STAT 515 fa 2023 Lec 18 slides

Simple linear regression

Multiple linear regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \varepsilon_i$$

Karl B. Gregory

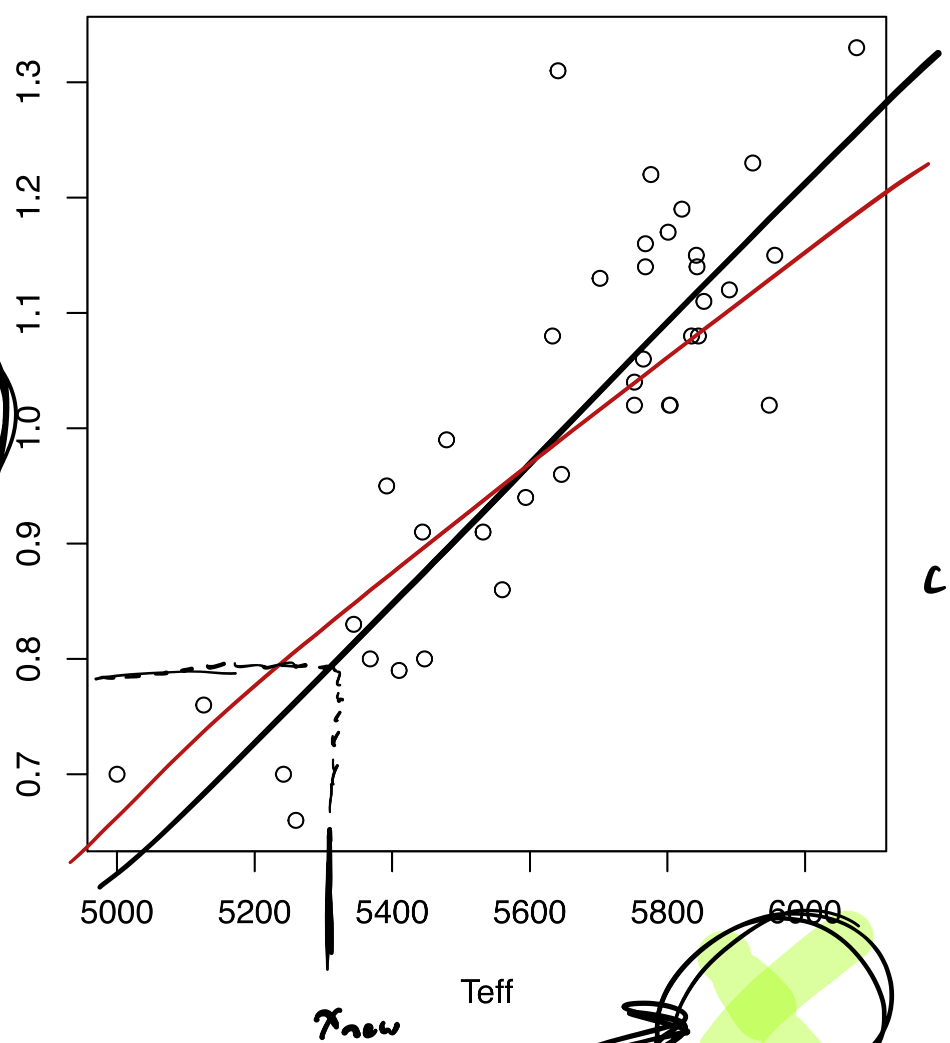
University of South Carolina

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

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]).



$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

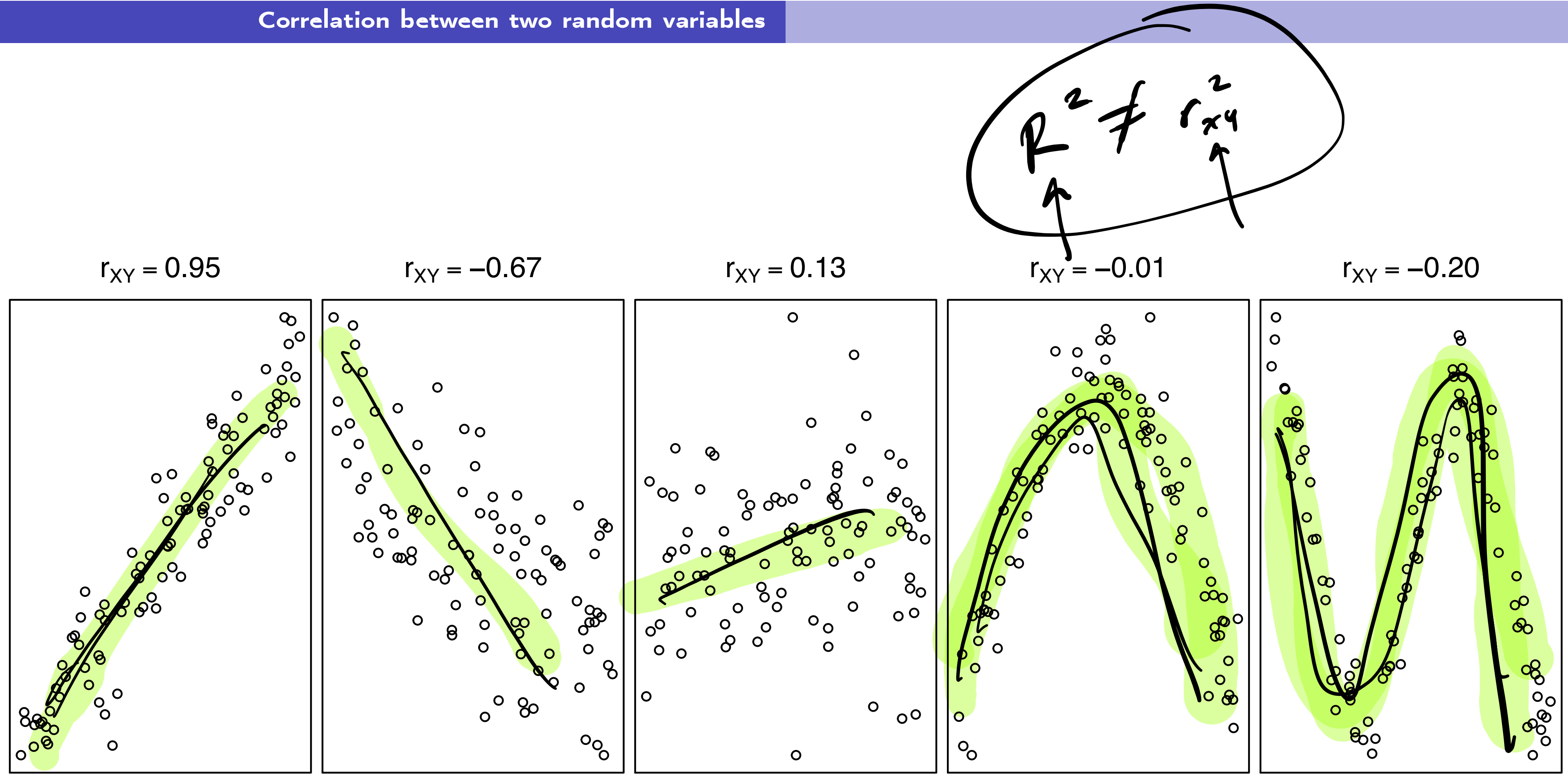
C.I. for β_1

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 indicates weak linear relationship.
- Can use `cor()` function in R.



