

Homework 5, Stat 541: Due Monday, Nov 7, 2011, 3pm

Linear Models and Normal Distributions

Student Name: (replace this with your name)

November 9, 2011

Instructions: Edit this LaTeX file by inserting your solutions after each problem statement. Generate a PDF file from it and e-mail the PDF in an attachment with filename "**hw05-Yourlastname-Yourfirstname.pdf**" to the usual class gmail address with "**Homework 5, 2011**" in the subject line:

stat541.at.wharton@gmail.com

You should not just answer the questions but give at least partial proofs where appropriate. You can discuss the homework with each other in general terms, but not with previous years' students of Stat 541. You must not consult solutions of similar homeworks of previous years. Finally, you must write your own solutions and not copy from anyone. Verbatim copying from others or unlisted sources, no matter how minor, will result in zero points for the whole homework. Report here who you collaborated with and what sources you used:

My collaborators: ... (replace this)

The complete list of my sources is as follows: ... (replace this)

The Problems

1. Let $\mathbf{X} = (X_1, X_2)^T$ and assume $\mathbf{E}[\mathbf{X}] = (\mu_1, \mu_2)^T$ and $\mathbf{C}[\mathbf{X}] = \sigma^2 \mathbf{I}_2$. Under what conditions are $Y_1 = \cos(\alpha)X_1 + \sin(\alpha)X_2$ and $Y_2 = \cos(\beta)X_1 + \sin(\beta)X_2$ (a) uncorrelated, (b) independent?

Answer: When calculating $\text{Cov}(Y_1, Y_2)$, the covariance term $\text{Cov}(X_1, X_2)$ disappears and

only the variance terms $\text{Var}[X_1] = \text{V}[X_2] = \sigma^2$ survive:

$$\text{Cov}(Y_1, Y_2) = (\cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta))\sigma^2 = \cos(\alpha - \beta)\sigma^2$$

The variables Y_1, Y_2 are uncorrelated if $\alpha - \beta = \pi/2 + \pi k$ for an arbitrary integer k , that is, if the coefficient unit vectors $(\cos(\alpha), \sin(\alpha))$ and $(\cos(\beta), \sin(\beta))$ are orthogonal. They are independent if in addition they are normally distributed.

2. If $(X_1, X_2) \sim \mathcal{N}((\mu, \mu), \sigma^2 I_2)$, what is the distribution of $(X_1 + X_2 - 2\mu)/|X_1 - X_2|$?

Answer: This statistic has a t_1 distribution or, equivalently, a Cauchy distribution. Reason: t_p is defined as the distribution of $Z_0/\sqrt{(Z_1^2 + \dots + Z_p^2)/p}$, where $Z_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. In the present situation we have: $Z_0 = (X_1 + X_2 - 2\mu) \sim \mathcal{N}(0, \sigma^2)$, $Z_1 = (X_1 - X_2) \sim \mathcal{N}(0, \sigma^2)$, they are independent due to the orthogonality of $(1, 1)$ and $(1, -1)$, and $\sqrt{Z_1^2/1} = |X_1 - X_2|$, which proves that the ratio has a t_1 distribution.

3. Let X_1, \dots, X_M and Y_1, \dots, Y_N be measurements that are all taken independently of each other. Make no other assumptions yet: $\mathbf{E}[X_i] = \mu_{i,X}$, $\mathbf{E}[Y_j] = \mu_{j,Y}$, $\text{Var}[X_i] = \sigma_{i,X}^2$, $\text{Var}[Y_j] = \sigma_{j,Y}^2$. Under what conditions on the means μ_{X_i} , μ_{Y_j} and the variances $\sigma_{i,X}^2$, $\sigma_{j,Y}^2$ are the grand mean $(X_1 + \dots + X_M + Y_1 + \dots + Y_N)/(M + N)$ and the mean difference $(X_1 + \dots + X_M)/M - (Y_1 + \dots + Y_N)/N$ uncorrelated?

