452
```

1. The following questions are on the data and model described in Q3 of HW3:
 - Examine the tolerances and variance inflation factors from this model. Do you think any collinearity is present based on the tolerance and VIF? Why or why not?

Table 1.1: Tolerances & Variance Inflation Factors

Variable	Label	DF	Tolerance	Variance Inflation
Intercept	Intercept	1	.	0
height	Height (cm)	1	0.00218	458.04764
weight	Weight (kg)	1	0.00142	703.44629
bmi		1	0.00564	177.45037
area	Body Surface Area (M**2)	1	0.00073266	1364.89752
age	Age (years)	1	0.92307	1.08334
avtrel	Average Treadmill Elevation (deg)	1	0.00172	580.38969
avtrsp	Average Speed of Treadmill (mph)	1	0.01276	78.39302
int		1	0.00126	795.44895
temp	Air Temperature (deg C)	1	0.03447	29.01379
barm	Barometric Pressure (mmHg)	1	0.94420	1.05910
hum	Relative Humidity %	1	0.03429	29.16692

- There is a good amount of evidence to suggest collinearity in:
 - BMI and body surface area can be predicted/derived from height and weight.
 - Average treadmill elevation, average treadmill speed, and their interaction are collinear.
 - Temperature and relative humidity can be somewhat “predicted” once one of them is known.
- Age and barometric pressure do not appear to be collinear. This makes sense since age cannot guarantee height and weight among adults. Barometric pressure would probably tend to stay constant or within a narrow range of values while humidity can vary much more.

- Conduct an eigenanalysis of the scaled SSCP and correlation matrices, presenting a table formatted like Table 8.6.2 in Muller and Fetterman.
 - Does there appear to be any collinearity between the intercept and the covariates? Why or why not? If so, list the variables?

Table 1.2: Eigenvalues & Condition Index – Scaled SSCP Matrix

Number	Eigenvalue	CondIndex
1	11.9049480	1.00
2	0.0360383	18.18
3	0.0293709	20.13
4	0.0161788	27.13
5	0.0066991	42.16
6	0.0049049	49.27
7	0.0015550	87.50
8	0.0002515	217.55
9	0.0000375	563.68
10	0.0000097	1109.52
11	0.0000049	1563.90
12	0.0000015	2772.26

Table 1.3: Eigenvectors – Scaled SSCP Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
inter	0.289612	0.003847	-.041736	0.011264	-.428489	0.169349
height	0.289592	-.012406	-.076762	-.085857	-.320422	-.333604
weight	0.288434	-.239953	-.360268	-.362654	0.422414	-.270732
bmi	0.288728	-.206145	-.292000	-.169993	0.215099	0.744305
area	0.289363	-.111646	-.201362	-.220653	0.010005	-.386236
age	0.287541	-.065062	-.310877	0.876080	0.184729	-.112124
avtrrel	0.288114	0.542971	0.114554	-.008826	0.189245	0.156326
avtrsp	0.289555	0.081696	-.006435	-.063918	-.344231	-.063169
int	0.287589	0.621228	0.147453	-.079888	0.268208	-.082800
temp	0.288619	-.262600	0.436206	0.056399	0.018763	0.043500
barm	0.289608	0.005253	-.047742	0.006020	-.425070	0.178686
hum	0.287332	-.355890	0.642383	0.046041	0.217990	-.043461
	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
inter	-.160576	0.078500	-.338231	-.161589	0.611844	-.393890
height	-.210883	0.038668	-.291386	0.338828	0.023636	0.665223
weight	-.018722	-.039185	0.238215	-.394929	0.334969	0.139220
bmi	0.123098	0.027757	-.194962	0.269726	-.143594	0.111429
area	-.132498	0.015069	-.118058	0.290762	-.440686	-.595672
age	0.045774	0.011125	0.001522	-.001560	-.002277	-.001034
avtrrel	-.507446	0.047211	-.121532	-.414212	-.312294	0.071559
avtrsp	0.689454	-.012728	-.115575	-.438188	-.311166	0.071916
int	0.359344	-.055730	0.120806	0.416970	0.313806	-.072700
temp	-.078823	-.802916	0.031595	0.002606	-.009520	-.007541
barm	-.155432	0.124637	0.804979	0.091474	-.071280	0.006178
hum	0.047328	0.569193	-.016930	0.000288	0.007634	0.005636

Several of the scaled SSCP condition indices are greater than 30 and multiple eigenvalues are close to zero; thus, indicating collinearity with the intercept.