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

Simple linear regression model

$\beta_0, \beta_1, \sigma^2$

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

noise

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

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

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

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

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

Least-squares estimators of simple linear regression coefficients

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

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

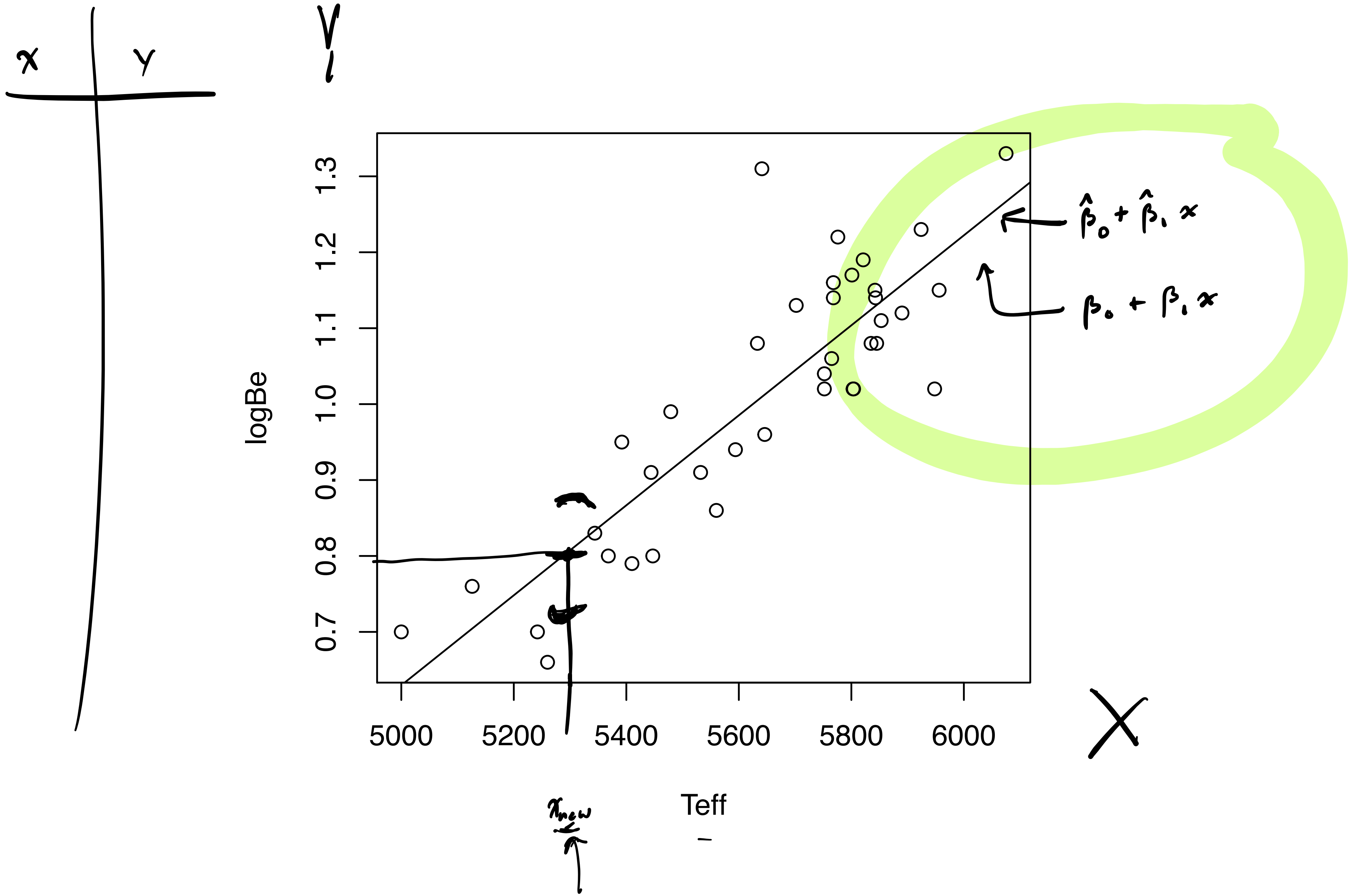
is (uniquely) minimized at

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = r_{xY} \cdot \frac{s_Y}{s_x}.$$

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)
```


x
Y
Y
logBe
1.3
1.2
1.1
1.0
0.9
0.8
0.7
5000
5200
5400
5600
5800
6000
Teff
X
$\hat{\beta}_0 + \hat{\beta}_1 x$
$\beta_0 + \beta_1 x$
x_{new}

- The *fitted values* are

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

- The *residuals* are

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

Our estimator of σ^2 will be $\hat{\sigma}^2 = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n} \hat{\varepsilon}_i^2$.

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

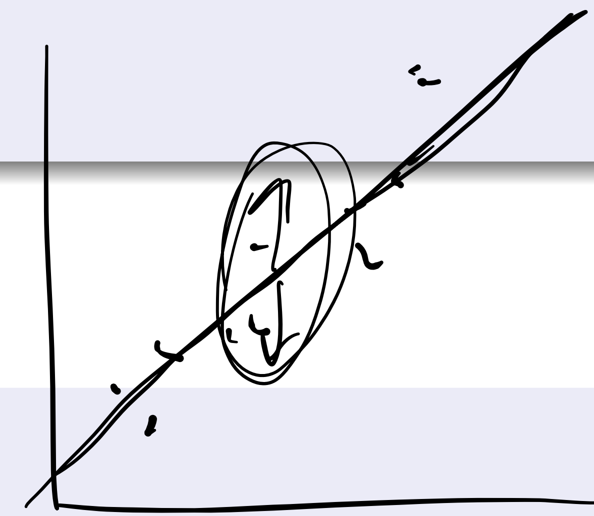
Sampling distribution of $\hat{\beta}_1$

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.

99%

