Introduction:

During the period from 1956 to 1968, several thousands of subjects participated in the Framingham Heart Study. These individuals attended three separate examination cycles, approximately 6 years apart. For this project, the data from a subset of 2,182 subjects is analyzed. First, baseline descriptive statistics are reported. Next, a Framingham Risk Score (FRS) is calculated for each subject at the third examination cycle using their measurements of sex, age, low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, systolic blood pressure, diastolic blood pressure, diabetes status, and current smoking status. These risk scores represent the ten-year probability of developing coronary heart disease (CHD). The primary outcome of interest for this project is pulse pressure (measured in mmHg), which is defined as systolic blood pressure minus diastolic blood pressure. To create representative components of the risk scores, the full set of repeated observations is used to calculate average cigarettes smoked per day, average body mass index (BMI), average heart rate, maximum serum glucose level, and maximum total cholesterol for each subject. A simple linear regression of pulse pressure as the outcome and FRS as the sole predictor is conducted. Then a multiple linear regression model of pulse pressure as the outcome and representative FRS components as predictors is selected using model selection methods. The main question of interest for this analysis is what variables are most important in explaining pulse pressure for study subjects during the third examination cycle.

Baseline Statistics:

Descriptive statistics for subjects in the Framingham Heart Study at baseline (i.e. during the first examination cycle) are provided below. Table 1A includes descriptive statistics for continuous variables, including sample size, mean, standard deviation, median, minimum, maximum, lower quartile (Q1), and upper quartile (Q3). Note that the interquartile range (IQR) can be calculated by taking the difference of the upper and lower quartiles. Table 1B includes descriptive statistics for categorical variables, including sample size, frequency (i.e. count), and percent frequency.

Table 1A: Baseline Descriptive Statistics for Continuous Variables

Variable	Sample Size	Mean	Standard Deviation	Median	Min.	Max.	Lower Quartile	Upper Quartile
Total Cholesterol (mg/dL)	2,158	229.9	41.86	226.5	113.0	464.0	200.0	256.0
Age (Years)	2,182	47.2	8.01	46.0	32.0	70.0	41.0	53.0
Systolic Blood Pressure (mmHg)	2,182	119.7	10.53	120.0	83.5	139.5	112.0	128.0
Diastolic Blood Pressure (mmHg)	2,182	76.8	7.21	78.0	50.0	89.5	72.0	82.0
Cigarettes per Day	2,160	9.2	11.62	2.0	0.0	70.0	0.0	20.0
Body Mass Index (kg/m^2)	2,176	24.9	3.45	24.6	15.5	40.6	22.5	27.0
Heart Rate (BPM)	2,182	74.1	11.22	74.0	44.0	125.0	66.0	80.0
Glucose (mg/dL)	1,980	79.9	17.82	77.0	40.0	386.0	71.0	85.0

Table 1B: Baseline Descriptive Statistics for Categorical Variables

Category	Frequency	Percent (%)
Sex (N = 2,182)		
Male	943	43.22
Female	1,239	56.78
Current Smoking Status (N = 2,182)		
Current Smoker	1,135	52.02
Not Current Smoker	1,047	47.98
Diabetes Status (N = 2,182)		
Diabetic	28	1.28
Not Diabetic	2,154	98.72
Blood Pressure Medication Status (N = 2,152)		
Currently Using BP Meds	12	0.56
Not Currently Using BP Meds	2,140	99.44

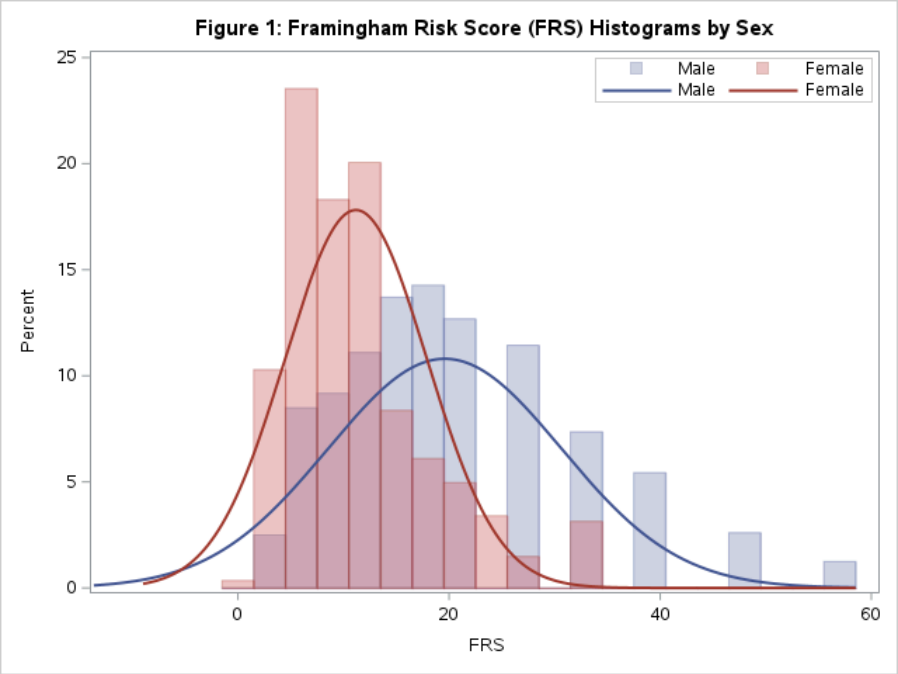
Observe that there are slightly more females than males in the sample population (56.78% vs. 43.22%). Just over half of the sample population was currently smoking during the first examination cycle (52.02%). The vast majority of study participants were not diabetic at baseline (98.72%). Only a tiny fraction of subjects were on blood pressure medications during the first period (0.56%).

