STAT 516 sp 2025 exam 01

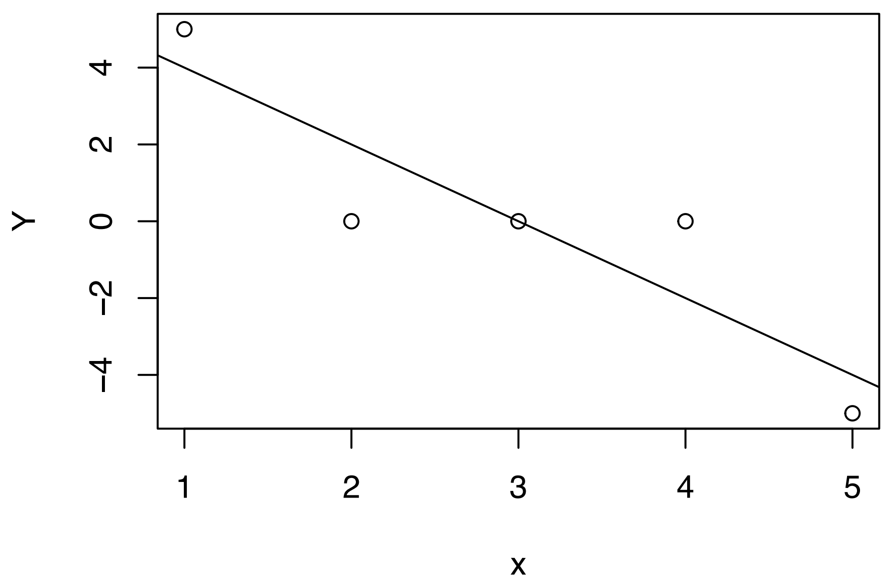
75 minutes, no calculators or notes allowed

Solutions

1. Simple linear regression (part 1)

Consider the data plotted below:

```r
x <- c(1,2,3,4,5)
Y <- c(5,0,0,0,-5)
plot(Y~x)
abline(6,-2)
```



The least-squares line has intercept $\hat{\beta}_0 = 6$ and slope $\hat{\beta}_1 = -2$.

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

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$= 6 - 2x_i$

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	-4	-1

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

(b) Compute $\mathrm{SS}_{\mathrm{Tot}}$.

$SS_{Tot} = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 = 5^2 + 5^2 = 50$. Note that $\bar{Y}_n = 0$.

(c) Compute $\mathrm{SS}_{\mathrm{Reg}}$ and $\mathrm{SS}_{\mathrm{Error}}$.

$SS_{Reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2 = 4^2 + 2^2 + (-2)^2 + (-4)^2 = 16 + 4 + 4 + 16 = 40$

$SS_{Error} = 1^2 + (-2)^2 + 2^2 + (-1)^2 = 1 + 4 + 4 + 1 = 10$

(d) Compute R^2.

$R^2 = \frac{40}{50} = 0.80$

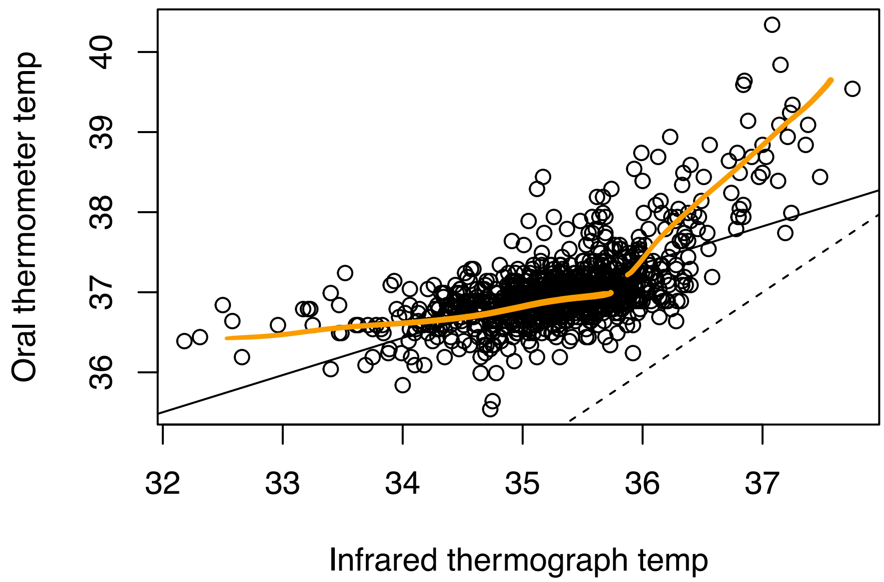
(e) Obtain $\hat{\sigma}^2$.

