

STAT 516 Lec 02

Review of simple linear regression

Karl Gregory

2024-01-18

Hemoglobin versus RBC count example

These data are a subset of the dataset Marcinkevičs et al. (2023)

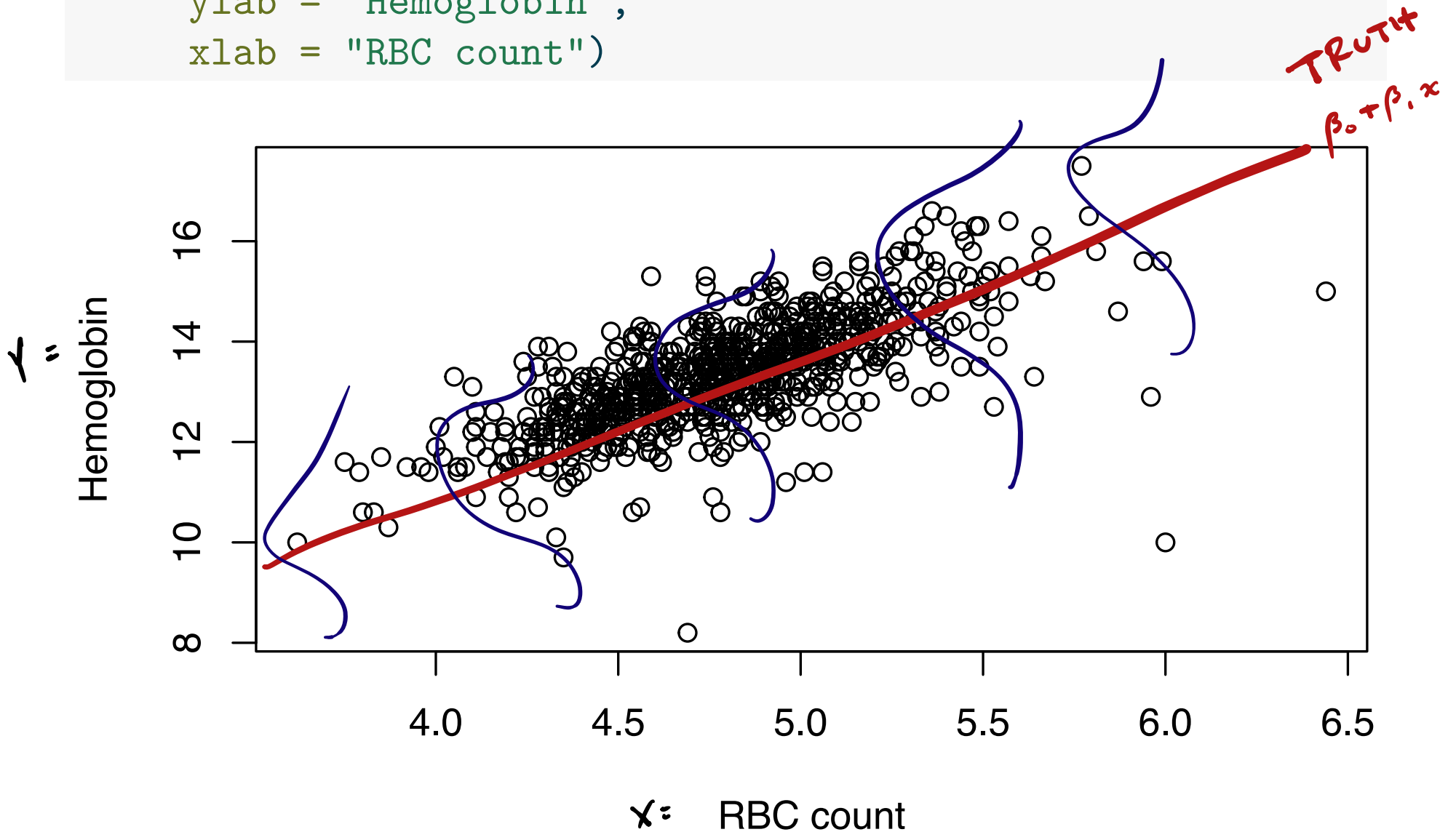
Some outliers and missing values were removed.

```
link <- url("https://people.stat.sc.edu/gregorkb/data/hrbc.csv")
data <- read.csv(link)
head(data)
```

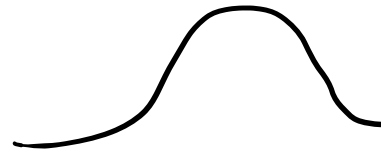
	hem	rbc	sex	age	diag
1	14.8	5.27	female	12.68	appendicitis
2	15.7	5.26	male	14.10	no appendicitis
3	11.4	3.98	female	14.14	no appendicitis
4	13.6	4.64	female	16.37	no appendicitis
5	12.6	4.44	female	11.08	appendicitis
6	12.5	4.96	male	11.05	no appendicitis

Hemoglobin level vs red blood cell count for $n = 762$ children.

```
plot(data$hem ~ data$rbc,  
      ylab = "Hemoglobin",  
      xlab = "RBC count")
```



Simple linear regression



For $(x_1, Y_1), \dots, (x_n, Y_n)$, the simple linear regression model is

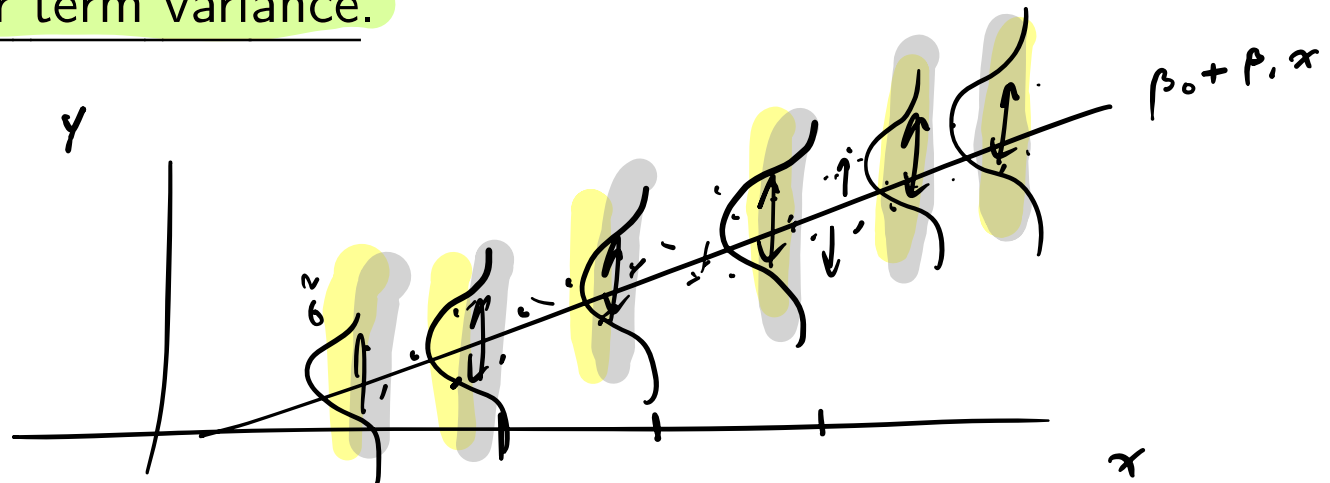
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

↑ hem ↑ rbc count

noise / "error"

where

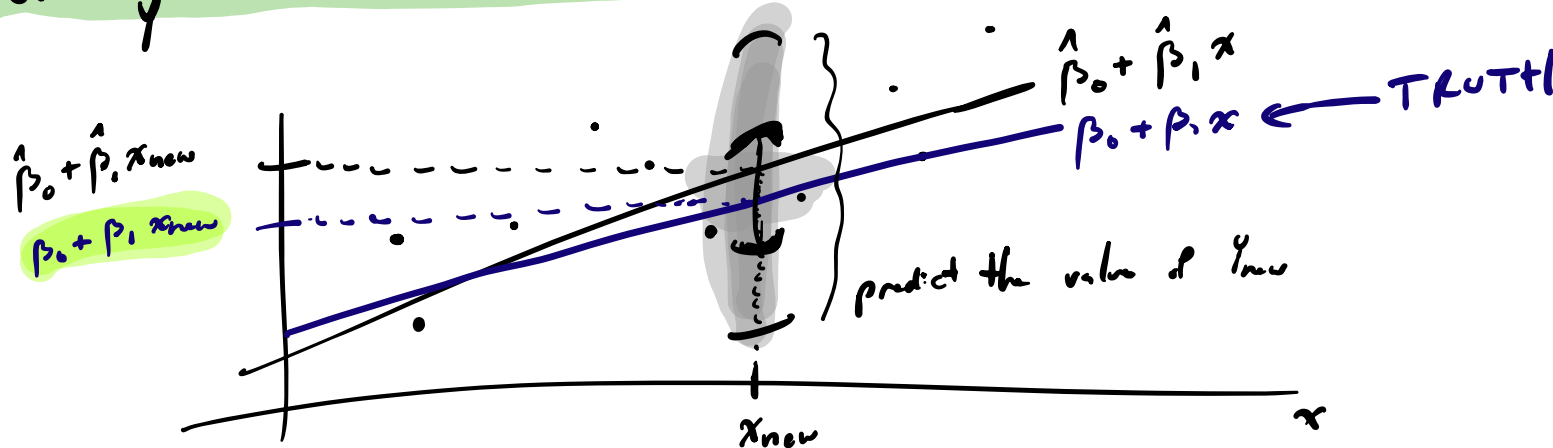
- ▶ x_1, \dots, x_n are the covariate or predictor values.
- ▶ Y_1, \dots, Y_n are the response values.
- ▶ β_0 and β_1 are the intercept and slope parameters, respectively.
- ▶ $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ error terms.
- ▶ σ^2 is the error term variance.



Goals in simple linear regression

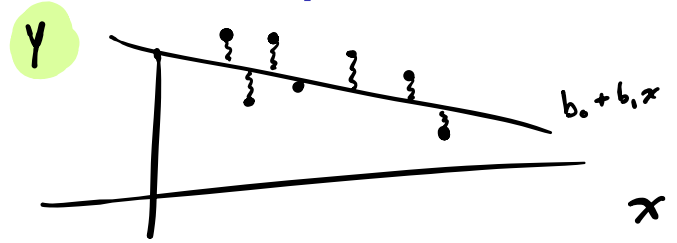
We wish to:

1. Estimate the intercept and slope parameters β_0 and β_1 .
2. Estimate the error term variance σ^2 .
3. Perform inference on β_1 . ← Build C.I. or test hypotheses.
4. Build a confidence interval for $\beta_0 + \beta_1 x_{\text{new}}$ at any x_{new} .
5. Build a prediction interval for Y at any x_{new} .
6. Decompose the variation in Y into (sums of) sums of squares.
7. Check whether the model assumptions are satisfied.
8. Understand influence of outliers.



Least-squares estimation of slope and intercept

The squared error criterion given by the sum

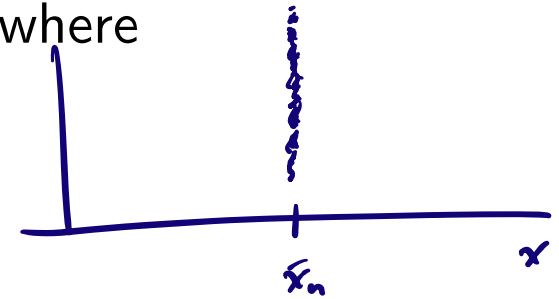


$$Q(b_0, b_1) = \sum_{i=1}^n (Y_i - (b_0 + b_1 x_i))^2$$

of squared vertical distances of Y_i from the line $y = b_0 + b_1 x$.

We find $Q(b_0, b_1)$ is minimized at $(b_0, b_1) = (\hat{\beta}_0, \hat{\beta}_1)$, where

- ▶ $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$
- ▶ $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$



provided $\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$.

If $\sum_{i=1}^n (x_i - \bar{x}_n)^2 = 0$
 it means x_1, \dots, x_n are all equal.

The least-squares line or fitted line is the line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

Pearson's correlation coefficient

Given $(x_1, Y_n), \dots, (x_n, Y_n)$, the quantity

$$r_{xY} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$$

is called Pearson's correlation coefficient.

- ▶ Describes strength and direction of *linear* relationships.
- ▶ Must satisfy $r_{xY} \in [-1, 1]$.
- ▶ Values close to zero indicate a weak linear relationship.
- ▶ Is related to $\hat{\beta}_1$ by

$$\hat{\beta}_1 = r_{xY} \sqrt{\frac{S_{YY}}{S_{xx}}}$$

with $S_{xx} = \sum_{i=1}^n (x_i - \bar{x}_n)^2$ and $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$.

Hemoglobin versus RBC count example (cont)

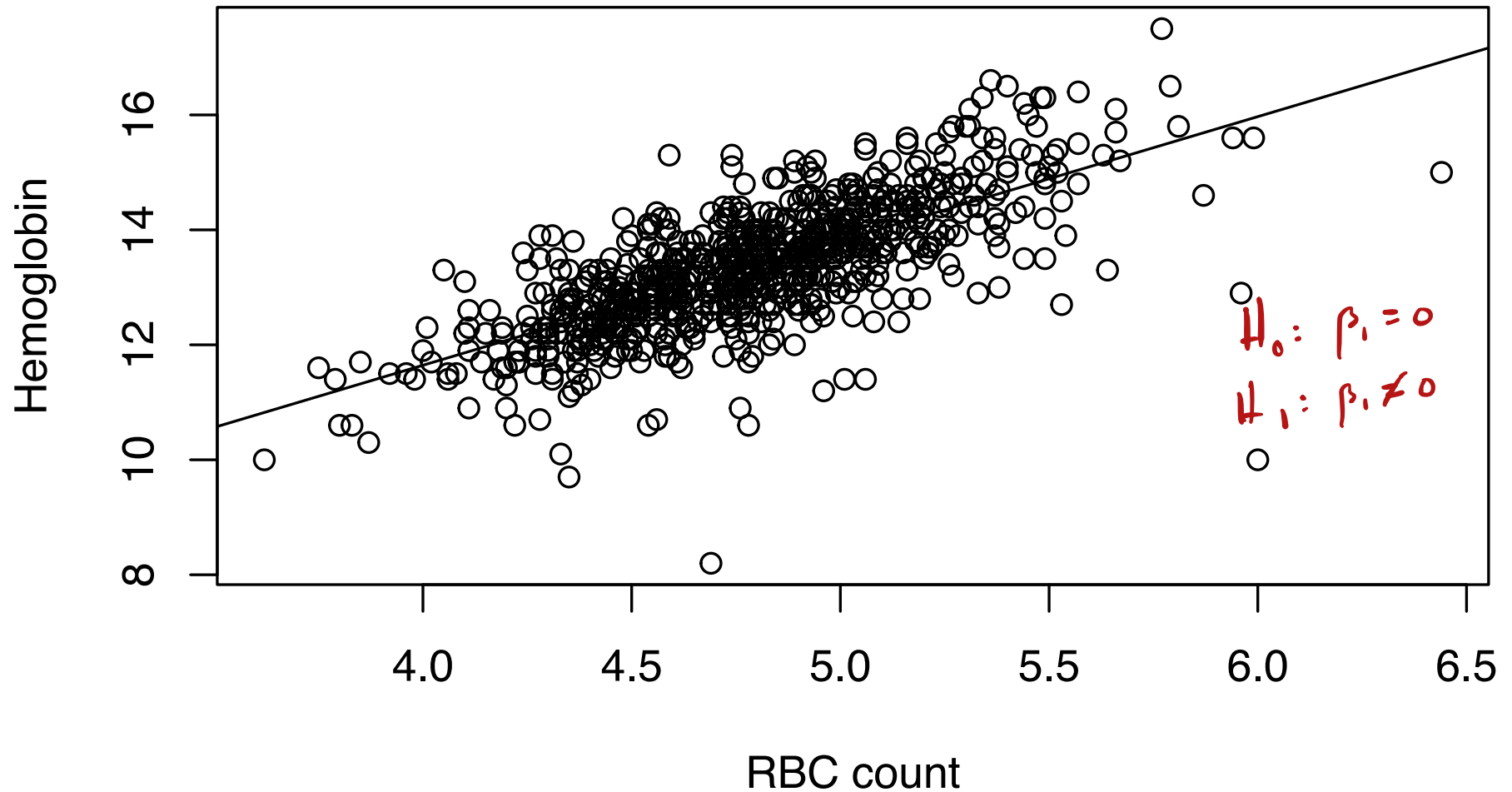
Find the least-squares line on the hemoglobin data.

```
Y <- data$hem
x <- data$rbc
n <- length(Y)
xbar <- mean(x)
Ybar <- mean(Y)
rxY <- cor(x, Y) # Pearson's correlation coefficient
Sxx <- sum( (x - xbar)^2 )
SYY <- sum( (Y - Ybar)^2 )
b1hat <- rxY * sqrt(SYY / Sxx)
b0hat <- Ybar - b1hat * xbar
```

