(a) Fill in the table with the fitted values $\hat{Y}_1, \ldots, \hat{Y}_5$ and the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_5$:

$\bar{Y}$ - Average $Y = 0/5 = 0$

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

i	Y_i	x_i	$\hat{Y}_i$	$\hat{\varepsilon}_i$
1	5	1	4	1
2	0	2	2	-2
3	0	3	0	0
4	0	4	-2	2
5	-5	5	-1	-1

(b) Compute SS_{Tot}.

$\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 \quad (5-0)^2 + (-5-0)^2 = 50$

(c) Compute SS_{Reg} and SS_{Error}.

$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2 \quad (4)^2 + (2)^2 + (-2)^2 + (-4)^2$

$SS_{Reg} = 40$

$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (1)^2 + (-2)^2 + (0)^2 + (2)^2 + (1)^2$

$SS_{Error} = 10$

(d) Compute R^2.

$\frac{SS_{Reg}}{SS_{Total}} \quad R^2 = \frac{40}{50} = 0.8$

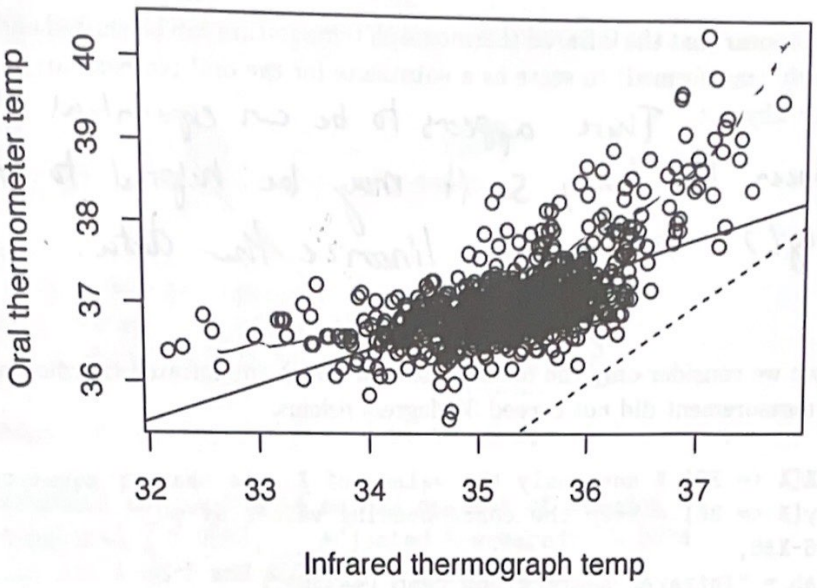
(e) Obtain $\hat{\sigma}^2$.

$\frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \quad \frac{1}{5-2} = \frac{1}{3} \quad (1)^2 + (-2)^2 + (2)^2 + (-1)^2 = \frac{1}{3}(10) = 3.\overline{33}$

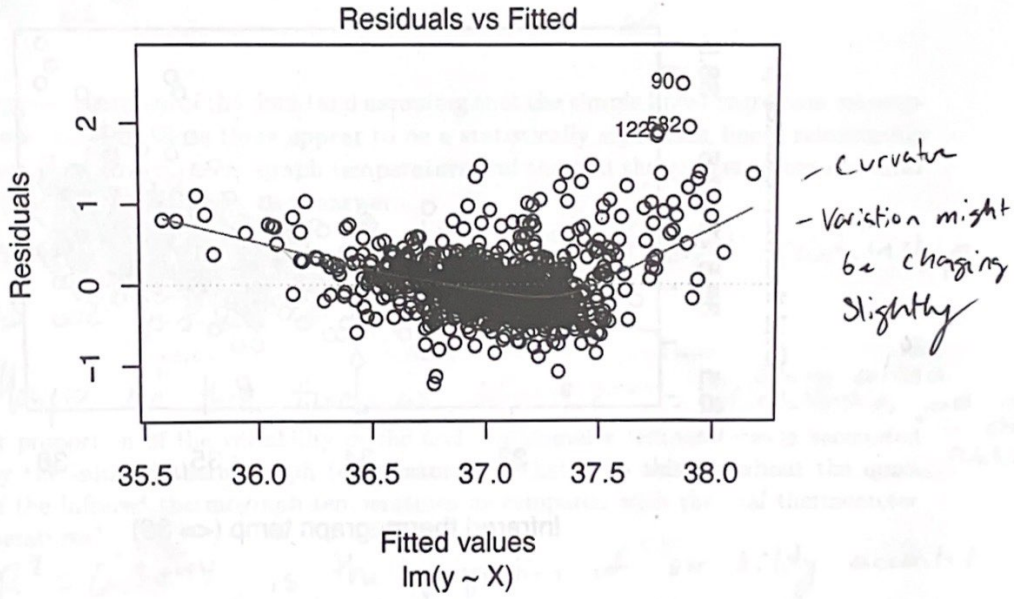
2. Simple linear regression (part 2)

