

Regression and correlation

Math 36b

May 7, 2009

Contents

1	Linear regression	2
1.1	Estimators	2
1.2	Residuals	3
1.3	Confidence intervals	5
1.4	Expected value of Y	8
	1.4.1 conditional expected value	8
	1.4.2 predicting Y	9
1.5	Comparison of two samples	10
1.6	Covariance and correlation	12
	1.6.1 Mathematical properties of covariance	12
	1.6.2 Statistical properties of covariance	13
	1.6.3 Correlation coefficient	14
	1.6.4 Pearson sample correlation coefficient	16
1.7	Comparing regression and correlation	18
	1.7.1 Is the correlation significant?	18
	1.7.2 Is the regression slope significantly different from zero?	18

1 Linear regression

The basic setup for linear regression is that Y should be a linear function of x with some error. For example we looked at the temperature/pressure relation for a gas. The theoretical relation is:

$$P = e^{a+b/T}$$

where P is pressure and T is absolute temperature. This means if we let $Y = \ln P$ and $x = 1/T$ then we get:

$$Y = a + bx$$

The way that the measurements are done is: We set the temperature and measure the pressure. So, x is not random. It is set by us. But $Y = \ln P$ has random errors of measurement which we assume to be normally distributed. So, for each temperature we get x_i which is not random and

$$Y_i = a + bx_i + \epsilon_i$$

where ϵ_i is assumed to be normal with mean 0:

$$\epsilon_i \sim N(0, \sigma^2)$$

What we want to do is to estimate the three parameters

$$\beta_0 = a, \quad \beta_1 = b, \quad \sigma$$

1.1 Estimators

The least squares estimators for $\beta_0 = a$ and $\beta_1 = b$ are given by minimizing the sum of squares of the error terms $\epsilon_i = Y_i - a - bx_i$:

$$L = \sum (Y_i - a - bx_i)^2$$

Setting $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$ we got:

$$\boxed{\hat{\beta}_1 = \hat{b} = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2}}$$

$$\boxed{\hat{\beta}_0 = \hat{a} = \bar{Y} - \hat{b}\bar{x}}$$

To simplify these equations we use the following notation:

1. $SS_x := \sum_1^n (x_i - \bar{x})^2 = n(\overline{x^2} - \bar{x}^2)$

$$2. S_{xy} := \sum(x_i - \bar{x})(Y_i - \bar{Y}) = n(\overline{xY} - \bar{x}\bar{Y})$$

SS_x is called the *x sum of squares* (as opposed to the “sum of the squares of x ” which is just $\sum x_i^2$.)

If we divide the second expression by the first the n 's cancel and we get:

$$\hat{b} = \frac{S_{xy}}{SS_x}$$

The formula for \hat{a} is easier to understand in the form:

$$\bar{Y} = \hat{a} + \hat{b}\bar{x}$$

1.2 Residuals

The *residuals* R_i are defined by

$$R_i = Y_i - \hat{a} - \hat{b}x_i$$

These are estimators for the error terms ϵ_i .

Theorem 1.1.

$$\frac{\sum R_i^2}{\sigma^2} = \frac{\sum (Y_i - \hat{a} - \hat{b}x_i)^2}{\sigma^2} \sim \chi_{n-2}^2$$

Proof. Each true error term is normal:

$$\epsilon_i = Y_i - a - bx_i \sim N(0, \sigma^2)$$

This means that

$$\sum \frac{\epsilon_i^2}{\sigma^2} = \sum \frac{(Y_i - a - bx_i)^2}{\sigma^2} \sim \chi_n^2$$

If we replace the two variables a, b with estimators \hat{a}, \hat{b} we lose two degrees of freedom which proves the theorem (ignoring for now some technical details). \square

