STOR 664: FALL 2023
Final Exam, December 7–8, 2023

The exam will be posted on gradescope and available from 6:00 pm Thursday, December 7 through 11:59 pm Friday, December 8.

This is an open-book, take-home exam. Access to course materials and standard computational tools (in particular, R) is allowed; communication with other students or with anybody via the internet, other than the instructor, is not. The university Honor Code is in effect at all times.

Answers may be given in any of the following formats (including combinations of more than one): R Markdown, Word, Latex or handwritten pages scanned or photographed; if handwritten, it is recommended you use blue or black ink on plain sheets of white paper. Hadwritted script on an iPad or tablet will also be accepted if uploaded in machine-readable (e.g. pdf) format. You are requested to submit your final solutions in gradescope but if for some reason that doesn't work, you may email tem to the instructor.

Each question will be graded out of 50 points for a total score of 100 for the whole exam. An error in one part-question will not prevent you gaining full credit in other parts of the same question even if you carry over the error. You may answer the questions (or parts within a question) in any order; answers out of sequence will not be penalized so long as you make clear which part of your answer refers to which part of the question sheet.

There is no official time limit for the exam but it is strongly recommended that you self-limit to 6 hours total; continuing to work beyond that time is unlikely to improve your grade. Any queries about the exam may be addressed directly to the instructor by email, text message or cellphone.

1. **Three-dimensional response surface design**. Consider an experiment with three variables $X_1,\ X_2,\ X_3$ laid out as follows:

```
X1: + + + + + + + + + 0 0 0 0 0 0 0 0 0 - - - - - - - - -
X2: + + + 0 0 0 - - - + + + 0 0 0 - - - + + + 0 0 0 - - -
X3: + 0 - + 0 - + 0 - + 0 - + 0 - + 0 - + 0 - + 0 - + 0 -
```

Here, the symbols $+$, 0 and - represent the numerical values $1,\ 0,\ -1$ and the order is intended to represent the actual order of covariates in the 27 observations in the experiment. (For example, the variable X_1 is represented through observations $x_{1,1} = x_{2,1} = \ldots = x_{9,1}, x_{10,1} = \ldots = x_{18,1} = 0, x_{19,1} = \ldots = x_{27,1} = -1$ with corresponding notation for $x_{i,2}$ and $x_{i,3},\ i = \ldots, 27$.) Also assume that each of the 27 responses is given by a formula of the form

$$y_i \quad = \quad \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,1}^2 + \beta_5 x_{i,2}^2 + \beta_6 x_{i,3}^2 + \epsilon_i$$

where $x_{i,1},\ x_{i,2},\ x_{i,3}$ represent the values of $X_1,\ X_2,\ X_3$ in the ith row of data and $\{\epsilon_i,\ i = 1, \ldots, 27\}$ represent independent normally distributed errors with mean 0 and common variance σ^2.

   (a) Write down the X matrix for this regression problem and calculate $X^T X$ and $(X^T X)^{-1}$. (*Note: Since this is an open-book computer-aided exam, you are allowed and encouraged to use R, matlab or other appropriate software to complete the calculations, but your final answer should be an explicit statement of the result.*) **[10 points]**

(b) Define:

$$\begin{aligned} T_0 &= \sum y_i, \\ T_1 &= \sum y_i x_{i,1}, \\ T_2 &= \sum y_i x_{i,2}, \\ T_3 &= \sum y_i x_{i,3}, \\ T_4 &= \sum y_i x_{i,1}^2, \\ T_5 &= \sum y_i x_{i,2}^2, \\ T_6 &= \sum y_i x_{i,3}^2 \end{aligned}$$

Show that

$$\hat{\beta}_0 = \frac{7}{27} T_0 - \frac{1}{9}(T_4 + T_5 + T_6)$$

and derive similar expressions for $\hat{\beta}_j,\ j = 1, \ldots, 6$. What are the variances of these estimators? **[10 points]**

(c) Now suppose the objective is to find the values $X_1^*,\ X_2^*,\ X_3^*$ (not necessarily integers) to maximize the expected response $\beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \beta_3 X_3^* + \beta_4 X_1^{*2} + \beta_5 X_2^{*2} + \beta_6 X_3^{*2}$. Assume $\beta_4,\ \beta_5,\ \beta_6$ are all < 0 so that the surface has a finite maximum. Derive theoretical expressions for $X_1^*,\ X_2^*,\ X_3^*$ and show how to calculate estimators $\widehat{X_1^*},\ \widehat{X_2^*},\ \widehat{X_3^*}$ in terms of the estimators calculated in part (b). **[5 points]**

(d) Are the estimators $\widehat{X_1^*},\ \widehat{X_2^*},\ \widehat{X_3^*}$ independent? — explain why or why not. Using whatever approximations you consider appropriate, state formulas for the standard errors $\widehat{X_1^*},\ \widehat{X_2^*},\ \widehat{X_3^*}$ in terms of the quantities $\hat{\beta}_0, \ldots, \hat{\beta}_6$ and s^2. (Assume s^2 is the usual unbiased sample estimator of σ^2; you don't need to give a separate formula for that.) **[10 points]**

(e) Suppose you want to find a 99% confidence interval for X_1^*. Show how to do this using (i) delta method, (ii) Fieller method. In particular, show that the endpoints of the Fieller confidence set are the solutions of the quadratic equation

$$(4\hat{\beta}_4^2 - 5.397305 s^2)x^2 + 4\hat{\beta}_1\hat{\beta}_4 x + \hat{\beta}_1^2 - 0.4497754 s^2 = 0,$$

and state the conditions under which the confidence set is an interval.

