STAT 578 - Assignment 1 - due date is on the course outline

Note: In all questions in this assignment it should be assumed that the design matrix $\mathbf{X}_{n\times p}$ has full rank p, and that the random errors ε are independent, equally varied and Normal. On this and other assignments you are welcome to come to me for assistance, or for guidance to helpful literature.

1. Consider the usual linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. Suppose that one anticipates making a new observation - call it Y_{new} - on the r.v. Y at the value $\mathbf{x}_0$ of $\mathbf{x}$. In class we derived a *confidence interval* on $E[Y|\mathbf{x}_0] = \mathbf{x}_0^T\boldsymbol{\theta}$. Here you are asked to derive a *prediction interval* for Y_{new}, i.e. an interval - PI, say - with the property that
$$1 - \alpha = P\left(Y_{new} \text{ will lie in PI}\right).$$

2. (Scheffé type simultaneous confidence intervals) In the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, consider the problem of constructing confidence intervals on all linear combinations $\mathbf{x}'\boldsymbol{\theta}$. One wants intervals of the form $\mathbf{x}'\hat{\boldsymbol{\theta}} \pm c \cdot \left(\text{est'd std. dev. of } \mathbf{x}'\hat{\boldsymbol{\theta}}\right)$ where the constant c is such that, before sampling, the simultaneous coverage probability is $1-\alpha$:
$$1 - \alpha \leq P\left(|\mathbf{x}'\hat{\boldsymbol{\theta}} - \mathbf{x}'\boldsymbol{\theta}| \leq c \cdot \left(\text{est'd std. dev. of } \mathbf{x}'\hat{\boldsymbol{\theta}}\right) \text{ for } \underline{\text{all}}\ \mathbf{x}\right).$$
Show that $c = \sqrt{pF^p_{n-p}(1-\alpha)}$ is the smallest value of c with this property. (The QR-decomposition will be helpful here.)

3. Canonical form of the linear model. In the linear model, with i.i.d. normally distributed errors, denote by $\boldsymbol{\xi}_{n\times 1}$ the mean vector $E[\mathbf{Y}]$. The model specifies that $\boldsymbol{\xi} \in \boldsymbol{\Pi}$, a $p-$dimensional vector space in $\mathbb{R}^n$. Consider testing the hypothesis
$$H\text{: } \boldsymbol{\xi} \in \boldsymbol{\Pi}_0, \text{ a subspace of } \boldsymbol{\Pi} \text{ of dimension } q < p.$$
To put this into "canonical" form, one starts with an $n \times q$ matrix $\mathbf{Q}_1$ whose columns form an orthonormal basis for $\boldsymbol{\Pi}_0$. By appending $p - q$ independent vectors from $\boldsymbol{\Pi}$ which are not in $\boldsymbol{\Pi}_0$, and applying Gram-Schmidt again, one extends this to an $n \times p$ matrix $\left(\mathbf{Q}_1 \vdots \mathbf{Q}_2\right)$ whose columns form an orthonormal basis for $\boldsymbol{\Pi}$. Finally, as in the application of Gram-Schmidt considered in class, one obtains an orthogonal matrix $\mathbf{Q} = \left(\mathbf{Q}_1 \vdots \mathbf{Q}_2 \vdots \mathbf{Q}_3\right)$. Now define
$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_1'\mathbf{y} \\ \mathbf{Q}_2'\mathbf{y} \\ \mathbf{Q}_3'\mathbf{y} \end{pmatrix} = \mathbf{Q}'\mathbf{y} \text{ and } \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_1'\boldsymbol{\xi} \\ \mathbf{Q}_2'\boldsymbol{\xi} \\ \mathbf{Q}_3'\boldsymbol{\xi} \end{pmatrix} = \mathbf{Q}'\boldsymbol{\xi}.$$

(a) Show that in the unrestricted model, we are observing independent random variables $Z_1, ..., Z_n$ where $Z_i \sim N(\eta_i, \sigma_\varepsilon^2)$ for $1 \leq i \leq p$ and $Z_i \sim N(0, \sigma^2)$ for $p+1 \leq i \leq n$. Then show that the hypothesis is *H:* $\eta_{q+1} = ... = \eta_p = 0$.