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

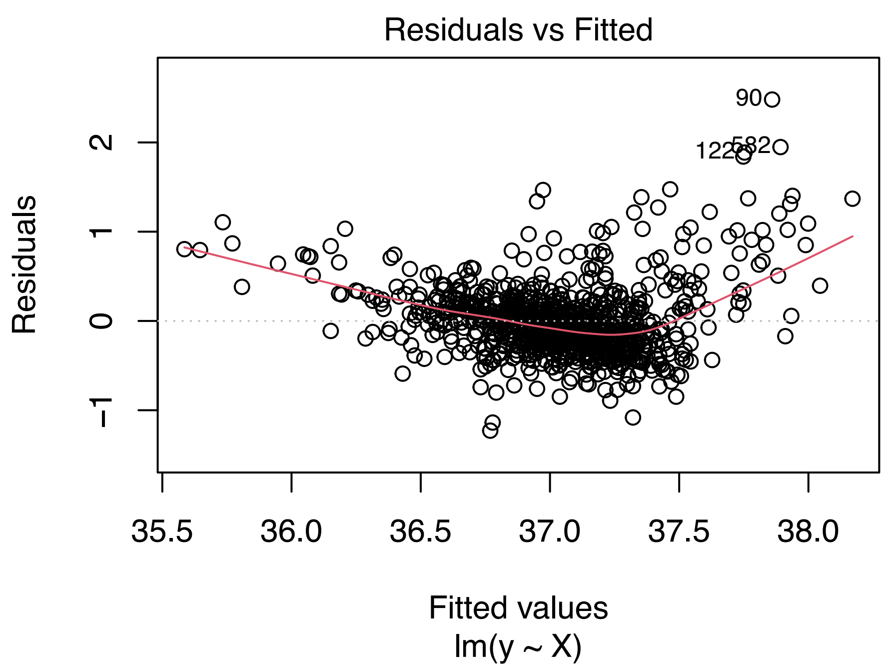
2. Simple linear regression (part 2)

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)
```




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



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

It looks like the infrared gives temperatures lower than the oral thermometer, since all the points are above the line $y=x$.

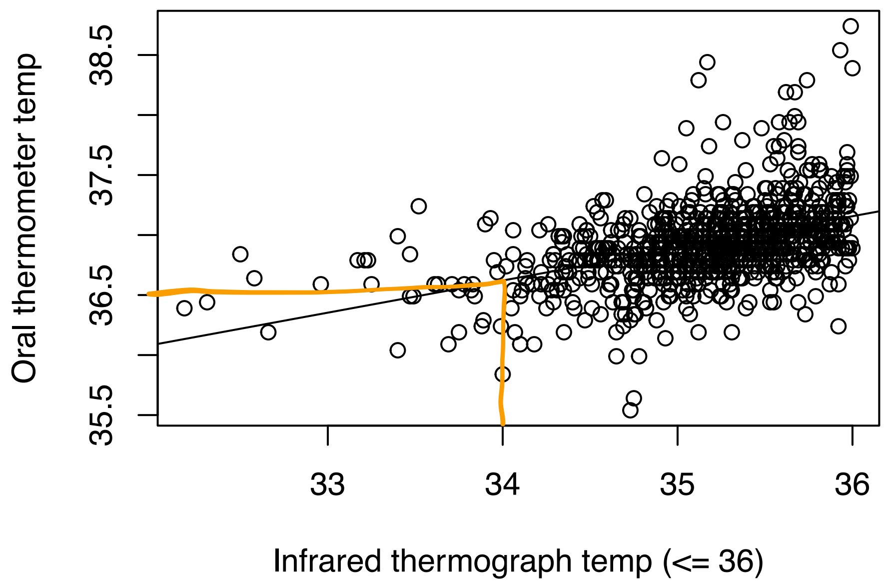
The relationship does not appear to have a constant slope.

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

No, the relationship appears to be nonlinear, as one cannot draw a single line nicely through the data points in the scatterplot. Moreover, the residuals vs fits plot shows a curved pattern in the residuals, which suggests a nonlinear relationship.

(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 ***  very small
---
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
                             = n-2
```

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

The sample size is $n = 824$, since the denominator df of the F statistic is given as 822, which is $n-2$ in simple linear regression.