(It is acceptable that you may get slightly different values for the numerical quantities 5.397305 and 0.4497754, but you should show where they come from. Since very similar formulas will apply for X_2^* and X_3^*, you are not asked to find those.) **[15 points]**

Question 2 on the next page.

2. **Computational Exercise (This part is expected to be completed using R)**. You may use any R packages but your final answer should include your R code including any packages that you use.

   Consider the `diabetes` dataset in Faraway's package. This may be loaded into R directly through `library(faraway)` followed by `data(diabetes)`. The dataset includes a response variable `glyhb` (Glycosolated Hemoglobin) and numerous predictors. Glycosolated hemoglobin greater than 7.0 is usually taken as a positive diagnosis of diabetes.

   Remove the variables `id` (not relevant for predictions), `bp.2s` and `bp.2d` (too many missing datapoints) and use `na.omit` to reduce the dataset to one where all observations are complete. (These operations are considered part of the initial data manipulation and not awarded formal credit. You should end up with a matrix with 366 observations and 16 variables including `glyhb`.)

   (a) Fit a linear regression model for `glyhb` on the other 15 variables. Use stepwise regression to reduce the number of variables and comment briefly on your results. [**5 points**]

   (b) Would you consider any transformation of `glyhb`? Use plots and any numerical diagnostics you consider appropriate, choose a suitable transformation, and repeat the analysis of part (a). (For the rest of this question, you should use the transformed values, unless your decision is to stick with the original values of `glyhb`.) [**5 points**]

   The next few parts should be based on your final model from parts (a) and (b).

   (c) Calculate the studentized residuals and comment on whether there appear to be outliers, justifying your answer with appropriate tests or plots. [**5 points**]

   (d) Plot the residuals against each of the covariate and the fitted values; hence comment on whether there is an evidence that the model dies not fit the data. [**5 points**]

   (e) Do the residuals appear to follow a normal distribution? Briefly justify your answer with suitable tests or plots. [**5 points**]

   (f) Calculate appropriate diagnostics for leverage and influence, and comment on the results. [**5 points**]

   (g) One observation appears to have very large leverage and another seems to have large influence. Identify these variables and repeat the main parts of steps (c) through (f) without those observations; do any of your conclusions substantially change? [**5 points**]

   (h) Returning to the original dataset with 366 observations and 16 variables, but still using any transformation you decided to use on `glyhb`, split the data into a training dataset and a test dataset where the test dataset consists of every tenth observations (i.e. rows 10, 20, ..., 360) and the training dataset is everything else. Based on the training dataset, repeat your model selection from part (b) and find alternative models using (i) principal components regression, (ii) PLS regression, (iii) ridge regression and (iv) LASSO. You may use any appropriate package(s) but please be explicit about your method of calculation.

   Then, for all five methods, use the best model you fitted to predict the values of `glyhb` on the test dataset. Use root mean squared error (on the *original* scale — this way we can compare estimators using different transformations) to decide which of the five methods is best on this example. Summarize your conclusions. [**15 points**]

Sample Solutions

Note: As with any exam of this nature, there are many possible variations on the solutions. Equivalent solutions that get the right answers will earn full credit, provided they are adequately explained.

1. (a) You can write this out longhand or use an R script such as

```
x0=rep(1,27)
x1=c(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,-1,-1,-1,-1,-1,-1,-1,-1,-1)
x2=c(1, 1, 1, 0, 0, 0,-1,-1,-1, 1, 1, 1, 0, 0, 0,-1,-1,-1, 1, 1, 1, 0, 0, 0,-1,-1,-1)
x3=c(1, 0,-1, 1, 0,-1, 1, 0,-1, 1, 0,-1, 1, 0,-1, 1, 0,-1, 1, 0,-1, 1, 0,-1, 1, 0,-1)
X=cbind(x0,x1,x2,x3,x1*x1,x2*x2,x3*x3)
S=t(X)%*%X
SI=solve(S)
```

where `S` represents $X^T X$ and `SI` represents $(X^T X)^{-1}$. You can also quickly check that $54 * SI$ is an integer array. Therefore,

