**Question 1.** A study was performed to examine the effect of diet on depression in 600 graduate students at one university.

Graduate students were randomized to diet groups such that 300 graduate students were assigned a standard American diet and 300 graduate students were assigned a plant based diet.

All food and beverages were provided by the study center for two months and no students dropped out of the study.

Information was also collected on the students' self-reported exercise for an average week.

After two months, the graduate students took an exam in order to determine if they were clinically depressed or not.

The following table provides the variable coding.

| Variable Coding | |
|---|---|
| diet | 0 = standard American diet<br>1 = plant based diet |
| exercise | 0 = exercise less than an average of 3 hours a week<br>1 = exercise greater than or equal to an average of 3 hours a week |
| depressed | 0 = Not depressed<br>1 = Depressed |

The following SAS programs were run and partial output is included on the next few pages.

Note: SAS gave the following output for all models:

Convergence criterion (GCONV=1E-8) satisfied.

Please note that this is not a real study. This example is just being used to test and illustrate concepts from the class.

# Model 1.A

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = diet / COVB;
 FREQ n;
RUN;
```

| Model Fit Statistics | | |
|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 830.547 | 772.283 |
| SC | 834.944 | 781.077 |
| -2 Log L | 828.547 | 768.283 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 60.2640 | 1 | <.0001 |
| Score | 59.2252 | 1 | <.0001 |
| Wald | 57.1256 | 1 | <.0001 |

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 0.4895 | 0.1189 | 16.9390 | <.0001 |
| diet | 1 | -1.3053 | 0.1727 | 57.1256 | <.0001 |

# Model 1.B

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = exercise / COVB;
 FREQ n;
RUN;
```

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 830.547 | 798.156 |
| **SC** | 834.944 | 806.950 |
| **-2 Log L** | 828.547 | 794.156 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 34.3913 | 1 | <.0001 |
| **Score** | 34.0805 | 1 | <.0001 |
| **Wald** | 33.4018 | 1 | <.0001 |

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | 0.3483 | 0.1192 | 8.5341 | 0.0035 |
| **exercise** | 1 | -0.9744 | 0.1686 | 33.4018 | <.0001 |

# Model 1.C

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = diet exercise / COVB;
 FREQ n;
RUN;
```

| Model Fit Statistics | | |
|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 830.547 | 736.602 |
| SC | 834.944 | 749.793 |
| -2 Log L | 828.547 | 730.602 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 97.9446 | 2 | <.0001 |
| Score | 92.7105 | 2 | <.0001 |
| Wald | 81.9487 | 2 | <.0001 |

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 1.0821 | 0.1619 | 44.6895 | <.0001 |
| diet | 1 | -1.3913 | 0.1809 | 59.1761 | <.0001 |
| exercise | 1 | -1.0832 | 0.1807 | 35.9537 | <.0001 |

| Estimated Covariance Matrix | | | |
|---|---|---|---|
| Parameter | Intercept | diet | exercise |
| Intercept | 0.0262 | -0.01832 | -0.01901 |
| diet | -0.01832 | 0.032712 | 0.005487 |
| exercise | -0.01901 | 0.005487 | 0.032637 |

# Model 1.D

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = diet exercise diet* exercise / COVB;
 FREQ n;
RUN;
```

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| AIC | 830.547 | 730.313 |
| SC | 834.944 | 747.900 |
| -2 Log L | 828.547 | 722.313 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| Likelihood Ratio | 106.2344 | 3 | <.0001 |
| Score | 101.0480 | 3 | <.0001 |
| Wald | 86.6121 | 3 | <.0001 |

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| Intercept | 1 | 1.3949 | 0.2074 | 45.2163 | <.0001 |
| diet | 1 | -1.9354 | 0.2700 | 51.3933 | <.0001 |
| exercise | 1 | -1.6034 | 0.2632 | 37.1051 | <.0001 |
| diet*exercise | 1 | 1.0454 | 0.3651 | 8.1959 | 0.0042 |

| Estimated Covariance Matrix | | | | |
|---|---|---|---|---|
| **Parameter** | **Intercept** | **diet** | **exercise** | **dietexercise** |
| Intercept | 0.043029 | -0.04303 | -0.04303 | 0.043029 |
| diet | -0.04303 | 0.072886 | 0.043029 | -0.07289 |
| exercise | -0.04303 | 0.043029 | 0.069287 | -0.06929 |
| dietexercise | 0.043029 | -0.07289 | -0.06929 | 0.133332 |

**Question 2.** An investigator ran the following code for a small study and was very confused. The study had a binary outcome (disease=1 for the disease and 0 otherwise), a binary exposure variable (exposure=1 for the exposure and 0 otherwise), and one binary covariate (covariate = 0 or 1). The SAS code and partial output is given below.

# Model 2.

```
DATA test;
INPUT covariate exposure disease n;
DATALINES;
0 1 1 6
0 1 0 3
0 0 1 1
0 0 0 6
;
RUN;

PROC LOGISTIC;
 MODEL disease (EVENT = '1') = exposure covariate / COVB;
 FREQ n;
RUN;
```

| Model Convergence Status |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 4.7312 | 1 | 0.0296 |
| **Score** | 4.3900 | 1 | 0.0361 |
| **Wald** | 3.7049 | 1 | 0.0543 |

**Note:** The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

| | |
|---|---|
| **covariate =** | 0 |

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -1.7918 | 1.0801 | 2.7518 | 0.0971 |
| **exposure** | 1 | 2.4849 | 1.2910 | 3.7049 | 0.0543 |
| **covariate** | 0 | 0 | . | . | . |

| Odds Ratio Estimates | | | |
|---|---|---|---|
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **exposure** | 12.000 | 0.956 | 150.688 |

**Question 1.1 [5 points]** For **Model 1.A**, is there a significant association between the odds of depression and diet using a Wald test? (Provide the odds ratio and corresponding 95% Wald CI.)

