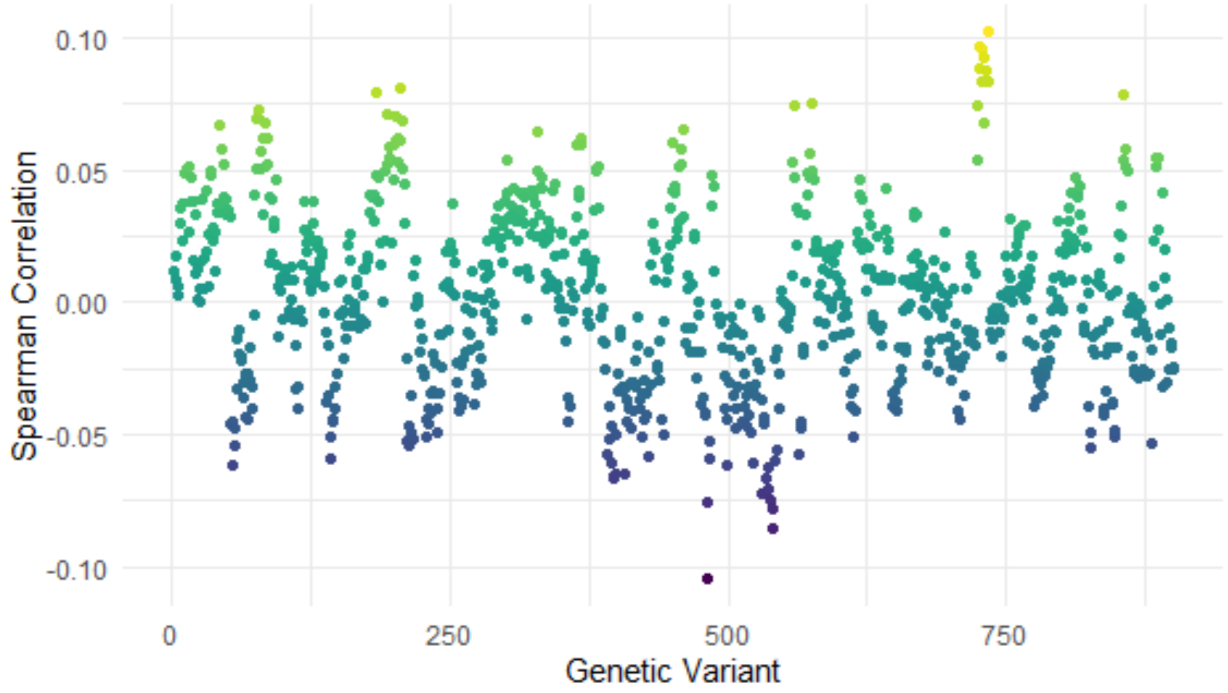
Problem 1A

Figure Q1.1 Pairwise correlation between genetic variants in matrices X and Z.



As seen in Figure Q1.1, the Spearman pairwise correlation oscillates between -0.104 and +0.104 with clustering high positive correlation ($\rho > 0.08$) for genetic variants 204 and 726-734 and clustering high negative correlation ($\rho < -0.07$) for genetic variants 481-482 and 531, 535, 537-540.

The oscillatory behavior of the correlation plot suggests that neighboring genetic variants are similarly correlated, i.e., if genetic variant 750 for cognition (Z) and structure (X) data have a correlation of 0.06, then we would expect a similar pairwise correlation between X and Z for genetic variant 751. This suggests that the hypothesis of linkage disequilibrium, or correlation between nearby variants, is justified for this data.

No, the estimates are not unbiased.

$$E\left[\frac{1}{n}X^T y_1\right] = \frac{1}{n}X^T E[y_1] = \frac{1}{n}X^T X\alpha \quad \text{(Q1.1)}$$

$$E\left[\frac{1}{n}X^T y_2\right] = \frac{1}{n}X^T E[y_2] = \frac{1}{n}X^T X\beta \quad \text{(Q1.2)}$$

$$E\left[\frac{1}{n}Z^T y_3\right] = \frac{1}{n}Z^T E[y_3] = \frac{1}{n}Z^T Z\eta \quad \text{(Q1.3)}$$

The only way for the above estimators to be unbiased would be if $\frac{1}{n}X^T X = I$ and $\frac{1}{n}Z^T Z = I$ in Equations (Q1.1) – (Q1.3). However, this is not the case for these data, as

evidenced by a quick matrix product which results in values of approximately 1 (0.999) on the main diagonal, but off diagonal entries which are non-zero.

Problem 1B

In this problem, the standard error and basic percentile bootstrap CI were computed using the following algorithm:

Algorithm to compute the standard error and basic 95% bootstrap CI.

1. Compute $\hat{\varphi}_{13}$ for the original $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ using $\varphi_{13} = \frac{\boldsymbol{\alpha}^T\boldsymbol{\eta}}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}\sqrt{\boldsymbol{\eta}^T\boldsymbol{\eta}}}$.
2. For b in 1 to B
    - 2.1.1. Sample from $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ with replacement. Call these vectors $\boldsymbol{\alpha}^*$ and $\boldsymbol{\eta}^*$.
    - 2.1.2. Compute $\hat{\varphi}^*_{13,b} = \frac{\boldsymbol{\alpha}^{*T}\boldsymbol{\eta}^*}{\sqrt{\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*}\sqrt{\boldsymbol{\eta}^{*T}\boldsymbol{\eta}^*}}$.
3. Compute estimate for standard error of $\hat{\boldsymbol{\varphi}}^*$ as $\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\varphi}^*_{13,b} - \frac{1}{B}\sum_{b=1}^{B}\hat{\varphi}^*_{13,b})^2}$.
4. Order the elements in $\hat{\boldsymbol{\varphi}}^*$ as $\hat{\boldsymbol{\varphi}}^*_{(1)}, \ldots, \hat{\boldsymbol{\varphi}}^*_{(B)}$.
5. Compute the basic 95% bootstrap CI as $(\hat{\boldsymbol{\varphi}}^*_{(\lfloor 0.025B \rfloor)}, \hat{\boldsymbol{\varphi}}^*_{(\lfloor 0.975B \rfloor)})$ where $\lfloor . \rfloor$ rounds down to the nearest integer.
6. Same process holds for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$.

Table Q1.1: Estimates of the genetic correlations and corresponding bootstrap estimators of the standard error and 95% CI.

	Estimate	*Standard Error**	*95% CI***
Genetic Correlations			
φ_{13}	0.228	0.031	(0.168, 0.288)
φ_{23}	0.549	0.026	(0.495, 0.598)

*Estimate for standard error computed by 5,000 bootstraps.
**Estimate for basic percentile CI computed by 5,000 bootstraps.

As seen in Table Q1.1, the estimated correlation between the thalamus and cognition was over twice that of the hippocampus and cognition (0.549 versus 0.228, respectively). There was also slightly greater precision in this estimate ($\widehat{SE}(\widehat{\varphi_{23}}) = 0.026$ versus $\widehat{SE}(\widehat{\varphi_{13}}) = 0.031$). However, I'm not sure I trust these correlations, as, from a biological perspective, they seem fairly inflated.

Since each 95% bootstrap CI covers the estimate without including zero, we can conclude that (for $\alpha = 0.05$) there is evidence that the estimated correlations are non-zero.

These results may suggest that, although both regions of the brain appear to be partly associated with cognition, the thalamus may play a bigger role in cognitive processing than the hippocampus. The hippocampus is known to both directly and indirectly affect thalamic function, so, in the causal pathway, it may hold that the thalamus is the mediator. However, the limbic system is complex, and there are *many* possible casual pathways that could explain these findings.

Problem 1C

We will use the formula for partial correlation given on page 41, and derived on pages 42-43, of T.W Anderson's book, "An Introduction to Multivariate Statistical Analysis."

Conditional on the thalamus, the genetic correlation between the hippocampus and cognition can estimated using

$$\varphi_{13.2} = \frac{(\varphi_{13} - \varphi_{12}\varphi_{23})}{\sqrt{(1-\varphi_{12}^2)(1-\varphi_{23}^2)}} \quad \text{(Q1.4)}$$

where $\varphi_{13} = \frac{\alpha^T \eta}{\sqrt{\alpha^T \alpha}\sqrt{\eta^T \eta}}$, $\varphi_{23} = \frac{\beta^T \eta}{\sqrt{\beta^T \beta}\sqrt{\eta^T \eta}}$, and $\varphi_{12} = \frac{\alpha^T \alpha}{\sqrt{\alpha^T \alpha}\sqrt{\beta^T \beta}}$.

Table Q1.2: Estimated correlation between the hippocampus and cognition after conditioning on the thalamus with corresponding bootstrap SE and 95% CI.

	Estimate	*Standard Error**	*95% CI***
Genetic Correlation $\varphi_{13.2}$	0.001	0.039	(-0.075, 0.077)

*Estimate for standard error computed by 5,000 bootstraps.
**Estimate for basic percentile CI computed by 5,000 bootstraps.

When we condition on the thalamus, as seen in Table Q1.2, there is no evidence of residual correlation ($\varphi_{13.2} = 0.001$, 95% CI: (-0.075, 0.077)) between the hippocampus and cognition. This suggests, as previously hypothesized, that the hippocampus indirectly influences cognitive function through its connections with the thalamus. Now, whether the thalamus acts directly on cognitive function or influences other organs in the brain (which then, in turn, make you a statistics wizard), cannot be determined from these data.

Problem 1D

Table Q1.3: Comparing estimates of genetic correlation using marginal and OLS estimators.