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & -1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ -1 & -1 & -1 \end{pmatrix}, \ X^T X = \begin{pmatrix} 27 & 0 & 0 & 0 & 18 & 18 & 18 \\ 0 & 18 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 18 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 18 & 0 & 0 & 0 \\ 18 & 0 & 0 & 0 & 18 & 12 & 12 \\ 18 & 0 & 0 & 0 & 12 & 18 & 12 \\ 18 & 0 & 0 & 0 & 12 & 12 & 18 \end{pmatrix}, \ (X^T X)^{-1} = \frac{1}{54} \begin{pmatrix} 14 & 0 & 0 & 0 & -6 & -6 & -6 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ -6 & 0 & 0 & 0 & 9 & 0 & 0 \\ -6 & 0 & 0 & 0 & 0 & 9 & 0 \\ -6 & 0 & 0 & 0 & 0 & 0 & 9 \end{pmatrix}.$$

(b) The vector $\begin{pmatrix} T_0 \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \end{pmatrix}$ represents $X^T\mathbf{y}$ and hence the estimators $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ lead to

$$\hat{\boldsymbol{\beta}} \ = \ \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \end{pmatrix} \ = \ \frac{1}{54} \begin{pmatrix} 14T_0 - 6(T_4 + T_5 + T_6) \\ 3T_1 \\ 3T_2 \\ 3T_3 \\ -6T_0 + 9T_4 \\ -6T_0 + 9T_5 \\ -6T_0 + 9T_6 \end{pmatrix}.$$

The top row is equivalent to the given expression for $\hat{\beta}_0$ and the remaining rows lead to $\hat{\beta}_1 = \frac{T_1}{18}, \ldots, \hat{\beta}_4 = \frac{T_4}{6} - \frac{T_0}{9}$,etc. Based on $(X^T X)^{-1}\sigma^2$, the variances of $\hat{\beta}_0, \ldots, \hat{\beta}_6$ are $\frac{7\sigma^2}{27}, \frac{\sigma^2}{18}, \frac{\sigma^2}{18}, \frac{\sigma^2}{18}, \frac{\sigma^2}{6}, \frac{\sigma^2}{6}, \frac{\sigma^2}{6}$.

(c) From basic calculus $X_1^* = -\frac{\beta_1}{2\beta_4}$, $X_2^* = -\frac{\beta_2}{2\beta_5}$, $X_3^* = -\frac{\beta_3}{2\beta_6}$. The obvious estimators will substitute $\hat{\beta}_j$ for β_j, e.g. $\widehat{X_1^*} = -\frac{\hat{\beta}_1}{2\hat{\beta}_4}$.

(d) If we omit the first row and column of $(X^TX)^{-1}$, the remaining 6×6 matrix is diagonal; hence, $\hat{\beta}_1, \ldots, \hat{\beta}_6$ are mutually independent. Therefore, the pairs $(\hat{\beta}_1, \hat{\beta}_4)$, $(\hat{\beta}_2, \hat{\beta}_5)$, $(\hat{\beta}_3, \hat{\beta}_6)$ are mutually independent and, hence, so are any functions of these pairs. Therefore, the three estimators $\widehat{X_1^*}$, $\widehat{X_2^*}$, $\widehat{X_3^*}$ *are* mutually independent. (Note that $\hat{\beta}_0$ is correlated with $\hat{\beta}_j$ for $j = 4, 5, 6$ but that does not affect your answer because none of $\widehat{X_1^*}$, $\widehat{X_2^*}$, $\widehat{X_3^*}$ depends on $\hat{\beta}_0$.)

Using the delta method, the variance of $-\frac{\hat{\beta}_1}{2\hat{\beta}_4}$ is approximately $\frac{1}{4\beta_4^2}\mathrm{Var}(\hat{\beta}_1) + \frac{\beta_1^2}{4\beta_4^4}\mathrm{Var}(\hat{\beta}_4)$ (no covariance term because $\hat{\beta}_1$ and $\hat{\beta}_4$ are uncorrelated and hence independent). Using the previously derived formulas for the variances, this reduces to $\frac{1}{4\beta_4^2}\frac{\sigma^2}{18} + \frac{\beta_1^2}{4\beta_4^4}\frac{\sigma^2}{18} = \frac{\sigma^2}{72\beta_4^4}\left(\beta_4^2 + 3\beta_1^2\right)$ for the approximate) variance of $\widehat{X_1^*}$. Similarly, the approximate variances of $\widehat{X_2^*}$ and $\widehat{X_3^*}$ are $\frac{\sigma^2}{72\beta_5^4}\left(\beta_5^2 + 3\beta_2^2\right)$ and $\frac{\sigma^2}{72\beta_6^4}\left(\beta_6^2 + 3\beta_3^2\right)$.

Therefore, suitable approximate formulas for the standard errors of $\widehat{X_1^*}$, $\widehat{X_2^*}$ and $\widehat{X_3^*}$ are $s\sqrt{\frac{\hat{\beta}_4^2+3\hat{\beta}_1^2}{72\hat{\beta}_4^4}}$, $s\sqrt{\frac{\hat{\beta}_5^2+3\hat{\beta}_2^2}{72\hat{\beta}_5^4}}$, $s\sqrt{\frac{\hat{\beta}_6^2+3\hat{\beta}_3^2}{72\hat{\beta}_6^4}}$.

