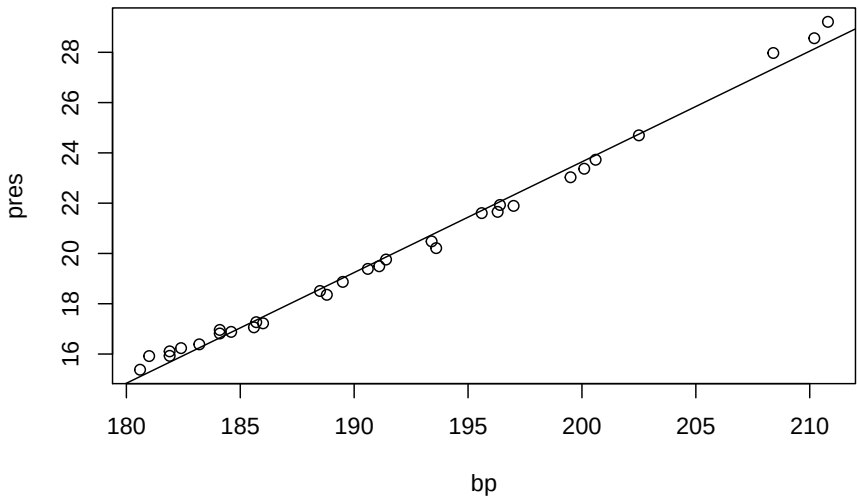
应用回归分析第一章作业

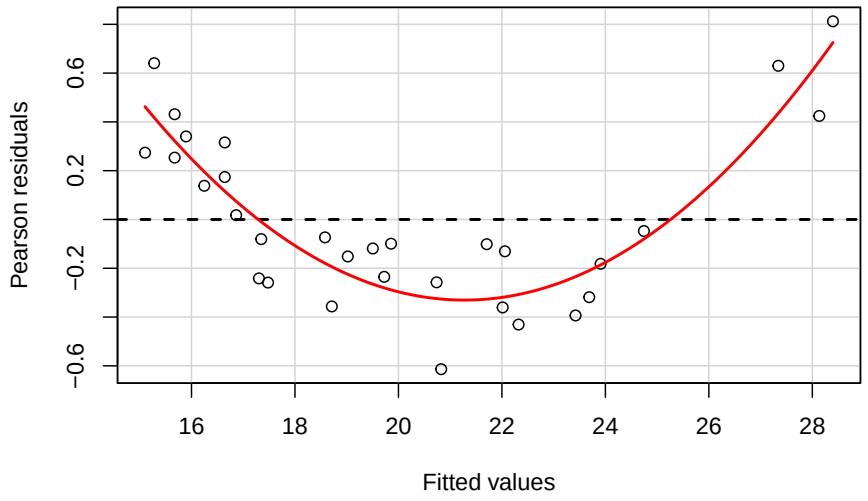
邵智轩　　1400012141　　物理学院

习题 1.2

```
library(alr4)  #Hooker 的数据在 package 'alr4' 中
attach(Hooker)
# 1.2.1
plot(bp, pres)  # 在第四版中, temp 名称改为 bp (boiling point)
L1 = lm(pres ~ bp)
abline(L1)
```



```
residualPlot(L1)
```