	Estimate	*Standard Error**	*95% CI***
From Marginal Estimators			
φ_{13}	0.228	0.031	(0.168, 0.288)
φ_{23}	0.549	0.026	(0.495, 0.598)
From OLS Estimators			
φ_{13}	0.098	0.034	(0.032, 0.165)
φ_{23}	0.083	0.032	(0.019, 0.145)

*Estimate for standard error computed by 5,000 bootstraps.
**Estimate for basic percentile CI computed by 5,000 bootstraps.

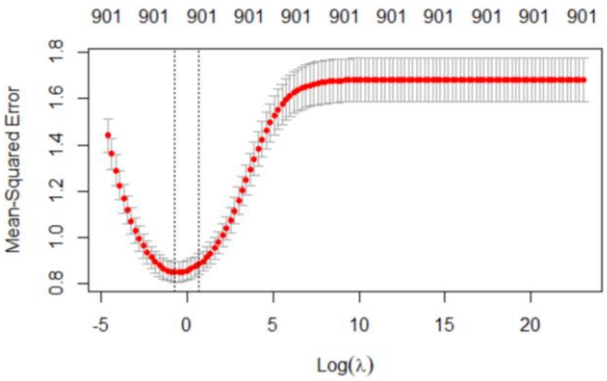
This exam has taught me the how important it is to use an unbiased estimator with small variance, whenever feasible. Recall that the OLS estimator is the best linear unbiased estimator (BLUE), by the Gauss Markov theorem. Thus, I am more inclined to trust the results derived using this estimator over the results derived using the marginal estimator. However, I am assuming that the main assumptions for fitting a linear regression hold, including linearity, homoskedasticity, independence, and normality of the errors. Otherwise, Gauss Markov does not apply.

It is interesting to note that there is now very little correlation ($\hat{\varphi}_{O,13} = 0.098$ and $\hat{\varphi}_{O,23} = 0.083$ compared to $\hat{\varphi}_{13} = 0.228$ and $\hat{\varphi}_{23} = 0.549$) between both limbic structures and cognitive function. However, despite the correlation change in magnitude, the standard errors did not decrease. Surprisingly, for the OLS estimators, they are slightly larger at $SE(\hat{\varphi}_{O,13}) = 0.034$ (versus the marginal estimate of 0.031) and $SE(\hat{\varphi}_{O,23}) = 0.032$ (versus the marginal estimate of 0.026). This makes me suspicious of the OLS estimates.

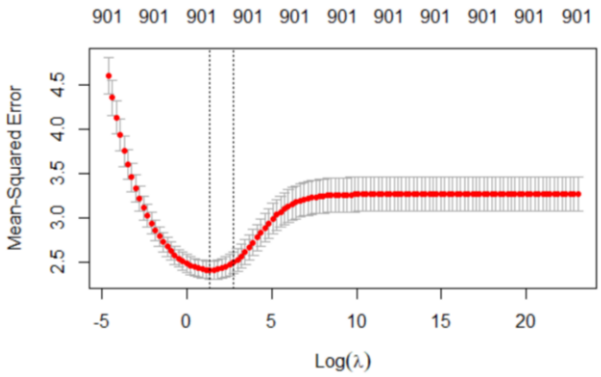
Problem 1E

Figure Q1.2: Tuning parameter profile plots for model in A) hippocampus, B) thalamus, and C) cognition.

A)



B)



C)

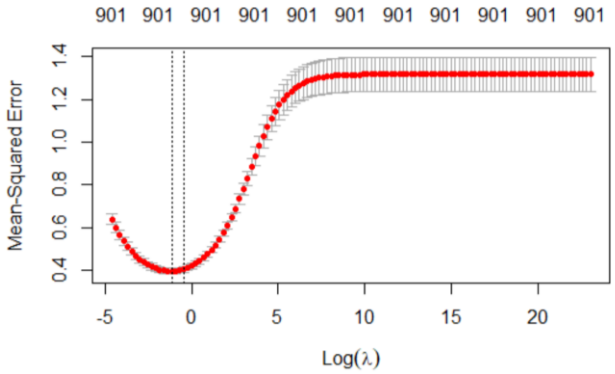


Table Q1.4: Estimates of genetic correlation using marginal, OLS, and ridge estimators.

	Estimate
Genetic Correlation	
From Marginal Estimators	
φ_{13}	0.228
φ_{23}	0.549
From OLS Estimators	
φ_{13}	0.098
φ_{23}	0.083
From Ridge Estimators	
φ_{13}	0.095
φ_{23}	0.401

The optimal tuning parameter (λ) was chosen as the largest lambda value within one standard error of the minimum lambda, as suggested by (Friedman, Hastie, and Tibshirani, 2010). Tuning parameters were sequenced from $\lambda = 10^{10}$ to $\lambda = 10^{-2}$. The optimal tuning parameters for models involving $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\eta}$ were 1.995, 15.849, and 0.631, respectively (Figure Q1.2 – note tuning parameters are plotted on a log scale).

As seen in Table Q1.4, the OLS estimators and ridge estimators for φ_{13} (the genetic correlation between the hippocampus and cognition) are very similar at 0.098 and 0.095, respectively. This suggests that, since the effects of other covariates were already removed by the investigators, this correlation is probably an accurate representation of the magnitude of the association between the hippocampus and cognition.

However, notice the similarity between the marginal and ridge estimator for φ_{23} of 0.549 and 0.401, respectively, which differ greatly from the OLS estimator of 0.083. I suspect these differences were due to multicollinearity. When multicollinearity exists in a dataset, OLS estimates (while still unbiased) can have inflated variances. Thus, ridge regression is often used to fit a multivariate regression with two or more predictors that are highly correlated. In the case of severe multicollinearity, X can be rank deficient, and thus X^TX will be singular. Adding a small penalty to X^TX will allow it to be invertible and so we can find a unique estimate.

Problem 1F

Table Q1.5: Median absolute error (MDAE) between simulated genetic effect $\boldsymbol{\alpha_s}$ and both marginal estimator $\hat{\boldsymbol{\alpha}}_s$ and OLS estimator $\hat{\boldsymbol{\alpha}}_{Os}$.

		*Marginal MDAE**	*OLS MDAE*
Distribution			
$\boldsymbol{\alpha_s} \sim N(0,1)$	$\boldsymbol{\epsilon_{\alpha s}} \sim N(0,1)$	2.46	0.29
$\boldsymbol{\alpha_s} \sim N(0,1)$	$\boldsymbol{\epsilon_{\alpha s}} \sim Unif(0,100)$	2.63	8.49
$\boldsymbol{\alpha_s} \sim \Gamma(1,100)$	$\boldsymbol{\epsilon_{\alpha s}} \sim N(0,1)$	0.203	0.295
$\boldsymbol{\alpha_s} \sim \Gamma(1,100)$	$\boldsymbol{\epsilon_{\alpha s}} \sim Unif(0,100)$	0.627	8.46

*Median absolute error (MDAE) across 100 simulations.

Algorithm to compare estimators under simulation:

1. Simulate $\boldsymbol{y_{s1}} = \boldsymbol{X\alpha_s} + \boldsymbol{\epsilon_{\alpha s}}$
2. Compute the marginal estimator $\hat{\boldsymbol{\alpha}}_s = \left(\frac{1}{n}\right)\boldsymbol{X^T y_{s1}}$
3. Compute the OLS estimator $\hat{\boldsymbol{\alpha}}_{\boldsymbol{Os}} = (\boldsymbol{X^TX})^{-1}\boldsymbol{X^T y_{s1}}$

4. Compute the median absolute error for each estimator. Namely, compute $median(|\alpha_{s,1} - \hat{\alpha}_{s,1}|, \dots, |\alpha_{s,p} - \hat{\alpha}_{s,p}|)$ and $median(|\alpha_{s,1} - \hat{\alpha}_{Os,1}|, \dots, |\alpha_{s,p} - \hat{\alpha}_{Os,p}|)$ where p is the length of the vector.

5. Generate $\boldsymbol{\alpha_s} \sim Dist1$ and $\boldsymbol{\epsilon_{\alpha s}} \sim Dist2$ where $Dist1$ and $Dist2$ are chosen by the user with dimensions $p \times 1$ and $n \times 1$, respectively. Repeat steps 1-4 for each new generation and then average the results to get an estimate of performance.

The performance of the estimators under simulation depends on the input distributions for $\boldsymbol{\alpha_s}$ and $\boldsymbol{\epsilon_{\alpha s}}$. The OLS estimator is the BLUE for the linear model under a set of distributional assumptions. If these assumptions are violated, then the BLUE may not be the BLUE, as was observed in the simulations presented in Table Q1.5.

For the wrong choice of error function (i.e., non-gaussian) or, in the case of severe outliers, the OLS estimator can perform poorly. By generating non-gaussian error and by generating severely-skewed data with long tails, we were able to generate cases for which the marginal estimator outperformed the OLS estimator in terms of median absolute error (MDAE). It is important to note that we had to generate severe scenarios before the marginal estimator was able to beat the OLS estimator. Generally, the OLS estimator was robust (compared to the marginal estimator) to departures in normality of the standard error and skewed distributions. However, notice that the marginal MDAE is more consistent across the different scenarios than the OLS MDAE.

For standard normal genetic effect ($\boldsymbol{\alpha_s}$) and standard normal error ($\boldsymbol{\epsilon_{\alpha s}}$), the OLS estimator outperformed the marginal estimator across 100 simulations (OLS MDAE of 0.29 vs Marginal MDAE of 2.46).

For standard normal genetic effect but $Unif(0{,}100)$ error, the marginal estimator well-outperformed the OLS estimator across 100 simulations (Marginal MDAE of 2.63 vs OLS MDAE of 8.49).