Yes, there is a significant association between depression and diet using a Wald test (p-value<0.0001). The OR=0.271 and the 95% Wald CI= (0.193,0.380).

OR: exp(-1.3053)= 0.2710912

95% Wald CI= (exp(-1.3053-1.96*0.1727), exp(-1.3053+1.96*0.1727))= (0.1932459, 0.3802950)

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 0.4895 | 0.1189 | 16.9390 | <.0001 |
| diet | 1 | -1.3053 | 0.1727 | 57.1256 | <.0001 |

| Odds Ratio Estimates | | | |
|---|---|---|---|
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| diet | 0.271 | 0.193 | 0.380 |

**Question 1.2 [5 points]** For **Model 1.A**, is there a significant association between the odds of depression and diet using a Likelihood Ratio test? (Provide the test statistic and corresponding p-value.)

Yes, there is a significant association between the odds of depression and diet using a Likelihood Ratio test (p-value<0.0001,Chi-square test statistic=60.2640).

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 60.2640 | 1 | <.0001 |
| **Score** | 59.2252 | 1 | <.0001 |
| **Wald** | 57.1256 | 1 | <.0001 |

**Question 1.3 [5 points]** For **Model 1.A**, despite a fairly large sample size of n=600 and a non-rare outcome in the sample with no sparsity in the cells of the contingency table, the chi-square test statistic for the association between the odds of depression and diet is smallest and the corresponding p-value is largest for the Wald test compared to the Score test and Likelihood Ratio test. Why is this the case in general?

In general, the Wald test has lower power compared to the Score and LRT.

**Question 1.4 [10 points]** For **Model 1D**, is there a significant association between the odds of depression and average weekly exercise among graduate students assigned to the standard American diet (diet=0)? (Provide an OR and 95% Wald CI.)

Yes, there is a significant association between exercise and the odds of depression among graduate students assigned to the standard American diet. (OR=0.2 and 95% Wald CI= (0.12,0.33))

Let

logit(P(depression$_i$=1))=$\beta_0 + \beta_{diet}\text{diet}_i + \beta_{exercise}\text{exercise}_i + \beta_{interaction}\text{diet}_i * \text{exercise}_i$

$$\widehat{OR} = \frac{exp(\hat{\beta}_0 + \hat{\beta}_{exercise})}{exp(\hat{\beta}_0)} = exp(\hat{\beta}_{exercise}) = \exp(-1.6034) = 0.2012112$$

95% Wald CI: (exp(-1.6034-1.96*0.2632), exp(-1.6034+1.96*0.2632))= (0.1201190,0.3370486)

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 1.3949 | 0.2074 | 45.2163 | <.0001 |
| diet | 1 | -1.9354 | 0.2700 | 51.3933 | <.0001 |
| exercise | 1 | -1.6034 | 0.2632 | 37.1051 | <.0001 |
| diet*exercise | 1 | 1.0454 | 0.3651 | 8.1959 | 0.0042 |

**Question 1.5 [10 points]** For **Model 1D**, is there a significant association between the odds of depression and average weekly exercise among graduate students assigned to the plant based diet (diet=1)? (Provide an OR and 95% Wald CI)

Yes, there is a significant association between exercise and the odds of depression among graduate students assigned to the plant based diet. (OR=0.57 & 95% Wald CI=(0.35,0.94))

Let

logit(P(depression$_i$=1))=$\beta_0 + \beta_{diet}\text{diet}_i + \beta_{exercise}\text{exercise}_i + \beta_{interaction}\text{diet}_i * \text{exercise}_i$

$$\widehat{OR} = \frac{exp\left(\hat{\beta}_0 + \hat{\beta}_{diet} + \hat{\beta}_{exercise} + \hat{\beta}_{interaction}\right)}{exp\left(\hat{\beta}_0 + \hat{\beta}_{diet}\right)} = exp\left(\hat{\beta}_{exercise} + \hat{\beta}_{interaction}\right)$$
$$= \exp(-1.6034 + 1.0454) = 0.5723526$$

$$\sqrt{Var\left(\hat{\beta}_{exercise} + \hat{\beta}_{interaction}\right)} = \sqrt{Var\left(\hat{\beta}_{exercise}\right) + Var\left(\hat{\beta}_{interaction}\right) + 2 * Cov\left(\hat{\beta}_{exercise}, \hat{\beta}_{interaction}\right)}$$
$$= \sqrt{0.069287 + 0.133332 + 2 * -0.06929} = \sqrt{0.064039} = 0.2530593$$

95% Wald CI: (exp((-1.6034+1.0454)-1.96*0.2530593),exp((-1.6034+1.0454)+1.96*0.2530593))
=(0.3485421,0.9398793)

