Mathematical notes - Week 9

1. Covariance and correlation

   (a) The covariance of random variables X and Y is defined as
   $\mathrm{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$
   Prove that $\mathrm{Cov}[X,Y] = E[XY] - E[X]E[Y]$.

$$E\left[(X-\mu_x)(Y-\mu_y)\right] = E[XY] - \mu_x E[Y] - \mu_Y E[X] + \mu_x\mu_Y \text{ (linearity)}$$
$$= E[XY] - \mu_x\mu_Y - \cancel{\mu_Y\mu_x} + \cancel{\mu_x\mu_Y}$$
$$= E[XY] - E[X]E[Y]$$

   (b) The correlation coefficient of random variables X and Y is defined as
$$\rho(X,Y) = \frac{\mathrm{Cov}[X,Y]}{\sigma_X\sigma_Y}.$$
   Prove that $|\rho(X,Y)| \leq 1$. $Z_X = \frac{X-\mu_X}{\sigma_X}$ $Z_Y = \frac{Y-\mu_Y}{\sigma_Y}$ both have expectation 0 and variance 1

$$\mathrm{Var}[Z_x \pm Z_Y] = E[(Z_x \pm Z_Y)^2] = E[Z_x^2] + E[Z_Y^2] \pm 2\,\mathrm{Cov}[Z_x, Z_Y] \geq 0$$
$$= 1 + 1 \pm 2\,\mathrm{Cov}[Z_x, Z_Y] \geq 0$$

So $1 + 1 \pm 2\frac{\mathrm{Cov}[X,Y]}{\sigma_x\sigma_y} \geq 0$; $2 \geq \mp\rho(X,Y)$ and $|\rho(X,Y)| \leq 1$

   (c) Prove that when calculating the sample correlation r, you can divide $\sum(x_i - \overline{x})(y_i - \overline{y})$ by n, $n-1$, or 1 in the numerator, as long as you do the same thing in the denominator.

$$r = \frac{\text{Sample covariance}}{\sqrt{(\text{Sample variances})}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Since the factor of $\frac{1}{n-1}$ cancels from numerator and denominator, we could have used $\frac{1}{n}$ or 1 instead.

2. Proof 1 - Least-squares regression

You have values x_i of a "predictor" and matching values y_i of a "response." Your goal is to minimize the sum of squares of the prediction errors,

$$g(a,b) = \sum_{i=1}^{n} (a + bx_i - y_i)^2.$$

Prove that

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \; a = \overline{y} - b\overline{x}.$$

$$\partial g/\partial a = 2\sum_{i=1}^{n}(a + bx_i - y_i) = 0$$

so $na + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0$

or $na + bn\overline{x}_i - n\overline{y}_i = 0$ and $\boxed{a = \overline{y} - b\overline{x}}$

$$\partial g/\partial b = 2\sum_{i=1}^{n} x_i(a + bx_i - y_i) = 0$$

From above $\sum_{i=1}^{n} \overline{x}(a + bx_i - y_i) = 0$ (multiply top line by the constant $\overline{x}$)

Subtract: $\sum_{i=1}^{n}(x_i - \overline{x})(a + bx_i - y_i) = 0$

Substitute for a: $\sum_{i=1}^{n}(x_i - \overline{x})(\overline{y} - b\overline{x} + bx_i - y_i) = 0$

$$\sum_{i=1}^{n}(x_i - \overline{x})\, b\, (x_i - \overline{x}) - \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 0$$

and so $\boxed{b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$

3. Proof 2 - Dividing up the variance of the observed y's

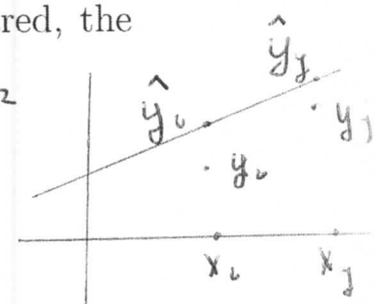
Define $ss_{xy} = \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})$; $ss_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$; $ss_y = \sum_{i=1}^{n}(y_i - \overline{y})^2$.

Correlation $r = \dfrac{ss_{xy}}{\sqrt{ss_x ss_y}}$; Slope of regression line $b = \dfrac{ss_{xy}}{ss_x}$.

Prove that $r^2 ss_y = b^2 ss_x$.

$$r^2 ss_y = \frac{(ss_{xy})^2}{(ss_x)(ss_y)} \cdot ss_y = \frac{(ss_{xy})^2}{ss_x} \qquad b^2 ss_x = \frac{(ss_{xy})^2}{(ss_x)^2} \cdot ss_x = \frac{(ss_{xy})^2}{ss_x} \checkmark$$

An observation is y_i; a predicted observation is $\hat{y}_i = a + bx_i$; $\overline{y} = a + b\overline{x}$. Prove that the ratio of the variance of the predicted y's to the variance of the observed y's equals R-squared, the square of the sample correlation r.