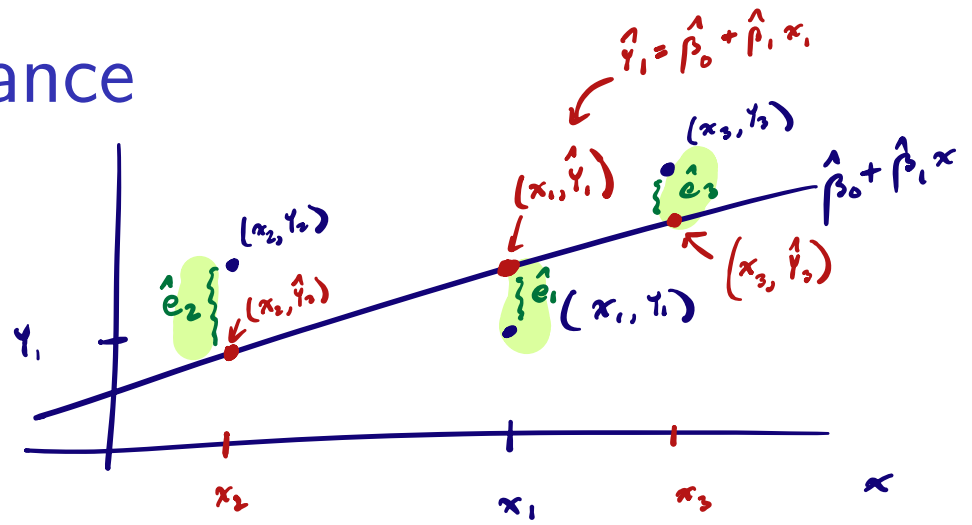
$\hat{\beta}_1$
 $\hat{\beta}_0$


```
plot(data$hem ~ data$rbc,  
      ylab = "Hemoglobin", xlab = "RBC count")  
abline(b0hat, b1hat)
```

\uparrow intercept \uparrow slope



Estimating the error term variance



After obtaining $\hat{\beta}_0$ and $\hat{\beta}_1$, define the

- ▶ fitted values as $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- ▶ residuals as $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

for $i = 1, \dots, n$.

Then an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$$

like a variance

mean of residuals is equal to zero.

To compute this, we must first estimate β_0 and β_1 with $\hat{\beta}_0$ and $\hat{\beta}_1$. So we lose 2 degrees of freedom.

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

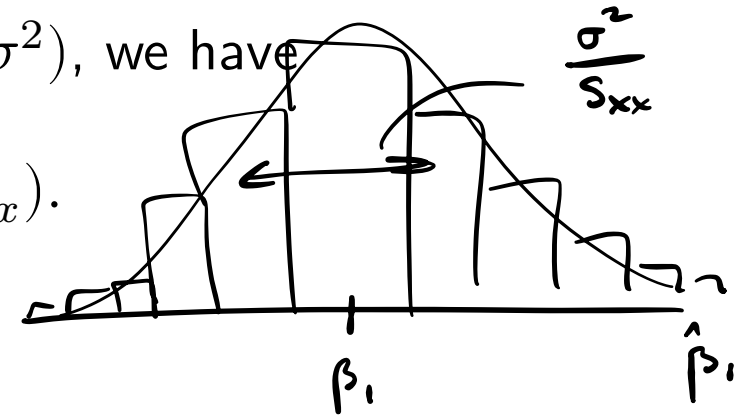
must estimate μ with \bar{x}_n so we "lose" one degree of freedom

Confidence interval for the slope parameter

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Provided $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$, we have

$$\hat{\beta}_1 \sim \text{Normal}(\beta_1, \sigma^2 / S_{xx}).$$

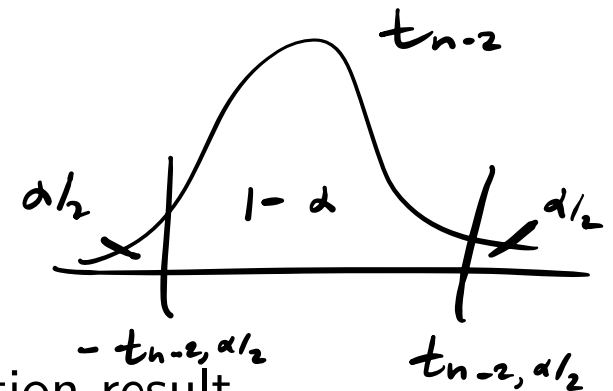


“Studentizing” the above gives

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} / \sqrt{S_{xx}}} \sim t_{n-2}.$$

So a $(1 - \alpha)100\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{n-2, \alpha/2} \hat{\sigma} / \sqrt{S_{xx}}.$$



Exercise: Justify the CI using the sampling distribution result.

$$P\left(-t_{n-2, \alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} / \sqrt{S_{xx}}} < t_{n-2, \alpha/2}\right) = 1 - \alpha$$

Hemoglobin versus RBC count example (cont)

Obtain an estimate of the error term variance.

← fitted values

```
Yhat <- b0hat + b1hat * x
```

```
ehat <- Y - Yhat ← residuals
```

```
sgsqhat <- sum(ehat^2) / (n-2) =  $\hat{\sigma}^2$ 
```

We obtain $\hat{\sigma}^2 = 0.626$.

← $\alpha = 0.05$

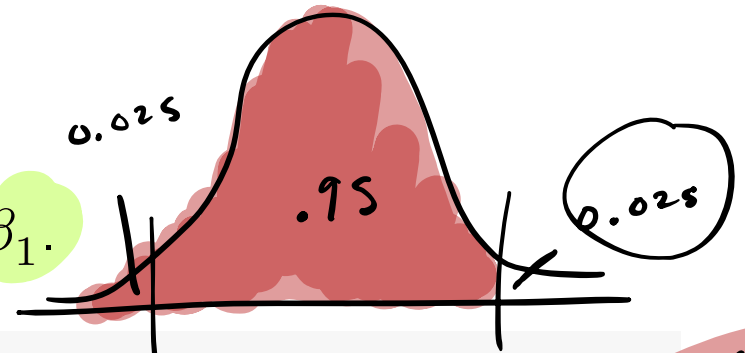
Now construct a 95% confidence interval for β_1 .

```
alpha <- 0.05
```

```
ta2 <- qt(1 - alpha/2, df = n - 2)
```

```
lo <- b1hat - ta2 * sqrt(sgsqhat / Sxx)
```

```
up <- b1hat + ta2 * sqrt(sgsqhat / Sxx)
```



← 0.975

← 762-2

← 1.963

← $t_{762-2, 0.025}$

← $qt(0.975, 760)$

← 1.963

The 95% CI is (2.011, 2.314).

$$\hat{\beta}_1 \pm t_{n-2, \alpha/2}$$

$$\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Tests of hypotheses about the slope

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
If $\beta_1 = 0$, then there is no linear relationship between x and Y .

We most often test hypotheses about β_1 of the form

$$\begin{array}{l} H_0: \beta_1 \geq 0 \quad \text{or} \quad H_0: \beta_1 = 0 \quad \text{or} \quad H_0: \beta_1 \leq 0 \\ H_1: \beta_1 < 0 \quad \quad \quad H_1: \beta_1 \neq 0 \quad \quad \quad H_1: \beta_1 > 0. \end{array}$$

Reject or fail to reject H_0 based on the value of the test statistic

$$T_{\text{stat}} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma} / \sqrt{S_{xx}}}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Rejection rules for the above at significance level α are

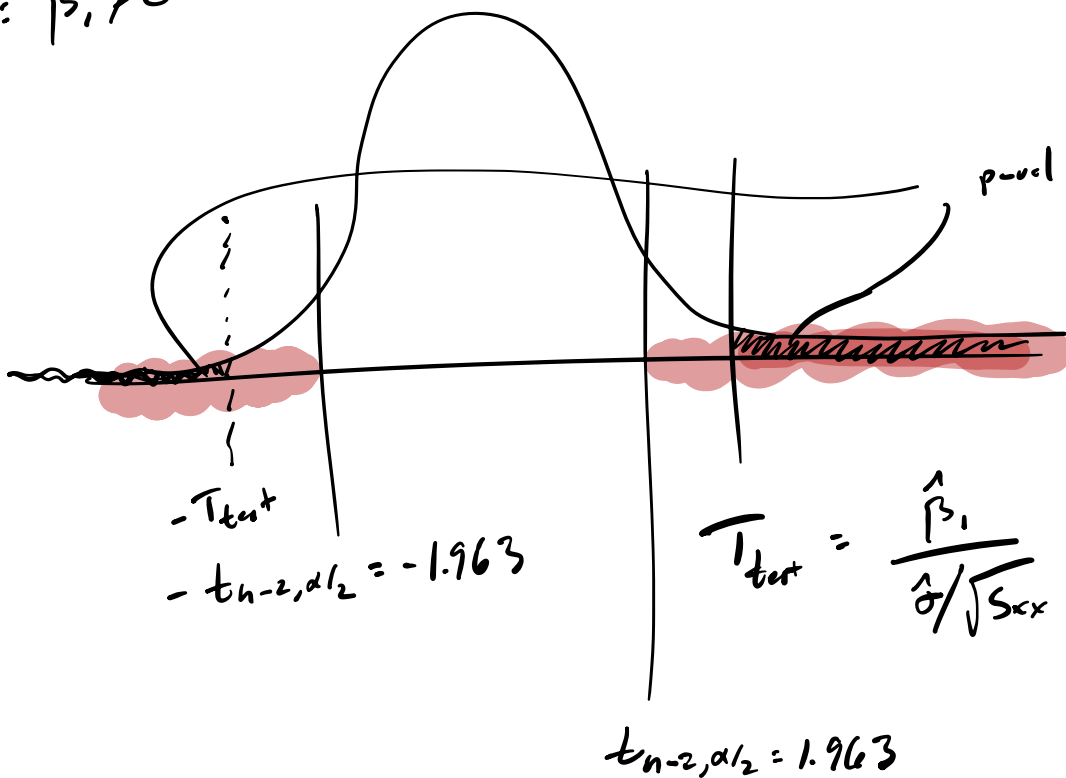
$$T_{\text{stat}} < -t_{n-2, \alpha} \quad \text{or} \quad |T_{\text{stat}}| > t_{n-2, \alpha/2} \quad \text{or} \quad T_{\text{stat}} > t_{n-2, \alpha}.$$

The corresponding p-values are, with $T \sim t_{n-2}$, the probabilities

$$P(T < T_{\text{stat}}) \quad \text{or} \quad 2 \times P(T > |T_{\text{stat}}|) \quad \text{or} \quad P(T > T_{\text{stat}}).$$

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$



$$p\text{-val} = 2 P(T > |T_{stat}|),$$

$$T \sim t_{n-2}$$

Hemoglobin versus RBC count example (cont)

Test the hypotheses $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ at $\alpha = 0.05$.

```

alpha <- 0.05
Tstat <- b1hat / sqrt(sgsqhat/Sxx) =  $\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}}$ 
rej <- abs(Tstat) > qt(1-alpha/2, df = n - 2) ←  $t_{n-2, \alpha/2} = 1.963$ 
pval <- 2*(1 - pt(abs(Tstat), df = n - 2))

```

$2(1 - P(T \leq |T_{stat}|))$, $T \sim t_{n-2}$

The p-value is 0; we reject H_0 .