**Analysis of Maximum Likelihood Estimates**

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | 1.3949 | 0.2074 | 45.2163 | <.0001 |
| diet | 1 | -1.9354 | 0.2700 | 51.3933 | <.0001 |
| exercise | 1 | -1.6034 | 0.2632 | 37.1051 | <.0001 |
| diet*exercise | 1 | 1.0454 | 0.3651 | 8.1959 | 0.0042 |

**Estimated Covariance Matrix**

| Parameter | Intercept | diet | exercise | dietexercise |
|---|---|---|---|---|
| Intercept | 0.043029 | -0.04303 | -0.04303 | 0.043029 |
| diet | -0.04303 | 0.072886 | 0.043029 | -0.07289 |
| exercise | -0.04303 | 0.043029 | 0.069287 | -0.06929 |
| dietexercise | 0.043029 | -0.07289 | -0.06929 | 0.133332 |

**Question 1.6 [10 points]** In **Model 1.D**, using a Likelihood Ratio Test, is the interaction between diet and exercise significantly associated with the odds of depression? (Provide a Likelihood Ratio Test Statistics to support your answer.)

Yes, the interaction between diet and exercise is significantly associated with the odds of depression (LRT statistic=16.578, p-value<0.05).

LRT: -2*(722.313-730.602) = 16.578 > $\chi^2_{1,\,0.95} = 3.841$ then p-value<0.05

Model 1C

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 830.547 | 736.602 |
| **SC** | 834.944 | 749.793 |
| **-2 Log L** | 828.547 | 730.602 |

Model 1D

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 830.547 | 730.313 |
| **SC** | 834.944 | 747.900 |
| **-2 Log L** | 828.547 | 722.313 |

**Question 1.7 [5 points]** For **Models 1.A, 1.B, 1.C**, and **1.D**, which is the best model based on AIC?

Model 1.D, the full model with the interaction, is the best model based on AIC since this model has the lowest AIC by at least 2 points.

**Question 1.8 [5 points]** For **Models 1.A, 1.B, 1.C**, and **1.D**, which is the best model based on BIC?

Model 1.C, the model with diet and exercise, is the best model based on BIC since this model has a BIC within 2 points of the full model (i.e Model 1D).

749.793-747.900= 1.893

**Question 1.9 [5 points]** In general, why do the models selected by AIC and BIC differ in Question 1.7 and 1.8?

Because BIC penalizes the additional covariates more extremely than AIC.

Model 1A

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 830.547 | 772.283 |
| **SC** | 834.944 | 781.077 |
| **-2 Log L** | 828.547 | 768.283 |

Model 1B

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 830.547 | 798.156 |
| **SC** | 834.944 | 806.950 |
| **-2 Log L** | 828.547 | 794.156 |

Model 1C

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 830.547 | 736.602 |
| **SC** | 834.944 | 749.793 |
| **-2 Log L** | 828.547 | 730.602 |

Model 1D

| Model Fit Statistics | | |
|---|---|---|
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 830.547 | 730.313 |
| **SC** | 834.944 | 747.900 |
| **-2 Log L** | 828.547 | 722.313 |

**Question 1.10 [5 points]** Which model (1.A, 1.B, 1.C, & 1.D) is the saturated model? Or is the saturated model not given?

Model 1.D is the saturated model.

**Question 1.11 [5 points]** Use the deviance to compare model 1.D and 1.C. If this question cannot be answered with the output provided, please state what output you would need to answer this question.

This question cannot be answered with the output provided. I would need -2 log likelihood from the SAS or R output for the data entered in group form.

**Question 1.12 [10 points]** In the study for question 1 (Models 1.A, 1.B, 1.C, 1.D), are exactly 300 graduate students (i.e. 50% of the 600 graduate students) depressed after the 2 month diet? Justify your answer for full credit.

No.

If 300 of the graduate students are depressed. The -2 log L for the null model is as follows.

-2*logL= -2*(300*log(300/600)+300*log(300/600))= 831.7766

But based on the SAS output, -2 log L for the null=828.547 which doesn't match -2 Log L from above.

**For Model 1A**

| Model Fit Statistics | | |
|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 830.547 | 772.283 |
| SC | 834.944 | 781.077 |
| -2 Log L | 828.547 | 768.283 |

| Response Profile | | |
|---|---|---|
| Ordered Value | depressed | Total Frequency |
| 1 | 0 | 322 |
| 2 | 1 | 278 |

**Question 2.1 [10 points]** For Question 2,

$$logit(\Pr(\text{disease}_i = 1)) = \beta_0 + \beta_E exposure_i + \beta_C covariate_i$$

In matrix form,

$$logit(\Pr(\mathbf{Y} = \mathbf{1})) = \boldsymbol{X\beta}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \; and \; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_E \\ \beta_C \end{bmatrix}$$

Give the dimensions of the matrix **X** and write $\boldsymbol{X}$ in matrix form with ALL the values inputted from Model 2 (i.e. elements of **X** should be 0s and 1s. Not $exposure_i$ or $covariate_i$). Use **Y** given above as a guide. Do NOT use dots (i.e. …) or arrows (i.e. ->). Give ALL of the elements of **X**.

