STAT 2132
Orly Olbum

How do you tell the difference between a male and female chromosome? You pull down their genes!

FINAL EXAM

4/29/21

$Y \in \mathbb{R}^{p \times n}$

$g = 1, \dots, p$ genes

$i = 1, \dots, n$ samples

$r(i) \in \{1, \dots, m = 46\}$ individuals who generate sample i

$Y_{g*} \in \mathbb{R}^n$, $Y_{*i} \in \mathbb{R}^p$, $Y_{gi} \in \mathbb{R}$

1. dose: factor of 5 ordered levels ($1=0$, $5=5$)

$$Y_{gi} = \sum_{d=1}^{5} 1\{\text{sample } i \text{ was given dose } d\}\, \beta_{gd} + \delta_{gr(i)} + e_{gi}$$
$$= x_i^T \beta_g + \delta_{gr(i)} + e_{gi}, \quad i = 1, \dots, n \qquad (1)$$

$\beta_g = (\beta_{g1}, \dots, \beta_{g5})^T \in \mathbb{R}^5$

$\delta_{gm} \overset{iid}{\sim} N(0, \tau_g^2)$

$e_{g1}, \dots, e_{gn} \overset{iid}{\sim} N(0, \sigma_g^2)$

$\Big\}$ $e_{gi} \perp\!\!\!\perp \delta_{gj}$ for $\forall\, i = 1, \dots, n$ and $j = 1, \dots, m$

$\beta_g, \tau_g^2, \sigma_g^2$ all depend on g, x_i are across genes

(a) what is the interpretation of $\beta_{gd}, \tau_g^2, \sigma_g^2$? Is β_{gd} identifiable in (1)?

- β_{gd} is the vector of fixed-effects (effect of dose) coefficients
- τ_g^2 is the variance between individuals
- σ_g^2 is the variance within samples of each individual
- Identifiability of β_{gd}: without constraints on the sums of coefficients, β_{gd} will not be identifiable

(b) Intraclass correlation for each gene g in terms of the parameters in Model (1).

Ratio of between-variance to total variance:

$$\text{corr}(Y_{ij}, Y_{ij'}) = \frac{\text{var}(\delta_{gm})}{\text{var}(Y_{gi})} = \frac{\tau_g^2}{\tau_g^2 + \sigma_g^2}$$

(1, continued)

(using only individuals with 5 dosage measurements)

(c) $\bar{\beta}_g = \frac{1}{5}\sum_{d=1}^{5} \beta_{gd}$ is mean for effect of gene g

what is GLS estimate for $\Delta_g = \beta_g - 1_5 \bar{\beta}_g$ wrt Y_{gi}'s?

Should not depend on τ_g^2 or σ_g^2.

If $\Delta_g = \beta_g - 1_5 \bar{\beta}_g$, we are mean-centering the data X and essentially re-fitting the model parameters

For just one gene, we have the model

$$Y_r = X\beta + Z\delta_r + e$$

we mean-center just one:

$$Y_r = (X - \bar{X})\beta^* + Z\delta_r^* + e$$

$$= X\beta^* - \bar{X}\beta^* + Z\delta_r^* + e = X\Delta + \delta_r^* + e$$

where Δ is the difference between our original β and the average across the dosage measurements, $\Delta = \beta - \bar{\beta}$

Now we do not need to rely on $\text{var}(\delta^*) = \tau^2$ because mean-centering potentially takes care

The estimate for Δ, $\hat{\Delta}$, relates to Y_r as: $\hat{\Delta} = (X^T \hat{V}^{*-1} X)^{-1} X^T \hat{V}^{*-1} Y_r$, where V only relies on σ^2 (of e).

To extend to multiple genes, we have

$$Y_{gi} = X\Delta_g + Z\delta_r^* + e_g$$, and the variance:

(d) $\text{var}(\hat{\Delta}_g) \rightarrow$ where $\hat{\Delta}_g = \hat{\beta}_g - 1_5 \hat{\bar{\beta}}_g$

$\text{var}(\hat{\Delta}_g) = (X^T \hat{V}^{*-1} X)^{-1}$ which, from (c), we know V^* only relies on σ_g^2 and not τ_g^2 after re-fitting the mean-centered model.

