STAT 515 Lec 18 slides

Simple linear regression

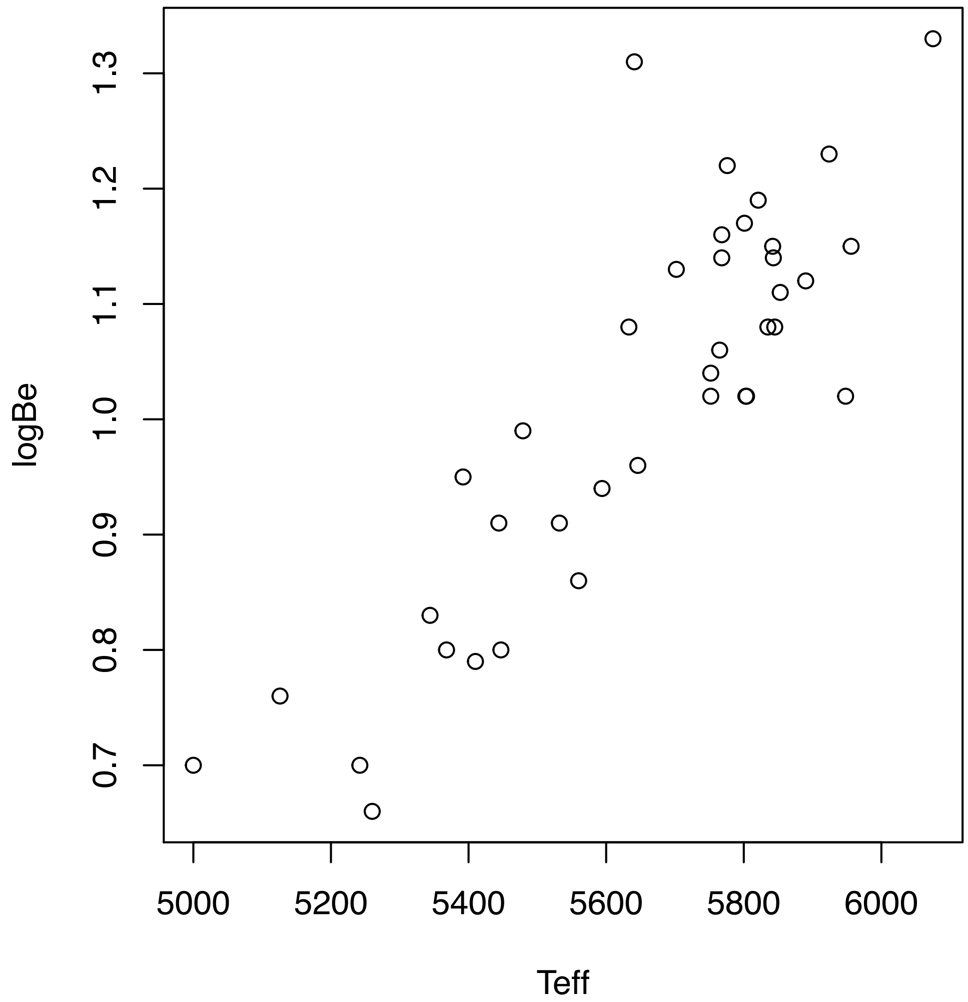
Karl Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Study relationship between two variables with data $(x_1, Y_1), \ldots, (x_n, Y_n)$.

Example: Log of beryllium abundance versus temperature of 38 stars (see [1]).



$r_{xy} = 0.862$

Pearson's correlation coefficient

For data pairs $(x_1, Y_1), \dots, (x_n, Y_n)$, the Pearson correlation coefficient is

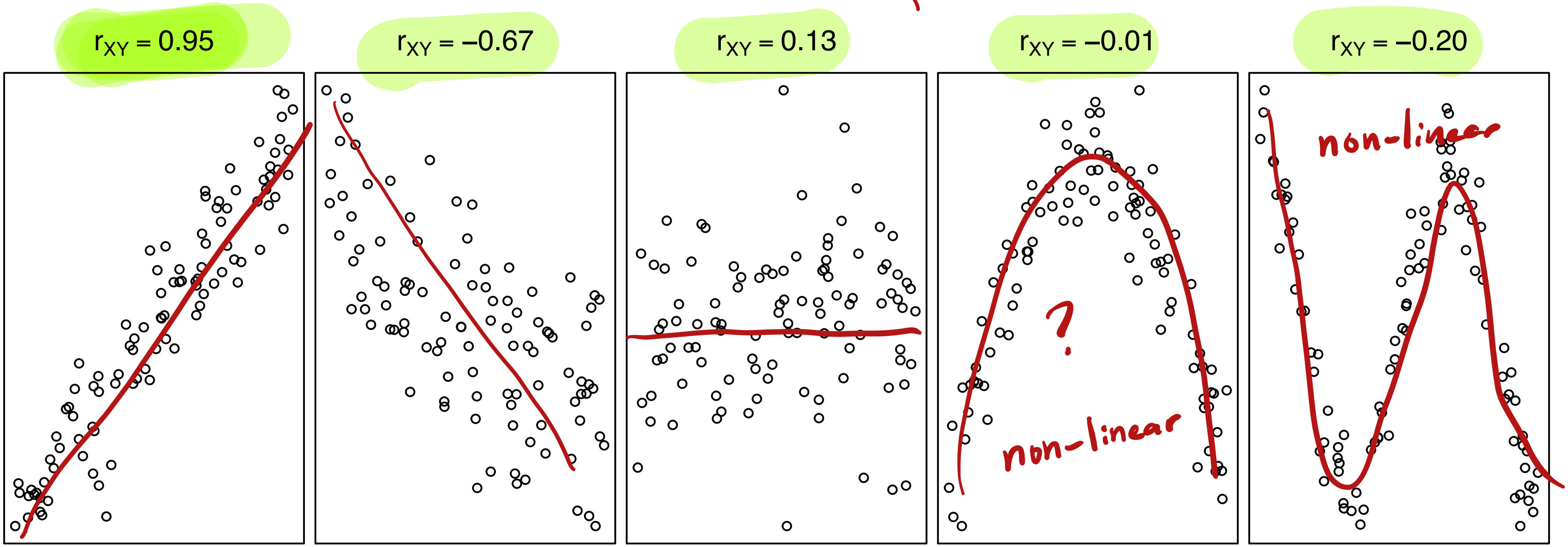
$$r_{xY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}.$$

- We have $r_{xY} \in [-1, 1]$.
- Values close to zero indicate weak linear relationship.
- Can use cor() function in R.