```
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))
```

On acorn data: Length = $\beta_0 + \beta_1$ Diam $+ \varepsilon$ ($\beta_1 \overset{?}{=} 0$)

95% C.I. for β_1 is $(-0.032, 0.891)$

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

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 of β_0 and β_1.
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

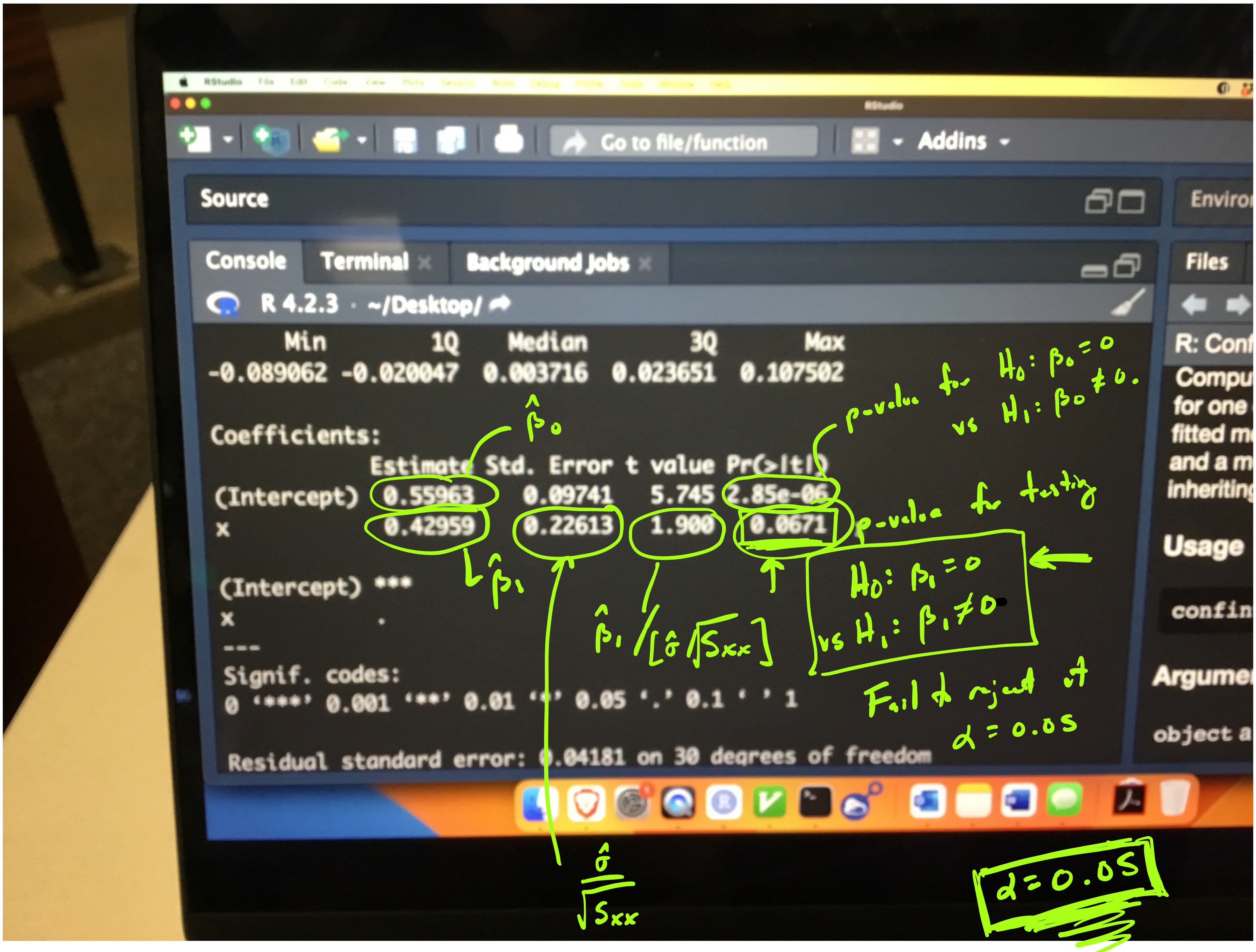
acorn

for the [redacted]n data.

Estimated model:

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

Length = 0.5596 + 0.4296 Diam


Go to file/function
Addins
Source
Console
Terminal
Background Jobs
R 4.2.3 · ~/Desktop/
Min 1Q Median 3Q Max
-0.089062 -0.020047 0.003716 0.023651 0.107502