X has 16 rows and 3 columns

Then there are 9 rows with the intercept=1, covariate =0 and exposure =1. So there are 9 rows of (110) in **X**. There are 7 rows with the intercept=1, covariate =0 and exposure=0. So there are 7 rows of (100) in **X**.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

```
DATA test;
INPUT covariate exposure disease n;
DATALINES;
0 1 1 6
0 1 0 3
0 0 1 1
0 0 0 6
;
RUN;
```

**Question 2.2 [5 points]** Explain why the following note was given by SAS for the **Model 2** output.

**Note:** The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

| | |
|---|---|
| **covariate =** | 0 |

The covariate is 0 for all subjects which is a linear combination of the intercept column in X (i.e. the covariate column is equivalent to the intercept column of X minus itself).

**Question 2.3 [5 points]** The OR for the exposure is very large (i.e. OR=12) and the 95% Wald CI is very wide (0.956,150.688), but the corresponding p-value is relatively modest (i.e p-value=0.0543). Explain why the Wald test is not performing well in this scenario.

The sample size is relatively small (n=16) and the Wald test does not perform well for small sample sizes.

| Odds Ratio Estimates | | | |
|---|---|---|---|
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| exposure | 12.000 | 0.956 | 150.688 |

**Study:** A study was performed to examine whether dietary fiber intake has an effect on HbA1c levels. The HbA1c test (hemoglobin A1c test) is a laboratory test used to estimate average blood glucose levels. Normal HbA1c levels are 4%-6%, but are commonly higher in cigarette smokers. Dietary fiber intake was measured as a continuous variable (grams/day) and vitamin C usage was measured as a categorical variable from a food frequency questionnaire. The study sample consisted of 125 smokers and 75 non-smokers, for a total of 200 participants.

The following variables are available for the analysis:

| | |
|---|---|
| hba1c: | hemoglobin A1c levels (%) |
| fiber : | dietary fiber intake (grams/day) |
| smoker: | current smoking status (0 = non-smokers; 1=smokers) |
| vitC: | supplement of vitamin C (2= large dose, 1= normal dose, 0=no dose) |

| *Vitamin C Usage* | *New indicator variables* | | |
|---|---|---|---|
| | vitC_none | vitC_normal | vitC_large |
| vitC_none | 1 | 0 | 0 |
| vitC_normal | 0 | 1 | 0 |
| vitC_large | 0 | 0 | 1 |

## Model 1:

You perform a simple linear regression of HbA1c (*hba1c*) on dietary fiber intake (*fiber*). The following SAS output was obtained.

```
PROC REG;
 MODEL hba1c = fiber;
RUN;
```

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | XXXXX | XXXXX | 2.60748 | XXXXX | 0.2135 |
| Error | XXXXX | XXXXX | XXXXX | | |
| Corrected Total | XXXXX | XXXXX | | | |

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 6.62683 | 0.23910 | 27.72 | <.0001 |
| fiber | 1 | -0.01855 | 0.01487 | -1.25 | 0.2135 |

# Model 2:

You perform a t-test and a simple linear regression of HbA1c (*hba1c*) on smoking status (smoker). The following SAS output was obtained and then sections were blanked out.

```
proc ttest;
 var hba1c;
 class smoker;
 run;
```

| smoker | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 0 | 125 | 5.5324 | 0.4719 | 0.0422 | 4.5300 | 7.5500 |
| 1 | 75 | 7.7157 | 1.0592 | 0.1223 | 5.5000 | 10.6600 |
| Diff (1-2) | | -2.1833 | 0.7475 | 0.1092 | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | XXXXX | XXXXX | <.0001 |
| Satterthwaite | Unequal | 91.9 | -16.87 | <.0001 |

```
PROC REG;
 MODEL hba1c = smoker / covb;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | Q2A | 223.45052 | XXXXX | Q2E | <.0001 |
| Error | Q2B | Q2D | 0.55881 | | |
| Corrected Total | Q2C | 334.09564 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | Q2F | 0.06686 | 82.74 | <.0001 |
| smoker | 1 | Q2G | XXXXX | Q2H | Q2I |

# Model 3:

You perform a linear regression of HbA1c (*hba1c*) on fiber, smoking status (smoker), and fiber*smoker (fiber_smoke). The following SAS output was obtained.

```
PROC REG;
MODEL hba1c = fiber smoker fiber_smoke / covb;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 3 | 229.16117 | 76.38706 | 142.68 | <.0001 |
| Error | 196 | 104.93447 | 0.53538 | | |
| Corrected Total | 199 | 334.09564 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.73170 | R-Square | 0.6859 |
| Dependent Mean | 6.35115 | Adj R-Sq | 0.6811 |
| Coeff Var | 11.52070 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 5.84243 | 0.16545 | 35.31 | <.0001 |
| fiber | 1 | -0.02118 | 0.01038 | -2.04 | 0.0427 |
| smoker | 1 | 2.43152 | 0.28710 | 8.47 | <.0001 |
| fiber_smoke | 1 | -0.01549 | 0.01773 | -0.87 | 0.3834 |

| Covariance of Estimates | | | | |
|---|---|---|---|---|
| Variable | Intercept | fiber | smoker | fiber_smoke |
| Intercept | 0.0273743493 | -0.001577284 | -0.027374349 | 0.0015772839 |
| fiber | -0.001577284 | 0.0001077386 | 0.0015772839 | -0.000107739 |
| smoker | -0.027374349 | 0.0015772839 | 0.0824244199 | -0.004724646 |
| fiber_smoke | 0.0015772839 | -0.000107739 | -0.004724646 | 0.0003144918 |

**Model 4:** You perform a linear regression of HbA1c (*hba1c*) on vitamin C usage where vitC_normal=1 for normal dosages of vitamin C & 0 otherwise, vitC_large=1 for large dosages of vitamin C & 0 otherwise, and vitC_none= 1 for no vitamin C dosage and 0 otherwise.