(1, continued)

(e) Derive an analytic, unbiased estimator for σ_g^2 and prove it's unbiased. Use to find an expression for $\hat{var}(\hat{\Delta}_g)$.

(Hint: consider $Y_{gr^*} = (Y_{gi})_{\{i : r(i) = r^*\}} \in \mathbb{R}^5$ for each $r^* \in \{1, \dots, m\}$, derive $Q_5 Y_{gr^*}$, $Q_5 \in \mathbb{R}^{5\times 5}$ is orth. proj. matrix for 1_5's compliment).

Our model: $Y = X\beta + Z\delta + e$

where $\delta \sim N(0, \tau^2)$ and $e \sim N(0, \sigma^2)$

For Q, orthogonal projection matrix for 1_5's compliment, $V^{-1} = Q^T Q$ where V is the variance of Y and is positive definite, as is V^{-1}

Transform Y:

$QY = QX\beta + QZ\delta + Qe$ → say $Qe = \tilde{e}$

where the GLS estimator for β is now

$\hat{\beta}_{gls} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ with variance

$V(\hat{\beta}_{gls}) = (X^T V^{-1} X)^{-1}$

If $var(Y) = \sigma^2 \Omega$ where $tr(\Omega) = n$, and we had Q such that $Q^T Q = \Omega^{-1}$, (gls), then $var(QY) = \sigma_g^2 I$ which can be estimated by the MSE which we know is unbiased → $\hat{\sigma}_g^2 = \tilde{e}^T \tilde{e} / (N - p)$

Now, $var(\hat{\Delta}_g)$ → $\hat{\Delta}_g = (X^T \hat{V}^{*-1} X)^{-1} X^T \hat{V}^{*-1} y$

where $var(\hat{\Delta}_g) = (X^T \hat{V}^{*-1} X)^{-1}$

and $\hat{V}^*$ is $\hat{\sigma}_g^2 I$, where $\hat{\sigma}_g^2$ is the estimate for σ_g^2 found above.

(1, continued)

ⓕ Using only individuals with 5 dosage measurements, and solutions to ⓒ & ⓔ, compute $\hat{\Delta}_g$ and $\hat{var}(\hat{\Delta}_g)$. Derive a p-value for null hypothesis $H_{0g}: \beta_{g1} = \ldots = \beta_{g5}$. Use Bonferroni to control FWER at $\alpha = 0.05$. How many H_{0g} are rejected? What does this say about the relationship between anthracycline dosage and gene expression?

$H_{0,1}: \beta_{11} = \ldots = \beta_{15} \rightarrow Y = \beta(X_1 + \ldots + X_5) + \varepsilon$

$H_{A,1}$: at least one $\neq \rightarrow Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$

I had R trouble making a loop to do this 12,317 times, so the output on the next page is just for gene 1 (first entry), with the knowledge that it can be expanded. We would want to control FWER at $\alpha = 0.05$, so for $\binom{5}{2} = 10$ comparisons, each p-value will have to be < 0.005. For V1, we reject 8 H_0's.

Since we have so many p-values leading to rejected H_0's (at least, for gene 1), there appears to be a significant relationship between dosage level and gene expression.

I also did not know how to mean-center the model (and after many days of searching could not figure it out) so the results on the next page are an attempt. The output shows the mean-centered coefficients after fitting a regular model.