Below is a scatterplot of several subjects' temperatures when measured with an oral thermometer (considered to be quite reliable) versus their temperatures when taken with an infrared thermograph (from a thermal image). It is of interest to see if the temperature measurement from the infrared thermograph is as reliable as that from the oral thermometer. The temperatures are recorded in degrees celcius. Overlaid on the plot is the line $y = x$ as well as the least-squares line. In addition, a residuals versus fitted values plot is shown.

```
plot(y~X,
     ylab = "Oral thermometer temp",
     xlab = "Infrared thermograph temp")
abline(0,1,lty = 2)
lm_out <- lm(y~X)
abline(lm_out)
```




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



- Curvature
- Variation might be changing slightly

(a) Describe the relationship between the temperatures recorded by the two measurement methods and comment on the accuracy and reliability of the infrared thermograph.

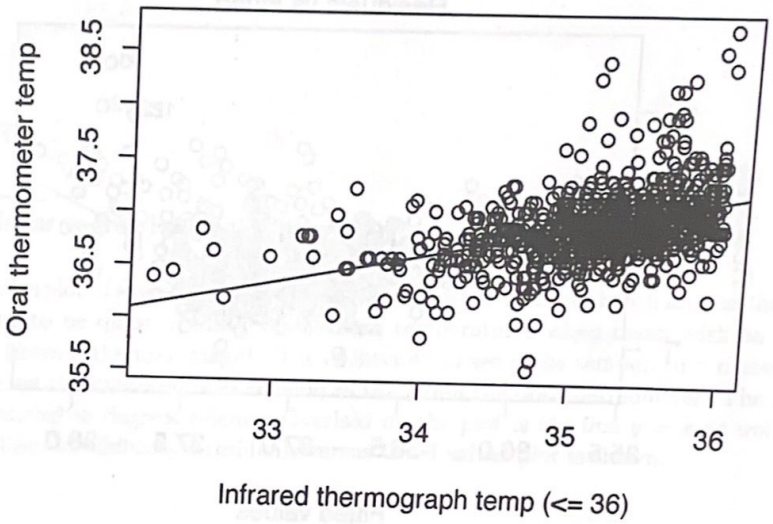
From the plot it seems there may be a positive correlation between infrared and oral, albeit not too strong. However, looking at the residual plot, and even the actual plot to an extent, there is clear evidence of a non-linear relationship. So it may not be the most reliable. Past ~ 36° the oral reads much higher

(b) Does it appear that the infrared thermograph temperature can be shifted and scaled (tha is linearly transformed) to serve as a substitute for the oral temperature measurement? Why or why not?

The infrared thermograph can not be linearly transformed because in order to be linear your graph must show no pattern on the residuals vs fitted plot. ✓

(c) Suppose we consider only the observations for which the infrared thermograph temperature measurement did not exceed 36 degrees celcius.

```
X36 <- X[X <= 36] # keep only the values of X less than or equal to 36
y36 <- y[X <= 36] # keep the corresponding values of y
plot(y36~X36,
     xlab = "Infrared thermograph temp (<= 36)",
     ylab = "Oral thermometer temp")
lm_out <- lm(y36~X36)
abline(lm_out)
```




```
summary(lm_out)
```

```
Call:
```

```
lm(formula = y36 ~ X36)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27662 -0.17215 -0.01855  0.14059  1.58592

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.51509    0.64008   42.99   <2e-16 ***
X36          0.26782    0.01821   14.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3044 on 822 degrees of freedom
Multiple R-squared:  0.2084,    Adjusted R-squared:  0.2074
F-statistic: 216.4 on 1 and 822 DF,  p-value: < 2.2e-16
```

(i) What is the sample size n after removing the observations for which the infrared thermograph temperature measurement did not exceed 36 degrees celcius?

