STAT 420

The (normal) simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ,$$

where ε_i's are independent Normal$(0, \sigma^2)$ (iid Normal$(0, \sigma^2)$).

β_0, β_1, and σ^2 are unknown model parameters.

$$SXX = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

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

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

Slope $\hat{\beta}_1 = \dfrac{SXY}{SXX}$ Y-intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Suppose x_i's are fixed (not random).

$\Rightarrow$ Y_i's are independent Normal$(\beta_0 + \beta_1 x_i, \sigma^2)$ random variables.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2} \sim N\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right) = N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)\right)$$

$$s_e^2 = \frac{1}{n-2} \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \qquad \frac{(n-2) s_e^2}{\sigma^2} \sim \chi^2(n-2)$$

1. The owner of *Momma Leona's Pizza* restaurant chain believes that if a restaurant is located near a college campus, then there is a linear relationship between sales and the size of the student population. Suppose data were collected from a sample of 10 *Momma Leona's Pizza* restaurants located near college campuses. For the ith restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars). The values of x_i and y_i for the 10 restaurants in the sample are summarized in the following table:

Restaurant	Student Population (1000s)	Quarterly Sales ($1000s)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

$\bar{x} = 14, \quad \bar{y} = 130$

$SXX = 568$

$SXY = 2{,}840$

$SYY = 15{,}730$

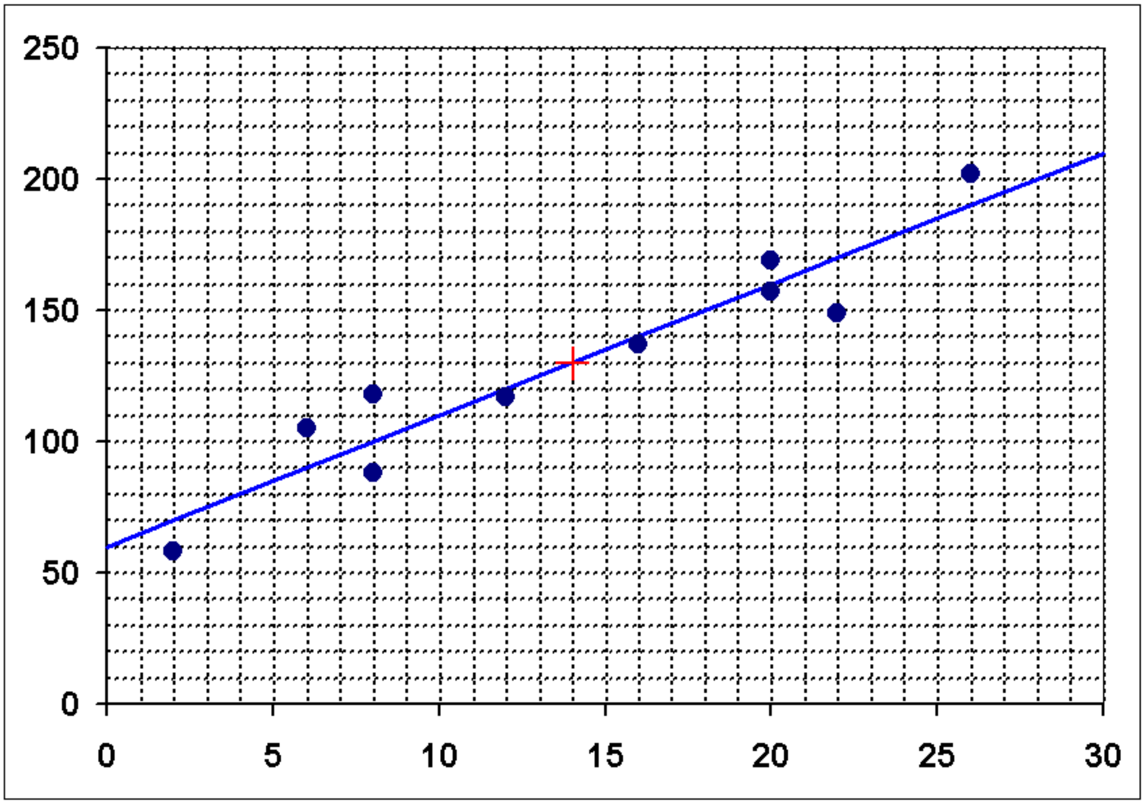
$\hat{\beta}_1 = 5, \quad \hat{\beta}_0 = 60$