Problem 1

Using only individuals with 5 dose measurements.
(f)

```
##
## Call:
## glm(formula = V1 ~ conc + line, data = sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6092  -0.8256  -0.1089   0.8444   3.0380
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.945260   0.293161   3.224  0.00150 **
## conc0.625    3.302772   0.284904  11.593  < 2e-16 ***
## conc1.25     0.841102   0.284904   2.952  0.00358 **
## conc2.5     -1.490695   0.284904  -5.232 4.65e-07 ***
## conc5       -1.751328   0.284904  -6.147 5.00e-09 ***
## line        -0.009952   0.005030  -1.978  0.04941 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.501651)
##
##     Null deviance: 894.14  on 184  degrees of freedom
## Residual deviance: 268.80  on 179  degrees of freedom
## AIC: 608.12
##
## Number of Fisher Scoring iterations: 2
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: glm(formula = V1 ~ conc + line, data = sub)
##
## Linear Hypotheses:
##                     Estimate Std. Error z value Pr(>|z|)
## 0.625 - 0 == 0       3.3028     0.2849  11.593   <0.001 ***
## 1.25 - 0 == 0        0.8411     0.2849   2.952   0.0262 *
## 2.5 - 0 == 0        -1.4907     0.2849  -5.232   <0.001 ***
## 5 - 0 == 0          -1.7513     0.2849  -6.147   <0.001 ***
## 1.25 - 0.625 == 0   -2.4617     0.2849  -8.640   <0.001 ***
## 2.5 - 0.625 == 0    -4.7935     0.2849 -16.825   <0.001 ***
## 5 - 0.625 == 0      -5.0541     0.2849 -17.740   <0.001 ***
## 2.5 - 1.25 == 0     -2.3318     0.2849  -8.185   <0.001 ***
## 5 - 1.25 == 0       -2.5924     0.2849  -9.099   <0.001 ***
## 5 - 2.5 == 0        -0.2606     0.2849  -0.915   0.8913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## (Adjusted p values reported -- single-step method)

## (Intercept)   conc0.625    conc1.25     conc2.5       conc5
##    0.575838    2.933349    0.471680   -1.860117   -2.120750
```

(1, continued)

(3) Substantial correlation between the p genes. If p-values are super-uniform for true H_{0g}, will FWER be controlled at α? If yes, prove it. If not, give a counter-example.

we should assume, under H_0, that the p-values are uniformly distributed, otherwise we have an inflated FWER. From HW7 we saw that the control is not guaranteed for dependence among p-values. the FWER says that $100(1-\alpha)\%$ of the time, our p-values should be under $\frac{\alpha}{10} \rightsquigarrow (\frac{5}{2})$ we are okay with super-uniform p-values because that is what's expected under H_0 to control FWER at α.

(Use all $n = 217$ observations)

2. Investigate expression of gene JADE1 ($g = 3884$) which is known to promote apoptosis in healthy patients.

(a) Estimate τ_g^2, σ_g^2, and intraclass correlation for $g = 3884$.

From the fitted model in R, the estimates are: ↳ see code and output attached

$$\hat{\tau}_g^2 = 0.1814$$
$$\hat{\sigma}_g^2 = 0.1917$$
$$\widehat{ICC} = \frac{0.1814}{0.1814 + 0.1917} = 0.4862$$

$(\Sigma = 0.1814 + 0.1917 = 0.3731)$

(b) Consider $H_0: \beta_{g1} = \dots = \beta_{g5}$, against H_A as unequal. Use LRT for $g = 3884$. What do you conclude?

With $\alpha = 0.05$, we get a LR test statistic of 231.76, which yields a p-value ≈ 0.000 and we reject H_0. some of β_{gi} $(i = 1, \dots, 5)$ are not equal to zero (ie, some doses have some effect on gene expression). Satterthwaite's method also garners a significant result to reject H_0.

(c) we assumed Y_{g*} was normally distributed for $g = 3884$. Reasonable? (see plots attached)

A histogram of $g = 3884$ shows a nice, somewhat symmetric bell-curve.

A qq-plot of $g = 3884$ shows some tail action away from the normal line.

To get rid of these tails, we could multiply the residuals by $A = V^{-1/2} \rightarrow A\Sigma A^T \rightarrow$ similar to what we will do in part (e)

Problem 2

Using all data.