Summarize strength / direction of a linear relationship

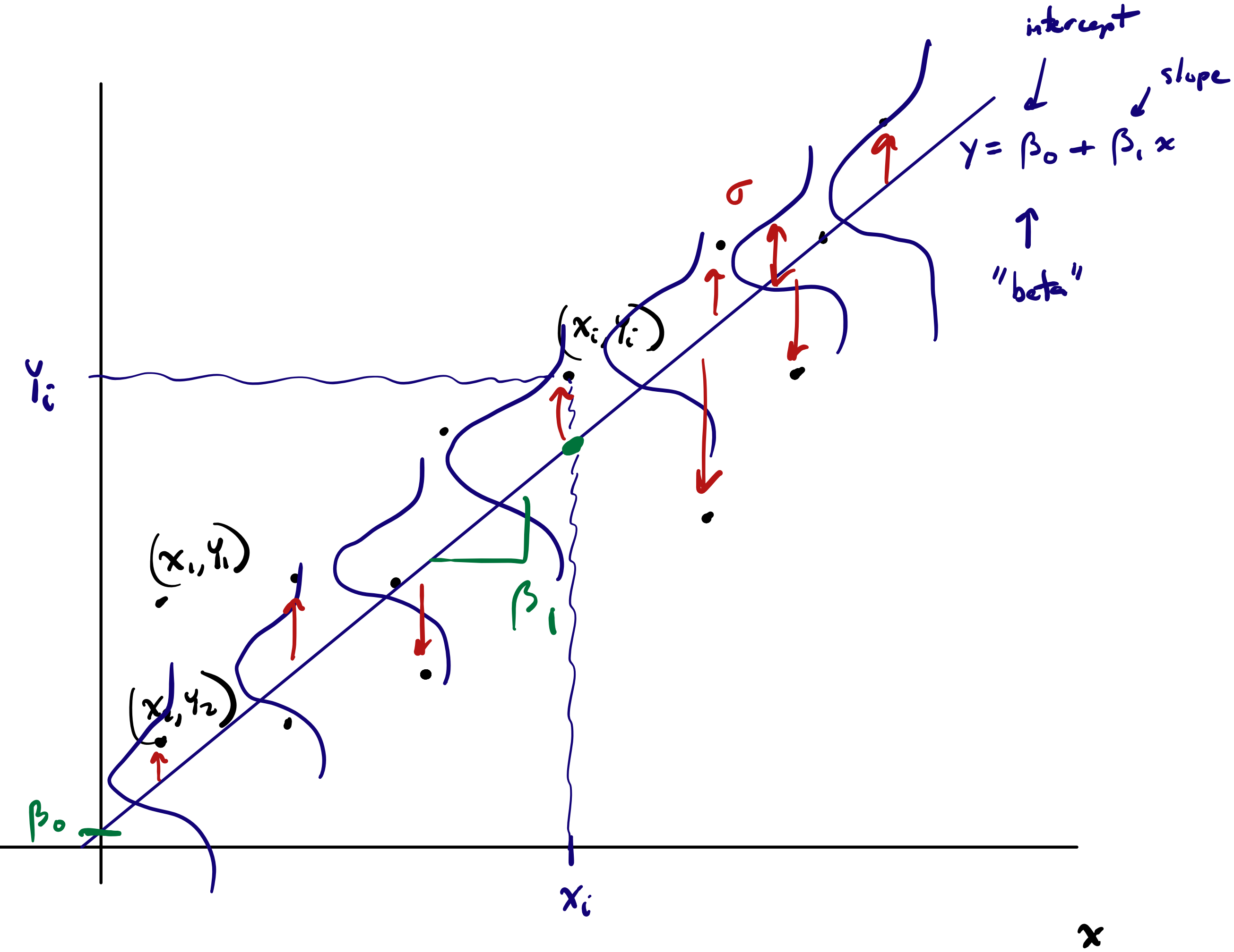
linear

r_{xy} : Pearson's correlation coefficient



Exercise: Compute Pearson's correlation coefficient on the beryllium data.

Simple Linear Regression Model (SLR)



$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

"response" variable

"covariate" "predictor" "explanatory variable" "regressor"

"epsilon" noise, error term, random

Simple linear regression model

For data pairs $(Y_1, x_1), \dots, (Y_n, x_n)$, suppose

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

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

- $x_1, \dots, x_n$ are fixed real numbers
- $Y_1, \dots, Y_n$ are independent random variables
- β_0 and β_1 are unknown constants
- $\varepsilon_1, \dots, \varepsilon_n$ are iid errors with
  - $\mathbb{E}\varepsilon_i = 0$
  - $\operatorname{Var} \varepsilon_i = \sigma^2$

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

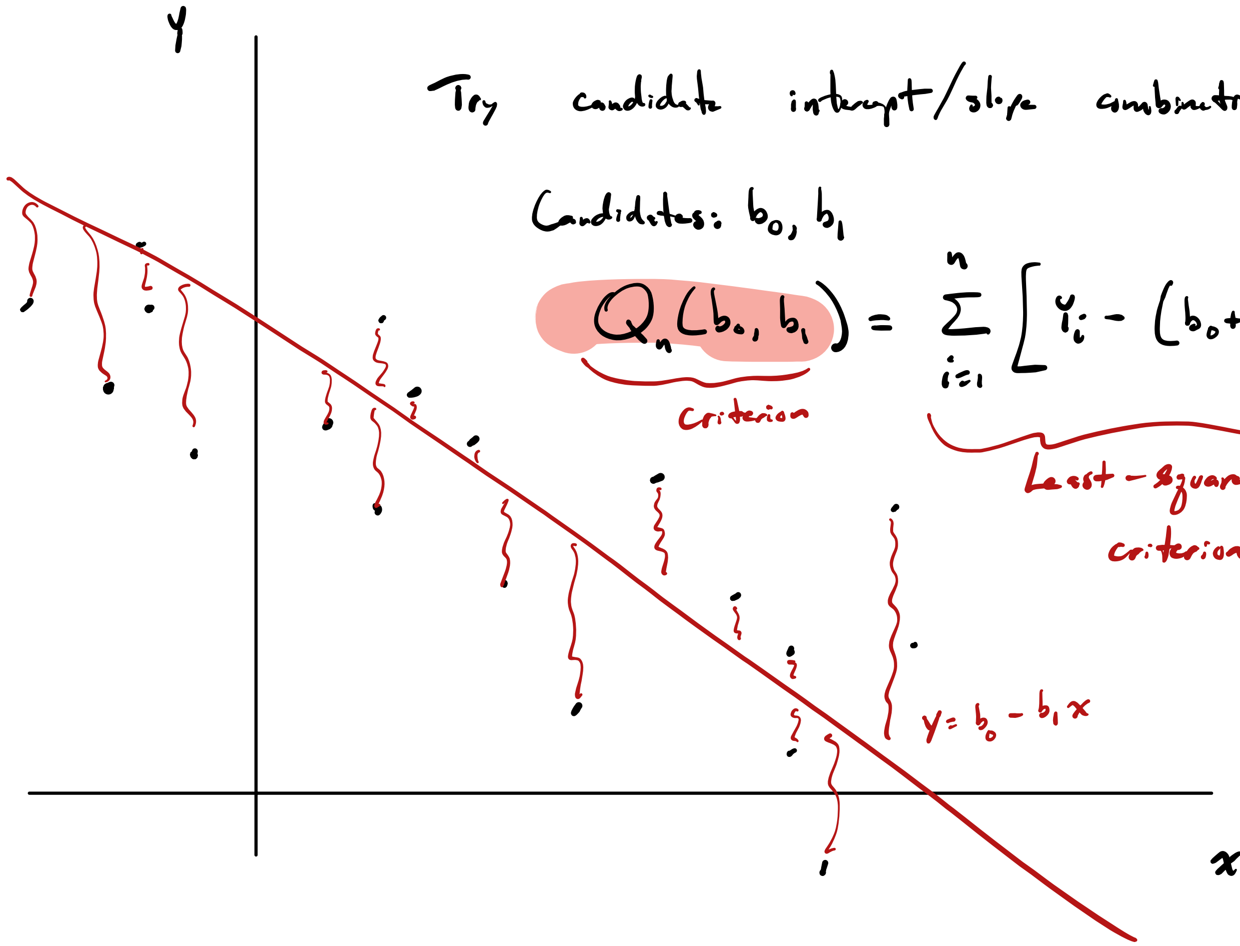
Goal: (1) Estimate the unknown constants β_0 and β_1 and the error variance σ^2.