Framingham Risk Score (FRS) Statistics:

Summary statistics for the Framingham Risk Score, stratified by sex, are presented below in Table 2. The FRS statistics include sample size, mean, standard deviation, median, minimum, maximum, lower quartile (Q1), and upper quartile (Q3). Note that these risk scores represent the ten-year probability of developing coronary heart disease (CHD). For example, the mean score in males is 19.6, which indicates that the average probability of developing CHD in the next ten years is 19.6% among males. The mean of the risk scores seems quite a bit higher in males compared with females (19.6 vs. 11.2). The standard deviation of risk scores appears much larger in males than females (11.08 vs. 6.72). Based on these descriptive statistics alone, Framingham Risk Scores are on average higher and have more variation in males compared with females.

Table 2: Framingham Risk Score (FRS) Descriptive Statistics by Sex

Sex	Sample Size	Mean	Standard Deviation	Median	Min.	Max.	Lower Quartile	Upper Quartile
Male	883	19.6	11.08	18.0	2.0	56.0	11.0	27.0
Female	1,147	11.2	6.72	9.0	1.0	32.0	6.0	15.0



Histograms of FRS values stratified by sex are presented on the left in Figure 1. The distribution of FRS for females is represented by the red color, and for males, the blue color. Observe that the blue histogram is centered to the right of the red histogram. This indicates that average FRS scores are higher in males than females (as seen in Table 1). Further, the distribution of FRS scores is more spread out for males than females, as indicated by the larger standard deviation in males than females displayed in Table 1. Note that a two (independent) sample t-test, with null hypothesis of no difference in FRS means between the sexes vs. alternative hypothesis of a difference in FRS means has an observed t statistic of 19.93 (with 1,367.6 degrees of freedom–using the Satterthwaite approximation due to unequal variances), and the associated p-value is < 0.0001. Therefore, the null hypothesis is rejected at the alpha = 0.05 significance level, which indicates that FRS values are statistically significantly higher in males than females. The Satterthwaite approximation was used because the p-value from the F-test for equality of two variances was < 0.0001. Rejecting the null hypothesis for the F-test indicates that the variances of FRS values differ between the sexes as observed in Table 2 and Figure 1.

Representative Components of FRS Statistics:

Summary statistics for five representative components of the FRS statistics are presented below in Table 3A. These components include average cigarettes smoked per day, average body mass index (BMI), average heart rate, maximum serum glucose level, and maximum total cholesterol calculated across the full set of repeated observations for each subject. The statistics for each variable include sample size, mean, standard deviation, median, minimum, maximum, lower quartile (Q1), and upper quartile (Q3).

Table 3A: Components of FRS Descriptive Statistics

Variable	Sample Size	Mean	Standard Deviation	Median	Min.	Max.	Lower Quartile	Upper Quartile
Average Cigarettes per Day	2,182	8.7	11.17	1.7	0.0	73.3	0.0	16.7
Average Body Mass Index (kg/m^2)	2,180	25.1	3.38	24.8	16.0	42.3	22.9	27.2
Average Heart Rate (BPM)	2,182	75.5	9.55	75.0	43.7	125.3	69.0	81.7
Maximum Glucose (mg/dL)	2,170	92.6	23.43	88.0	43.0	386.0	80.0	98.0
Maximum Total Cholesterol (mg/dL)	2,179	256.7	43.36	253.0	155.0	625.0	226.0	281.0