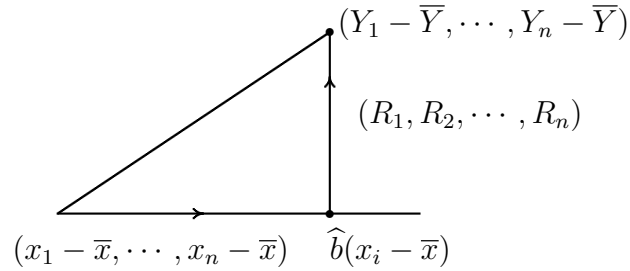
The average value of R_i is always zero:

$$\bar{R} = \bar{Y} - \hat{a} - \hat{b}\bar{x} = 0$$

If we subtract this (zero) from R_i the \hat{a} terms cancel:

$$R_i = (Y_i - \bar{Y}) - \hat{b}(x_i - \bar{x})$$

This is a vector equation which we can draw like this: The sum of squares



of R_i is the length squared of the vector (R_1, R_2, \dots, R_n) :

$$\|R\|^2 = \sum R_i^2$$

Remember that we minimized this sum of squares! In fact, the projection of a vector A to a vector B is given by the formula:

$$Proj_B A = \frac{A \cdot B}{\|B\|^2} B$$

But the dot product is $\sum (Y_i - \bar{Y})(x_i - \bar{x}) = S_{xy}$ and the length squared of $(x_i - \bar{x})$ is $\sum (x_i - \bar{x})^2 = SS_x$. The ratio is

$$\frac{(Y_i - \bar{Y}) \cdot (x_i - \bar{x})}{\|(x_i - \bar{x})\|^2} = \frac{S_{xy}}{SS_x} = \hat{b}$$

This means the point with coordinates $\hat{b}(x_i - \bar{x})$ is the projection of the point $(Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})$ to the line spanned by the vector $(x_i - \bar{x})$. So, the vector R is perpendicular to the vector $(x_i - \bar{x})$ and Pythagoras says:

$$\begin{aligned} \|R\|^2 + \|\hat{b}(x_i - \bar{x})\|^2 &= \|(Y_i - \bar{Y})\|^2 \\ \sum R_i^2 + \hat{b}^2 \sum (x_i - \bar{x})^2 &= \sum (Y_i - \bar{Y})^2 \\ \sum R_i^2 + \frac{S_{xy}^2}{SS_x} &= SS_y \end{aligned}$$

This gives the book's formula for the χ^2 statistic:

$$\frac{\sum R_i^2}{\sigma^2} = \frac{1}{\sigma^2} \left(SS_y - \frac{S_{xy}^2}{SS_x} \right) \sim \chi_{n-2}^2$$

Since the expected value of χ_{n-2}^2 is $n - 2$, an unbiased estimator for σ^2 is given by

$$S^2 := \frac{1}{n-2} \sum R_i^2$$

1.3 Confidence intervals

We want to know how accurate these estimators are. First we want to get a t -distribution out of the estimator for $\beta_1 = b$. We need a definition and a theorem. The definition is:

$$\hat{\sigma}^2 := \frac{1}{n} \sum R_i^2 = \frac{1}{n} \left(SS_y - \frac{S_{xy}^2}{SS_x} \right)$$

This is the MLE biased estimator for σ^2 .

Theorem 1.2.

$$\frac{\hat{b} - b}{\hat{\sigma}} \sqrt{\frac{(n-2)SS_x}{n}} \sim t_{n-2}$$

I proved this in class started with the lemma:

Lemma 1.3.

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{SS_x}\right)$$

Proof of Theorem. Since \hat{b} is normally distributed, we get $Z \sim N(0, 1)$ by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{\hat{b} - b}{\sigma / \sqrt{SS_x}}$$

To get the t -distribution we replace σ with the χ^2 estimator $S = \sqrt{\frac{1}{n-2} \sum R_i^2}$:

$$t_{n-2} = \frac{Z}{\sqrt{U/(n-2)}} = \frac{\hat{b} - b}{S/\sqrt{SS_x}}$$

where $U = (n-2)S^2/\sigma^2 = \frac{1}{\sigma^2} \sum R_i^2 \sim \chi_{n-2}^2$. □

Proof of Lemma. This uses the fact that the sum of the numbers $x_i - \bar{x}$ is zero.

$$\begin{aligned} \hat{b} &= \frac{\sum (Y_i - \bar{Y})(x_i - \bar{x})}{SS_x} = \sum Y_i \frac{x_i - \bar{x}}{SS_x} - \bar{Y} \frac{\sum (x_i - \bar{x})}{SS_x} \\ &= \sum Y_i \frac{x_i - \bar{x}}{SS_x} \end{aligned}$$