For $\Gamma(1{,}100)$ genetic effect but standard normal error, the marginal estimator slightly outperformed the OLS estimator across 100 simulations (Marginal MDAE of 0.203 vs OLS MDAE of 0.295).

Finally, for $\Gamma(1{,}100)$ genetic effect and $Unif(0{,}100)$ error, the marginal estimator well-outperformed the OLS estimator across 100 simulations (Marginal MDAE of 0.627 vs OLS MDAE of 8.46).

Problem 1G

Note: For this problem, $\boldsymbol{\alpha_s} \sim N(0,1)$, $\boldsymbol{\eta_s} \sim N(0,1)$, $\boldsymbol{\epsilon_{\alpha s}} \sim N(0,0.1)$, and $\boldsymbol{\epsilon_{\eta s}} \sim N(0,0.1)$ to generate $\boldsymbol{y_{s1}}$ and $\boldsymbol{y_{s3}}$ via the algorithm described in f).

Table Q1.6: Estimated genetic correlation obtained via marginal estimators of the simulated genetic effect $\boldsymbol{a_s}$ with corresponding bootstrap SE and 95%CI.

	Estimate	*Standard Error**	*95% CI***
Genetic Correlation			
φ_{s13}	0.244	0.028	(0.199, 0.298)

*Estimate for standard error computed by 5,000 bootstraps.
**Estimate for basic percentile CI computed by 5,000 bootstraps.

If the estimator for the genetic correlation is unbiased, then the bias should be approximately zero. Bias was computed using

$$bias = \frac{1}{B}\sum_{b=1}^{B}(\hat{\varphi}_{s13,b} - \varphi_{s13,b}) \quad \text{(Q1.5)}$$

where B is the number of simulations, $\hat{\varphi}_{s13,b}$ is the estimated genetic correlation at the b^{th} simulation using the marginal estimators, and $\varphi_{s13,b}$ is the truth at the b^{th} simulation.

A total of 100 simulations using Equation (Q1.5) revealed a bias of 0.006 for $\hat{\varphi}_{s13,b}$, which is reasonably large considering that the mean magnitude of the true and estimated correlations (for these simulations) was 0.12 and 0.03, respectively. We already know from Table Q1.5 in part f) that the marginal estimator for $\boldsymbol{\alpha_s}$ consistently over- or under-estimates $\boldsymbol{\alpha_s}$. The same would be expected for $\boldsymbol{\eta_s}$**.** The formula for the genetic correlation requires repeat products of two biased estimators (one product in the numerator and two products in the denominator), thereby, introducing more error into the estimation.

A factor that further amplifies the bias present in $\hat{\boldsymbol{\alpha}}_{\boldsymbol{s}}$ and $\hat{\boldsymbol{\eta}}_{\boldsymbol{s}}$ are the distributions used to model the coefficient and error terms. Table Q1.5 from part f) displays differences in median absolute error when these distributions are altered.

Problem 2A

Table Q2.1: Treatment group composition by patient status, medical condition, and cost of treatment.

	Treatment A (N=110)	*Treatment B (N=243)*	*Treatment C (N=87)*	*Overall (N=440)*
Patient Status				
Stable	95 (21.6%)	121 (27.5%)	44 (10.0%)	260 (59.1%)
Unstable	15 (3.4%)	122 (27.7%)	43 (9.8%)	180 (40.9%)
Medical Condition				
Cardiovascular Disease	15 (3.4%)	54 (12.3%)	47 (10.7%)	116 (26.4%)
Liver Disease	47 (10.7%)	101 (23.0%)	21 (4.8%)	169 (38.4%)
Kidney Disease	48 (10.9%)	88 (20.0%)	19 (4.3%)	115 (35.2%)
Treatment Cost				
< $1k	78 (17.7%)	105 (23.9%)	25 (5.7%)	208 (47.3%)
$1k-3k	17 (3.9%)	82 (18.6%)	17 (3.9%)	116 (26.4%)
> $3k	15 (3.4%)	56 (12.7%)	45 (10.2%)	116 (26.4%)

Note: Values expressed as N (%). Percentages may not sum to exactly 100 due to rounding error. There were no missing values in this dataset.

Problem 2B

Fixed-Effects Polytomous Logistic Regression Model:

$$\log\left\{\frac{P(T_i=1|\boldsymbol{X})}{P(T_i=0|\boldsymbol{X})}\right\} = \alpha_1 + \sum_{j=1}^{3}\beta_{1j}X_{ij} \quad \text{(Q2.1)}$$

$$\log\left\{\frac{P(T_i=2|\boldsymbol{X})}{P(T_i=0|\boldsymbol{X})}\right\} = \alpha_2 + \sum_{j=1}^{3}\beta_{2j}X_{ij} \quad \text{(Q2.2)}$$

where we wish to consider the effects of patient status (X_{i1}), medical condition (X_{i2}), and treatment cost (X_{i3}) on therapeutic strategy T_i where $T_i = 1, 2, 3$ corresponds to treatment A, B, and C, respectively. Here, i is the index for subject and j is the index for parameter.

Distribution: The distribution of T is $multinomial(n = 440, \boldsymbol{\pi})$ where $\boldsymbol{\pi}$ is a vector containing the probabilities of success for each treatment type where $\boldsymbol{\pi} = \big(P(T=0), P(T=1), P(T=2)\big)$ such that $\sum_{i=1}^{3}\pi_i = 1$.

Fitting the Model: Fit using a polytomous logistic regression where treatment categories are assumed to be nominal and unordered with more than two levels. The link function used is the generalized logit (glogit).

All reference categories are chosen as the lowest numeric level (e.g., the reference category for therapeutic strategy is 0). All observations are assumed independent.

Table Q2.2: Parameter estimates and standard errors for Models Q2.1 and Q2.2.

Parameter	*Level*	*Therapeutic Strategy*	*Estimate*	*Standard Error*
Intercept	-	1	0.052	0.354
Intercept	-	2	-0.459	0.405
Status	2	1	1.865	0.314
Status	2	2	1.853	0.376
Condition	2	1	-0.268	0.380
Condition	2	2	-1.404	0.441
Condition	3	1	-0.507	0.369
Condition	3	2	-1.715	0.429
Cost	2	1	1.310	0.328
Cost	2	2	1.024	0.439
Cost	3	1	1.036	0.357
Cost	3	2	1.866	0.416

Table Q2.3: Odds ratio estimates and corresponding 95% Wald confidence intervals generated using estimates from Models Q2.1 and Q2.2.

Effect	*Therapeutic Strategy*	*OR*	*95% Wald Confidence Limits*	
Status 2 vs 1	1	6.46	3.49	11.95
Status 2 vs 1	2	6.38	3.05	13.34
Condition 2 vs 1	1	0.77	0.36	1.61
Condition 2 vs 1	2	0.25	0.10	0.58
Condition 3 vs 1	1	0.60	0.29	1.24
Condition 3 vs 1	2	0.18	0.08	0.42
Cost 2 vs 1	1	3.71	1.95	7.04
Cost 2 vs 1	2	2.79	1.18	6.58
Cost 3 vs 1	1	2.82	1.40	5.67
Cost 3 vs 1	2	6.46	2.86	14.60

Brief Interpretation of select parameter estimates:
The relative log odds of receiving therapeutic strategy B versus A will increase by 1.87 if a patient shifts from stable to unstable. The remaining Status parameter is interpreted similarly.

The relative log odds of receiving therapeutic strategy B versus A will decrease by 0.27 if the official diagnosis changes from cardiovascular disease to liver disease. The remaining Condition parameters are interpreted similarly.

The relative log odds of receiving therapeutic strategy B versus A will increase by 1.31 if the monthly cost of the treatment increases from <$1k to $1k-3k. The remaining Cost parameters are interpreted similarly.

Brief Interpretation of select ORs:
The odds ratio for patients who switch from stable to unstable is 6.46 (95% CI: 3.49 to 11.95) for receiving therapeutic strategy B versus A. In other words, patients who are deemed unstable by the diagnosing physician are more likely to be recommended to therapy B versus A. The remaining OR for status is interpreted similarly.

The odds ratio for patients whose official diagnosis changes from cardiovascular disease to liver disease is 0.77 (95% CI: 0.36 to 1.61) for receiving therapeutic strategy B versus A. In other words, patients who suffer from cardiovascular disease are less likely to be recommended to therapy B than A. The remaining ORs for Condition are interpreted similarly.

The odds ratio for patients whose therapy cost increases from <$1k to $1k-3k is 3.71 (95% CI: 1.95 to 7.04) for receiving therapeutic strategy B versus A. In other words, therapy B is likely more costly than therapy A. The remaining ORs for Cost are interpreted similarly.

Problem 2C

Mixed-Effects Polytomous Logistic Regression Model:

$$\log\left\{\frac{P(T_i=1|\boldsymbol{X})}{P(T_i=0|\boldsymbol{X})}\right\} = \alpha_1 + \sum_{j=1}^{3}\beta_{1j}X_{ij} + b_{i1} \quad \text{(Q2.3)}$$

$$\log\left\{\frac{P(T_i=2|\boldsymbol{X})}{P(T_i=0|\boldsymbol{X})}\right\} = \alpha_2 + \sum_{j=1}^{3}\beta_{2j}X_{ij} + b_{i2} \quad \text{(Q2.4)}$$

where we wish to consider the effects of status (X_{i1}), condition (X_{i2}), and cost (X_{i3}) on therapeutic strategy T_i where $T_i = 1, 2, 3$ corresponds to treatment A, B, and C, respectively. Here, i is the index for subject and j is the index for parameter.

