STAT 512: Statistical Inference **Autumn 2020**

Lecture 5: Correlation, prediction, and regression

Instructor: Yen-Chi Chen

5.1 Correlation

The (Pearson's) correlation is a common measure between the association of two random variables. Formally, for random variables X and Y, their correlation is

$$\rho_{XY} = \mathsf{Cor}(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}.$$

It has three nice properties:

- $\mathsf{Cor}(X, Y) = \mathsf{Cor}(Y, X)$.
- *Location-scale:* $\mathsf{Cor}(aX + b, cY + d) = \mathsf{sign}(ac)\mathsf{Cor}(X, Y)$.
- $-1 \leq \mathsf{Cor}(X, Y) \leq 1$. $\mathsf{Cor}(X, Y) = \pm 1$ if and only if they are perfectly linear, i.e., $X = aY + b$ for some constant a, b.

You can think of correlation as a measure of the *linear* relationship between two random variables. A large correlation implies a strong linear relationship. However, *a low correlation does not imply the two random variables are not related with each other!*

Example (0 correlation but perfectly related). Consider a random variable X take three possible values $-1, 0, 1$ with a probability $P(X = -1) = P(X = 1) = 1/4$ and $P(X = 0) = 1/2$. Let $Y = X^2$. You can see that X and Y are deterministically related. It is easy to see that $\mathbb{E}[X] = 0$ and $\mathbb{E}[XY] = 0$ and $\mathsf{Var}(X), \mathsf{Var}(Y) > 0$. However, the covariance between them will be

$$\mathsf{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 = 0,$$

which implies that $\mathsf{Cor}(X, Y) = 0$. So they are uncorrelated but perfectly related with each other.

5.2 Mean-square error prediction

A classical problem in statistics is prediction. The prediction problem is the case where we are given a random variable X and we want to use it to predict another random variable Y. Thus, we can think of using a quantity $g(X)$ as a prediction of Y.

To measure how good the predictor $g(X)$ is, we often use the *mean-square error (MSE)*:

$$R(g) = \mathbb{E}((Y - g(X))^2).$$

Namely, the MSE is the expected squared deviation from our predictor $g(X)$ to the target Y.

Ideally, we want to choose g that minimizes $R(g)$. Formally, we want to find

$$g^* = \mathsf{argmin}_g R(g).$$

We now take a deeper look at the MSE $R(g) = \mathbb{E}((Y - g(X))^2)$. Using the law of total expectation,

$$\mathbb{E}((Y - g(X))^2) = \mathbb{E}[\mathbb{E}[(Y - g(X))^2|X]].$$

Using the fact that for any fixed constant c,

$$\mathbb{E}[(Y - c)^2] = \mathbb{E}[(Y - \mathbb{E}[Y] + E[Y] - c)^2] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] + (\mathbb{E}[Y] - c)^2 = \mathsf{Var}(Y) + (\mathbb{E}[Y] - c)^2,$$

we can rewrite the MSE as

$$R(g) = \mathbb{E}[\mathbb{E}[(Y - g(X))^2|X]] = \mathbb{E}[\mathsf{Var}(Y|X) + (\mathbb{E}[Y|X] - g(X))^2] = \mathbb{E}[\mathsf{Var}(Y|X)] + \mathbb{E}[(\mathbb{E}[Y|X] - g(X))^2].$$

The first quantity is independent of g so it does not matter in the selection of g. The second quantity involves

$$(\mathbb{E}[Y|X] - g(X))^2 \geq 0.$$

The only case that the equality holds is $g(X) = \mathbb{E}[Y|X]$. As a result, to minimize the MSE, we should use the conditional expectation $\mathbb{E}[Y|X]$ as our predictor. The conditional expectation $\mathbb{E}[Y|X = x] = m(x)$ is also known as the *regression function* or the *best predictor*.

With the regression function, we can decompose Y as

$$Y = \underbrace{\mathbb{E}[Y|X]}_{\text{best predictor}} + \underbrace{(Y - \mathbb{E}[Y|X])}_{\text{residuals}}. \tag{5.1}$$

Here are some interesting properties of the decomposition in equation (5.1):