## Model 4a:

```
PROC REG;
MODEL hba1c = vitC_none vitC_normal vitC_large  / noint covb;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | Q4A | XXXXX | XXXXX | XXXXX | <.0001 |
| **Error** | Q4B | XXXXX | Q4D | | |
| **Uncorrected Total** | Q4C | XXXXX | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **vitC_none** | 1 | 6.52479 | 0.11740 | 55.58 | <.0001 |
| **vitC_normal** | 1 | 6.26842 | 0.20775 | 30.17 | <.0001 |
| **vitC_large** | 1 | 5.94372 | 0.19530 | 30.43 | <.0001 |

| Covariance of Estimates | | | |
|---|---|---|---|
| **Variable** | **vitC_none** | **vitC_normal** | **vitC_large** |
| **vitC_none** | 0.0137827786 | 0 | 0 |
| **vitC_normal** | 0 | 0.0431618594 | 0 |
| **vitC_large** | 0 | 0 | 0.0381430386 |

## Model 4b:

```
PROC REG;
MODEL hba1c =vitC_normal vitC_large;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 10.98596 | 5.49298 | 3.35 | 0.0371 |
| **Error** | 197 | 323.10968 | 1.64015 | | |
| **Corrected Total** | 199 | 334.09564 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 1 | Q4E | Q4G | XXXXX | <.0001 |
| **vitC_normal** | 1 | Q4F | Q4H | XXXXX | 0.2840 |
| **vitC_large** | 1 | -0.58107 | 0.22787 | -2.55 | 0.0115 |

# Model 5:

You perform a simple linear regression of HbA1c (*hba1c*) on dietary fiber intake (*fiber*) and smoking status (*smoker*). The following SAS output was obtained.

```
PROC REG;
 MODEL hba1c = fiber smoker;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 228.75255 | 114.37628 | 213.89 | <.0001 |
| **Error** | 197 | 105.34308 | 0.53474 | | |
| **Corrected Total** | 199 | 334.09564 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 1 | 5.92014 | 0.13943 | 42.46 | <.0001 |
| **fiber** | 1 | **XXXXX** | **XXXXX** | **Q5A** | 0.0019 |
| **smoker** | 1 | 2.19877 | 0.10692 | 20.56 | <.0001 |

***Question 1***. *10 points*. For **Model 1**, provide a brief interpretation of the association between dietary fiber intake and HbA1c levels, including the **95% CI**, point estimate, p-value, and decision.

**95% CI: -0.01855±1.972*(0.01487) = (-0.04787364, 0.01077364)**

**There is not a significant association between fiber intake and HbA1c levels in this study (p = 0.2135). For every 1 gram/day increase in dietary fiber intake, HbA1c levels decrease by 0.01855 percentage points (95% CI: -0.048 to 0.011).**

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 6.62683 | 0.23910 | 27.72 | <.0001 |
| fiber | 1 | -0.01855 | 0.01487 | -1.25 | 0.2135 |

```
PROC GLM;
 MODEL hba1c = fiber / clparm;
RUN;
```

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|
| Intercept | 6.626829255 | 0.23909756 | 27.72 | <.0001 | 6.155324684 | 7.098333825 |
| fiber | -0.01855 | 0.01487 | -1.25 | 0.2135 | -0.04787364 | 0.01077364 |

***Question 2A***. *10 points*. Fill in the missing values for **Model 2** (parts **Q2A-Q2E**). **Justify your answers for full credit.**

***Q2A.***
**DF model =p =1**

***Q2B.***
**DF error =n-p-1 =200-1-1=198**

***Q2C.***
**DF total = DF model+ DF error= n-1 =200-1= 199**

***Q2D.***
**SS error= SS total –SS model = 334.09564- 223.45052=110.6451**

***Q2E.***
**MS model = SS model / DF model =223.45052/1 =223.45052**
**F-stat= MS model/ MS error= 223.45052/ 0.55881=399.8685**

```
proc ttest;
 var hba1c;
 class smoker;
 run;
```

| smoker | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 0 | 125 | 5.5324 | 0.4719 | 0.0422 | 4.5300 | 7.5500 |
| 1 | 75 | 7.7157 | 1.0592 | 0.1223 | 5.5000 | 10.6600 |
| Diff (1-2) | | -2.1833 | 0.7475 | 0.1092 | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | **XXXXX** | **XXXXX** | <.0001 |
| Satterthwaite | Unequal | 91.9 | -16.87 | <.0001 |

```
PROC REG;
 MODEL hba1c = smoker / covb;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | **1 Q2A** | 223.45052 | 223.45052 **XXXXX** | **399.87 Q2E** | <.0001 |
| Error | **198 Q2B** | **110.64511 Q2D** | 0.55881 | | |
| Corrected Total | **199 Q2C** | 334.09564 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | **5.5324 Q2F** | 0.06686 | 82.74 | <.0001 |
| smoker | 1 | **2.1833 Q2G** | 0.10918 **XXXXX** | **20.00 Q2H** | **Q2I** |

***Question 2B**. 10 points.* Fill in the missing values for **Model 2** (parts **Q2F-Q2J**). **Justify your answers for full credit.**

