Applied Statistics - Homework 9

Sai Nithil

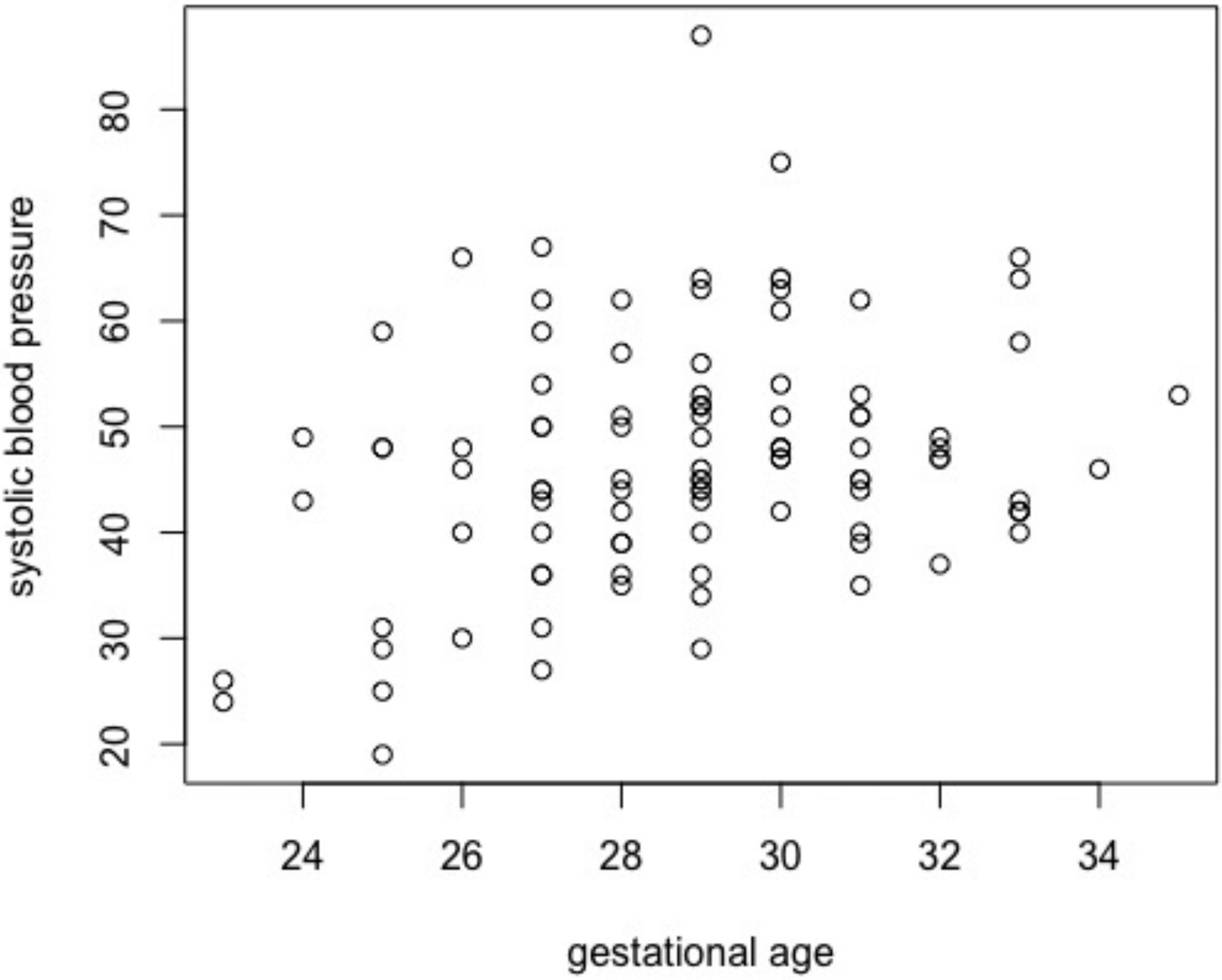
NUID: 001564864

T. Sai Nithil

18.5.9

(a)

```
> data <- read.table(file="lowbwt.txt", header = TRUE)
> plot(data$gestage, data$sbp, xlab = "gestational age", ylab = "systolic blood pressure")
>
```



The dispersion of scatter plot shows that there is huge variance in the distribution of gestational age with systolic blood pressure. This is probably not a lineer dependence.