Since each Y_i is normal: $Y_i \sim N(a + bx_i, \sigma^2)$, each term in the sum for \hat{b} is normal:

$$Y_i \frac{x_i - \bar{x}}{SS_x} \sim N\left(\frac{(a + bx_i)(x_i - \bar{x})}{SS_x}, \frac{\sigma^2(x_i - \bar{x})^2}{SS_x^2}\right)$$

Add this up for all i to get

$$\hat{b} \sim N \left(\underbrace{\sum (a + bx_i)(x_i - \bar{x}) / SS_x}_{bSS_x}, \sigma^2 \underbrace{\sum (x_i - \bar{x})^2 / SS_x^2}_{SS_x} \right)$$

since $a \sum (x_i - \bar{x}) = 0$ and $b \sum x_i(x_i - \bar{x}) = bSS_x$. □

From this I derived the formula for the 95% confidence interval for b :

$$b = \hat{b} \pm t_{n-2, .975} S / \sqrt{SS_x}$$

$$b = \frac{S_{xy}}{SS_x} \pm t_{(n-2), .975} \sqrt{\frac{SS_y - \frac{S_{xy}^2}{SS_x}}{(n-2)SS_x}}$$

Which we could simplify to:

$$b = \frac{S_{xy}}{SS_x} \pm \frac{t_{(n-2), .975}}{\sqrt{n-2}} \sqrt{\frac{SS_y}{SS_x} - \frac{S_{xy}^2}{SS_x^2}}$$

This formula is designed to be easy to calculate with. For example, if we had data of the form:

x_i	Y_i
1	2
2	7
3	3
4	5

we can calculate $x_i^2, Y_i^2, x_i Y_i$ and add up the rows to get:

x_i	Y_i	x_i^2	Y_i^2	$x_i Y_i$
1	2	1	4	2
2	7	4	49	14
3	3	9	9	9
4	5	16	25	20
10	17	30	87	45
$\sum x_i$	$\sum Y_i$	$\sum x_i^2$	$\sum Y_i^2$	$\sum x_i Y_i$

Divide by n to get $\bar{x}, \bar{Y}, \overline{x^2}, \overline{Y^2}, \overline{xY}$. Then use the formulas:

1. $SS_x = n(\overline{x^2} - \bar{x}^2) = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = 30 - \frac{10^2}{4} = 5$
2. $SS_Y = n(\overline{Y^2} - \bar{Y}^2) = \sum Y_i^2 - \frac{1}{n}(\sum Y_i)^2 = 87 - \frac{17^2}{4} = 14.75$
3. $S_{xy} = n(\overline{xY} - \bar{x}\bar{Y}) = \sum x_i Y_i - \frac{1}{n}(\sum x_i)(\sum Y_i) = 45 - \frac{170}{4} = 2.5$

Take the variance of \hat{b} (from the Lemma):

$$\text{Var}(\hat{b}) = \frac{\sigma^2}{SS_x}$$

Lemma 1.4.

$$\text{Var}(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)$$

Proof. This is easy. Since $\hat{a} = \bar{Y} - \hat{b}\bar{x}$

$$\begin{aligned} \text{Var}(\hat{a}) &= \text{Var}(\bar{Y}) + \text{Var}(\hat{b}\bar{x}) \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \underbrace{\text{Var}(\hat{b})}_{\sigma^2/SS_x} \end{aligned}$$

□

Theorem 1.5.

$$\hat{a} \sim N \left(a, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right) \right)$$

This means that

$$Z = \frac{\hat{a} - a}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}}$$

Replacing σ with $S = \frac{1}{n-2} \sum R_i^2$ we get:

Corollary 1.6.

$$\frac{\hat{a} - a}{S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}} \sim t_{n-2}$$

This means the 95% confidence interval for a is

$$a = \hat{a} \pm t_{n-2, .975} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$$