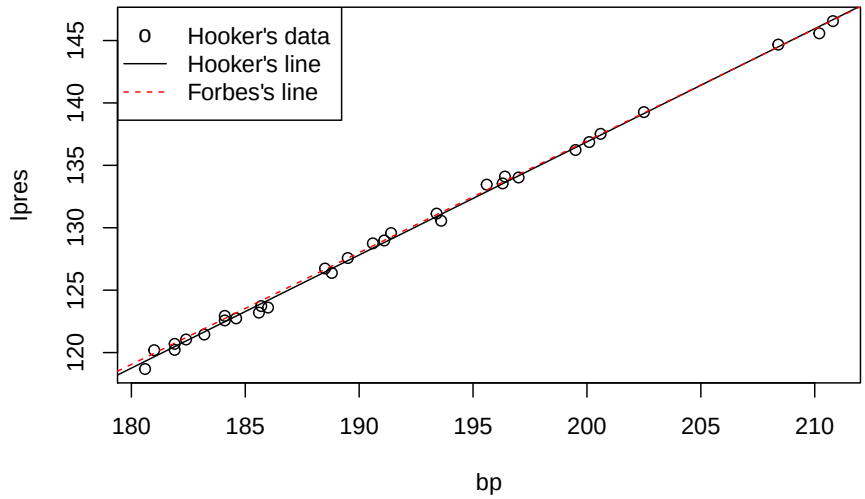
粗看起来，拟合的直线与数据匹配得比较密切，$R^2 = 0.992$。但是从 Residual Plot 中能看到明显的 non-random 的 “U 型”pattern，这暗示我们应该对 `pres` 做非线性的变换。

```
#1.2.2
plot(bp,lpres)
L2=lm(lpres~bp)
abline(L2)
attach(Forbes)
```

```
## The following objects are masked from Hooker:
##
##     bp, lpres, pres
```

```
lm.Forbes=lm(lpres~bp)
detach(Forbes)
abline(lm.Forbes,col='red',lty=2)
```

```
legend('topleft',col=c('black','black','red'),pch=c('o',NA,NA),lty=c(NA,1,2),
       legend=c("Hooker's data", "Hooker's line","Forbes's line"))
```



与 “pres~bp” 图相比，线性程度稍好。

```
# 1.2.3
kable(summary(L2)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-44.3908637	1.4611821	-30.3801	0
bp	0.9063623	0.0076111	119.0838	0

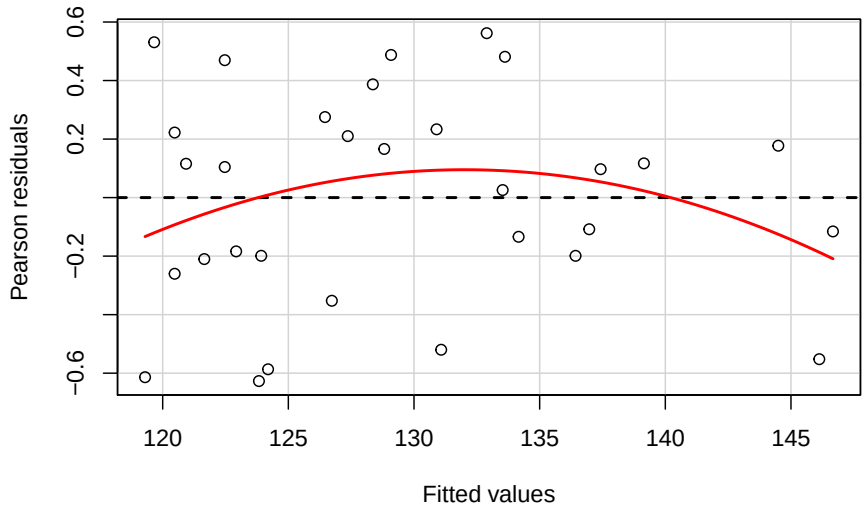
`summary` 给出了参数估计，t 检验量，相应 P 值。通过 P 值可以充分的理由拒绝 $H_0 : \beta_1 = 0$。另外从 summary 中也可看到 $R^2 = 0.998$ 也说明 $\log(pres)$ 与 bp 满足很好的线性关系。

```
kable(anova(L2))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bp	1	1882.298253	1882.2982531	14180.96	0
Residuals	29	3.849292	0.1327342	NA	NA