The subject-specific random intercepts are such that $b_{i1} \sim N(0, \sigma_{b1}^2)$ and $b_{i2} \sim N(0, \sigma_{b2}^2)$, where σ_{b1}^2 and σ_{b2}^2 are unknown variances.

Distribution: The distribution of T is $multinomial(n = 440, \boldsymbol{\pi})$ where $\boldsymbol{\pi}$ is a vector containing the probabilities of success for each treatment type where $\boldsymbol{\pi} = \big(P(T=0), P(T=1), P(T=2)\big)$ such that $\sum_{i=1}^{3}\pi_i = 1$.

Fitting the Model: Fit using a mixed-effects polytomous logistic regression where treatment categories are assumed to be nominal and unordered with more than two levels. The link function used is still the generalized logit (glogit).

Observations from the same physician were **not** assumed to be independent. Thus, a random intercept model was fit assuming a compound symmetric covariance structure. Since there are only two repeated observations per subject, the simplest covariance structure is suitable.

Model parameters were first estimated using Gauss-Hermite quadrature with 7 quadrature points. Note that setting the number of quadrature points to 7 requires $7^1 = 7$ (1 for a random intercept) conditional log likelihoods to be evaluated at each pass through the data. Convergence failed for 8 or more quadrature points using a relative function convergence criterion of 2.2E-6.

Due to the instability of the estimates obtained via the Gauss-Hermite quadrature, the model was refit using residual log pseudo-likelihood. A refit of the model after examining the parameter estimates requires a multiplicity adjustment on the type I error rate. We recommend that the reader keep this in mind when interpreting subsequent OR confidence intervals, as these intervals are presented as $(1-\alpha)100\%$ for $\alpha = 0.05$.

Table Q2.4: Parameter estimates and standard errors for Models Q2.3 and Q2.4.

Parameter	*Level*	*Therapeutic Strategy*	*Estimate*	*Standard Error*
Intercept	-	1	0.049	0.348
Intercept	-	2	-0.461	0.410
Status	2	1	1.854	0.311
Status	2	2	1.838	0.377
Condition	2	1	-0.238	0.375
Condition	2	2	-1.379	0.443
Condition	3	1	-0.475	0.364
Condition	3	2	-1.690	0.431
Cost	2	1	1.272	0.323
Cost	2	2	1.006	0.440
Cost	3	1	1.000	0.351
Cost	3	2	1.842	0.418

Table Q2.5: Odds ratio estimates and corresponding 95% Wald confidence intervals generated using estimates from Models Q2.3 and Q2.4.

Effect	*Therapeutic Strategy*	*OR*	*95% Wald Confidence Limits*	
Status 2 vs 1	1	6.39	3.47	11.77
Status 2 vs 1	2	6.28	3.00	13.17
Condition 2 vs 1	1	0.79	0.38	1.65
Condition 2 vs 1	2	0.25	0.11	0.60
Condition 3 vs 1	1	0.62	0.30	1.27
Condition 3 vs 1	2	0.19	0.08	0.43
Cost 2 vs 1	1	3.57	1.89	6.73
Cost 2 vs 1	2	2.74	1.15	6.49
Cost 3 vs 1	1	2.72	1.36	5.42
Cost 3 vs 1	2	6.31	2.77	14.35

Brief Interpretation of select parameter estimates:
The relative log odds of receiving therapeutic strategy B versus A will increase by 1.85 if a patient shifts from stable to unstable. The remaining Status parameter is interpreted similarly.

The relative log odds of receiving therapeutic strategy B versus A will decrease by 0.24 if the official diagnosis changes from cardiovascular disease to liver disease. The remaining Condition parameters are interpreted similarly.

The relative log odds of receiving therapeutic strategy B versus A will increase by 1.27 if the monthly cost of the treatment increases from <$1k to $1k-3k. The remaining Cost parameters are interpreted similarly.

Brief Interpretation of select ORs:
The odds ratio for patients who switch from stable to unstable is 6.39 (95% CI: 3.47 to 11.77) for receiving therapeutic strategy B versus A. In other words, patients who are deemed unstable by the diagnosing physician are more likely to be recommended to therapy B versus A. The remaining OR for Status is interpreted similarly.

The odds ratio for patients whose official diagnosis changes from cardiovascular disease to liver disease is 0.79 (95% CI: 0.38 to 1.65) for receiving therapeutic strategy B versus A. In other words, patients who suffer from cardiovascular disease are less likely to be recommended to therapy B than A. The remaining ORs for Condition are interpreted similarly.

The odds ratio for patients whose therapy cost increases from <$1k to $1k-3k is 3.57 (95% CI: 1.89 to 6.73) for receiving therapeutic strategy B versus A. In other words, therapy B is likely more costly than therapy A. The remaining ORs for Cost are interpreted similarly.

Problem 2D

Compare: The estimates for the fixed effects and corresponding standard errors changed very little between the two models. For example, Status in the fixed-effects model Q2.1 was estimated at 1.865 (SE: 0.314) and Status in the mixed-effects model Q2.3 was estimated at 1.854 (SE: 0.311). Similarly, Cost in the fixed-effects model Q2.1 was estimated at 1.310 (SE: 0.328) and Cost in the mixed-effects model Q2.3 was estimated at 1.272 (SE: 0.323).

Contrast: The fixed-effects model converged smoothly and in very few (5 total) iterations using maximum likelihood. However, as previously addressed, the mixed-effects model had difficulty with convergence when using maximum likelihood with Gauss-Hermite quadrature for the likelihood approximation. A maximum of 7 quadrature points could be used before the model failed to converge (43 total iterations before failure) within a specified (reasonable) tolerance. Thus, the model was refit using residual log pseudo-likelihood.

Differing Interpretation?

The fixed-effects model does not take into account potential correlation among repeated observations from the same physician. Failure to control for this intra-physician correlation can result in misleadingly small estimates for the standard error. Thus, the interpretations that result from a fixed-effects model are different from those that result from a mixed-effects model, since the former does not account for this correlation while the latter does.

Recommendation:

It is highly probable that physicians are personally more inclined to select one treatment over another. There can be several reasons for this, including kickback (renumeration for prescribing a certain treatment), success rates (the physician has witnessed one treatment performing better than others), or familiarity (one therapy is the typical “go-to” for that physician). In cases like these, I would typically err on the safe side and account for this possible correlation with a mixed-effects model.

However, in this case I recommend *against* fitting a mixed-effects model for a couple of reasons.

i. The estimated G matrix was not positive definite. This can imply one of two things:
   a. After controlling for the fixed effects, there is not enough variation remaining in the response to attribute it to the random intercept (Kiernen, Tao, and Gibbs, 2012, p. 10).

b. The model is mis-specified and either the covariance structure needs to be simplified or we should consider a marginal model that does not contain a random intercept. (Kiernen, 2018, p. 18).

ii. There were obvious convergence problems, so the parameter estimates and corresponding standard errors may not be accurate.

The random intercept overly complicates a model that does not require it. We are already using a compound symmetry covariance structure, which is the simplest covariance structure available (outside of independence, which would just reduce the fit to a glm) and only requires estimation of two parameters. We are also not estimating any random slopes - only random intercepts. Thus, there is really no way to further simplify the model unless we eliminate one or more of the fixed effects or drop the random intercept. Thus, I recommend that the investigator used the fixed-effect model (Models Q2.1 and Q2.2).

Problem 2E

In this study, physician preference for a therapeutic strategy was assessed as it related to patient status (stable vs unstable), medical condition (cardiovascular, liver, or renal disease) and treatment cost (<$1k, $1-3k, >$3k). Two hundred and twenty physicians were enrolled in the study, and each physician provided prescriptions to two patients.

The stability of the patient's medical condition was an important factor ($p<0.0001$, F-test) to physicians in the appropriate choice of therapy. When a patient's medical status shifted from stable to unstable, the physician was more likely to be recommend therapy B versus A (OR: 6.46, 95% CI: 3.49 to 11.95) or therapy C versus A (OR: 6.38, 95% CI: 3.05 to 13.34). This *may* suggest that therapies B and C are more aggressive than therapy A.

The official medical diagnosis was also an important factor ($p=0.0001$, F-test) to physicians in the appropriate choice of therapy. If the diagnosis was changed from cardiovascular to liver disease, the physician was less likely to recommend therapy B versus A (OR: 0.77, 95% CI: 0.36 to 1.61) or therapy C versus A (OR: 0.25, 95% CI: 0.10 to 0.58). If the diagnosis was changed from cardiovascular to renal disease, the physician was less likely to recommend therapy B versus A (OR: 0.60, 95% CI: 0.29 to 1.24) or therapy C versus A (OR: 0.18, 95% CI: 0.08 to 0.42).

Finally, treatment cost was also an important factor ($p<0.0001$, F-test) to physicians in the appropriate choice of therapy. If the monthly cost of therapy increased from <$1k to $1k-3k, the physician was more likely to recommend therapy B versus A (OR: 3.71, 95% CI: 1.95 to 7.04) or therapy C versus A (OR: 2.79, 95% CI: 1.18 to 6.58). If the monthly cost of therapy increased from <$1K to >$3k, the physician was more likely to recommend therapy B versus A (OR: 2.82, 95% CI: 1.40 to 5.67) or therapy C versus A (OR: 6.46, 95% CI: 2.86 to 14.60). This *may* suggest that therapies B and C are more expensive than therapy A.