By looking at the elements of the 12th Principal Component, the elements corresponding to the intercept, height, weight, bmi, and area have values farther from zero relative to the other variables.

This would suggest that height, weight, bmi, and area span the intercept.

(b) Does there appear to be any collinearity among the covariates? Why or why not? If so, list the variables?

Table 1.4: Eigenvalues & Condition Index - Correlation

Number	Eigenvalue	CondIndex
1	3.00984215	1.0000
2	2.44782689	1.1089
3	2.02476481	1.2192
4	1.11320127	1.6443
5	1.01325013	1.7235
6	0.80927943	1.9285
7	0.56109511	2.3161
8	0.01770758	13.0374
9	0.00187374	40.0791
10	0.00070517	65.3317
11	0.00045372	81.4475

Table 1.5: Eigenvectors – Correlation

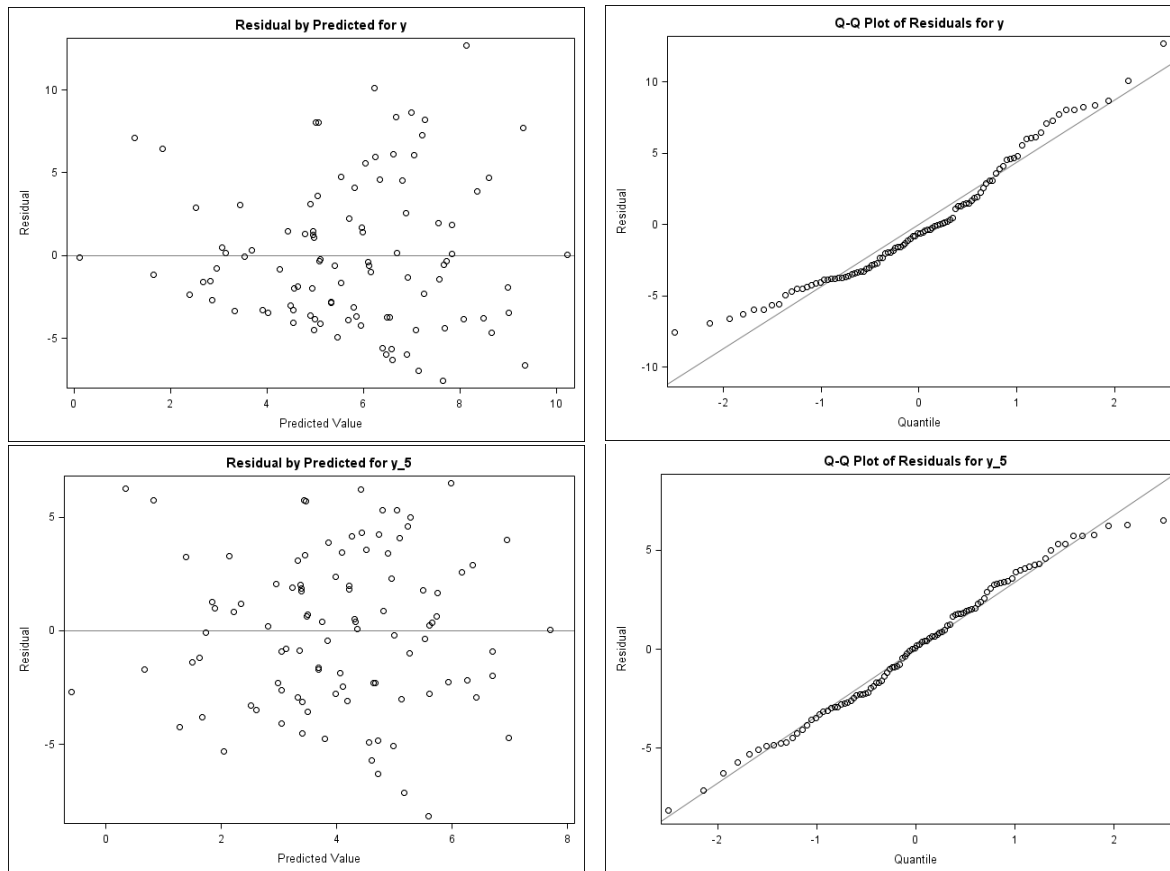
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
height	0.392316	0.296786	0.032498	-.431698	0.052791	0.294700
weight	0.558767	0.051052	-.092125	0.096942	0.000376	-.170215
bmi	0.367229	-.183577	-.143312	0.493420	-.041432	-.472648
area	0.549273	0.157720	-.048934	-.111776	0.019229	-.000896
age	0.092890	-.065940	-.101323	0.368850	0.755491	0.497666
avtrrel	-.199140	0.514774	-.034223	0.298870	0.116311	-.185008
avtrsp	0.074811	0.479606	0.086722	-.079394	-.115421	0.054060
int	-.148534	0.589598	-.004523	0.228684	0.068080	-.138848
temp	0.091766	-.044660	0.676385	0.168590	-.008712	0.038468
barm	0.065437	0.035488	-.176120	0.457649	-.626029	0.594893
hum	0.093072	-.022211	0.678255	0.162613	-.029775	0.036111
	Prin7	Prin8	Prin9	Prin10	Prin11	
height	-.259751	0.018476	0.520184	-.036531	0.375786	
weight	-.058566	0.003991	-.634360	-.174066	0.449619	
bmi	0.168277	0.006166	0.554743	0.038138	0.069126	
area	-.143725	-.017816	-.095960	0.175724	-.772725	
age	0.148534	-.013550	-.001968	0.003505	-.002419	
avtrrel	-.397123	-.000850	-.061237	0.611781	0.146172	
