Lecture 31

31.1 Statistical inference in simple linear regression.

Let us first summarize what we proved in the last two lectures. We considered a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where ε has distribution $N(0, \sigma^2)$ and given the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we found the maximum likelihood estimates of the parameters of the model and showed that their joint distribution is described by

$$\hat{\beta}_1 \sim N\Big(\beta_1, \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)}\Big), \;\; \hat{\beta}_0 \sim N\Big(\beta_0, \Big(\frac{1}{n} + \frac{\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\sigma^2\Big)$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}\sigma^2}{n(\overline{X^2} - \bar{X}^2)}$$

and $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$ and

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Suppose now that we want to find the confidence intervals for unknown parameters of the model β_0, β_1 and σ^2. This is straightforward and very similar to the confidence intervals for parameters of normal distribution.

For example, using that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$, if we find the constants c_1 and c_2 such that

$$\chi^2_{n-2}(0, c_1) = \frac{\alpha}{2} \text{ and } \chi^2_{n-2}(c_2, +\infty) = \frac{\alpha}{2}$$

then with the remaining probability $1-\alpha$

$$c_1 \le \frac{n\hat{\sigma}^2}{\sigma^2} \le c_2.$$

Solving this for σ^2 we find the $1-\alpha$ confidence interval:

$$\frac{n\hat{\sigma}^2}{c_2} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{c_1}.$$

Next, we find the $1-\alpha$ confidence interval for β_1. We will use that

$$\xi_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{n(\overline{X^2}-\bar{X}^2)}}} \sim N(0,1) \text{ and } \frac{n\hat{\sigma}^2}{\sigma^2} = \xi_1^2 + \ldots + \xi_{n-2}^2$$

where $\xi_0, \ldots, \xi_{n-2}$ are i.i.d. standard normal. Therefore,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{n(\overline{X^2}-\bar{X}^2)}}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} = \frac{\xi_0}{\sqrt{\frac{1}{n-2}(\xi_1^2 + \ldots + \xi_{n-2}^2)}} \sim t_{n-2}$$

has Student distribution with $n-2$ degrees of freedom and, simplifying, we get

$$(\hat{\beta}_1 - \beta_1)\sqrt{\frac{(n-2)(\overline{X^2}-\bar{X}^2)}{\hat{\sigma}^2}} \sim t_{n-2}.$$

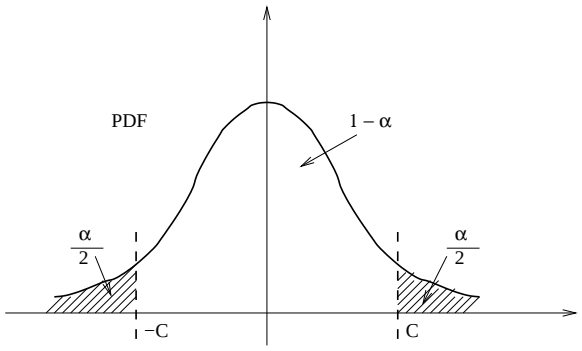

Figure 31.1: Confidence Interval.

Therefore, if we find c such that

$$t_{n-2}(-c, c) = 1 - \alpha$$

as shown in figure 31.1 then with probability $1-\alpha$:

$$-c \le (\hat\beta_1 - \beta_1)\sqrt{\frac{(n-2)(\overline{X^2}-\bar X^2)}{\hat\sigma^2}} \le c$$

and solving for β_1 gives the $1-\alpha$ confidence interval:

$$\hat\beta_1 - c\sqrt{\frac{\hat\sigma^2}{(n-2)(\overline{X^2}-\bar X^2)}} \le \beta_1 \le \hat\beta_1 + c\sqrt{\frac{\hat\sigma^2}{(n-2)(\overline{X^2}-\bar X^2)}}.$$