Problem 2F
There are three main possible missing data mechanisms that we have studied:

Let $R_{ij} = \begin{cases} 1, & Y_{ij}\ observed \\ 0, & Y_{ij}\ not\ observed \end{cases}$

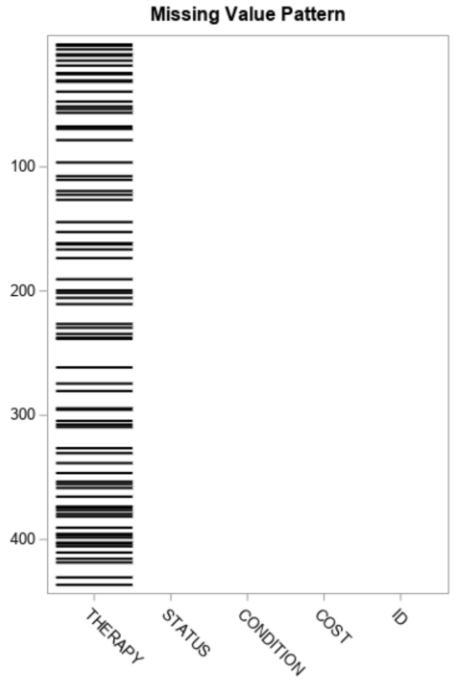
(i) Missing Completely at Random (MCAR): $P(R_{ij}|Y_i, X_{ij}) = P(R_{ij}|X_{ij})$
(ii) Missing at Random (MAR): $P(R_{ij}|Y_i, X_{ij}) = P(R_{ij}|Y_i^{obs}, X_{ij})$
(iii) Missing Not at Random (MNAR): Otherwise

As seen in Figure Q2.1, it appears that the data is Missing at Random (MAR). Almost 88% of the missingness is related to patient status of unstable. I suspect that it is more difficult to make a decision to prescribe a patient one therapy over another if that patient is experiencing disease progression, and, thus, the physician would be more inclined to leave that variable blank.

I would not recommend discarding the missing cases since there is a considerable amount of missingness (~17%). A 2016 study (Karangwa *et al.*, 2016) assessed the performance of two popular imputation methods, multivariate normal imputation (MVNI) and multiple imputation by chained equations (MICE), and found that MVNI produced more accurate estimates for unordered, categorical data missing at random. Thus, I would impute the missing values using MVNI and then I would fit the selected, fixed effects model described in d) using the imputed data.

Fig Q2.1 Distribution of missingness in logit.csv. A) Y-axis represents row in dataset (n=440). Modified code (Wicklin, 2016). **B)** Frequency of missing outcome (therapy) for each predictor.

A)



B)

	Therapy Missing N (%)
Patient Status	
Stable	9 (12.2%)
Unstable	65 (87.8%)
Medical Condition	
Cardiovascular Disease	22 (29.7%)
Liver Disease	33 (44.6%)
Kidney Disease	19 (25.7%)
Treatment Cost	
< $1k	31 (41.9%)
$1k-$3k	25 (33.8%)
> $3k	18 (24.3%)

Problem 3A

Table Q3.1: Treatment group composition by pathologic complete response (pCR) rate.

	Treatment Arm			
	Antibody Only	*Kinase Inhibitor*	*Combination Therapy*	*FDR-Adj. p-value**
Trial 1				
N	33	23	31	
pCR rate, 95% CI	75.8 (57.7, 88.9)	43.5 (23.2, 65.5)	93.6 (78.6, 99.2)	0.002
Trial 2				
N	9	15	8	
pCR rate, 95% CI	55.6 (21.2, 86.3)	66.7 (38.4, 88.2)	75.0 (34.9, 96.8)	0.752
Trial 3				
N	45	45	40	
pCR rate, 95% CI	64.4 (48.8, 78.1)	44.4 (29.6, 60.0)	67.5 (50.9, 81.4)	0.086
Trial 4				
N	146	176	175	
pCR rate, 95% CI	52.1 (43.6, 60.4)	57.4 (49.7, 64.8)	78.9 (72.1, 84.7)	<.0001
Trial 5				
N	122	95	124	
pCR rate, 95% CI	48.4 (39.2, 57.6)	43.2 (33.0, 53.7)	91.1 (84.7, 95.5)	<.0001
Trial 6				
N	43	53	53	
pCR rate, 95% CI	44.2 (29.1, 60.1)	69.8 (55.7, 81.7)	62.3 (47.9, 75.2)	0.069
Trial 7				
N	114	102	115	
pCR rate, 95% CI	69.3 (60.0, 77.6)	51.0 (40.9, 61.0)	78.3 (69.6, 85.4)	0.0004
Trial 8				
N	39	33	55	
pCR rate, 95% CI	15.4 (5.9, 30.5)	12.1 (3.4, 28.2)	18.9 (9.1, 30.9)	0.752
Trial 9				
N	50	64	81	
pCR rate, 95% CI	70.0 (55.4, 82.1)	75.0 (62.6, 85.0)	86.4 (77.0, 93.0)	0.086

Note: Percentages may not sum to exactly 100 due to rounding error. There were no missing values in this dataset. Rates are expressed as percentages.

*P-values generated using a Wald Chi-square test on 2 degrees of freedom and corrected for false discovery rate (FDR).

Summary of individual effectiveness of therapies across trials:

Without adjusting for genes possibly correlated with HER2 status, combination therapy increased the pCR rate substantially. In 8 of the 9 trials, combination therapy demonstrated the best efficacy in treating HER2-positive breast cancer, with response measured by absence of residual invasive (and/or *in situ*) cancer. Kinase inhibitor and antibody alone appear to compete in terms of efficacy across the trials, with antibody alone having a higher pCR rate in 5 of the 9 trials.

Summary of the relative effectiveness of the combination therapy compared to the antibody only arm across trials.

Note: To reduce alpha spending, between-arm testing was only performed for trials with FDR-adjusted p-value < 0.05. Recall that the Wald chi-square test looks for a difference between **any** of the arms. So, if p-adj > 0.05 for a test on 2 df, then no further testing is required.

As seen in the FDR-adjust p-values in Table Q3.1, there was no evidence of differing efficacy (measured by pCR rate) between combination therapy and antibody alone in four of the trials, namely Trials 2 3, 6, 8, and 9. When the remaining 4 trials were examined, while there was no evidence of a difference in efficacy between combination therapy and antibody alone in Trials 1 and 8 (FDR-adjusted p-values: 0.09 and 0.12, respectively), there was substantial evidence of a difference in efficacy in Trials 4 and 5 (FDR-adjusted p-values: <.0001 and <.0001, respectively), with combination therapy outperforming antibody alone in both trials.

Problem 3B

Table Q3.2: Results from a Chi-square test (on 2 degrees of freedom) in each trial comparing pCR between arms adjusting for gene 1.

	Trial #								
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
Wald Chi-Square	15.39	0.96	5.38	27.55	52.07	6.29	18.74	0.48	6.11
p-value	0.0005	0.62	0.07	<.0001	<.0001	0.04	<.0001	0.79	0.05

Table Q3.3: Estimated effect of gene 1 and adjusted effect for arm measured by odds ratio (OR).

	Kinase Inhibitor vs Antibody Alone	*Combination vs Antibody Alone*	*Gene 1 Expression*
Trial 1			
OR pCR, 95% CI	0.14 (0.04, 0.56)	3.63 (0.65, 20.35)	2.78 (1.39, 5.56)
Trial 2			
OR pCR, 95% CI	1.90 (0.32, 11.23)	2.78 (0.33, 23.24)	1.57 (0.73, 3.37)
Trial 3			
OR pCR, 95% CI	0.45 (0.19, 1.06)	1.17 (0.47, 2.90)	1.09 (0.78, 1.52)
Trial 4			
OR pCR, 95% CI	1.24 (0.80, 1.94)	3.44 (2.11, 5.59)	1.01 (0.84, 1.23)
Trial 5			
OR pCR, 95% CI	0.82 (0.48, 1.41)	10.89 (5.33, 22.24)	1.14 (0.91, 1.42)
Trial 6			
OR pCR, 95% CI	2.89 (1.24, 6.71)	2.03 (0.89, 4.62)	1.18 (0.84, 1.67)
Trial 7			
OR pCR, 95% CI	0.45 (0.25, 0.79)	1.63 (0.89, 2.98)	1.43 (1.11, 1.85)
Trial 8			
OR pCR, 95% CI	0.72 (0.18, 2.90)	1.13 (0.36, 3.55)	1.98 (1.17, 3.37)
Trial 9			
OR pCR, 95% CI	1.31 (0.57, 3.04)	2.96 (1.21, 7.22)	1.43 (1.00, 2.04)