Test the hypotheses $H_0: \beta_1 \leq 2$ vs $H_1: \beta_1 > 2$.

```

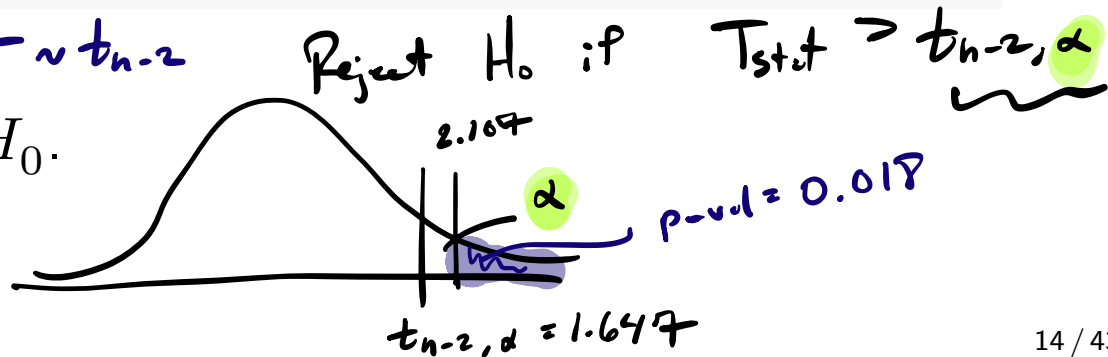
Tstat <- (b1hat - 2) / sqrt(sgsqhat/Sxx)
rej <- Tstat > qt(1-alpha, df = n - 2) =  $t_{n-2, \alpha} = 1.647$ 
pval <- 1 - pt(Tstat, df = n - 2)

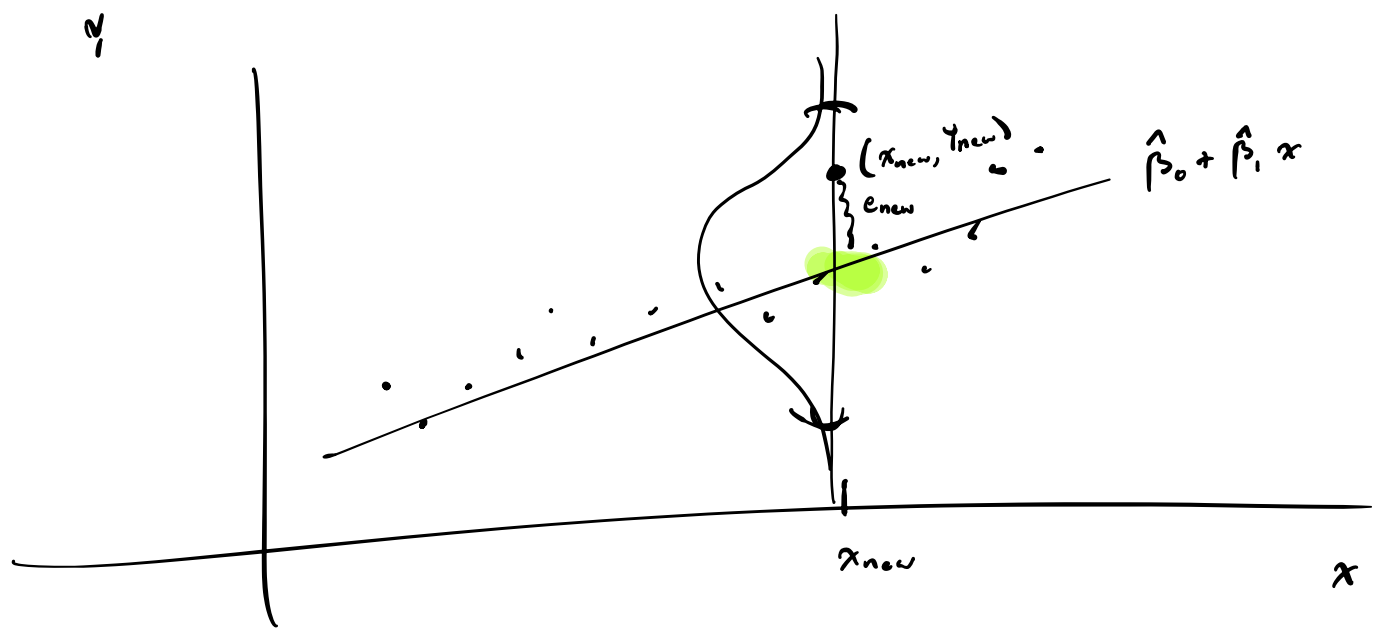
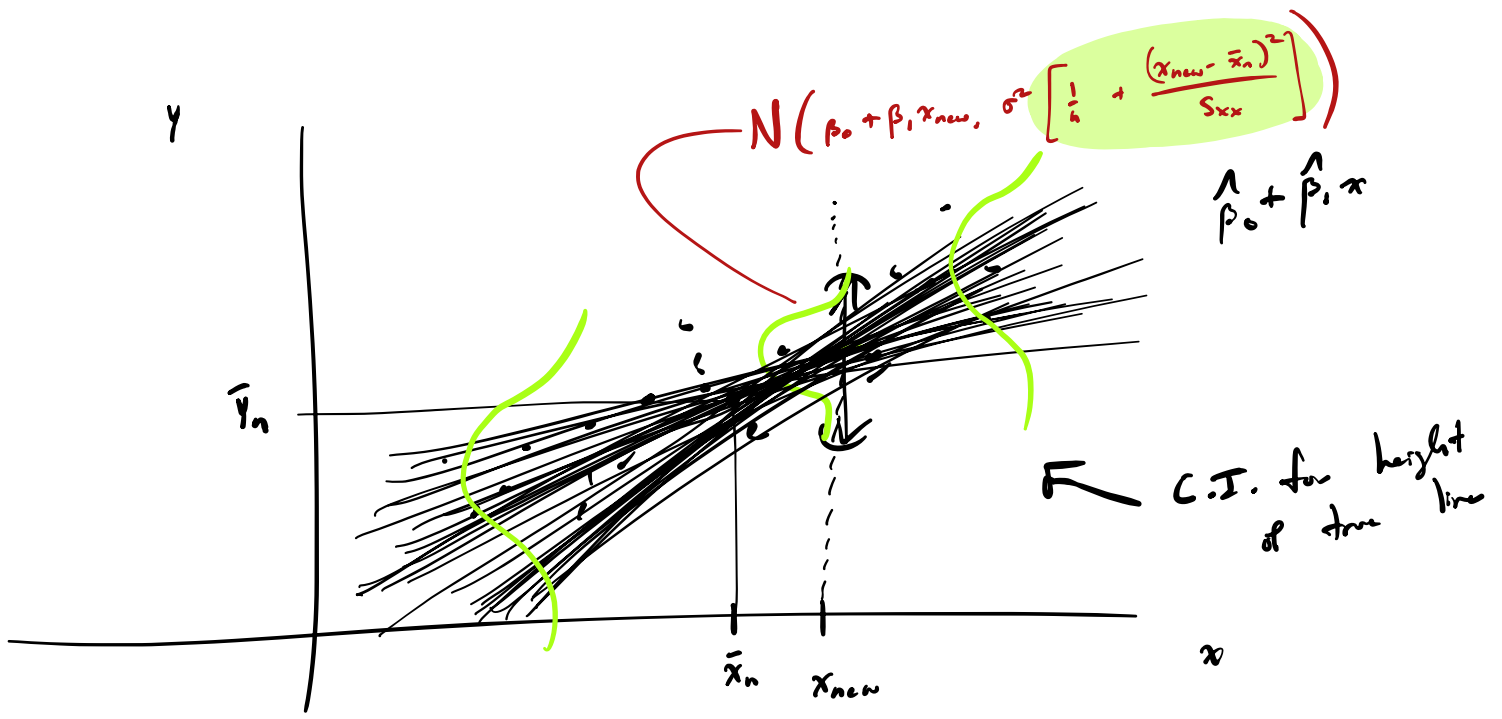
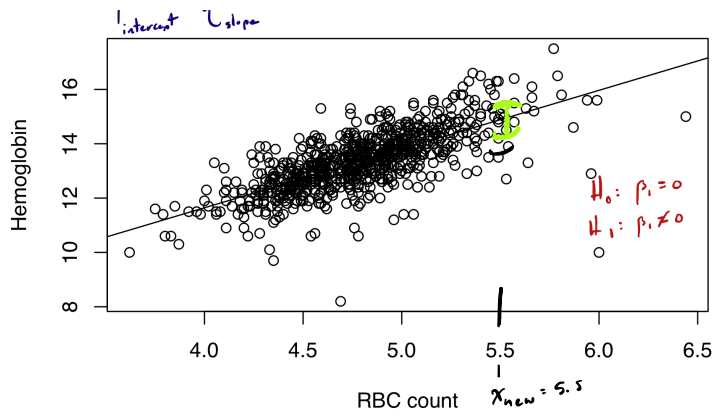
```

$T_{stat} = \frac{\hat{\beta}_1 - 2}{\hat{\sigma} / \sqrt{S_{xx}}} = 2.107$

$P(T > T_{stat})$, $T \sim t_{n-2}$

The p-value is 0.018; we reject H_0 .





Confidence interval for the height of the line

$$\text{lev}_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{S_{xx}}$$

Provided $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$, we have

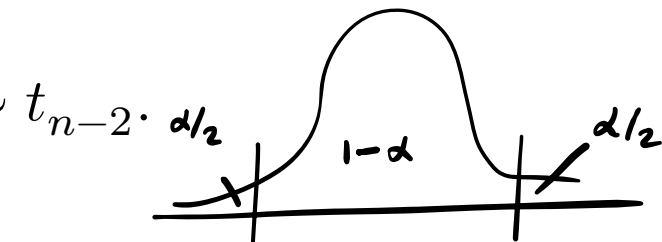
$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \sim \text{Normal} \left(\beta_0 + \beta_1 x_{\text{new}}, \sigma^2 \left[\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}} \right] \right).$$

↑ estimated height of line at x_{new}

“Studentizing” the above gives

subtract mean, divide by (estimated) standard deviation.

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} - (\beta_0 + \beta_1 x_{\text{new}})}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}}$$



So a $(1 - \alpha)100\%$ confidence interval for $\beta_0 + \beta_1 x_{\text{new}}$ is $t_{n-2, \alpha/2}$ $t_{n-2, \alpha/2}$

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{n-2, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}$$

$$P \left(-t_{n-2, \alpha/2} < \frac{\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} - (\beta_0 + \beta_1 x_{\text{new}})}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}} < t_{n-2, \alpha/2} \right) = 1 - \alpha$$

Hemoglobin versus RBC count example (cont)

Give a 95% CI for the mean hemoglobin level of individuals with RBC count 5.5.

$$x_{\text{new}} = 5.5$$

```
alpha <- 0.05
```

```
xnew <- 5.5
```

$$\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}$$

```
xnew_se <- sqrt(sgsqhat)*sqrt(1/n+(xnew-xbar)^2/Sxx) ←
```

```
ta2 <- qt(1-alpha/2, n-2) =  $t_{n-2, \alpha/2} = 1.963$ 
```

```
lo <- b0hat + b1hat * xnew - ta2 * xnew_se
```

```
up <- b0hat + b1hat * xnew + ta2 * xnew_se
```

$$\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}}_{\text{point estimate}} \pm \underbrace{t_{n-2, \alpha/2}}_{\text{critical value}} \underbrace{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}}_{\text{standard error}}$$

We are 95% confident that the mean hemoglobin level of individuals with RBC count 5.5 lies in the interval (14.767, 15.011).

Prediction interval for a new value of the response

let $(x_{\text{new}}, Y_{\text{new}})$ be the new data point.

Provided $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$, we have

$$\hat{e}_{\text{new}} = Y_{\text{new}} - \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \sim \text{Normal} \left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}} \right] \right).$$

$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

↑ this is the extra

“Studentizing” the above gives

$$\frac{Y_{\text{new}} - \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}} \sim t_{n-2}.$$

So a $(1 - \alpha)100\%$ prediction interval for Y_{new} is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{n-2, \alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}.$$

Hemoglobin versus RBC count example (cont)

Give a 95% prediction interval for the hemoglobin level when the RBC count is 5.5.

```
alpha <- 0.05
```

```
xnew <- 5.5
```

```
xnew_pse <- sqrt(sgsqhat)*sqrt(1+1/n+(xnew-xbar)^2/Sxx)
```

```
ta2 <- qt(1-alpha/2,n-2)
```

```
lo <- b0hat + b1hat * xnew - ta2 * xnew_pse
```

```
up <- b0hat + b1hat * xnew + ta2 * xnew_pse
```

$$\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(\bar{x}_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}$$

We are 95% confident that an individual with RBC count 5.5 will have a hemoglobin level in the interval (13.331, 16.447).

wider than the C.I.
for $\beta_0 + \beta_1 \bar{x}_{\text{new}}$.

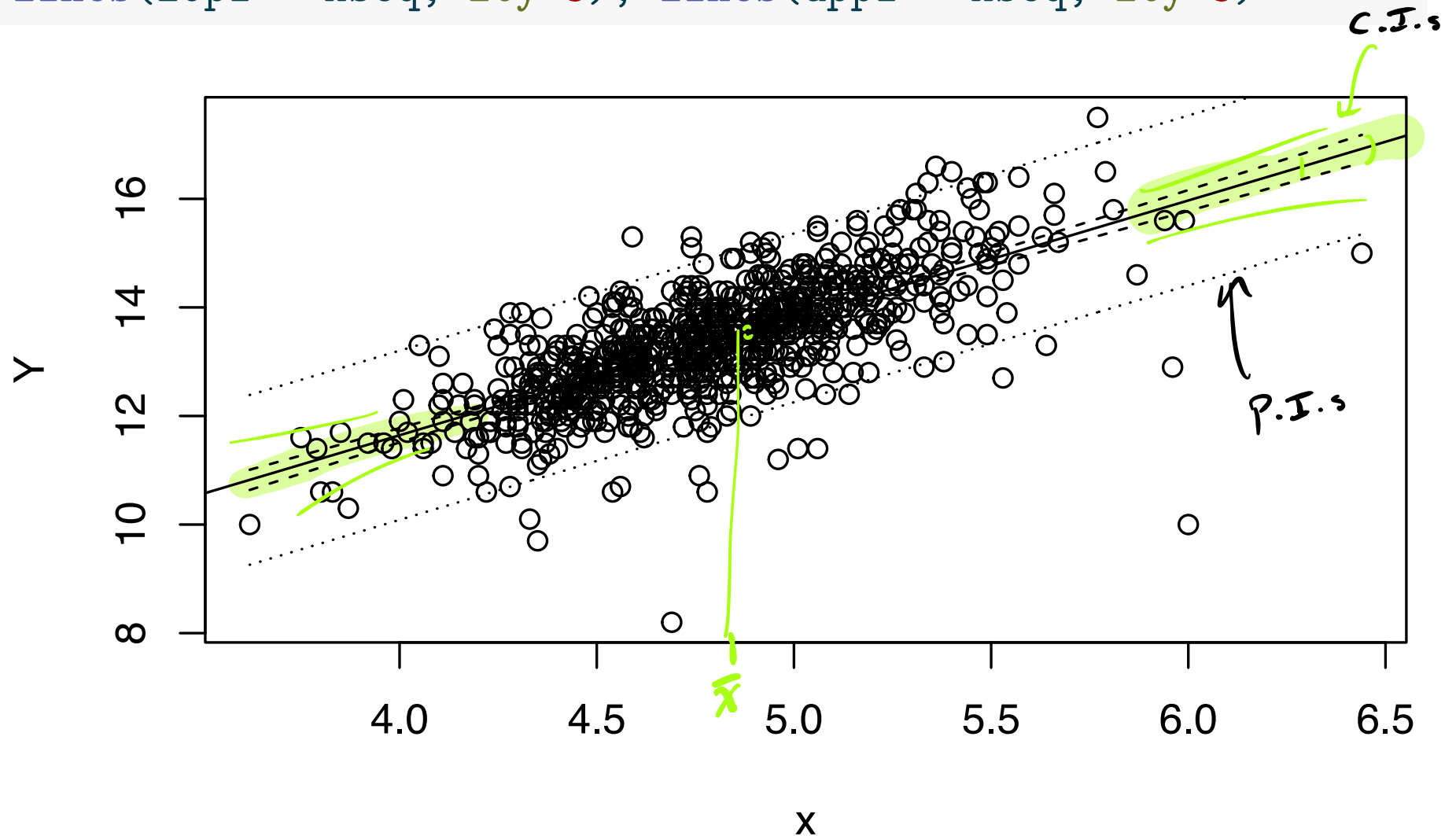
Plot confidence and prediction limits over the range of RBC counts.