$\hat{y} = 60 + 5 \cdot x$

$RSS = 1{,}530$

$R^2 = 0.9027$

$s_e^2 = 191.25$



Confidence interval for β_1: $\hat{\beta}_1 \pm t_{\alpha/2} \cdot \dfrac{s_e}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}}$ $\hat{\beta}_1 \pm t_{\alpha/2} \cdot \dfrac{s_e}{\sqrt{SXX}}$

where $t_{\alpha/2}$ is the appropriate value of t-distribution

with $n-2$ degrees of freedom.

Test statistic for $H_0: \beta_1 = \beta_{10}$:

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{s_e \Big/ \sqrt{\sum\left(x_i - \bar{x}\right)^2}} = \frac{\hat{\beta}_1 - \beta_{10}}{s_e \Big/ \sqrt{SXX}} \qquad (\, n-2 \text{ degrees of freedom} \,)$$

a) Construct a 90% confidence interval for β_1.

$\hat{\beta}_1 \pm t_{\alpha/2} \cdot \dfrac{s_e}{\sqrt{SXX}}$ $10-2=8$ degrees of freedom, $t_{0.05} = 1.860$.

$5 \pm 1.860 \cdot \dfrac{13.83}{\sqrt{568}}$ $\mathbf{5 \pm 1.08}$ $\mathbf{(\,3.92\,,\,6.08\,)}$

b) Test the assumption that students do not affect the sales. That is, test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$ (the significance of regression test). Use $\alpha = 0.01$.

Test Statistic:

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{s_e \Big/ \sqrt{SXX}} = \frac{5-0}{\sqrt{191.25} \Big/ \sqrt{568}} = 8.616.$$

Rejection Region:

Reject H_0 if $T < -t_{0.005}(\,10-2=8 \text{ df}\,)$ or $T > t_{0.005}(\,8 \text{ df}\,)$

$\pm t_{0.005}(\,8 \text{ df}\,) = \pm 3.355.$

Reject $\mathbf{H_0}$

c) That is, test $H_0 : \beta_1 = 4$ vs. $H_1 : \beta_1 > 4$. Use $\alpha = 0.05$.

Test Statistic:

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{s_e \big/ \sqrt{SXX}} = \frac{5 - 4}{\sqrt{191.25} \big/ \sqrt{568}} = 1.723.$$

Rejection Region:

Reject H_0 if $T > t_{0.05}(\,8\text{ df}\,)$

$t_{0.05}(\,8\text{ df}\,) = 1.860.$

Do NOT Reject H_0 $(\,0.05 < \text{p-value} < 0.10\,)$

_ _

Confidence interval for β_0:

$$\hat{\beta}_0 \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}$$

where $t_{\alpha/2}$ is the appropriate value of t-distribution with $n - 2$ degrees of freedom.

Test statistic for H_0: $\beta_0 = \beta_{00}$:

$$T = \frac{\hat{\beta}_0 - \beta_{00}}{s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}} \qquad (\, n - 2 \text{ degrees of freedom }\,)$$

d) Construct a 90% confidence interval for β_0.

$$60 \pm 1.860 \cdot \sqrt{191.25} \cdot \sqrt{\frac{1}{10} + \frac{14^2}{568}} \qquad \mathbf{60 \pm 17.16}$$

e) Test $H_0: \beta_0 = 75$ vs. $H_1: \beta_0 < 75$. Use a 5% level of significance.