- **Unbiased.** $\mathbb{E}[\text{best predictor}] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ and $\mathbb{E}[\text{residual}] = 0$.
- **Uncorrelated.** $\mathsf{Cov}(\mathbb{E}[Y|X], Y - \mathbb{E}[Y|X]) = 0$.
- **Residual variance.** $\mathsf{Var}(Y - \mathbb{E}[Y|X]) = \mathbb{E}[\mathsf{Var}(Y|X)]$. To see this,

$$\begin{aligned}
\mathsf{Var}(Y - \mathbb{E}[Y|X]) &= \mathsf{Var}(Y) - 2\mathsf{Cov}(Y, \mathbb{E}[Y|X]) + \mathsf{Var}(\mathbb{E}[Y|X]) \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 - 2(\mathbb{E}[Y\mathbb{E}[Y|X]] - \mathbb{E}[Y]\mathbb{E}[\mathbb{E}[Y|X]]) + \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 - 2\mathbb{E}[\mathbb{E}[Y|X]^2] + 2\mathbb{E}[Y]^2 + \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2 \\
&= \mathbb{E}[Y^2] - \mathbb{E}[\mathbb{E}[Y|X]^2] \\
&= \mathbb{E}[\mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2] \\
&= \mathbb{E}[\mathsf{Var}(Y|X)].
\end{aligned}$$

- **Variance decomposition.** With the above properties, we obtain

$$\mathsf{Var}(Y) = \mathsf{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathsf{Var}(Y|X)].$$

  Although this is the same formula as the law of total variance, it now can be interpreted as:

$$\mathsf{Var}(Y) = \underbrace{\mathsf{Var}(\mathbb{E}[Y|X])}_{\mathsf{Var}(\text{best predictor})} + \underbrace{\mathbb{E}[\mathsf{Var}(Y|X)]}_{\text{average } \mathsf{Var}(\text{residuals})}.$$

5.3 Linear prediction (linear regression)

In the above analysis, we see that the best predictor is the conditional expectation. However, this is attainable if we consider all possible function $g(x)$. In reality, we may only restrict ourselves to some set of simple functions. One canonical example is the linear functions. Namely, suppose we want to find α, β such that the MSE

$$R(\alpha, \beta) = \mathbb{E}((Y - \alpha - \beta X)^2)$$

is minimized (this is also known as the *least square* approach). How will we choose α and β?

To solve this problem, we first expand $R(\alpha, \beta)$:

$$\begin{aligned} R(\alpha, \beta) &= \mathbb{E}((Y - \alpha - \beta X)^2) \\ &= \mathbb{E}(Y^2 + \alpha^2 + \beta^2 X^2 - 2Y\alpha - 2XY\beta + 2\alpha\beta X) \\ &= \mathbb{E}[Y^2] + \alpha^2 + \beta^2 \mathbb{E}[X^2] - 2\alpha\mathbb{E}[Y] - 2\beta\mathbb{E}[XY] + 2\alpha\beta\mathbb{E}[X], \end{aligned}$$

which is a quadratic function of α, β. Thus,

$$\alpha^*, \beta^* = \mathsf{argmin}_{\alpha,\beta} R(\alpha, \beta)$$

solves the gradient equation (known as the first order equation):

$$\begin{aligned} 0 &= \frac{\partial}{\partial \alpha} R(\alpha^*, \beta^*) \\ &= 2\alpha^* - 2\mathbb{E}[Y] - 2\beta^* \mathbb{E}[X] \\ 0 &= \frac{\partial}{\partial \beta} R(\alpha^*, \beta^*) \\ &= 2\beta^* \mathbb{E}[X^2] - 2\mathbb{E}[XY] + 2\alpha^* \mathbb{E}[X] \\ \Rightarrow \quad & \beta^* \mathsf{Var}(X) = \mathsf{Cov}(X, Y) \\ \Rightarrow \quad & \beta^* = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(X)} \\ \Rightarrow \quad & \alpha^* = \mathbb{E}[Y] - \mathbb{E}[X]\beta^*. \end{aligned}$$

With these, the *best linear predictor (BLP)* is

$$\begin{aligned} m^*(x) &= \alpha^* + \beta^* x \\ &= \mathbb{E}[Y] + \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(X)}(x - \mathbb{E}[X]) \\ &= \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X}(x - \mu_X), \end{aligned}$$

where $\mu_X = \mathbb{E}[X], \mu_Y = \mathbb{E}[Y], \sigma_X^2 = \mathsf{Var}(X), \sigma_Y^2 = \mathsf{Var}(Y)$ and ρ_{XY} is the Pearson's correlation.