(a)

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y.sub ~ as.factor(conc) + (1 | cell_line)
##    Data: data
##
## REML criterion at convergence: 346.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.75730 -0.52300 -0.03407  0.67182  2.66692
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  cell_line (Intercept) 0.1814   0.4259
##  Residual              0.1917   0.4378
## Number of obs: 217, groups:  cell_line, 46
##
## Fixed effects:
##                       Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)            3.07072    0.09268 121.17280  33.132   <2e-16 ***
## as.factor(conc)0.625  -0.15819    0.09425 167.63610  -1.678   0.0951 .
## as.factor(conc)1.25   -1.41025    0.09588 168.94233 -14.708   <2e-16 ***
## as.factor(conc)2.5    -1.50336    0.09585 168.82599 -15.684   <2e-16 ***
## as.factor(conc)5      -1.51355    0.09516 168.47648 -15.905   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) a.()0. a.()1. a.()2.
## as.f()0.625 -0.529
## as.fc()1.25 -0.524  0.512
## as.fct()2.5 -0.524  0.512  0.512
## as.fctr(c)5 -0.527  0.516  0.513  0.510
```

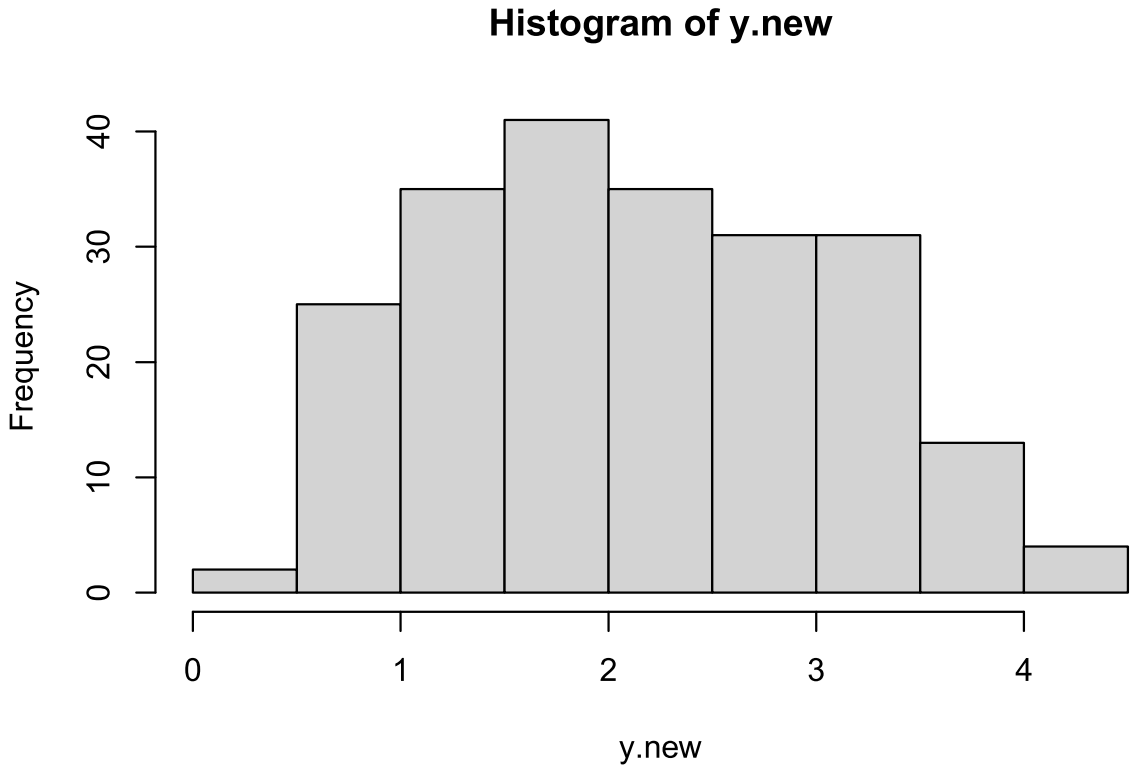
(b)

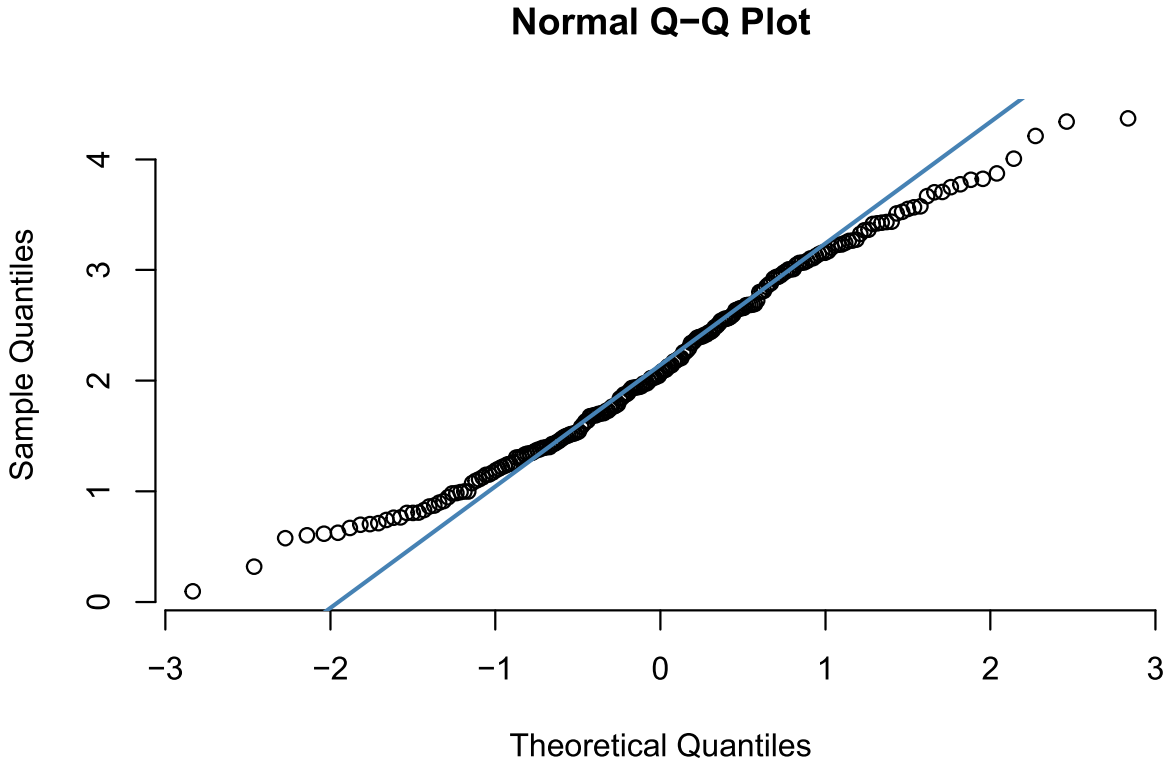
```
## Single term deletions using Satterthwaite's method:
##
## Model:
## y.sub ~ as.factor(conc) + (1 | cell_line)
##                 Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
## as.factor(conc) 99.935  24.984     4 168.18  130.34 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 231.7571
```

