Stats305a Problem Set 1
Due: Tuesday, October 26 by 11:59pm on Gradescope.

**Question 2.1** (Regression with random samples and best linear approximations): Say that \((X, Y) \in \mathbb{R}^d \times \mathbb{R}\) come from a joint probability distribution \(P\), where \(\mathbb{E}[XX^T] = C > 0\) and \(X, Y\) both have finite second moments. Assume that the first coordinate of \(X\) is constant, with \(X_1 = 1\). For \(x \in \mathbb{R}^d\), define the regression function

\[
f^*(x) := \mathbb{E}[Y \mid X = x].
\]

(a) Show that \(f\) minimizes \(\mathbb{E}[(Y - f(x))^2]\) over all functions \(f : \mathbb{R}^d \to \mathbb{R}\).

(b) Instead of fitting a model of \(Y \mid X\) over the space of all functions, consider fitting one over all linear predictors, and choosing \(\beta^*\) to minimize the expected squared loss

\[
L(b) := \frac{1}{2} \mathbb{E}[(Y - X^T b)^2]
\]

over \(b \in \mathbb{R}^d\). Characterize the solution \(\beta^*\) (i.e., give a formula for it), and show that the linear function \(\varphi(x) = \beta^*^T x\) is the best linear approximation to \(f\) (in mean-squared distance). *Note:* in case you are worried about it, it is fine to exchange expectation and differentiation in this case; you definitely don’t need to show that, though it is true (see, for example, Bertsekas [1]).

(c) Say that we have an i.i.d. sample \((x_i, y_i), i = 1, \ldots, n\) from \(P\), with \(y_i = f(x_i) + \varepsilon_i\), and

\[
\hat{\beta} = \arg\min_b L_n(b) := \frac{1}{2n} \sum_{i=1}^n (x_i^T b - y_i)^2
\]

is the ordinary least-squares estimator, and assume \(n \geq d\). Is \(\hat{\beta}\) unbiased for \(\beta^*\)?

**Question 2.2** (T statistics, F statistics, and linear algebra): Consider the model \(y = X\beta + \varepsilon\), \(\varepsilon \sim \mathcal{N}(0, \sigma^2 I)\), \(X \in \mathbb{R}^{n \times d}\) with rank \(d\). The t-statistic for a coordinate \(j\) is

\[
t_j = \frac{\hat{\beta}_j}{\hat{s}\hat{e}(\hat{\beta})},
\]

where \(\hat{s}\hat{e} = \hat{\sigma}\sqrt{\hat{e}_j^T (X^T X)^{-1} \hat{e}_j}\) is the usual standard error estimate for \(\hat{\beta}_j\). For example, \(\texttt{R}\) reports p-values for these t-statistics when using \texttt{lm} and \texttt{summary}. Let \(X\) have columns \(x^{(j)}\), \(j = 1, \ldots, d\), and \(X_{\setminus j}\) be \(X\) with column \(j\) removed, i.e.

\[
X_{\setminus j} = \begin{bmatrix} x^{(1)} & \cdots & x^{(j-1)} & x^{(j+1)} & \cdots & x^{(d)} \end{bmatrix},
\]

which is a submodel as we have discussed in class. The F-statistic for coordinate \(j\) is then

\[
F_j = \frac{\|(H - H_j)y\|_2^2}{\frac{1}{n-d} \|(I - H)y\|_2^2},
\]

where \(H = X(X^T X)^{-1}X^T\) is the usual hat matrix (projection onto range(\(X\))) and \(H_j\) is the projection matrix onto range(\(X_{\setminus j}\)). Show that \(t_j^2 = F_j\).
**Hint.** Assume without loss of generality that $j = d$, the $d$th component. (One can do so by permuting the columns of $X$.) Consider the QR factorization of $X$, i.e., $X = QR$ where $Q \in \mathbb{R}^{n \times n}$ is orthogonal and $R \in \mathbb{R}^{n \times d}$ has the form

$$R = \begin{bmatrix} T & \mathbf{0}_{n-d\times d} \end{bmatrix}$$

for an upper triangular (invertible) matrix $T$ with entries $T_{ij} = R_{ij}$ for all $1 \leq i, j \leq d$.