$$\frac{\text{Predicted variance}}{\text{Observed variance}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(a + bx_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\frac{\text{Pred.}}{\text{Obs.}} = \frac{\sum_{i=1}^{n}(\cancel{\bar{y}} - b\bar{x} + bx_i - \cancel{\bar{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{b^2(ss_x)}{ss_y} = \frac{r^2 ss_y}{ss_y} = \boxed{r^2}$$

Prove that the ratio of the variance of the residuals $y - \hat{y}$ to the variance of the observed y's equals $1 - r^2$.

$$\frac{\text{Residual}}{\text{Observed}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1} ss_y} = \frac{\sum_{i=1}^{n}(y_i - a - bx_i)^2}{ss_y} = \frac{\sum_{i=1}^{n}(y_i - \bar{y} + b\bar{x} - bx_i)^2}{ss_y}$$

$$= \sum_{i=1}^{n}\left[(y_i - \bar{y})^2 - 2b(y_i - \bar{y})(x_i - \bar{x}) + b^2(x_i - \bar{x})^2\right]\frac{1}{ss_y}$$

$$= \frac{ss_y - 2b(ss_{xy}) + b^2 ss_x}{ss_y} = \frac{ss_y - 2b(b\,ss_x) + b^2 ss_x}{ss_y}$$

$$= 1 - \frac{b^2 ss_x}{ss_y} = 1 - \frac{r^2 ss_y}{ss_y} = \boxed{1 - r^2}$$

4. Proof 3 - Maximum likelihood regression

You have a fixed set of values, x_i, of a "predictor" variable.

For each x_i, the response Y_i is a random variable whose expectation is $\mu_i = \alpha + \beta x_i$ and whose variance is σ^2. The residuals $Y_i - \mu_i$ are independent.

Given a set of pairs of values $(x_1, Y_1), (x_2, Y_2), \cdots (x_n, Y_n)$, prove that the maximum-likelihood estimates of α and β satisfy the equations

$$\sum_{i=1}^{n}(\hat{\alpha} - \hat{\beta}x_i - Y_i) = 0, \sum_{i=1}^{n} x_i(\hat{\alpha} - \hat{\beta}x_i - Y_i) = 0.$$

and that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

From Proof 1, $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$.

Assumptions: Normal, independent Constant σ

$$P(Y_1, Y_2, \cdots Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2}}$$

$$-\log P = \sum_{i=1}^{n}\left(\log\sqrt{2\pi} + \log\sigma + \frac{1}{\sigma^2}\left(\frac{(\alpha + \beta x_i - Y_i)^2}{2}\right)\right)$$

$$\partial/\partial\alpha(-\log P) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\alpha + \beta x_i - Y_i) = 0$$

$$\partial/\partial\beta(-\log P) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(\alpha + \beta x_i - Y_i) = 0$$

same as for least-squares regression in proof 1

New result. $$\frac{\partial}{\partial\sigma}(-\log P) = \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$$

So $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$$

This has a χ^2 distribution with n-2 degrees of freedom

Unbiased estimator is $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$

5. Logistic regression

You have a fixed set of values, x_i, of a "predictor" variable. Each "response" variable Y_i is a Bernoulli random variable with parameter p_i.

Assume that

$$p_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}.$$

(a) Prove that $\alpha + \beta x_i$ is equal to the "log odds" $\ln \frac{p_i}{1-p_i}$.

(b) Prove that $0 < p_i < 1$.

(c) Given a set of pairs of values $(x_1, Y_1), (x_2, Y_2), \cdots (x_n, Y_n)$, form the likelihood function $L(\alpha, \beta)$ and express its logarithm in terms of α and β. Do not attempt to maximize it!

a. "Odds in favor": $\frac{p_i}{1-p_i} = \frac{e^{\alpha+\beta x_i}/(1+e^{\alpha+\beta x_i})}{1/(1+e^{\alpha+\beta x_i})} = e^{\alpha+\beta x_i}$

"Log odds": $\log \frac{p_i}{1-p_i} = \alpha + \beta x_i$

b. The log odds goes from $-\infty$ to $+\infty$ as p_i goes from 0 to 1

so we can not predict $p_i \leq 0$ or $p_i \geq 1$.

As $\alpha + \beta x_i \to -\infty$ $\quad p_i \to \frac{0}{1+0} = 0$

As $\alpha + \beta x_i \to +\infty$ $\quad p_i = \frac{1}{e^{-(\alpha+\beta x_i)}+1} \to 1$

c. $L(\alpha,\beta) = \prod_{i=1}^{\infty} p_i^{Y_i}(1-p_i)^{1-Y_i}$ note: each Y_i is either 0 or 1

$$\begin{aligned}\log L(\alpha,\beta) &= \sum_{i=1}^{n}\left(Y_i \log p_i + (1-Y_i)\log(1-p_i)\right) \\ &= \sum_{i=1}^{n}\left[Y_i \log \frac{p_i}{1-p_i} + \log(1-p_i)\right] \\ &= \sum_{i=1}^{n}\left[Y_i(\alpha+\beta x_i) - \log\left(1+e^{\alpha+\beta x_i}\right)\right]\end{aligned}$$