avtrsp	0.825059	0.015953	-.032894	0.221914	0.054851	
int	-.081491	0.009784	0.071012	-.715838	-.172852	
temp	-.055085	0.706224	-.008134	0.004308	-.015415	
barm	-.090203	0.006020	0.003518	-.000885	-.001557	
hum	-.042697	-.707081	0.009525	-.007438	0.015737	

From the eigenanalysis of the correlation matrix, there appears to be other collinearity within the variables besides the intercept.

By looking at the elements of the 11th Principal Component, the elements corresponding to the height, weight, bmi, area, average treatmill elevation, average treadmill speed, the interaction, temperature, and humidity have values farther from zero relative to age and barometric pressure.

- Find the Box-Cox transformation of the simulated data (BoxCox.dat) and compare the residual plots of the raw and transformed data.

Box-Cox Transformation Information for y				
Lambda		R-Square	Log Like	
0.00		0.19	-147.31	
0.25		0.19	-124.38	*
0.50	+	0.18	-122.59	<
0.75		0.17	-131.76	
1.00		0.16	-147.64	
< - Best Lambda				
* - 95% Confidence Interval				
+ - Convenient Lambda				



The residual plot for the non-transformed y-values shows a slight pattern with the residuals being clustered closer together for smaller predicted values of y with the dispersion increasing as the predicted values of y increase. This indicates that the assumptions of homogeneity of variance and linearity might be violated. Due to curvature, the Q-Q plot for the non-transformed y-values also indicates that the assumption of Gaussian errors might be violated.

After transforming the y-values using a Box-Cox transformation with $\lambda=0.50$, the residual plot is scattered equally regardless of the predicted value of y₅ and the Q-Q plot no longer contains a curve in the middle. This suggests the Box-Cox transformation of y does not violate the assumptions of homogeneity, linearity, nor Gaussian errors.

