Suggested Solutions to Midterm 1
-Econ 140, Fall 2002-
(Tanguy Brachet)

Problem 1. Assume that the ‘per capita’ income of residents in a country is normally distributed with a mean ($\mu$) of $1000 and variance ($\sigma^2$) equal to 10000.

So $Y \sim N(1000, 10000)$, which implies that the new variable $\frac{Y - \mu}{\sigma} = \frac{Y - 1000}{\sqrt{10000}}$ is standard normally distributed, i.e. $\frac{Y - 1000}{100} \sim N(0,1)$. Throughout this problem, as has been the convention in lecture, $Z$ will denote a standard normally distributed random variable.

i) what is the probability that the per capita income lies between $800 and $1200?

\[
\Pr(800 < Y < 1200) = \Pr \left( \frac{800 - 1000}{\sqrt{10000}} < \frac{Y - 1000}{\sqrt{10000}} < \frac{1200 - 1000}{\sqrt{10000}} \right) \\
= \Pr \left( \frac{-200}{100} < \frac{Y - 1000}{100} < \frac{200}{100} \right) \\
= \Pr (-2 < Z < 2) \\
= 2 \times \Pr(0 < Z < 2) \text{ by symmetry of the } N(0,1) \text{ distribution} \\
= 2 \times 0.4772 \\
= 0.9544
\]

ii) what is the probability that it exceeds $1140?

\[
\Pr(1140 < Y) = \Pr \left( \frac{1140 - 1000}{\sqrt{10000}} < \frac{Y - 1000}{\sqrt{10000}} \right) \\
= \Pr \left( \frac{140}{100} < \frac{Y - 1000}{100} \right) \\
= \Pr(1.4 < Z) \\
= 1 - \Pr(Z < 1.4) \\
= 1 - [\Pr(Z < 0) + \Pr(0 < Z < 1.4)] \\
= 1 - [0.5 + \Pr(0 < Z < 1.4)] \\
= 1 - 0.5 - 0.4192 \\
= 0.0808
\]

iii) what is the probability that it is less than $760?

\[
\Pr(Y < 760) = \Pr \left( \frac{Y - 1000}{\sqrt{10000}} < \frac{760 - 1000}{\sqrt{10000}} \right) \\
= \Pr \left( \frac{Y - 1000}{100} < \frac{-240}{100} \right) \\
= \Pr(Z < -2.4) \\
= \Pr(Z > 2.4) \text{ by symmetry of the } N(0,1) \text{ distribution} \\
= \Pr(0 < Z) - \Pr(0 < Z < 2.4) \\
= 0.5 - \Pr(0 < Z < 2.4) \\
= 0.5 - 0.4918 \\
= 0.0082
\]
iv) Suppose a sample of 1000 residents is taken and the sample mean is found to be $8900. What is the probability that the sample mean lies between $8800 and $12000?

Note that we are now asked to calculate a probability for a new random variable: A sample of 1000 residents $(Y_1, ..., Y_{1000})$ is taken and, from it, the sample mean, $\bar{Y} = \frac{\sum_{i=1}^{1000} Y_i}{1000}$, is constructed. The important thing in this question is to recognize that the distribution of $\bar{Y}$ differs from that of each of its component $Y_i$'s. In particular, as shown in lecture and in section, the variance of $\bar{Y}$ is 1000 times smaller than that of any of the $Y_i$'s: $Var(\bar{Y}) = \frac{\sigma^2}{1000}$.