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.55963 0.09741 5.745 2.85e-06
x 0.42959 0.22613 1.900 0.0671
(Intercept) ***
x .

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04181 on 30 degrees of freedom
$\hat\beta_0$
$\hat\beta_1$
p-value for $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$.
p-value for testing
$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$
$\hat\beta_1 / [\hat\sigma / \sqrt{S_{xx}}]$
Fail to reject at $\alpha = 0.05$
$\hat\sigma / \sqrt{S_{xx}}$
$\alpha = 0.05$
Environ
Files
R: Conf
Compu
for one
fitted m
and a m
inheriting
Usage
confin
Argume
object a

Consider the assumptions of the model

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

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()`.

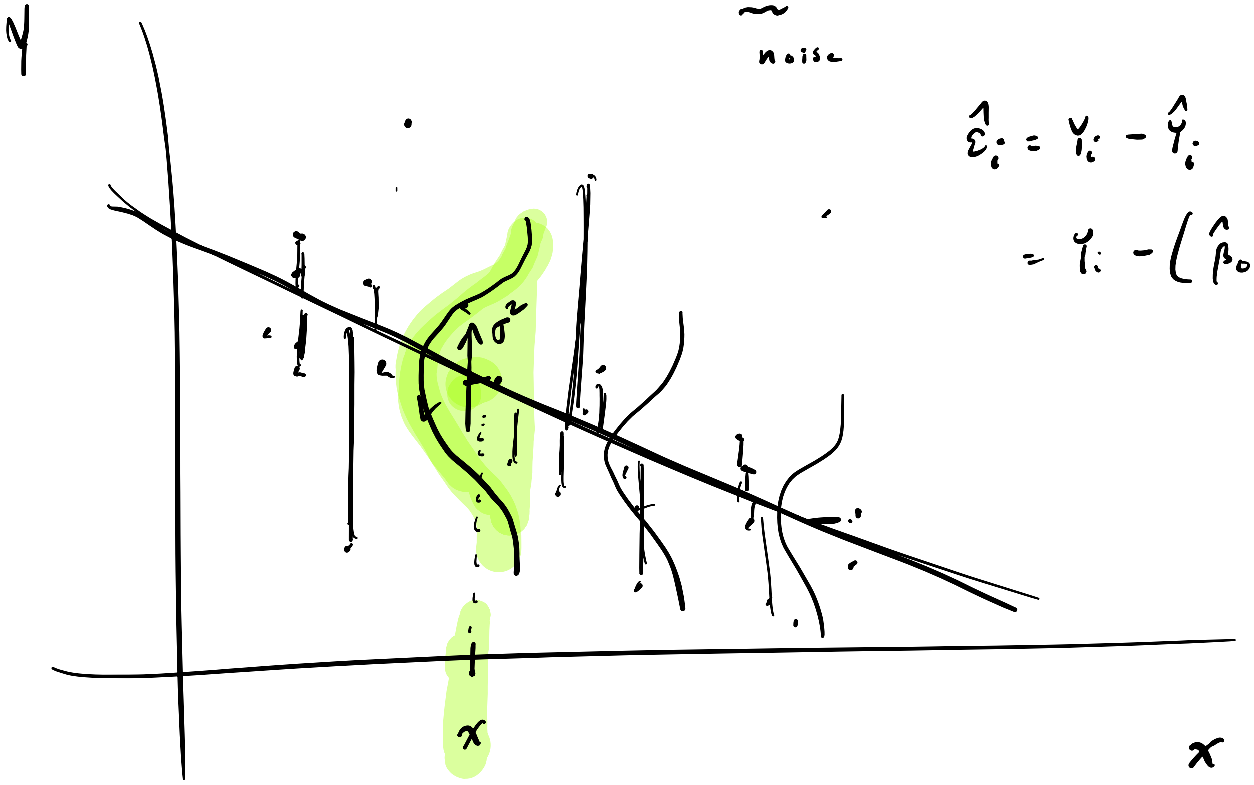