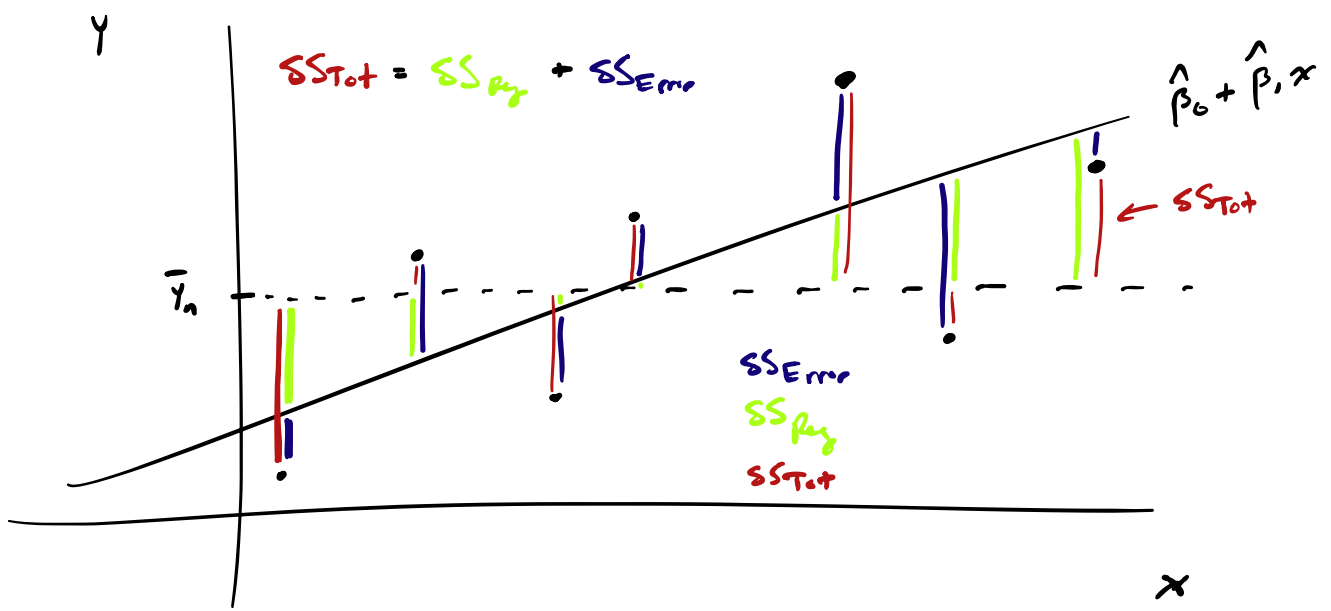
```
alpha <- 0.05
ta2 <- qt(1-alpha/2,n-2)
xseq <- seq(min(x),max(x),length = 500)

xseq_se <- sqrt(sgsqhat)*sqrt(1/n+(xseq-xbar)^2/Sxx)
loci <- b0hat + b1hat * xseq - ta2 * xseq_se
upci <- b0hat + b1hat * xseq + ta2 * xseq_se

xseq_pse <- sqrt(sgsqhat)*sqrt(1+1/n+(xseq-xbar)^2/Sxx)
lopi <- b0hat + b1hat * xseq - ta2 * xseq_pse
uppi <- b0hat + b1hat * xseq + ta2 * xseq_pse
```

```
plot(Y~x)
abline(b0hat,b1hat)
lines(loci ~ xseq, lty=2); lines(upci ~ xseq, lty=2)
lines(lopi ~ xseq, lty=3); lines(upper ~ xseq, lty=3)
```





Sums of squares in simple linear regression

We decompose the variation in Y_1, \dots, Y_n by defining the:

- ▶ Total sum of squares: $SS_{\text{Tot}} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \leftarrow S_{yy}$
- ▶ Regression sum of squares: $SS_{\text{Reg}} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$
- ▶ Error sum of squares: $SS_{\text{Error}} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{e}_i^2$

We have $SS_{\text{Tot}} = SS_{\text{Reg}} + SS_{\text{Error}}$.

Variation "explained" by x \uparrow $\underbrace{\hspace{10em}}_{\text{Variation due to random noise}}$

The coefficient of determination is defined as $R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Tot}}}$.

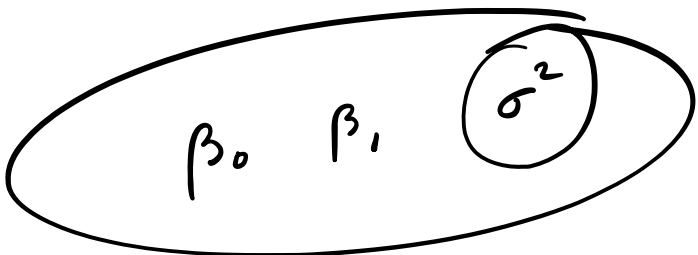
- ▶ $R^2 \in [0, 1]$
- ▶ Proportion of variation in Y "explained" by the covariate x .

▶ In simple linear regression, $R^2 = r_{xy}^2$ (tedious to show this).
 \uparrow Pearson's correlation

The mean squares in simple linear regression

The SS, appropriately scaled, follow chi-square distributions:

$$\begin{aligned}
 & \left. \begin{aligned}
 & \text{▶ } \frac{SS_{\text{Tot}}}{\sigma^2} \sim \chi_{n-1}^2(\phi_{\text{Total}}) \\
 & \text{▶ } \frac{SS_{\text{Reg}}}{\sigma^2} \sim \chi_1^2(\phi_{\text{Reg}}) \\
 & \text{▶ } \frac{SS_{\text{Error}}}{\sigma^2} \sim \chi_{n-2}^2
 \end{aligned} \right\} \chi_{n-1}^2(\phi)
 \end{aligned}$$



The quantities ϕ_{Tot} and ϕ_{Reg} are called noncentrality parameters.

Dividing SS_{Reg} and SS_{Error} by their dfs, we define:

$$\begin{aligned}
 & \text{▶ } \underline{\text{Regression mean square}}: MS_{\text{Reg}} = \frac{SS_{\text{Reg}}}{1} \\
 & \text{▶ } \underline{\text{Error mean square}}: MS_{\text{Error}} = \frac{SS_{\text{Error}}}{n-2} = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \hat{\sigma}^2
 \end{aligned}$$

The Analysis of Variance (ANOVA) table

We often present the SS, df, and MS values in a table like this:

Source	Df	SS	MS	F value	p-value
x Regression	1	SS_{Reg}	MS_{Reg}	F_{stat}	$P(F > F_{\text{stat}})$
Error	$n - 2$	SS_{Error}	MS_{Error}		
Total	$n - 1$	SS_{Tot}			

This is an example of an ANOVA table.

$$F_{\text{stat}} = \frac{MS_{\text{Reg}}}{MS_{\text{Error}}} \quad \text{in SLR.} \quad = \quad T_{\text{stat}}^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{xx}}} \right)^2$$

Can use to test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$.

p-value of this test.

Overall F test

In addition to the SS, df, and MS value, the ANOVA table presents

- ▶ $F_{\text{stat}} = \frac{MS_{\text{Reg}}}{MS_{\text{Error}}}$
- ▶ $P(F > F_{\text{stat}})$, where this is computed under $F \sim F_{1, n-2}$

These are the test statistic and p-value of the overall F test.

In simple linear regression this p-value is the same as the one for testing $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ with the t test, and $F_{\text{stat}} = T_{\text{stat}}^2$.

We will discuss the overall F test in greater detail later.

Turns out

$$F_{\text{stat}} = \frac{(n-2) r_{xy}^2}{1 - r_{xy}^2}$$

in simple linear regression.

Building the ANOVA table

```

SST <- sum((Y - Ybar)^2)
SSR <- sum((Yhat - Ybar)^2)
SSE <- sum((Y - Yhat)^2)
MSR <- SSR / 1
MSE <- SSE / (n-2)
Fstat <- MSR / MSE
pval <- 1 - pf(Fstat, 1, n-2)

```

$$n = 762$$

$$n - 2 = 760$$

$$\frac{491.37}{0.63}$$

Source	Df	SS	MS	F value	p-value
x	1	491.37	491.37	785.15	0
Error	760	475.63	0.63		
Total	761	967			

$$\frac{475.63}{760}$$

The `lm()`, `summary()`, and `anova()` functions in R

↑ linear model

```
lm_out <- lm(Y~x)
lm_out
```

Call:

```
lm(formula = Y ~ x)
```

Coefficients:

(Intercept)	x
2.994	2.163
$\hat{\beta}_0$	$\hat{\beta}_1$

summary(lm_out)

Call:

lm(formula = Y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-5.9702	-0.4232	0.0074	0.4645	2.3791

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.99447 $\hat{\beta}_0$	0.37065	8.079	2.56e-15 ***
x	2.16263 $\hat{\beta}_1$	0.07718	28.021	< 2e-16 ***

$$T_{stat} = \frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{xx}}}$$

Default:

$H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$

For $H_0: \beta_0 = 0$
 vs $H_1: \beta_0 \neq 0$.
 ~ Usually unimportant.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7911 on 760 degrees of freedom

Multiple R-squared: 0.5081, Adjusted R-squared: 0.5075

F-statistic: 785.1 on 1 and 760 DF, p-value: < 2.2e-16

$F_{stat} \rightarrow R^2 = \frac{SS_{Reg}}{SS_{Total}}$

```
anova(lm_out)
```

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<i>Regression</i> x	1	491.37	491.37	785.15	< 2.2e-16 ***
<i>Error</i> Residuals	760	475.63	0.63		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The predict() function in R

from lm()

```
predict(lm_out, newdata = data.frame(x = 5.5), int = "conf")
```

$x_{new} = 5.5$

$\hat{\beta}_0 + \hat{\beta}_1 x_{new}$

	fit	lwr	upr
1	14.88892	14.76725	15.01059

same as before

(14.767, 15.011).

```
predict(lm_out, newdata = data.frame(x = 5.5), int = "pred")
```

	fit	lwr	upr
1	14.88892	13.33117	16.44667

(13.331, 16.447).

$\hat{\beta}_0 + \hat{\beta}_1 x_{new}$

same

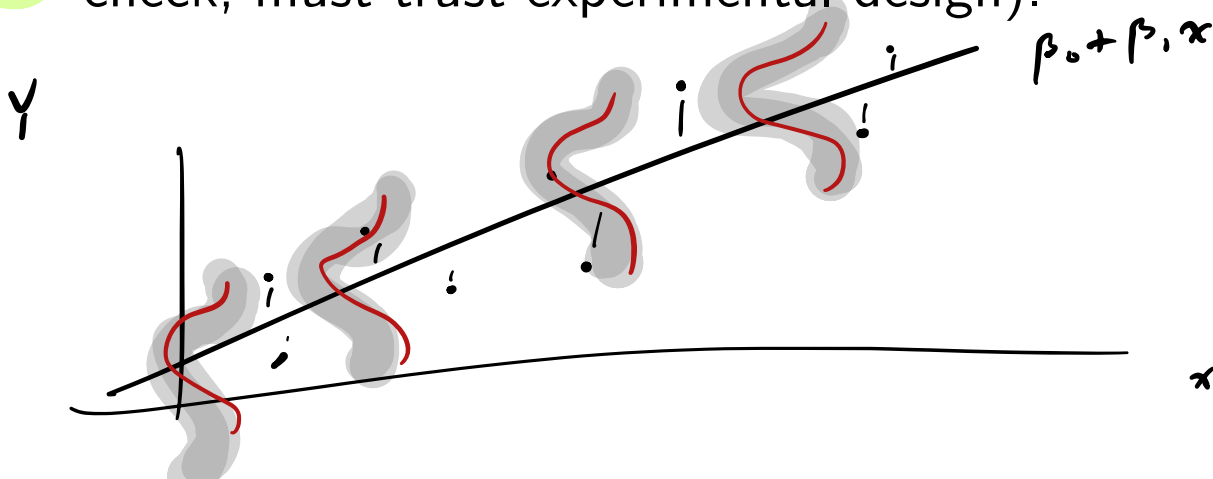
Checking model assumptions

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad i=1, \dots, n.$$

\uparrow independent

Validity of the foregoing analyses depends on these assumptions:

1. The responses are normally distributed around the regression line (Check QQ plot of residuals). *If n is large this doesn't matter.*
2. The response has the same variance for all values of the covariate (Check residuals vs fitted values plot).
3. The covariate and the response are linearly related (Check residuals vs fitted values plot).
4. The response values are independent of each other (No way to check; must trust experimental design).