(b) Show that, in this notation, the F-statistic resulting from the likelihood ratio test is

$$F = \frac{\|\mathbf{z}_2\|^2 / (p-q)}{\|\mathbf{z}_3\|^2 / (n-p)};$$

conclude from this that $F \sim F_{n-p}^{p-q}$ under H. [For this you will first want to derive that the likelihood function is $L(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \sigma_\varepsilon^2) = (2\pi\sigma_\varepsilon^2)^{-n/2} e^{-\frac{\|\mathbf{z}_1 - \boldsymbol{\eta}_1\|^2 + \|\mathbf{z}_2 - \boldsymbol{\eta}_2\|^2 + \|\mathbf{z}_3\|^2}{2\sigma_\varepsilon^2}}$, and then maximize $\log L$ with and without the restrictions dictated by the hypothesis. Then look at the likelihood ratio.]

(c) Suppose that H is false. Define $\boldsymbol{\xi}_0$ to be the closest vector, in $\boldsymbol{\Pi}_0$, to $\boldsymbol{\xi}$. Show that

(i) $||\boldsymbol{\xi} - \boldsymbol{\xi}_0||^2 = \min_{\mathbf{t} \in \boldsymbol{\Pi}_0} ||\boldsymbol{\xi} - \mathbf{t}||^2 = ||\boldsymbol{\eta}_2||^2$;

(ii) $E\left[\frac{\|\mathbf{z}_3\|^2}{\sigma_\varepsilon^2 (n-p)}\right] = 1$ but $E\left[\frac{\|\mathbf{z}_2\|^2}{\sigma_\varepsilon^2 (p-q)}\right] = 1 + \frac{\delta^2}{p-q} > 1$; here δ^2 is the 'non-centrality parameter' defined by $\delta^2 = \frac{\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|^2}{\sigma_\varepsilon^2}$.
Although you are not being asked to do so, it is an easy matter to show that when the null hypothesis is false, so that $\boldsymbol{\eta}_2 \neq \mathbf{0}$, then the power of the test is an increasing function of the non-centrality parameter δ^2. In this sense the probability of rejection increases as the null hypothesis becomes 'less true' - a property that we should expect of any reasonable test.

4. Assume the usual linear model.

(a) Show that the ridge estimator $\hat{\boldsymbol{\theta}}_R$ is the solution to the problem:

$$\min_{\theta} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||^2 + k||\boldsymbol{\theta}||^2 \right\}.$$

Use this to explain why $\hat{\boldsymbol{\theta}}_R$ is sometimes called a *shrinkage* estimator.

(b) Show that, for all sufficiently small values of the biasing constant k, the MSE of $\hat{\boldsymbol{\theta}}_R$ is less than that of $\hat{\boldsymbol{\theta}}_{OLS}$. (After verifying that the MSE is as given in class, it will be helpful to apply the spectral decomposition to $\mathbf{X}'\mathbf{X}$, as in question 5 below.)

5. An alternative to ridge regression is *Principal Components Regression.* Take the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. Decompose $\mathbf{X}'\mathbf{X}$ as $\mathbf{X}'\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$, where $\boldsymbol{\Gamma}$ is orthogonal and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues, in decreasing order. Put $\boldsymbol{\alpha} = \boldsymbol{\Gamma}'\boldsymbol{\theta}$ and $\mathbf{Z} = \mathbf{X}\boldsymbol{\Gamma}$, so that the model becomes $\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$. The columns of $\mathbf{Z}$ are known as the *principal components.*

(a) Verify that the covariance matrix of the OLSE $\hat{\boldsymbol{\alpha}}$ is $\sigma_\varepsilon^2\boldsymbol{\Lambda}^{-1}$ and that the i^{th} principal component has Euclidean norm $\sqrt{\lambda_i}$. This can be thought of as justifying the dropping of the variables (in the $(\mathbf{Z},\boldsymbol{\alpha})$ parameterization) corresponding to the smallest of the λ_i.