(2) Make inferences about β_1 in particular (3) Make predictions about Y.

Try candidate intercept/slope combinations

Candidates: b_0, b_1

$$\underbrace{Q_n(b_0, b_1)}_{\text{criterion}} = \underbrace{\sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 x_i) \right]^2}_{\text{Least-squares criterion.}}$$



Least-squares estimators of simple linear regression coefficients

Provided $\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$, the function

$$Q_n(\beta_0, \beta_1) := \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 x_i)]^2$$

is (uniquely) minimized at

Pearson's correlation

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$$

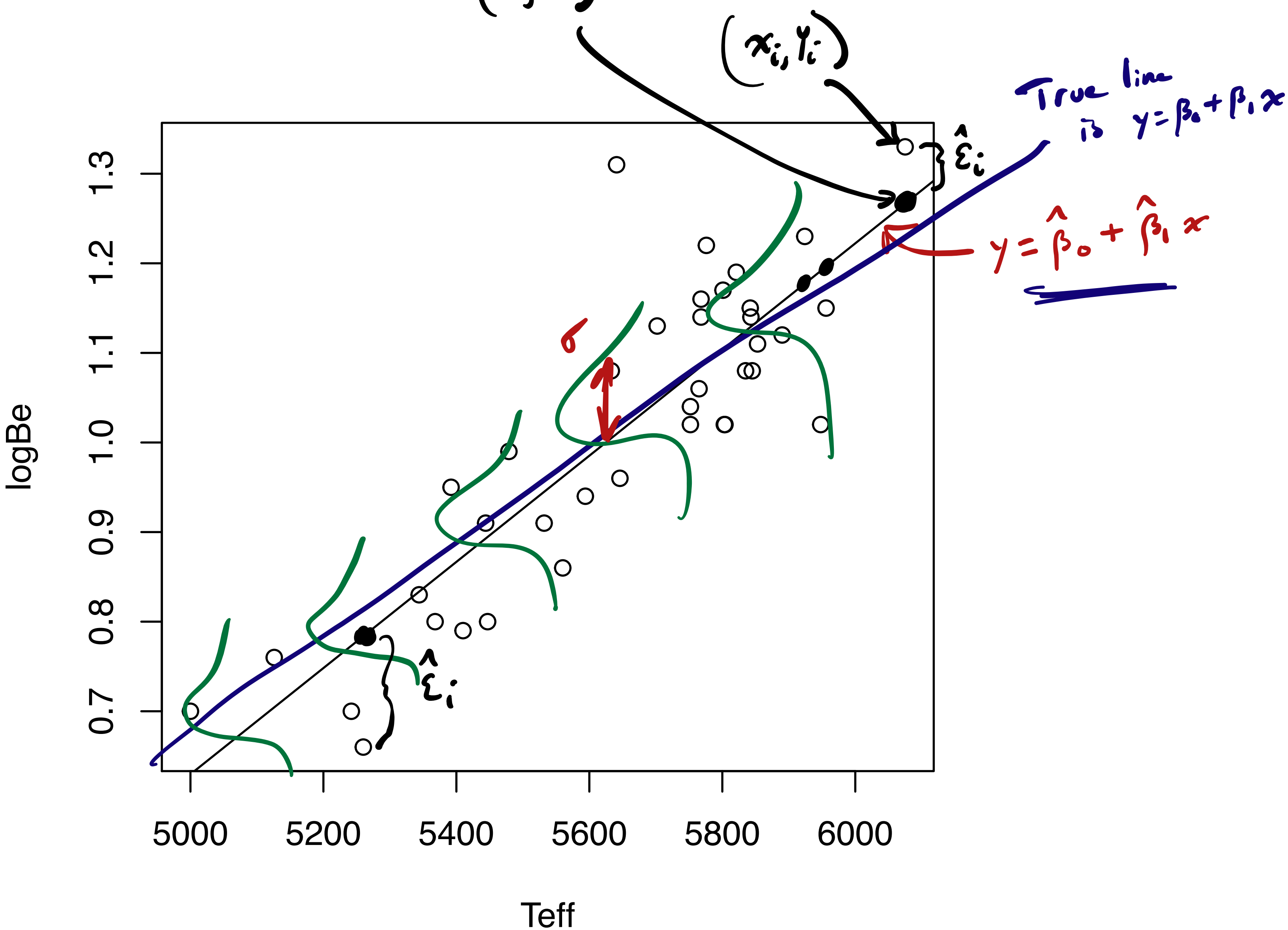
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = r_{xY} \cdot \frac{s_Y}{s_x}.$$

In above $s_Y^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ and $s_x^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$.

Exercise: Compute $\hat{\beta}_0$ and $\hat{\beta}_1$ for the beryllium data and plot the LS line.

```
# load the data
load(url("https://people.stat.sc.edu/gregorkb/data/beryllium.Rdata"))

# pull x and Y from the beryllium data frame
x <- beryllium$Teff
Y <- beryllium$logN_Be

# compute the least-squares regression coefficients
x_bar <- mean(x)
b1 <- cor(x,Y) * sd(Y) / sd(x)
b0 <- mean(Y) - b1*x_bar

# make a scatterplot with the least-squares line overlaid
plot(Y ~ x , xlab="Teff", ylab = "logBe")
abline(b0,b1)
```

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$



$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

- The *fitted values* are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{for } i = 1, \dots, n.$$

height of estimated line at observed x_i

- The *residuals* are

$$\hat{\varepsilon}_i = \underline{Y_i} - \underline{\hat{Y}_i} \quad \text{for } i = 1, \dots, n.$$

Our estimator of σ^2 will be $\hat{\sigma}^2 = \dfrac{1}{n-2} \displaystyle\sum_{i=1}^{n} \hat{\varepsilon}_i^2 . = \frac{1}{n-2} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}_n \right)^2$

$\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Draw pictures: Illustrate what the residuals and fitted values are.

Make inferences on β_1.

Sampling distribution of $\hat{\beta}_1$

Provided $\varepsilon_1, \dots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$, we have

$$\hat{\beta}_1 \sim \text{Normal}(\beta_1, \sigma^2/S_{xx}) \quad \text{and} \quad (n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$$

from which follows

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

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

A $(1-\alpha)100\%$ CI for β_1 is given by

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

Exercise: Build a 95% CI for β_1 for the beryllium data.

```
n <- length(Y)
Sxx <- sum((x - x_bar)^2)
sigma_hat <- sqrt( sum(e_hat^2)/(n-2))

lo <- b1 - qt(.975,n-2) * sigma_hat / sqrt(Sxx)
up <- b1 + qt(.975,n-2) * sigma_hat / sqrt(Sxx)

# easy way:
confint(lm(Y ~ x))
```

Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \dots, n$, where $\varepsilon_1, \dots, \varepsilon_n \stackrel{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$.

Tests about β_1

Define the test statistic

$$T_{\text{test}} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}}.$$

Then the following tests have $P(\text{Type I error}) \leq \alpha$.