Exercise: Check the diagnostic plots for the beryllium data.

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \overset{\text{independent}}{\sim} N(0, \sigma^2)$

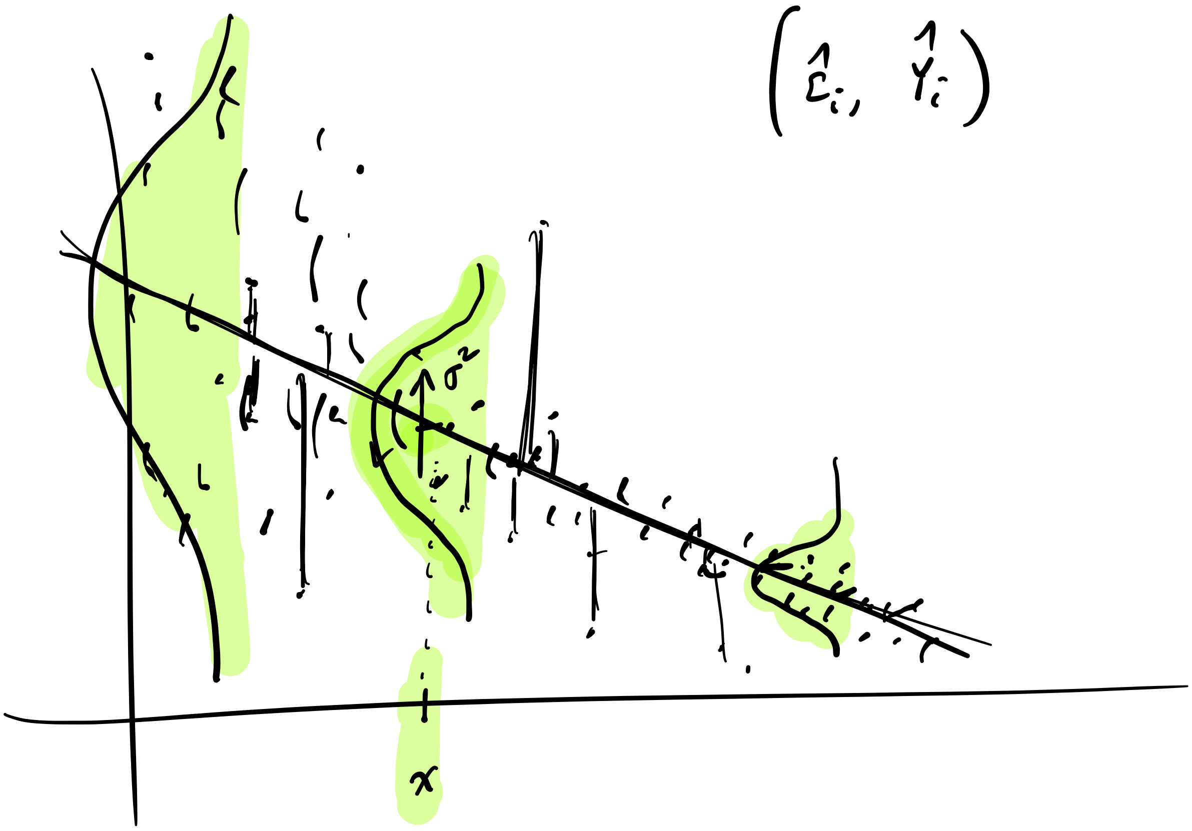
(ε_i: noise)

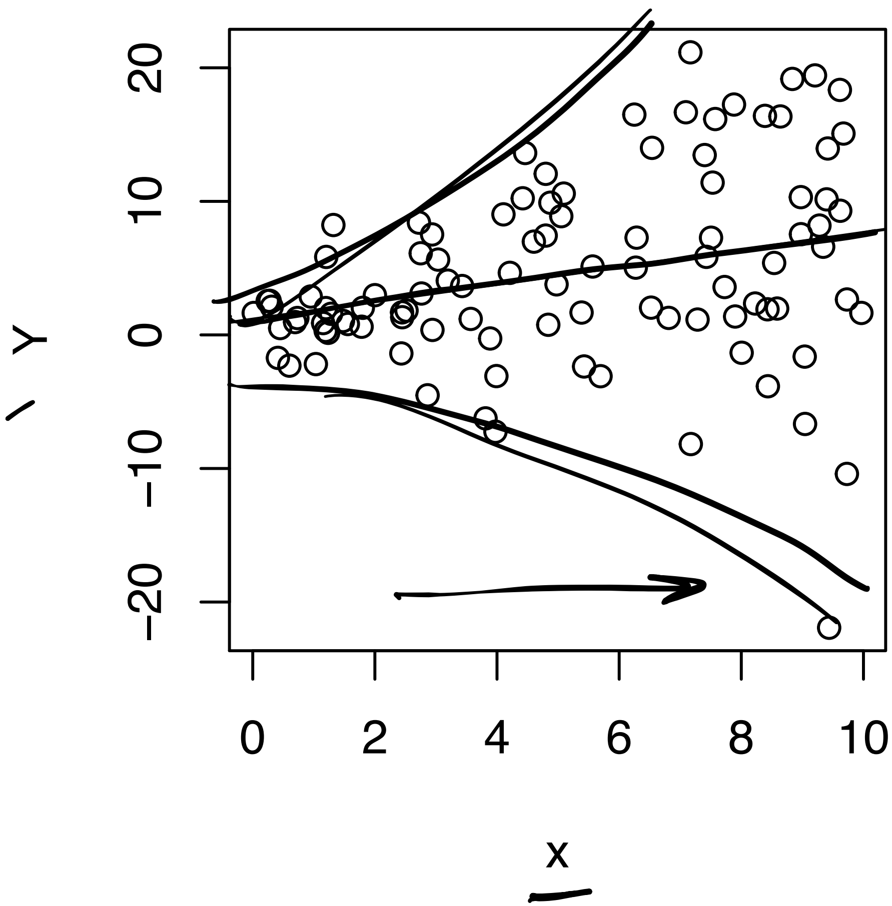
$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

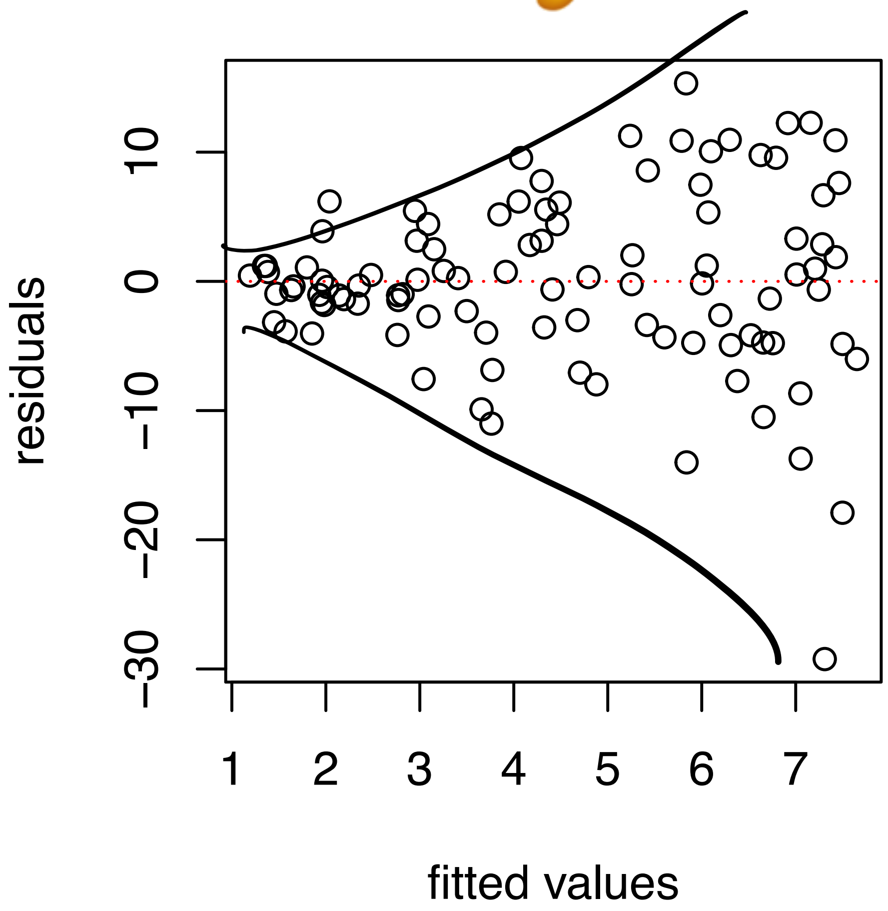
$= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$



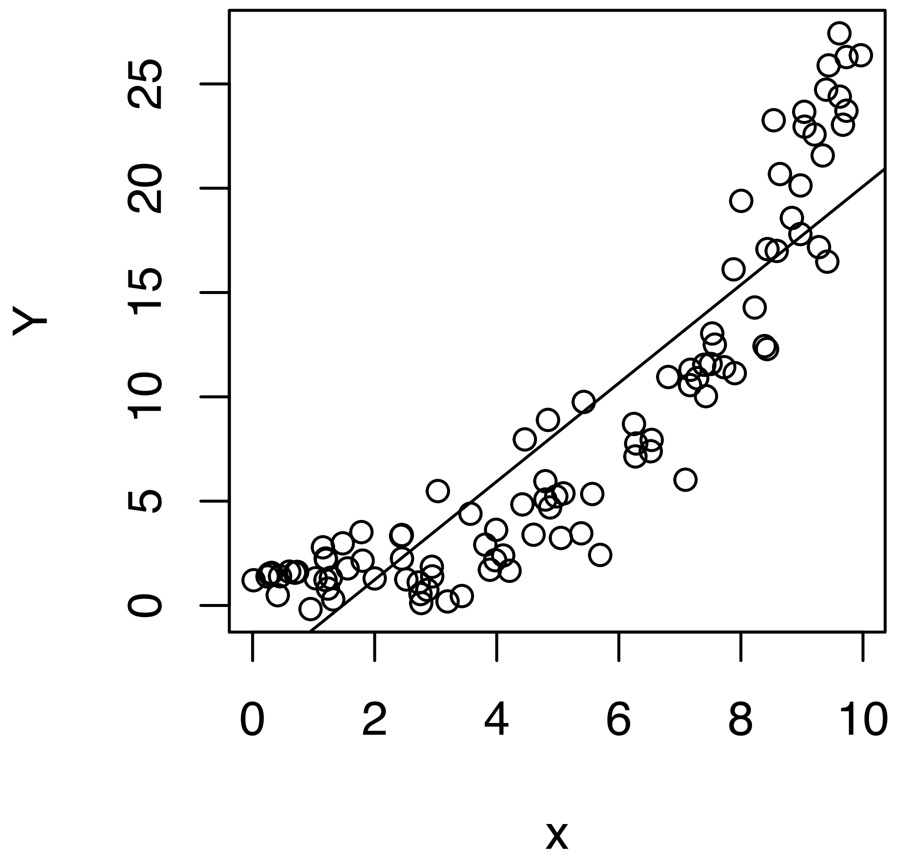
$(\hat{\varepsilon}_i, \hat{Y}_i)$

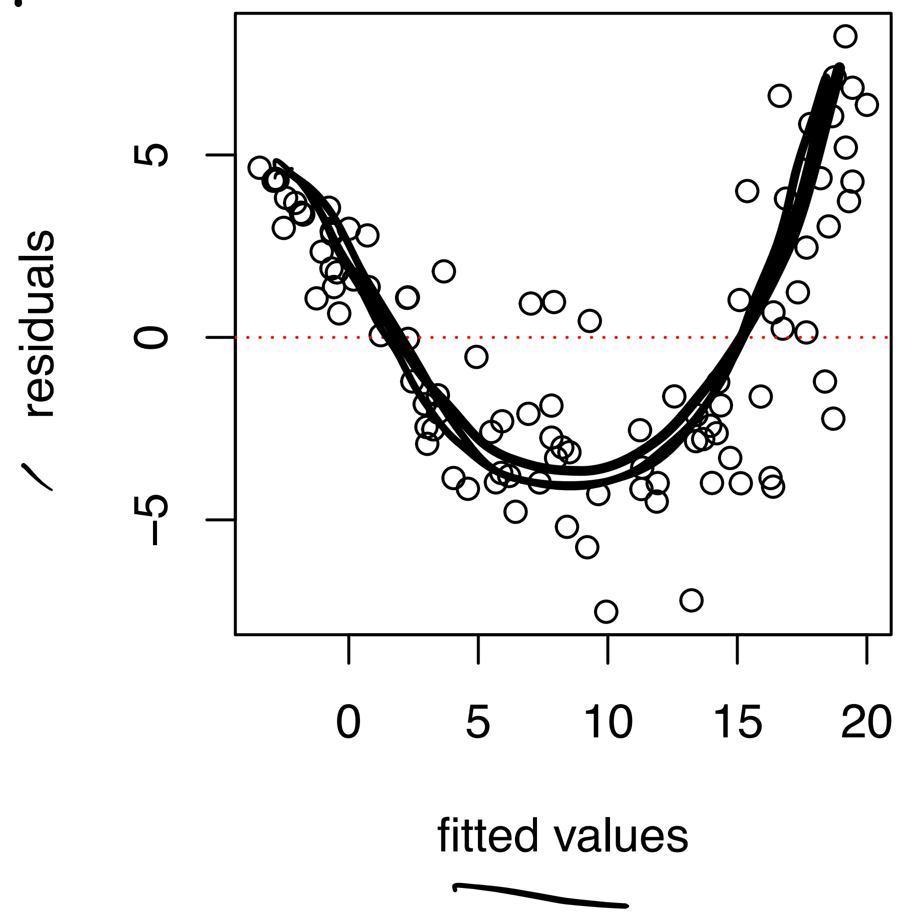


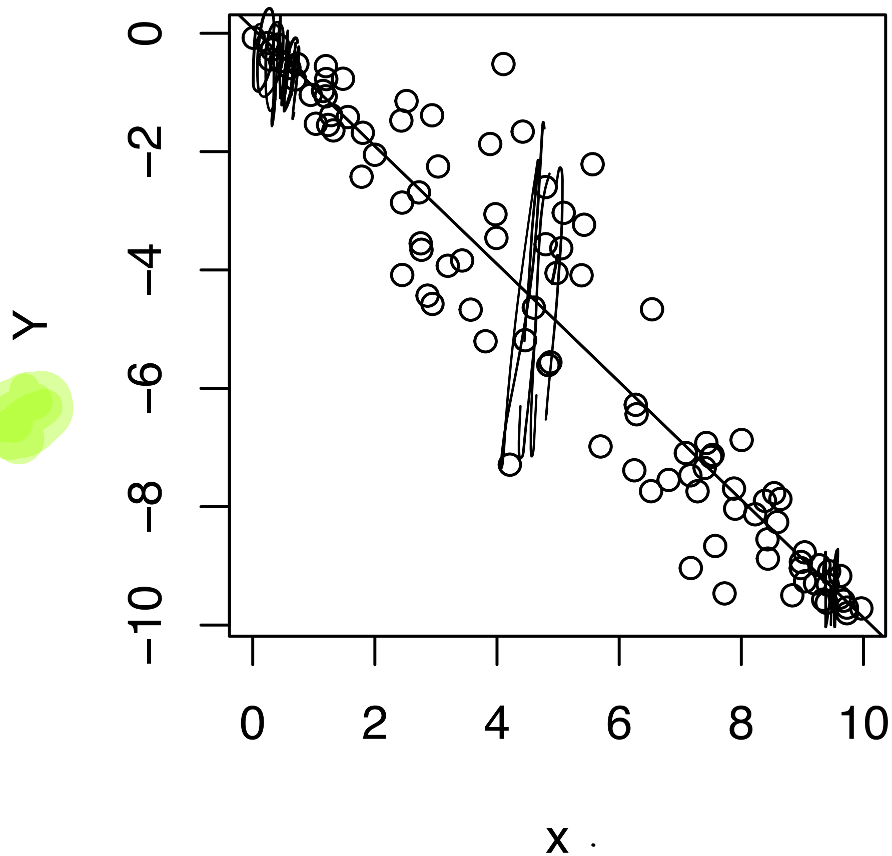



residuals
fitted values
10
0
−10
−20
−30
1
2
3
4
5
6
7

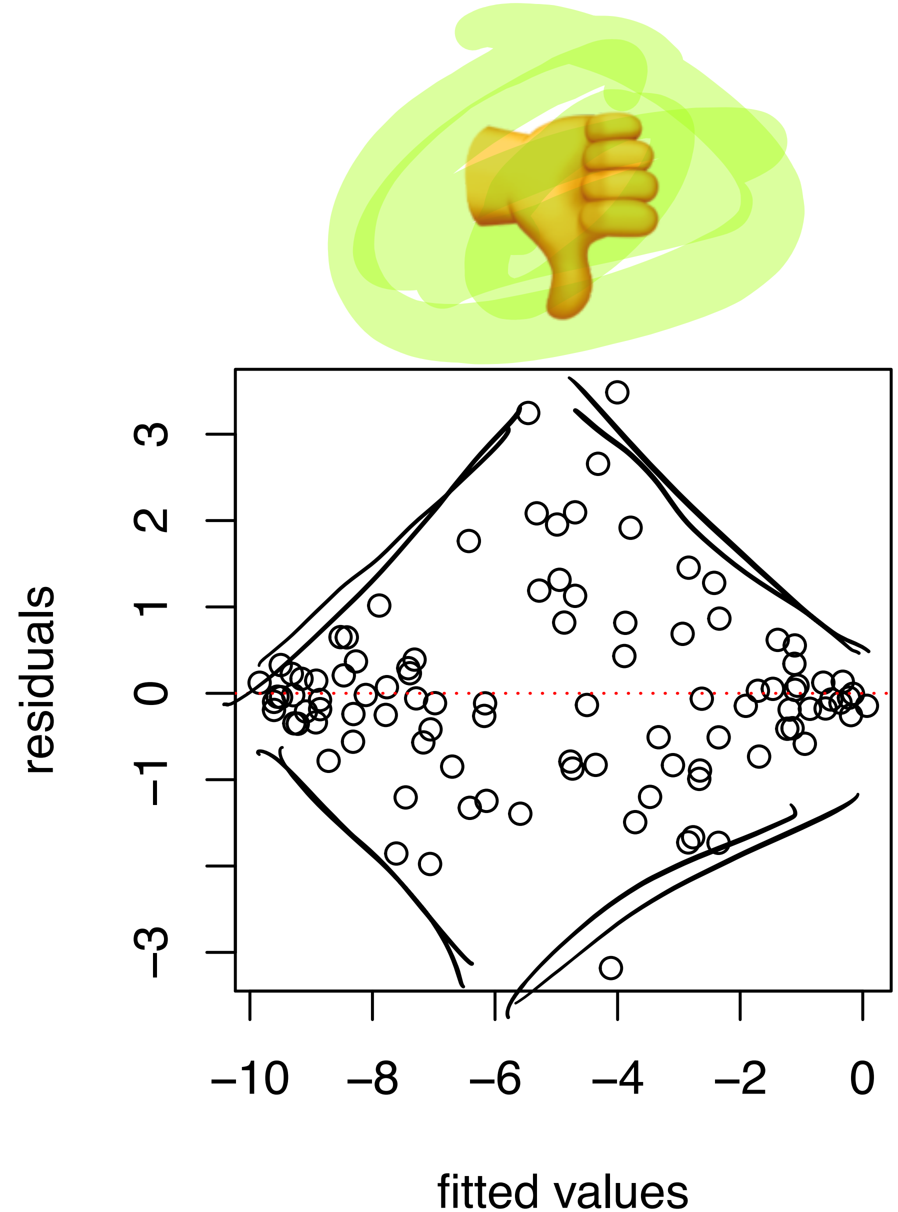
relationship is non linear.



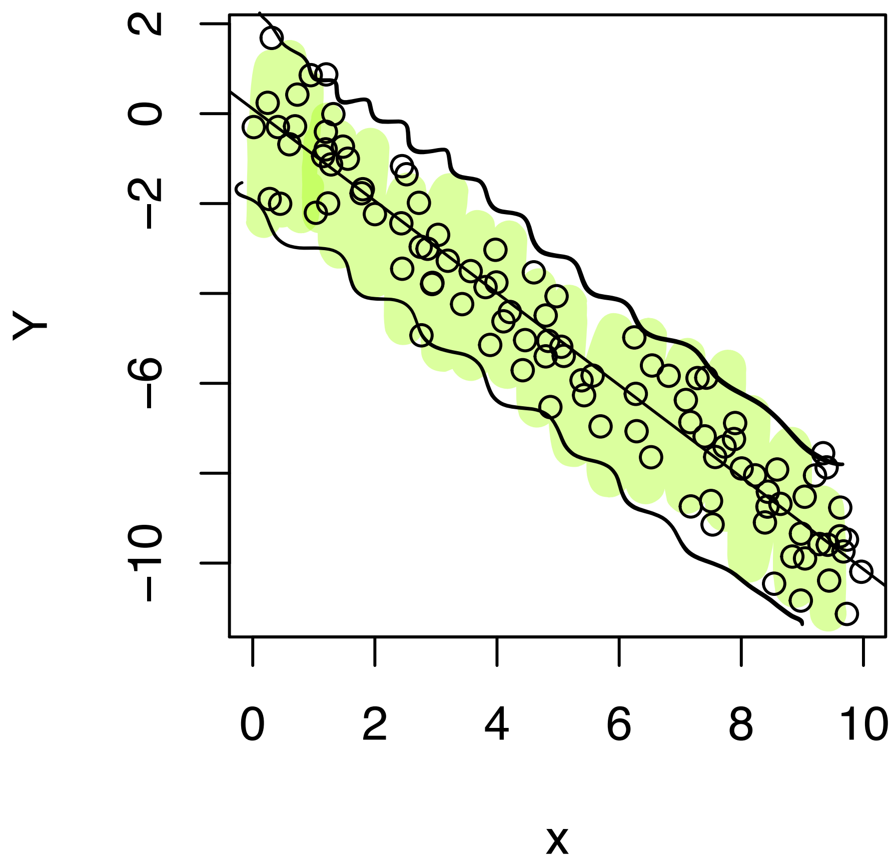


