Conditional distribution for Multivariate Gaussian

A. Rafiq

Consider the p-dimensional multivariate normal random variable $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Partition $\mathbf{Y}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \boldsymbol{\Lambda} = \Sigma^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}$$

where $\mathbf{Y}_1$ and $\boldsymbol{\mu}_1$ have length p_1, $\boldsymbol{\Sigma}_{11}$ has dimension $p_1 \times p_1$ and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^T$.
Show that the conditional distribution of $\mathbf{Y}_2$, given $\mathbf{Y}_1 = \mathbf{y}_1$ can be written as

$$\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1 \sim N_{p_2}(\boldsymbol{\mu}_2 - \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(\mathbf{y}_{-}\boldsymbol{\mu}_1),\ \boldsymbol{\Lambda}_{22}^{-1}),$$

where $p_2 = p - p_1$. *Hint: start from the definition* $p(\mathbf{Y}_2|\mathbf{Y}_1) = \frac{P(\mathbf{Y}_1, \mathbf{Y}_2)}{p(\mathbf{Y}_1)}$. *Discard all the terms that do not depend on* $\mathbf{Y}_2$ *and show that the remaining expression is the exponential of a quadratic form with the appropriate mean and precision matrix.*

The marginal density of $Y_2|Y_1$ is given by:

$$p(y_2|y_1) = \frac{p(y_1, y_2)}{p(y_1)} \tag{1}$$

Here the joint distribution $p(y_1, y_2) = p(\mathbf{y})$, which is given by:

$$\begin{aligned} p(\mathbf{y}) &= \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)} && (2) \\ &\propto e^{-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)} && (3) \end{aligned}$$

We shall focus on the terms in the exponent and to simplify things will use $\delta\mathbf{y} = (\mathbf{y} - \mu) = \begin{bmatrix} \delta\mathbf{y}_1 \\ \delta\mathbf{y}_2 \end{bmatrix}$ and will use $\Lambda = \Sigma^{-1}$.

$$\begin{aligned} -\frac{1}{2}(\mathbf{y} - \mu)^T\Sigma^{-1}(\mathbf{y} - \mu) &= -\frac{1}{2}\delta\mathbf{y}^T\Lambda\delta\mathbf{y} && (4) \\ &= -\frac{1}{2}\begin{bmatrix} \delta\mathbf{y}_1^T & \delta\mathbf{y}_2^T \end{bmatrix} \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} \delta\mathbf{y}_1 \\ \delta\mathbf{y}_2 \end{bmatrix} && (5) \\ &= -\frac{1}{2}\left[\delta\mathbf{y}_1^T\Lambda_{11}\delta\mathbf{y}_1 + \delta\mathbf{y}_1^T\Lambda_{12}\delta\mathbf{y}_2 + \underbrace{\delta\mathbf{y}_2^T\Lambda_{21}\delta\mathbf{y}_1}_{=\delta\mathbf{y}_1^T\Lambda_{12}\delta\mathbf{y}_2} + \delta\mathbf{y}_2^T\Lambda_{22}\delta\mathbf{y}_2\right] && (6) \\ &= -\frac{1}{2}\delta\mathbf{y}_1^T\Lambda_{11}\delta\mathbf{y}_1 - \frac{1}{2}\delta\mathbf{y}_2^T\Lambda_{22}\delta\mathbf{y}_2 - \delta\mathbf{y}_1^T\Lambda_{12}\delta\mathbf{y}_2 && (7) \end{aligned}$$

Note that on line 6 of the equation, we use the property of of scalar results from matrix multiplication in order to equate $\delta\mathbf{y}_2^T{}_{21}\Lambda\delta\mathbf{y}_1 = \delta\mathbf{y}_1^T{}_{12}\Lambda\delta\mathbf{y}_2$. At this stage, we need to **complete the square**.