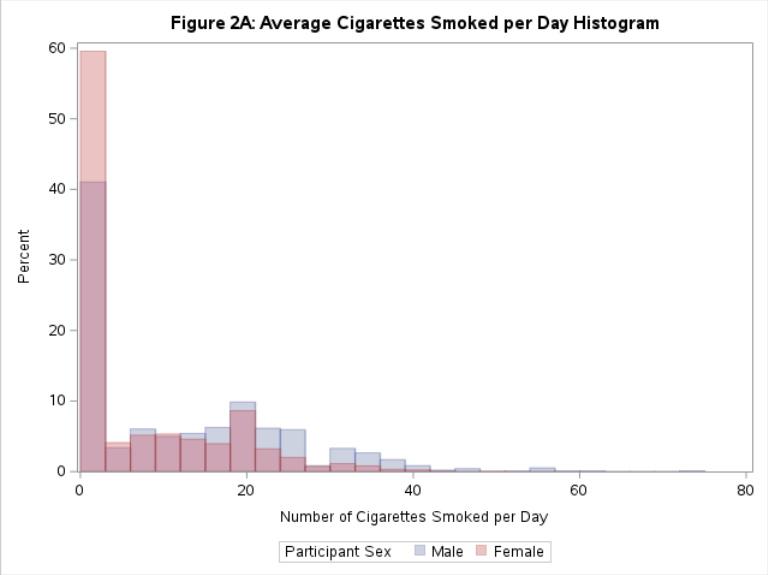
Summary statistics for the five representative components of the FRS statistics stratified by sex are displayed below in Table 3B. The statistics for each variable include the mean in males, standard deviation in

males, mean in females, standard deviation in females, mean difference (male mean – female mean), 95% confidence interval for the mean difference, t statistic from a two-sample t-test, and the associated p-value from the two-sample t-test. Observe that there is a statistically significant difference at the alpha = 0.01 significance level (the Bonferroni correction is used to adjust for multiple comparisons, so alpha = 0.05 / 5 = 0.01) between males and females for all five components (as measured by two-sample t-tests). Note that four of the components (average cigarettes per day, average BMI, maximum glucose, and maximum total cholesterol) had unequal variances (as measured by rejecting the null hypothesis for the F-test for equality of two variances) between the sexes, so the Satterthwaite approximation was used when performing the t-tests. For the average heart rate component, the null hypothesis of the F-test for equality of two variances was not rejected, so the pooled variance was used. Average cigarettes smoked per day, average body mass index, and maximum serum glucose levels were significantly higher in males compared with females. On the other hand, average heart rate and maximum total cholesterol were significantly lower in males compared with females.

Table 3B: Components of FRS Stratified by Sex

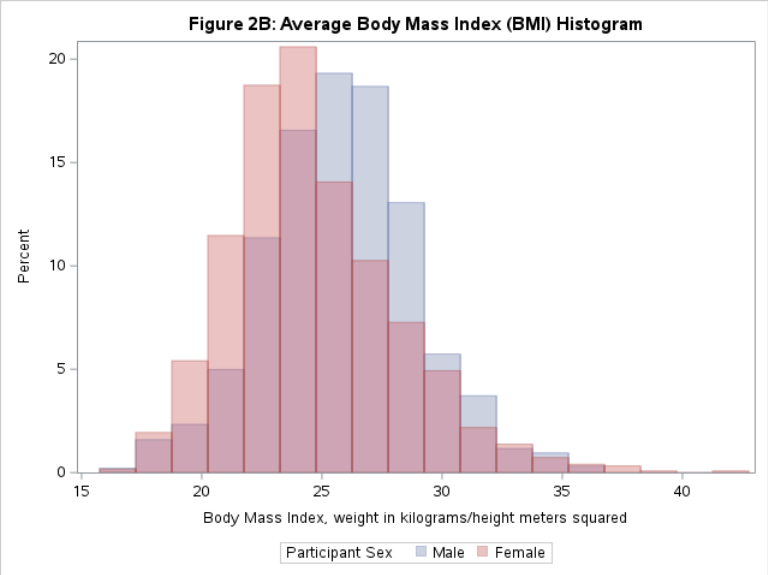
Variable	Male Mean	Male Standard Deviation	Female Mean	Female Standard Deviation	Mean Difference	95% CI Mean Difference*	t Value*	Pr > |t|
Average Cigarettes per Day	11.72	12.6366	6.05	9.2956	5.22	(4.26, 6.18)	10.67	< 0.0001
Average Body Mass Index (kg/m^2)	25.79	3.1572	24.63	3.4524	1.15	(0.87, 1.43)	8.11	< 0.0001
Average Heart Rate (BPM)	73.91	9.5197	76.62	9.4093	–2.71	(–3.51, –1.91)	–6.63	< 0.0001
Maximum Glucose (mg/dL)	95.14	25.8765	90.64	21.1761	4.50	(2.47, 6.54)	4.34	< 0.0001
Maximum Total Cholesterol (mg/dL)	250.6	39.3594	261.3	45.6527	–10.76	(–14.34, –7.18)	–5.90	< 0.0001