(e) The delta method will treat the standard errors of part (e) as exact, leading to a confidence interval for X_1^* of the form

$$-\frac{\hat{\beta}_1}{2\hat{\beta}_4} \pm 2.84534s\sqrt{\frac{\hat{\beta}_4^2 + 3\hat{\beta}_1^2}{72\hat{\beta}_4^4}}.$$

The number 2.84534 comes from `qt(0.995,20)` in R.

The Fieller method for a confidence interval will include all values x for which a test of $H_0:\ X_1^* = x$ versus $H_1:\ X_1^* \neq x$ will accept H_0 at the two-sided significance level 0.01. Since H_0 is equivalent to the hypothesis $\beta_2 + 2\beta_4 x = 0$, this hypothesis is accepted for all x such that

$$\left|\frac{\hat{\beta}_1 + 2\hat{\beta}_4 x}{s\sqrt{\frac{1}{18} + \frac{4x^2}{6}}}\right| \leq 2.84534$$

where the number 2.84534 is again derived from `qt(0.995,20)`.

$$(\hat{\beta}_1 + 2\hat{\beta}_4 x)^2 \leq s^2(0.4497754 + 5.397305x^2).$$

The numbers 0.4497754 and 5.397305 may be formally derived in R as

```
c(qt(0.995,20)^2/18,2*qt(0.995,20)^2/3)
```

The endpoints of the confidence set are derived by replacing the $\leq$ in the last equation by $=$, which is quickly rearranged to the equation given in the question. The condition for this confidence set to be an interval is that the coefficient of x^2 is positive, i.e. $4\hat{\beta}_4^2 > 5.397305$.

2. (a) Some possible initial code for this example is

```
library(faraway)
data(diabetes)
diab1=diabetes[,-c(1, 15, 16)]
diab2=na.omit(diab1)
lm1=lm(glyhb˜., diab2)
lm2=step(lm1)
```

which leads to the following (edited) tables for `lm1` and `lm2`:

```
summary(lm1)
....
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.1567074  2.3918851  -0.902   0.3679
chol            0.0050363  0.0033584   1.500   0.1346
stab.glu        0.0275617  0.0015274  18.045   <2e-16 ***
hdl            -0.0034033  0.0104283  -0.326   0.7444
ratio           0.0851633  0.1142484   0.745   0.4565
locationLouisa -0.2337873  0.1608593  -1.453   0.1470
age             0.0134974  0.0060629   2.226   0.0266 *
genderfemale    0.0850168  0.2517108   0.338   0.7358
height          0.0227918  0.0306269   0.744   0.4573
weight         -0.0041222  0.0052581  -0.784   0.4336
framemedium     0.0518612  0.1985753   0.261   0.7941
framelarge     -0.2281141  0.2642373  -0.863   0.3886
bp.1s           0.0027597  0.0048761   0.566   0.5718
bp.1d          -0.0014471  0.0075742  -0.191   0.8486
waist           0.0295509  0.0308031   0.959   0.3380
hip             0.0157799  0.0349919   0.451   0.6523
time.ppn        0.0005652  0.0002494   2.266   0.0241 *
Residual standard error: 1.439 on 349 degrees of freedom
Multiple R-squared:  0.6025,Adjusted R-squared:  0.5843
....
summary(lm2)
....
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.3766599  0.4038739   0.933   0.3516
chol            0.0046921  0.0019880   2.360   0.0188 *
stab.glu        0.0276382  0.0014935  18.506   <2e-16 ***
ratio           0.1201150  0.0503588   2.385   0.0176 *
locationLouisa -0.2232598  0.1518282  -1.470   0.1423
age             0.0147647  0.0048871   3.021   0.0027 **
time.ppn        0.0005037  0.0002439   2.065   0.0396 *
Residual standard error: 1.429 on 359 degrees of freedom
Multiple R-squared:  0.5967,Adjusted R-squared:   0.59
```

The model `lm2` drops several insigificant variables from `lm1`. The most significant variables for predicting `glyhb` appear to be `stab.glu` (Stabilized Glucose) and `age`.

(b) See, for example,

```
par(mfrow=c(2,2))
library(MASS)
boxcox(lm2)
hist(diab2$glyhb, breaks=20)
hist(log(diab2$glyhb), breaks=20)
hist(1/diab2$glyhb, breaks=20)
```

The Box-Cox plot suggests $\lambda \approx -1$ so we show the three histograms for the original y variable (diab2$glyhb) and for $\log y$ and $1/y$. All three histograms are somewhat asymmetrical but that for $1/y$ is arguably closest to normal.