Elsewhere, we know that $\bar{Y}$ is normally distributed: if $n$ observations are drawn from a normal distribution with mean $\mu_Y$ and variance $\sigma_Y^2$, then the sample mean of those draws will be $N(\mu_Y, \frac{\sigma^2_Y}{n})$. Thus, $\bar{Y} \sim N(\mu_Y = 1000, \frac{\sigma^2_Y}{n} = \frac{10000}{1000} = 10)$ and, therefore, after standardization, $\frac{\bar{Y} - 1000}{\sqrt{10}} \sim N(0,1)$. We are now equipped to proceed.

\[
Pr\left(800 < \bar{Y} < 1200\right) = Pr\left(\frac{800 - 1000}{\sqrt{10}} < \frac{\bar{Y} - 1000}{\sqrt{10}} < \frac{1200 - 1000}{\sqrt{10}}\right)
\]
\[
= Pr\left(\frac{-200}{\sqrt{10}} < \frac{\bar{Y} - 1000}{\sqrt{10}} < \frac{200}{\sqrt{10}}\right)
\]
\[
= Pr\left(-63.25 < Z < 63.25\right)
\]
\[
\approx 2 \times Pr\left(0 < Z < 63.25\right) \quad \text{by symmetry of the } N(0,1) \text{ distribution}
\]
\[
\approx 2 \times 0.5 \quad \text{since 63.25 is WAY into the tail of the } N(0,1)
\]
\[
= \frac{1}{2}
\]

v) Compare your answer in iv) with i).

The results in iv) exceed those in i) precisely because of the above remarks: the distribution of $\bar{Y}$ is much more concentrated around the mean than that of each of the $Y_i$'s. This point is illustrated in the chart at the top of the page, which graphs the pdfs of $Y$ and $\bar{Y}$. Notice how much faster the pdf of $\bar{Y}$ tails off to zero than that of $Y$. 

Figure 1:
Problem 2. If \( \hat{Y} \) is a point estimator of \( \mu_Y \) (the population mean), where \( \hat{Y} = \sum_{i=1}^{n} c_i Y_i \) and where \( c_i = \frac{1}{\sqrt{n}} \), and we have a random sample \((Y_1, \ldots, Y_n)\) drawn with mean \( \mu_Y \) and variance \( \sigma^2 \).

Note that we can summarize the information so far in the following way:

\[
\hat{Y} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i \quad \text{where}
\]
\[
E(Y_i) = \mu_Y \quad \text{for } i = 1, \ldots, n
\]
\[
V(Y_i) = \sigma^2 \quad \text{for } i = 1, \ldots, n
\]

and each \( Y_i \) is independently and identically distributed.

i) compute \( E(\hat{Y}) \) and compute \( Var(\hat{Y}) \).

\[
E \left( \hat{Y} \right) = E \left( \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i \right)
\]
\[
= \sum_{i=1}^{n} E \left( \frac{1}{\sqrt{n}} Y_i \right) \quad \text{since the expectation of the sum is the sum of the expectations}
\]
\[
= \sum_{i=1}^{n} \frac{1}{\sqrt{n}} E(Y_i) \quad \text{since } \frac{1}{\sqrt{n}} \text{ is a constant}
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E(Y_i) \quad \text{since } \frac{1}{\sqrt{n}} \text{ is a constant not indexed by } i
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mu_Y
\]
\[
= \frac{1}{\sqrt{n}} (n \times \mu_Y)
\]
\[
= \sqrt{n} \times \mu_Y
\]

\[
Var \left( \hat{Y} \right) = Var \left( \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i \right)
\]
\[
= \sum_{i=1}^{n} Var \left( \frac{1}{\sqrt{n}} Y_i \right) + \sum_{i \neq j} \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \times \frac{1}{\sqrt{n}} \times \underbrace{Cov(Y_i, Y_j)}_{=0 \text{ by random sampling}}
\]
\[
= \sum_{i=1}^{n} \left( \frac{1}{\sqrt{n}} \right)^2 Var(Y_i) \quad \text{since } \frac{1}{\sqrt{n}} \text{ is a constant}
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} Var(Y_i)
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \sigma^2
\]
\[
= \frac{1}{n} (n \times \sigma^2)
\]
\[
= \sigma^2
\]
ii) if the sample mean \( \bar{Y} = \frac{\sum Y}{n} \) has a variance of \( \frac{\sigma^2}{n} \), is \( \bar{Y} \) BLUE? Explain your answer.

Recall what BLUE means: an estimator is BLUE for \( \mu_Y \) if it is Linear in the data, Unbiased for \( \mu_Y \), and among all linear and unbiased estimators of \( \mu_Y \), it has smallest variance (i.e. it is Best).

