均值 $E(Y|X)$ 的置信区间，最外面的两条虚线是个体值的置信区间。分析 Y 的总体均值 $E(Y|X)$ 与个体值的置信区间及其图形，我们发现：

① 样本容量n越大，预测精度越高，反之则预测精度越低；

② 样本容量一定时，置信带的宽度在 X 均值处最小，其附近进行插值预测精度越大；X 越远离其均值，置信带越宽，预测可信度下降。

10.2.5　一元线性回归综合案例

随着医疗技术的提高、医疗保障水平的提高和覆盖范围的扩大，对居民潜在医疗服务需求产生了较大的影响。各地在制定、实施规划的过程中，要按照医疗资源利用情况，对于医疗资源利用率低的医疗机构要适当缩小规模，或与其他医疗机构进行重组和调整；扩大医疗机构规模。要以提高医疗服务工作效率和医疗系统整体功能为主要手段满足增长的医疗服务需求。

为了给制定医疗机构的规划提供依据，需要分析比较医疗机构与人口数量的关系，建立卫生医疗机构数与人口数的回归模型。以人口数量为自变量，以医疗机构数为因变量，则一元线性回归模型为：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

其中 Y_i 表示卫生医疗机构数，X_i 表示人口数。变量采用四川省2000年各地的截面数据，如表10–4所示。这里 β_2 为人口每增加一万人医疗机构增加的数量，u_i 为随机误差项，即除了人口数以外，影响医疗机构数量的其他次要的、随机的因素。

表10–4　四川省2000年各地区医疗机构数与人口数

地　区	人口数（万人）	医疗机构数（个）
成都	1013.3	6304
自贡	315	911
攀枝花	103	934
泸州	463.7	1297
德阳	379.3	1085
绵阳	518.4	1616
广元	302.6	1021
遂宁	371	1375
眉山	339.9	827
宜宾	508.5	1530
广安	438.6	1589
达州	620.1	2403
雅安	149.8	866
巴中	346.7	1223
资阳	488.4	1361
阿坝	82.9	536

读入数据后，为了对解释变量和被解释变量之间的关系进行初步了解，我们采用描述统计方法对变量之间的关系进行探索。变量之间的pearson相关系数为0.8826382，说明变量之间有较强的相关关系。通过图10–5的散点图，并添加线性趋势直线可以看出，变量之间大都在一条直线附近波动，说明两变量之间存在线性关系。

```
> data=read.table("medicine.txt",head=T)
> attach(data)
> plot(x,y,xlab="人口数",ylab="医疗机构数")
> abline(lm(y~x))
> cor(x,y)
[1] 0.8826382
```

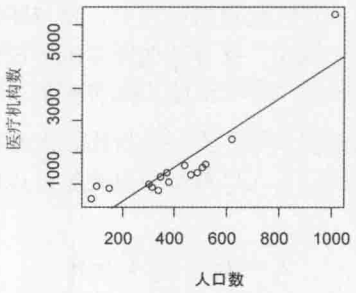

图10–5　散点图

为了进一步分析随着人口数的增加，所需的医疗机构数增加的数量，建立一元线性回归模型，并通过OLS估计方法对参数进行求解，使用summary()函数获取参数表及检验结果。在summary()函数的结果中，提供了残差的描述性统计量、参数值、参数的标准误差、t值和t检验p值、残差标准误、可决系数R^2和修正后的可决系数$\overline{R}^2$、F值和F检验的p值。其得到的模型为：

$$Y_i = -587.2682 + 5.3211X_i + u_i$$

```
> lm=lm(y~x)
> lm.summary=summary(lm)
> lm.summary

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-650.6 -434.6 -167.7  162.6 1499.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -587.2682   345.6418  -1.699    0.111
x              5.3211     0.7574   7.026    6e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 651.3 on 14 degrees of freedom
Multiple R-squared:  0.7791,    Adjusted R-squared:  0.7633
F-statistic: 49.36 on 1 and 14 DF,  p-value: 5.998e-06
```