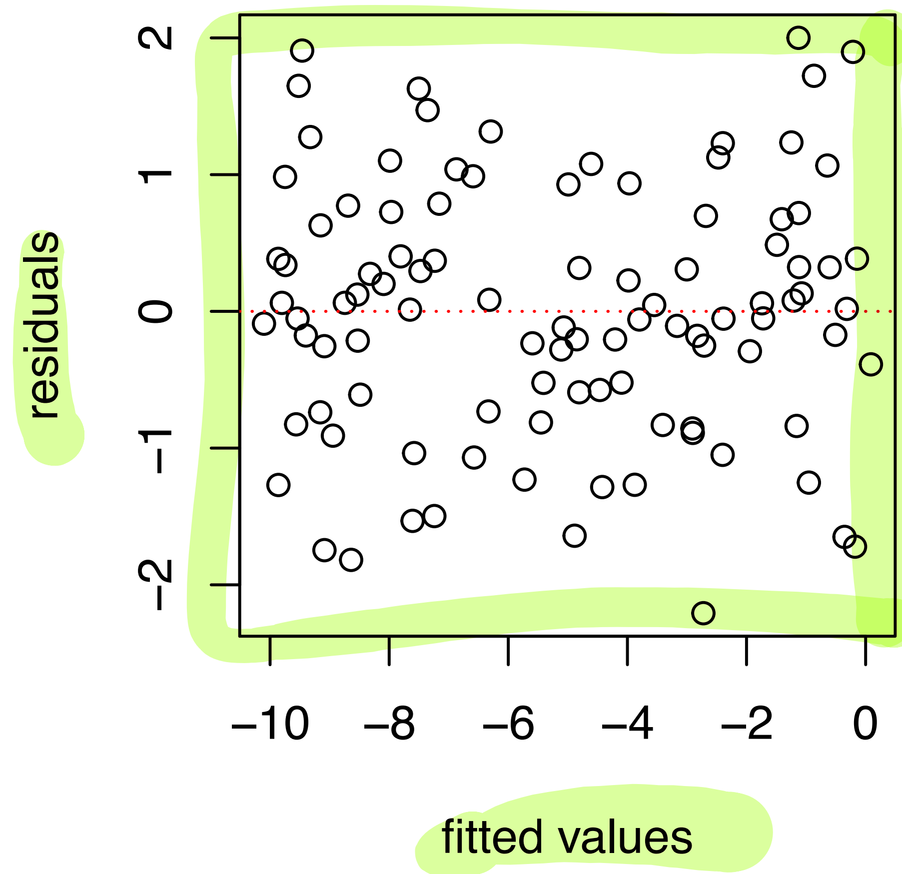


Y
0
−2
−4
−6
−8
−10
0
2
4
6
8
10
x



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


Y
2
0
−2
−6
−10
0
2
4
6
8
10
x


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

Coefficient of determination

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

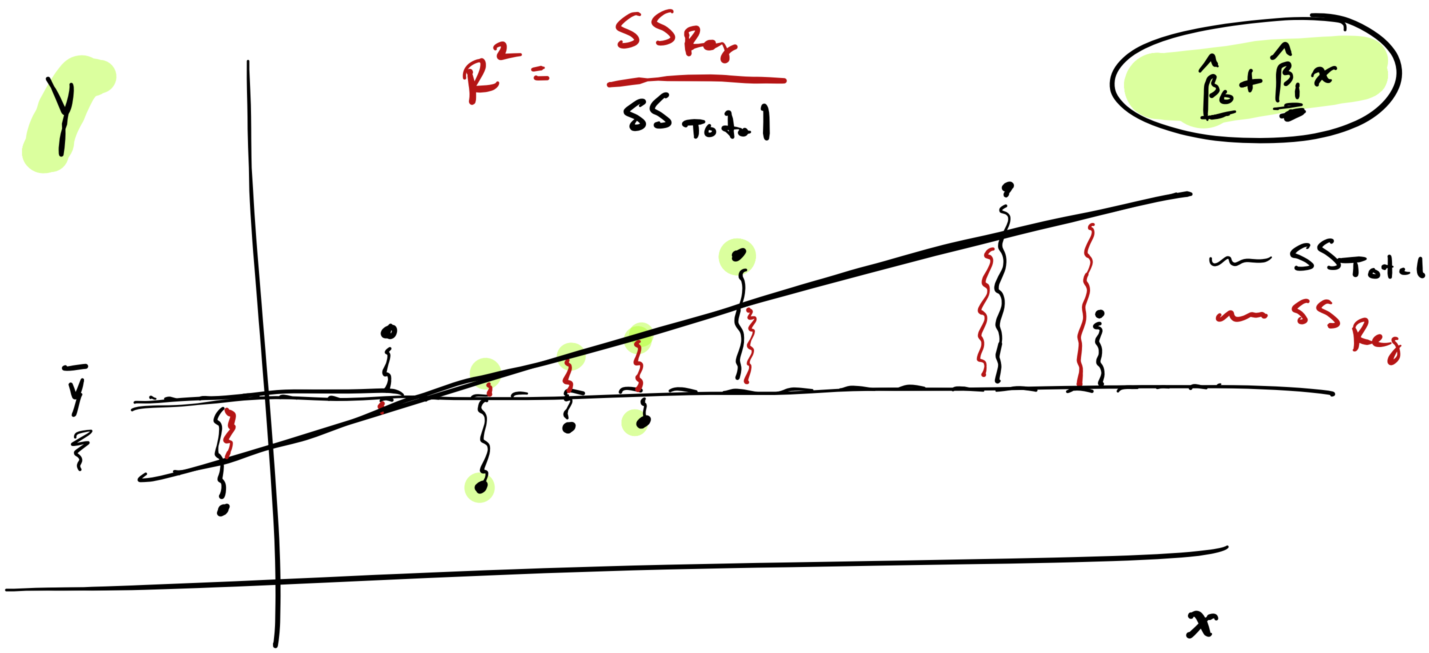
$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$$

In the above

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

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

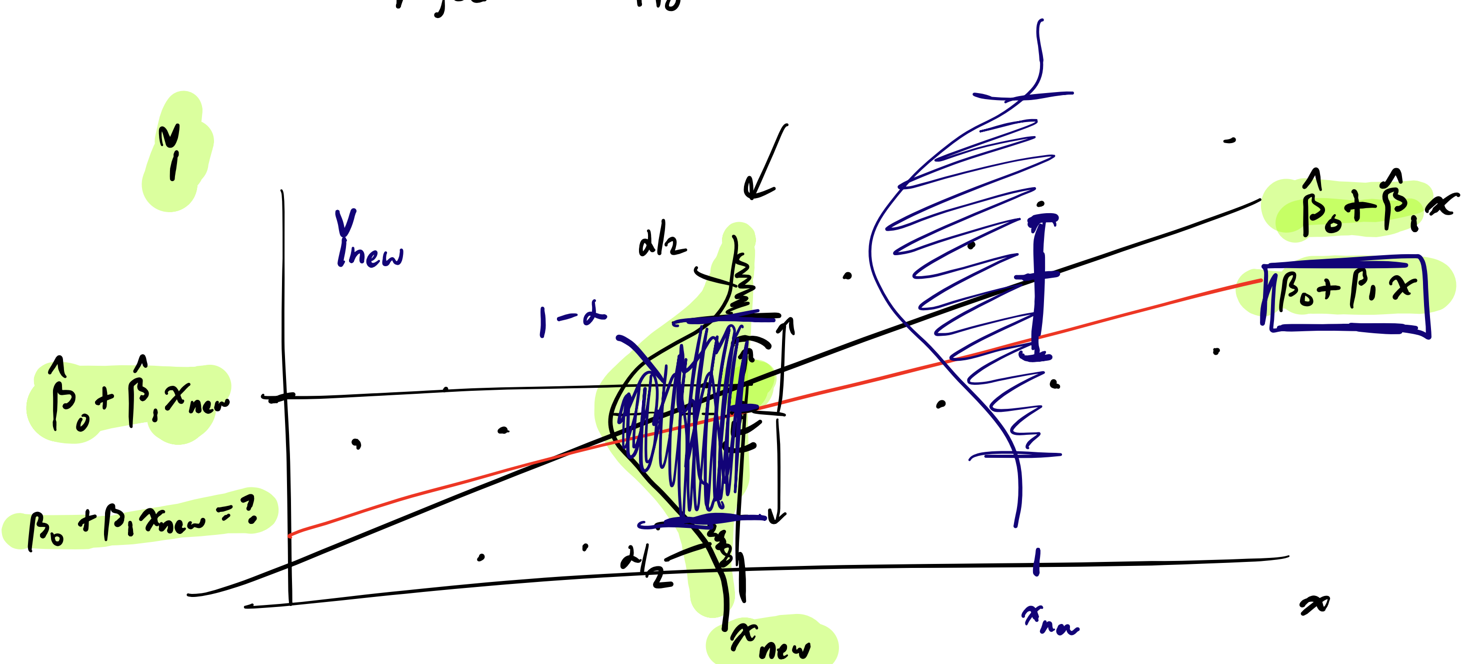
If x is a good predictor of Y, then we have R^2 close to 1.



Caution: Value of R^2 does not automatically tell you whether to reject

$H_0: \beta_1 = 0.$