Test Statistic: $$T = \frac{60 \; - \; 75}{\sqrt{191.25}\sqrt{\dfrac{1}{10} + \dfrac{14^2}{568}}} = -1.626.$$

Rejection Region: Reject H_0 if $T < -t_{0.05}(8 \text{ df}) = -1.860.$

Do NOT Reject H_0

– –

Confidence interval for σ^2:

$$\left(\frac{(n-2)s_e^2}{\chi^2_{\alpha/2}}, \frac{(n-2)s_e^2}{\chi^2_{1-\alpha/2}} \right) \qquad \left(\frac{n\hat{\sigma}^2}{\chi^2_{\alpha/2}}, \frac{n\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \right)$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are the appropriate values of χ^2 distribution with $n-2$ degrees of freedom.

f) Construct a 95% confidence interval for σ^2.

$\chi^2_{0.025}(8 \text{ df}) = 17.54,$ $\chi^2_{0.975}(8 \text{ df}) = 2.180.$

$$\left(\frac{8 \cdot 191.25}{17.54}, \frac{8 \cdot 191.25}{2.180} \right) \qquad \mathbf{(87.229, 701.835)}$$

– –

Mean response (y) for a fixed value of x: $\mu(x) = \mu_{y|x} = \beta_0 + \beta_1 x.$

To estimate $\mu(x)$, use $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$

$E(\hat{Y}) = \mu(x) = \beta_0 + \beta_1 x.$ $\qquad \operatorname{Var}(\hat{Y}) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{SXX} \right).$

Confidence interval for $\mu(x)$: $$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

where $t_{\alpha/2}$ is the appropriate value of t-distribution with $n-2$ degrees of freedom.

Prediction interval for a future value of y corresponding to a given value of x:

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$ (limits of prediction)

where $t_{\alpha/2}$ is the appropriate value of t-distribution with $n-2$ degrees of freedom.

g) Construct a 95% confidence interval for $\mu(x = 10)$.

$$110 \pm 2.306 \cdot \sqrt{191.25} \cdot \sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}}$$

110 ± 11.42

h) Construct a 95% confidence interval for $\mu(x = 38)$.

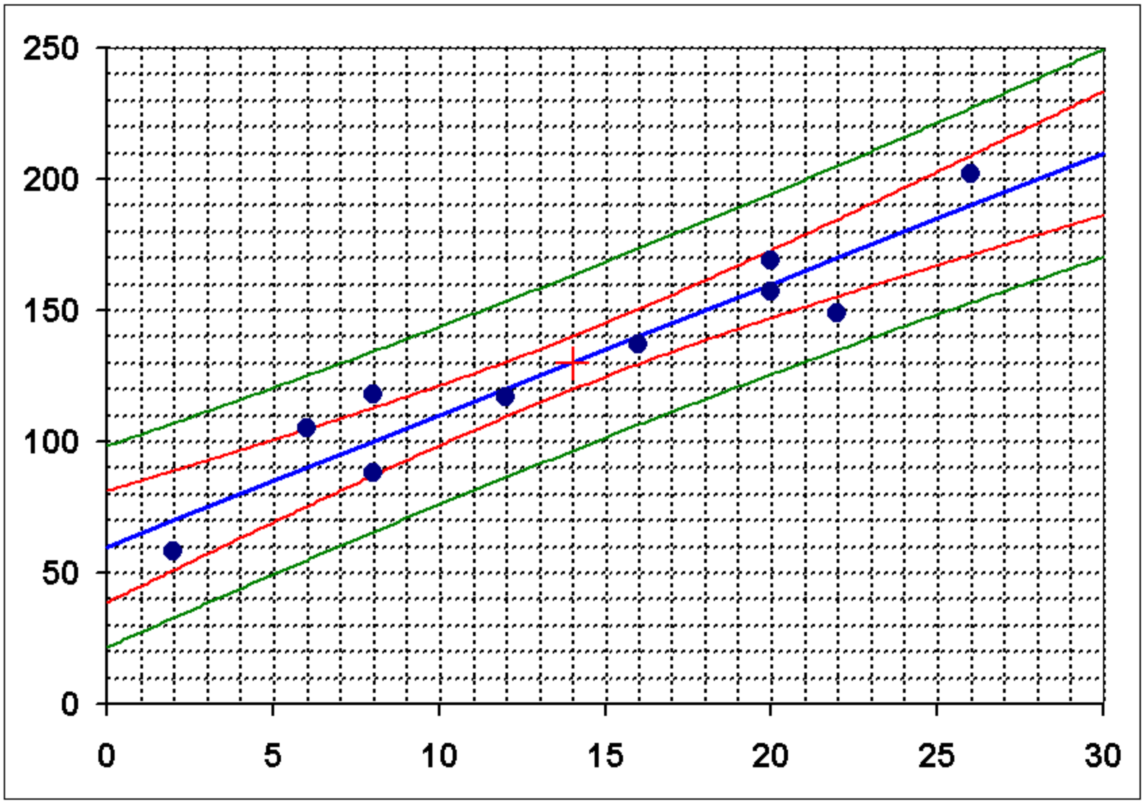
$$250 \pm 2.306 \cdot \sqrt{191.25} \cdot \sqrt{\frac{1}{10} + \frac{(38-14)^2}{568}}$$