得到回归模型后，对模型进行检验。进行拟合优度检验，提取可决系数 R^2 和修正后的可决系数 $\overline{R}^2$。提取可决系数 R^2 为0.7790501，修正后的可决系数 $\overline{R}^2$ 为0.763268。这说明样本回归直线的解释能力为77.91%，它代表医疗机构数量 Y_t 的总变差中，由解释变量人口数 X_t 解释的部分占77.91%，或者说医疗机构数量的总变动的77.91%由样本回归直线做出解释。

```
> lm.summary$r.squared
[1] 0.7790501
> lm.summary$adj.r.squared
[1] 0.763268
```

进行方程的整体显著性检验。F检验的结果在summary()函数所得结果的下方，其中F检验统计量为49.36，两个自由度为1和14，对应的p值为0，说明方程整体是显著的。

对单个变量进行显著性检验。t检验的结果在summary()函数所得的参数表中，其中截距项t检验的检验统计量，即t值为-1.699，p值为0.111，不能拒绝原假设，结果为不显著。人口数量的t值为7.026，p值为0，结果显著。

t检验的结果说明，人口数量对医疗机构数量的影响是显著的，而截距项的存在却不是显著的，这与事实相符，因为截距项的意义为在 X_i 为0时 Y_i 的平均数量，而在没有人口的情况下是不需要建立医疗机构的。因此我们应该剔除截距项后重新对回归模型进行求解，采用的模型为：

$$Y_i = \beta_1 X_i + u_i$$

在lm(y~x)命令中的x前加0即可得到不带截距项的回归，从summary()函数得到的结果可以看出，可决系数和修正后的可决系数都有所提高，F检验的p-value为0，方程整理显著，t检验的结果p值也为0，说明人口数量对医疗结构数的影响显著。

```
> lm2.summary=summary(lm2)
> lm2=lm(y~0+x)
> lm2.summary=summary(lm2)
> lm2.summary

Call:
lm(formula = y ~ 0 + x)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-683.44 -564.47 -246.33  -86.26 2062.32

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x   4.1860     0.3785   11.06 1.31e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 691 on 15 degrees of freedom
Multiple R-squared:  0.8907,    Adjusted R-squared:  0.8835
F-statistic: 122.3 on 1 and 15 DF,  p-value: 1.309e-08
```

方程的结果为：

$$Y_i = 4.1860X_i + u_i$$

说明人口数量每提高一万人，医疗机构的数量大致上升4.1860。

为了更好地规划医疗结构，人们常常需要根据人口数量预测医疗机构数。预测可以通过predict()函数来实现，并且可将观测值、回归直线、均值预测区间和个值预测区间画在同一张图上，如图10-6所示。

```
> par(mfrow=c(1,1))
> sx=sort(x) #把自变量先从小到大排序
> conf = predict(lm2,data.frame(x=sx),interval="confidence") #求均值的预测区间
> pred = predict(lm2,data.frame(x=sx),interval="prediction")#求个值的预测区间
> plot(x,y);  #画散点图
> abline(lm2)              #添加回归线
> lines(sx,conf[,2]); lines(sx,conf[,3])#用实线表示预测均值的95%置信带
> lines(sx,pred[,2],lty=3); lines(sx,pred[,3],lty=3)#用虚线表示预测均值的95%置信带
```

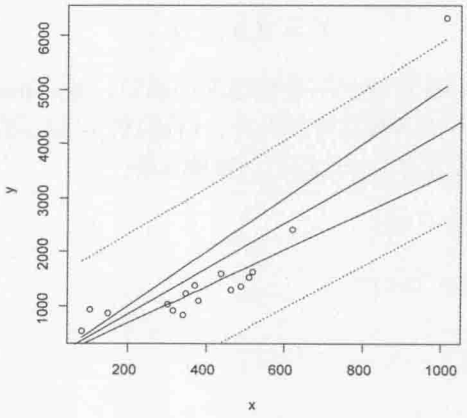

图10-6　预测区间