(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 for testing $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ is very small, indicating strong evidence against the null hypothesis of no linear relationship.

(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?

This is the value R^2, which is shown in the output to be 0.2084. This seems much too small if one wishes to make accurate substitutions of infrared thermometer temperatures for oral temperatures.

(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?

The prediction interval would be more appropriate, since we are concerned with the temperature of a single individual.

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

This appears to be about 36.5°.

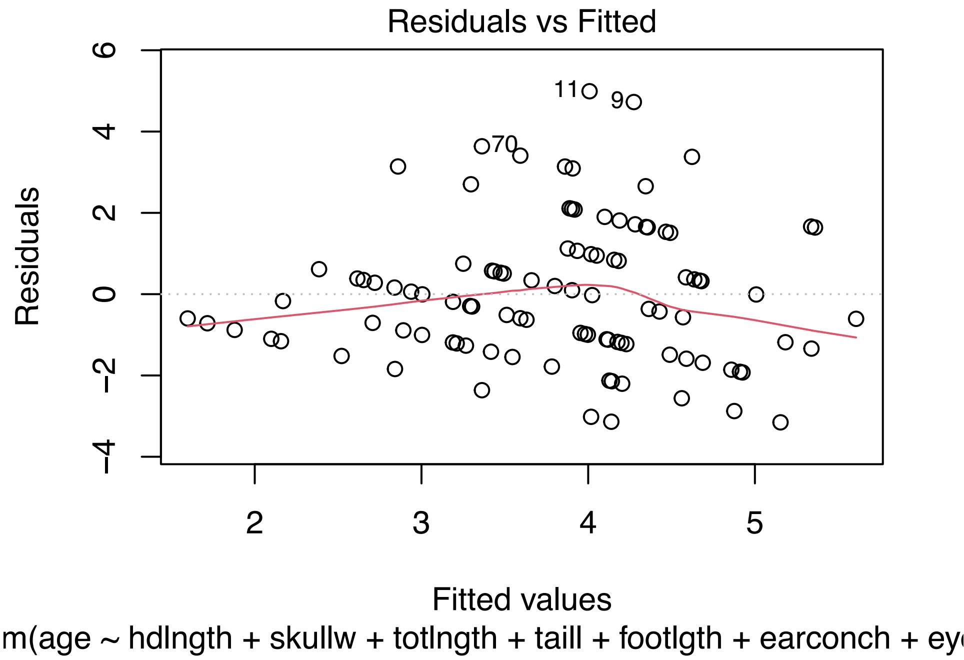
(vi) Identify which of the below intervals is the CI for the height of the true regression function and which is the PI for an individual oral thermometer temperature when the infrared thermograph temperature is 34 degrees celcius. Circle the correct interval.

PI (wider) (36.0217917, 37.2204201)   CI (36.5750806, 36.6671312)