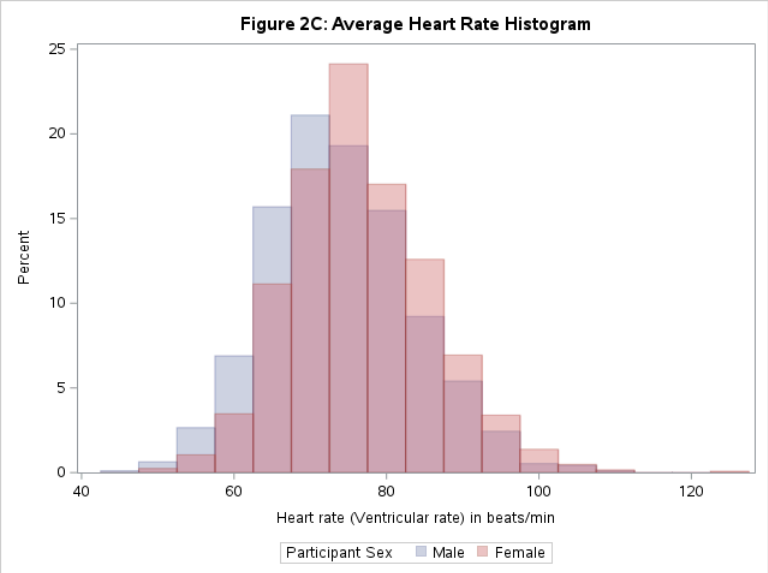
*All variables except for Average Heart Rate use the Satterthwaite method due to unequal variances

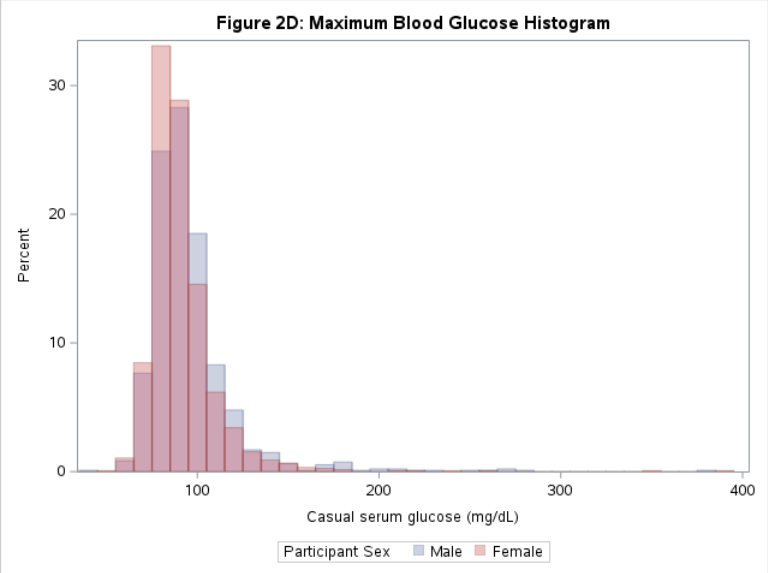


The distribution of average cigarettes smoked is more spread out in males than females (especially the longer upper tail in males), which provides visual evidence supporting the rejection of the null hypothesis for the F-test of the equality of two variances.

In Figure 2A, overlaid histograms of average cigarettes smoked per day across the three examination cycles stratified by sex are displayed. Both histograms are skewed to the right. Approximately 60% of females smoked fewer than 3 cigarettes per day on average compared with 40% of males. The distribution in males is centered to the right of the distribution in females, which agrees with the rejection of the null hypothesis for the two-sample t-test of the equality of two means.