824

$822 = n-(p+1)$ $n = 822+(p+1) = 822+(1+1) = 824$ ✓

(ii) Looking at this range of the data (and assuming that the simple linear regression assumptions are satisfied), does there appear to be a statistically significant linear relationship between the infrared thermograph temperature and the oral thermometer temperature? Use the R output to justify your answer.

Yes. the p-value (2e-16) is very small. This means that if there was no relationship, there would be an almost 0% chance we observed this data. So a relationship is extremely likely. ✓

(iii) What proportion of the variability in the oral thermometer temperatures is accounted for by the infrared thermograph temperatures? What does this say about the quality of the infrared thermograph temperatures as compared with the oral thermometer temperatures?

R^2

.2084 is accounted for by the infrared thermometer temps. Because this proportion is low, the infrared temps are likely not very reliable as opposed to the oral temps ✓

(iv) Suppose one stations an infrared camera at the entrance to a doctor's office and uses this fitted model to get an approximate temperature of each individual entering. Which interval would be more appropriate for expressing uncertainty about an individual's temperature—a confidence interval for the height of the true regression function at the observed infrared thermograph temperature, or a prediction interval for the new response value?

Prediction interval (Bc its for an individual)

(v) From the scatterplot, give (appoximately) the estimated oral thermometer temperature corresponding to an infrared thermograph temperature of 34 degrees celcius.

36.6°C

(vi) Identify which of the below intervals is the CI for the height of the true regression function and which is the PI for an individual oral thermometer temperature when the infrared thermograph temperature is 34 degrees celcius.

PI
(36.0217917, 37.2204201)

CI
(36.5750806, 36.6671312)

3. Multiple linear regression

The R code below fits a multiple linear regression model on a data set of measurements taken on possums. The response variable is the age of the possum and the predictors are various measurements taken on the possums.

```
lm_out <- lm(age ~ hdlngth + skullw + totlngth + taill + footlgth +
             earconch + eye + chest + belly, data = possum)
plot(lm_out,which = 1)
```

```
earconch     9.813e-02  8.528e-02   1.151    0.253
eye          2.524e-01  1.894e-01   1.333    0.186
chest        1.512e-01  1.417e-01   1.067    0.289
belly        1.437e-01  8.936e-02   1.608    0.111
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.802 on 91 degrees of freedom
Multiple R-squared:  0.1941,    Adjusted R-squared:  0.1144
F-statistic: 2.436 on 9 and 91 DF,  p-value: 0.01574
```

(a) What is the purpose of the Cook's distance plot? What is the Cook's distance of an observation?

The purpose of the Cook's distance plot is to show how much each data point changes the fit if removed, and how much it has an impact on the MLR. A larger Cook's distance, for example, will be different horizontally (leverage) and vertically. Observation 11 has Cook's Distance of about 0.19.

(b) Based on the R output, does it appear that one can accurately estimate the age of a possum based on the various measurements included as covariates? Explain your answer in detail.

No, as the R^2 value is close to zero showing little statistically significant correlation b/w the covariates and Age.

(c) Give your conclusion regarding the hypotheses H_0: $\beta_j = 0$ for all j (that is, *none* of the covariates is linearly related to **age**) versus H_1: $\beta_j \neq 0$ for at least one j (that is, at least one covariate is linearly related to **age**).

I reject H_0 with p-value 0.01574 and conclude that at least one covariate is linearly related to age.

(d) Suppose you wish to run a single test to check whether the covariates **hdlngth**, **skullw**, **totlngth**, **taill**, **footlgth**, **earconch**, and **eye** all have regression coefficients equal to zero. Fill in the missing degrees of freedom in the expression for the test statistic of the full-reduced model F test:

$$F_{\text{stat}} = \frac{\text{SS}_{\text{Error}}(\text{reduced}) - \text{SS}_{\text{Error}}(\text{full})/(\ ?\ ) \Rightarrow 5 = 7}{\text{SS}_{\text{Error}}(\text{full})/(\ ?\ ) \Rightarrow n-(p+1) = 91}$$

(e) This full-reduced model F test ends up having a p-value of 0.6569133. What does this mean?

