1. We wish to fit paired sample \((x_1, y_1), \ldots, (x_n, y_n)\) with a linear regression model \(Y = cx + \epsilon\). We assume that \(\epsilon\) follows a normal distribution with zero mean and variance \(\sigma^2\).

(a) Estimate \(c\) by minimizing the sum of the squared residuals (5pts).

(b) Write the log-likelihood function as a function of \(c\) and \(\sigma\) (5pts).

(c) Find the likelihood estimator of \(c\) by differentiating the log-likelihood function in (b). Derive everything (5pts).

(d) Either prove or disprove unbiasedness of the estimator you computed in (c) (5pts, no point given if (c) is incorrect).

(e) Compute the variance of the estimator you computed in (a) (5pts, no point given if (a) is incorrect).

(f) If the sample correlation coefficient of the above paired data is 0.5, what is the sample correlation coefficient of data \((x_1 - \bar{x}, y_1 - \bar{y}), \ldots, (x_n - \bar{x}, y_n - \bar{y})\)? \(\bar{x}\) and \(\bar{y}\) are the respective sample means of \(x_i\)’s and \(y_i\)’s. Prove your result (5pts).

Solution. (a) The sum of squared residuals \(SSE = \sum_{i=1}^{n}(y_i - cx_i)^2\) (2pts). Letting \(\partial SSE/\partial c = 0\), we get \(\hat{c} = \sum_{i=1}^{n}x_i(y_i - cx_i) = 0\). Solving this, \(\hat{c} = \sum_{i=1}^{n}x_iy_i/\sum_{i=1}^{n}x_i^2\) (3pts). (b) Note that \(\mathbb{E}Y_i = cx_i\) and \(\text{Var}Y_i = \sigma^2\) for some \(\sigma\). So \(Y_i \sim N(cx_i, \sigma^2)\). Then the likelihood function is given by

\[
L(c, \sigma) = \frac{\text{const.}}{\sigma^n} \exp \left( \frac{\sum_{i=1}^{n}(y_i - cx_i)^2}{2\sigma^2} \right).
\]

The log-likelihood function is then

\[
\log L(c, \sigma) = \text{const.} - n \log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - cx_i)^2.
\]

(c) Letting \(\partial \log L/\partial c = 0\), we get \(\sum_{i=1}^{n}x_i(y_i - cx_i) = 0\) so we get the same answer. (d) \(\mathbb{E}\hat{c} = \sum_{i=1}^{n}x_i/\sum_{i=1}^{n}x_i^2 = c\). (e) \(\text{Var}\hat{c} = \sum_{i=1}^{n}x_i^2 \text{Var}Y_i/\left[\sum_{i=1}^{n}x_i^2\right]^2 = \sigma^2/\sum_{i=1}^{n}x_i^2\). (f) The sample correlation is given by \(r = S_{xy}/\sqrt{S_{xx}S_{yy}}\), where \(S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\). Let \(z_i, w_i = (x_i - \bar{x}, y_i - \bar{y})\). Note that \(\bar{z} = \bar{w} = 0\). So the sample correlation for \((z_i, w_i)\) would be \(\hat{p} = S_{zw}/\sqrt{S_{zz}S_{ww}}\) where \(S_{zw} = \sum_{i=1}^{n}(z_i - \bar{z})(w_i - \bar{w}) = \sum_{i=1}^{n}z_iw_i = S_{xy}\). Also \(S_{zz} = S_{xx}\) and \(S_{ww} = S_{yy}\). So the sample correlation is the same.

2. When you throw a coin 10 times, you observed 3 heads. Test if the coin is biased at \(\alpha = 0.2\). Clearly state appropriate parameters, hypotheses, a test statistic. You may use the following R output (15pts).

```
> pbinom(0:5,10,0.5)
[1] 0.00 0.01 0.06 0.17 0.38 0.62
```