1.4 Expected value of Y

The regression line gives a prediction as to the values of Y for x not in the data set:

$$\hat{Y} = \hat{a} + \hat{b}x$$

But \hat{a}, \hat{b} are NOT independent. \bar{Y} and \hat{b} are independent. So we have to substitute:

$$\begin{aligned}\hat{a} &= \bar{Y} - \hat{b}\bar{x} \\ \hat{Y} &= \bar{Y} - \hat{b}\bar{x} + \hat{b}x = \bar{Y} + \hat{b}(x - \bar{x})\end{aligned}$$

1.4.1 conditional expected value

is given by:

$$\begin{aligned}\mathbb{E}(Y|x) &= \mathbb{E}(\hat{Y}) = a + bx \\ \text{Var}(\hat{Y}) &= \text{Var}(\bar{Y}) + (x - \bar{x})^2 \text{Var}(\hat{b}) \\ &= \frac{\sigma^2}{n} + (x - \bar{x})^2 \frac{\sigma^2}{SS_x} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x} \right)\end{aligned}$$

This gives:

$$\begin{aligned}Z &= \frac{\hat{Y} - (a + bx)}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}} \\ t_{n-2} &\sim \frac{\hat{Y} - (a + bx)}{S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}}\end{aligned}$$

where

$$S = \sqrt{\frac{\sum R_i^2}{n - 2}}$$

This means the 95% confidence interval for the “extrapolated” value $\mathbb{E}(Y|x) = a + bx$ is

$$a + bx = \hat{a} + \hat{b}x \pm t_{n-2, .975} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$$

1.4.2 predicting Y

The true value of Y will have an error: $Y = a + bx + \epsilon$ with $\text{Var}(\epsilon) = \sigma^2$ which is independent of the predicted value \hat{Y} . So, the difference $\hat{Y} - Y$ has variance

$$\text{Var}(\hat{Y} - Y) = \text{Var}(\hat{Y}) + \text{Var}(Y) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x} + 1 \right)$$

and $\mathbb{E}(\hat{Y} - Y) = 0$. So, the 95% confidence interval for Y is

$$\hat{Y} \pm t_{n-2, .975} S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$$

1.5 Comparison of two samples

We did an example of comparison of two sample. This was problem 11.3.22 in the book. The two samples are approval ratings of two mayors one Democrat and one Republican. A linear regression chart shows that the approval rating dropped linearly for both and we want to compare the two slopes at $\alpha = 5\%$.

We have:

$$Y_i = a + bx_i + \epsilon_i \quad i = 1, 2, \dots, n \quad (n = 5)$$

$$Y_j^* = a^* + b^*x_j^* + \epsilon_j^* \quad j = 1, 2, \dots, m \quad (m = 6)$$

We get two χ^2 variables which are independent:

$$\frac{\sum_{i=1}^n R_i^2}{\sigma^2} \sim \chi_{n-2}^2, \quad \frac{\sum_{j=1}^m R_j^{*2}}{\sigma^{*2}} \sim \chi_{m-2}^2$$

We assume that $\sigma = \sigma^*$. Then, we can add these and get:

$$\frac{\sum R_i^2 + \sum R_j^{*2}}{\sigma^2} \sim \chi_{n+m-4}^2$$

Since this has expected value equal to $n+m-4$ we get an unbiased estimator of σ^2 :

$$S_P^2 := \frac{\sum R_i^2 + \sum R_j^{*2}}{n+m-4}$$

This is the *pooled variance*, the best estimator we have for σ^2 using all the data.

Next we need the variance of $\widehat{b} - \widehat{b}^*$.

$$\text{Var}(\widehat{b} - \widehat{b}^*) = \text{Var}(\widehat{b}) + \text{Var}(\widehat{b}^*)$$

(using the rule that $\text{Var}(cX) = c^2 \text{Var}(X)$ to $c = -1$). This is equal to:

$$= \frac{\sigma^2}{SS_x} + \frac{\sigma^{*2}}{SS_x^*}$$

Assuming that $\sigma = \sigma^*$ this is:

$$= \sigma^2 \left(\frac{1}{SS_x} + \frac{1}{SS_x^*} \right)$$

So,

$$Z = \frac{(\widehat{b} - \widehat{b}^*) - (b - b^*)}{\sigma \sqrt{\frac{1}{SS_x} + \frac{1}{SS_x^*}}}$$

