BB503/BB602 - R Training - Week XII

Ege Ulgen

Power Analysis/Sample Size Calculation

```
# install.packages("pwr")
library(pwr)

### Effect size: magnitude of the effect under the alternative hypothesis
# The larger the effect size, the easier it is to detect an effect and require fewer samples
### Power: probability of correctly rejecting the null hypothesis if it is false
# The higher the power, the more likely it is to detect an effect if it is present and the more samples
# Standard setting for power is 0.80
### Significance level (alpha): probability of falsely rejecting the null hypothesis even though it is
# The lower the significance level, the more likely it is to avoid a false positive and
# the more samples needed
# Standard setting for alpha is 0.05

?cor.test

### Correlation
?pwr.r.test
# r=correlation
# sig.level=significant level
# power=power of test

# effect size >>> 0.1=small, 0.3=medium, and 0.5 large

# Is there a correlation between hours studied and test score?
# assuming large correlation
pwr.r.test(r=0.5, sig.level=0.05, power=0.80)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 28.248
##               r = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```
# calculating power
pwr.r.test(n = 50, r=0.5, sig.level=0.05)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
```

```
##                n = 50
##                r = 0.5
##        sig.level = 0.05
##            power = 0.96698
##      alternative = two.sided
```

```
pwr.r.test(n = 10, r=0.5, sig.level=0.05)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##                n = 10
##                r = 0.5
##        sig.level = 0.05
##            power = 0.32907
##      alternative = two.sided
```

```
### Two-sample t-test
?pwr.t.test
# d=effect size
# sig.level=significant level
# power=power of test
# type=type of test

# effect size >>> 0.2=small, 0.5=medium, and 0.8 large
# effect size calculation >>> Cohen's D

# Are the average body temperatures of women and men different?
# assuming medium effect size
pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##                n = 63.766
##                d = 0.5
##        sig.level = 0.05
##            power = 0.8
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# small effect size
pwr.t.test(d=0.2, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##                n = 393.41
##                d = 0.2
##        sig.level = 0.05
##            power = 0.8
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# large effect size
pwr.t.test(d=0.8, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 25.525
##               d = 0.8
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# Is the average body temperature higher in women than in men?
pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="greater")
```

```
##
##      Two-sample t test power calculation
##
##               n = 50.151
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
##
## NOTE: n is number in *each* group
```

```
### One-way ANOVA
?pwr.anova.test
# k=number of groups
# f=effect size
# sig.level=significant level
# power=power of test


# effect size >>> 0.1=small, 0.25=medium, and 0.4 large

# Is there a difference in disease incidence across 6 different cities?
# assuming small effect size
pwr.anova.test(k = 6 , f = 0.1 , sig.level = 0.05 , power = 0.80)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 6
##               n = 214.72
##               f = 0.1
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```
### Chi-squared test
?pwr.chisq.test
# w=effect size
# df=degrees of freedom
# sig.level=significant level
# power=power of test


# effect size >>> 0.1=small, 0.3=medium, and 0.5 large


# Does the observed proportions of phenotypes from a genetics experiment different from the expected 9:

# aassuming medium effect
# df = 4 (phenotypes) - 1 = 3
pwr.chisq.test(w=0.3, df=3, sig.level=0.05, power=0.80)
```

```
##
##      Chi squared power calculation
##
##               w = 0.3
##               N = 121.14
##              df = 3
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: N is the number of observations
```

```
### Linear Regression
?pwr.f2.test