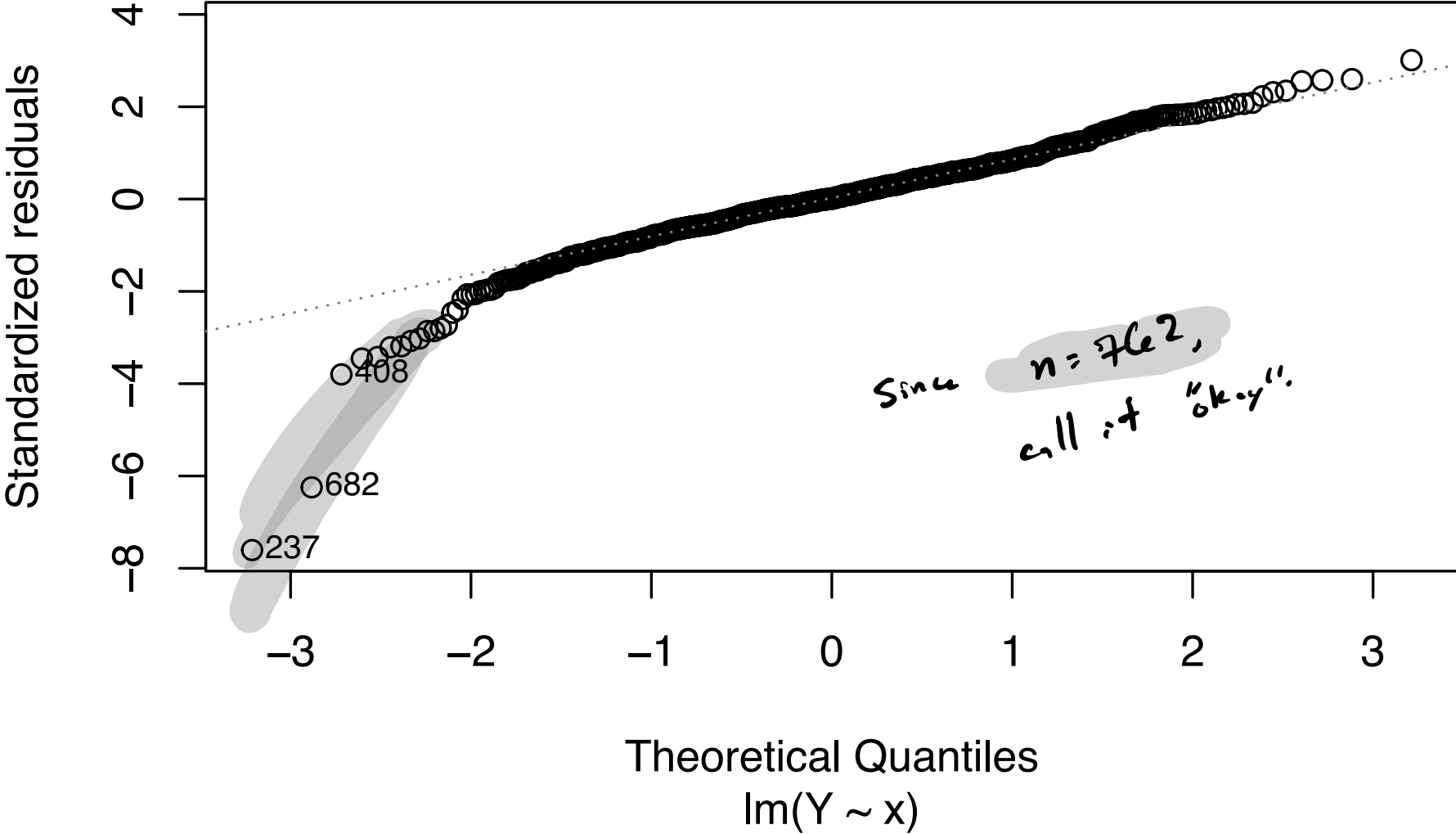


Generating diagnostic plots from `lm()` with `plot()`

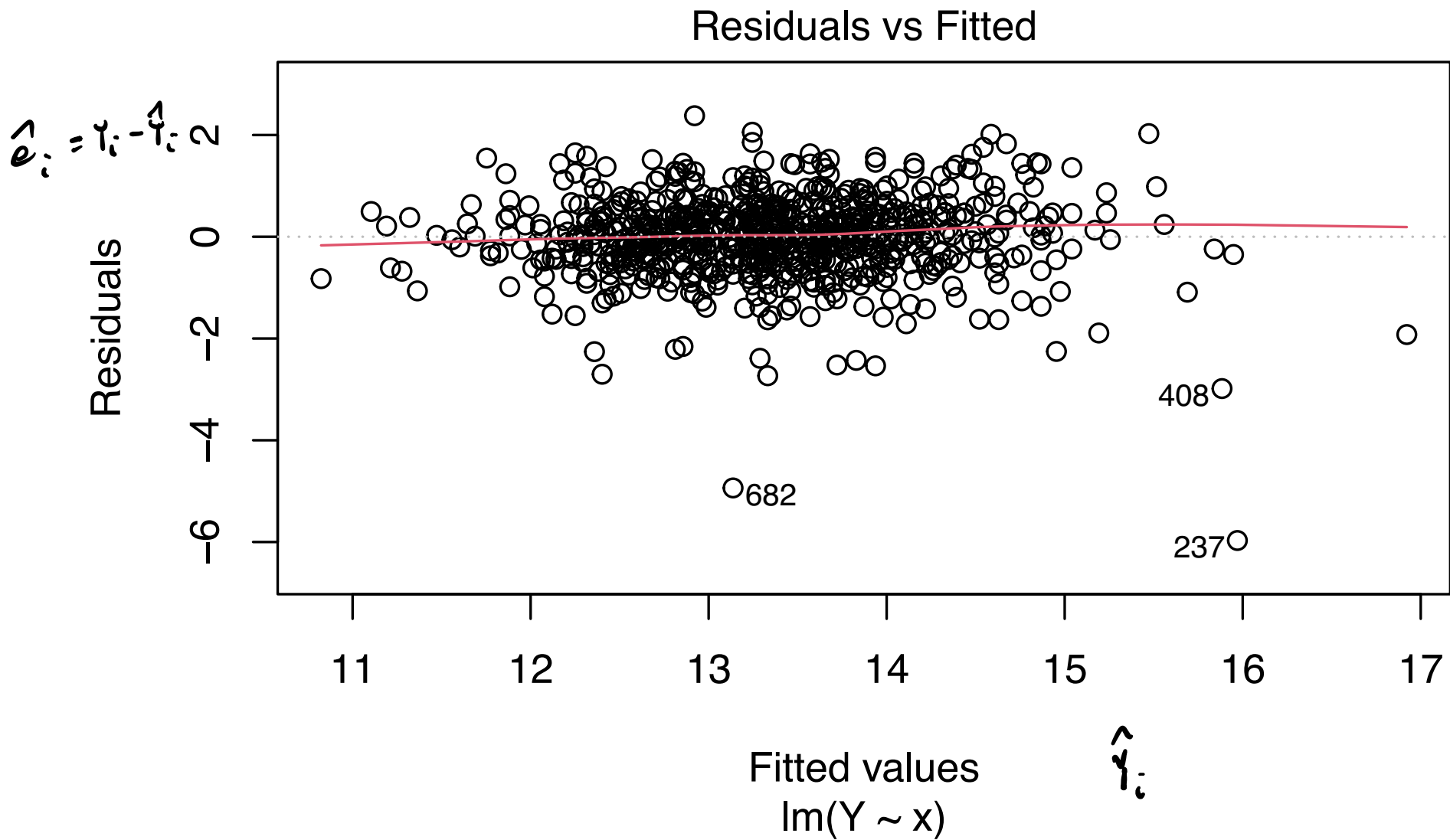
```
plot(lm_out, which = 2)
```

$\hookrightarrow \text{lm}(Y \sim x)$

Q-Q Residuals



```
plot(lm_out, which = 1)
```



Abalone data example

Predict shucked weight of an abalone by its length.

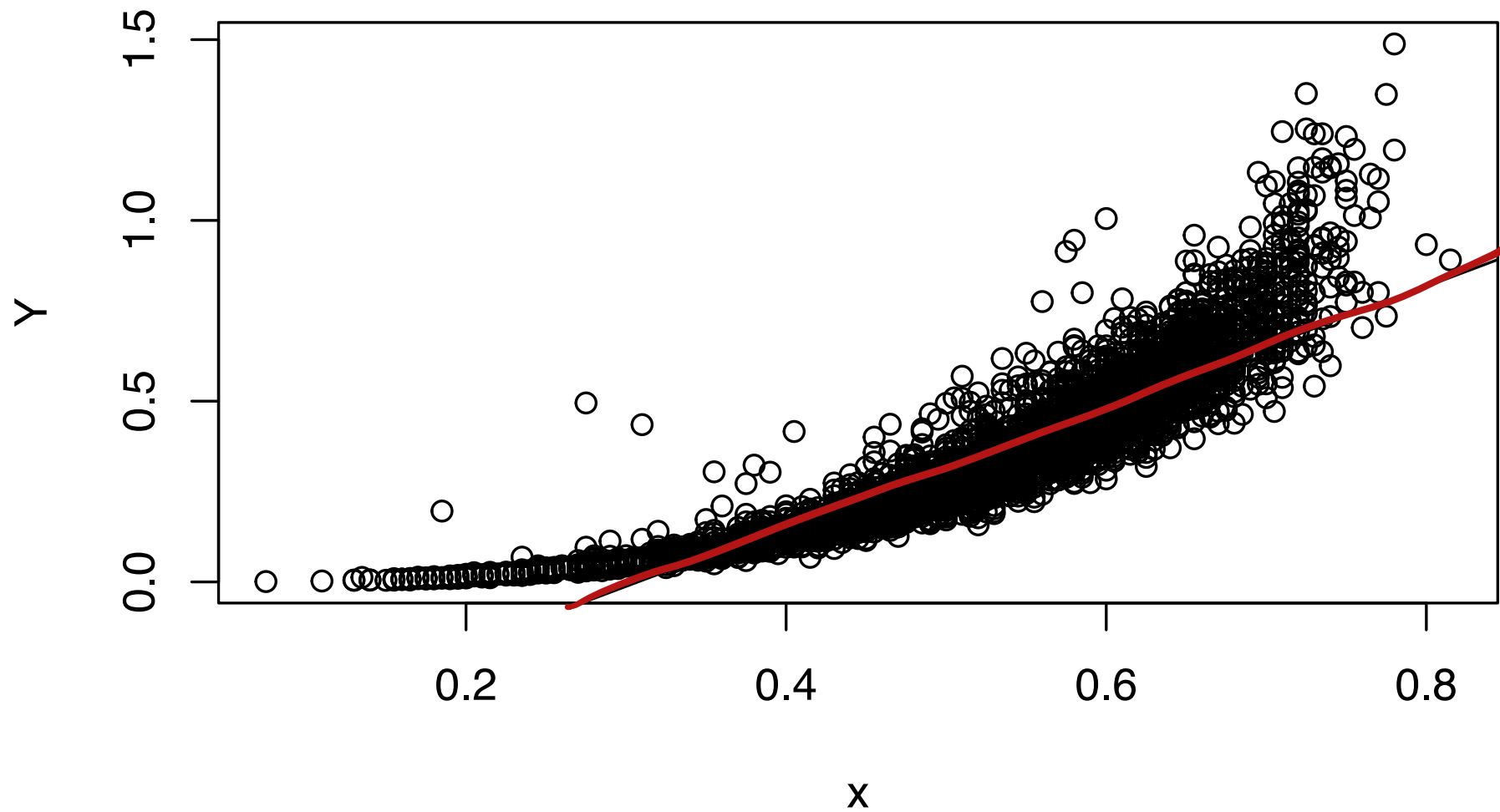
```
l <- url("https://people.stat.sc.edu/gregorkb/data/abalone.csv")
abalone <- read.csv(l, col.names = c("Sex",
                                     "Length",
                                     "Diameter",
                                     "Height",
                                     "Whole_Wt",
                                     "Shucked_Wt",
                                     "Viscera_Wt",
                                     "Shell_Wt",
                                     "Rings"))

Y <- abalone$Shucked Wt
x <- abalone$Length
n <- length(Y)
```

There are $n = 4176$ records. Data come from Nash and Ford (1995).

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

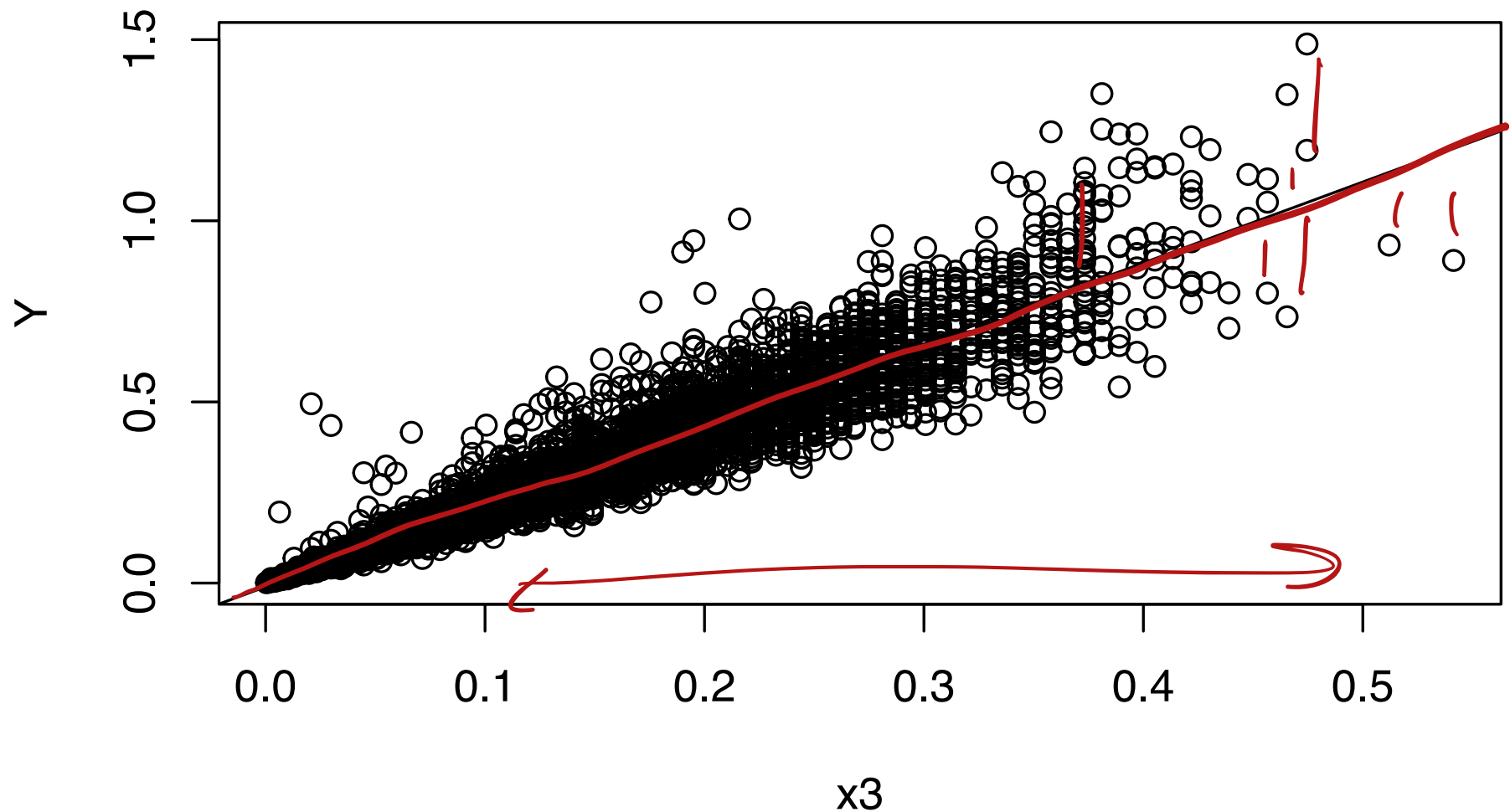
```
lm1 <- lm(Y~x)  
plot(Y~x)  
abline(lm1)
```