Theorem 1 (Completing the Square)**.** *For symmetric* $A \in \mathbb{R}^{NxN}$ *and* $\mathbf{x}, \mathbf{b} \in \mathbb{R}^N$*:*

$$\frac{1}{2}\mathbf{x}^T A\mathbf{x} + 2\mathbf{b}^T\mathbf{x} = \frac{1}{2}(\mathbf{x} + A^{-1}\mathbf{b})^T A(\mathbf{x} + A^{-1}\mathbf{b}) - \frac{1}{2}\mathbf{b}^T A^{-1}\mathbf{b} \tag{8}$$

Proof. For scalars we have the following formula for completing the square:

$$\frac{a}{2}x^2 + bx = \frac{a}{2}\left(x + \frac{b}{a}\right)^2 - \frac{b^2}{2a} \tag{9}$$

In higher dimensions, we want to get it into a similar form:

$$\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^t\mathbf{x} = \frac{1}{2}(\mathbf{x} + \mathbf{m})^T A(\mathbf{x} + \mathbf{m}) - \mathbf{c} \tag{10}$$

where $\mathbf{m}$ and $\mathbf{c}$ are vectors to be determined. By expanding the brackets on the left hand side:

$$\begin{aligned}\frac{1}{2}(\mathbf{x} + \mathbf{m})^T A(\mathbf{x} + \mathbf{m}) - \mathbf{c} &= \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \frac{1}{2}\mathbf{x}^T A\mathbf{m} + \frac{1}{2}\mathbf{m}^T A\mathbf{x} + \frac{1}{2}\mathbf{m}^T A\mathbf{m} - \mathbf{c} \\ &= \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \underbrace{\mathbf{m}^T A}_{=\mathbf{b}^T}\mathbf{x} + \underbrace{\frac{1}{2}\mathbf{m}^T A\mathbf{m} - \mathbf{c}}_{=0}\end{aligned}$$

Here we have $\mathbf{m}^T A = \mathbf{b}^T$ and $\frac{1}{2}\mathbf{m}^T A\mathbf{m} - \mathbf{c} = 0$. From the first equality, we see that $\mathbf{m} = A^{-1}\mathbf{b}$ and we substitute this into the latter to get:

$$\begin{aligned}\mathbf{c} &= \frac{1}{2}\mathbf{m}^T A\mathbf{m} \\ &= \frac{1}{2}\mathbf{b}^T A^{-1} A A^{-1}\mathbf{b} \\ &= \frac{1}{2}\mathbf{b}^T A^{-1}\mathbf{b}\end{aligned}$$

By substituting these back into equation (3) we obtain our desired solution. □

Now going back to our equation, we pick out the terms we want to use the completing the square method on:

$$\begin{aligned}&-\frac{1}{2}\delta\mathbf{y}_1^T\Lambda_{11}\delta\mathbf{y}_1 - \frac{1}{2}\underbrace{\delta\mathbf{y}_2^T}_{\mathbf{x}^T}\underbrace{\Lambda_{22}}_{A}\underbrace{\delta\mathbf{y}_2}_{\mathbf{x}} - \underbrace{\delta\mathbf{y}_1^T\Lambda_{12}}_{\mathbf{b}^T}\underbrace{\delta\mathbf{y}_2}_{\mathbf{x}} && (11)\\ &= -\frac{1}{2}\delta\mathbf{y}_1^T\Lambda_{11}\delta\mathbf{y}_1 - \frac{1}{2}(\delta\mathbf{y}_2 + \Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1)^T\Lambda_{22}(\delta\mathbf{y}_2 + \Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1) + \frac{1}{2}\delta\mathbf{y}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1 && (12)\\ &= -\frac{1}{2}\delta\mathbf{y}_1^T(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})\delta\mathbf{y}_1 - \frac{1}{2}(\delta\mathbf{y}_2 + \Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1)^T\Lambda_{22}(\delta\mathbf{y}_2 + \Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1) && (13)\end{aligned}$$