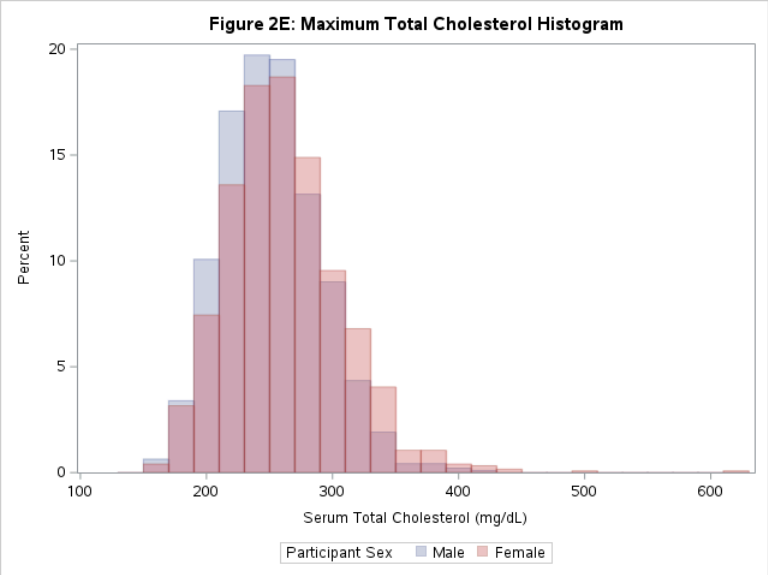




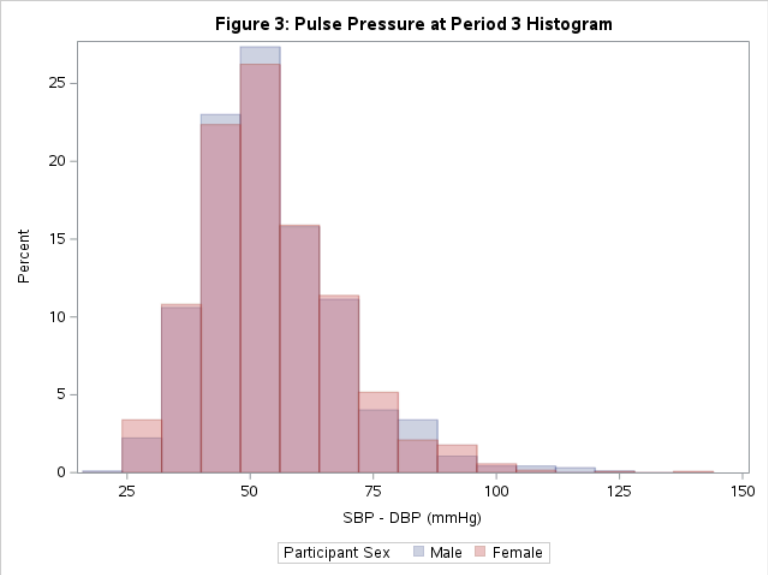
In Figure 2B, histograms for the average BMI across the three examination cycles stratified by sex are displayed. Both histograms are fairly normally distributed. The distribution of average BMI in males is clearly centered to the right of the distribution of average BMI in females, which agrees with the rejection of the null hypothesis for the two-sample t-test of the equality of two means. The male distribution appears a bit more spread out than the female distribution, which is supported by the rejection of the null hypothesis for the F-test of the equality of two variances.

In Figure 2C, average heart rate across the three examination cycles is visualized in histograms stratified by sex. Both histograms are approximately normally distributed. The distribution of average heart rate in males is centered to the left of the distribution in females, which agrees with the rejection of the null hypothesis for the two-sample t-test of the equality of two means. Both distributions have similar spread, which explains why the null hypothesis for the F-test of the equality of two variances was not rejected.

In Figure 2D, histograms for the maximum serum glucose levels across the three periods stratified by sex are displayed. Both histograms are skewed to the right, with several potential outlier observations in the upper tail. The male distribution is centered to the right of the female distribution, which visually supports the rejection of the null hypothesis for the two-sample t-test of the equality of two means and the finding that average maximum glucose levels are higher in males compared with females.



In Figure 2E, maximum total cholesterol across the three examination cycles is visualized in histograms stratified by sex. Both histograms are fairly normally distributed, with perhaps a bit of skew to the right due to a handful of potential outlier observations in the right tail. The male distribution is centered to the left of the female distribution, which visually agrees with the rejection of the null hypothesis for the two-sample t-test of the equality of two means and the conclusion that average maximum total cholesterol levels are lower in males compared with females.



In Figure 3, overlaid histograms of pulse pressure during the third examination cycle (outcome variable) stratified by sex are displayed. Both histograms are slightly skewed to the right. This indicates that a natural logarithm transformation of the outcome variable may be considered when conducting linear regression.

Simple Linear Regression:

In order to study the relationship between examination cycle three risk scores and examination cycle three pulse pressure, a simple linear regression of the outcome period three pulse pressure and the continuous predictor period three FRS is conducted. The model is summarized in Tables 4A & 4B below.

Table 4A: Simple Linear Regression of Pulse Pressure and FRS Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	76,583	76,583	471.19	< 0.0001
Error	2,028	329,612	162.53034		
Corrected Total	2,029	406,195			

Table 4B: Simple Linear Regression of Pulse Pressure and FRS Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	43.89972	0.51314	85.55	< 0.0001
FRS	1	0.62616	0.02885	21.71	< 0.0001
R-Square = 0.1885; Adjusted R-Square = 0.1881; AIC = 12,368; BIC = 10,338; C(p) = 2.00					

The global null hypothesis states that the FRS beta coefficient is equal to zero. The alternative global hypothesis posits that the FRS beta coefficient differs from zero. For the global test, the observed F statistic is 471.19 (with 1 numerator and 2,028 denominator degrees of freedom), and the associated p-value is < 0.0001 (Table 4A). Therefore, we reject the global null hypothesis at the alpha = 0.05 significance level. We conclude that there is evidence to reject the null hypothesis that the beta coefficient for the FRS predictor is equal to zero.