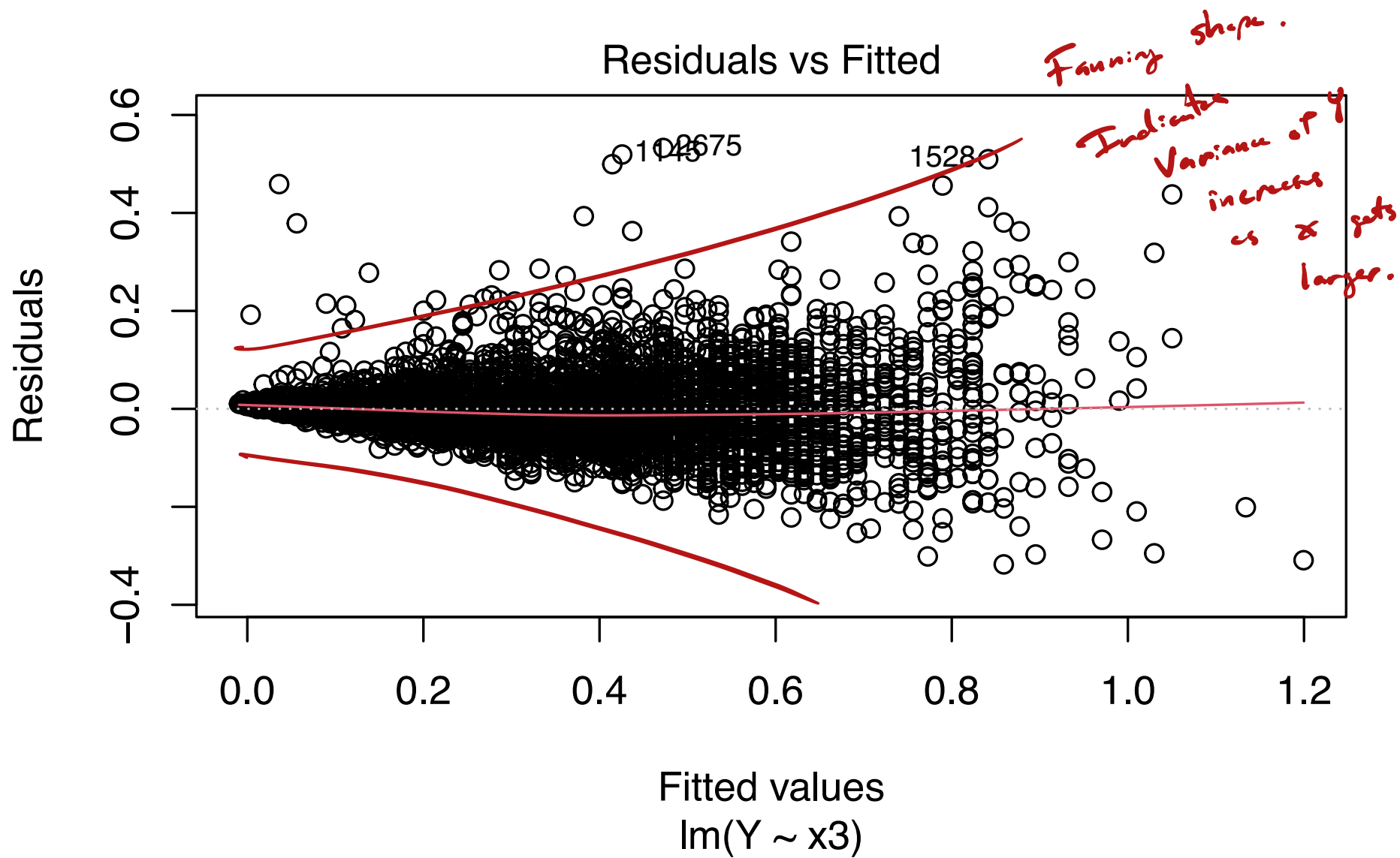
Try transforming x:

$$Y_i = \beta_0 + \beta_1 x_i^3 + \varepsilon_i$$

```
x3 <- x^**3  
lm2 <- lm(Y ~ x3)  
plot(Y ~ x3)  
abline(lm2)
```



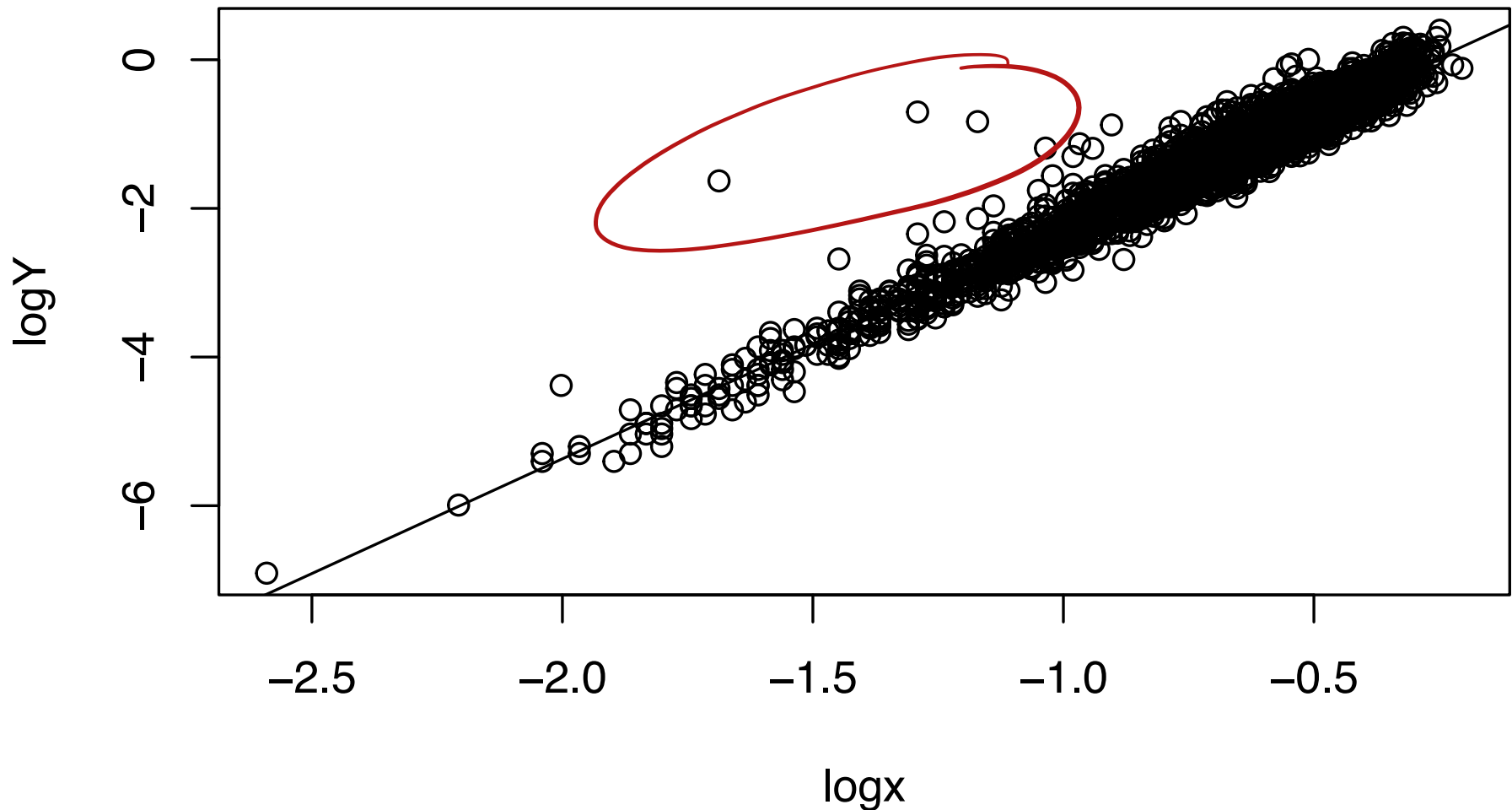
```
plot(lm2, which = 1)
```



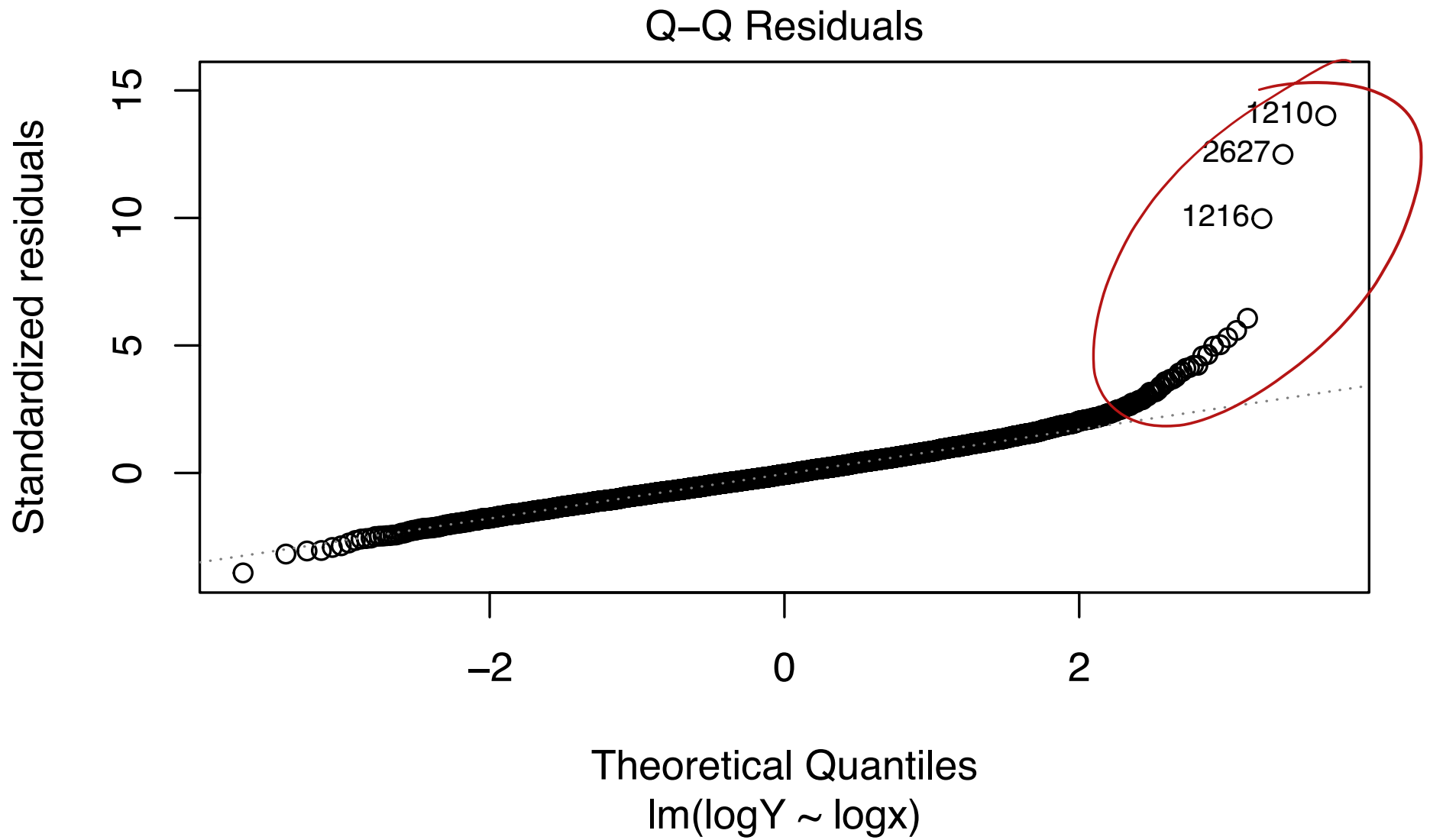
Could transform both Y and x:

$$\log Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$$

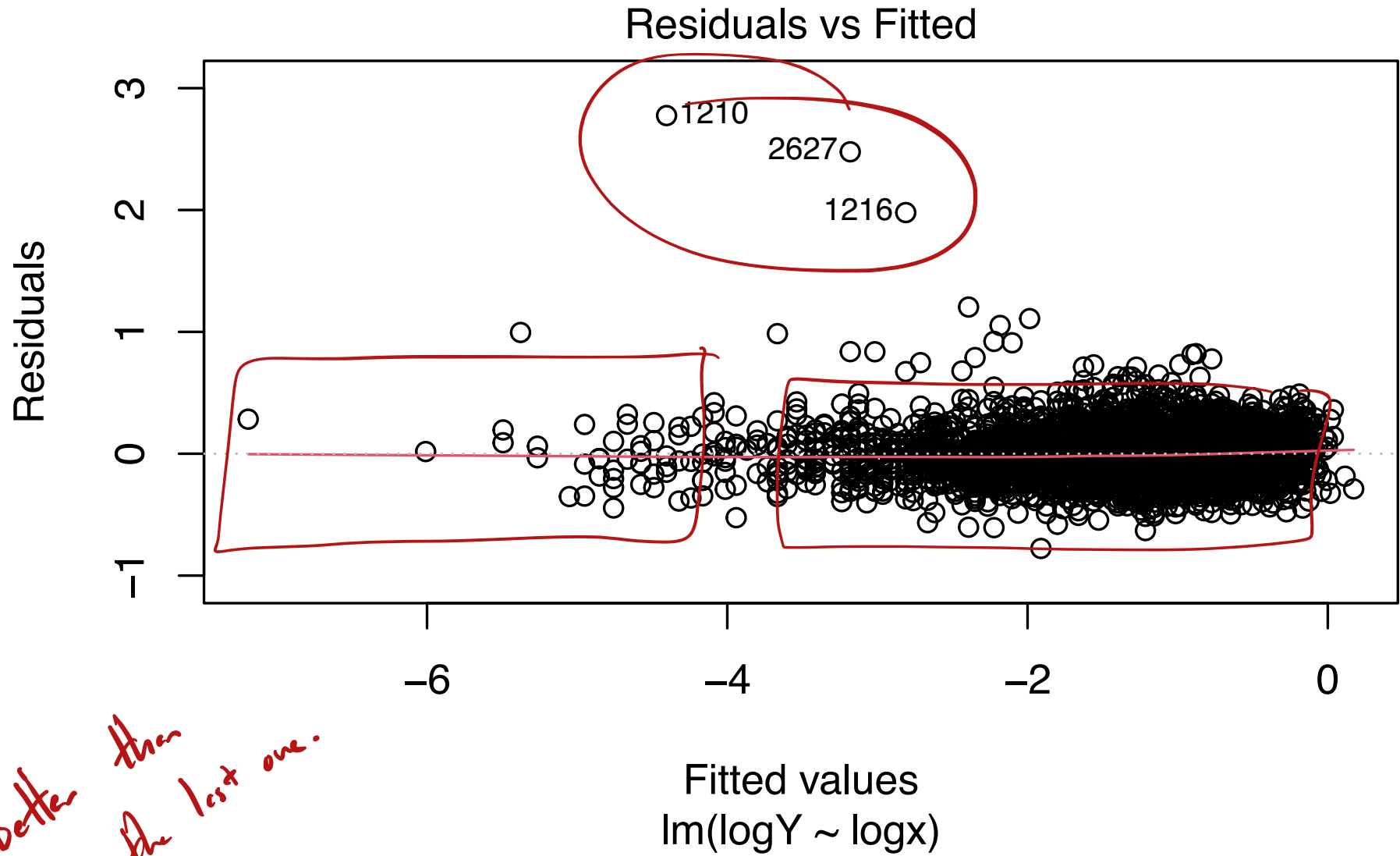
```
logY <- log(Y); logx <- log(x) % change Y_i = beta_1 * % change in X_i  
lm3 <- lm(logY ~ logx)  
plot(logY~logx); abline(lm3)
```




```
plot(lm3, which = 2)
```



```
plot(lm3, which = 1)
```



Transforming variables to obtain a linear relationship

Take care how to interpret β_1 after transforming the data.