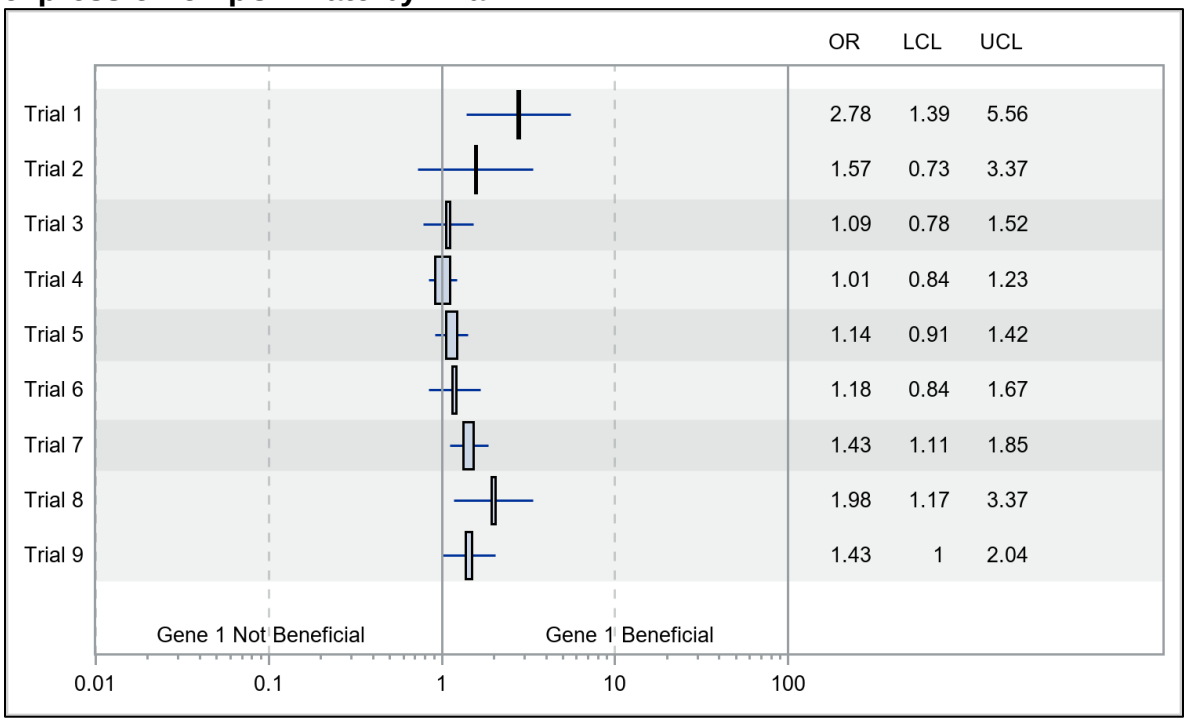
The estimated effect of gene 1 expression is not comparable across studies because each study enrolled a different sample size, and these sample sizes differ across arms. A specific genotype may contribute to the variability in drug response, and so the effects would only be comparable if the sample sizes were identical in each treatment arm (e.g., N combination therapy in Trial 1 equals N combination therapy in Trial 2). This phenomenon is displayed in Figure Q3.1, where trials with larger weight have tighter confidence intervals and thus **should** contribute more to the final analysis.

While it is apparent that increased expression of gene 1 is beneficial for pathologic complete response, conclusions regarding this relationship would be biased towards small trials. Most of the presented trials are relatively balanced across treatment arms, so one potential solution would be to perform the variance method for dichotomous outcomes (DiMaggio, 2011):

1. Compute $\log(OR_i)$ where i indexes trials 1-9.
2. Compute the weights as the inverse variance $\hat{w}_i = 1/\widehat{se}_i^2$ where $\widehat{se}_i^2$ is the variance estimate for $\log(OR_i)$.

3. Compute the weighted mean OR by using $\overline{WOR} = \frac{\sum_{i=1}^{9} \hat{w}_i log(OR_i)}{\sum_{i=1}^{9} \hat{w}_i}$ and exponentiate the final result.
4. Calculate the estimated standard error of the weighted mean OR as $\widehat{se}(\overline{WOR}) = (\sum_{i=1}^{9} \hat{w}_i)^{-1/2}$.
5. Compute a 95%CI for the weighted mean OR using $\overline{WOR} \pm 1.96\big(\widehat{se}(\overline{ES})\big)$

Figure Q3.1 Unadjusted odds ratios (OR) and 95% CIs for impact of Gene 1 expression on pCR Rate by Trial.



The area of each rectangle is proportional to the total sample size of all combined trials.
Code for forest plot modified from (Mantage, 2012).

Problem 3C

Table Q3.4: Estimated adjusted odds ratio (OR) for AK vs A.

	OR AK vs A, 95% CI
Model	
Fixed-Effects	2.50 (1.97, 3.17)
Mixed-Effects	2.67 (1.54, 4.63)

Fixed-Effects model

Interpret OR AK vs A: The adjusted odds ratio for participants who receive the combination therapy versus antibody alone is 2.50 (95% CI: 1.97 to 3.17) for responding to therapy measured by pCR. In other words, adjusting for gene 1 expression, participants who receive the combination therapy are more likely to

respond positively to treatment than those who receive antibody alone.

Potential concerns: These results should be interpreted with caution because the fixed-effects model does not account for the potential within-trial correlation across trials. For example, pathologic complete responses due to combination therapy are likely to be more similar among patients in one trial versus patients in another. This can be due to differences in access to healthcare, site location, clinical staff, patient population, and/or drug administration route.

Mixed-effects model

Compare to previous: Parameter estimates for the mixed-effects model, which includes a trial-level random intercept and trial-level random slope for arm, were obtained using Gauss-Hermite quadrature with 4 quadrature points (> 4 was too computationally intensive) and a Cholesky root parameterization for the blocks in G.

We know that pCR results within a trial may be correlated, so we want to control for this potential correlation by using a mixed effects model. By specifying a random intercept, we allow the effect of trial on pCR to differ across trials. By specifying a random slope, we allow the effect of treatment arm to vary across trials. With this setup, we can estimate any needed variance components and make valid inferences regarding differences in treatment effect for the pooled data.

Interpret Coefficients: The relative log odds of pCR will decrease by 0.13 if the patient receives kinase inhibitor instead of antibody alone.

Similarly, the relative log odds of pCR will increase by 0.98 if the patient receives combination therapy instead of antibody alone.

Finally, a one-unit increase in gene 1 expression will result in a 0.18 increase in the relative log odds of pCR.

Interpret OR AK vs A: The adjusted odds ratio for participants who receive the combination therapy versus antibody alone is 2.67 (95% CI: 1.54 to 4.63) for responding to therapy measured by pCR. In other words, adjusting for gene 1 expression, participants who receive the combination therapy are more likely to respond positively to treatment than those who receive antibody alone.

Compare OR estimates: Notice that the estimated OR of 2.67 in the mixed-effects model does not differ much from the estimated OR of 2.50 in the fixed effects model. However, the confidence interval for the OR estimate from the mixed-effects model is over twice as wide (width: 3.09) as the confidence interval from the fixed-effects model (width: 1.20). This is because a fixed-effects model does not take into account potential within-trial correlation. Failure to control for this correlation can result in misleadingly small estimates for the standard error.

LRT for within-trial correlation:

Hypothesis: $H_0: \tau^2 = 0$ vs. $H_1: \tau^2 > 0$ for $b_i \sim N(0, \tau^2)$
where b_i is the random effect where $i = 1, \ldots, N$ indexes the number of trials

Test Type: Likelihood Ratio Test (LRT) on 0 and 2 degrees of freedom

Interpretation of Null: Within-trial variance is 0 (i.e., perfect correlation within observations from the same trial).

Test Statistic:
$$LR = (-2\ Log\ Likelihood\ in\ Fixed\ Effects\ Model) - (-2\ Log\ Likelihood\ in\ Random\ Effects\ Model)$$
$$\approx\ 2420.21 - 2253.73 \approx 166.5$$

Distn Under H_0: As $N \to \infty$ (number trials goes to infinity), then
$$LR \underset{d}{\to} 0.5\chi_0^2 + 0.5\chi_2^2$$
for χ_0^2 a point mass at 0 and χ_2^2 the chi-square distribution on 2 df. This mixture results because H_0 is on the boundary of the parameter space.

P-value: Average the two p-values generated from the 50:50 mixture:
$$p = \frac{\chi_0^2(\text{stat}=166.5)+\chi_2^2(\text{stat}=166.5)}{2} \approx \frac{0+6.87\times10^{-125}}{2} \approx 3.5\ \times 10^{-37}$$

Decision: Assume $\alpha = 0.05$. Reject H_0 since $p = 3.43\ \times 10^{-125} < 0.05$

Conclusion: There is evidence of correlation between observations within the same trial.

Problem 3D

Table Q3.5: Test Gene 1's relation to pCR using the random effects model from c)

	*LRT Statistic**	*Raw p-value*	*FDR-adjusted p-value**	*Examine?*
Gene 1	12.98	0.0003	0.004	Yes
Gene 2	68.46	<.0001	<.0001	Yes
Gene 3	0.81	0.37	0.68	No
Gene 4	0.07	0.79	0.82	No
Gene 5	0.10	0.75	0.81	No
Gene 6	1.11	0.29	0.63	No
Gene 7	0.41	0.52	0.70	No
Gene 8	1.08	0.30	0.63	No
Gene 9	1.08	0.30	0.63	No
Gene 10	0.01	0.92	0.92	No
Gene 11	0.10	0.75	0.81	No
Gene 12	1.11	0.29	0.63	No
Gene 13	0.18	0.67	0.81	No
Gene 14	0.77	0.38	0.68	No
Gene 15	1.20	0.27	0.63	No
Gene 16	1.07	0.30	0.63	No
Gene 17	1.31	0.25	0.63	No
Gene 18	0.20	0.65	0.63	No
Gene 19	1.11	0.29	0.70	No
Gene 20	0.44	0.51	0.63	No
Gene 21	1.06	0.30	0.71	No
Gene 22	0.35	0.55	0.70	No
Gene 23	0.49	0.48	0.63	No
Gene 24	1.64	0.20	0.71	No
Gene 25	0.44	0.51	0.70	No
Gene 26	0.63	0.43	0.63	No
Gene 27	0.13	0.72	0.70	No
Gene 28	1.90	0.17	0.70	No
Gene 29	3.46	0.06	0.81	No
Gene 30	0.00	1.00	0.63	No

*LRT statistic calculated by taking -2LogL(reduced model) – (-2LogL(full model)). The reduced model is the model from part d) without a fixed effect for gene, and the full model is the model from part d) with a fixed effect for gene.