方差分析给出的结果仍是以充足的理由拒绝 $H_0 : \beta_1 = 0$（与 t 检验等价）。

```
residualPlot(L2)
```



与 1.2.1 中的残差图相比，pattern 并不明显，残差更接近一条水平线。

```
#1.2.4
kable(confint(L2))# 截距和斜率 95% 的置信区间
```

	2.5 %	97.5 %
(Intercept)	-47.3793167	-41.4024108
bp	0.8907958	0.9219288

```
#1.2.5
new <- data.frame(bp=c(185,212))
kable(predict.lm(L2,new,interval = 'prediction',level=0.9))
```

fit	lwr	upr
123.2862	122.6511	123.9212
147.7579	147.0768	148.4390

```
#1.2.6
kable(summary(lm.Forbes)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-42.1377793	3.3401989	-12.61535	0
bp	0.8954937	0.0164518	54.43147	0

```
kable(predict.lm(lm.Forbes,new,interval = 'prediction',level=0.9))
```

fit	lwr	upr
123.5285	122.6709	124.3862
147.7069	146.9751	148.4387

将简要对比总结如下表:

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$
Hooker	-44.4	0.9064	0.364
Forbes	-42.1	0.8955	0.379

对比 Forbes 和 Hooker 的数据的拟合结果，以及在 1.2.2 的图中对比拟合直线，可以发现两者的结果非常接近。而 Forbes 的数据由于有一个可疑的 outlier，残差估计量比 Hooker 大一些，导致对 1.2.6 中温度预测的 90%

置信区间也宽一些。

习题 1.4

1.4.1

RSS 取最小值的充要条件是:

$$\frac{\partial \mathrm{RSS}}{\partial \beta_1} = -2\sum(y_i - \beta_1 x_i)x_i = 0$$

解得 $\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$

$$\mathrm{E}(\hat{\beta}_1) = \frac{\sum x_i \mathrm{E}(Y_i)}{\sum x_i^2} = \frac{\sum x_i(\beta_1 x_i)}{\sum x_i^2} = \beta_1$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{1}{(\sum x_i^2)^2}(\sum x_i^2\sigma^2) = \frac{1}{(\sum x_i^2)^2}(\sum x_i^2)\sigma^2 = \frac{\sigma^2}{\sum x_i^2}$$

模型中有 1 个估计参数（β_1），所以 $\hat{\sigma}^2$ 的自由度为 $n-1$，其表达式为：

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}_0}{n-1}$$

$$\mathrm{RSS}_0 = \sum(y_i - \hat{\beta}_1 x_i)^2 = \sum y_i^2 - (\sum x_i y_i)^2 / \sum x_i^2$$

满足 $n-1$ 个自由度的 χ^2 分布，即 $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$，且为 σ^2 的无偏估计。

1.4.2

导出由 (1.21) 给出的较大模型的方差分析表，但用的是 (1.39) 的较小的模型。（由于 $\sum \hat{e}_i \neq 0$，平方和分解公式并不成立。）

Source of Variation	Degree of freedom	Sum of squares	Mean squares	F statistic
截距	1	$\mathrm{SSint} = \mathrm{RSS}_0 - \mathrm{RSS} = \sum[(\tilde{\beta}_1 - \hat{\beta}_1)x_i + \tilde{\beta}_0]^2$	$\mathrm{SSint}/1$	$\frac{\mathrm{SSint}}{\mathrm{RSS}/(n-2)}$

Source of Variation	Degree of freedom	Sum of squares	Mean squares	F statistic
全模型残差	$n-2$	$\mathrm{RSS}=\sum(y_i-\tilde{\beta}_1x_i-\tilde{\beta}_0)^2$	$\mathrm{RSS}/(n-2)$	