$H_0\colon \beta_1 \geq 0$ $H_1\colon \beta_1 < 0$	$H_0\colon \beta_1 = 0$ $H_1\colon \beta_1 \neq 0$	$H_0\colon \beta_1 \leq 0$ $H_1\colon \beta_1 > 0$
Reject H_0 if $T_{\text{test}} < -t_{n-2,\alpha}$	Reject H_0 if $\lvert T_{\text{test}} \rvert > t_{n-2,\alpha/2}$	Reject H_0 if $T_{\text{test}} > t_{n-2,\alpha}$
$p\text{-val} = P(T < T_{\text{test}})$	$p\text{-val} = 2 \cdot P(T > \lvert T_{\text{test}} \rvert)$	$p\text{-val} = P(T > T_{\text{test}})$

Discuss: Draw pictures of how to get the p-values.

Exercise: Get the p-value for testing $H_0\colon \beta_1 = 0$ for the beryllium data.

Exercise: Use the `lm()`, `summary()`, the `confint()` functions in R to obtain

1. the least-squares estimators
2. the p-value for testing H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$
3. confidence intervals for β_0 and β_1

for the beryllium data.

Consider the assumptions of the model

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

where $\varepsilon_1, \dots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$.

(A.1) The responses are Normally distributed around the regression line.

To check: Look at a QQ plot of the residuals.

(A.2) The responses have the same variance for all values of the covariate.

To check: Look at the residuals versus fitted values plot.

(A.3) The covariate and the response are linearly related.

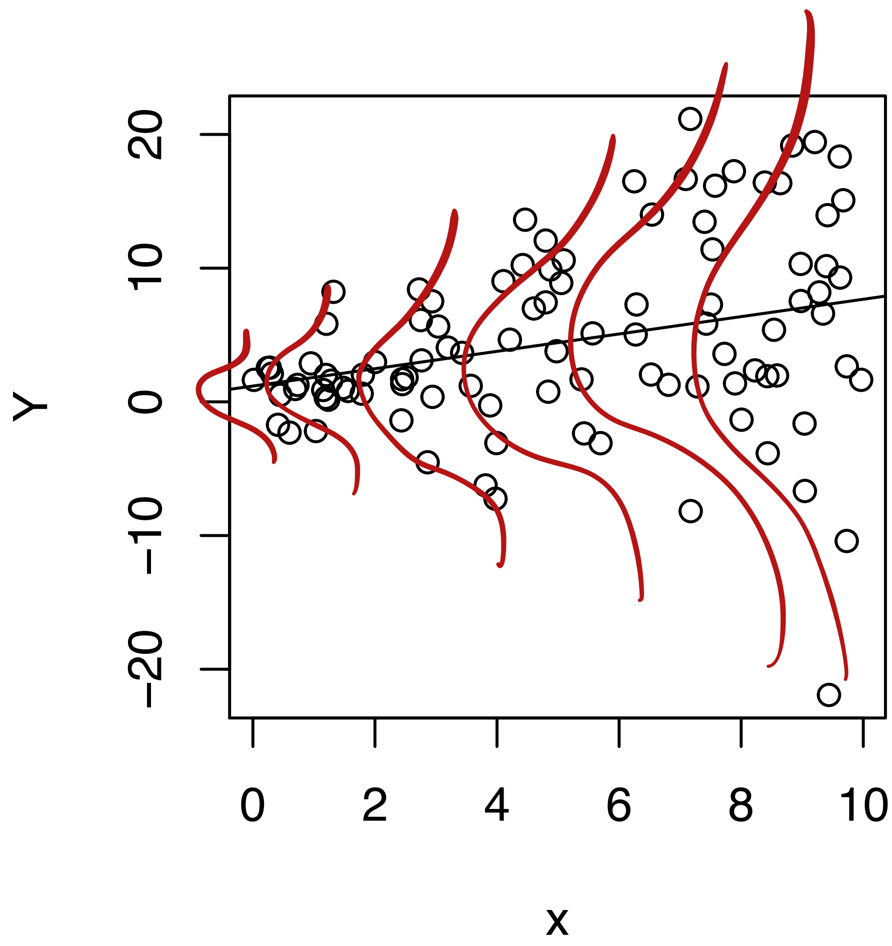
To check: Look at the residuals versus fitted values plot.

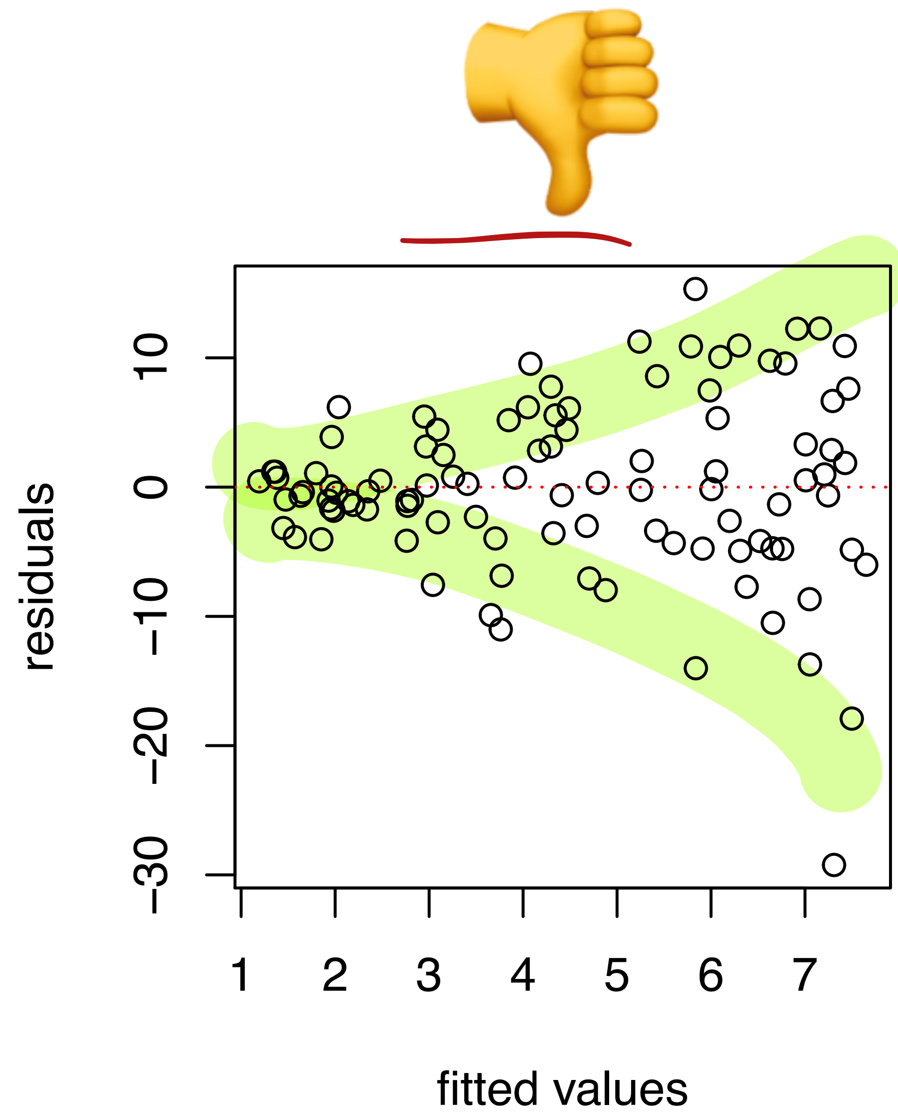
(A.4) The responses are independent from each other.