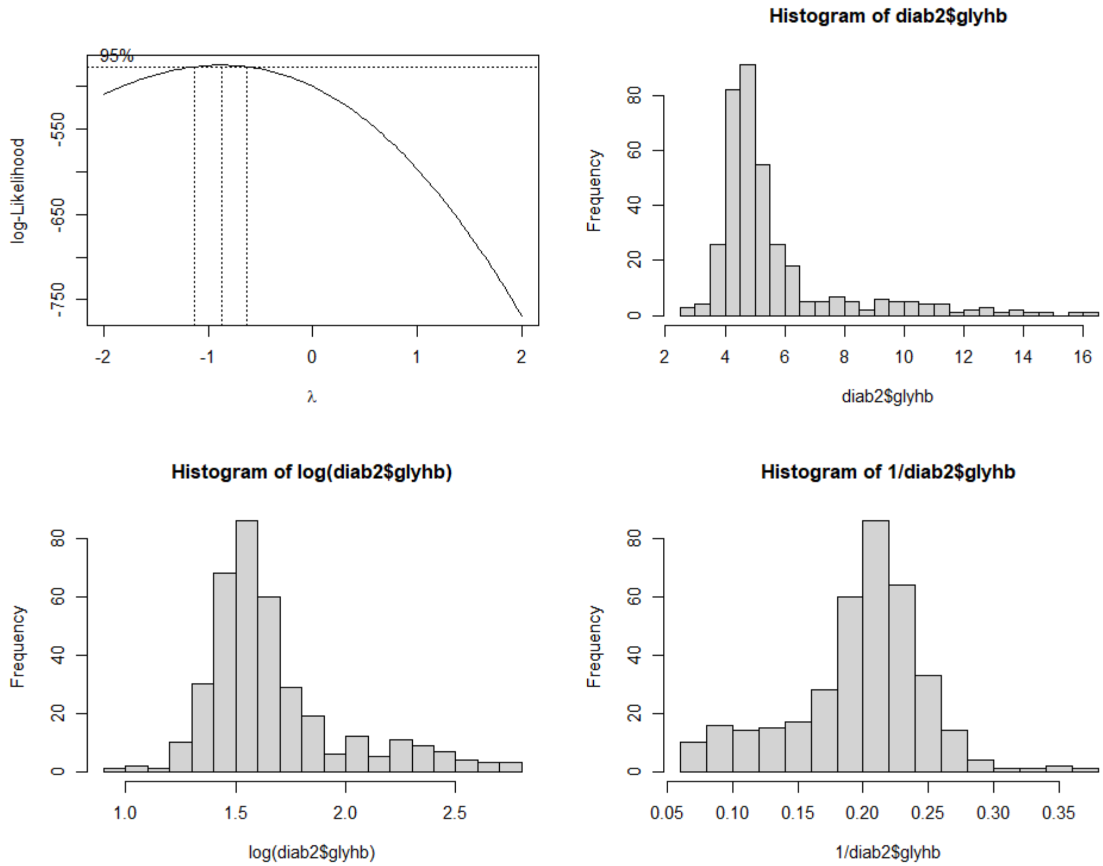

Figure 1: Box-Cox plot for Model `lm2` and Three Histograms.

In this analysis, I transform `glyhb` (without changing its name) and repeat the analysis of before:

```
diab2$glyhb=1/diab2$glyhb
lm3=lm(glyhb˜., diab2)
```

```
lm4=step(lm3)
summary(lm4)
```