Example: Log transforming x and Y gives β_1 a %-change interpretation:

$$\log y = \beta_0 + \beta_1 \log x \iff \frac{d \log y}{dx} = \beta_1 \frac{1}{x} \iff \frac{dy}{y} = \beta_1 \frac{dx}{x}$$

Abalone data example (cont)

We must back-transform prediction intervals if we have transformed Y .

predict w/ of abalone with length 0.5

```
xnew <- 0.5  
newdata <- data.frame( logx = log(xnew))  
pi_logY <- predict(lm3, newdata = newdata, int = "pred")  
pi_logY
```

$\sim \text{lm}(\log Y \sim \log x)$

```
          fit          lwr          upr  
1 -1.335636 -1.724831 -0.946441
```

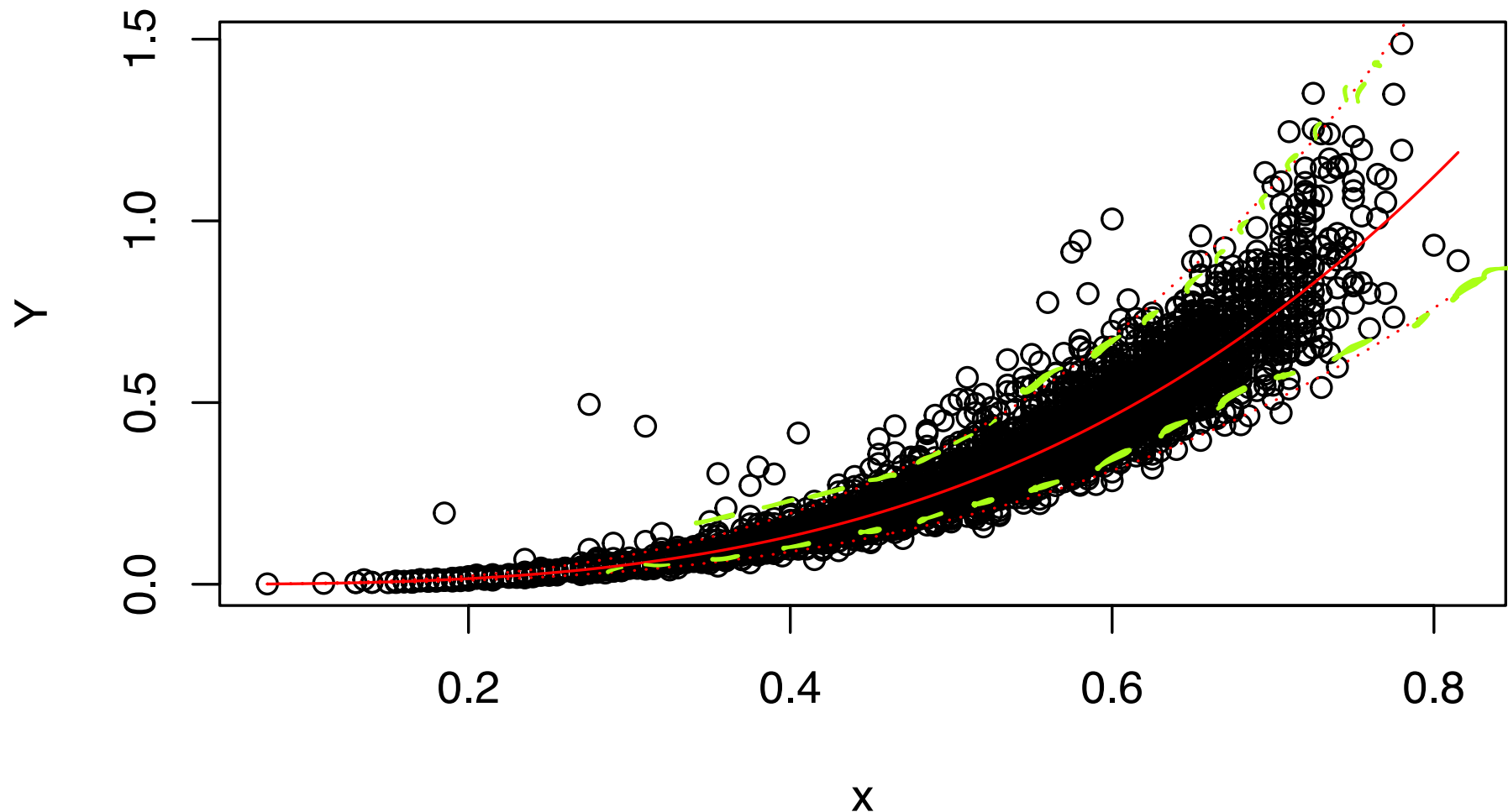
$\log Y \in (a, b)$

```
pi <- exp(pi_logY)  
pi
```

$Y \in (e^a, e^b)$

```
          fit          lwr          upr  
1  0.2629909  0.1782032  0.3881199
```

```
plot(Y~x); logx <- seq(min(logx),max(logx),length=500)
newdata <- data.frame(logx = logx)
logy_hat <- predict(lm3,newdata = newdata,int = "pred")
lines(exp(logy_hat[,1]) ~ exp(logx), col = "red")
lines(exp(logy_hat[,2]) ~ exp(logx), col = "red", lty = 3)
lines(exp(logy_hat[,3]) ~ exp(logx), col = "red", lty = 3)
```



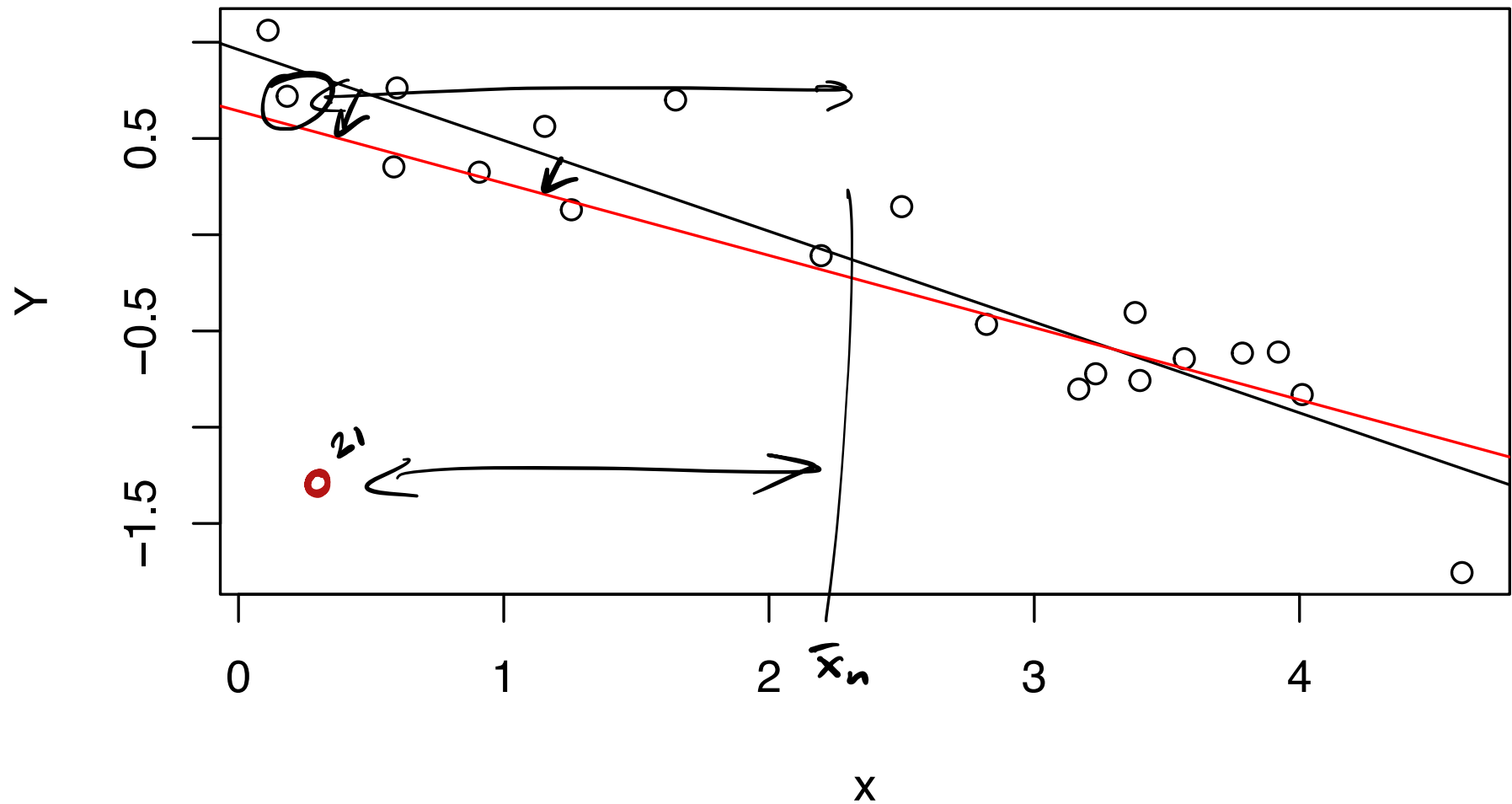
Outliers in simple linear regression

Outlying data points can have a large influence on the estimated regression function.

Let's generate some data and then add an outlier:

```
n <- 20
b0 <- 1
b1 <- -1/2
sg <- .2
x0 <- runif(n,0,5)
e <- rnorm(n,0,sg)
Y0 <- b0 + b1 * x0 + e
x <- c(x0,.3)
Y <- c(Y0,-1.3)
```

```
plot(Y~x);points(Y[n+1]~x[n+1], col = "red")
abline(lm(Y0~x0))
abline(lm(Y~x), col = "red")
```



The red data point appears to exert an undue influence over the fit.

Leverage and Cook's distance

The leverage of a point (x_i, Y_i) among $(x_1, Y_1), \dots, (x_n, Y_n)$ is

$$\text{lev}_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{S_{xx}} \quad \leftarrow \text{how far is } x_i \text{ from } \bar{x}_n?$$

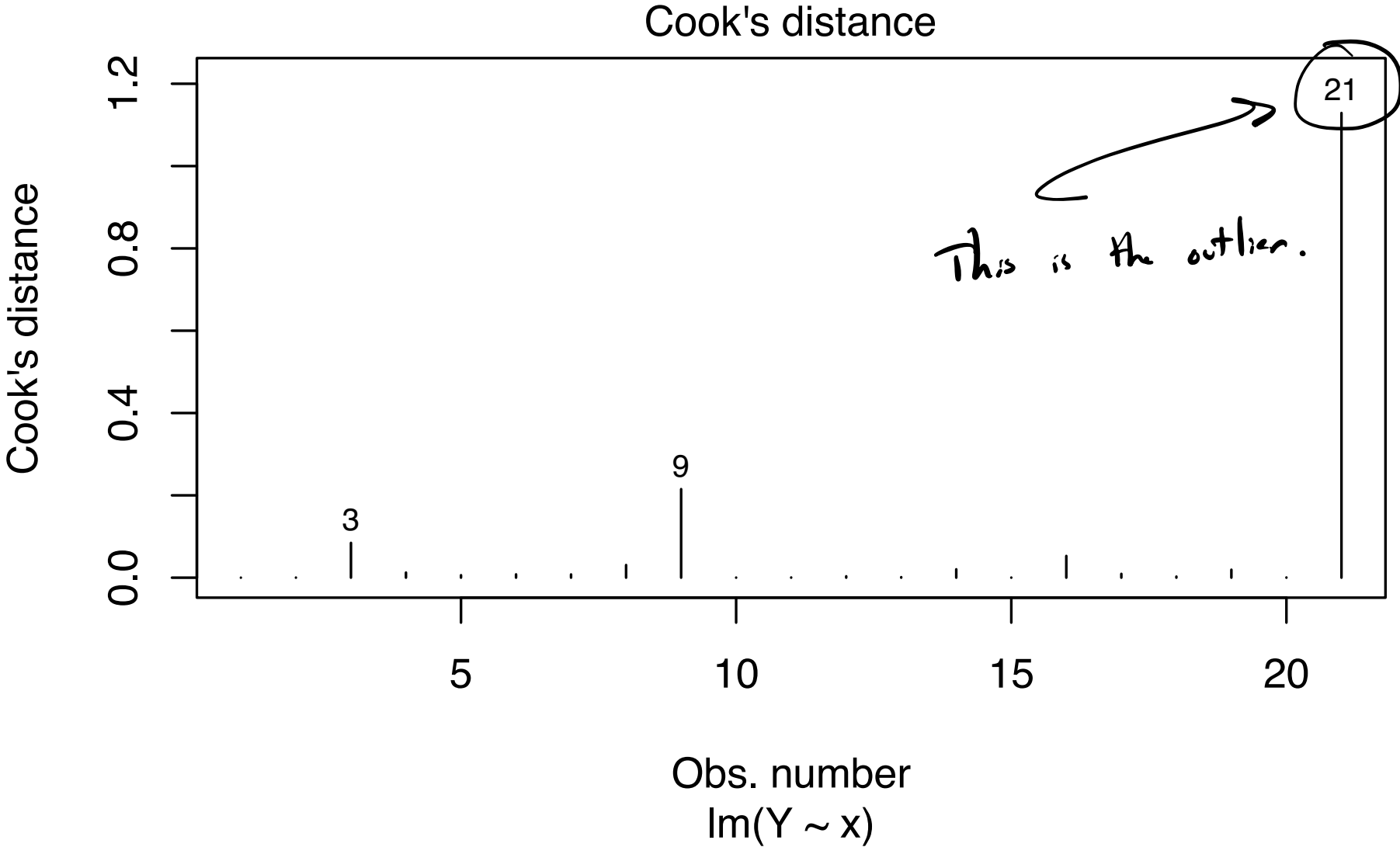
Leverage only shows outlying-ness in the x direction.