The simple linear regression model has an R^2 value of 0.1885 (Table 4B). This indicates that approximately 19% of the variability in period three pulse pressure can be explained by the period three Framingham Risk Score predictor. Therefore, this model is quite poor at predicting period three pulse pressure.

For the FRS predictor, the parameter estimate is 0.63 (SE = 0.029). This means that for each one point increase in FRS (i.e. each one percentage point increase in the ten-year probability of developing CHD), period three pulse pressure increases by an average of 0.63 (95% CI: 0.57, 0.68) mmHg. Since the parameter estimate is positive, higher examination cycle three risk scores are associated with higher examination cycle three pulse pressures. Note that a simple linear regression with the transformed outcome of pulse pressure using the natural logarithm (since pulse pressure is skewed to the right) was run, but this model appeared to be roughly equivalent to the model presented above. This model is not presented due to space constraints.

Multiple Linear Regression:

The next step of the analysis is to determine a multiple linear regression model of examination cycle three pulse pressure as the outcome and representative FRS components as predictors. First, the full model with all possible predictors (period three age, sex, average cigarettes smoked, average BMI, average heart rate, maximum serum glucose, and maximum total cholesterol) is analyzed prior to performing model selection. The full model is presented below in Tables 5A & 5B.

Table 5A: Full Multiple Linear Regression Model of Pulse Pressure Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	80,457	11,494	68.65	< 0.0001
Error	2,159	361,485	167.43181		
Corrected Total	2,166	441,943			

Table 5B: Full Multiple Linear Regression Model of Pulse Pressure Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Type II Semi-Partial R^2	VIF
Intercept	1	-12.84582	4.13967	-3.10	0.0019	0	.	0
Age (Period 3)	1	0.73322	0.03680	19.93	< 0.0001	0.41035	0.15043	1.119
Sex (Female vs. Male)	1	1.42027	0.61421	2.31	0.0209	0.04928	0.00203	1.199
Average Cigarettes Smoked	1	0.05650	0.02708	2.09	0.0371	0.04407	0.00165	1.178

Average BMI	1	0.40639	0.08464	4.80	< 0.0001	0.09586	0.00873	1.052
Average Heart Rate	1	0.07212	0.03020	2.39	0.0170	0.04830	0.00216	1.079
Maximum Glucose	1	0.04020	0.01211	3.32	0.0009	0.06598	0.00418	1.043
Maximum Total Cholesterol	1	0.00269	0.00660	0.41	0.6836	0.00818	0.00006296	1.063
R-Square = 0.1821; Adjusted R-Square = 0.1794; AIC = 13,273; BIC = 11,106; C(p) = 8.00								

Due to space constraints, the results of the full model are not reported in detail. However, note that the model is statistically significant (F-statistic = 68.65, p-value < 0.0001), and all predictors except for maximum total cholesterol are statistically significant at the alpha = 0.05 significance level. Model selection criteria, including AIC, BIC, Mallows' Cp, and Adjusted R^2 are reported at the bottom of Table 5B.

Multiple Linear Regression Model Selection:

To determine the final multiple linear regression model (with pulse pressure during period three as the outcome variable), stepwise selection using several different criteria (AIC, BIC, Adjusted R^2, Mallows' Cp) was undertaken. The selected model includes six predictors (Period Three Age, Sex, Average Cigarettes Smoked, Average BMI, Average Heart Rate, and Maximum Glucose), which is presented in Tables 6A & 6B. All four metrics (AIC, BIC, Adjusted R^2, and Mallows' Cp) resulted in this model being selected during stepwise selection. The value of Mallows' Cp for the final model is 6.17, which is close to the number of predictors plus one (7). In the full model, all but one of the predictors (maximum total cholesterol) was statistically significant at the alpha = 0.05 significance level; maximum total cholesterol ends up not being included in the final model.

Table 6A: Final Multiple Linear Regression Model of Pulse Pressure Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	80,422	13,404	80.11	< 0.0001
Error	2,161	361,560	167.31137		
Corrected Total	2,167	441,982			

Table 6B: Final Multiple Linear Regression Model of Pulse Pressure Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Type II Semi-Partial R^2	VIF
Intercept	1	-12.51556	4.05200	-3.09	0.0220	0	.	0
Age (Period 3)	1	0.73589	0.03610	20.38	< 0.0001	0.41183	0.15728	1.078
Sex (Female vs. Male)	1	1.45074	0.60739	2.39	0.0170	0.05035	0.00216	1.174
Average Cigarettes Smoked	1	0.05661	0.02705	2.09	0.0365	0.04416	0.00166	1.176
Average BMI	1	0.41030	0.08436	4.86	< 0.0001	0.09686	0.00896	1.048
Average Heart Rate	1	0.07282	0.03014	2.42	0.0158	0.04876	0.00221	1.076
Maximum Glucose	1	0.04020	0.01210	3.32	0.0009	0.06598	0.00418	1.042
R-Square = 0.1820; Adjusted R-Square = 0.1797; AIC = 13,271; BIC = 11,104; C(p) = 6.17								