leading to

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.494e-01  1.520e-02  22.982  < 2e-16 ***
chol           -7.209e-05  5.027e-05  -1.434  0.15237
stab.glu       -4.952e-04  3.806e-05 -13.011  < 2e-16 ***
ratio          -2.910e-03  1.311e-03  -2.220  0.02703 *
locationLouisa  6.523e-03  3.845e-03   1.697  0.09063 .
age            -6.216e-04  1.241e-04  -5.010 8.56e-07 ***
waist          -1.093e-03  3.496e-04  -3.127  0.00191 **
time.ppn       -1.177e-05  6.174e-06  -1.907  0.05735 .
Residual standard error: 0.03611 on 358 degrees of freedom
Multiple R-squared:  0.5134,Adjusted R-squared:  0.5038
```

The variables `stab.glu` and `age` are still the most significant but also `waist` is identified. It is somewhat disturbing that both forms of R^2 are substantially lower than in the original model, which raises the question of whether the $y \to 1/y$ transformation has really improved the model. For the rest of this solution, I shall stick with this transformation but will also give credit for the original or some alternative transformation if it is consistently followed through.

(c) You can calculate the studentized residuals by `rstudent(lm4)`; the range goes from –4.205 to 4.371 whih suggests that there are indeed outliers (we know that studentized residuals have the distribution $t_{n-p-1} = t_{357}$ if the model is correct; the 2-sided p-value associated with 4.371 is about 1.6×10^{-5} which is extremely low even allowing that this is the most extreme out of 366 observations). We conclude that there are outliers in the data.

(d) None of the residual plots (Figure 2) suggests anything obviously wrong with the model.

(e) Applying `qqnorm` followed by `qqline` to the studentized residuals gives the left-hand plot in Figure 3 (the regular residuals would produce something very similar). The plot does not appear consistent with a normal distribution. For a goodness of fit test, you could load the R function `gofsim` from https://rls.sites.oasis.unc.edu/faculty/rs/source/Data/Rcode-gof.txt, then `gofsim(diab2$glyhb,model.matrix(lm4),0,1000)` shows p-values very close to 0 in all four tests. (You could, of course, use anybody else's implementations of the goodness of fit tests and you should come to the same conclusion.)

(f) `sort(hatvalues(lm4))` identifies observation 56 (numbered 63 in the original database) as very high leverage, and `sort(cooks.distance(lm4))` shows observation 176 (original database 195) as influential (see the right-hand plot in Figure 3; although not marked, the observation on the extreme right is number 63).