Because it's so large, it means none of the 7 tested covariates have a linear relationship with age.

The R code below fits a model with only the covariates chest and belly.

```
lm_out <- lm(age ~ chest + belly, data = possum)
summary(lm_out)
```

```
Call:
lm(formula = age ~ chest + belly, data = possum)

Residuals:
   Min     1Q Median     3Q    Max
-3.266 -1.305 -0.349  1.173  4.953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.58479    2.54158  -2.591   0.0110 *
chest        0.17353    0.11128   1.559   0.1221
belly        0.17495    0.08244   2.122   0.0363 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.782 on 98 degrees of freedom
Multiple R-squared:  0.1512,    Adjusted R-squared:  0.1339
F-statistic: 8.732 on 2 and 98 DF,  p-value: 0.0003238
```

(f) Comment on whether either of the covariates chest and belly appear to have a significant linear relationship with the possum age. Use the R output to justify your answer.

Yes, at least one does because of a higher F-st
but more importantly an even lower p-value of 0.000324

(g) State whether a 95% confidence interval for the regression coefficient of the chest covariate would contain zero.

It would

4. Inference on the mean of a Normal distribution

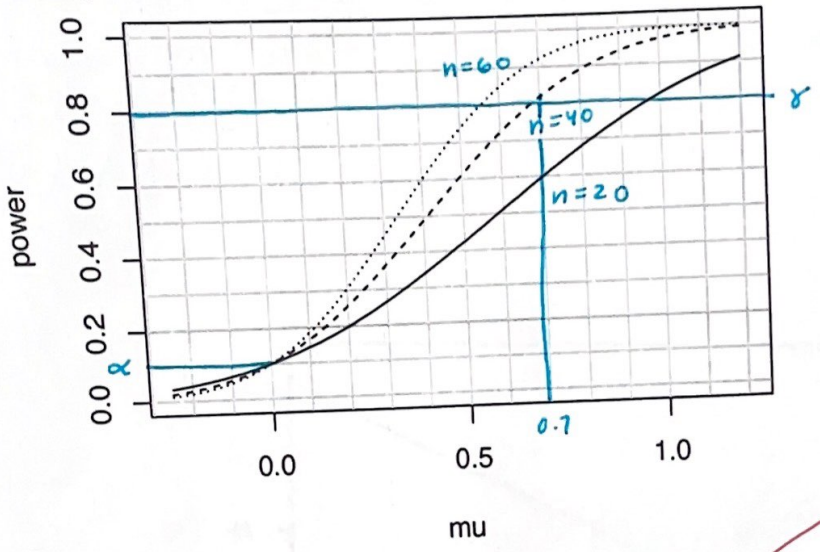
Suppose you draw a random sample from a distribution with unknown mean μ and unknown variance σ^2. You assume the distribution is a Normal distribution and you would like to test

$$H_0\colon\ \mu \leq 0 \text{ versus } H_1\colon\ \mu > 0.$$

Suppose the test which rejects H_0 when

$$\frac{\bar{X}_n - 0}{S_n/\sqrt{n}} > t_{n-1,\alpha}$$

has the power curves plotted below under the sample sizes $n = 20$, $n = 40$, and $n = 60$.



(a) What significance level α is being used?

$0.1 = \alpha$

(b) Label the power curves as corresponding to the sample sizes $n = 20$, $n = 40$, and $n = 60$.

As n increases, power increases

(c) Suppose you wish to reject H_0 with probability at least 0.80 when the true mean is 0.70. Which of the three sample sizes do you recommend? Explain why.

I recommend n=40, based on the plot, when $\mu = 0.70$, the power is 0.8 when n=40; and at n=20 it's 0.6 and at n=60 it's ~0.91.

(d) If H_0 is false, which of the three sample sizes gives the smallest probability of a Type II error?

Type II error - when H_0 is false, but we accept H_0.

Smaller n → more TII Error

Larger n → less TII Error

So, $n = 60$

(e) Describe the effect on the power curves of a larger variance σ^2. You may draw additional curves on the plot to illustrate your answer.

A larger variance decreases power.