3. Investigators are interested in the effect of dermal nicotine exposure in a population of Latino tobacco workers in North Carolina. (Nicotine can be absorbed from tobacco leaves through the skin and can cause nicotine poisoning, which is characterized by nausea, vomiting, headache, and dizziness.) Data were collected on tobacco work tasks and risk factors for exposure to nicotine during a summer tobacco work season. Nicotine exposure was measured by levels of cotinine, a nicotine metabolite, contained in saliva. Other covariates of interest include age, body mass index, education, work conditions (working in wet conditions is believed to increase nicotine absorption), type of tobacco work ("priming" refers to picking or harvesting the tobacco and is expected to result in highest nicotine exposures, "barning" refers to putting the harvested tobacco into a barn for curing, "topping" refers to breaking the flower off the top of the plant, and "other" refers to farm work that does not involve tobacco contact, such as driving a truck), and smoking (smokers would also have nicotine exposure through cigarettes, and it is not known whether exposure to tobacco leaves would increase cotinine levels to a similar extent in both smokers and non-smokers).

The variables are available in the file tobacco.dat and listed in the following order.

- COTININE: salivary cotinine concentration (in ng/mL)
- AGE: age (in years)
- BMI: body mass index (in kg/m²)
- EDUC: years of education
- WET: takes value 1 if work conditions on day of measurement were wet and takes value 0 otherwise
- TASK: takes value 1 for priming, 2 for barning, 3 for topping, and 4 for other work not involving tobacco contact
- LNNSMOKE: natural logarithm of (1 + number of cigarettes smoked per day)

To report a test, provide H_0 , the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter.

- One-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task as the only predictor.
 - Report a test of whether all cell means are equal.

H_0 : All cell-means equal ($\mu_{\text{Priming}} = \mu_{\text{Barning}} = \mu_{\text{Topping}} = \mu_{\text{Other}}$)

H_1 : At least one cell-mean differs from another cell-mean.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Usual Overall Test	3	790.4453988	263.4817996	116.20	<.0001

$F(3, 690) = 116.20$ with $p < 0.0001$.

Reject H_0 .

There is evidence to suggest that at least two of the cell-means differ.

- If your overall test of the task effect was significant, examine all pairwise comparisons using the Scheffe correction. Summarize your findings in a table including columns for the estimated mean difference, degrees of freedom, F statistic, p -value, and a Scheffe confidence interval for the mean difference. Explain your findings in language the investigator can understand.

Table 3.1: Pairwise Comparisons

	Estimated Mean Difference	Degrees of Freedom	F	p-value	Scheffe 95% Confidence Interval
Priming v. Barning	0.9208	1, 690	19.58	0.0002	[0.3376, 1.5040]
Priming v. Topping	1.6738	1, 690	131.87	<0.0001	[1.2654, 2.0823]
Priming v. Other	2.6993	1, 690	332.68	<0.0001	[2.2845, 3.1140]
Barning v. Topping	0.7531	1, 690	12.99	0.0049	[0.1675, 1.3386]
Barning v. Other	1.7785	1, 690	71.38	<0.0001	[1.1885, 2.3684]
Topping v. Other	1.0254	1, 690	47.26	<0.0001	[0.6074, 1.4434]

The average values for the log of salivary cotinine concentration (in mg/mL) found in Latino tobacco workers in North Carolina differs significantly depending on the task workers performed.

- Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding, and give the interpretations of parameters in both coding schemes. In addition, provide the \mathbf{C} and $\boldsymbol{\theta}_0$ matrices used to test the hypothesis that average cotinine levels for workers involved in priming are greater than the average cotinine levels for all other workers.

Table 3.2: Parameter Estimates

Cell Mean Coding			
	Estimates	Standard Errors	
Priming	4.5086	0.1022	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.5086 mg/mL.
Barning	3.5878	0.1813	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.5878 mg/mL.
Topping	2.8347	0.1039	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.8347 mg/mL.
Other	1.8093	0.1070	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8093 mg/mL.
Reference Cell Coding – Solution 1			
	Estimates	Standard Errors	
Intercept	1.8093	0.1070	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8093 mg/mL.
Priming	2.6993	0.1480	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 2.6993 mg/mL higher than those whose task is not priming, barning, or topping.