We can now put these terms into the exponential to get:

$$\begin{aligned}p(y_1, y_2) &\propto e^{-\frac{1}{2}\delta\mathbf{y}_1^T(\Lambda_{11}-\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})\delta\mathbf{y}_1 - \frac{1}{2}(\delta\mathbf{y}_2+\Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1)^T\Lambda_{22}(\delta\mathbf{y}_2+\Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1)} && (14)\\ &= e^{-\frac{1}{2}\delta\mathbf{y}_1^T(\Lambda_{11}-\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})\delta\mathbf{y}_1} e^{-\frac{1}{2}(\delta\mathbf{y}_2+\Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1)^T\Lambda_{22}(\delta\mathbf{y}_2+\Lambda_{22}^{-1}\Lambda_{21}\delta\mathbf{y}_1)} && (15)\\ &= \underbrace{e^{-\frac{1}{2}(\mathbf{y}_1-\boldsymbol{\mu_1})^T(\Lambda_{11}-\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})(\mathbf{y}_1-\boldsymbol{\mu_1})}}_{1}\underbrace{e^{-\frac{1}{2}((\mathbf{y}_2-\boldsymbol{\mu_2})+\Lambda_{22}^{-1}\Lambda_{21}(\mathbf{y}_1-\boldsymbol{\mu_1}))^T\Lambda_{22}((\mathbf{y}_2-\boldsymbol{\mu_2})+\Lambda_{22}^{-1}\Lambda_{21}(\mathbf{y}_1-\boldsymbol{\mu_1}))}}_{2} && (16)\end{aligned}$$

This is now the product of two Gaussian distributions:

1. In this section, we have $\mathbf{y_1} \sim \mathcal{N}_{p_1}(\boldsymbol{\mu_1}, (\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})^{-1})$.
2. In this section, we have $\mathbf{y_2}|\mathbf{y_1} \sim \mathcal{N}_{p_2}(\boldsymbol{\mu_2} - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{y}_1 - \boldsymbol{\mu_1}), \Lambda_{22}^{-1})$.

Since we have $p(y_2|y_1) = \frac{p(y_1,y_2)}{p(y_1)}$, we can drop the first term. Hence we get:

$$p(y_2|y_1) \sim \mathcal{N}_{p_2}(\mu_{\mathbf{2}} - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{y}_1 - \mu_{\mathbf{1}}), \Lambda_{22}^{-1}) \tag{17}$$

as required.

> Using the expression for the conditional distribution for multivariate Gaussian derived in the previous question and the block inverse formula above, prove that (using the same notation as before) the conditional distribution of $\mathbf{Y}_2$, given $\mathbf{Y}_1 = \mathbf{y}_1$, is
>
> $$\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1 \sim N_{p_2}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}),$$
>
> where $p_2 = p - p_1$.

in order to reach this expression, we need to evaluate the **block inverse** of Σ:

$$\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1} = \begin{bmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{bmatrix} \tag{18}$$

Here, we use the block inverse formula $\begin{bmatrix}A & B\\ C & D\end{bmatrix}^{-1} = \begin{bmatrix}M & \text{-}ABN\\ \text{-}NC^{-1}A & N\end{bmatrix}$, where $M = (A - BD^{-1}C)^{-1}$ and $N = (D - CA^{-1}B)^{-1}$. This is of a slightly different form to what was provided with in the question, but it is equivalent and tidier to use [Source: Matrix Cookbook, P46], we can see that:

$$\Lambda_{21} = -(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}^{-1}\Sigma_{11} \tag{19}$$

$$\Lambda_{22} = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \tag{20}$$

By substituting this into our derived expression we reach our final solution.