***Q2F.***

**Beta0 =mean of HbAc1 for smoker 0= 5.5324**

***Q2G.***

**Beta0+ Beta1=Mean of hbac1 for smoker 1**
**Beta1= Mean of hbac1 for smoker 1- Beta0**
**Beta1= Mean of hbac1 for smoker 1- Mean of hbac1 for smoker 0**
**Beta1=7.7157-5.5324=2.1833**

***Q2H.***

**t-stat= sqrt(F-stat from ANOVA table Q2E)= sqrt(**399.87**) = 19.99675 approx. = 20.00**

***Q2I.***

**p-value for overall F-stat from ANOVA table: p-value** <.0001

```
proc ttest;
 var hba1c;
 class smoker;
 run;
```

| smoker | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 0 | 125 | 5.5324 | 0.4719 | 0.0422 | 4.5300 | 7.5500 |
| 1 | 75 | 7.7157 | 1.0592 | 0.1223 | 5.5000 | 10.6600 |
| Diff (1-2) | | -2.1833 | 0.7475 | 0.1092 | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | **XXXXX** | **XXXXX** | <.0001 |
| Satterthwaite | Unequal | 91.9 | -16.87 | <.0001 |

```
PROC REG;
 MODEL hba1c = smoker / covb;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 Q2A | 223.45052 | 223.45052 **XXXXX** | 399.87 Q2E | <.0001 |
| Error | 198 Q2B | 110.64511 Q2D | 0.55881 | | |
| Corrected Total | 199 Q2C | 334.09564 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | **5.5324** Q2F | 0.06686 | 82.74 | <.0001 |
| smoker | 1 | **2.1833** Q2G | 0.10918 **XXXXX** | **20.00** Q2H | Q2I |

***Question 3A**. 10 points*. Provide an interpretation of the relationship between HbA1c and fiber for non- smokers in **Model 3** (include a point estimate, test statistic and decision).

**E[HbA1c]=Beta0 + Beta_fiber * fiber+ Beta_smoker * smoker +Beta_int fiber * smoker**

**For smoker=0**
**E[HbA1c]=Beta0 + Beta_fiber * fiber**

**Point estimate: Beta_fiber= -0.02118**
**Test stat: -2.04**

**There is a significant relationship between HbA1c and fiber for non-smokers (p-value=0.0427).**
**On average, HbA1c decreases by 0.02118 units for non-smokers.**

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 5.84243 | 0.16545 | 35.31 | <.0001 |
| fiber | 1 | -0.02118 | 0.01038 | -2.04 | 0.0427 |
| smoker | 1 | 2.43152 | 0.28710 | 8.47 | <.0001 |
| fiber_smoke | 1 | -0.01549 | 0.01773 | -0.87 | 0.3834 |

| Covariance of Estimates | | | | |
|---|---|---|---|---|
| Variable | Intercept | fiber | smoker | fiber_smoke |
| Intercept | 0.0273743493 | -0.001577284 | -0.027374349 | 0.0015772839 |
| fiber | -0.001577284 | 0.0001077386 | 0.0015772839 | -0.000107739 |
| smoker | -0.027374349 | 0.0015772839 | 0.0824244199 | -0.004724646 |
| fiber_smoke | 0.0015772839 | -0.000107739 | -0.004724646 | 0.0003144918 |

***Question 3B***. *10 points*. Provide an interpretation of the relationship between HbA1c and fiber for smokers in **Model 3** (include a point estimate, test statistic and decision).

**E[HbA1c]=Beta0 + Beta_fiber * fiber+ Beta_smoker * smoker +Beta_int fiber * smoker**

**For smoker=1**
**E[HbA1c]=[Beta0 + Beta_smoker] + [Beta_fiber+ Beta_int] * fiber**

**Point estimate: Beta_fiber+ Beta_int = -0.02118+-0.01549= -0.03667**

**Var(Beta_fiber+ Beta_int)= Var(Beta_fiber)+ Var(Beta_int)+ 2 cov(Beta_fiber, Beta_int)**
**=0.0001077386+0.0003144918+2*-0.000107739**
**=0.0002067524**

**Test stat: -0.03667/sqrt(0.0002067524) = -2.550267 > $t_{196,0.975}$ = 1.9723 So p-value<0.05**

**There is a significant relationship between HbA1c and fiber for smokers (p-value<0.05).**
**On average, HbA1c decreases by 0.03667 units for smokers.**

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 5.84243 | 0.16545 | 35.31 | <.0001 |
| fiber | 1 | -0.02118 | 0.01038 | -2.04 | 0.0427 |
| smoker | 1 | 2.43152 | 0.28710 | 8.47 | <.0001 |
| fiber_smoke | 1 | -0.01549 | 0.01773 | -0.87 | 0.3834 |

| Covariance of Estimates | | | | |
|---|---|---|---|---|
| Variable | Intercept | fiber | smoker | fiber_smoke |
| Intercept | 0.0273743493 | -0.001577284 | -0.027374349 | 0.0015772839 |
| fiber | -0.001577284 | 0.0001077386 | 0.0015772839 | -0.000107739 |
| smoker | -0.027374349 | 0.0015772839 | 0.0824244199 | -0.004724646 |
| fiber_smoke | 0.0015772839 | -0.000107739 | -0.004724646 | 0.0003144918 |