(b)

```
> lm(formula = data$sbp ~ data$gestage)

Call:
lm(formula = data$sbp ~ data$gestage)

Coefficients:
 (Intercept)  data$gestage
      10.552         1.264
```

$\therefore\ y = 10.552 + 1.264x$

Slope $\rightarrow 1.264 > 0$

Increase in gestational age by 1 $\Rightarrow$ increase in systolic blood pressure by 1.264

y-intercept $\rightarrow 10.552$

systolic bloodpressure when gestational age $= 0$

(c)

```
> summary(lm(formula = data$sbp ~ data$gestage))

Call:
lm(formula = data$sbp ~ data$gestage)

Residuals:
    Min      1Q  Median      3Q     Max
-23.162  -7.828  -1.483   5.568  39.781

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.5521    12.6506   0.834  0.40625
data$gestage   1.2644     0.4362   2.898  0.00463 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11 on 98 degrees of freedom
Multiple R-squared:  0.07895,Adjusted R-squared:  0.06956
F-statistic: 8.401 on 1 and 98 DF,  p-value: 0.004628
```

t-value $\rightarrow$ 2.989 P-value - 0.046 < 0.005

$\Rightarrow$ reject H_0

$\therefore$ systolic blood pressure increases as gestational age increases

(d)

$y_{mean} = 10.552 + 1.264 \times 31 = 49.748$ mmHg

(mean sbp)

(e)

```
> fit <- lm(sbp~gestage, data=data)
> predict(fit, newdata=data.frame(gestage=31),interval='confidence')
       fit      lwr      upr
1 49.74784 46.90159 52.59409
```

95% CI

(f)

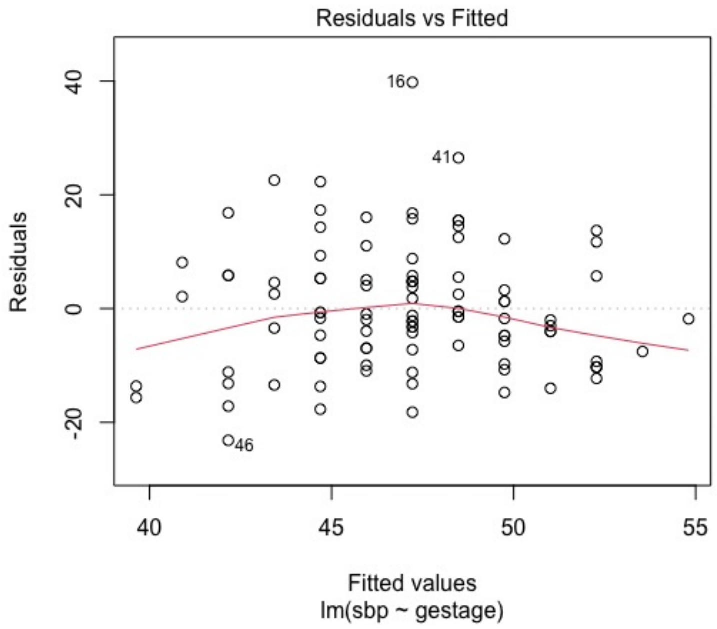
predicted systolic blood pressure for randomly selected new infant $\}$ = 49.748 mmHg

(g)

```
> predict(fit, newdata=data.frame(gestage=31),interval='prediction')
       fit      lwr     upr
1 49.74784 27.73488 71.7608
>
```

95% prediction interval. This is wider than confidence interval

(h)