Estimation with Gaussians

Conditional expectation as "best" predictor

Theorem 1 (Mean Squared Error)**.** *Let X and Y be two jointly distributed random variables with density function $f(x, y)$. Suppose we want to find a prediction function $g(X)$ that minimises the mean square error $MSE = \mathbb{E}[(Y - g(X))^2]$. Show that $g(X) = \mathbb{E}[Y|X]$.*

Proof. Let's denote $\mathbb{E}[Y|X] = \mu_{Y|X}$

$$\begin{aligned}
MSE &= \int\int (y - g(x))^2 f(y|x)dyf(x)dx \\
&= \int\int ((y - \mu_{Y|X}) + (\mu_{Y|X} - g(x)))^2 f(y|x)dyf(x)dx \\
&= \int\int (\mu_{Y|X} - g(x))^2 f(y|x)dyf(x)dx + 2\int\int (y - \mu_{Y|X})(\mu_{Y|X} - g(x))f(y|x)dyf(x)dx \\
&+ \quad \underbrace{\int\int (y - \mu_{Y|X})^2 f(y|x)dyf(x)dx}_{\text{We can drop this term as it is independent of } g(x)} \\
&\propto \int\int (\mu_{Y|X} - g(x))^2 f(y|x)dyf(x)dx + 2\int\int (y - \mu_{Y|X}) \underbrace{(\mu_{Y|X} - g(x))}_{\text{Independant of } Y} f(y|x)dyf(x)dx \\
&= \int\int (\mu_{Y|X} - g(x))^2 f(y|x)dyf(x)dx + 2\int (\mu_{Y|X} - g(x)) \underbrace{\int (y - \mu_{Y|X})f(y|x)dy}_{\substack{= \mathbb{E}[Y - \mu_{Y|X}|X] \\ = \mathbb{E}[Y|X] - \mu_{Y|X} = 0}} f(x)dx \\
\implies MSE &= \int\int (\mu_{Y|X} - g(x))^2 f(y|x)dyf(x)dx \geq 0
\end{aligned}$$

This final expression has a minimum of 0. the only way this is obtained is when the function on the inside of the integral is equal to 0. this is only possible when $g(x) = \mu_{Y|X} = \mathbb{E}[Y|X]$, as required. □

MLE of σ^2

Theorem 2. *Consider the linear model $Y \sim \mathcal{N}_N(X\beta, \sigma^2 I)$. Show that the maximum likelihood estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{N}(Y - X\hat{\beta})^T(Y - X\hat{\beta})$ and this is a biased estimator.*

Proof. The pdf of the multivariate normal distribution is given by:

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\sigma^2 I|^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}$$

In this scenario, we can simply use our pdf as just the Likelihood function:[1]

$$\mathcal{L}(\beta, \sigma^2; \mathbf{Y}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\sigma^2 I|^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}$$

By taking the log likelihood we get:

$$\begin{aligned} \ell(\beta, \sigma^2; \mathbf{Y}) &= \log[\mathcal{L}(\beta, \sigma^2; \mathbf{Y})] \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) \end{aligned}$$

By differentiating in respect to σ^2, we get:

$$\begin{aligned} \frac{\partial \ell}{\partial \sigma^2} &= -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) = 0 \\ \implies \hat{\sigma}^2 &= \frac{1}{N}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) \end{aligned}$$

Now we have our estimator, we must show that it is biased. Here we let $dim(X) = p$