Use the abbreviations $\bar{\mu}_X = \sum_{i=1}^M \mu_{i,X}/M$, $\bar{\mu}_Y = \sum_{j=1}^N \mu_{j,Y}/N$, $\bar{\sigma}_X^2 = \sum_{i=1}^M \sigma_{i,X}^2/M$, and $\bar{\sigma}_Y^2 = \sum_{j=1}^N \sigma_{j,Y}^2/N$, if convenient.

Answer: The stated independence assumption implies that the covariance terms $\text{Cov}(X_i, X_j)$ ($i \neq j$), $\text{Cov}(Y_k, Y_l)$ ($k \neq l$), and $\text{Cov}(X_i, Y_k)$ disappear. It follows that, when calculating the covariance between the grand mean and the mean difference, only the “diagonal” terms survive, that is, the terms $\text{Cov}(X_i/(M + N), X_i/M) = (1/(M + N))(1/M)\sigma_{i,X}^2$ and $\text{Cov}(Y_j/(M + N), -Y_j/N) = -(1/(M + N))(1/N)\sigma_{j,Y}^2$:

$$\begin{aligned} & \text{Cov}\left(\frac{1}{M + N}(X_1 + \dots + X_M + Y_1 + \dots + Y_N), \frac{1}{M}(X_1 + \dots + X_M) - \frac{1}{N}(Y_1 + \dots + Y_N)\right) \\ &= \frac{1}{(M + N)M} \sum_{i=1}^M \sigma_{i,X}^2 - \frac{1}{(M + N)N} \sum_{j=1}^N \sigma_{j,Y}^2 \\ &= \frac{1}{(M + N)} (\bar{\sigma}_X^2 - \bar{\sigma}_Y^2) \end{aligned}$$

Therefore, the grand mean and the mean difference are uncorrelated iff $\bar{\sigma}_X^2 = \bar{\sigma}_Y^2$. The expectations of the variables are irrelevant for the question and can be arbitrary.

4. Making the usual normal assumptions in a linear model, is the IQR (= InterQuartile Range) of the residuals independent of the estimated regression coefficients or not?

Answer: Yes, it is: Residuals and coefficient estimates are independent of each other because they are computed from linear combinations of \mathbf{y} with orthogonal coefficient vectors. Using the assumptions $\mathcal{N}(\dots, \sigma^2 \mathbf{I})$ we obtain not only vanishing correlations but also independence.

The second step of the argument is to state that any functions $f(\mathbf{r})$ and $g(\mathbf{b})$ must also be stochastically independent. In our particular case $f(\mathbf{r}) = \text{IQR}(\mathbf{r})$ and $g(\mathbf{b}) = \mathbf{b}$. QED

5. Argue as best as you can that the distribution of $(b_j - \beta_j)/\text{IQR}(\mathbf{r})$ is free of β_k and σ (that is, it is a so-called “pivotal” statistic).

Answer: Two stages:

- Distribution of the numerator: $(b_j - \beta_j) \sim \mathcal{N}(0, \sigma^2 / \|\mathbf{x}_{j\cdot\text{adj}}\|^2)$ (equivalently $(b_j - \beta_j) \sim \mathcal{N}(0, \sigma^2 (X^T X)_{j,j}^{-1})$), hence $(b_j - \beta_j)/\sigma \sim \mathcal{N}(0, 1/\|\mathbf{x}_{j\cdot\text{adj}}\|^2)$. This is a single distribution that does not involve β or σ .
- Distribution of the denominator: $\mathbf{r} \sim \mathcal{N}(0, (I - P)\sigma^2)$, hence $\mathbf{r}/\sigma \sim \mathcal{N}(0, I - P)$. Furthermore observe that $\text{IQR}(\mathbf{r}/\sigma) = \text{IQR}(\mathbf{r})/\sigma$ since IQR is a measure of dispersion. It follows that $\text{IQR}(\mathbf{r})/\sigma$ has a single distribution that does not involve β or σ .
- Joint distribution of numerator and denominator: In the previous problem we found that \mathbf{b} and $\text{IQR}(\mathbf{r})$, and hence $(b_j - \beta_j)/\sigma$ and $\text{IQR}(\mathbf{r})/\sigma$, are stochastically independent under the usual normal linear models assumptions. Since the two components have distributions that do not involve any β_k or σ , it follows that the joint distribution does not either.
- Distribution of the ratio: Forming the ratio of $(b_j - \beta_j)/\sigma$ and $\text{IQR}(\mathbf{r})/\sigma$ produces a random variable that does not involve any β_k or σ either.