(b) Now consider the acetylene data discussed in class. First make the correlation transform so that $\mathbf{y}$ and each of the nine columns of $\mathbf{X}$ have unit length. Four of the nine eigenvalues of $\mathbf{X}'\mathbf{X}$ are very small, with condition numbers exceeding 1,000. Define $\hat{\boldsymbol{\alpha}}_{OLS}(j)$ to be the OLS estimator after the elimination of the principal components corresponding to the j^{th} smallest eigenvalues. The Principal Component Estimate is then $\hat{\boldsymbol{\alpha}}_{PC}(j) = \left(\hat{\boldsymbol{\alpha}}'_{OLS}(j), \mathbf{0}'_j\right)'$. Determine $\hat{\boldsymbol{\alpha}}_{PC}(j)$ for $j = 1, ..., 4$. Compare the corresponding estimates $\hat{\boldsymbol{\theta}}(j)$ $(= \boldsymbol{\Gamma}\hat{\boldsymbol{\alpha}}_{PC}(j))$ with the ridge estimate (still without an intercept) using $k = .03$.

6. Write an R function to carry out the test for Lack of Fit. The first line should be something like lof = function(x, y, ...) where x is the usual design matrix without a column of ones, y is the data vector, and ... represents any other input you think will be useful. The value of the function should be a matrix (with the rows and columns labelled appropriately) consisting of the sums of squares, degrees of freedom and mean squares for Lack of Fit and for Pure Error, and the p-value of the test (under normality). Plots of the data versus the regressors, with the fitted line superimposed, should be included as well. Apply your function to the PCB data (available from the course website) in two cases - $y = PCB$, $x = age$ and $y = \ln(PCB)$, $x = \sqrt[3]{age}$. Comment on the results. An example of R code that will prepare a matrix of output, once the inputs are computed, is:

```
out = matrix(nrow=3,ncol=5)
dimnames(out) = list(c("Pure Error", "Lack of Fit", "Error"),
c("Sum of Squares", "df", "Mean Square", "F", "p"))
out[1,] = c(SSPE, df.SSPE, MSPE, NA, NA)
out[2,] = c(SSLOF, df.SSLOF, MSLOF, round(F.lof,3), round(p.lof,3))
out[3,] = c(SSE, df.SSE, MSE, NA, NA)
```

7. (a) Show that $(n-p-1)\,S^2_{(i)} = (n-p)\,S^2 - \frac{e_i^2}{1-h_{ii}}$, where $S^2_{(i)}$ is the estimate of σ_ε^2 computed from a sample after deleting the i^{th} case.

(b) The “stackloss” data set contains figures from 21 days of operation of a plant oxidizing ammonia to nitric acid. A complete description can be obtained by entering `help(stackloss)` in R. The data are called `stack.loss` (the dependent variable) and `stack.x` (the matrix of independent variables). Determine which, if any, observations are highly influential, or outlying, on the basis of any of (i) a high leverage, (ii) a large studentized residual, (iii) a large deleted fit value, or (iv) a large value of Cook’s statistic. Prepare plots of the studentized residuals against the fitted values, and against each independent variable, with the indices of the flagged observations displayed. What seems to make these cases aberrant? An example of R code that will do the plotting (if the relevant indices are in "flagged") is:

```
fits = predict(fit)
plot(fits, stud.res)
points(x = fits[flagged], y = stud.res[flagged], pch=19)
text(x = fits[flagged], y = stud.res[flagged], labels = flagged, pos=4)
```

8. Recall the example of logistic regression discussed in class. Re-analyze these data, but now fit a *probit* model, in which the logistic transformation $L^{-1}(\pi)$ is replaced by the probit transformation $\Phi^{-1}(\pi)$ (Φ is the $N(0,1)$ d.f.). Compare the two fits.