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\Big[\frac{1}{N}(Y - X\hat{\beta})^T(Y - X\hat{\beta})\Big] \\ &= \frac{1}{N}\mathbb{E}\Big[Y^TY - Y^TX\hat{\beta} - \hat{\beta}^TX^TY + \hat{\beta}^TX^TX\hat{\beta}\Big] \\ &= \frac{1}{N}\mathbb{E}\Big[Y^TY - Y^TX(X^TX)^{-1}X^TY - Y^TX(X^TX)^{-1}X^TY + Y^TX(X^TX)^{-1}X^TX(X^TX)^{-1}X^TY\Big] \\ &= \frac{1}{N}\mathbb{E}\Big[Y^TY - Y^TX(X^TX)^{-1}X^TY - Y^TX(X^TX)^{-1}X^TY + Y^TX(X^TX)^{-1}X^TY\Big] \\ &= \frac{1}{N}\mathbb{E}\Big[Y^TY - Y^TX(X^TX)^{-1}X^TY\Big] \\ &= \frac{1}{N}\mathbb{E}\Big[Y^TY\Big] - \frac{1}{N}\mathbb{E}\Big[Y^TX(X^TX)^{-1}X^TY\Big] \\ &= \frac{1}{N}\text{trace}\left(\mathbb{E}\Big[Y^TY\Big]\right] - \frac{1}{N}\text{trace}\left(\mathbb{E}\Big[X(X^TX)^{-1}X^TYY^T\Big]\right) \\ &= \frac{1}{N}\text{trace}\left(\mathbb{E}\Big[Y^TY\Big]\right) - \frac{1}{N}\text{trace}\left(\mathbb{E}\Big[X(X^TX)^{-1}X^TYY^T\Big]\right) \\ &= \frac{1}{N}\text{trace}\big(\sigma^2 I_N\big) - \frac{1}{N}\text{trace}\left(X(X^TX)^{-1}X^T\mathbb{E}\Big[YY^T\Big]\right) \\ &= \frac{\sigma^2}{N}\text{trace}\big(I_N\big) - \frac{\sigma^2}{N}\text{trace}\left(X(X^TX)^{-1}X^T\right) \\ &= \frac{\sigma^2}{N}N - \frac{\sigma^2}{N}\text{trace}\left(X^TX(X^TX)^{-1}\right) \\ &= \frac{\sigma^2}{N}N - \frac{\sigma^2}{N}\text{trace}\big(I_p\big) \\ &= \frac{\sigma^2}{N}N - \frac{\sigma^2}{N}p = \frac{N-p}{N}\sigma^2 \neq \sigma^2 \end{aligned}$$

Hence we have a biased estimator □

[1] For anyone interested in seeing a derivation of the MLE of a Multivariate Normal Distribution with a given sample check out this proof

Obtaining an unbiased estimator for σ^2
Even though $\hat{\sigma}^2_{MLE}$ is a biased estimator, we can still use it to obtain an unbiased Estimator:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_{MLE}] &= \frac{N-p}{N}\sigma^2 \\
\mathbb{E}\Big[\frac{1}{N}(Y - X\hat{\beta})^T(Y - X\hat{\beta})\Big] &= \frac{N-p}{N}\sigma^2 \\
\frac{N}{N-p}\mathbb{E}\Big[\frac{1}{N}(Y - X\hat{\beta})^T(Y - X\hat{\beta})\Big] &= \sigma^2 \\
\mathbb{E}\underbrace{\Big[\frac{1}{N-p}(Y - X\hat{\beta})^T(Y - X\hat{\beta})\Big]}_{=S^2} &= \sigma^2
\end{aligned}$$

Here we have an estimator $S^2 = \frac{1}{N-p}(Y - X\hat{\beta})^T(Y - X\hat{\beta})$, such that $\mathbb{E}[S^2] = \sigma^2$. Hence S^2 is an unbiased estimator of σ^2.

Weighted Data Points

Consider a data set in which each data point $(y_i, \mathbf{x}_i)$ has a weight $w_i > 0$ associated with it, so that the sum of squares error function becomes:

$$S = \frac{1}{2}\sum_{i=1}^{N} w_i(y_i - \beta^T\phi(\mathbf{x}_i))^2 \tag{1}$$

We shall derive the parameter vector which minimises the error function, β^*.

$$\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{N} w_i(y_i - \beta^T\phi(\mathbf{x}_i))^2 \\
&= \frac{1}{2}(Y - \begin{bmatrix}\phi(\mathbf{x}_1)\\ \vdots \\ \phi(\mathbf{x}_N)\end{bmatrix}\beta)^T \begin{bmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_N\end{bmatrix}(Y - \begin{bmatrix}\phi(\mathbf{x}_1)\\ \vdots \\ \phi(\mathbf{x}_N)\end{bmatrix}\beta) \\
&= \frac{1}{2}(Y - X\beta)^T W (Y - X\beta)
\end{aligned}$$

We now take first order conditions of the equation:

$$\begin{aligned}
\frac{\partial S}{\partial \beta} &= -X^T W(Y - X\beta) = 0 \\
&\implies \beta^* = (X^T W X)^{-1} X^T W Y
\end{aligned}$$

But what are the advantages of using these weights?