For the final model, the global null hypothesis states that all six beta coefficients are equal to zero. The alternative global hypothesis posits that at least one of the beta coefficients differs from zero. For the global test, the observed F statistic is 80.11 (with 6 numerator and 2,161 denominator degrees of freedom), and the associated p-value is < 0.0001 (Table 6A). Therefore, we reject the global null hypothesis at the alpha = 0.05 significance level. We conclude that there is evidence to reject the null hypothesis that the beta coefficients are all equal to zero. In other words, the model is statistically significant, which indicates that at least one predictor has a statistically significant beta coefficient.

The final model has an R^2 value of 0.1820 (Table 6B). This indicates that approximately 18% of the variability in period three pulse pressure can be explained by the predictors age, sex, average cigarettes smoked, average BMI, average heart rate, and maximum glucose. Since this is a multiple linear regression, a more relevant quantity is the adjusted R^2 value (which adjusts for the number of predictors). The adjusted R^2 value is 0.1797 (Table 6B). As a result, this model is quite poor at predicting examination cycle three pulse pressure.

Observe that the variance inflation factor (VIF) values are close to one (largest value is 1.176) for all six predictors in the final model (Table 6B). Thus, there are no issues regarding collinearity.

For the age variable, the observed t statistic is 20.38 (with 2,161 degrees of freedom), and the associated p-value is < 0.0001 (Table 6B). Therefore, we reject the null hypothesis at the alpha = 0.05 significance level. We conclude that the age predictor is statistically significant (when adjusting for the other five predictors). The parameter estimate for the age variable is 0.74 (SE = 0.036). This means that for each one year increase in age during period three, pulse pressure during period three increases by an average of 0.74 (95% CI: 0.67, 0.81) mmHg when all other predictors are held fixed. Since the parameter estimate is positive, older age during period three is associated with higher pulse pressure when adjusting for all other variables in the model.

The standardized parameter estimate for age is 0.41, which is the largest standardized beta estimate in the model by a factor of four (average BMI is the second largest, with an estimate of 0.10). In addition, the Type II semi-partial R^2 value for the age predictor is 0.16, which is the largest semi-partial R^2 value in the model by a factor of seventeen (average BMI is the second largest, with a semi-partial R^2 value of 0.009). This indicates that period three age is the most important predictor of period three pulse pressure in the model.

The sex variable has an observed t statistic of 2.39 (with 2,161 degrees of freedom), and the associated p-value is 0.0170 (Table 6B). Therefore, we reject the null hypothesis at the alpha = 0.05 significance level. We conclude that the sex predictor is statistically significant (when adjusting for the other five predictors). The parameter estimate for the sex variable is 1.45 (SE = 0.607) [note that male sex is the reference group]. This demonstrates that pulse pressure during period three is an average of 1.45 (95% CI: 0.26, 2.64) mmHg higher in females compared with males when all other predictors in the model are held fixed.

For the average cigarettes smoked variable, the observed t statistic is 2.09 (with 2,161 degrees of freedom), and the associated p-value is 0.0365 (Table 6B). Therefore, we reject the null hypothesis at the alpha = 0.05 significance level. We conclude that the average cigarettes smoked predictor is statistically significant (when adjusting for the other five predictors). The parameter estimate for the cigarettes smoked variable is 0.057 (SE = 0.027). This indicates that for each one cigarette increase in average cigarettes smoked, pulse pressure during period three increases by an average of 0.057 (95% CI: 0.0036, 0.11) mmHg when all other predictors are held fixed. Since the parameter estimate is positive, higher cigarette usage is associated with higher period three pulse pressure when adjusting for all other variables in the model.

The average body mass index variable has an observed t statistic of 4.86 (with 2,161 degrees of freedom), and the associated p-value is < 0.0001 (Table 6B). Therefore, we reject the null hypothesis at the alpha = 0.05 significance level. We conclude that the average BMI predictor is statistically significant (when adjusting for the other five predictors). The parameter estimate for the average BMI variable is 0.41 (SE = 0.084). This signifies that for each one kg/m^2 increase in average body mass index, pulse pressure during period three increases by an average of 0.41 (95% CI: 0.24, 0.58) mmHg when all other predictors are held fixed. Since the parameter estimate is positive, higher body mass index is associated with higher pulse pressure when adjusting for all other variables.