Interestingly, the MSE under the best linear predictor will be

$$\begin{aligned} R(\alpha^*, \beta^*) &= \mathbb{E}((Y - \alpha^* - \beta^* X)^2) \\ &= \mathbb{E}[(Y - \mu_Y - \rho_{XY} \frac{\sigma_Y}{\sigma_X}(X - \mu_X))^2] \\ &= \sigma_Y^2 - 2\rho_{XY} \frac{\sigma_Y}{\sigma_X} \mathbb{E}[(Y - \mu_Y)(X - \mu_X)] + \rho_{XY}^2 \frac{\sigma_Y^2}{\sigma_X^2} \sigma_X^2 \\ &= \sigma_Y^2 (1 - \rho_{XY}^2). \end{aligned}$$

An important feature of the above analysis is that *we did NOT assume the linear model to be correct!* We can always find a best linear predictor regardless of what the true regression function looks like.

5.3.1 Multivariate linear prediction

Suppose that the covariate $X = (X_1, \cdots, X_p)$ is now multivariate. The linear prediction approach still works and the MSE will be

$$R(\alpha, \beta) = \mathbb{E}((Y - \alpha - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_p X_p)^2),$$

where $\beta = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$. Let $Z = (1, X_1, \cdots, X_p)^T \in \mathbb{R}^{p+1}$ be a *data vector* and $\gamma = (\alpha, \beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$ be a coefficient vector. Then the MSE has an elegant form:

$$R(\gamma) = R(\alpha, \beta) = \mathbb{E}((Y - \gamma^T Z)^2).$$

A direct expansion shows that

$$\begin{aligned} R(\gamma) &= \mathbb{E}(Y^T Y) - 2\mathbb{E}[Y Z^T \gamma] + \mathbb{E}[\gamma^T Z Z^T \gamma] \\ &= \mathbb{E}(Y^T Y) - 2\gamma^T \mathbb{E}[ZY] + \gamma^T \mathbb{E}[ZZ^T]\gamma, \end{aligned}$$

which is a quadratic function of γ.

Differentiating this with respect to the coefficient vector γ leads to

$$0 = -2\mathbb{E}[ZY] + 2\mathbb{E}[ZZ^T]\gamma.$$

Thus, the least square solution will be

$$\gamma^* = \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZY].$$

Note that $\mathbb{E}[ZZ^T]$ is a matrix and $\mathbb{E}[ZZ^T]^{-1}$ is the matrix inverse.

With $\gamma^* = (\alpha^*, \beta^*)^T$, we can easily write down the BLP:

$$m^*(x) = \gamma^{*T} z = \alpha^* + \beta^{*T} x = \alpha^* + \sum_{j=1}^{p} \beta_j^* x_j.$$

5.3.2 Correctness of model

In linear prediction, we did NOT assume the linear model to be correct. In the case of the linear model being correct, we have some really nice properties. We use the notation from the multivariate case for simplicity.

When the linear model is incorrect. The coefficient from the least square approach is $\gamma^* = \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZY]$. In general, this quantity will change if the distribution of the covariate X (Z) changes. So the coefficient depends on the distribution of the covariates.

When the linear model is correct. Suppose that the linear model is correct, i.e., $Y = \bar{\gamma}^T Z + \epsilon$ for some $\bar{\gamma} \in \mathbb{R}^{p+1}$, and ϵ is a noise such that $\epsilon \perp Z$ and $\mathbb{E}[\epsilon|Z] = 0$.

Then the coefficient that minimizes the MSE will be

$$\begin{aligned} \gamma^* &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZY] \\ &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[Z(\bar{\gamma}^T Z + \epsilon)] \\ &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[Z(Z^T\bar{\gamma} + \mathbb{E}[\epsilon|Z])] \\ &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZZ^T\bar{\gamma}] \\ &= \bar{\gamma}. \end{aligned}$$

Thus, the least square coefficient is the same as the true coefficient. This also implies that the least square coefficient will be *invariant* to the distribution of the covariate when the linear model is correct.

5.4 Prediction of categorical outcomes: classification

Classification is one of the most important data analysis problems. Much early work on this topic was done by statisticians but in the past 20 years, computer science and machine learning communities have made much much more progress on this topic.

Here are some classical applications of classification.

- **Email spam.** Email service provider such as Google often faced the problem of classifying a new email. The problem they want to address is like: given an email, how do you decide if it is a spam, an ordinary email, or an important one?
- **Image classification.** If you have used Facebook, you may notice that whenever your photo contains a picture of one of your friends, Facebook may ask you if you want to tag your friend, even if you did not manually tell the computer that this is your friend. How do they know that picture is a human and that guy is your friend?

We consider a simple scenario – binary classification. Namely, there are only two possible classes that we will consider. We will just denote the two classes as 0 and 1.