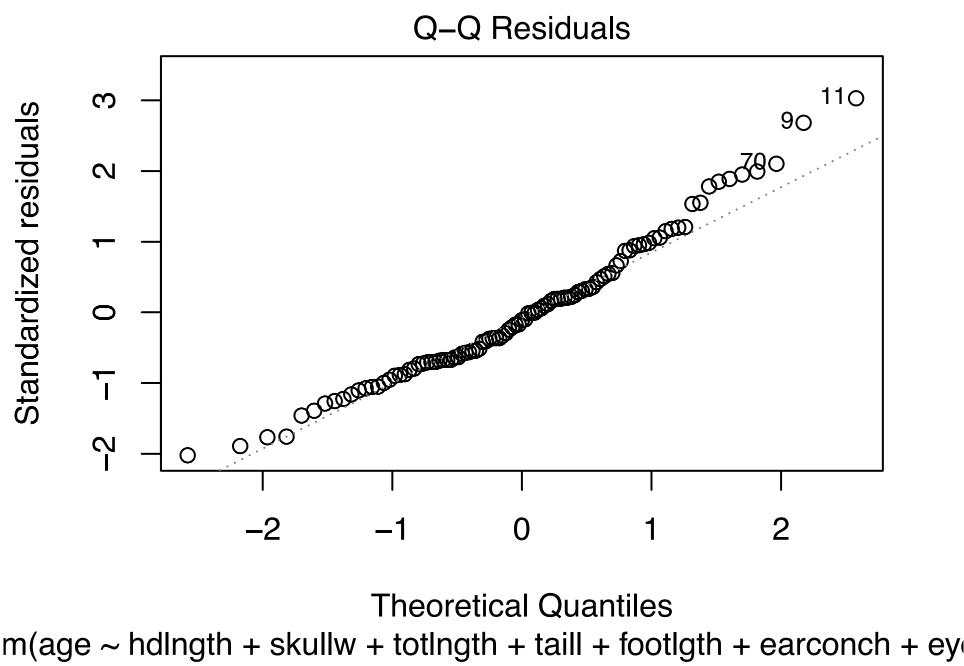
3. Multiple linear regression

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.

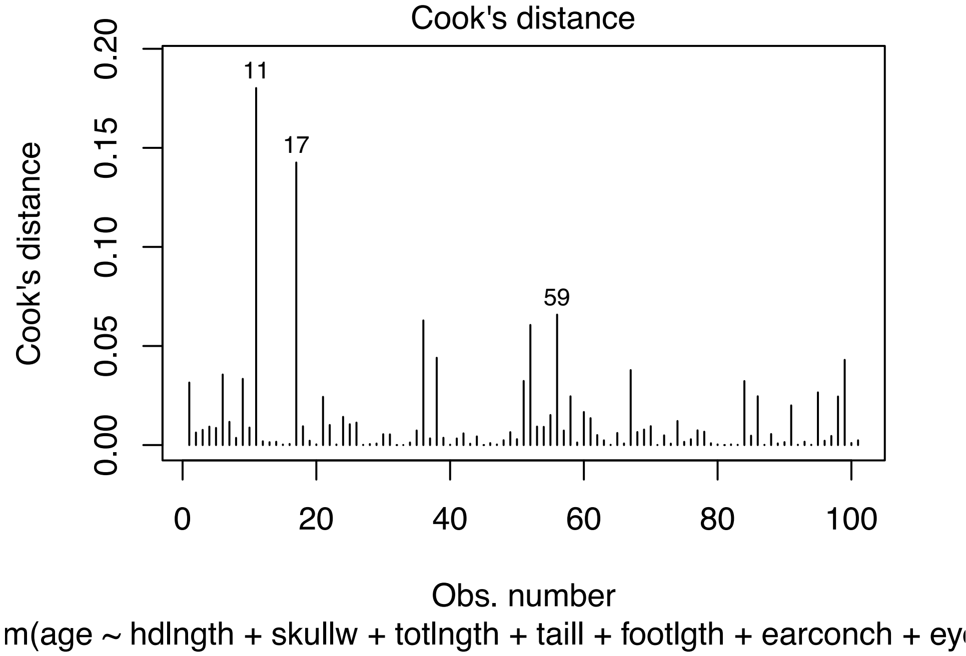
```
lm_out <- lm(age ~ hdlngth + skullw + totlngth + taill + footlgth +
               earconch + eye + chest + belly, data = possum)
plot(lm_out,which = 1)
```