Barning	1.7785	0.2105	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 1.7785 mg/mL higher than those whose task is not priming, barning, or topping.
Topping	1.0254	0.1492	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 1.0254 mg/mL higher than those whose task is not priming, barning, or topping.
Reference Cell Coding – Solution 2			
	Estimates	Standard Errors	
Intercept	4.5086	0.1022	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.5086 mg/mL.
Barning	-0.9208	0.2081	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 0.9208 mg/mL lower than those whose task is priming.
Topping	-1.6738	0.1458	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 1.6738 mg/mL lower than those whose task is priming.
Other	-2.6993	0.1480	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 2.6993 mg/mL lower than those whose task is priming.

$$H_0: \mu_{\text{priming}} = (\mu_{\text{barning}} + \mu_{\text{topping}} + \mu_{\text{other}}) / 3$$

Cell Mean Coding: $H_0: \beta_{\text{priming}} - \frac{\beta_{\text{barning}} + \beta_{\text{topping}} + \beta_{\text{other}}}{3} = 0$

$$C = [1 \quad -1/3 \quad -1/3 \quad -1/3] \quad \theta_0 = 0$$

Reference Cell Coding – Solution 1:

$$H_0: \beta_0 + \beta_{\text{priming}} = \frac{(\beta_0) + (\beta_0 + \beta_{\text{barning}}) + (\beta_0 + \beta_{\text{topping}})}{3} \equiv$$

$$H_0: \beta_0 + \beta_{\text{priming}} = \beta_0 + \frac{1}{3}(\beta_{\text{barning}} + \beta_{\text{topping}}) \equiv H_0: \beta_1 - \frac{1}{3}(\beta_2 + \beta_3) = 0$$

$$C = [0 \quad 1 \quad -1/3 \quad -1/3] \quad \theta_0 = 0$$

Reference Cell Coding – Solution 2:

$$H_0: \beta_0 = \frac{(\beta_0 + \beta_{\text{other}}) + (\beta_0 + \beta_{\text{barning}}) + (\beta_0 + \beta_{\text{topping}})}{3} \equiv$$

$$H_0: \beta_0 = \beta_0 + \frac{1}{3}(\beta_{\text{barning}} + \beta_{\text{topping}} + \beta_{\text{other}}) \equiv H_0: \frac{1}{3}(\beta_1 + \beta_2 + \beta_3) = 0$$