Well, \( \bar{Y} \) is certainly (trivially) linear in the \( Y_i \)'s, but as established in part i), it is NOT unbiased: \( E(\bar{Y}) = \sqrt{n} \mu_Y \neq \mu_Y \). Thus, \( \bar{Y} \) can't possibly be BLUE for \( \mu_Y \). (Not to mention that even if it were unbiased, it has larger variance than \( \bar{Y} \): \( \sigma^2 \geq \frac{\sigma^2}{n} \).

**Problem 3.** For the straight line model: \( Y_i = a + bX_i + \epsilon_i \), derive the estimator for \( b \) and show that it is equivalent to:

\[
\hat{b} = \frac{\sum_i (X_i - \bar{X}) Y_i}{\sum_i (X_i - \bar{X})^2}
\]

Assume that you are finding a minimum (that is, that the second derivative is positive).

Recall that the general method we’ve adopted to find estimators has been the minimization of the Sum of Squared Residuals (SSR) with respect to the parameters of interest. In this problem, we’ll want to minimize the SSR with respect to \( \hat{a} \) and \( \hat{b} \) to get an estimate of \( b \). Note that the problem doesn’t specifically request that you derive \( a \), but we’ll have to do it in order to get a formula for \( b \) in terms only of \( X \)'s and \( Y \)'s.

So, the problem we need to solve is:

\[
\min_{\hat{a}, \hat{b}} \sum_{i=1}^{n} e_i^2 = \min_{\hat{a}, \hat{b}} \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2
\]

To solve the minimization problem, we take partial derivatives with respect to \( \hat{a} \) and \( \hat{b} \) and set them to zero (these are the “**First Order Conditions**”). Let’s first handle the intercept:

\[
\frac{\partial SSR}{\partial \hat{a}} = \frac{\partial \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2}{\partial \hat{a}} = 0
\]

\[
\sum_{i=1}^{n} \frac{\partial}{\partial \hat{a}} \left( (Y_i - \hat{a} - \hat{b}X_i)^2 \right) = 0 \quad \text{since the derivative of the sum is the sum of the derivatives}
\]

\[
\sum_{i=1}^{n} 2 \times (Y_i - \hat{a} - \hat{b}X_i) \times (-1) = 0
\]

\[
(-2) \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i) = 0
\]

\[
\sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i) = 0
\]

\[
\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{a} - \sum_{i=1}^{n} \hat{b}X_i = 0
\]
$\sum_{i=1}^{n} Y_i - n \hat{a} - \hat{b} \sum_{i=1}^{n} X_i = 0$

$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} X_i$

$\hat{a} = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{b} \frac{\sum_{i=1}^{n} X_i}{n}$

where $\frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}$ and $\frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$

$\hat{a} = \bar{Y} - \hat{b} \bar{X}$

Now for the estimate of the slope:

$\frac{\partial SSR}{\partial \hat{b}} = \frac{\partial \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2}{\partial \hat{b}} = 0$

$\sum_{i=1}^{n} \frac{\partial}{\partial \hat{b}} \left( Y_i - \hat{a} - \hat{b}X_i \right)^2 = 0$ since the derivative of the sum is the sum of the derivatives

$\sum_{i=1}^{n} 2 \times (Y_i - \hat{a} - \hat{b}X_i) \times (-X_i) = 0$

$(-2) \times \sum_{i=1}^{n} X_i \left( Y_i - \hat{a} - \hat{b}X_i \right) = 0$

$\sum_{i=1}^{n} X_i \left( Y_i - \hat{a} - \hat{b}X_i \right) = 0$

$\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \hat{a} - \sum_{i=1}^{n} \hat{b}X_i^2 = 0$

$\sum_{i=1}^{n} X_i Y_i - \hat{a} \sum_{i=1}^{n} X_i - \hat{b} \sum_{i=1}^{n} X_i^2 = 0$