过原点模型残差	$n-1$	$\mathrm{RSS}_0=\sum(y_i-\hat{\beta}_1x_i)^2$		

假设检验 $H_0:y=\beta_1x+e$，$H_1:y=\beta_1x+\beta_0+e$,，使用 F 检验，

$$F=\frac{\mathrm{SSint}}{\mathrm{RSS}/(n-2)}=\frac{\mathrm{RSS}_0-\mathrm{RSS}}{\tilde{\sigma}^2}$$

使用 $\beta_0=0$ 时的 t 检验

$$T=\frac{\tilde{\beta}_0-0}{\mathrm{se}(\tilde{\beta}_0)}$$

$$T^2=\frac{(\bar{y}-\tilde{\beta}_1\bar{x})^2}{\tilde{\sigma}^2(1/n+\bar{x}^2/\mathrm{SXX})}=(\bar{y}-\tilde{\beta}_1\bar{x})^2\frac{n\cdot\mathrm{SXX}}{\tilde{\sigma}^2\sum x_i^2}$$

将 SXX、SXY、SYY、RSS、RSS_0 展开成 $\sum x_i^2$、$\sum y_i^2$、$\sum x_iy_i$、$\bar{x}$、$\bar{y}$ 的函数：

$$\mathrm{SYY}=\sum(y_i-\bar{y})^2=\sum y_i^2-n\bar{y}^2$$

$$\mathrm{SXX}=\sum x_i^2-n\bar{x}^2$$

$$\mathrm{SXY}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_iy_i-n\bar{x}\bar{y}$$

$$\mathrm{RSS}=\mathrm{SYY}-\frac{\mathrm{SXY}^2}{\mathrm{SXX}}=\sum y_i^2-n\bar{y}^2-\frac{(\sum x_iy_i-n\bar{x}\bar{y})^2}{\sum x_i^2-n\bar{x}^2}$$

$$\mathrm{RSS}_0=\sum y_i^2-\frac{(\sum x_iy_i)^2}{\sum x_i^2}$$

$$F=\frac{\mathrm{RSS}_0-\mathrm{RSS}}{\tilde{\sigma}^2}=\frac{n(\bar{x}\sum x_iy_i-\bar{y}\sum x_i^2)^2}{\tilde{\sigma}^2\sum x_i^2\cdot\mathrm{SXX}}$$

$$
\begin{aligned}
T^2 &= \frac{n}{\tilde{\sigma}^2 \sum x_i^2 \cdot \mathrm{SXX}} \cdot (\bar{y} - \frac{\mathrm{SXY}}{\mathrm{SXX}} \bar{x})^2 \mathrm{SXX}^2 \\
&= \frac{n}{\tilde{\sigma}^2 \sum x_i^2 \cdot \mathrm{SXX}} \cdot (\mathrm{SXX} \cdot \bar{y} - \mathrm{SXY} \cdot \bar{x})^2 \\
&= \frac{n}{\tilde{\sigma}^2 \sum x_i^2 \cdot \mathrm{SXX}} \cdot [(\sum x_i^2 - n\bar{x}^2) \cdot \bar{y} - (\sum x_i y_i - n\bar{x}\bar{y}) \cdot \bar{x}]^2 \\
&= \frac{n}{\tilde{\sigma}^2 \sum x_i^2 \cdot \mathrm{SXX}} \cdot (\bar{y} \sum x_i^2 - \bar{x} \sum x_i y_i)^2 \\
&= F
\end{aligned}
$$

从而 $F = T^2$，在数值上等价。

1.4.3

```
attach(snake)
m0<-lm(Y~X-1)
summary(m0)
```

```
##
## Call:
## lm(formula = Y ~ X - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4207 -1.4924 -0.1935  1.6515  3.0771
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## X  0.52039    0.01318   39.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 16 degrees of freedom
## Multiple R-squared:  0.9898, Adjusted R-squared:  0.9892
## F-statistic:  1559 on 1 and 16 DF,  p-value: < 2.2e-16
```

斜率 $\hat{\beta}_1 = 0.5204$，标准差估计 $\hat{\sigma} = 1.7$，量纲均为 [1]。