$$C = [0 \quad 1/3 \quad 1/3 \quad 1/3] \quad \theta_0 = 0$$

Note all these solutions assume equal sample sizes in the groups. You could factor the group sample sizes into these calculations.

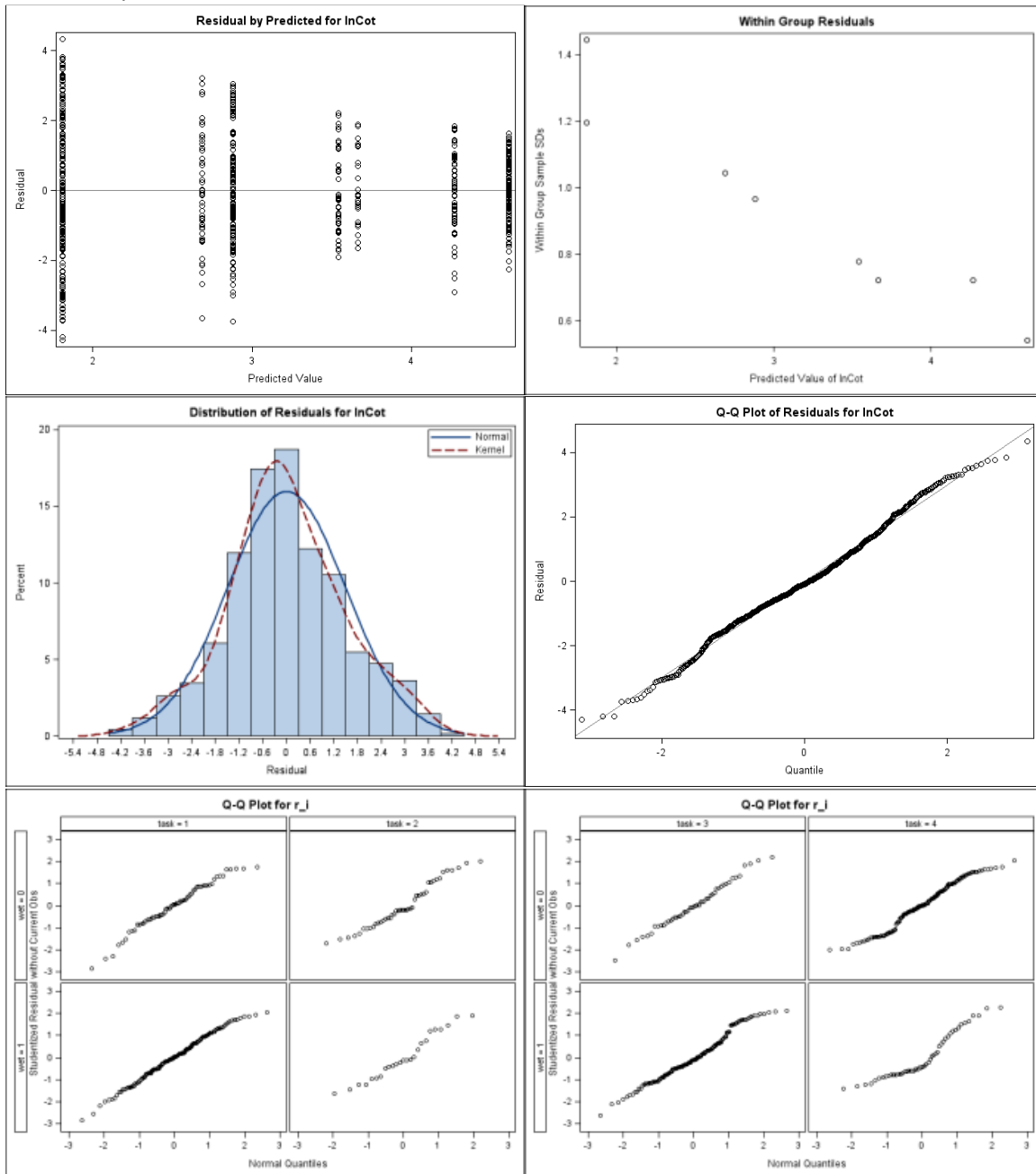
- Two-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task and wet as predictors.
 - Fit the two-way ANOVA model with full interaction, and interpret all parameter estimates in your model, clearly stating which coding scheme you used. Discuss the validity of the HILE Gauss assumptions for this model.

Table of task by wet				
	wet		Total	
	No	Yes		
task				
Priming	66	151	217	
Barning	44	25	69	
Topping	49	161	210	
Other	148	50	198	
Total	307	387	694	

Complete & Not Balanced.

Cell Mean Coding -			
$\ln(\text{cotinine}) = w0t1 \hat{\beta}_1 + w1t1 \hat{\beta}_2 + w0t2 \hat{\beta}_3 + w1t2 \hat{\beta}_4 + w0t3 \hat{\beta}_5 + w1t3 \hat{\beta}_6 + w0t4 \hat{\beta}_7 + w1t4 \hat{\beta}_8$			
	Estimates	Standard Errors	
Priming & Not Wet	4.2693	0.1855	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.2693 mg/mL in non-wet working conditions.
Priming & Wet	4.6131	0.1226	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.6131 mg/mL in wet working conditions.
Barning & Not Wet	3.5427	0.2272	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.5427 mg/mL in non-wet working conditions.
Barning & Wet	3.6670	0.3014	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.6670 mg/mL in wet working conditions.
Topping & Not Wet	2.6882	0.2153	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.6882 mg/mL in non-wet working conditions.
Topping & Wet	2.8793	0.1188	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.8793 mg/mL in wet working conditions.
Other & Not Wet	1.8089	0.1239	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8089 mg/mL in non-wet working conditions.