# u=numerator degrees of freedom
# v=denominator degrees of freedom
# f2=effect size
# sig.level=significant level
# power=power of test

# effect size >>> 0.02=small, 0.15=medium, and 0.35 large

# Can height, age, and time spent at the gym, predict weight in adult males?
# assuming medium effect size
(res <- pwr.f2.test(u = 3, f2 = 0.15, sig.level = 0.05, power = 0.8))
```

```
##
##      Multiple regression power calculation
##
##               u = 3
##               v = 72.706
##              f2 = 0.15
##       sig.level = 0.05
##           power = 0.8
```

```
res$v + 4 #(tot. num. of variables)### Correlation
```

```
## [1] 76.706
```

```
?pwr.r.test
# r=correlation
# sig.level=significant level
# power=power of test
# effect size >>> 0.1=small, 0.3=medium, and 0.5 large

# Is there a correlation between hours studied and test score?
# assuming large correlation
pwr.r.test(r=0.5, sig.level=0.05, power=0.80)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 28.248
##               r = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```
# calculating power
pwr.r.test(n = 50, r=0.5, sig.level=0.05)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 50
##               r = 0.5
##       sig.level = 0.05
##           power = 0.96698
##     alternative = two.sided
```

```
pwr.r.test(n = 10, r=0.5, sig.level=0.05)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 10
##               r = 0.5
##       sig.level = 0.05
##           power = 0.32907
##     alternative = two.sided
```

```
### Two-sample t-test
?pwr.t.test
# d=effect size
# sig.level=significant level
# power=power of test
# type=type of test

# effect size >>> 0.2=small, 0.5=medium, and 0.8 large
# effect size calculation >>> Cohen's D

# Are the average body temperatures of women and men different?
# assuming medium effect size
pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.766
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# small effect size
pwr.t.test(d=0.2, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 393.41
##               d = 0.2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# large effect size
pwr.t.test(d=0.8, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 25.525
##               d = 0.8
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# Is the average body temperature higher in women than in men?
pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="greater")
```

```
##
##      Two-sample t test power calculation
##
##               n = 50.151
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
##
## NOTE: n is number in *each* group
```

```
### One-way ANOVA
?pwr.anova.test
# k=number of groups
# f=effect size
# sig.level=significant level
# power=power of test
```

```r
# effect size >>> 0.1=small, 0.25=medium, and 0.4 large

# Is there a difference in disease incidence across 6 different cities?
# assuming small effect size
pwr.anova.test(k = 6 , f = 0.1 , sig.level = 0.05 , power = 0.80)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 6
##               n = 214.72
##               f = 0.1
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```r
### Chi-squared test
?pwr.chisq.test
# w=effect size
# df=degrees of freedom
# sig.level=significant level
# power=power of test


# effect size >>> 0.1=small, 0.3=medium, and 0.5 large


# Does the observed proportions of phenotypes from a genetics experiment different from the expected 9:

# aassuming medium effect
# df = 4 (phenotypes) - 1 = 3
pwr.chisq.test(w=0.3, df=3, sig.level=0.05, power=0.80)
```

```
##
##      Chi squared power calculation
##
##               w = 0.3
##               N = 121.14
##              df = 3
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: N is the number of observations
```

```r
### Linear Regression
?pwr.f2.test
# u=numerator degrees of freedom
# v=denominator degrees of freedom
# f2=effect size
# sig.level=significant level
# power=power of test

# effect size >>> 0.02=small, 0.15=medium, and 0.35 large
# Can height, age, and time spent at the gym, predict weight in adult males?
# assuming medium effect size
(res <- pwr.f2.test(u = 3, f2 = 0.15, sig.level = 0.05, power = 0.8))
```

```
##
##      Multiple regression power calculation
##
##               u = 3
##               v = 72.706
##              f2 = 0.15
##       sig.level = 0.05
##           power = 0.8
```

```
res$v + 4 #(tot. num. of variables)
```

```
## [1] 76.706
```