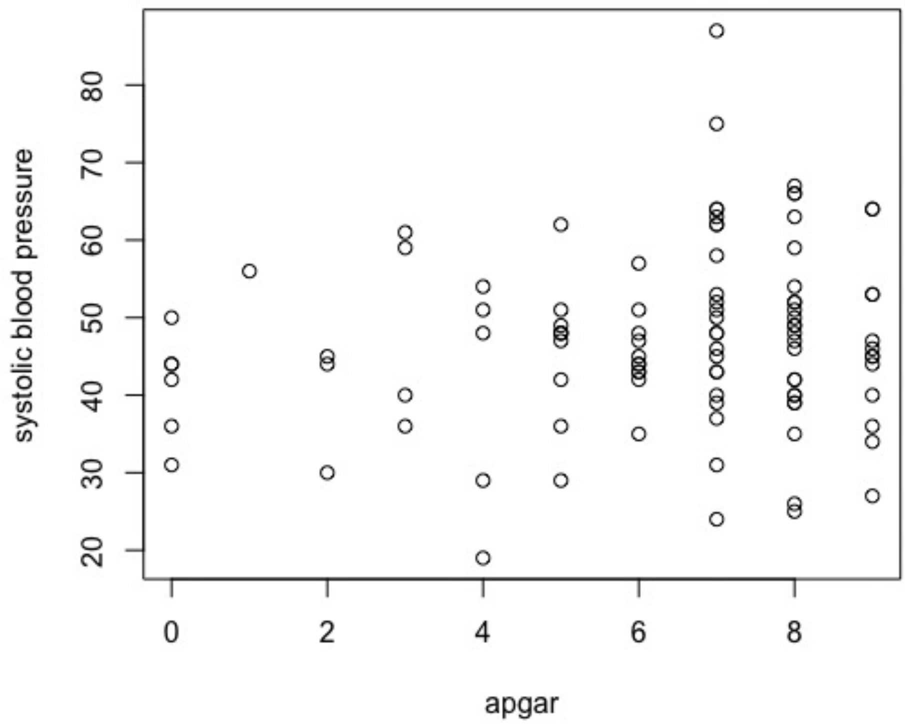
∴ From above plot, we expect that linear regression model is not a great fit for the data.

Amount of variability in sbP ~ 7.9%.

This is very low. There is no evidence that assumption of homoscedasticity has been violated or transformation of either response or one of the explanation is necessary.

19.4.8

(a)



The variance is quite high. A linear relationship is highly unlikely.

(b)

```
> summary(lm(formula = data$sbp ~ data$gestage+ data$apgar5,data=data))

Call:
lm(formula = data$sbp ~ data$gestage + data$apgar5, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.374  -8.180  -1.088   4.985  39.424 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)    9.8034    12.6629   0.774   0.4407   
data$gestage   1.1848     0.4424   2.678   0.0087 **
data$apgar5    0.4875     0.4613   1.057   0.2932   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.99 on 97 degrees of freedom
Multiple R-squared:  0.08944,Adjusted R-squared:  0.07066 
F-statistic: 4.764 on 2 and 97 DF,  p-value: 0.01063
```

$\beta_1 = 1.1848$

$\therefore$ 1 week increase in gestational age $\Rightarrow$ 1.1848 mm Hg increase in sbp

$\beta_2 = 0.4875$

$\therefore$ 1 unit increase in 5 minute apgar score $\Rightarrow$ 0.4875 mm Hg increase in sbp

(c)

$y_{mean} = 9.8034 + 1.1848(31) + 0.4875(7)$

$\simeq 49.9447$

mean sbp

(d)

```
> fit<-lm(sbp~gestage+apgar5, data=data)
> predict(fit,  newdata=data.frame(gestage = 31, apgar5 = 7), interval='confidence')
       fit      lwr      upr
1 49.94562 47.07655 52.81469
```