Other & Wet	1.8105	0.2131	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8105 mg/mL in wet working conditions.

Assumptions:



Homogeneity – Within cells isn't as important as between cells. The scatter plot of the standard deviations of the residuals for each group reveals a potential pattern to the data. This assumption could be violated and since we have inequality of sample sizes between groups this could impact the testing accuracy.

Independence – given through the sample design.

Linearity – This is okay given the design.

Existence – finite sample satisfies this.

Gaussian Errors – The overall histogram of residuals and overall and individual QQ plots appear to support this assumption.

- Based on this model, create a table of the estimated mean log cotinine levels, associated standard errors, and how each estimated mean is obtained from the model parameters (e.g., $\hat{\beta}_0 + \hat{\beta}_1$) for each task-wet combination.

Cell Mean Coding			
	Estimates	Standard Errors	$w0t1 \hat{\beta}_1 + w1t1 \hat{\beta}_2 + w0t2 \hat{\beta}_3 + w1t2 \hat{\beta}_4 + w0t3 \hat{\beta}_5 + w1t3 \hat{\beta}_6 + w0t4 \hat{\beta}_7 + w1t4 \hat{\beta}_8$
Grand Mean	3.1599	0.0699	$(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5 + \hat{\beta}_6 + \hat{\beta}_7 + \hat{\beta}_8)/8$
Priming	4.4412	0.1112	$(\hat{\beta}_1 + \hat{\beta}_2)/2$
Barning	3.6049	0.1887	$(\hat{\beta}_3 + \hat{\beta}_4)/2$
Topping	2.7837	0.1229	$(\hat{\beta}_5 + \hat{\beta}_6)/2$
Other	1.8097	0.1232	$(\hat{\beta}_7 + \hat{\beta}_8)/2$
Wet	3.2425	0.1017	$(\hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_6 + \hat{\beta}_8)/4$
Not Wet	3.0773	0.0961	$(\hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_5 + \hat{\beta}_7)/4$

Cell Mean Coding			
	Estimates	Standard Errors	
Priming & Not Wet	4.2693	0.1855	$\hat{\beta}_1$
Priming & Wet	4.6131	0.1226	$\hat{\beta}_2$
Barning & Not Wet	3.5427	0.2272	$\hat{\beta}_3$
Barning & Wet	3.6670	0.3014	$\hat{\beta}_4$
Topping & Not Wet	2.6882	0.2153	$\hat{\beta}_5$
Topping & Wet	2.8793	0.1188	$\hat{\beta}_6$
Other & Not Wet	1.8089	0.1239	$\hat{\beta}_7$
Other & Wet	1.8105	0.2131	$\hat{\beta}_8$

- The Full Model in Every Cell: For these questions, use the log of salivary cotinine as the response and task, and lnnsmoke as predictors.
 - Fit the full model in every cell. Provide and interpret estimates of all parameters for this model.

Cell Mean Coding -			
$\ln(\text{cotinine}) = \hat{\beta}_{0t1}t1 + \hat{\beta}_{0t2}t2 + \hat{\beta}_{0t3}t3 + \hat{\beta}_{0t4}t4 + \hat{\beta}_{1t1}(t1\ln\text{nsmoke}) + \hat{\beta}_{1t2}(t2\ln\text{nsmoke}) + \hat{\beta}_{1t3}(t3\ln\text{nsmoke}) + \hat{\beta}_{1t4}(t4\ln\text{nsmoke})$			
	Estimates	Standard Errors	
Priming	4.3344	0.0932	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.3344 mg/mL.
Barning	3.1134	0.1666	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.1134 mg/mL.
Topping	2.0123	0.0985	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.0123 mg/mL.
Other	0.9137	0.0957	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 0.9137 mg/mL.
Priming & LNNSMOKE	0.2946	0.0887	Latino tobacco workers in NC whose task is priming average log of salivary cotinine concentration increases by 0.2946 mg/mL for every 1 unit increase in lnnsmoke.