**P-value calculated on 1 df and corrected for false discovery rate (FDR).

Based on the phrasing of this question, I anticipated statistical significance in more than 2 of the 30 genes. Despite this, and cognizant of future analyses, I would recommend that the investigator adjust for multiple comparisons across the 30 hypothesis tests. This recommendation could involve two possible courses of action:

1. Adjust the alpha level and select the genes to further examine using this new cutoff for the p-value. For example, the new cutoff might be $\alpha = 0.001$ instead of the usual type 1 error rate of $\alpha = 0.05$.

2. Adjust the p-values for multiple comparisons and pick the genes with p-values < 0.05.

Here, I chose the second option and adjusted all p-values that resulted from the chi-square test for family-wise error rate via the false discovery rate (FDR) approach. The results from this correction (Table Q3.5) results in two genes, Genes 1 and 2, with FDR-adjusted p-values < 0.05. Thus, I would recommend that the investigators perform further testing on this subset of genes.

Problem 3E

Algorithm for obtaining an exact p-value using a permutation test:

1. For p in 1, …, P
    (1) Subset data by trial.
        (a) Further subset data by arm.
        (b) Shuffle gene_3 values within subsetted data. Call this column of shuffled values gene3_swap.
    (2) Calculate the LRT statistic between PCR and gene3_swap.
3. Calculate the LRT statistic between PCR and gene_3 using the original dataset.
4. Count how many times the permutation test p-values exceed the observed p-value and divide this result by P (the number of permutations).

The resulting p-value from the permutation test for independence was 0.55 (for P=1000 permutations), which was larger than the LRT statistic raw p-value of 0.37 for Gene 3.

While these tests do not yield different results (at $\alpha = 0.05$) in terms of the decision to fail to reject, the p-values still differ by a magnitude of 0.18.

The above algorithm is appropriate for this data because it works for nested clusters (arm within trial), does not depend on balance between the arms, and works for generalized linear models with logit link functions. If there are, in fact, distributional violations for the random effects (i.e., non-normality), then, as the consulting statistician, I would recommend the permutation test over the mixed-effects model.

Problem 4A

Note: For this problem, we will refer to study as the UPBEAT study (upbeat meaning happy and beat for heartbeat).

Table Q4.1: Select baseline characteristics in the UPBEAT cohort

	Overall (N=754)
Baseline Characteristics	
Age (yr), median (25^{th}, 75^{th})	49.0 (41.0, 56.0)
Female	493 (65.4%)
Depressive Symptoms - Yes	207 (27.5%)
Physical Activity Guidelines Met - Yes	471 (62.5%)
Healthy Eating Index Score (0-100), median (25^{th}, 75^{th})*	60.5 (51.3, 70.3)
Body Mass Index (kg/m^2), median (25^{th}, 75^{th})**	29.0 (26.1, 32.8)
Education Status	
Less than High School	271 (35.9%)
High School	209 (27.7%)
More than High School	274 (36.3%)

Note: Values expressed as N (%) or median (25^{th},75^{th} percentiles). Percentages may not sum to exactly 100 due to rounding error. There were no missing values in this study.
* Healthy eating index score ranges from 0 to 100 with higher score indicative of healthier diet quality.
** One participant had a baseline BMI of 70.3 kg/m^2. This value was replicated at subsequent visits, so this extreme value is assumed correct and not due to user entry error.

Table Q4.2: Baseline depression status and five major cardiovascular risk factors of the UPBEAT cohort across 10 years of follow-up.

	Visit 1 (Baseline) (N=754)	*Visit 2 (5 yr) (N=754)*	*Visit 3 (10 yr) (N=754)*	*p-value**
Characteristics Across Time				
Depressed at Baseline	207 (27.5%)	-	-	-
High Cholesterol	330 (43.8%)	365 (48.4%)	369 (48.9%)	0.02
High Blood Pressure	229 (30.4%)	310 (41.1%)	296 (39.3%)	<.0001
Obese (BMI > 30 kg/m^2)	318 (42.2%)	343 (45.5%)	380 (50.4%)	<.0001
Alcohol Use	337 (44.7%)	435 (57.7%)	378 (50.1%)	<.0001
Smoker	145 (19.2%)	112 (14.9%)	106 (14.1%)	<.0001

Note: Values expressed as N (%). Percentages may not sum to exactly 100 due to rounding error. There were no missing values in this study.
* P-value comparisons across time are generated using a Type III (Wald) test from fitting a generalized estimating equation for the marginal model, $logit\big(E[Y_{ij}]\big) = \gamma_1 + \gamma_2 Visit_{ij}$, where i indexes subject, j indexes visit, and the binary outcome is coded 0=no and 1=yes.

Problem 4B

Problem 4B i.

Hypothesis: $H_0: \beta_3 = 0$ vs. $H_1: \beta_3 \neq 0$

Test: Likelihood Ratio Test (LRT) on 1 degree of freedom.

Test Statistic: $LR = 2\{\ell_n(Full) - \ell_n(Reduced)\} \approx 2(7214.6 - 7220.8) \approx 12.5 \sim \chi_1^2$

P-value: $p = \chi_1^2(statistic = 12.5) \approx 0.0004$

Decision: Assume $\alpha = 0.05$. Reject H_0 since $p = 0.0004 < 0.05$

Conclusion: There is sufficient evidence to suggest that there is an association between baseline age and body mass index in participants enrolled in the UPBEAT study.

Problem 4B ii.

Hypothesis: $H_0: \beta_2 = 0$ vs. $H_1: \beta_2 \neq 0$

Test: Likelihood Ratio Test (LRT) on 1 degree of freedom.

Test Statistic: $LR = 2\{\ell_n(Full) - \ell_n(Reduced)\} \approx 2(7214.6 - 7218.3) \approx 7.4 \sim \chi_1^2$

P-value: $p = \chi_1^2(statistic = 7.4) \approx 0.007$

Decision: Assume $\alpha = 0.05$. Reject H_0 since $p = 0.007 < 0.05$

Conclusion: There is sufficient evidence to suggest that, adjusting for age at baseline, there is an association between changes in age and changes in body mass index over time in participants enrolled in the UPBEAT study.

Problem 4B iii.

Table Q4.3: Table of parameter estimates for parameters tested in parts i. and ii.

Parameter	*Estimate*	*Standard Error*
$\boldsymbol{\beta_2}$	0.075	0.028
$\boldsymbol{\beta_3}$	-0.104	0.029

Problem 4B iv.

Hypothesis: $H_0: \beta_2 - \beta_3 = 0$ vs. $H_1: \beta_2 - \beta_3 \neq 0$

Test: F-test with numerator df = 1 and denominator df = 2259.

Test Statistic: $F = \frac{SS/df_{num}}{SS/df_{den}} = \frac{352.36/1}{78047.32/2259} \approx 10.2 \sim F(1, 2259)$

P-value: $p = F(1, 2259, statistic = 10.2) \approx 0.001$

Decision: Assume $\alpha = 0.05$. Reject H_0 since $p = 0.001 < 0.05$

Conclusion: There is sufficient evidence to suggest that the parameter on age (across all time points) and the parameter on baseline age are not equal in participants enrolled in the UPBEAT study.

Problem 4C i.

Model: $logit(E[Y_{i1}]) = \beta_0 + \beta_1 Depr1_{i1} + \beta_2 Age_{i1} + \beta_3 Educ_{i1} + \beta_4 Female_{i1} + \beta_5 hei2010_{i1} + \beta_6 pag2008_{i1}$

where Y_{i1} is the CVD risk factor of interest at baseline.

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

Test: Wald test on 1 degree of freedom.

Distribution of Test Statistics: $W_n \sim \chi_1^2$

Table Q4.4: Table of unadjusted Wald test statistics and Holm-Bonferroni adjusted p-values for models requested in part c) i.

Risk Factor	*Unadj. Test Statistic on Depression**	*Adj. p-value***
High Cholesterol	1.89	0.55
High Blood Pressure	2.22	0.55
Obese (BMI > 30 kg/m^2)	0.11	1.00
Alcohol Use	0.43	1.00
Smoker	18.4	<.0001

*Raw (unadjusted for multiple comparisons) Wald test statistics on baseline depressive symptoms.
**P-values adjusted for multiple comparisons using Holm-Bonferroni.

Decision: Assume $\alpha = 0.05$. Fail to reject H_0 for first four rows of table Q4.4 since $p > 0.05$. Reject H_0 for final row (smoker) since $p = 0.0001 < 0.05$.

Conclusion: After adjusting for the baseline effects of age, education status, gender, healthy eating index score, and physical activity status, there is insufficient evidence to suggest that those who display depression symptoms (at baseline) are at increased risk of cardiovascular disease due to high cholesterol, high blood pressure, obesity, or alcohol consumption (at baseline).

However, there is sufficient evidence to suggest that those who display depression symptoms (at baseline) are at increased risk of cardiovascular disease due to smoking (at baseline).

Problem 4C ii.