```r
kable(confint(m0))# 斜率的置信区间
```

	2.5 %	97.5 %
X	0.492451	0.548337

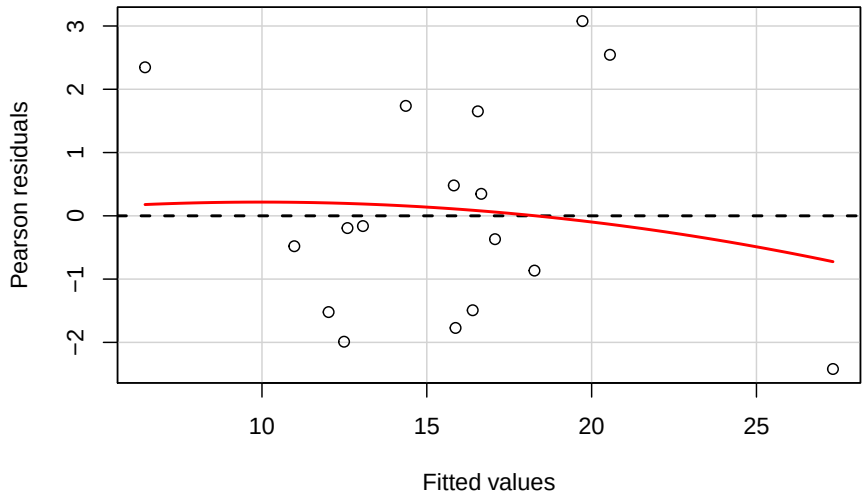
```r
kable(summary(lm(Y~X))$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7253804	1.5488161	0.4683451	0.6462711
X	0.4980812	0.0495217	10.0578308	0.0000000

在零假设 $H_0 : \beta_0 = 0$ 下 $P(|T| \geq |\frac{\hat{\beta}_0}{\text{se}(\hat{\beta}_0)}|) = 0.646$，故不拒绝原假设 $H_0 : \beta_0 = 0$。

1.4.4

```r
residualPlot(m0)# 作残差关于拟合值的图
```



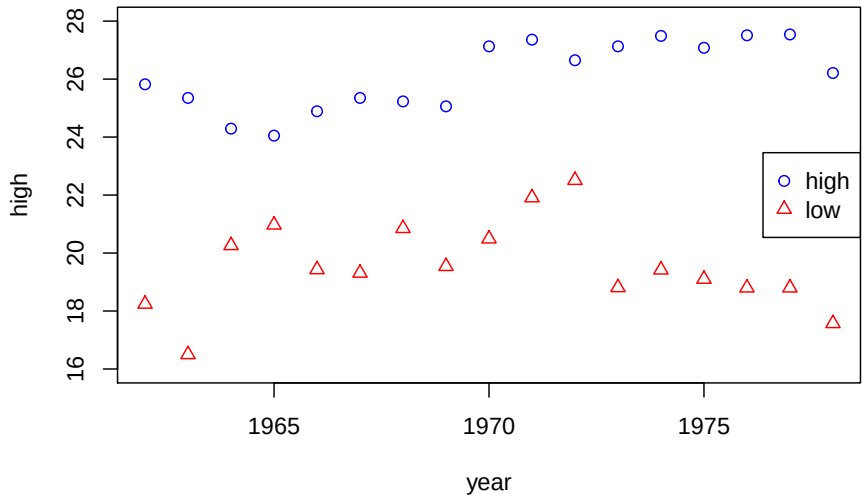
两端的 X 取值偏离直线较多；残差没有明显的 pattern，可以认为过原点的模型是比较合适的。

```
sum(m0$residuals)# 验证残差项之和不为 0
```

```
## [1] 0.9184594
```