Replacing σ with the χ^2 estimators S_P we get

$$t_{n+m-4} \sim \frac{(\widehat{b} - \widehat{b}^*) - (b - b^*)}{S_P \sqrt{\frac{1}{SS_x} + \frac{1}{SS_x^*}}}$$

To do the actual problem, first I checked that the variances could be equal:

$$F_{n-2, m-2} = F_{3,4} = \frac{S^2}{S^{*2}} = \frac{\sum R_i^2/3}{\sum R_j^{*2}/4} = 0.536401733$$

which is not significant. So, we assume that $\sigma = \sigma^*$ and use the pooled variance

$$S_P^2 = \frac{\sum R_i^2 + \sum R_j^{*2}}{7} = 1.225740361$$

This gives the estimate for the standard deviation of $\widehat{b} - \widehat{b}^*$:

$$S_P \sqrt{\frac{1}{SS_x} + \frac{1}{SS_x^*}} = 0.279833434$$

Assuming the null hypothesis that $b = b^*$ we get

$$t_7 = \frac{\widehat{b} - \widehat{b}^*}{S_P \sqrt{\frac{1}{SS_x} + \frac{1}{SS_x^*}}} = \frac{-3.461538462 - (-2.737288136)}{0.279833434}$$

$$t_7 = -2.588147944$$

Since this is greater (in absolute value) than the critical value:

$$t_{7,.975} = 2.36462256$$

we reject the null hypothesis and conclude that the slopes of these two lines are significantly different. In other words, the rate at which the two mayors lose popularity is different. Surprisingly, it would not be correct to conclude that the Democratic mayor lost popularity at a faster rate than the Republican mayor even though this seems to be the obvious conclusion.

1.6 Covariance and correlation

Suppose that (X, Y) is a random point in the plane. If we take a sample we get several points $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. Each pair of numbers is taken from one person or thing. So, X_1, Y_2 are independent but X_1, Y_1 are usually not independent.

Definition 1.7. $\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y))$

1.6.1 Mathematical properties of covariance

Theorem 1.8. *Cov is symmetric and bilinear.*

Proof. *Symmetric* means $\text{Cov}(X, Y) = \text{Cov}(Y, X)$. This is obvious.

Bilinear means symmetric in each variable. By symmetry we only need to show linearity in X . This means that, if $X = aX_1 + bX_2$ then

$$\text{Cov}(X, Y) = a \text{Cov}(X_1, Y) + b \text{Cov}(X_2, Y)$$

The proof of this is also easy: We start with the formula for the expected value of X :

$$\mu_X = \mathbb{E}(X) = a\mathbb{E}(X_1) + b\mathbb{E}(X_2) = a\mu_{X_1} + b\mu_{X_2}$$

Subtracting this from $X = aX_1 + bX_2$ we get:

$$X - \mu_X = a(X_1 - \mu_{X_1}) + b(X_2 - \mu_{X_2})$$

So,

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \mathbb{E}([a(X_1 - \mu_{X_1}) + b(X_2 - \mu_{X_2})](Y - \mu_Y)) \\ &= a\mathbb{E}((X_1 - \mu_{X_1})(Y - \mu_Y)) + b\mathbb{E}((X_2 - \mu_{X_2})(Y - \mu_Y)) \\ &= a \text{Cov}(X_1, Y) + b \text{Cov}(X_2, Y) \end{aligned}$$

□

Theorem 1.9. *Addition of a constant to X or Y does not change the covariance:*

$$\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$$

Proof. Adding a constant to X adds the same constant to its average value μ_X and similarly for Y :

$$\mu_{X+a} = \mu_X + a, \quad \mu_{Y+b} = \mu_Y + b$$

Then

$$X - \mu_X = X + a - (\mu_X + a) = X - \mu_X$$

and the same for $Y + B$. So,

$$\begin{aligned} \text{Cov}(X + a, Y + b) &= \mathbb{E}((X + a - (\mu_X + a))(Y + b - (\mu_Y + b))) \\ &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \text{Cov}(X, Y) \end{aligned}$$

□

1.6.2 Statistical properties of covariance

The main point is:

Theorem 1.10. *If X and Y are independent then $\text{Cov}(X, Y) = 0$.*

Proof. When two random variables are independent, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. So,

$$\mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \underbrace{\mathbb{E}(X - \mu_X)}_0 \underbrace{\mathbb{E}(Y - \mu_Y)}_0 = 0$$

Since $\mathbb{E}(X - \mu_X) = \mathbb{E}(X) - \mu_X = \mu_X - \mu_X = 0$. □

When X, Y are not independent, we can still cancel many of the terms in the expansion:

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \mathbb{E}(XY - \mu_X Y - X \mu_Y + \mu_X \mu_Y) \\ 0 &= \mathbb{E}((X - \mu_X)\mu_Y) = \mathbb{E}(X\mu_Y - \mu_X \mu_Y) \\ 0 &= \mathbb{E}(\mu_X(Y - \mu_Y)) = \mathbb{E}(\mu_X Y - \mu_X \mu_Y) \end{aligned}$$

Adding these up and canceling terms we get:

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y$$

In the special case when $X = Y$, we get:

Theorem 1.11. $\text{Cov}(X, X) = \text{Var}(X)$

Proof. Both formulas give the variance of X :

$$\text{Cov}(X, X) = \mathbb{E}((X - \mu_X)^2) = \mathbb{E}(X^2) - \mu_X^2 = \text{Var}(X)$$

□

There is an obvious estimator for the covariance (given by MOM: replacing expected values by sample means):

$$\widehat{\text{Cov}}(X, Y) = \overline{XY} - \overline{X}\overline{Y} = \frac{1}{n}S_{xy}$$

However, this is biased. An unbiased estimator is given by $\frac{1}{n-1}S_{xy}$

Theorem 1.12. $\mathbb{E}(S_{xy}) = (n - 1) \text{Cov}(X, Y)$

Proof. This follows from the fact that Cov is bilinear and X_i, Y_j are independent for $i \neq j$ and $\text{Cov}(X_i, Y_i)$ is equal to the same number C for all i . So,

$$\mathbb{E}(S_{xy}) = \sum_{i=1}^n \mathbb{E}((X_i - \overline{X})(Y_i - \overline{Y})) = \sum_{i=1}^n \text{Cov}(X_i - \overline{X}, Y_i - \overline{Y})$$

$$= \sum \text{Cov} \left(X_i - \frac{1}{n} \sum X_j, Y_i - \frac{1}{n} \sum Y_k \right)$$

Since $\text{Cov}(X_i, Y_j) = 0$, all the cross terms vanish and we get only the terms where $j = k$. These include the term where $j = k = i$ and the $n - 1$ terms where $j = k \neq i$:

$$\begin{aligned} &= \sum \text{Cov} \left(X_i - \frac{1}{n} X_i, Y_i - \frac{1}{n} Y_i \right) + \sum \sum_{j \neq i} \text{Cov} \left(\frac{1}{n} X_j, \frac{1}{n} Y_j \right) \\ &= \left[n \left(1 - \frac{1}{n} \right)^2 + n(n - 1) \left(\frac{1}{n} \right)^2 \right] \text{Cov}(X, Y) \end{aligned}$$

This is equal to $(n - 1) \text{Cov}(X, Y)$. □

1.6.3 Correlation coefficient

We start with a simple theorem and interpret it in terms of vectors.

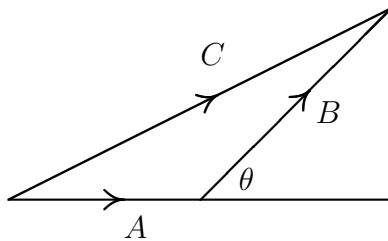
Theorem 1.13. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$

Proof. This just follows from the bilinearity of covariance:

$$\begin{aligned} \text{Var}(X + Y) &= \text{Cov}(X + Y, X + Y) \\ &= \text{Cov}(X, X) + \text{Cov}(X, Y) + \text{Cov}(Y, X) + \text{Cov}(Y, Y) \\ &= \text{Var}(X) + 2 \text{Cov}(X, Y) + \text{Var}(Y) \end{aligned}$$

□

Compare this with the Law of Cosines:



$$\begin{aligned} \vec{C} &= \vec{A} + \vec{B} \\ \|\vec{C}\|^2 &= \|\vec{A}\|^2 + \|\vec{B}\|^2 \\ &\quad + 2\|\vec{A}\| \cdot \|\vec{B}\| \cos \theta \end{aligned}$$