```
## [1] 5.524979e-49
```

(c)


Histogram of y.new
Frequency
0
10
20
30
40
0
1
2
3
4
y.new


Normal Q−Q Plot
Sample Quantiles
0
1
2
3
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

(2, continued)

(d) Estimate & find 95% CI for $\beta_{g5} - \beta_{g1}$ for $g = 3884$. what can we conclude about the relationship between anthracycline dosage and JADE1 expression? Does it support the hypothesis of anthracycline inducing ACT by inhibiting apoptosis?

$\hat{d} = \hat{\beta}_{g5} - \hat{\beta}_{g1} = -1.5135$
$\hat{se} = 0.09516$
95% CI: $(-1.773, -1.254)$

Since the CI does not include zero (all < 0), this supports the hypothesis that anthracycline induces ACT by inhibiting apoptosis.

(e) $S = \{ c^T \beta_g : \sum_{i=1}^{5} c_i = c^T 1_5 = 0 \}$ for $g = 3884$.

(i) $Y_{g*} \sim N(X\beta_g, \Sigma_g)$, Σ_g is known, pd matrix. Show that matrix A exists such that

$$AY_{g*} \sim N((AX)\beta_g, I_n)$$

$$\Sigma = \Sigma_g = \tau_g^2 + \sigma_g^2 = \sigma_g^2 I_n + \begin{bmatrix} \tau_g^2 & \tau_g^2 & \cdots & \tau_g^2 \\ \tau_g^2 & & & \\ \vdots & & & \vdots \\ \tau_g^2 & & & \tau_g^2 \end{bmatrix}$$

If, for some S, $SS^T = \Sigma$,

and $S^{-1}Y = S^{-1}X\beta + S^{-1}\delta + S^{-1}\epsilon$,

then $E(S^{-1}Y) = S^{-1}X\beta$

$E(S^{-1}\delta) = 0, \quad E(S^{-1}\epsilon) = 0$

and $\text{var}(S^{-1}Y) = S^{-1}\,\text{var}(Y)(S^{-1})^T$

$= S^{-1}\Sigma(S^{-1})^T$

$= S^{-1} S S (S^{-1})^T$

$= I_n$

so A can be S^{-1}, and gets rid of correlation/noise

(d)

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lmerTest::lmer(formula = y.sub ~ conc + (1 | cell_line), data = data)
##
## Linear Hypotheses:
##                     Estimate Std. Error z value Pr(>|z|)
## 0.625 - 0 == 0     -0.15819    0.09425  -1.678    0.447
## 1.25 - 0 == 0      -1.41025    0.09588 -14.708   <1e-04 ***
## 2.5 - 0 == 0       -1.50336    0.09585 -15.684   <1e-04 ***
## 5 - 0 == 0         -1.51355    0.09516 -15.905   <1e-04 ***
## 1.25 - 0.625 == 0 -1.25206    0.09389 -13.335   <1e-04 ***
## 2.5 - 0.625 == 0  -1.34517    0.09387 -14.330   <1e-04 ***
## 5 - 0.625 == 0     -1.35535    0.09318 -14.545   <1e-04 ***
## 2.5 - 1.25 == 0    -0.09311    0.09467  -0.984    0.863
## 5 - 1.25 == 0      -0.10329    0.09426  -1.096    0.809
## 5 - 2.5 == 0       -0.01019    0.09450  -0.108    1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)

##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lmerTest::lmer(formula = y.sub ~ conc + (1 | cell_line), data = data)
##
## Quantile = 2.7279
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                     Estimate lwr      upr
## 0.625 - 0 == 0     -0.15819 -0.41530  0.09892
## 1.25 - 0 == 0      -1.41025 -1.67180 -1.14870
## 2.5 - 0 == 0       -1.50336 -1.76483 -1.24189
## 5 - 0 == 0         -1.51355 -1.77313 -1.25396
## 1.25 - 0.625 == 0 -1.25206 -1.50819 -0.99593
## 2.5 - 0.625 == 0  -1.34517 -1.60123 -1.08911
## 5 - 0.625 == 0     -1.35535 -1.60955 -1.10116
## 2.5 - 1.25 == 0    -0.09311 -0.35134  0.16513
## 5 - 1.25 == 0      -0.10329 -0.36042  0.15383
## 5 - 2.5 == 0       -0.01019 -0.26797  0.24760

## Computing profile confidence intervals ...

##                   2.5 %      97.5 %
```

```
## .sig01       0.3278742  0.54546958
## .sigma       0.3905788  0.48291626
## (Intercept)  2.8897781  3.25180430
## conc0.625   -0.3417747  0.02541229
## conc1.25    -1.5971409 -1.22359107
## conc2.5     -1.6900953 -1.31668937
## conc5       -1.6988193 -1.32808451
```