For the heart rate variable, the observed t statistic is 2.42 (with 2,161 degrees of freedom), and the associated p-value is 0.0158 (Table 6B). Therefore, we reject the null hypothesis at the alpha = 0.05 significance level. We conclude that the average heart rate predictor is statistically significant (when adjusting for the other five predictors). The parameter estimate for the average heart rate variable is 0.073 (SE = 0.030). This means that for each one beat per minute increase in average heart rate, pulse pressure during period three increases by an average of 0.073 (95% CI: 0.014, 0.13) mmHg when all other predictors are fixed. Since the parameter estimate is positive, faster heart rate is associated with higher period three pulse pressure when adjusting for all other variables in the model.

The glucose variable has an observed t statistic of 3.32 (with 2,161 degrees of freedom), and the associated p-value is 0.0009 (Table 6B). Therefore, we reject the null hypothesis at the alpha = 0.05 significance level. We conclude that the maximum glucose predictor is statistically significant (when adjusting for the other five predictors). The parameter estimate for the maximum glucose variable is 0.040 (SE = 0.012). This suggests that for each one mg/dL increase in maximum serum glucose, pulse pressure during period three increases by an average of 0.040 (95% CI: 0.016, 0.064) mmHg when all other predictors are fixed. Since the parameter estimate is positive, higher glucose levels are associated with higher pulse pressure when adjusting for all other variables.

Comparison and Summary:

When comparing the simple linear regression model with the multiple linear regression model, observe that the simple model yields better model fit metrics (AIC: 12,368 vs. 13,271; BIC: 10,338 vs. 11,104; Adjusted R^2: 0.1881 vs. 0.1797). Using a single predictor as opposed to several predictors can be desirable in reducing the cost of analysis. However, since the Framingham Risk Score relies on several components, a subject missing data from *any* of the components during period three will result in a missing value for FRS, which reduces the number of data points in the regression (there are 2,030 observations in the simple linear regression vs. 2,168 in the final multiple linear regression). Recall that the risk score only depends on period three data while the representative FRS components depend on data across all three visits. Therefore, the interpretation for the simple linear regression model only relates to period three; specifically, the effect of the period three ten year probability of developing CHD on pulse pressure. On the other hand, the interpretation for the multiple linear regression model relates to the entire set of repeated observations over the three examination cycles. In both cases, the adjusted R^2 values are very poor, indicating that neither model is particularly good at predicting pulse pressure. Nevertheless, directly comparing both models isn't exactly fair because the subjects used to estimate the models differ (i.e. simple linear regression model has fewer data points). Due to the fact that both models are not based on the exact same set of observations, direct comparisons of model fit are not appropriate. Since the simple linear regression model relies on only period three observations and is more prone to issues relating to missing data, the multiple linear regression model is the better approach. As a result, the multiple linear regression model is preferred over the simple linear regression model. Future analysis could involve imputation methods to address missing data issues, analysis of potential interaction between variables, or use of higher order polynomials as predictors (e.g. age^2, age^3, etc.). In addition, other outcome variables could be studied, such as binary hypertension status (using logistic regression) or time to coronary heart disease/other cardiovascular events (using survival analysis).

Conclusion:

The primary question of interest for this project was to determine what factors are most predictive of period three pulse pressure among subjects in the Framingham Heart Study. Framingham Risk Scores (FRS) are calculated for the third examination cycle, which take into account age, sex, low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, systolic blood pressure, diastolic blood pressure, diabetes status, and current smoking status. These risk scores signify the ten-year probability of developing coronary heart disease (CHD). The primary outcome of interest for this project is examination cycle three pulse pressure (measured in mmHg), which is defined as systolic blood pressure minus diastolic blood pressure. Risk scores were found to be statistically significantly higher in males than females, which indicates that males have a higher risk of developing CHD than females. Using data from across all three visits for all subjects, representative components of FRS were constructed. For all five components, there was a statistically significant difference in means between males and females. The components average cigarettes smoked per day, average body mass index, and maximum serum glucose levels were significantly higher in males compared with females. On the contrary, the components average heart rate and maximum total cholesterol were significantly lower in males compared with females. In addition, four of the five components showed a statistically significant difference in the variance between males and females. A simple linear regression with the outcome variable period three pulse pressure and a single predictor period three FRS demonstrated that higher risk scores are associated with increased examination cycle three pulse pressure. A multiple linear regression with the outcome variable period three pulse pressure and five components of FRS as predictors suggested that older period three age, female sex, greater average cigarettes smoked, higher average heart rate, and higher maximum glucose levels are all associated with increased examination cycle three pulse pressure. However, period three age is by far the most important predictor of pulse pressure. Note that there are several limitations of this work. For instance, the sample population of Framingham, Massachusetts consisted of mostly white individuals from the 1950s and 1960s, which may not be representative of the general United States population in the 2020s.