1.9

```
Amazon=read.csv("Amazon.csv",header = TRUE)
attach(Amazon)
#1.9.1
plot(year,high,col='blue',ylim=c(16,28))# 作 high 关于 year, low 关于 year 的散点图
points(year,low,col='red',pch=2)
legend('right',col=c('blue','red'),pch=c(1,2),legend=c('high','low'))
```




```
#1.9.2
kable(summary(lm(high~year))$coef)#high 关于 year 的回归
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-330.2123529	78.0331898	-4.231691	0.0007250
year	0.1808824	0.0396106	4.566510	0.0003708

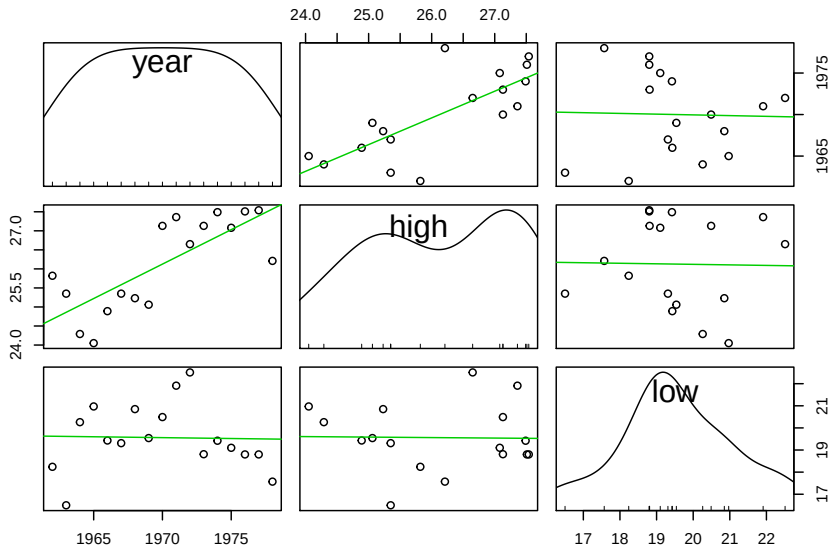
```
kable(summary(lm(low~year))$coef)#low 关于 year 的回归
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.1069608	151.723912	0.2313871	0.8201408
year	-0.0078922	0.077017	-0.1024730	0.9197387

```
kable(summary(lm(high~low))$coef)#high 关于 low 的回归
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.4008796	4.0247814	6.5595810	0.0000090
low	-0.0140596	0.2051998	-0.0685166	0.9462794

```
scatterplotMatrix(Amazon,smooth = F,spread = F)
```



以上回归系数中，high 关于 year，low 关于 year 的斜率单位都为“m/year”，即平均每年水位变化多少米（上升为正，下降为负）。high 关于 low 的回归斜率单位为“m/m”，即 low 每上升 1 米，high 平均变化多少米。

通过散点图矩阵和三组回归的 P 值，“low 与 high” 和 “low 与 year” 的相关性很弱，可以认为不相关。而 “high 与 year” 表现出一定的相关性（P 值显著地小于 0.05），斜率为 0.181m/year > 0，为亚马逊河径流量增大的假设给予了支持。

1.9.3

假设发展前高水位 High_before~$N(\mu_1, \sigma_1^2)$,发展后 High_after~$N(\mu_2, \sigma_2^2)$。可以作假设检验：NH:$\mu_1 \geq \mu_2$ vs. AH:$\mu_1 < \mu_2$。

```r
t.test(high[1:8],y=high[9:17],alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  high[1:8] and high[9:17]
## t = -8.3314, df = 13.014, p-value = 7.095e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.667223
## sample estimates:
## mean of x mean of y
##  25.00500  27.12222
```

有充足的理由接受备择假设，认为发展后（1970 年后）的平均水位确实比发展前升高了（接近 2 米）。

然而，我们无法将水位上升完全归因与森林砍伐。High 对 year 的线性回归中，$R^2 = 0.582$。也就是说年份的变化（森林的破坏）虽然能解释很大一部分原因，但可能还有其他因素的贡献。