(g) I am not showing all the estimates and plots here, but very little changes if those two observations are omitted; no new points emerge as having high leverage or Cook's distance, but there are still outliers and strong evidence against the fit of a normal distribution. (Also, although I'm not showing all those estimates and plots either, if you repeat the results of (c) through (g) with no transformation of `glyhb`, you again get similar results, in fact even worse for the fit of a normal distribution. I conclude that there is no straightforward transformation that solves the residual normality problem.)

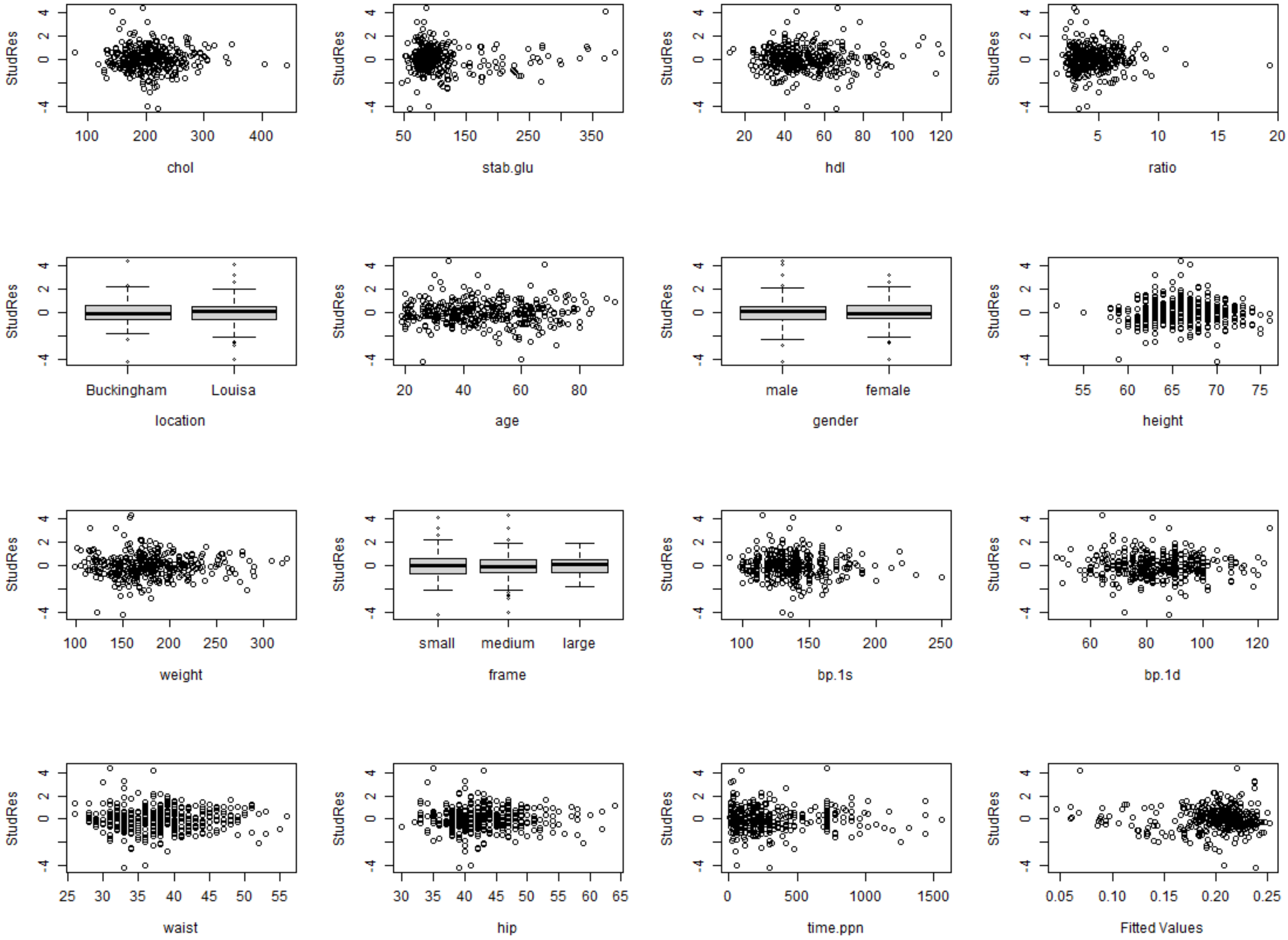

Figure 2: Studentized residuals plotted against covariates and fitted values.

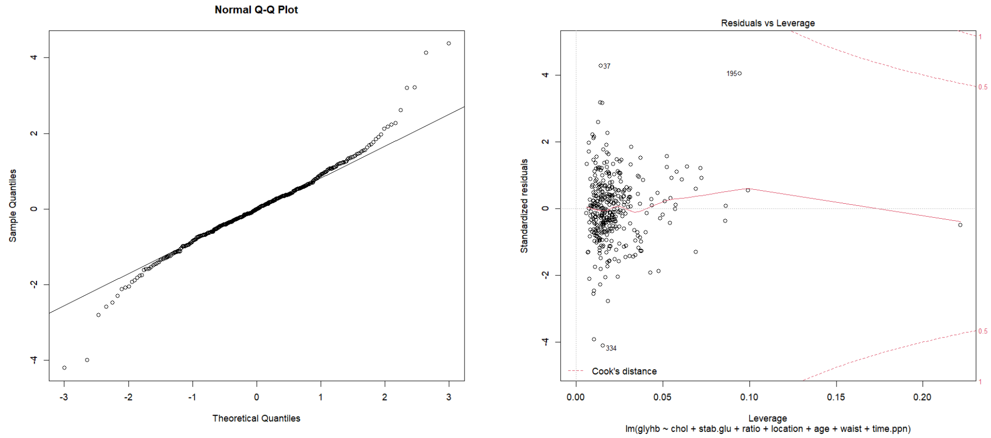

Figure 3: Normal probability plot for studentized residuals; Residuals versus Leverage plot.

(h) With the dataset first split into a training dataset and a test dataset as described, the first analysis repeated the stepwise variable selection given earlier, and the resulting model used to predict the values of `glhyb` in the test dataset. Since these values are the reciprocals of the original values, and the question explicitly asked to compute the RMSE on the original scale, the predictions were evaluated as $\sqrt{\sum_i \left(\frac{1}{\widehat{y_i^*}} - \frac{1}{y_i^*}\right)^2}$ where y_i^* is the ith test value and $\widehat{y_i^*}$ its predictor. Next, PCR and PLS regression were fitted using the `pls` library, the optimal model size being selected by leave-one-out cross-validation. [Side comment: the default 10-fold cross-validation would also be acceptable but, because that means a random split of the data, would give slightly different values.] The PCR method selected a model with 7 components and the PLS method selected a model with 5 components. The predictions on the test dataset were evaluated by the same RMSE function as used for the stepwise selection method. Finally, the LASSO and ridge regression methods were implemented using the `glmnet` function with $\alpha = 1$ and $\alpha = 0$ respectively. Leave-one-out cross-validation (`nfolds=330`) was used to select λ; the diagram illustrating this process for LASSO is shown as Figure 4. The resulting models with optimal λ were used to predict the response values on the test dataset. The RMSEs for the five methods were 1.917, 1.936, 1.939, 1.940 and 1.935 respectively, which imply that the stepwise regression method worked best, though all five methods give similar results. Some possible R code is as follows (if you looked up the 2022 final exam, you will see that very similar code was used for the corresponding question there):