Linear Regression

Rationale

We'll use the pre/post dataset for this exercise. This dataset contains simulated data for pre-intervention measurements (`pre`) for 20 individuals together with their post-intervention measurements (`post`).

```
pre_post_df <- read.csv("../data/pre_post_data.csv")
dim(pre_post_df)
```

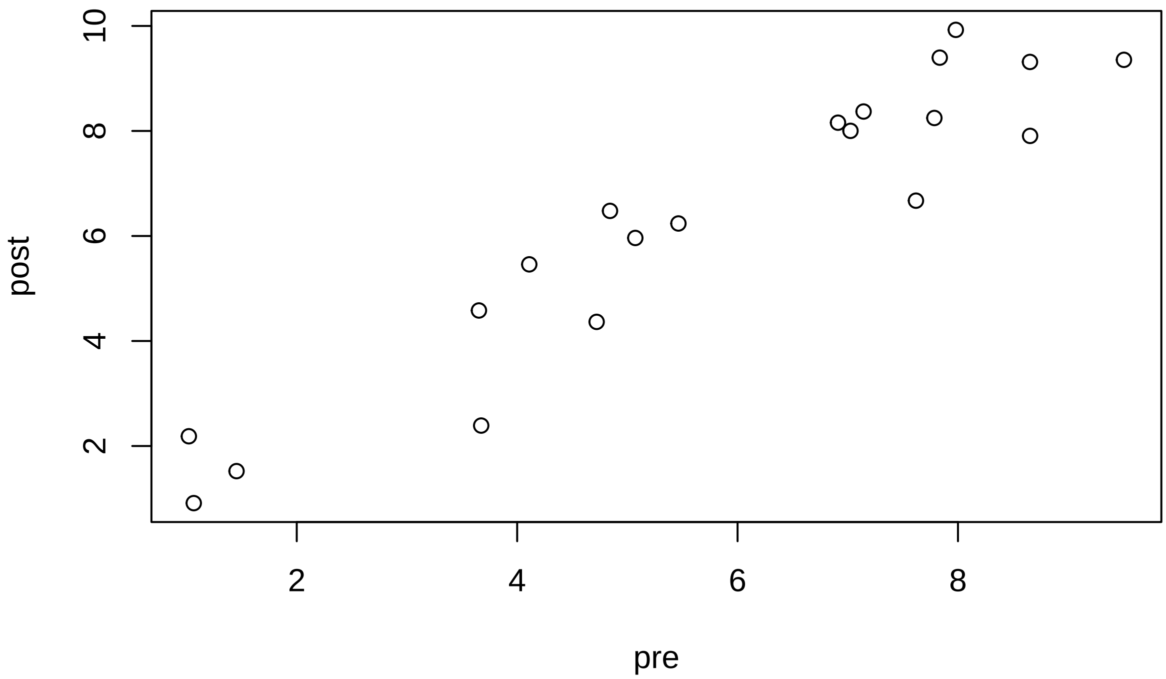
```
## [1] 20  2
```

```
head(pre_post_df, 3)
```

```
##      pre   post
## 1 7.7858 8.2474
## 2 7.6186 6.6725
## 3 8.6532 9.3145
```

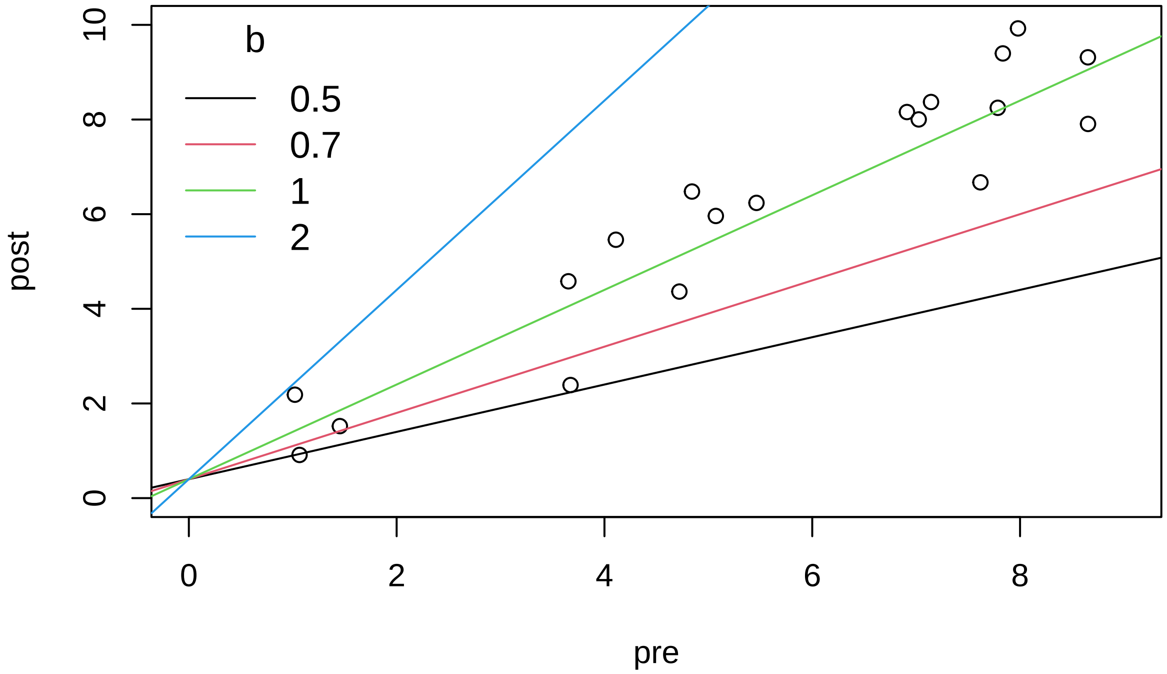
Let's visualize the scatter plot to investigate any relationship between `pre` and `post` measurements:

```
plot(post~pre, data = pre_post_df)
```