Solution. Let \(p\) be the probability of getting head. Then \(H_0: p = 0.5\) vs. \(H_1: p \neq 0.5\). Assign \(X_i = 1\) if \(i\)-th tossing yields head and \(X_i = 0\) otherwise. Note that \(P(X_i = 1) = p, P(X_i = 0) = 1 - p\). The point estimator for \(p\) would be \(\hat{p} = \sum_{i=1}^{10}X_i/10 = \bar{X}\). We will take \(\sum_{i=1}^{10}X_i\) rather than \(\bar{X}\) as the test statistic because it will be much easier to figure out the distribution of the test statistic. Note that \(\sum_{i=1}^{10}X_i \sim \text{Binomial}(10, p)\). Under \(H_0\), \(\sum_{i=1}^{10}X_i \sim \text{Binomial}(10, 0.5)\). We reject \(H_0\) if \(\sum_{i=1}^{10}x_i\) is either too small or too large. Note that \(P(\sum_{i=1}^{10}X_i \leq 3) = P(\sum_{i=1}^{10}X_i \geq 7) = 0.17\) from symmetry. So the \(P\)-value is 0.34. We do not reject the null hypothesis at \(\alpha = 0.2\).
3. A manufacturer of automatic washers offers a particular model in one of three colors: white, green and blue. Of the 100 washers sold, 40 were white, 35 green, 25 blue. In each of the subsequent questions, clearly state appropriate parameters, hypotheses, a test statistic and its distribution under the null assumption.

(a) Would you conclude that customers have no preference? Test it at \( \alpha = 0.05 \) (10 pts).

(b) Would you conclude that customers have a preference for white color? Test it at \( \alpha = 0.05 \) (10 pts).

Solution. Let \( p_w, p_r, p_b \) be the proportion of customers who prefer white, green and blue respectively. (a) If there is no preference, we expect \( H_0 : p_w = p_r = p_b = 1/3 \). Under the null, the expected numbers of washers customers choose are 33.3 for each color. The chi-square statistic value is \( \chi^2 = (40 - 33.3)^2/33.3 + (35 - 33.3)^2/33.3 + (25 - 33.3)^2/33.3 = 3.50 \). There are three categories. So we compare it with the cutoff value 4.60 from \( \chi^2 \). Do not reject \( H_0 \). (b) If there is no preference for white, we expect \( p_w = 1/3 = 0.333 \). So the hypotheses of interest would be \( H_0 : p_w = 1/3 \) vs. \( H_1 : p_w > 1/3 \). The test statistic is based on \( z \)-statistic of large sample proportion. \( z \)-value is \( z = (0.4 - 0.333)/\sqrt{0.3330.666/1000} = 0.450 \). Since the rejection region is \( z > 1.64 \), we reject \( H_0 \) and conclude that the customers prefer white color.

4. A quality control engineer has measured the numbers of defectives per day from a certain production process for 50 days and recorded below. Test if the number of defectives follows a binomial distribution at \( \alpha = 0.05 \). (20pts).

<table>
<thead>
<tr>
<th>number of defects</th>
<th>frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

Solution. We test if data follow Binomial(3, \( p \)). First we estimate \( p \) by matching the sample mean and the population mean, which gives the maximum likelihood estimation of \( p \). \( \bar{x} = (0 \cdot 10 + 1 \cdot 24 + 2 \cdot 10 + 3 \cdot 6)/50 = 1.24 = 3p. \) \( \hat{p} = 0.41 \). Let’s see if data follow \( P(X = x) = \binom{3}{x} \hat{p}^x (1 - \hat{p})^{3-x} \). The null hypothesis we need to test is \( p_x = P(X = x) \). \( H_0 : p_0 = (1 - \hat{p})^3 = 0.20, p_1 = 3\hat{p}(1 - \hat{p})^2 = 0.41, p_2 = 3\hat{p}^2(1 - \hat{p}) = 0.30, p_3 = \hat{p}^3 = 0.07 \). The expected numbers of defectives are 10.1, 20.7, 15, 3.5. Then \( \chi^2 = (10 - 10.1)^2/10.1 + (24 - 20.7)^2/20.7 + (10 - 15)^2/15 + (6 - 3.5)^2/3.5 = 3.97 \). The 95% cut-off value for \( \chi^2_{1-0.05} = 5.99 \) so we do not reject \( H_0 \).

5. Problem on the coefficient of determination.

(a) The following it the R output of a regression analysis based on linear model \( Y = \beta_0 + \beta_1 x + \epsilon \) for 11 paired data \((x_i, y_i)\). Compute the coefficient of determination and the coefficient of correlation (5pts).

(b) It is shown during the lecture that the coefficient of determination is always between 0 and 1. Prove it (15pts).

Solution. (a) Note that \( t \)-value = 3 = \( r\sqrt{n-2}/\sqrt{1-r^2} \). Solving this we get \( r^2 = 0.5 \). The correlation is then \( r = \sqrt{0.5} \).

(b) The coefficient of determination is given by \( r^2 = 1 - \frac{SSE}{SST} \), where \( SSE = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \), while \( SST = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \). It can be shown that \( SSE \leq SST \) (see lecture note) so 0 \( \leq r^2 \leq 1 \).