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



```r
plot(lm_out,which = 4)
```




```r
summary(lm_out)
```

```
Call:
lm(formula = age ~ hdlngth + skullw + totlngth + taill + footlgth +
    earconch + eye + chest + belly, data = possum)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1535 -1.1886 -0.1893  0.9489  4.9917

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.369e+01  6.501e+00  -2.106    0.038 *
hdlngth      4.851e-02  8.969e-02   0.541    0.590
skullw       2.127e-02  8.870e-02   0.240    0.811
totlngth     2.249e-02  8.280e-02   0.272    0.787
taill       -3.169e-04  1.436e-01  -0.002    0.998
footlgth    -1.090e-01  8.316e-02  -1.311    0.193
```

```
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 plot of the Cook's distances helps us to identify outliers. The Cook's distance of an observation is a measurement of how much the least squares line would change if the observation were removed. A large value of Cook's distance indicates outlyingness.

(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.

The value R^2 is quite low, so estimates of possum age based on these covariates are not likely to be very accurate.

(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`).

The p-value for the overall F test is fairly small (0.01574), suggesting that at least one covariate is linearly related to the response.

(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:

7 (7 variables removed)

$$F_{\text{stat}} = \frac{\text{SS}_{\text{Error}}(\text{reduced}) - \text{SS}_{\text{Error}}(\text{full})/(\ ?\ )}{\text{SS}_{\text{Error}}(\text{full})/(\ ?\ )}$$

91 (n-(p+1) from R output)

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

This means there is little evidence that any of these 7 covariates is linearly related to the possum age.

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

```r
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.

The test of $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$ for "belly" has a fairly small p-value (0.0363), so this covariate may be linearly related to age. For "chest", however, the p-value is rather large, so there is no strong evidence that it is related to age.

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

Since the p-value for testing $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$ is 0.1221 > 0.05, a 95% C.I. for β_j would contain zero.

4. Inference on the mean of a Normal distribution

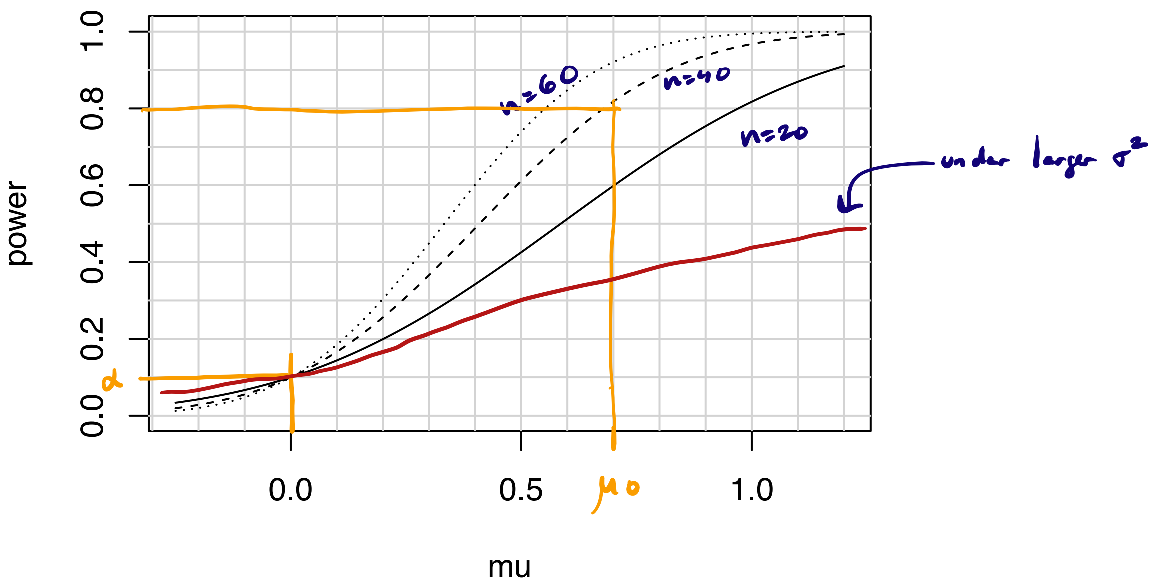
Suppose you draw a random sample from a distribution with unknown mean μ and unknown variance σ^2. You assume the distribution is a Normal distribution and you would like to test

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

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?

We have $\alpha = 0.10$. This is the height of the power curves at the null value $\mu_0 = 0$.

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

* See plot *

(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.

We should take $n = 40$. This is the smallest of the sample sizes under which the power is at least 0.80

when $\mu = 0.70$.

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

This is $n=60$. It maximizes the power when $\mu>0$ (so minimizing the probability of falsely failing to reject H_0).

(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.

Under a larger variance σ^2 the power curves would decrease for $\mu>0$. The curve would still have height equal to $\alpha=0.10$ at $\mu=0$.