**Question 2.3** (Non-independent noise and testing challenges): A subtle but problematic situation occurs in linear models when noise is correlated instead of independent—indeed, this is often much worse than non-normality of noise, which the central limit theorem more or less addresses. To make this a bit more concrete, we consider a 2-group ANOVA model,

$$y_{1j} = \mu + \alpha_1 + \varepsilon_{1j}, \quad y_{2j} = \mu + \alpha_2 + \varepsilon_{2j},$$

where we assume we observe a sample of size $n$ for each group (i.e. $j = 1, \ldots, n$). The standard assumption is that $\varepsilon_i \sim N(0, \sigma^2 I_n)$, where we use $\varepsilon_i = [\varepsilon_{i1} \cdots \varepsilon_{in}]^T \in \mathbb{R}^n$ for shorthand, and we have the null model

$$H_0 : \alpha_1 = \alpha_2, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

In this case, for $\overline{y}_i = \frac{1}{n} \sum_{j=1}^{n} y_{ij}$ and standard error estimate

$$S_n^2 := \frac{1}{2n-2} \left[ \sum_{j=1}^{n} (y_{1j} - \overline{y}_1)^2 + \sum_{j=1}^{n} (y_{2j} - \overline{y}_2)^2 \right],$$

the usual t-statistic is

$$t := \frac{\overline{y}_1 - \overline{y}_2}{S_n \sqrt{2/n}} \sim T_{2n-2},$$

the t-distribution with $2n - 2$ degrees of freedom, or equivalently,

$$\tilde{F} := \frac{n \left( \overline{y}_1 - \overline{y}_2 \right)^2}{S_n^2} \sim F_{1,2n-2}.$$

We will show that a test that rejects when $F$ is large (i.e., the standard ANOVA) may reject unrealistically frequently when the errors are correlated.

To that end, consider the situation that $\varepsilon_1, \varepsilon_2$ are independent, but

$$\varepsilon_i \sim N\left(0, \sigma^2 (1 - \rho) I_n + \sigma^2 \rho \mathbf{1}\mathbf{1}^T\right),$$

where $\rho \in [0, 1]$ indicates correlation within the group. Such correlation may be reasonable, e.g., when (hidden) confounding relates members of a group. Through the remainder of this question, let $C_\rho = (1 - \rho) I_n + \rho \mathbf{1}\mathbf{1}^T$ be a shorthand for the covariance.

(a) Show that if $Z \sim N(0, C_\rho)$, then $(I - \frac{1}{n} \mathbf{1}\mathbf{1}^T)Z = Z - \mathbf{1}\mathbf{Z}_n \sim N(0, (1 - \rho) (I_n - \frac{1}{n} \mathbf{1}\mathbf{1}^T))$.

(b) Show that $\frac{1}{\sqrt{1-\rho}} S_n^2 \sim \chi^2_{2n-2} / \chi^2_{2n-2}$ under the correlation structure (2.2).

(c) Show that if $\alpha_1 = \alpha_2$ and the correlation (2.2) holds,

$$\overline{y}_1 - \overline{y}_2 \sim N\left(0, \sigma^2 \left(2 \frac{1-\rho}{n} + 2\rho\right)\right).$$
(d) Argue that \( \bar{y}_1 - \bar{y}_2 \) is independent of \( S_n^2 \) even with correlation (2.2).

(e) Show that under correlation (2.2),

\[
\hat{F}_\rho := \frac{1 - \rho}{1 - \rho + \rho n} \cdot \hat{F} = \frac{n}{2(1 - \rho) + 2\rho n} (\bar{y}_1 - \bar{y}_2)^2 \sim F_{1, 2n-2},
\]

so that the valid test is to reject when \( \hat{F}_\rho \) is large.

(f) Argue that as \( n \) grows large, the standard ANOVA will falsely reject the null hypothesis \( H_0 \) too frequently when the correlation (2.2) holds.

**Question 2.4 (Intuition for correlated rejections via simulation):** Here we revisit question 2.3, except that we perform some simulations and corrections. First, we describe a general strategy for eliminating correlations in the noise, making the prima facie ridiculous assumption that we know the noise covariance.

(a) Let \( y = X\beta + \varepsilon \) where \( \varepsilon \sim (0, \Sigma) \). Show that \( \Sigma^{-1/2} y = \Sigma^{-1/2} X\beta + \xi \), where \( \xi \sim (0, I_n) \).

By part (a), if we knew \( \Sigma \), we could make the substitutions

\[
\tilde{y} = \Sigma^{-1/2} y, \quad \tilde{X} = \Sigma^{-1/2} X
\]

and perform ordinary least squares on \((\tilde{X}, \tilde{y})\); all of the distributional results and tests we have developed would then work.