Cook's Distance measures how much each data point changes the fit:

$$D_i = \frac{\hat{e}_i^2}{2\hat{\sigma}^2} \frac{\text{lev}_i}{(1 - \text{lev}_i)^2}.$$

Make a plot of the cooks distances

```
plot(lm(Y~x), which = 4)
```



Code to compute Cook's distances

```
n <- length(Y)
xbar <- mean(x)
Ybar <- mean(Y)
rxY <- cor(x,Y) # Pearson's correlation coefficient
Sxx <- sum((x - xbar)^2)
SYY <- sum((Y - Ybar)^2)
b1hat <- rxY * sqrt(SYY / Sxx)
b0hat <- Ybar - b1hat * xbar
lev <- 1/n + (x - xbar)^2/Sxx
Yhat <- b0hat + b1hat * x
ehat <- Y - Yhat
sgsqhat <- sum(ehat^2)/(n-2)

cooksD <- ehat^2 / (2*sgsqhat) * lev / (1 - lev)^2
```

ADDENDUM: Two-sample t-test by SLR

Let $Y_{ij} \stackrel{\text{ind}}{\sim} \text{Normal}(\mu_i, \sigma^2)$, $j = 1, \dots, n_i$, $i = 1, 2$ and consider

$$H_0: \mu_2 - \mu_1 = 0 \text{ versus } H_1: \mu_2 - \mu_1 \neq 0.$$

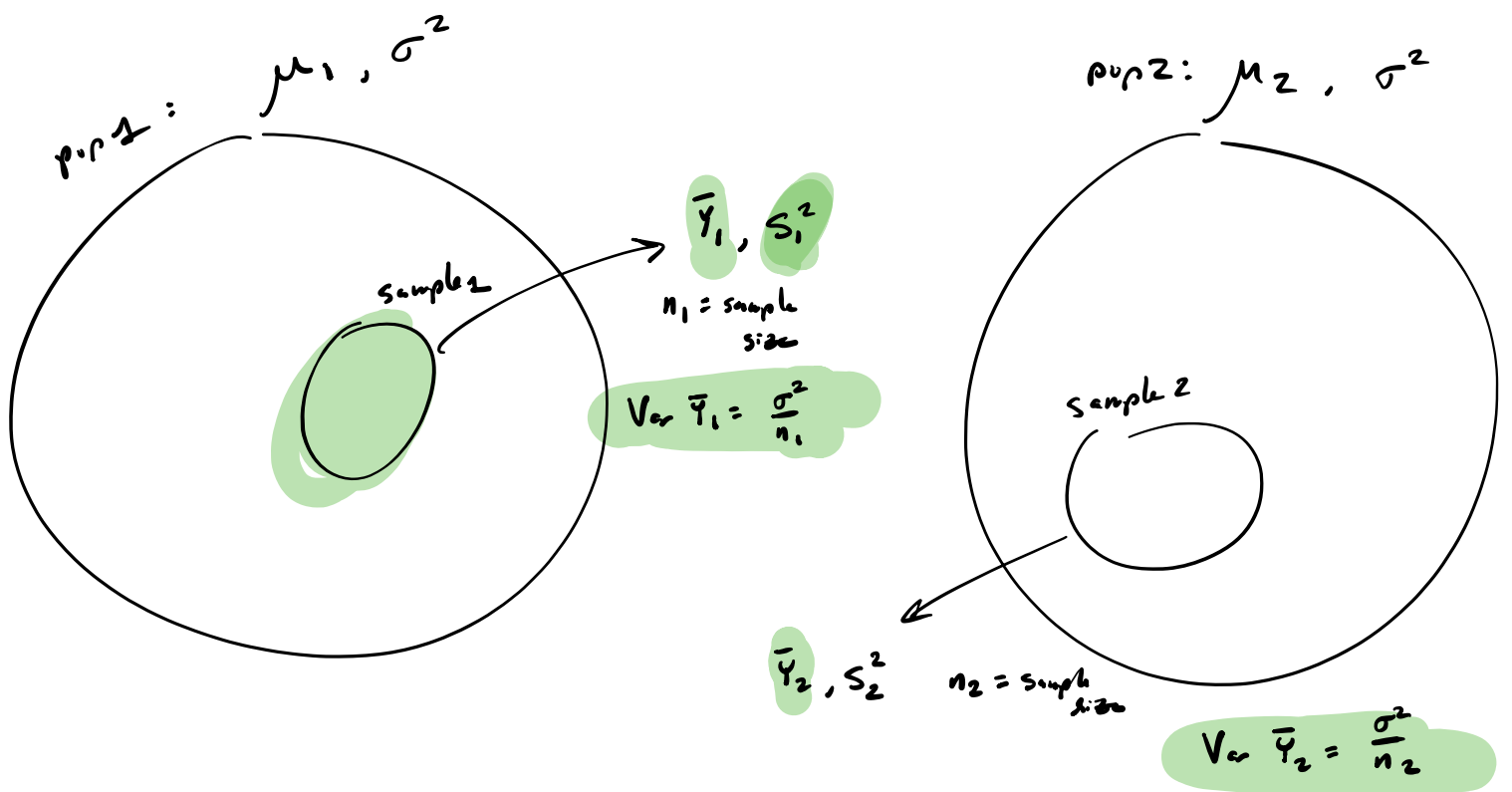
The (equal-variances) two-sample t-test uses the test statistic

$$T_{\text{stat}} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$, $i = 1, 2$ and

$$S_{\text{pooled}}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \quad S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

We reject H_0 at significance level α if $|T_{\text{stat}}| > t_{n-2, \alpha/2}$.



$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$



$$H_0: \mu_2 - \mu_1 = 0$$

$$H_1: \mu_2 - \mu_1 \neq 0$$

$$\bar{y}_2 - \bar{y}_1 \sim \text{Normal} \left(\mu_2 - \mu_1, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \right)$$

$$T_{\text{stat}} = \frac{\bar{y}_2 - \bar{y}_1 - (0)}{\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}}$$

estimate

$$= \frac{\bar{y}_2 - \bar{y}_1}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\hat{\sigma}^2 = S^2_{\text{pooled}} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$\sim t_{n-2}$ under H_0 . So reject H_0 when $|T_{\text{stat}}| \geq t_{n-2, \alpha/2}$

Appendicitis example

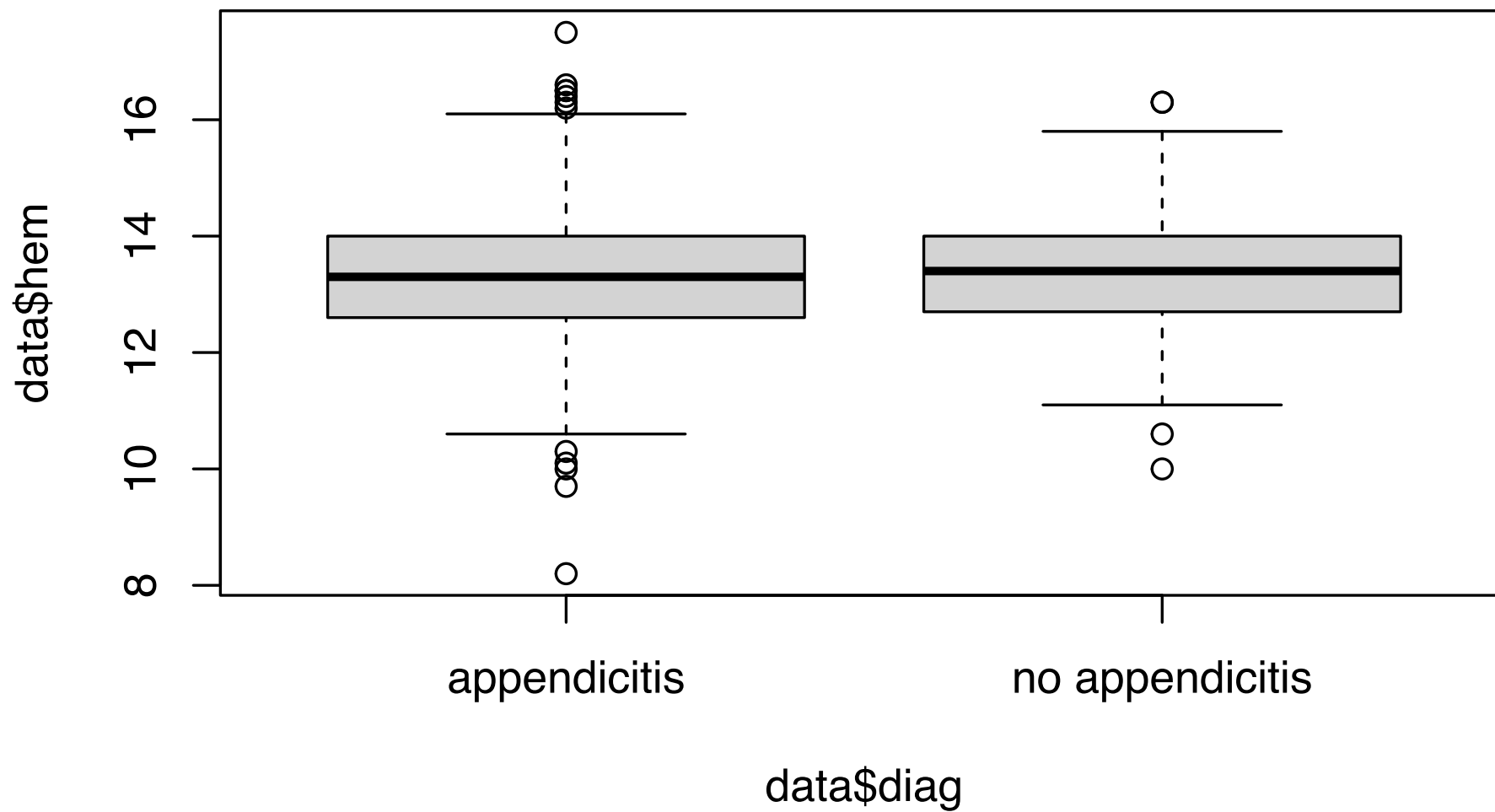
Look again at the data from Marcinkevičs et al. (2023).

```
link <- url("https://people.stat.sc.edu/gregorkb/data/hrbc.csv")
data <- read.csv(link)
head(data)
```

	hem	rbc	sex	age		diag
1	14.8	5.27	female	12.68		appendicitis
2	15.7	5.26	male	14.10	no	appendicitis
3	11.4	3.98	female	14.14	no	appendicitis
4	13.6	4.64	female	16.37	no	appendicitis
5	12.6	4.44	female	11.08		appendicitis
6	12.5	4.96	male	11.05	no	appendicitis

Is the mean hemoglobin level the same in children with and without appendicitis?

```
boxplot(data$hem ~ data$diag)
```



```
t.test(data$hem ~ data$diag, var.equal = TRUE)
```

Tst.t

$n_1 + n_2 - 2$

$$H_0: \mu_{app} - \mu_{no\ app} = 0$$

$$H_1: \mu_{app} - \mu_{no\ app} \neq 0$$

Two Sample t-test

data: data\$hem by data\$diag

t = -0.49212, df = 760, p-value = 0.6228

← same p-value as in SLR

alternative hypothesis: true difference in means between group appendicitis

95 percent confidence interval:

-0.2038964 0.1221585

sample estimates:

mean in group appendicitis mean in group no appendicitis

13.33229

13.37316

$\hat{\beta}_0$

↑
appendicitis

↑
no appendicitis

Two-sample t-test by simple linear regression

$$(x_i, Y_i), i = 1, \dots, n$$

"dummy" variable

Let the Y_i be the hemoglobin values and define an indicator variable as

$$x_i = \begin{cases} 0 & \text{if no appendicitis} \\ 1 & \text{if appendicitis} \end{cases} \quad \text{for } i = 1, \dots, n.$$

$\leftarrow n_1$ values of 0 for x_i $\leftarrow n_1 + n_2$

$\leftarrow n_2$ values of 1 for x_i

Then in the SLR model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ we have

- ▶ $\beta_0 = \mu_{\text{no app}}$
- ▶ $\beta_0 + \beta_1 = \mu_{\text{app}}$
- ▶ $\beta_1 = \mu_{\text{app}} - \mu_{\text{no app}}$

$$Y_i = \begin{cases} \beta_0 + \varepsilon_i & \text{if no appendicitis} \\ \beta_0 + \beta_1 & \text{if appendicitis} \end{cases}$$

The t test in the simple linear regression setup of

$$H_0: \beta_1 = 0 \text{ versus } H_1: \beta_1 \neq 0$$

will give the same p value as the equal-variances two-sample t test of

$$H_0: \mu_{\text{app}} - \mu_{\text{no app}} = 0 \text{ versus } H_1: \mu_{\text{app}} - \mu_{\text{no app}} \neq 0. \quad \text{Cool!}$$

Exercise: Show that in the above setup we have

$$T_{\text{stat}} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{xx}}}$$

$\hookrightarrow S_{xx} = \sum_{i=1}^n (x_i - \bar{x}_n)^2$

Do it in steps, showing:

1. $\hat{\beta}_0 = \bar{Y}_1 - \bar{Y}_n = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$
2. $\hat{\beta}_1 = \bar{Y}_2 - \bar{Y}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \quad n = n_1 + n_2$
3. $\hat{\sigma} = S_{\text{pooled}}$
4. $1 / \sqrt{S_{xx}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$\frac{1}{\hat{\sigma}^2} = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n_1+n_2-2} \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2$$

$$= \frac{1}{n_1+n_2-2} \sum_{i=1}^n \left(y_i - \left(\bar{y}_1 + (\bar{y}_2 - \bar{y}_1) x_i \right) \right)^2$$

$$= \frac{1}{n_1 + n_2 - 2} \left[\sum_{i: x_i = 0} (y_i - \bar{y}_1)^2 + \sum_{i: x_i = 1} (y_i - \bar{y}_2)^2 \right]$$

$$= \frac{1}{n_1 + n_2 - 2} \left[(n_1 - 1) \underbrace{\frac{\sum_{i: x_i = 0} (y_i - \bar{y}_1)^2}{n_1 - 1}}_{S_1^2} + (n_2 - 1) \underbrace{\frac{\sum_{i: x_i = 1} (y_i - \bar{y}_2)^2}{n_2 - 1}}_{S_2^2} \right]$$

$$= \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

$$= S_{\text{pooled}}^2$$

Appendicitis example (cont)

Prepare the data:

```
Y <- data$hem  
x <- as.numeric(data$diag == "appendicitis")  
head(cbind(Y,x))
```

TRUE/FALSE

	Y	x	
[1,]	14.8	1	app
[2,]	15.7	0	no app...
[3,]	11.4	0	
[4,]	13.6	0	
[5,]	12.6	1	
[6,]	12.5	0	

summary(lm(Y~x))

← Hemoglobin
↑ 0s and 1s

Call:
lm(formula = Y ~ x)

Residuals:
Min 1Q Median 3Q Max
-5.1323 -0.7323 -0.0323 0.6677 4.1677

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.37316	0.06375	209.782	<2e-16 ***
x	-0.04087	0.08305	-0.492	0.623

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.128 on 760 degrees of freedom
Multiple R-squared: 0.0003186, Adjusted R-squared: -0.0009968
F-statistic: 0.2422 on 1 and 760 DF, p-value: 0.6228

$\hat{\beta}_0 = 13.37316$
 $\hat{\beta}_1 = -0.04087 = \bar{Y}_{app} - \bar{Y}_{no-app}$

$\frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{xx}}} = T_{stat} = \frac{\bar{Y}_{app} - \bar{Y}_{no-app}}{S_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

p-value for testing
 $H_0: \beta_1 = 0$
↑
 $\mu_{app} - \mu_{no-app}$

Automatic if we designate the predictor as a "factor" (but watch sign!).

```
x <- as.factor(data$diag)
summary(lm(data$hem ~ x))
```

is a "factor" type object

Call:
lm(formula = data\$hem ~ x)

$$x_i = \begin{cases} 1 & \text{no app} \\ 0 & \text{app} \end{cases}$$

Residuals:

Min	1Q	Median	3Q	Max
-5.1323	-0.7323	-0.0323	0.6677	4.1677

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.33229	0.05322	250.490	<2e-16 ***
xno appendicitis	0.04087	0.08305	0.492	0.623

\bar{y}_{app}

$\bar{y}_{no\ app} - \bar{y}_{app}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.128 on 760 degrees of freedom

Multiple R-squared: 0.0003186, Adjusted R-squared: -0.0009968

F-statistic: 0.2422 on 1 and 760 DF, p-value: 0.6228

References

- Marcinkevičs, Ričards, Patricia Reis Wolfertstetter, Ugne Klimiene, Ece Ozkan, Kieran Chin-Cheong, Alyssia Paschke, Julia Zerres, et al. 2023. “Regensburg Pediatric Appendicitis Dataset.” Zenodo. <https://doi.org/10.5281/zenodo.7711412>.
- Nash, Sellers, Warwick, and Wes Ford. 1995. “Abalone.” UCI Machine Learning Repository. <https://doi.org/10.24432/C55C7W>.