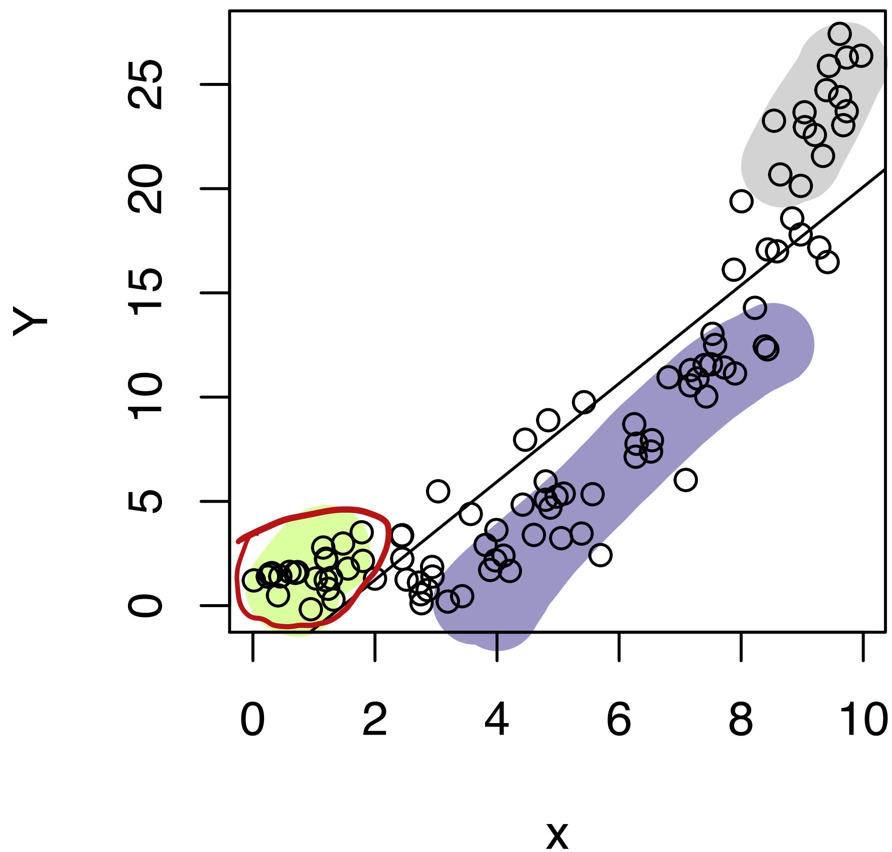
Cannot check: Trust the experimental design/beyond scope of course.

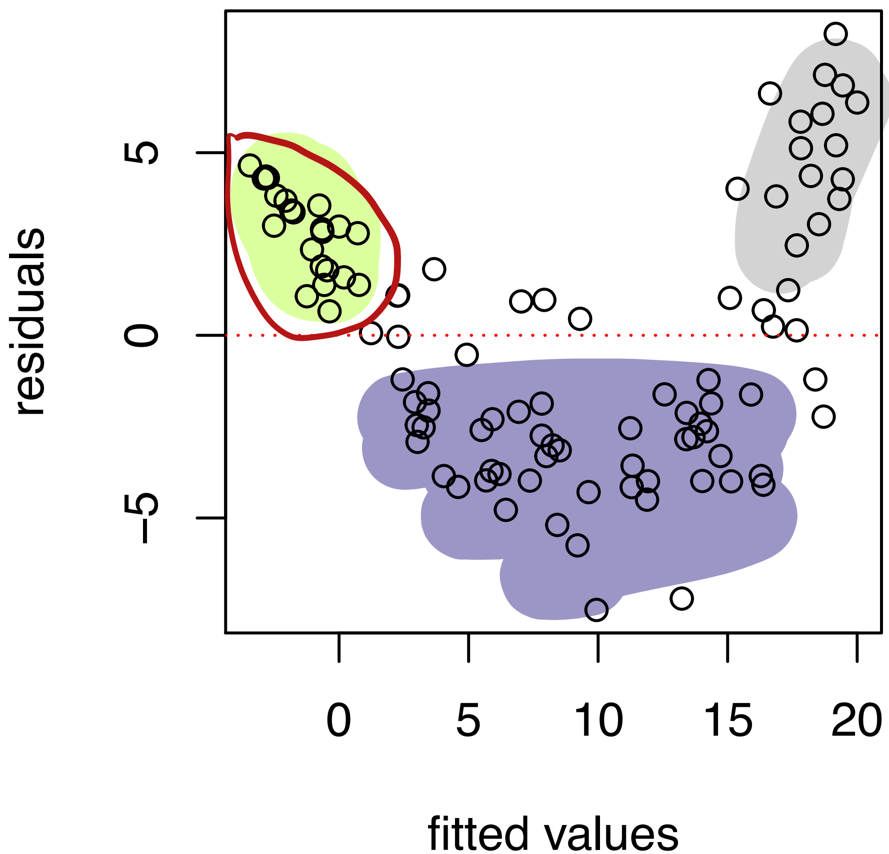
Use `plot()` on the output of `lm()`.

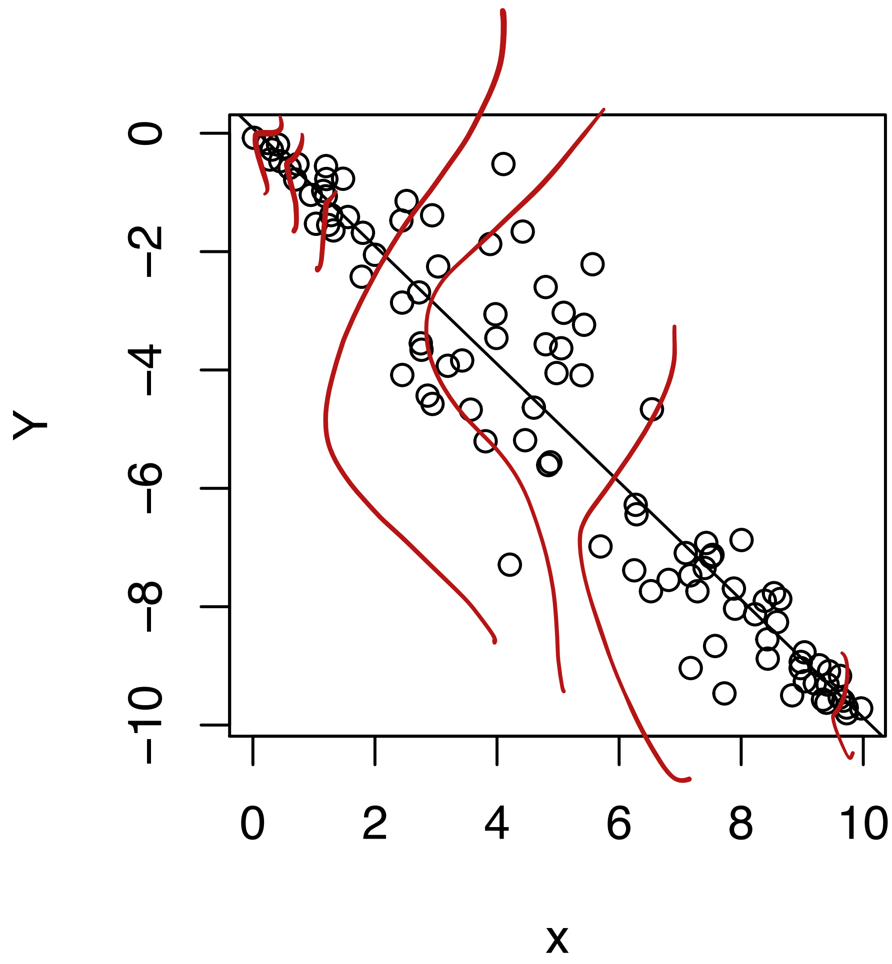
Exercise: Check the diagnostic plots for the beryllium data.