$\sum_{i=1}^{n} X_i Y_i - \hat{a} \sum_{i=1}^{n} X_i = -\hat{b} \sum_{i=1}^{n} X_i^2 = 0$ since $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ so that $\sum_{i=1}^{n} X_i = n \bar{X}$

$\sum_{i=1}^{n} X_i Y_i - \left( \bar{Y} - \hat{b} \bar{X} \right) \times (n \bar{X}) - \hat{b} \sum_{i=1}^{n} X_i^2 = 0$ plugging in the formula for $\hat{a}$

$\sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y} - n \hat{b} \bar{X}^2 - \hat{b} \sum_{i=1}^{n} X_i^2 = 0$

$\hat{b} \left( \sum_{i=1}^{n} X_i^2 - n \bar{X}^2 \right) = \sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}$

$\hat{b} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$
seeing if it boils down to the formula for $\hat{b}$ that we just derived:

$$\frac{\sum_i (X_i - \bar{X}) Y_i}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i Y_i - \bar{X} \bar{Y})}{\sum_i (X_i^2 - 2X_i \bar{X} + \bar{X}^2)}$$

$$= \frac{\sum_i X_i Y_i - \sum_i \bar{X} Y_i}{\sum_i X_i^2 - \sum_i 2X_i \bar{X} + \sum_i \bar{X}^2}$$

$$= \frac{\sum_i X_i Y_i - \bar{X} \sum_i Y_i}{\sum_i X_i^2 - 2\bar{X} \sum_i X_i + n\bar{X}^2}$$

where $\sum_i Y_i = n\bar{Y}$ and $\sum_i X_i = n\bar{X}$

$$= \frac{\sum_i X_i Y_i - n\bar{X} \bar{Y}}{\sum_i X_i^2 - n\bar{X}^2} = \hat{b}$$

(Q.E.D.)

**Problem 4.** You are given the following figures on log wages ($Y$) and education ($X$) in a matrix of products and cross-products.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>50</td>
</tr>
<tr>
<td>45</td>
<td></td>
</tr>
</tbody>
</table>

(Remember that the variables $x$ and $y$ are in coded form: $x = (X - \bar{X})$; $y = (Y - \bar{Y})$. For example, for the cell $y,y$, the product is $\sum_{i=1}^n y_i^2$.) $N = 15.$

i) **Estimate the slope coefficient for the model:** $Y_i = a + bX_i + e_i$.

Recall that the formula for the estimate of the slope coefficient that we derived in Problem 3 can be re-expressed in terms of coded variables:

$$\hat{b} = \frac{\sum_i X_i Y_i - n\bar{X} \bar{Y}}{\sum_i X_i^2 - n\bar{X}^2} = \frac{\sum_i (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

$$= \frac{50}{65}$$

$$= 0.7692$$

ii) **Estimate the standard error for $\hat{b}$**.

Recall that the variance of $\hat{b}$ is given by:

$$Var(\hat{b}) = \frac{\sigma^2}{\sum_i x_i^2}$$

If $\sigma^2$ were known, then we’d basically be done. However, $\sigma^2$ is not known, so it needs to be estimated by $s^2$, where:

$$s^2 = \frac{RSS}{n-k} = \frac{\sum_{i=1}^n e_i^2}{n-k}$$

where $k$ is the number of parameters in the model; so here, $k = 2$.

But what is the RSS? To determine the RSS, use the Sums of Squares Identity: $TSS = ESS + RSS \iff RSS = TSS - ESS$ where $TSS$ is the Total Sum of Squares and $ESS$ is the Explained Sum of Squares.
Elsewhere, we know the formulas for $TSS$ and $ESS$:

\[
TSS = \sum_i (Y_i - \bar{Y})^2 = \sum_i y_i^2 = 45
\]

\[
ESS = \sum_i \left( \hat{Y}_i - \bar{Y} \right)^2 = \sum_i \left( \left( \hat{a} + \hat{b}X_i \right) - \bar{Y} \right)^2
\]

\[
= \sum_i \left( \left( Y - \hat{b}X \right) + \hat{b}X_i - \bar{Y} \right)^2
\]

\[
= \sum_i \left( \hat{b} (X_i - \bar{X}) \right)^2
\]

\[
= \sum_i \hat{b}^2 x_i^2
\]

\[
= \hat{b}^2 \sum_i x_i^2
\]

\[
= \left( \frac{50}{65} \right)^2 \times (65) = 38.46^1
\]