95% CI

(e)

```
> summary(lm(formula = sbp ~ gestage + apgar5, data = data))

Call:
lm(formula = sbp ~ gestage + apgar5, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.374  -8.180  -1.088   4.985  39.424 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   9.8034    12.6629   0.774   0.4407   
gestage       1.1848     0.4424   2.678   0.0087 **
apgar5        0.4875     0.4613   1.057   0.2932   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.99 on 97 degrees of freedom
Multiple R-squared:  0.08944,Adjusted R-squared:  0.07066 
F-statistic: 4.764 on 2 and 97 DF,  p-value: 0.01063
```

t-statistic $\rightarrow$ 1.057

p-value $\rightarrow$ $0.2932 > 0.05$

$\therefore$ accept $H_0 \Rightarrow \beta_2 = 0$

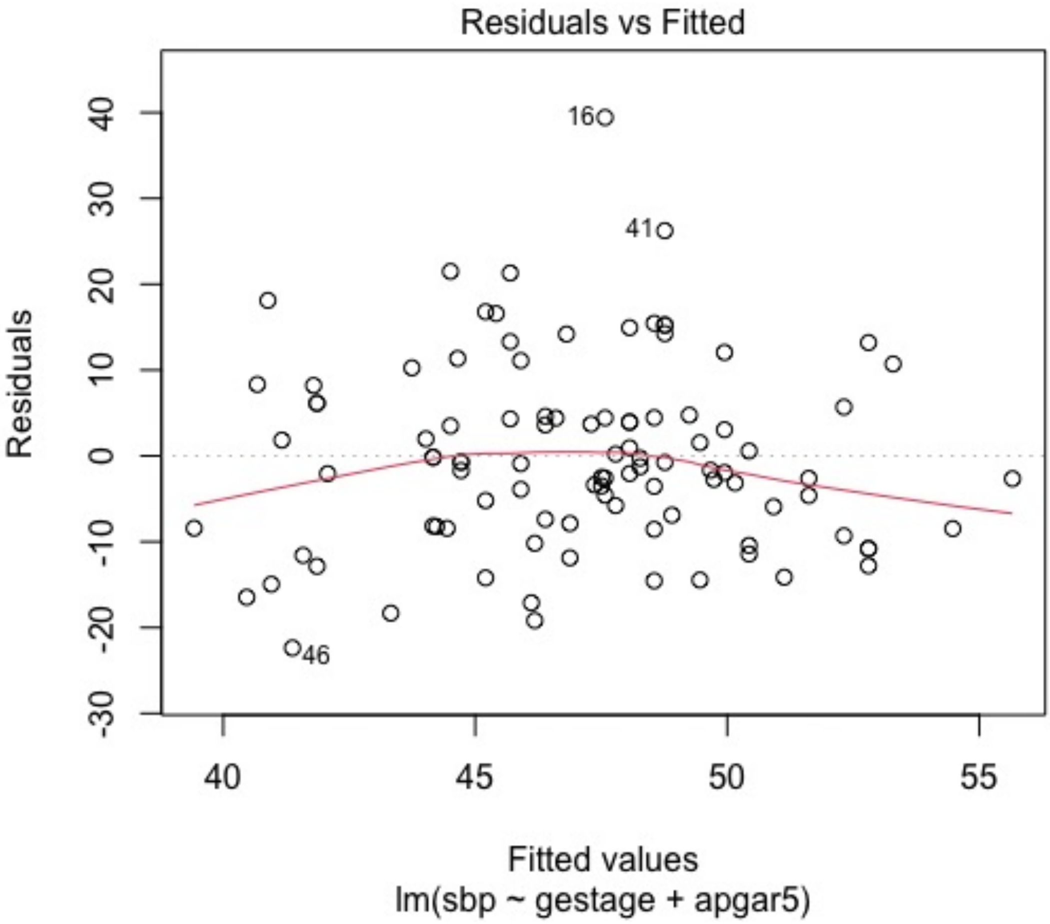
(f)

$R^2 = 0.08944 \approx 8.9\%$

$\Rightarrow$ Change is small

$\therefore$ Therefore inclusion of 5-minute apgar score in model already containing gestational age does not improve ability to predict sbp since we found $\beta_2 = 0$ from part (e)

(g)



The above plot doesn't seem like a good linear fit. There is a significant variance on both sides of $y=0$. Although, there are outliers, there is no evidence that assumption of homoscedasticity has been violated or that a transformation of either response or one of explanatory variable is necessary.