(b) Repeat the following experiment several (say, 100) times for values of \( n = 2, 4, 8, 16, 32, 64, 128, 256, 512 \). Generate data from the ANOVA model (2.1), except that the noise is correlated (2.2) with \( \rho = .1 \), with \( \mu = \alpha_1 = \alpha_2 = 0 \). Perform an F-test of significance at level \( \alpha = .05 \), rejecting when \( \hat{F} \) is large. (As the null \( \alpha_1 = \alpha_2 \) holds, any rejections of equality of means is false, though a rejection of the ANOVA model with independent noise is sensible.) Plot the frequency of false rejections against sample size \( n \).

It is of interest to correct an estimate for possible correlations, thereby achieving a test whose nominal level is closer to accurate. In general, one never has enough data to estimate \( \text{Cov}(\varepsilon) \) in a linear regression model except under assumptions on the noise model. In the ANOVA model (2.1), it may be reasonable to assume that within a group, the noises are all equally correlated, that is, the noise model (2.2) holds, and we can approximate \( \text{Cov}(\varepsilon_i) \).

Note that \( \mathbb{E}[(y_{1j} - y_{2l})^2] = 2\sigma^2 \) and that \( y_1 \perp \perp y_2 \) under model (2.2) and the null \( \alpha_1 = \alpha_2 \). Define the estimates

\[
\hat{\sigma}^2 := \frac{1}{2n^2} \sum_{j,l} (y_{1j} - y_{2l})^2 \quad \text{and} \quad \hat{\rho} = 1 - \frac{S_n^2}{\hat{\sigma}^2},
\]

so that \( \hat{\rho} \to \rho \) as \( n \to \infty \) (you do not need to show this! We are simply asserting it). Then the plug-in test uses the statistic \( \hat{F}_\rho \) from Question 2.3, except we replace \( \rho \) with \( \hat{\rho} \).

(c) Repeat your experiment from part (b), except that you use the statistic \( \hat{F}_{\hat{\rho}} \) in place of \( \hat{F} \). Plot your frequency of false rejections against sample size \( n \). *Hint.* You should truncate \( \hat{\rho} \) so that \( \hat{\rho} \geq 0 \), as this will keep the problem better-conditioned.
Question 2.5 (Clumpy testing errors): In the data file abalone.data we have data on abalone (a type of mollusc) age, where the dataset is explained in file abalone.names. The goal is to predict the age of an abalone (given by the count of rings in its shell) from other characteristics. Here we use this dataset to investigate false discoveries and whether they come alone or in groups by adding additional complete noise variables to the data, then regressing a linear model including these noise variables.

Write code to perform the following: first, load the abalone data. Then

i. Add two columns (call them $x^{(1)}$ and $x^{(2)}$, say) to the data, where their entries are i.i.d.

$$
\begin{bmatrix}
x_i^{(1)} \\
x_i^{(2)}
\end{bmatrix} \sim N \left( 0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right),
$$

that is, i.i.d. normal random variables with correlation $\rho \in (-1, 1)$.

ii. Fit a linear model for the response $y = \text{Rings}$ against all other variables (including the noise variables $x^{(1)}$ and $x^{(2)}$).

iii. Perform a t-test for association of variable $x^{(1)}$ (adjusting for all other variables) and $x^{(2)}$ (again, adjusting for all other variables) with $y$, rejecting at the level $\alpha = .05$.

For the values $\rho \in \{-.9, -.8, - .4, 0, .4, .8, .9\}$, repeat the experiment in steps i–iii $N = 1000$ times. Across the experiments, record the number of times there is a false discovery of $x^{(1)}$, a false discovery for $x^{(2)}$, and a false discovery of both simultaneously. Report your false discovery counts and describe them. (Include your code in your solution.)

Hints and pointers. You will want to represent the abalone’s sex as a factor, that is, instead of the raw character M, F, or I (infant), represent it in a 0-1 encoding over 3 levels. That is, if $S \in \{M, F, I\}$ represents the sex of the abalone, transform it into

$$
\phi(S) = \begin{bmatrix} 1 \{S = M\} \\ 1 \{S = F\} \\ 1 \{S = I\} \end{bmatrix} \in \{0,1\}^3.
$$

In R this is achieved by using the method factor. Also, the t-test in step iii is simply the standard t-test we have developed in class and is that performed by R’s summary method of a linear model.

References