(2, continued)

(ii) By plugging in the estimate for Σ_g from (i) for true Σ_g, use (i) and Scheffe to find 95% CI's for all points in S.

$\hat{\Sigma} = 0.3731$ from (i)

now, variance $= I_n$ and $se = 0.61092$

$0.625 - 0$: $-0.15819 \pm 1.667 \rightarrow (-1.82519, 1.50881)$

$1.25 - 0$: $-1.41025 \pm 1.667 \rightarrow (-3.07725, 0.25675)$

$2.5 - 0$: $-1.50336 \pm 1.667 \rightarrow (-3.17036, 0.16364)$

$5 - 0$: $-1.51355 \pm 1.667 \rightarrow (-3.18355, 0.15645)$

$1.25 - 0.625$: $-1.25206 \pm 1.667 \rightarrow (-2.91906, 0.41494)$

$2.5 - 0.625$: $-1.34517 \pm 1.667 \rightarrow (-3.01217, 0.32183)$

$5 - 0.625$: $-1.35535 \pm 1.667 \rightarrow (-3.02235, 0.31165)$

$2.5 - 1.25$: $-0.09311 \pm 1.667 \rightarrow (-1.76011, 1.57389)$

$5 - 1.25$: $-0.10329 \pm 1.667 \rightarrow (-1.77029, 1.56371)$

$5 - 2.5$: $-0.01019 \pm 1.667 \rightarrow (-1.67719, 1.65681)$

with real estimate for se and multiplier 2.7286

(iii) Part (ii) ignores uncertainty in estimate for Σ_g. would properly accounting return wider or narrower CI's than in (ii)?

After accounting for uncertainty, we would see wider intervals to account mathematically for the additional estimation.

(2, continued)

(f) Prefer $E(Y_{g*})$ to be linear in dose.
Use LRT to test $E(Y_{g*}) = \mu_g + \text{dose}(i)\gamma_g$ for $g = 3884$ and $i = 1, \ldots, n$, $\text{dose}(i) \in \{0, 0.625, 1.25, 2.5, 5\}$.
What is the conclusion?

(see next page for R output)

Essentially we're testing

$H_0: E(Y_g) = x_i^T \beta$

$H_A: E(Y_g) = \mu_g + \text{dose}(i)\gamma_g$

returns a statistically significant p-value, so we may be better off modeling the mean model as linear in dose.

(f)

```
## [1] 145.8022
```

```
## [1] 1.61479e-30
```

3. Assume Model (1) is correct.

$$Y_{g*} \sim N(X\beta_g, \tau_g^2 B + \sigma_g^2 I_n), \quad g = 1, \dots, p$$

(a) Derive an expression for B

If B is our partition matrix, then

$$B_{rs} = \begin{cases} 1, & r,s \text{ come from the same individual} \\ 0, & \text{otherwise} \end{cases}$$

(b) Constructing computationally efficient algorithm to estimate τ_g^2, σ_g^2 for all $g = 1, \dots, p$.

(i) Show that there exists a non-random matrix $Q \in \mathbb{R}^{n \times (n-5)}$ with orthonormal columns such that

$$\tilde{Y}_{g*} = Q^T Y_{g*} \sim N(0, \tau_g^2 \tilde{B} + \sigma_g^2 I_{n-5}), \quad g = 1, \dots, p$$

what is $\tilde{B}$ in terms of Q and B?

If $\tilde{Y}_{g*} = Q^T Y_{g*}$,

$$\tilde{Y}_{g*} = Q^T(X\beta) + Q^T(\delta) + Q^T(e)$$

Since Q has orthonormal columns, $Q^T X = 0$

$$= Q^T\delta + Q^T e$$

where $Q^T e \sim N(0, \sigma_g^2 I_{n-5})$

which gives us

$$E(\tilde{Y}_{g*}) = E[Q^T(X\beta) + Q^T\delta + Q^T e] = 0$$

$$\text{var}(\tilde{Y}_{g*}) = \text{var}[Q^T(X\beta) + Q^T\delta + Q^T e]$$

$$= 0 + \tau_g^2 \tilde{B} + \sigma_g^2 I_{n-5}$$

because $\tilde{B} = Q^T B Q = QBQ$ because, from homework, $Q^T = Q$

(3, continued)

(b) (ii) Find a non-random unitary matrix $U \in \mathbb{R}^{(n-5)\times(n-5)}$ s.t. $U^T \tilde{\tilde{Y}}_{g*} \in \mathbb{R}^{n-5} \sim N$ with independent entries for all $g = 1, \ldots, p$. Derive expressions for $E(U^T \tilde{\tilde{Y}}_{g*})$ and $\text{Var}(U^T \tilde{\tilde{Y}}_{g*})$ (latter should be a diagonal matrix).