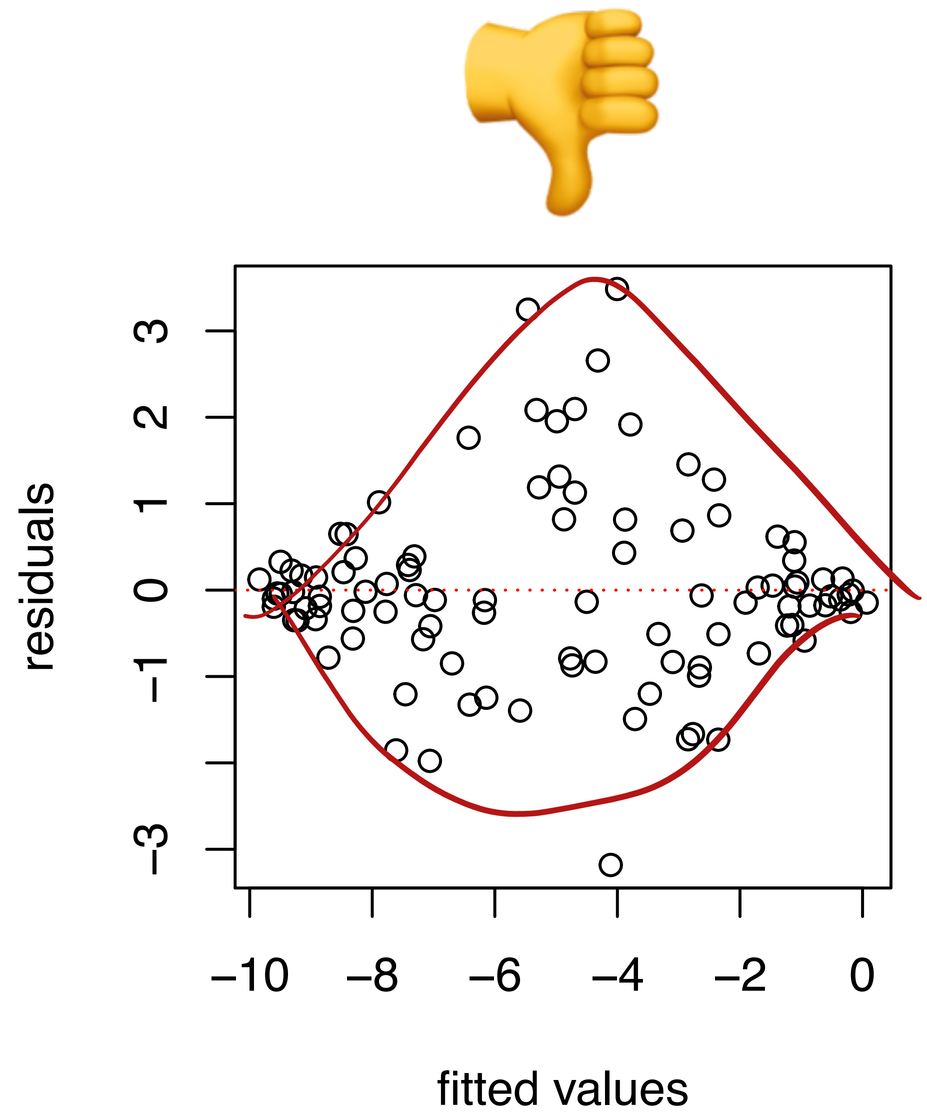

Y
x
20
10
0
−10
−20
0
2
4
6
8
10

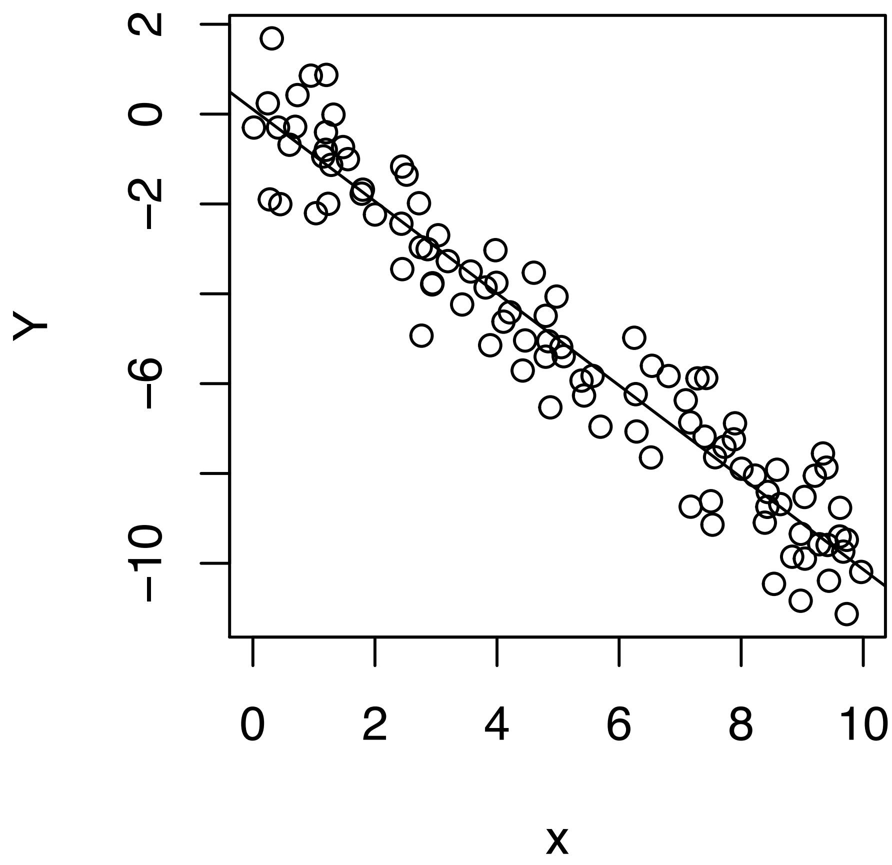

residuals
fitted values
10
0
−10
−20
−30
1
2
3
4
5
6
7

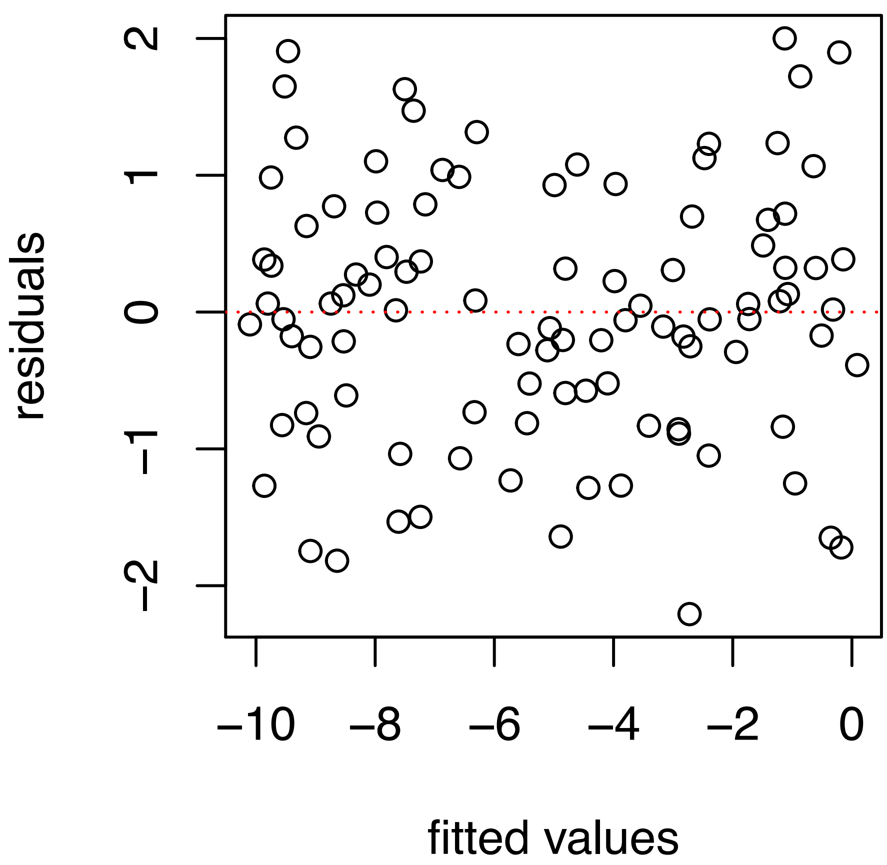

Y
x
0
2
4
6
8
10
0
5
10
15
20
25


residuals
fitted values
0
5
10
15
20
−5
0
5


Y
x
0
−2
−4
−6
−8
−10
0
2
4
6
8
10


residuals
fitted values
3
2
1
0
−1
−3
−10
−8
−6
−4
−2
0

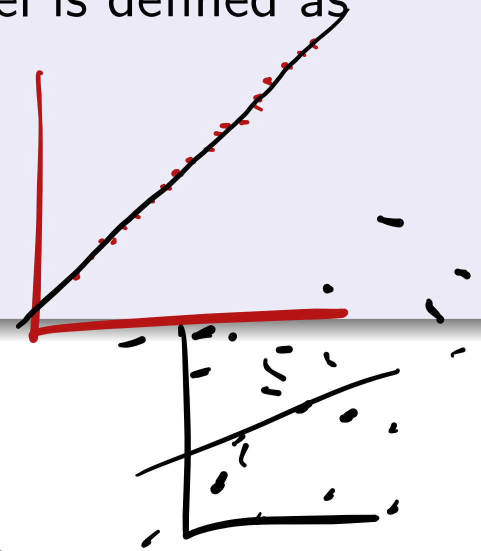

2
0
−2
−6
−10
Y
0
2
4
6
8
10
x


2
1
0
−1
−2
residuals
−10
−8
−6
−4
−2
0
fitted values

Coefficient of determination

The *coefficient of determination* for a linear regression model is defined as

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$$



In the above