Possible to have small R^2 but still reject H_0:



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

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

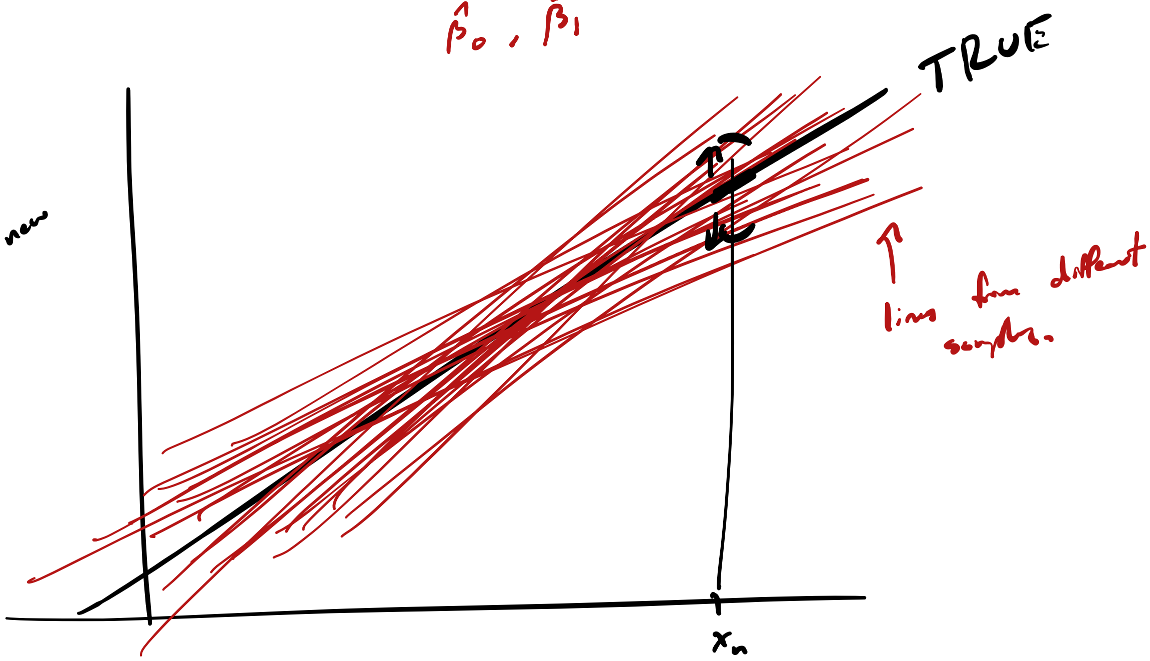
- A $(1-\alpha)100\%$ *prediction interval* for Y_{new} at x_{new} is given by

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

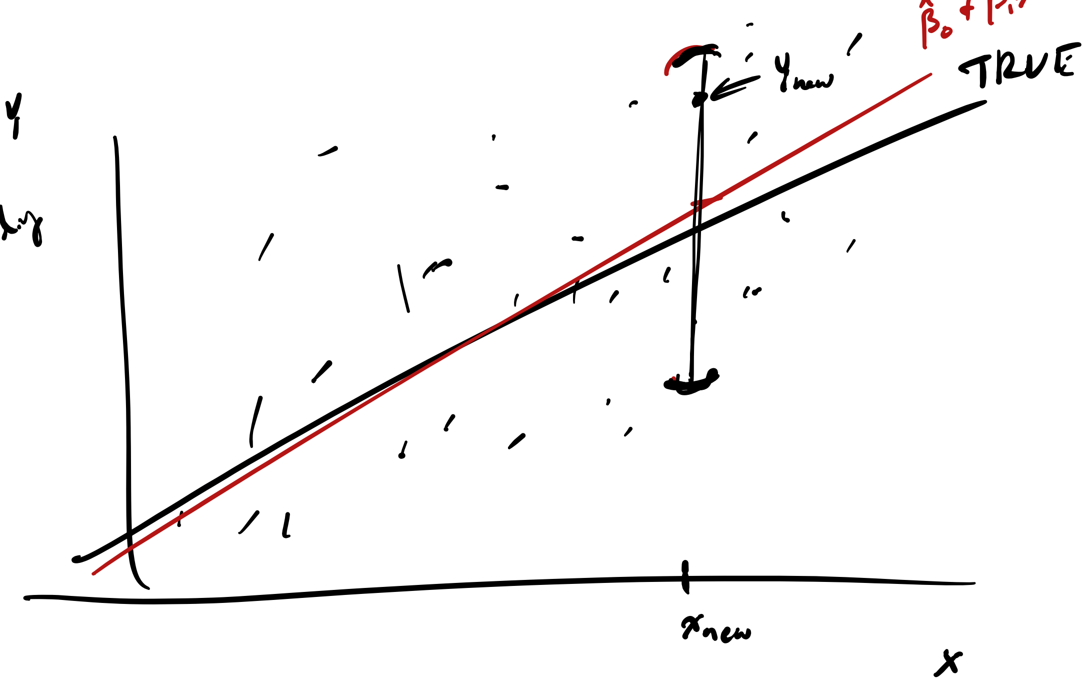
Exercise: Give the following intervals

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.
3. Make these intervals over a sequence of `Teff` values and plot the bounds.
4. Illustrate `predict()` function on `lm()` output.

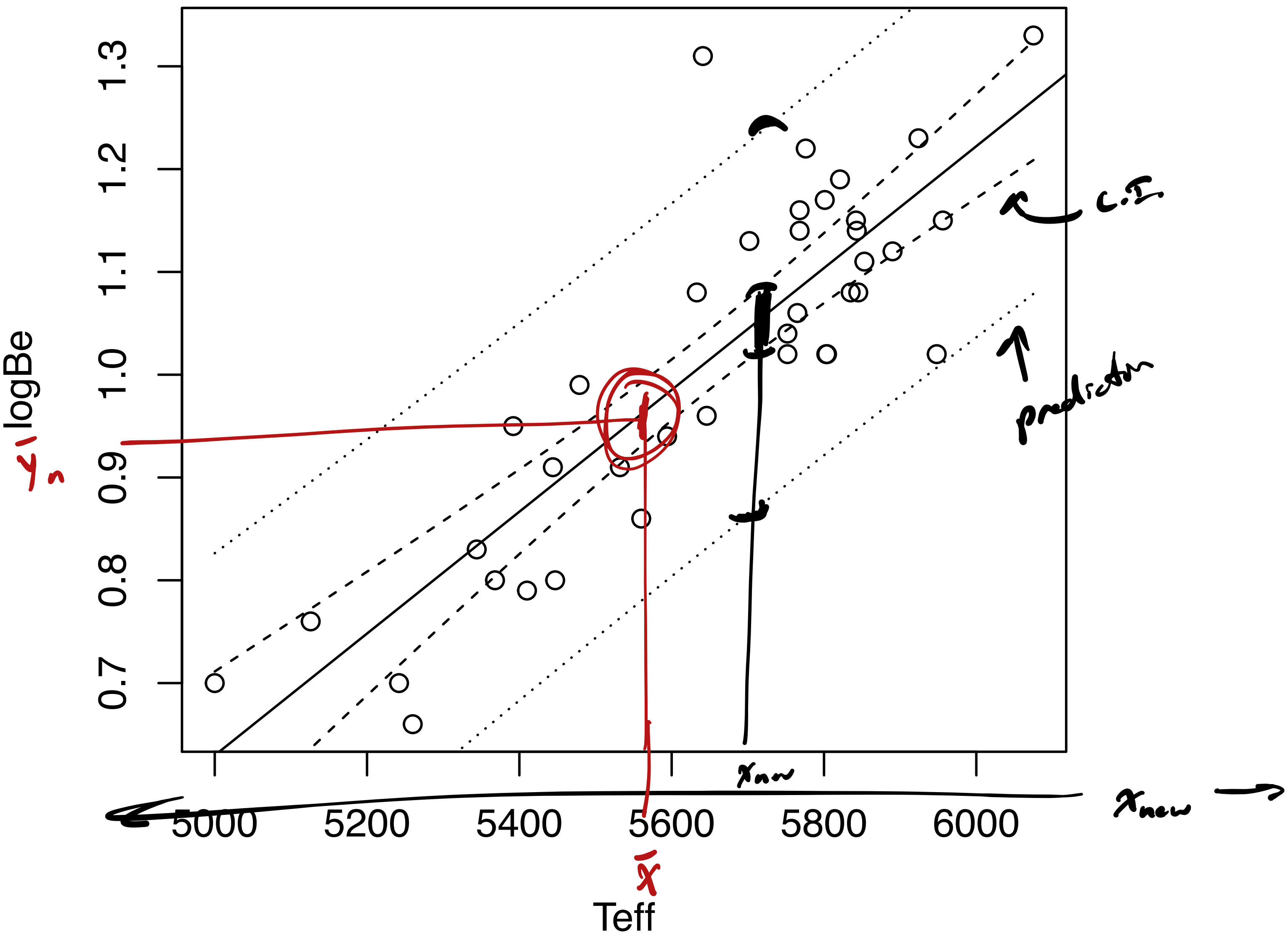
Confidence interval for $\beta_0 + \beta_1 x_{new}$



Prediction interval for Y_{new} corresponding to x_{new}.