we know $UU^T = I$ and $U^T = U^{-1}$

we need U to change the variance of Y

$\tilde{\tilde{Y}}_{g*} = Q^T Y_{g*}$

$V = \text{cov}(U^T \tilde{\tilde{Y}}_{g*}) = U^T \tilde{\tilde{B}} U + I$, where we need $U^T \tilde{\tilde{B}} U$ to be diagonal

If $\tilde{\tilde{B}} = Q^T B Q$, and $\text{cov}(U^T \tilde{\tilde{Y}}_{g*}) = U^T \tilde{\tilde{B}} U + I$,

Then $U = Q^T B$ will diagonalize V

$E(U^T Y_{g*}) = 0$

$\text{var}(U^T Y_{g*}) = U^T \tilde{\tilde{B}} U + U^T I U$

$= U^T \tilde{\tilde{B}} U + I$

$= (Q^T B)^T \tilde{\tilde{B}} Q^T B$

$= B^T Q \tilde{\tilde{B}} Q^T B$ which is diagonal

(iii) If $p \gg n$, why is estimating τ_g^2, σ_g^2 using $U^T \tilde{\tilde{Y}}_{g*}$ more efficient than Y_{g*} or $\tilde{\tilde{Y}}_{g*}$? Consider how many operations it requires to calculate the determinant of $\text{var}(U^T \tilde{\tilde{Y}}_{g*})$.

- we know $\text{var}(Y_{g*})$ takes n^3 operations
- and estimating τ_g^2, σ_g^2 has pn^3 run-times
- reduce to just n operations
  → because our variance is now a diag matrix
- $U^T \tilde{\tilde{Y}}_{g*}$ will be more efficient because we only need to consider n operations, and with a diagonal variance matrix there are less "transactions"

R Code for STAT 2132 Final Exam

```
load("Genes.RData")
data = read.csv("Data.csv")
y.t = t(Y) # switch rows and columns - transpose of Y

library(nlme)
library(lme4)
library(lmerTest)
library(multcomp)
# library(dplyr)
# library(GLSME)


## Problem 1
# (f)

all = cbind(data, y.t)
sub = all[all$return == 1,]
sub$conc = as.factor(sub$conc)

# do with just one gene, show method, can be expanded to all genes
fit = glm(V1 ~ conc + line, data = sub)
summary(fit)
summary(glht(fit, linfct = mcp(conc = "Tukey")))

# mean centering??
coef = fit$coefficients
avg = mean(fit$coefficients[1:5])
mean_cent = coef[1:5] - avg
mean_cent # mean-centered coefficients (???)


## Problem 2
# (a)

y.new = Y[3884,] # just gene JADE1: g = 3884
data$y.sub = y.new
fit2 = lmerTest::lmer(y.sub ~ as.factor(conc) + (1|cell_line), data = data)
summary(fit2)


# (b)
drop1(fit2, test = "Chisq")
fit.null = lmerTest::lmer(y.sub ~ 1 + (1|cell_line), data = data)
fit.alt = lmerTest::lmer(y.sub ~ as.factor(conc) + (1|cell_line), data = data)
testStat = as.numeric(2*(logLik(fit.alt) - logLik(fit.null)))
```

```
testStat
pval = pchisq(testStat, 4, lower = FALSE)
pval
# anova(fit.null, fit.alt, refit = FALSE)


# (c)
hist(y.new)
qqnorm(y.new, pch = 1, frame = FALSE)
qqline(y.new, col = "steelblue", lwd = 2)


# (d)
data$conc = as.factor(data$conc)
fit2 = lmerTest::lmer(y.sub ~ conc + (1|cell_line), data = data)
summary(glht(fit2, linfct = mcp(conc = "Tukey")))
confint(glht(fit2, linfct = mcp(conc = "Tukey"))) # CI for difference in dose effects
confint(fit2) # CI for each estimate


# (f)
# fit model with regular covariate, fit model linearly, run LRT
fit.h0 = lmer(y.sub ~ conc + (1|cell_line), data = data)
fit.h1 = lm(y.sub ~ conc + cell_line, data = data)
test.stat = as.numeric(2*(logLik(fit.h1) - logLik(fit.h0)))
test.stat
p = pchisq(test.stat, 4, lower = FALSE)
p
```