250 ± 33.66

i) Construct a 95% prediction interval for a future value of y corresponding to $x = 38$.
(Construct 95% limits of prediction if $x = 38$.)

$$250 \pm 2.306 \cdot \sqrt{191.25} \cdot \sqrt{1 + \frac{1}{10} + \frac{(38-14)^2}{568}}$$

250 ± 46.37



j) University of Illinois at Urbana-Champaign has 38 thousand students. The owner of *Momma Leona's Pizza* restaurant chain would agree to open a restaurant near the UIUC campus, but only if there is enough evidence that the average quarterly sales would be over $225,000. Test $H_0 : \mu(x = 38) = 225$ vs. $H_1 : \mu(x = 38) > 225$. Use $\alpha = 0.05$.

Test Statistic:
$$T = \frac{250 \;-\; 225}{\sqrt{191.25}\,\sqrt{\dfrac{1}{10} + \dfrac{(38-14)^2}{568}}} = 1.713.$$

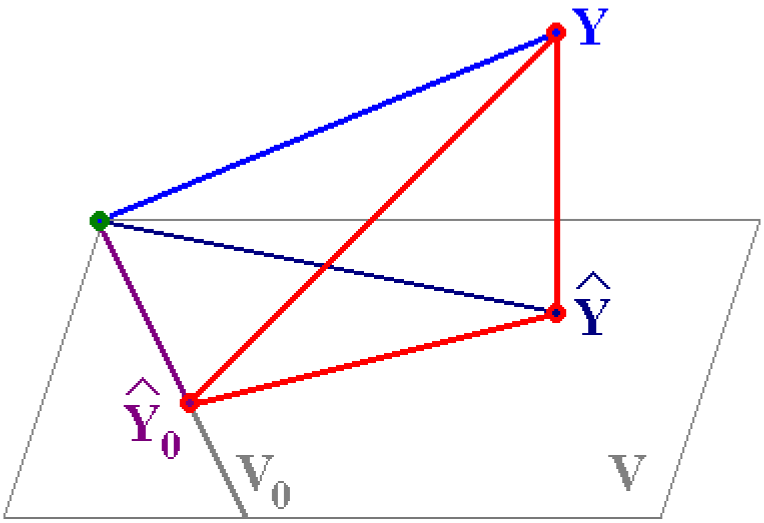
Rejection Region: Reject H_0 if $T > t_{0.05}(8 \text{ df}) = 1.860.$

Do NOT Reject H_0

Note that $\beta_0 = \mu(x = 0).$

$\hat{\beta}_0 = \hat{y}$ if $x = 0.$

b) Test the assumption that students do not affect the sales. That is, test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ (the significance of regression test).
Use $\alpha = 0.01$.