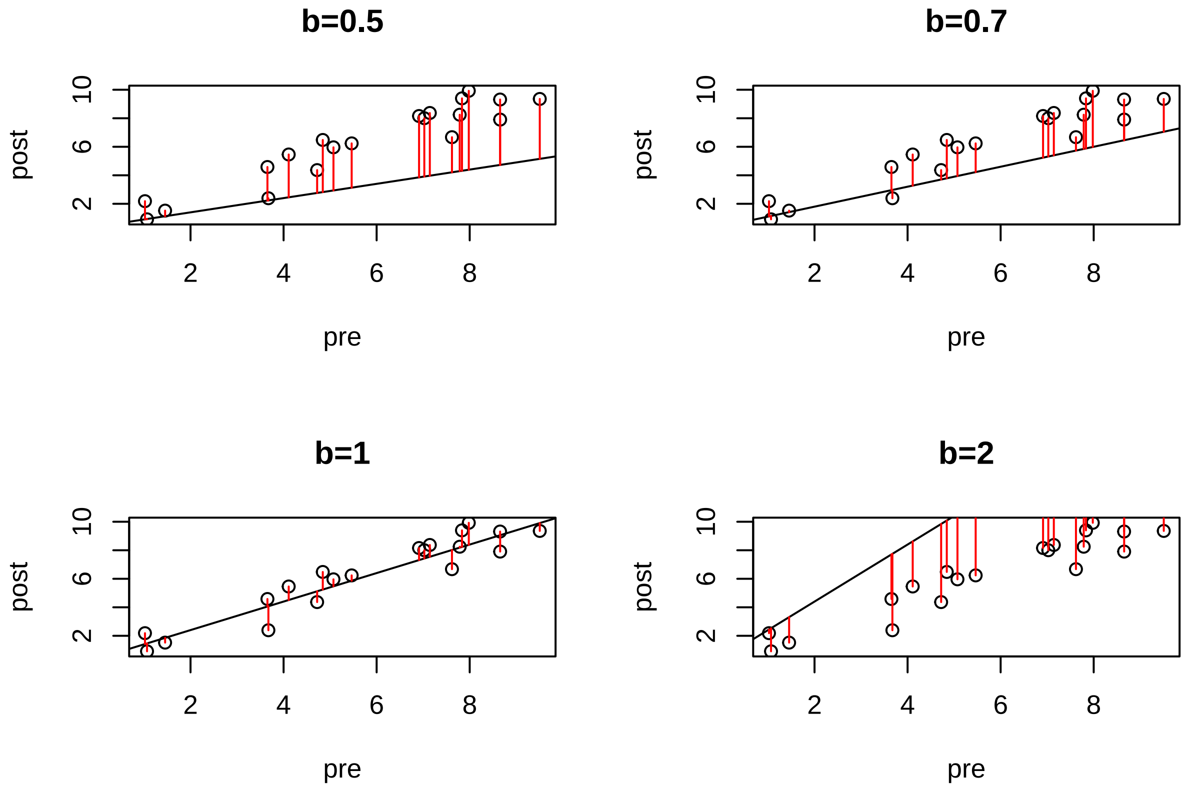
Using the function `abline()`, we can add different lines, where the argument `b` is the slope and `a` is the intercept

```
plot(post~pre, data = pre_post_df, xlim = c(0, 9), ylim = c(0, 10))
# y = bx + a
abline(a = .4, b = .5, col = 1)
abline(a = .4, b = .7, col = 2)
abline(a = .4, b = 1, col = 3)
abline(a = .4, b = 2, col = 4)

legend("topleft", legend = c(.5, .7, 1, 2), title = "b", col = 1:4, cex = 1.2, bty = "n", lty = 1)
```