Thus, $RSS = TSS - ESS = 45 - 38.46 = 6.54$. Hence,

\[
\hat{SE}(\hat{b}) = \sqrt{\frac{s^2}{\sum_i x_i^2}} = \sqrt{\frac{RSS}{n-k}}
\]

\[
= \sqrt{\frac{6.54}{15-2}}
\]

\[
= 0.0880
\]

iii) Form a 99 percent confidence interval for $b$ (the unknown population parameter).

Recall that whatever the true parameter, $b$, is the standardized random variable \( \frac{\hat{b} - b}{SE(\hat{b})} \) is distributed $t$ with $n - k = 15 - 2 = 13$ degrees of freedom. Thus,

\[
Pr \left( -t_{13,0.005} < \frac{\hat{b} - b}{SE(\hat{b})} < t_{13,0.005} \right) = 99%
\]

\[
Pr \left( -t_{13,0.005} \times \hat{SE}(\hat{b}) < \hat{b} - b < t_{13,0.005} \times \hat{SE}(\hat{b}) \right) = 99%
\]

\[
Pr \left( -b - t_{13,0.005} \times \hat{SE}(\hat{b}) < \hat{b} - b < -b + t_{13,0.005} \times \hat{SE}(\hat{b}) \right) = 99%
\]

\[
Pr \left( \hat{b} + t_{13,0.005} \times \hat{SE}(\hat{b}) > b > \hat{b} - t_{13,0.005} \times \hat{SE}(\hat{b}) \right) = 99%
\]

\[
Pr \left( 0.7692 - 3.012 \times 0.0880 < b < 0.7692 + 3.012 \times 0.0880 \right) = 99%
\]

\[
Pr \left( 0.504 < b < 1.034 \right) = 99%
\]

So we can be 99% confident that $b$ lies in the interval $[0.504, 1.034]$.

iv) Test the hypothesis that: $H_0 : b = 0.7$ at the $\alpha = 0.5$ significance level. If you are trying to minimize Type I errors, is this a sensible significance level to select?
Note that the first thing we should do before carrying out a hypothesis test is specify the alternative hypothesis. Let’s specify the two-tailed alternative, $H_1 : b \neq 0.7$. As alluded to in part iii), under the null hypothesis that the true parameter $\hat{b}$ is equal to 0.7, the standardized random variable $t^* = \frac{\hat{b} - 0.7}{SE(\hat{b})}$ is distributed $t$ with $n - k = 15 - 2 = 13$ degrees of freedom. Here,

$$t^* = \frac{\hat{b} - 0.7}{SE(\hat{b})} = \frac{0.7692 - 0.7}{0.0880} = 0.7864$$

To determine whether or not we reject the null at the 50% (not 5%!) level, we need to compare $|t^*|$ to the two-tailed 50% cutoff of the $t_{13}$, $t_{13, 0.25} = t_{13, 0.025} = 0.694$ because the rejection region for this test is $|t^*| > t_{13, 0.025} = 0.694$. So

$$0.7864 = |t^*| > t_{13, 0.025} = 0.694$$

and thus we reject $H_0$ at the 50% level.

Recall what the level of a test is: $\alpha$ is the probability of committing a Type I error in your test, i.e. the probability of rejecting the null hypothesis when it is in fact true. Thus, $\alpha = 0.5$ means that whenever the null hypothesis is true, we have a 50% chance of falsely rejecting it. Equivalently, when $H_0$ is true, we are as likely to reject it as we are to accept it. This hardly seems like a sensible rule for accepting and rejecting hypotheses. Therefore, we should probably reduce $\alpha$ so as to commit less false negatives.