```
rmse=function(x1,x2){sqrt(mean((1/x1-1/x2)^2))}
# rmse defined to calculate on original scale
diabte=diab2[10*1:36,]
diabtr=diab2[-10*1:36,]
lm5=lm(glyhb~.,diabtr)
lm6=step(lm5)
rmse(predict(lm6,diabte),diabte$glyhb)
# result 1.91741
#
# PCR and PLS regression
library(pls)
pcrmod=pcr(glyhb~.,data=diabtr,ncomp=15,validation=’LOO’)
```

```
pcrCV=RMSEP(pcrmod,estimate=’CV’)
which.min(pcrCV$val) # result is 8, corresponds to 7 components
rmse(predict(pcrmod,diabte,ncomp=7),diabte$glyhb)
# result 1.936387
pcrmod=plsr(glyhb~.,data=diabtr,ncomp=15,validation=’LOO’)
pcrCV=RMSEP(pcrmod,estimate=’CV’)
which.min(pcrCV$val) # result is 6, corresponds to 5 components
rmse(predict(pcrmod,diabte,ncomp=5),diabte$glyhb)
# result 1.938924
#
# ridge and lasso reression correspond to alpha=0 and alpha=1 in glmnet
library(glmnet)
par(mfrow=c(1,2),cex=1.3)
# note that the X matrix in glmnet must be numerical, so we have to take
# the model.matrix from previous analyses
X=model.matrix(lm1)[,2:17]
Xte=X[10*1:36,]
Xtr=X[-10*1:36,]
ytr=diabtr$glyhb
yte=diabte$glyhb
# lasso first
fit=glmnet(Xtr,ytr,alpha=1)
cvfit=cv.glmnet(Xtr,ytr,alpha=1,nfolds=330)
cvfit$lambda.min # result 0.0006689415
which.min(cvfit$cvm) # result 43
coef(cvfit, s = "lambda.min")
sum(abs(coef(cvfit, s = "lambda.min")[2:17])) # result 0.01521026
plot(fit)
mtext(side=3,line=2,’Diabetes Training Dataset’,cex=2)
lines(c(0.0152,0.0152),c(-100,100))
text(0.0152,0.004,expression(lambda==0.00067))
text(0.0152,0.0035,expression(sum(abs(beta[j]))==0.0152))
plot(cvfit)
mtext(side=3,line=2,’Cross-Validation’,cex=2)
rmse(predict(fit,Xte)[,43],yte)
# result 1.939561
#
# same for ridge
fit=glmnet(Xtr,ytr,alpha=0)
cvfit=cv.glmnet(Xtr,ytr,alpha=0,nfolds=330)
cvfit$lambda.min # result 006385371
which.min(cvfit$cvm) # result 93
coef(cvfit, s = "lambda.min")
sum(abs(coef(cvfit, s = "lambda.min")[2:18])) # result 0.01762758
plot(fit)
mtext(side=3,line=2,’Diabetes Training Dataset’,cex=2)
lines(c(0.0176,0.0176),c(-100,100))
text(0.0176,0.004,expression(lambda==0.0064))
text(0.0176,0.0035,expression(sum(abs(beta[j]))==0.0176))
plot(cvfit)
mtext(side=3,line=2,’Cross-Validation’,cex=2)
```

```
rmse(predict(fit,Xte)[,93],yte)
# result 1.934743
```

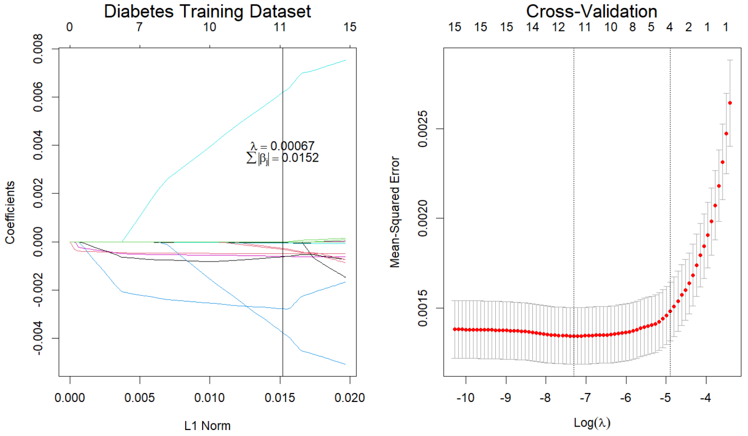

Figure 4: Plots for the LASSO fit.

Comments on student solutions. On the whole I was satisfied with the scores on the final exam; the mean was 86.4, median 89, first and third quartiles were 79.5 and 93.5 as computed in R.

Question 1 (mean score 43.9/50) scored slightly higher than Question 2 (mean 42.5). On Question 1, many students were confused by the independence question in (d); some students misstated the t distribution for parts (d) and (e); and some forgot or just didn't answer the question about when the Fieller confidence set is a confidence interval.

In Question 2, there seemed to be a lot of confusion about the transformation of the response variable. It surprised me that some students did the Box-Cox analysis correctly, but failed to draw the correct interpretation that some λ near –1 was called for (you could get away with a different transformation, such as a logarithmic transformation, but then you should have explained exactly why you chose that).

Parts (c) through (g) were fairly mechanical, though many students simply gave a long list of observations that were supposedly either outliers or influential without attempting much interpretation. Part (h) was of course the trickiest part, but you had plenty of notice (e.g. last year's exam) that something like this was going to come up, and the answer to this is also fairly mechanical once you figure out how the software packages work. In many cases, however, I couldn't figure out how you got your final answer, so a lot of the grade was based on how close you got to what I consider the correct answer, which is an RMSE between 1.9 and 2 for all five methods. The model answer given above was based on $\lambda = -1$, but since the question asked you to compute RMSEs on the original scale of the data, you should get answers of the same magnitude whatever your initial transformation. (For the record, I tried $\lambda = 1,\ 0.5,\ 0$ and –0.5 as well as $\lambda = -1$, and all of my final RMSEs were between 1.9 and 2.)

Overall, I felt that your understanding of the methods and mathematical techniques was quite good, but quite a few answers were poorly written. Given that for large parts of an exam like this, there is no single "right answer", so a lot of the credit went to how well you explained the answers you gave. I deducted many more points for poor explanations (or, in some cases, no explanation at all) than I did for technical errors!