Our aim is to minimize the distance to the line (residuals = errors):

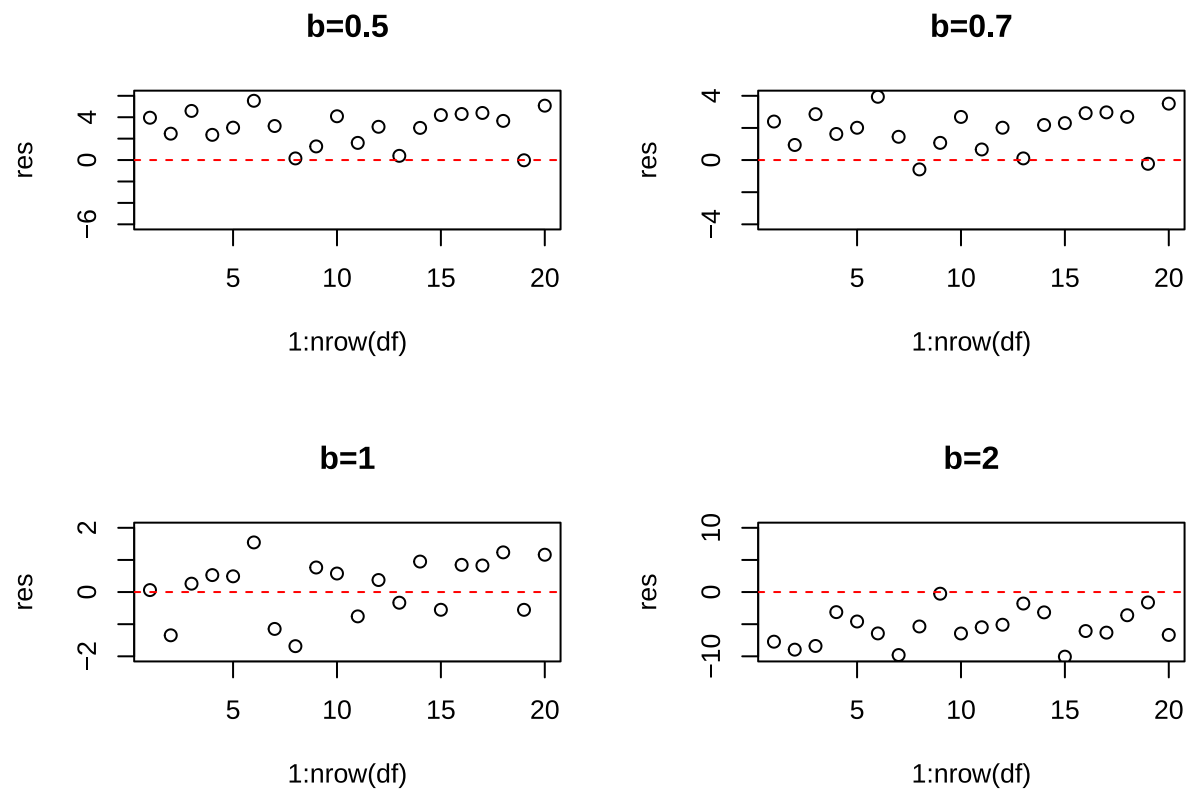
```r
plot_residual_dist <- function(df, b, a =.4, col = 1) {
    plot(post~pre, data = df, main = paste0("b=", b))
    abline(a = a, b = b, col = col)
    segments(x0 = df$pre, y0 = df$post, x1 = df$pre, y1 = b * df$pre + a, col = "red")
}

par(mfrow = c(2, 2))
plot_residual_dist(pre_post_df, .5)
plot_residual_dist(pre_post_df, .7)
plot_residual_dist(pre_post_df, 1)
plot_residual_dist(pre_post_df, 2)
```



```
par(mfrow = c(1, 1))
```