$$SS_{\text{Regression}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2 \quad \text{and} \quad SS_{\text{Total}} = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

- $\boxed{R^2 \in [0, 1].}$
- R^2 is the proportion of variability in the response "explained" by the covariate.

Predicting the value of Y_{new} of the pair $(Y_{\text{new}}, x_{\text{new}})$.

- A $(1-\alpha)100\%$ confidence interval for $\beta_0 + \beta_1 x_{\text{new}}$ is given by

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

leverage

- A $(1-\alpha)100\%$ *prediction interval* for Y_{new} at x_{new} is given by

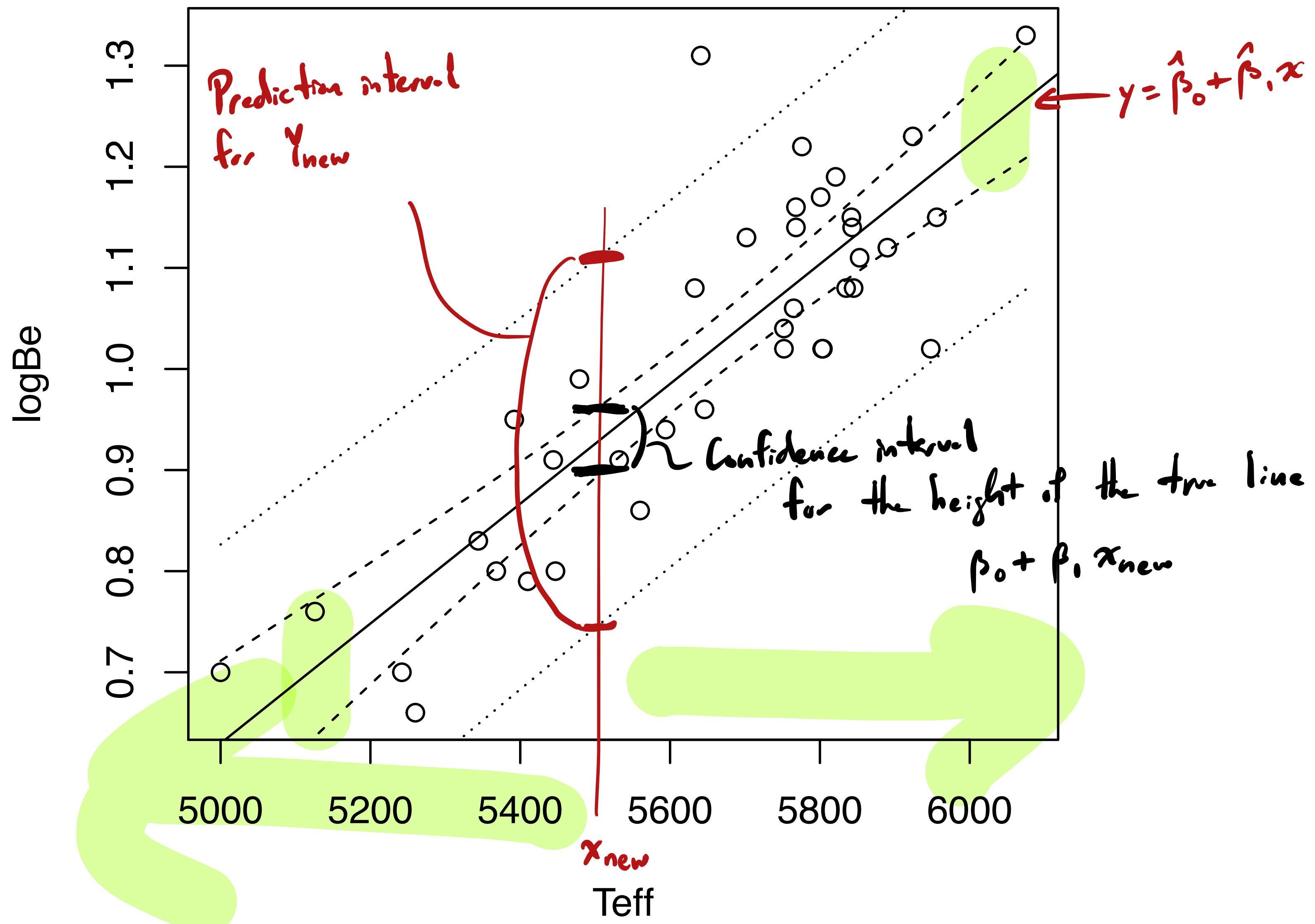
$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{n-1,\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}.$$