Barning & LNNSMOKE	0.7221	0.1450	Latino tobacco workers in NC whose task is barning average log of salivary cotinine concentration increases by 0.7221 mg/mL for every 1 unit increase in lnnsmoke.
Topping & LNNSMOKE	1.2305	0.0894	Latino tobacco workers in NC whose task is topping average log of salivary cotinine concentration increases by 1.2305 mg/mL for every 1 unit increase in lnnsmoke.
Other & LNNSMOKE	1.7789	0.1022	Latino tobacco workers in NC whose task is not priming, barning, or topping average log of salivary cotinine concentration increases by 1.7789 mg/mL for every 1 unit increase in lnnsmoke.

- Report an appropriate test of whether task is related to cotinine levels. If this test is significant, report step-down tests to determine exactly where differences lie. For all tests reported, be sure to state H_0 clearly and give explicit justification for which tests were used and why.

Use a test of coincidence; the hypothesis indicates that the slopes and intercepts are all equal regardless of task.

$$H_0: \beta_{0t1} = \beta_{0t2} = \beta_{0t3} = \beta_{0t4} \text{ and } \beta_{1t1} = \beta_{1t2} = \beta_{1t3} = \beta_{1t4}$$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
TEST OF COINCIDENCE	6	922.1700624	153.6950104	119.29	<.0001

$F(6, 687) = 119.29$ p-value < 0.001 Reject H_0
 Task is related either by slope or intercept to cotinine levels at the 0.01 level.

Step down to determine if the differences are in the slopes or intercepts.

$$H_0: \beta_{1t1} = \beta_{1t2} = \beta_{1t3} = \beta_{1t4}$$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
STEPDOWN: EQUAL SLOPES	3	169.6441128	56.5480376	43.89	<.0001

$F(3,690) = 43.89$ p-value < 0.0001 Reject H_0
 The slopes are significantly different from each other at the 0.01 level.

$$H_0: \beta_{0t1} = \beta_{0t2} = \beta_{0t3} = \beta_{0t4}$$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
STEPDOWN: EQUAL INTERCEPTS	3	902.0132306	300.6710769	233.36	<.0001

$F(3,690) = 233.36$ p-value < 0.0001 Reject H_0
 The intercepts are significantly different from each other at the 0.01 level.

Use pair-wise tests to determine exactly which intercepts/slopes are different.

$$\alpha = 0.05 / 6 = 0.0083 \quad \sim F(1, 692)$$

$$H_0: \beta_{0ti} = \beta_{0tj}$$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
PAIRWISE INTERCEPTS T1 V T2	1	52.7322641	52.7322641	40.93	<.0001
PAIRWISE INTERCEPTS T1 V T3	1	377.7620263	377.7620263	293.20	<.0001
PAIRWISE INTERCEPTS T1 V T4	1	845.0460533	845.0460533	655.87	<.0001
PAIRWISE INTERCEPTS T2 V T3	1	41.7145812	41.7145812	32.38	<.0001
PAIRWISE INTERCEPTS T2 V T4	1	168.9622476	168.9622476	131.14	<.0001
PAIRWISE INTERCEPTS T3 V T4	1	82.4489524	82.4489524	63.99	<.0001

Reject H_0 for all ij pairs
 All intercepts are significantly different from each other at the 0.01 level.

$$H_0: \beta_{1ti} = \beta_{1tj}$$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
PAIRWISE SLOPES T1 V T2	1	8.1529052	8.1529052	6.33	0.0121
PAIRWISE SLOPES T1 V T3	1	71.1683903	71.1683903	55.24	<.0001
PAIRWISE SLOPES T1 V T4	1	155.0229026	155.0229026	120.32	<.0001
PAIRWISE SLOPES T2 V T3	1	11.4843308	11.4843308	8.91	0.0029
PAIRWISE SLOPES T2 V T4	1	45.7480465	45.7480465	35.51	<.0001
PAIRWISE SLOPES T3 V T4	1	21.0166241	21.0166241	16.31	<.0001

Reject H_0 for all ij pairs except i=1 and j=2
 Except for task 1 and 2, the slopes are significantly different from each other at the 0.01 level.