The residuals should be around 0:

```
plot_residual_vals <- function(df, b, a = .4) {
    preds <- b * df$pre + a
    res <- df$post - preds
    tmp <- round(max(abs(res)))
    plot(1:nrow(df), res, ylim = c(-tmp, tmp), main = paste0("b=", b))
    abline(h = 0, col = "red", lty = 2)
}

par(mfrow = c(2, 2))
plot_residual_vals(pre_post_df, 0.5)
plot_residual_vals(pre_post_df, 0.7)
plot_residual_vals(pre_post_df, 1)
plot_residual_vals(pre_post_df, 2)
```



```
par(mfrow = c(1, 1))
```

Examples

Simple Linear Regression

```
fit_simple <- lm(post~pre, pre_post_df)
summary(fit_simple)
```

```
##
## Call:
## lm(formula = post ~ pre, data = pre_post_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.811 -0.670  0.278  0.668  1.342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4616     0.5165    0.89     0.38
## pre           1.0177     0.0826   12.32  3.3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.94 on 18 degrees of freedom
## Multiple R-squared:  0.894,  Adjusted R-squared:  0.888
## F-statistic:  152 on 1 and 18 DF,  p-value: 3.32e-10
```

```
# prediction
new_data <- data.frame(pre = c(3.2, 1.8, 8.2))
predict(fit_simple, new_data)
```

```
##      1      2      3 
## 3.7182 2.2935 8.8067
```

Multiple Linear Regression

We'll use the prostate cancer dataset for this exercise. The main aim of collecting this data set was to inspect the associations between prostate-specific antigen (PSA) and prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies.

```
prca_df <- read.csv("../data/prostate_cancer.csv")

dim(prca_df)
```

```
## [1] 97  8
```

```
head(prca_df, 3)
```

```
##     PSA    vol     wt age BPH invasion penetration Gleason
## 1 0.651 0.5599 15.959  50   0        0           0       6
## 2 0.852 0.3716 27.660  58   0        0           0       7
## 3 0.852 0.6005 14.732  74   0        0           0       7
```

```
# turn categorical variables into factors
prca_df$invasion <- as.factor(prca_df$invasion)
prca_df$Gleason <- as.factor(prca_df$Gleason)

summary(prca_df)
```