Model:	$logit\big(E\big[Y_{ij}\big]\big) = \beta_0 + \beta_1 Depr1_{ij} + \beta_2 Age_{ij} + \beta_3 Educ_{ij} + \beta_4 Female_{ij} + \beta_5 hei2010_{ij} + \beta_6 pag2008_{ij}$ where $(Y_{i1}, \dots, Y_{i5})$ is a binary covariate representing the CVD risk factor of high cholesterol, high blood pressure, obesity, alcohol use, and smoking at baseline for the i^{th} subject.
Odds Ratio (95% CI):	$OR \approx \exp\big(\hat{\beta}_1\big) \approx \exp\,(0.2033) \approx 1.23$ $95\%\ CI(OR) \approx (\exp(0.0489)\,, \exp(0.3576)) \approx (1.05, 1.43)$
Hypothesis:	$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$
Test:	Wald test on 1 degree of freedom.
Test Statistic:	$W_n \approx 6.66 \sim \chi_1^2$
P-value:	$p = \chi_1^2(statistic = 6.66) \approx 0.01$
Decision:	Assume $\alpha = 0.05$. Reject H_0 since $p = 0.01 < 0.05$
Conclusion (Odds ratio):	Participants with depressive symptoms at baseline have 1.23 times the odds (95% CI: 1.05 to 1.43) of increased risk of cardiovascular disease, measured by the 5 risk factors.
Conclusion (Wald test):	After adjusting for the baseline effects of age, education status, gender, healthy eating index score, and physical activity status, there is sufficient evidence to suggest that those who display depressive symptoms (at baseline) are at increased risk of cardiovascular disease measured by the 5 risk factors measured at baseline.

Problem 4C iii.

Have

$\boldsymbol{p} = \big(P(T_i = 0), P(T_i = 1), P(T_i = 2), P(T_i = 3), P(T_i = 4), P(T_i = 5)\big) = (0.122, 0.298, 0.320, 0.180, 0.074, 0.005)$

where $\boldsymbol{p}$ represents the baseline probabilities indicated above.

Then,

$$E(T) = \sum_{i=1}^{5} T_i * P(T_i = t_i) = 0(0.122) + 1(0.298) + 2(0.320) + 3(0.180) + 4(0.074) + 5(0.005) \approx 1.8$$

$$Var(T) = E[T_i^2] - E[T_i]^2$$
$$= \sum_{i=1}^{5} T_i^2 * P(T_i = t_i) - 1.8^2 = 0^2(0.122) + 1^2(0.298) + 2^2(0.320)$$
$$+ 3^2(0.180) + 4^2(0.074) + 5^2(0.005) - 1.8^2 \approx 1.27$$

However, if $Z \sim Binom(5, p)$ with $p = \frac{E(T)}{n} \approx \frac{1.8}{5} = 0.36$

Then, $Var(Z) = 5p(1-p) \approx 5(0.36)(1-0.36) \approx 1.15$

Since $Var(T) > Var(Z)$, we expect T_i to have extra-binomial variation.

Problem 4D i.

As noted in Chapter 21 (pg 622) of Fitzmaurice, if we were to specify the entire joint distribution for the Y_{ijk}, we could use maximum likelihood estimation. However, a GEE is easier to implement and understand.

We consider a model of the form (similar to pg 624 of Fitzmaurice)

$$logit(\mu_{ijk}) = \beta_0 + \beta_1 Depr1_{ijk} + \beta_2 Age\_shift_{ijk} + \beta_3 Female_{ijk}$$
$$+\beta_4 obese\_flag_{ijk} + \beta_5 alcuse\_flag_{ijk} + \beta_6 smoker\_flag_{ijk}$$
$$+\beta_7 age\_shift_{ijk} * obese\_flag_{ijk}$$
$$+\beta_8 age\_shift_{ijk} * alcuse\_flag_{ijk}$$
$$+\beta_9 age\_shift_{ijk} * smoker\_flag_{ijk} \quad \text{(Q4.1)}$$

where $(Y_{ij1}, \dots, Y_{ij3})$ is a binary covariate representing the CVD risk factor of obesity ($k = 1$), alcohol use ($k = 2$), and smoking ($k = 3$) for the i^{th} subject at the j^{th} visit. Age_Shift_{ijk} is calculated as $(Age_{ijk} - 50)/5$. Each of the flag variables is 1 if the outcome is associated with that flag and 0 otherwise. For example, if the outcome is 0 or 1 corresponding for the obesity risk factor then $obese\_flag_{ijk} = 1$ and all other flags are 0. Notice that interactions are not needed for depression status or sex because these are binary variables (and including an interaction would only add to the existing intercept, which is redundant since a flag is present).

Interpretation of any associated results is more understandable if we write three separate models for each risk factor as:

$$logit(\mu_{ij1}) = (\beta_0 + \beta_4) + \beta_1 Depr1_{ijk} + (\beta_2 + \beta_7) Age\_shift_{ijk} + \beta_3 Female_{ijk} \quad \text{(Q4.2)}$$

$$logit(\mu_{ij2}) = (\beta_0 + \beta_5) + \beta_1 Depr1_{ijk} + (\beta_2 + \beta_8) Age\_shift_{ijk} + \beta_3 Female_{ijk} \quad \text{(Q4.3)}$$

$$logit(\mu_{ij3}) = \beta_0 + \beta_1 Depr1_{ijk} + \beta_2 Age\_shift_{ijk} + \beta_3 Female_{ijk} \quad \text{(Q4.4)}$$

Notice that, for estimability, $\beta_6 = 0$ and $\beta_9 = 0$ (thus, simplifying the final equation).

The above three models were NOT fit; they are just listed to improve understanding. Fitting the models separately does not allow us to quantify the differences between the risk factors due to their potential correlation. As addressed on pg 625 of Fitzmaurice, the estimated coefficients for model (Q4.1), obtained using GEE with an unstructured (non-zero) working correlation, are expected to be different than those obtained with separate regressions due to the non-linearity of the logistic regression. These parameter estimates are described in Table Q4.5.

Table Q4.5: Table of parameter estimates, standard errors, and 95% confidence intervals for parameters in Model Q4.1.

Parameter	*Estimate*	*SE*	*95% CI Upper*	*95% CI Lower*
Intercept	-1.470	0.079	-1.625	-1.315
depr1	0.196	0.081	0.037	0.355
age_shift	-0.092	0.024	-0.139	-0.045
female	-0.312	0.072	-0.453	-0.171
obese_flag	1.366	0.108	1.167	1.565
alcuse_flag	1.660	0.083	1.497	1.823
smoker_flag	0	0	0	0
age_shift * obese_flag	0.189	0.031	0.128	0.251
age_shift * alcuse_flag	0.076	0.029	0.018	0.133
age_shift * smoker_flag	0	0	0	0

Problem 4D ii.

Sex and Alcohol Use:

For a participant of average age (~ 50 years of age, so $Age\_shift = 0$) who is not depressed at baseline, being female reduces the odds of alcohol use by a factor of $\exp(\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_5) \approx 0.89$. Thus, the odds of alcohol use are reduced by 11% in females, which makes sense since females tend to drink less than males in most cultures. However, if the participant is a female (of average age) but also depressed at baseline, then the odds of alcohol use increase by a factor of $\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_5) \approx 1.08$. Thus, the odds of alcohol use are increased by 8% in females who are depressed. If the participant is both male (of average age) and depressed, his odds of alcohol use are increased by a factor of $\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_5) \approx 1.47$ or 47%!

Age and Alcohol Use:

Note to self: $age = 5(age\_shift) + 50$ so every 1 unit increase in age_shift increases age by 5 years.

Every five-years of age increases the odds of alcohol use by a factor of $\exp(0.076) \approx 1.08)$. Thus, a participant's odds of alcohol use are increased by 8% for every 5 years of age.

Notice that the interpretation of sex on alcohol use requires selecting certain values for the other covariates. However, it is easier in the interpretation of age on alcohol use because we can just examine the interaction.

Problem 4D iii.

To test that $\beta_1^{obsese} = \beta_1^{alcuse} = \beta_1^{smoker} = 0$ requires using select entries from $Cov(\hat{\beta})$ (empirical) from the GEE output.

This is easily done with a Wald Chi-Square test. The null is equivalent to fitting Model (Q4.1) without the term $Depr1_{ijk}$, as this forces all coefficients on depression status, as seen in the separate models Q4.2-Q4.4, to zero.

Hypothesis:	$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$
Test:	Wald Chi-square test on 1 degree of freedom.
Test Statistic:	$W_n \approx 5.83 \sim \chi_1^2$
P-value:	$p = \chi_1^2(statistic = 5.83) \approx 0.02$
Decision:	Assume $\alpha = 0.05$. Reject H_0 since $p = 0.02 < 0.05$
Conclusion:	After accounting for potential correlation between risk factors and adjusting for age and gender, there is sufficient evidence to suggest that baseline depression status affects the three risk factors (obesity, alcohol use, and smoking) differently.

Problem 4E

The primary aim of the UPBEAT study was to examine the longitudinal relationship between baseline depression and five major cardiovascular risk factors: high cholesterol, high blood pressure, obesity, alcohol use, and smoking. A total of 28% (207 of 743) participants displayed depressive symptoms at baseline. These participants had 1.23 times the odds (95% CI: 1.05 to 1.43) of increased risk of CVD, measured by the 5 major risk factors.

However, it was also of interest to examine the longitudinal association between baseline depression and increased cardiovascular risk. Since obesity is frequently associated with high blood pressure and high cholesterol, analyses were focused on 3 of the 5 risk factors (obesity, alcohol use, and smoking).

Males who were depressed at baseline were the most likely to drink, and females who were not depressed at baseline were the least likely to drink. Thus, the UPBEAT study results suggest that the average male of age 50 who suffers from depression is at increased risk for cardiovascular disease due to alcohol consumption. For every 5 years a participant ages, his/her odds of alcohol use increase by 8%. Thus, the older a participant, the more likely s/he is for increased risk of CVD due to alcohol consumption.

Finally, there is evidence to suggest that baseline depression status affects the three risk factors differently (p=0.02, Wald test) when assessed longitudinally. Thus, this must be considered when planning interventions to target specific behaviors that are known to increase risk of cardiovascular disease.