The classification problem can be formalized as follows. Given a feature vector x_0, we want to create a **classifier** c that maps x_0 into 0 or 1. Namely, we want to find a function $c(x_0)$ that outputs only two possible number: 0 and 1. Moreover, we want to make sure that our *classification error* is small. Let y_0 be the actual label of x_0. We measure the classification error using a **loss function** L such that the the loss of making a prediction $c(x_0)$ when the actual class label is y_0 is $L(c(x_0), y_0)$. A common loss function is the **0-1 loss**, which is $L(c(x_0), y_0) = I(c(x_0) \neq y_0)$. Namely, when we make a wrong classification, we loss 1 point and we do not lose anything if we make the correct classification.

How do we find the classifier c? A good news is: often we have a labeled sample (data) $(X_1, Y_1), \cdots, (X_n, Y_n)$ available. Then we will find c using this dataset.

In statistics, we often model the data as an IID random sample from a distribution. We now define several useful distribution functions:

$$\begin{aligned} p_0(x) = p(X = x|Y = 0): &\quad \text{the density of } X \text{ when the actual label is 0,} \\ p_1(x) = p(X = x|Y = 1): &\quad \text{the density of } X \text{ when the actual label is 1,} \\ P(y|x) = P(Y = y|X = x): &\quad \text{the probability of being in the class } y \text{ when the feature is } x, \\ P_Y(y) = P(Y = y): &\quad \text{the probability of observing the class } y, \text{ regardless of the feature value.} \end{aligned} \tag{5.2}$$

Using a probability model, we will define the **risk function**, which is the expected value of the loss function when the input is random. The risk of a classifier c is

$$R(c) = \mathbb{E}(L(c(X), Y)).$$

Ideally, we want to find a classifier that minimizes the risk because such a classifier will minimize our *expected losses*. In the linear prediction problem, the loss function is the square loss $L(a, b) = (a-b)^2$ and the resulting risk function is the MSE.

Assume that we know the 4 quantities in equation (5.2), what class label will you predict when seeing a feature $X = x$? An intuitive choice is that we should predict the value y that maximizer $P(y|x)$. Namely, we predict the label using the one with highest probability. Such classifier can be written as

$$c_*(x) = \mathsf{argmax}_{y=0,1} P(y|x) = \begin{cases} 0, & \text{if } P(0|x) \geq P(1|x), \\ 1, & \text{if } P(1|x) > P(0|x). \end{cases} \tag{5.3}$$

Is this classifier good in the sense of the classification error (risk)? The answer depends on the loss function. A good news is: this classifier is the optimal classifier for the $0-1$ loss. Namely,

$$R(c_*) = \min_c R(c)$$

when using a $0-1$ loss. However, if we are using other loss function, this classifier will not be the best one (with the smallest expected loss).

Derivation of c_* is optimal under $0-1$ loss. Given a classifier c, the risk function $R(c) = \mathbb{E}(L(c(X), Y))$. Using the property of expectation, we can further write it as

$$R(c) = \mathbb{E}(L(c(X), Y)) = \mathbb{E}(\underbrace{\mathbb{E}(L(c(X), Y)|X)}_{(A)}).$$

For the quantity (A), we have

$$\begin{aligned} \mathbb{E}(L(c(X), Y)|X) &= L(c(X), 1)p(Y = 1|X) + L(c(X), 0)p(Y = 0|X) \\ &= I(c(X) \neq 1)p(Y = 1|X) + I(c(X) \neq 0)p(Y = 0|X) \\ &= \begin{cases} p(Y = 1|X) & \text{if } c(X) = 0 \\ p(Y = 0|X) & \text{if } c(X) = 1. \end{cases} \end{aligned}$$

Thus, seeing a feature X, the expected loss we have when predicting $c(X) = 0$ is $P(Y = 1|X)$ whereas when prediction $c(X) = 1$ is $P(Y = 0|X)$. The optimal choice is predicting $c(X) = 0$ if $P(Y = 1|X) \leq P(Y = 0|X)$ and $c(X) = 1$ if $P(Y = 1|X) > P(Y = 0|X)$ (the equality does not matter), which is the classifier c_*.

When a classifier attains the optimal risk (i.e., having a risk of $\min_c R(c)$), it is called a **Bayes classifier**. Thus, the classifier c_* is the Bayes classifier in $0-1$ loss.

For a classifier c, we define its **excess risk (regret)** as

$$\mathcal{E}(c) = R(c) - \min_c R(c).$$

The excess risk is a quantity that measures how the quality of c is away from the optimal/Bayes classifier. If we cannot find the Bayes classifier, we will at least try to find a classifier whose excess risk is small.