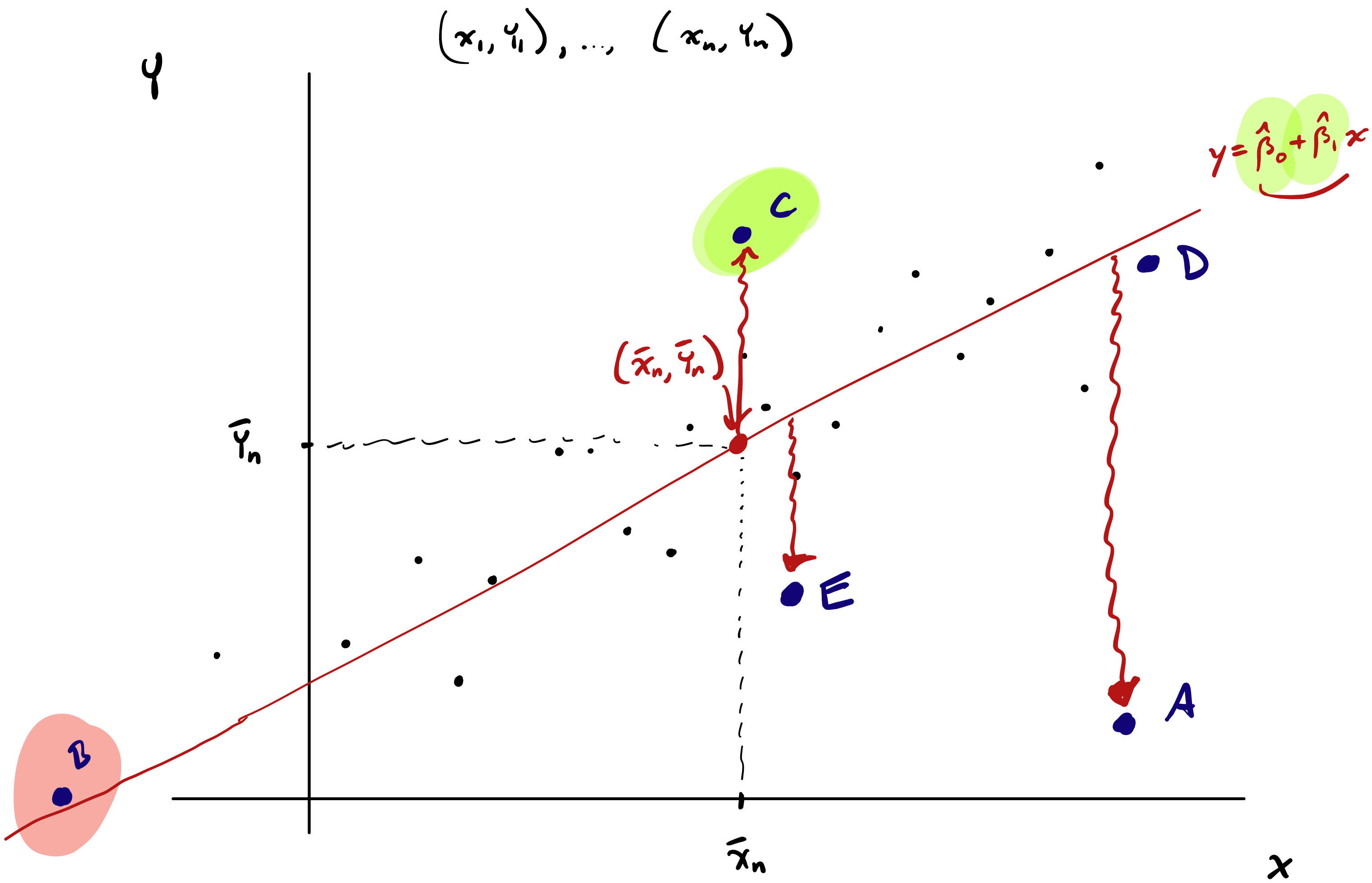
Exercise: Use `predict()` function on the `lm()` output to build a

1. CI for the mean `logBE` of stars with `Teff` equal to 5700.
2. PI for the `logBE` of a star with `Teff` equal to 5700.

```
# built-in way to obtain confidence or prediction intervals
lm.out <- lm(Y~x)
predict(lm.out, newdata = data.frame(x = 5700), interval = "confidence")
predict(lm.out, newdata = data.frame(x = 5700), interval = "prediction")
```

```
plot(Y ~ x , xlab="Teff",ylab = "logBe")
abline(beta0.hat,beta1.hat)

alpha <- .05
tval <- qt(1-alpha/2,n-2)

x.seq <- seq(min(x),max(x),length=99)
se.Y.hat.new <- sigma.hat * sqrt( 1/n + (x.seq - x.bar)^2/Sxx)
loconf <- beta0.hat+beta1.hat*x.seq - tval * se.Y.hat.new
upconf <- beta0.hat+beta1.hat*x.seq + tval * se.Y.hat.new

lines(loconf~x.seq,lty=2)
lines(upconf~x.seq,lty=2)

sd.e.hat.new <- sigma.hat *sqrt(1 + 1/n + (x.seq - x.bar)^2/Sxx)
lopred <- beta0.hat + beta1.hat * x.seq - tval * sd.e.hat.new
uppred <- beta0.hat + beta1.hat * x.seq + tval * sd.e.hat.new

lines(lopred~x.seq,lty=3)
lines(uppred~x.seq,lty=3)
```


Prediction interval
for Y_{new}
$y = \hat\beta_0 + \hat\beta_1 x$
Confidence interval
for the height of the true line
$\beta_0 + \beta_1 x_{new}$
logBe
0.7
0.8
0.9
1.0
1.1
1.2
1.3
5000
5200
5400
5600
5800
6000
x_{new}
Teff



leverage: $\quad \underline{\mathrm{lev}_i} = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2}$

Consider the effects of outliers on the estimated regression function.

Points can be outlying in x or Y direction.

Leverage

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

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

Points with high leverage have a large influence on the fitted regression line.

Consider fact: Least-squares line passes through the point $(\bar{x}_n, \bar{Y}_n)$.

Draw pictures.

$\boxed{\textit{Extrapolation}}$ is making predictions beyond the range of the observed data.

Specifically, if x_{new} is outside the range of observed $x_1, \dots, x_n$, it is extrapolation to build a CI for the expected value of Y_{new} or a PI for the realized value of Y_{new}.

A linear relationship may not continue beyond the range of observed $x_1, \dots, x_n$.

Nuno C Santos, G Israelian, RJ García López, M Mayor, R Rebolo, S Randich, A Ecuvillon, and C Domínguez Cerdeña.
Are beryllium abundances anomalous in stars with giant planets?
Astronomy & Astrophysics, 427(3):1085–1096, 2004.