- **Focusing accuracy** We may care very strongly about predicting the response for certain values of the input — ones we expect to see often again, ones where mistakes are especially costly or embarrassing or painful, etc. than others. If we give the points x_i near that region big weights w_i, and points elsewhere smaller weights, the regression will be pulled towards matching the data in that region. This will help us in cases when we do have replicated data points.
- **Discounting imprecision.** Ordinary least squares is the maximum likelihood estimate when the ϵ in $Y = X\beta + \epsilon$ is IID Gaussian white noise. This means that the variance of ϵ has to be constant, and we measure the regression curve with the same precision elsewhere. This situation, of constant

noise variance, is called **homoskedasticity**. Often however the magnitude of the noise is not constant, and the data are heteroskedastic. When we have **heteroskedasticity**, even if each noise term is still Gaussian, ordinary least squares is no longer the maximum likelihood estimate, and so no longer efficient. If however we know the noise variance σ_i^2at each measurement i, and set $w_i = 1/\sigma_i^2$, we get the heteroskedastic MLE, and recover efficiency. To say the same thing slightly differently, there's just no way that we can estimate the regression function as accurately where the noise is large as we can where the noise is small. Trying to give equal attention to all parts of the input space is a waste of time; we should be more concerned about fitting well where the noise is small, and expect to fit poorly where the noise is big.

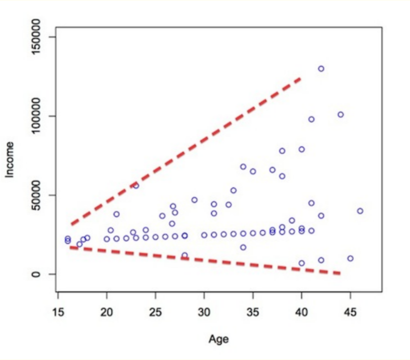
- **Doing something else.** There are a number of other optimization problems which can be transformed into, or approximated by, weighted least squares. The most important of these arises from generalized linear models, where the mean response is some nonlinear function of a linear predictor.[2]

Heteroskedasticity: An overview

Heteroskedasticity is a hard word to pronounce, but it doesn't need to be a difficult concept to understand. Put simply, heteroskedasticity refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

A scatterplot of these variables will often create a cone-like shape, as the variance of the dependent variable widens or narrows as the value of the independent variable increases. The inverse of heteroskedasticity is homoscedasticity, which indicates that a dependant variable's variance is equal across values of an independent variable.

For example: annual income might be a heteroskedastic variable when predicted by age, because most teenagers aren't flying around in jets that they bought from their own income. More commonly, teen workers earn close to the minimum wage, so there isn't a lot of variability during the teen years. However, as teens turn into 20-somethings, and 20-somethings into 30-somethings, some will tend to shoot-up the tax brackets, while others will increase more gradually (or perhaps not at all, unfortunately). Put simply, the gap is likely to widen with age. If the above where true and I had a random sample of earners across all ages, a plot of the association between age and income would demonstrate heteroskedasticity, like this:



Heteroskedasticity is most frequently discussed in terms of the assumption of parametric analyses (e.g. linear regression). More specifically, it is assumed that the error of a regression model is homoskedastic across all values of the predicted value of the dependant variable. Put more simply, a test of homoskedasticity of error terms determines whether a regression model's ability to predict a variable is consistent across all values. If a regression model is consistently accurate when it predicts low values, but highly inconsistent in accuracy when it predicts high values, then the results of that regression should not be trusted.

I want to re-iterate that the concern about heteroskedasticity, in the context of regression and other parametric analyses, is specifically related to error terms and NOT between two individual variables. This is a common misconception, similar to the misconception about normality (Variables need not be normally distributed, as long as the residuals of the regression model are normally distributed).

[2] Lecture 18: Extending Linear Regression: Weighted Least Squares, Heteroskedasticity, Local Polynomial Regression, *Carnegie Mellon University: 36-350, Data Mining*