[Note that you really do need independence of numerator and denominator, otherwise it could be the case that $(b_j - \beta_j)/\sigma$ and $\text{IQR}(\mathbf{r})/\sigma$ are correlated or otherwise dependent in

a way that is a function of β and/or σ even though their marginal distributions aren't.]

6. The concept of conditioning a distribution on a given statistic: Literally, this means holding the statistic fixed and sampling from the joint distribution under this constraint.

We are not into calculating conditional densities and conditional distributions, but we are into sampling from conditional distributions. Operationalizing a conditional distribution in terms of a sampling scheme is often quite different and sometimes simpler than writing down analytical formulas for densities.

You can learn some fundamentals without theory, just with formal intuitions and some simple principles. Here is a most fundamental example: if X and Y are independent, then the conditional distribution of (X, Y) given $X = x$ is obtained by sampling Y and forming pairs (x, Y) , where x is the fixed value we condition X on. You should be able to grasp the triviality of this example, but you may not fully grasp its reach. Why? Because this situation can often be forced by transforming the original situation to one that is really just this example.

- (a) Leading Example: Assume a normal linear model, $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_N)$, where we know β and σ . We know how to sample/simulate \mathbf{y} in the R language:

```
rnorm(N, m = X%*%beta, s = sigma)
```

However, in class we will need to be able to sample from the conditional distribution of \mathbf{y} given $\hat{\mathbf{y}}$!!! How shall we go about this? Consider: $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{r}$. What do you know about the joint distribution of $\hat{\mathbf{y}}$ and \mathbf{r} ? Figure the rest. Explain and create R code. It should have the form

`yhat + ...`

because you are supposed to condition on $\hat{\mathbf{y}}$.

Answer: We know that $\hat{\mathbf{y}}$ and \mathbf{r} are independent. Hence for conditioning on $\hat{\mathbf{y}}$ we can simulate \mathbf{r} independently, and then we simply add it to the given $\hat{\mathbf{y}}$. What is the distribution of \mathbf{r} ? We know it is $\mathcal{N}(\mathbf{0}, (\mathbf{I} - \mathbf{H})\sigma^2)$. But this is the distribution of $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$ mapped with $\mathbf{I} - \mathbf{H}$; in other words, draw y -vectors from $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$, regress

on \mathbf{X} and pull out the residuals. Hence the R code can be as simple as this:

```
yhat + resid(lm(rnorm(N, m = 0, s = sigma) ~ X))
```

- (b) Specialize the previous to the simplest of all linear models: $\mathbf{y} \sim \mathcal{N}(\mu\mathbf{e}, \mathbf{I}\sigma^2)$, that is, y_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. What is the conditional distribution of \mathbf{y} given $\text{mean}(\mathbf{y}) = M$? Explain and give R code.

Answer: The decomposition is $\mathbf{y} = \text{mean}(\mathbf{y}) + (\mathbf{y} - \text{mean}(\mathbf{y}))$ (where adding a scalar to a vector means adding the scalar to all components of the vector, like in R). The mean is independent of the residual vector $(\mathbf{y} - \text{mean}(\mathbf{y})) \sim \mathcal{N}(\mathbf{0}, (\mathbf{I} - \mathbf{e}\mathbf{e}^T/N)\sigma^2)$. Hence:

```
M + scale(rnorm(N, m = 0, s = sigma), center = T, scale = F)
```

or

```
tmp <- rnorm(N, m = 0, s = sigma); M + (tmp - mean(tmp))
```