If we imagine that $\|\vec{A}\|^2 = \text{Var}(A) = \sigma_X^2$ and $\|\vec{B}\|^2 = \text{Var}(B) = \sigma_Y^2$ then the covariance becomes $\sigma_X \sigma_Y \cos \theta$ and we get the following definition:

Definition 1.14. The *coefficient of correlation* between X and Y is

$$\rho(X, Y) = \cos \theta := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

In order for this to make sense we need to know that this fraction is ≤ 1 in absolute value:

Theorem 1.15. (a) $|\rho(X, Y)| \leq 1$

(b) $|\rho(X, Y)| = 1$ if and only if $Y = aX + b$ almost surely (with probability one).

The proof of this theorem uses the following lemma:

Lemma 1.16. $\rho(X, Y) = \rho(X^*, Y^*)$ where

$$X^* := \frac{X - \mu_X}{\sigma_X}, \quad Y^* := \frac{Y - \mu_Y}{\sigma_Y}$$

Proof. This just follows from the fact that covariance is bilinear and X^*, Y^* have mean 0 and variance 1:

$$\begin{aligned} \rho(X^*, Y^*) &= \text{Cov}(X^*, Y^*) = \text{Cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) \\ &= \frac{\text{Cov}(X - \mu_X, Y - \mu_Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \rho(X, Y) \end{aligned}$$

□

Proof of Theorem. The theorem follows from the following two facts:

1. The variance of any random variable is ≥ 0 . (And it could be infinite.)
2. $\text{Var}(X) = 0$ if and only if X is a constant (a.s.).

First, compute $\text{Var}(X^* \pm Y^*)$:

$$\text{Var}(X^* \pm Y^*) = \underbrace{\text{Var}(X^*)}_1 + \underbrace{\text{Var}(Y^*)}_1 \pm 2 \text{Cov}(X^*, Y^*)$$

This would be negative if $\text{Cov}(X^*, Y^*)$ is either greater than 1 or less than -1. Therefore,

$$1 \geq \rho(X, Y) = \rho(X^*, Y^*) = \text{Cov}(X^*, Y^*) \geq -1$$

Finally, if $\rho(X, Y) = \text{Cov}(X^*, Y^*) = 1$ then $\text{Var}(X^* - Y^*) = 0$ making $X^* = Y^*$ almost surely. Similarly, if $\rho(X, Y) = \text{Cov}(X^*, Y^*) = -1$ then $X^* = -Y^*$ almost surely. In either case, Y would be a linear function of X a.s. □

1.6.4 Pearson sample correlation coefficient

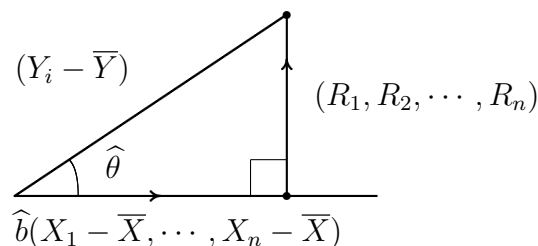
Definition 1.17. The *sample correlation coefficient* is defined to be:

$$\hat{\rho} = R = \frac{S_{xy}}{\sqrt{SS_x SS_y}}$$

This is a good estimator for $\rho(X, Y)$ since the expected value of S_{xy} is $(n - 1) \text{Cov}(X, Y)$ and the expected values of SS_x and SS_y are $(n - 1) \text{Var}(X)$ and $(n - 1) \text{Var}(Y)$ respectively. So,

$$R = \frac{S_{xy}}{\sqrt{SS_x SS_y}} \approx \frac{(n - 1) \text{Cov}(X, Y)}{\sqrt{(n - 1)^2 \text{Var}(X) \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \rho(X, Y)$$

Sample correlation is very similar to the regression slope estimator $\hat{b} = \frac{S_{xy}}{SS_x}$. The relation can be seen from the triangle:



We saw this right triangle before. The base has length

$$\hat{b} \sqrt{SS_x} = \frac{S_{xy}}{\sqrt{SS_x}}$$

The hypotenuse has length

$$\sqrt{\sum (Y_i - \bar{Y})^2} = \sqrt{SS_y}$$

So, the sample correlation is

$$R = \frac{S_{xy}}{\sqrt{SS_x SS_y}} = \cos \hat{\theta}$$

We also need to notice that:

$$1 - R^2 = \frac{\sum R_i^2}{SS_y} = \sin^2 \hat{\theta}$$

Theorem 1.18. *Suppose that X, Y are independent normal. Then*

$$t_{n-2} \sim \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

Note that

$$\frac{R}{\sqrt{1-R^2}} = \frac{\cos \hat{\theta}}{\sin \hat{\theta}} = \cot \hat{\theta}$$

So, the theorem says

$$\cos \hat{\theta} \sqrt{n-2} \sim t_{n-2}$$

1.7 Comparing regression and correlation

(From Quiz 3): We are given the height and serum cholesterol measurements for 16 men. The null hypothesis is that these two factors are not related.

Aside from $n = 16$ the only three numbers we need are:

$$SS_x = \sum x_i^2 - n\bar{x}^2 = 2583.234375$$

$$SS_y = \sum y_i^2 - n\bar{y}^2 = 670.9375$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 580.78125$$

1.7.1 Is the correlation significant?

$$r = \frac{S_{xy}}{\sqrt{SS_x SS_y}} = 0.441153418$$

$$t_{n-2} = t_{14} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 1.839298678$$

This is less than the critical value of t which is

$$t_{14,.975} = 2.144788596$$

Therefore, the result is not significant. The sample does not show any relation between height and cholesterol.

1.7.2 Is the regression slope significantly different from zero?

$$\hat{b} = \frac{S_{xy}}{SS_x} = 0.224827161$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 45.34304742$$

To test significance we find the 95% confidence interval for $b = \beta_1$. If 0 is in this interval, the slope is not significant.

$$\sum R_i^2 = SS_y - \frac{S_{xy}^2}{SS_x} = 540.3621006$$

$$S^2 = \frac{1}{n-2} \sum R_i^2 = 38.5972929$$

$$S = 6.212671961$$

Under the null hypothesis $b = 0$, we have

$$t_{n-2} = \frac{\hat{b}}{S/\sqrt{SS_x}} = 1.839298678$$

So, we accept. This is exactly the same test statistic as in the correlation analysis!

The 95% confidence interval is given by:

$$\begin{aligned} \hat{b} \pm \frac{S}{\sqrt{SS_x}} t_{n-2, .975} = \\ (-0.037341641, 0.486995962) \end{aligned}$$

Since 0 is in this interval we see again that we should accept the null hypothesis. This is the same test done a third time.

Theorem 1.19. *The t -statistics for the significance of the regression slope β_1 and of the sample correlation R are equal.*

Proof. The t -statistic for correlation is:

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{\frac{S_{xy}\sqrt{n-2}}{\sqrt{SS_x SS_y}}}{\sqrt{1 - \frac{S_{xy}^2}{SS_x SS_y}}} = \frac{S_{xy}\sqrt{n-2}}{\sqrt{SS_x SS_y - S_{xy}^2}} \sim t_{n-2}$$

The t -statistic for regression under the null hypothesis that $\beta_1 = 0$ is:

$$\frac{\hat{b}}{S/\sqrt{SS_x}} = \frac{S_{xy}/SS_x}{S/\sqrt{SS_x}} = \frac{S_{xy}}{S\sqrt{SS_x}} \sim t_{n-2}$$

But,

$$S\sqrt{SS_x} = \frac{\sqrt{SS_x}}{\sqrt{n-2}} \sqrt{SS_y - \frac{S_{xy}^2}{SS_x}} = \frac{1}{\sqrt{n-2}} \sqrt{SS_x SS_y - S_{xy}^2}$$

So,

$$\frac{S_{xy}}{S\sqrt{SS_x}} = \frac{S_{xy}}{\frac{1}{\sqrt{n-2}} \sqrt{SS_x SS_y - S_{xy}^2}} = \frac{S_{xy}\sqrt{n-2}}{\sqrt{SS_x SS_y - S_{xy}^2}}$$

□