```
PROC GLM;
 MODEL hba1c = fiber  smoker fiber_smoke;
 ESTIMATE 'no smoker' fiber 1;
 ESTIMATE 'smoker' fiber 1 fiber_smoke 1;
RUN;
```

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| no smoker | -0.02117731 | 0.01037972 | -2.04 | 0.0427 |
| smoker | -0.03667010 | 0.01437892 | -2.55 | 0.0115 |

***Question 3C**. 10 points*. Does the relationship between HbA1c and fiber significantly depend on smoking status? Give a p-value to support this decision.

**The relationship between HbA1c and fiber does not significantly depend on smoking status (p-value=0.3834).**

**Model 4**

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 1 | 5.84243 | 0.16545 | 35.31 | <.0001 |
| **fiber** | 1 | -0.02118 | 0.01038 | -2.04 | 0.0427 |
| **smoker** | 1 | 2.43152 | 0.28710 | 8.47 | <.0001 |
| **fiber_smoke** | 1 | -0.01549 | 0.01773 | -0.87 | 0.3834 |

***Question 4A**. 10 points*. Fill in the missing values for **Model 4a** (parts **Q4A-Q4D**). **Justify your answers for full credit.**

***Q4A.***

***DF model= p= 3 b/c no intercept***

***Q4B.***

***DF error = n-p-0 =200-3 =197***

***Q4C.***

***DF total= DF model+ DF error= 3+197 =200***

***Q4D.***

***MS error from cell means model= MS error from reference cell model= 1.64015***

## Model 4a:

```
MODEL hba1c = vitC_none vitC_normal vitC_large  / noint covb;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | Q4A 3 | XXXXX 8078.40722 | XXXXX 2692.80241 | XXXXX 1641.80 | <.0001 |
| Error | Q4B 197 | XXXXX 323.10968 | Q4D 1.64015 | | |
| Uncorrected Total | Q4C 200 | XXXXX 8401.51690 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| vitC_none | 1 | 6.52479 | 0.11740 | 55.58 | <.0001 |
| vitC_normal | 1 | 6.26842 | 0.20775 | 30.17 | <.0001 |
| vitC_large | 1 | 5.94372 | 0.19530 | 30.43 | <.0001 |

| Covariance of Estimates | | | |
|---|---|---|---|
| Variable | vitC_none | vitC_normal | vitC_large |
| vitC_none | 0.0137827786 | 0 | 0 |
| vitC_normal | 0 | 0.0431618594 | 0 |
| vitC_large | 0 | 0 | 0.0381430386 |

## Model 4b:

```
MODEL hba1c =vitC_normal vitC_large;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 10.98596 | 5.49298 | 3.35 | 0.0371 |
| Error | 197 | 323.10968 | 1.64015 | | |
| Corrected Total | 199 | 334.09564 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.28068 | R-Square | 0.0329 |
| Dependent Mean | 6.35115 | Adj R-Sq | 0.0231 |
| Coeff Var | 20.16459 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | Q4E 6.52479 | Q4G 0.11740 | XXXXX 55.58 | <.0001 |
| vitC_normal | 1 | Q4F -0.25637 | Q4H 0.23863 | XXXXX -1.07 | 0.2840 |
| vitC_large | 1 | -0.58107 | 0.22787 | -2.55 | 0.0115 |

***Question 4B**. 10 points*. Fill in the missing values for **Model 4** (parts **Q4E-Q4H**). **Justify your answers for full credit.**

**Cell means model:**

**E[hbac1]=Beta_none * VitC_none+ Beta_normal * VitC_normal+ Beta_large * VitC_large**

**Reference model:**

**E[hbac1]=Gamma_0 + Gamma _normal * VitC_normal+ Gamma _large * VitC_large**

***Q4E.***

**Gamma0= Beta_none=6.52479**

***Q4F.***

**Gamma _normal= Beta_normal- Beta_none= 6.26842-6.52479= -0.25637**

***Q4G.***

**SE(Gamma0)= SE(Beta_none)= 0.11740**

***Q4H.***

**SE(Gamma _normal)= sqrt( var[Beta_normal- Beta_none] )**
**= sqrt( var[Beta_normal]+var[ Beta_none]-2cov[Beta_normal, Beta_none] )**
**=sqrt(0.0137827786+0.0431618594-2*0) = 0.2386308**

# Model 4a:

```
MODEL hba1c = vitC_none vitC_normal vitC_large  / noint covb;
```

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| vitC_none | 1 | 6.52479 | 0.11740 | 55.58 | <.0001 |
| vitC_normal | 1 | 6.26842 | 0.20775 | 30.17 | <.0001 |
| vitC_large | 1 | 5.94372 | 0.19530 | 30.43 | <.0001 |

| Covariance of Estimates | | | |
|---|---|---|---|
| Variable | vitC_none | vitC_normal | vitC_large |
| vitC_none | 0.0137827786 | 0 | 0 |
| vitC_normal | 0 | 0.0431618594 | 0 |
| vitC_large | 0 | 0 | 0.0381430386 |

# Model 4b:

```
MODEL hba1c =vitC_normal vitC_large;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 10.98596 | 5.49298 | 3.35 | 0.0371 |
| Error | 197 | 323.10968 | 1.64015 | | |
| Corrected Total | 199 | 334.09564 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.28068 | R-Square | 0.0329 |
| Dependent Mean | 6.35115 | Adj R-Sq | 0.0231 |
| Coeff Var | 20.16459 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | **Q4E** 6.52479 | **Q4G** 0.11740 | **XXXXX** 55.58 | <.0001 |
| vitC_normal | 1 | **Q4F** -0.25637 | **Q4H** 0.23863 | **XXXXX** -1.07 | 0.2840 |
| vitC_large | 1 | -0.58107 | 0.22787 | -2.55 | 0.0115 |

***Question 4C.*** *10 points.* Using **Model 4a**, test whether HbA1c is the same for those taking normal doses of vitamin C (vitC_normal) versus those taking large doses of vitamin C (vitC_large). **Provide only the null hypothesis and test statistic.**

**Ho: Beta_normal – Beta_large =0**

$$t = \frac{\beta_{normal} - \beta_{large}}{SE\left(\beta_{normal} - \beta_{large}\right)} = \frac{6.26842 - 5.94372}{\sqrt{0.0431618594 + 0.0381430386}} = \frac{0.3247}{0.2851401} = 1.138738$$

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| normal-large | 0.3247 | 0.2851401 | 1.14 | 0.2562 |

# Model 4a:

```
PROC REG;
MODEL hba1c = vitC_none vitC_normal vitC_large  / noint covb;
RUN;
```

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| vitC_none | 1 | 6.52479 | 0.11740 | 55.58 | <.0001 |
| vitC_normal | 1 | 6.26842 | 0.20775 | 30.17 | <.0001 |
| vitC_large | 1 | 5.94372 | 0.19530 | 30.43 | <.0001 |

| Covariance of Estimates | | | |
|---|---|---|---|
| Variable | vitC_none | vitC_normal | vitC_large |
| vitC_none | 0.0137827786 | 0 | 0 |
| vitC_normal | 0 | 0.0431618594 | 0 |
| vitC_large | 0 | 0 | 0.0381430386 |

***Question 5****. 10 points*. Give the absolute value of the t statistic for fiber in **Model 5** (part **Q5A**). **Show your work for full credit.** Hint: this question requires a partial F-test.

**t-stat for H0: Beta_fiber=0 in Model 5 equals the square root of the partial F-test between model 4 with smoker and fiber and model 2 with just smoker.**

**Partial F-test: [ (SS model (full)- SS model (reduced) )/k] / MSE(full)**
**=((228.75255-223.45052)/1)/ 0.53474=9.915155**

**Absolute value of t-stat= sqrt(9.915155)= 3.148834**

## Model 5:

```
PROC REG;
 MODEL hba1c = fiber smoker;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 228.75255 | 114.37628 | 213.89 | <.0001 |
| **Error** | 197 | 105.34308 | 0.53474 | | |
| **Corrected Total** | 199 | 334.09564 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 1 | 5.92014 | 0.13943 | 42.46 | <.0001 |
| **fiber** | 1 | **XXXXX** -0.02648 | **XXXXX** 0.00841 | **Q5A** -3.15 | 0.0019 |
| **smoker** | 1 | 2.19877 | 0.10692 | 20.56 | <.0001 |

## Model 2:

```
PROC REG;
 MODEL hba1c = smoker / covb;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | **1 Q2A** | 223.45052 | 223.45052 **XXXXX** | **399.87 Q2E** | <.0001 |
| **Error** | **198 Q2B** | **110.64511 Q2D** | 0.55881 | | |
| **Corrected Total** | **199 Q2C** | 334.09564 | | | |

***Question 6 Extra Credit**: 5 points.* Using the output for **Model 1** calculate the correlation between fiber and HbAc1. **Justify your answers for full credit.**

**In simple linear regression,** $(correlation\ coefficent)^2 = R^2 = \frac{SS_{Model}}{SS_{Total}}$

**Then we need the SS model and SS total.**

**DF model=1**
**Then SS model = MS model * DF model =2.60748*1=2.60748**

**The F stat= t stat for fiber squared. F stat= (-1.25)^2=1.5625**
**F stat = MS model /MS error**
**Then MS error = MS model/ F stat =2.60748 /1.5625=1.668787**
**MS error DF= n-p-1=200-1-1=198**
**SS error= MS error * DF= 1.668787*198=330.4198**

**SS total= SS model +SSerror= 2.60748+330.4198=333.0273**

$$(correlation\ coefficent) = -\sqrt{R^2} = -\sqrt{\frac{SS_{Model}}{SS_{Total}}} = -\sqrt{\frac{2.60748}{333.0273}} = -0.08848519$$

```
PROC REG;
 MODEL hba1c = fiber;
RUN;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | XXXXX 1 | XXXXX 2.60748 | **2.60748** | XXXXX 1.56 | **0.2135** |
| **Error** | XXXXX 198 | XXXXX 331.48816 | XXXXX 1.67418 | | |
| **Corrected Total** | XXXXX 199 | XXXXX 334.09564 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 1 | 6.62683 | 0.23910 | 27.72 | <.0001 |
| **fiber** | 1 | -0.01855 | 0.01487 | **-1.25** | 0.2135 |

**Proc corr;**
**Var hba1c fiber;**
**run;**

| Pearson Correlation Coefficients, N = 200<br>Prob > \|r\| under H0: Rho=0 | | |
|---|---|---|
| | **hba1c** | **fiber** |
| **hba1c** | 1.00000 | -0.08834<br>0.2135 |
| **fiber** | -0.08834 | 1.00000 |