Here $V_0 = \{ a\,\mathbf{1},\ \ a \in \mathbf{R} \}$, $\quad \dim(V_0) = 1$, $\quad \hat{\mathbf{Y}}_\mathbf{0} = \left[\overline{Y}, \overline{Y}, \ldots, \overline{Y}\right]^T$,

$V = \{ a_0\,\mathbf{1} + a_1\,\boldsymbol{x},\quad a_0, a_1 \in \mathbf{R} \}$, $\quad \dim(V) = 2$.

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$$

Since $\hat{y}_i = \hat{\alpha} + \hat{\beta}\,x_i = \left(\bar{y} - \hat{\beta}\,\bar{x}\right) + \hat{\beta}\,x_i = \bar{y} + \hat{\beta}\left(x_i - \bar{x}\right)$,

SSRegression $= \sum\left(\hat{y}_i - \bar{y}\right)^2 = \sum\hat{\beta}^2\left(x_i - \bar{x}\right)^2 = \hat{\beta}^2\sum\left(x_i - \bar{x}\right)^2 = \hat{\beta}^2$ SXX.

ANOVA table:

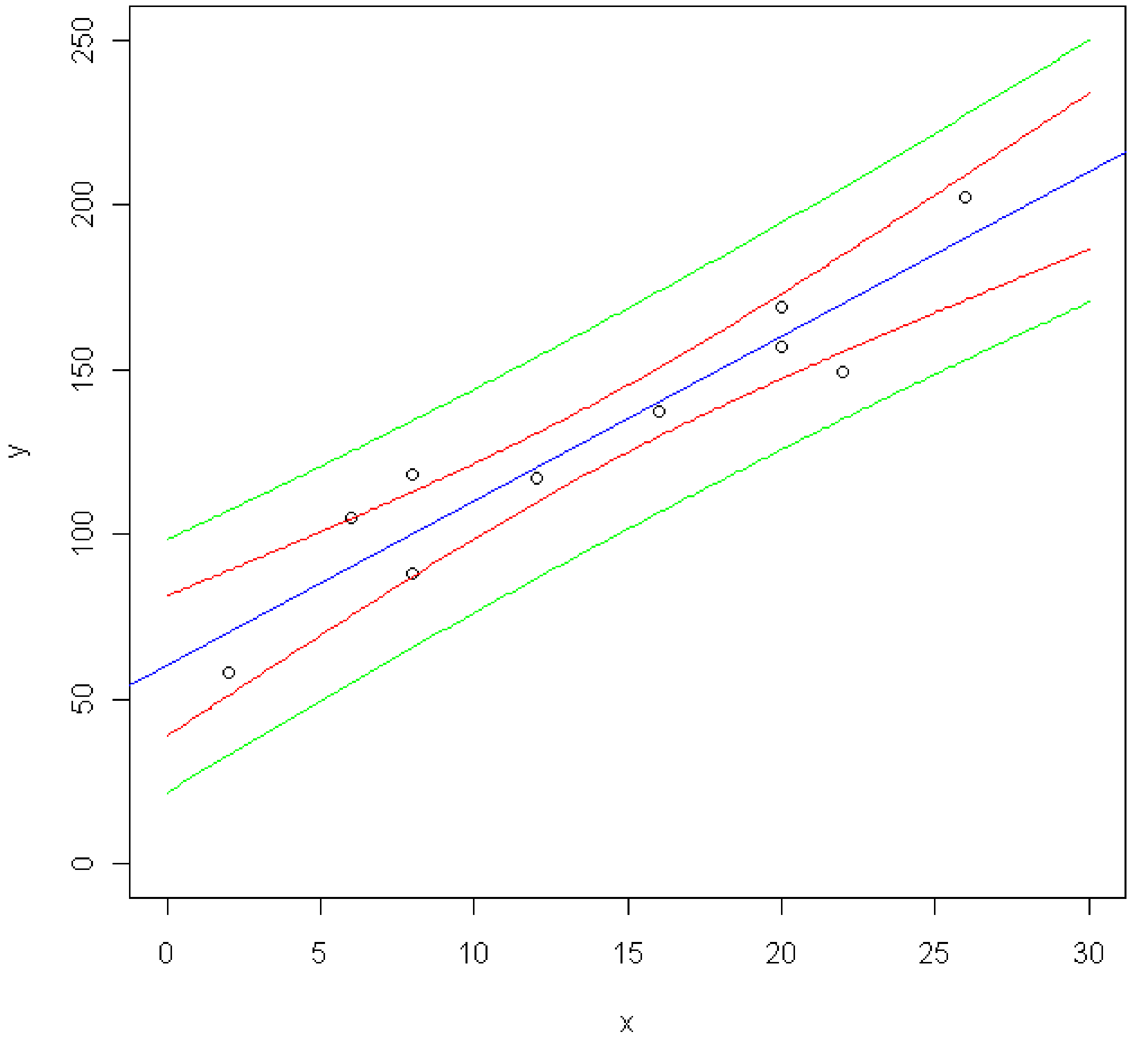
Source	**SS**		**DF**	**MS**	**F**
Regression	$\sum(\hat{y}_i - \bar{y})^2$	= 14,200	1	14,200	74.248366
Error	$\sum(y_i - \hat{y}_i)^2$	= 1,530	$n - 2 = 8$	191.25	
Total	$\sum(y_i - \bar{y})^2$	= 15,730	$n - 1 = 9$		