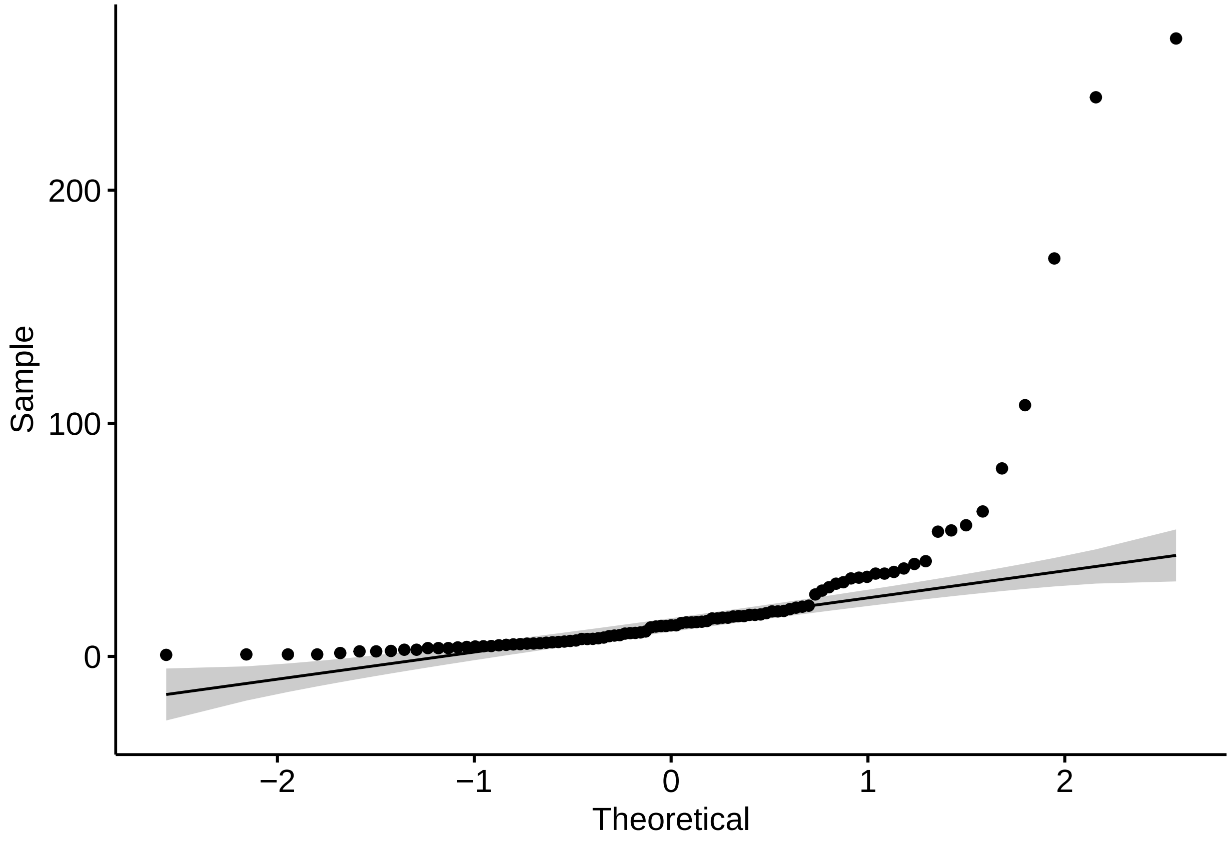
```
##       PSA               vol               wt             age      
##  Min.   :  0.651   Min.   : 0.259   Min.   : 10.7   Min.   :41.0  
##  1st Qu.:  5.641   1st Qu.: 1.665   1st Qu.: 29.4   1st Qu.:60.0  
##  Median : 13.330   Median : 4.263   Median : 37.3   Median :65.0  
##  Mean   : 23.730   Mean   : 6.999   Mean   : 45.5   Mean   :63.9  
##  3rd Qu.: 21.328   3rd Qu.: 8.415   3rd Qu.: 48.4   3rd Qu.:68.0  
##  Max.   :265.072   Max.   :45.604   Max.   :450.3   Max.   :79.0  
##       BPH        invasion  penetration     Gleason
##  Min.   : 0.00   0:76     Min.   : 0.000   6:33   
##  1st Qu.: 0.00   1:21     1st Qu.: 0.000   7:43   
##  Median : 1.35            Median : 0.449   8:21   
##  Mean   : 2.53            Mean   : 2.245          
##  3rd Qu.: 4.76            3rd Qu.: 3.254          
##  Max.   :10.28            Max.   :18.174          
```

```
## check normality of dependent variable
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
ggqqplot(prca_df$PSA)