Similarly, to find the confidence interval for β_0 we use that

$$\frac{\hat\beta_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar X^2}{n(\overline{X^2}-\bar X^2)}\right)\sigma^2}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat\sigma^2}{\sigma^2}} \sim t_{n-2}$$

and $1-\alpha$ confidence interval for β_0 is:

$$\hat\beta_0 - c\sqrt{\frac{\hat\sigma^2}{n-2}\Big(1 + \frac{\bar X^2}{\overline{X^2}-\bar X^2}\Big)} \le \beta_0 \le \hat\beta_0 + c\sqrt{\frac{\hat\sigma^2}{n-2}\Big(1 + \frac{\bar X^2}{\overline{X^2}-\bar X^2}\Big)}.$$

Prediction Interval.

Suppose now that we have a new observation X for which Y is unknown and we want to predict Y or find the confidence interval for Y. According to simple regression model,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

and it is natural to take $\hat Y = \hat\beta_0 + \hat\beta_1 X$ as the prediction of Y. Let us find the distribution of their difference $\hat Y - Y$. Clearly, the difference will have normal distribution so we only need to compute the mean and the variance. The mean is

$$\mathbb{E}(\hat Y - Y) = \mathbb{E}\hat\beta_0 + \mathbb{E}\hat\beta_1 X - \beta_0 - \beta_1 X - \mathbb{E}\varepsilon = \beta_0 + \beta_1 X - \beta_0 - \beta_1 X - 0 = 0.$$

Since a new pair (X, Y) is independent of the prior data we have that Y is independent of $\hat Y$. Therefore, since the variance of the sum or difference of independent random variables is equal to the sum of their variances, we get

$$\mathrm{Var}(\hat Y - Y) = \mathrm{Var}(\hat Y) + \mathrm{Var}(Y) = \sigma^2 + \mathrm{Var}(\hat Y),$$

where we also used that $\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon) = \sigma^2$. Let us compute the variance of $\hat Y$:

$$\mathrm{Var}(\hat Y) \;=\; \mathbb{E}(\hat\beta_0 + \hat\beta_1 X - \beta_0 - \beta_1 X)^2 = \mathbb{E}((\hat\beta_0 - \beta_0) + (\hat\beta_1 - \beta_1)X)^2$$

$$
\begin{aligned}
&= \underbrace{\mathbb{E}(\hat{\beta}_0 - \beta_0)^2}_{\text{variance of } \hat{\beta}_1} + X^2 \underbrace{\mathbb{E}(\hat{\beta}_1 - \beta_1)^2}_{\text{variance of } \hat{\beta}_0} + 2 \underbrace{\mathbb{E}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)}_{\text{covariance}} X \\
&= \Big(\frac{1}{n} + \frac{\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\sigma^2 + X^2 \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)} - 2X \frac{\bar{X}\sigma^2}{n(\overline{X^2} - \bar{X}^2)} \\
&= \sigma^2 \Big(\frac{1}{n} + \frac{(\bar{X} - X)^2}{n(\overline{X^2} - \bar{X}^2)}\Big).
\end{aligned}
$$

Therefore, we showed that

$$
\hat{Y} - Y \sim N\Big(0, \sigma^2\Big(1 + \frac{1}{n} + \frac{(\bar{X} - X)^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\Big).
$$

As a result, we have:

$$
\frac{\hat{Y} - Y}{\sqrt{\sigma^2\Big(1 + \frac{1}{n} + \frac{(\bar{X} - X)^2}{n(\overline{X^2} - \bar{X}^2)}\Big)}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} \sim t_{n-2}
$$

and the $1 - \alpha$ prediction interval for Y is

$$
\hat{Y} - c\sqrt{\frac{\sigma^2}{n-2}\Big(n + 1 + \frac{(\bar{X} - X)^2}{\overline{X^2} - \bar{X}^2}\Big)} \le Y \le \hat{Y} + c\sqrt{\frac{\sigma^2}{n-2}\Big(n + 1 + \frac{(\bar{X} - X)^2}{\overline{X^2} - \bar{X}^2}\Big)}.
$$