Rejection Region: Reject H_0 if $F > F_{0.01}(1, 8) = 11.26$.

Reject H_0

```
> x <- c(2,6,8,8,12,16,20,20,22,26)
> y <- c(58,105,88,118,117,137,157,169,149,202)

> fit <- lm(y ~ x)

> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max 
-21.00  -9.75  -3.00  11.25  18.00 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  60.0000     9.2260   6.503 0.000187 ***
x             5.0000     0.5803   8.617 2.55e-05 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1 

Residual standard error: 13.83 on 8 degrees of freedom
Multiple R-Squared: 0.9027,     Adjusted R-squared: 0.8906 
F-statistic: 74.25 on 1 and 8 DF,  p-value: 2.549e-05 

> anova(fit)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x          1 14200.0 14200.0  74.248 2.549e-05 ***
Residuals  8  1530.0   191.3                      
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1 

> confint(fit, level=0.90)
                  5 %      95 %
(Intercept) 42.843745 77.156255
x            3.920969  6.079031

> new <- data.frame(x=10)
> predict.lm(fit,new,interval=c("confidence"),level=0.95)
     fit    lwr     upr
[1,] 110 98.583 121.417

> new <- data.frame(x=38)
> predict.lm(fit,new,interval=c("confidence"),level=0.95)
     fit      lwr      upr
[1,] 250 216.3396 283.6604
> predict.lm(fit,new,interval=c("prediction"),level=0.95)
     fit      lwr      upr
[1,] 250 203.6316 296.3684
```

```
> plot(x,y,xlim=c(0,30),ylim=c(0,250))
> abline(fit$coefficients,col="blue")
>
> xx = seq(0,30,by=0.1)
>
> int1 = predict.lm(fit,data.frame(x=xx),interval=c("confidence"),level=0.95)
> int2 = predict.lm(fit,data.frame(x=xx),interval=c("prediction"),level=0.95)
>
> lines(xx,int1[,2],col="red")
> lines(xx,int1[,3],col="red")
> lines(xx,int2[,2],col="green")
> lines(xx,int2[,3],col="green")
```


y
0
50
100
150
200
250
0
5
10
15
20
25
30
x