```
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)
```


logBe
0.7
0.8
0.9
1.0
1.1
1.2
1.3
5000
5200
5400
5600
5800
6000
Teff

```
# 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")
```

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

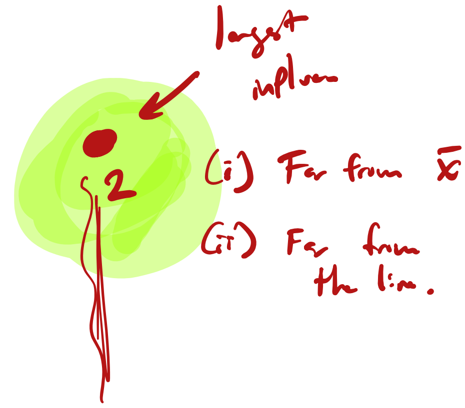
$$\text{lev}_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{S_{xx}}$$

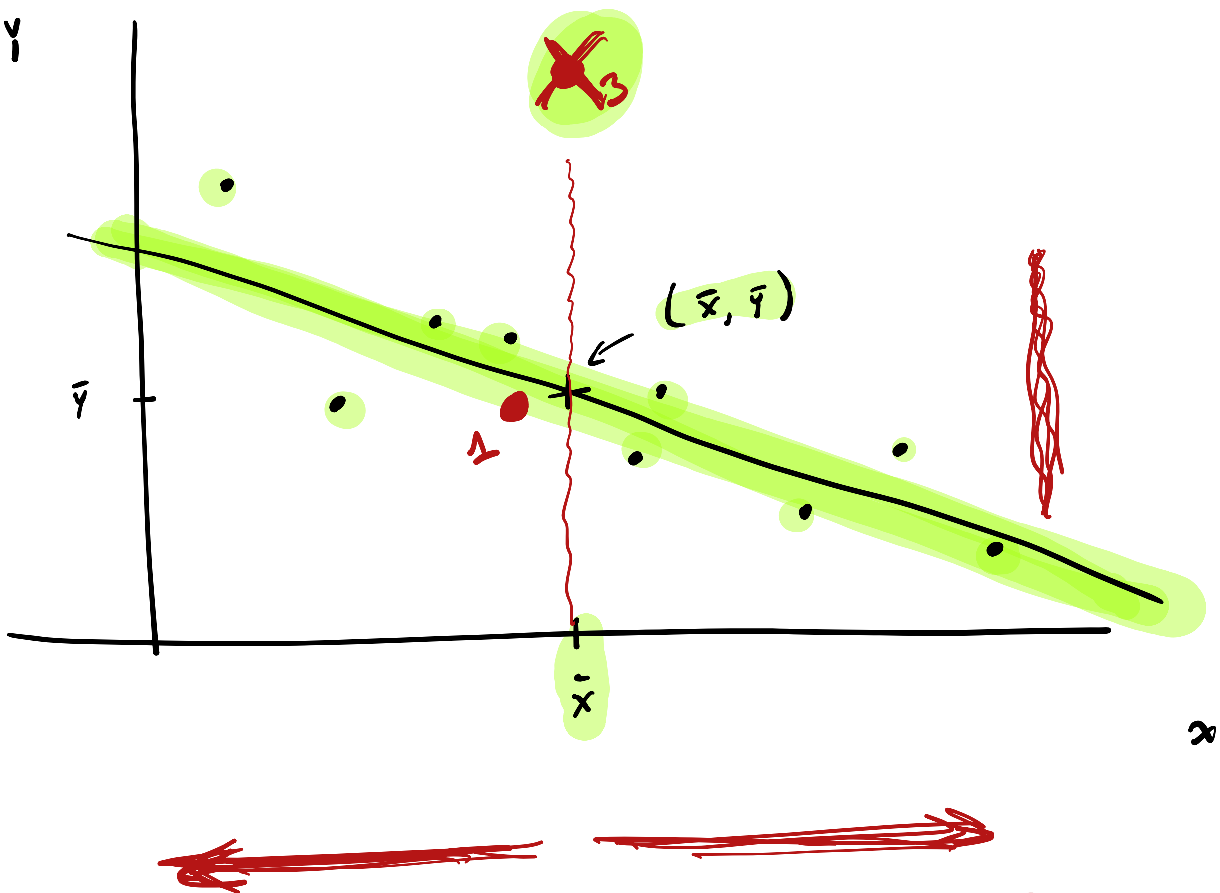
Points with high leverage have a large influence on the fitted regression line.

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

Draw pictures.

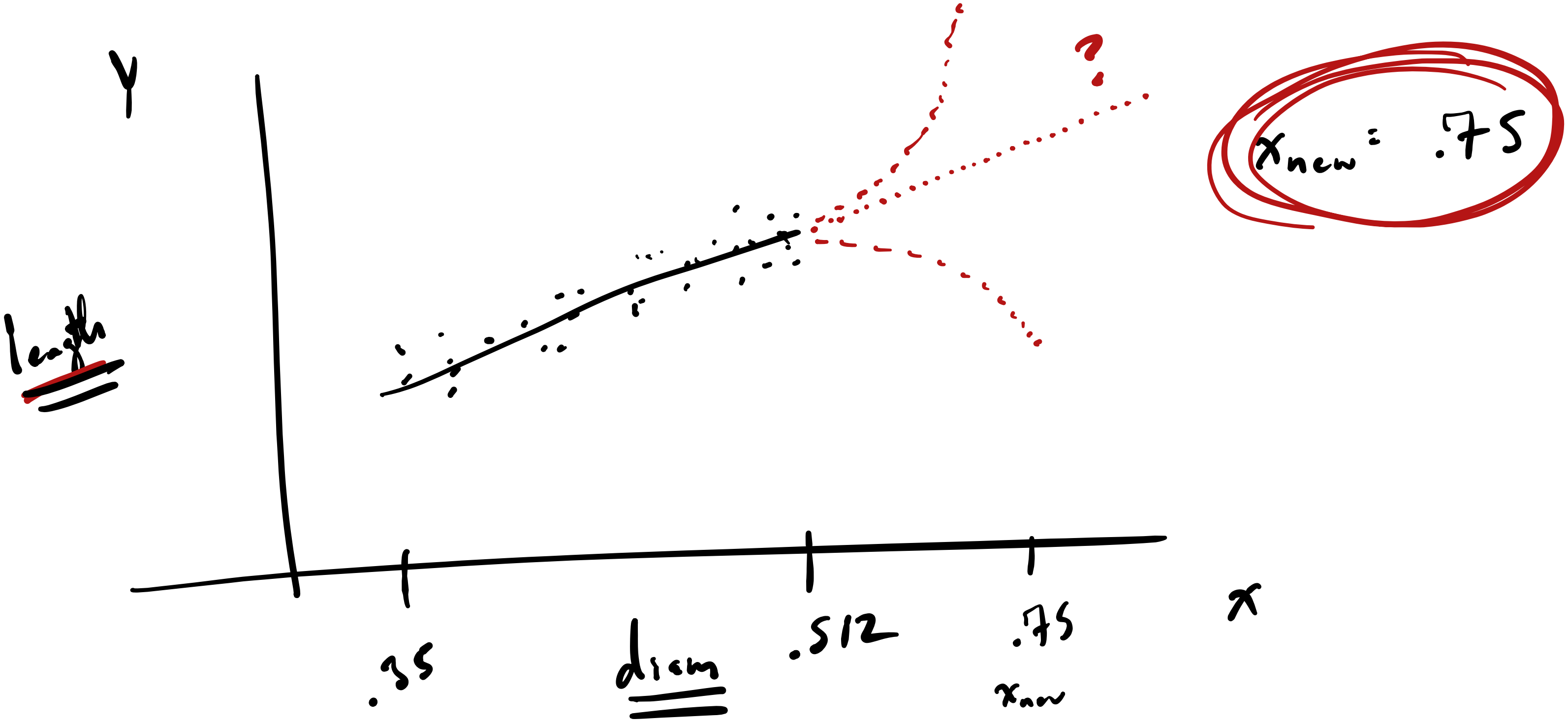
Leverage asks: What is the influence of a data point on the least-squares line?





leverage measures distance from $\bar{x}_n$.

Final admonition: Do not extrapolate.



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.