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

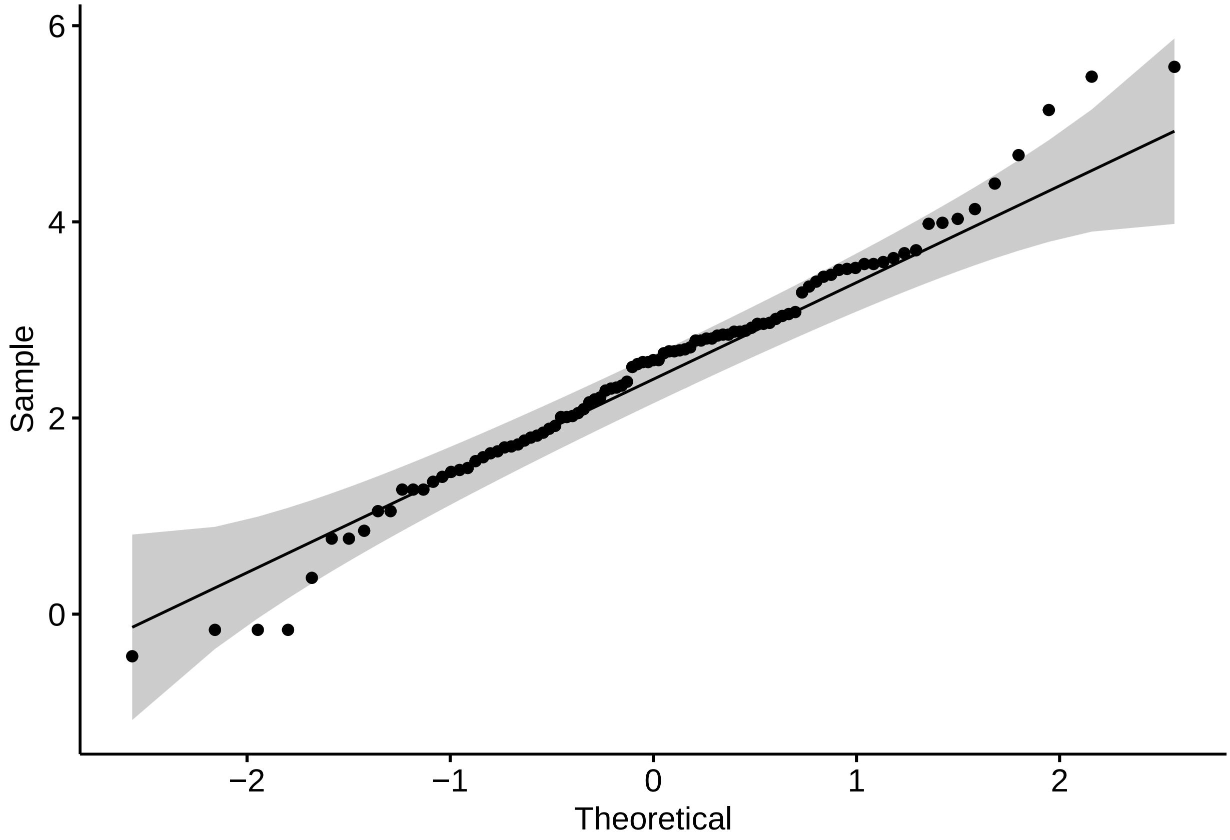


```
# the data quantiles deviates from the normal distribution quantiles
shapiro.test(prca_df$PSA)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  prca_df$PSA
## W = 0.479, p-value <2e-16
```

```
# For this reason, take the natural log values of the PSA levels, and again test for normality
prca_df$log_PSA <- log(prca_df$PSA)
ggqqplot(prca_df$log_PSA)
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



```
shapiro.test(prca_df$log_PSA)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  prca_df$log_PSA
## W = 0.984, p-value = 0.31
```

Let's check the effect of `Gleason` score on log(PSA) levels:

```
fit_gleason <- lm(log_PSA~Gleason, data = prca_df)
summary(fit_gleason)
```

```
##
## Call:
## lm(formula = log_PSA ~ Gleason, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5462 -0.6000  0.0052  0.5840  2.1600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.870      0.168   11.13  < 2e-16 ***
## Gleason7       0.516      0.223    2.31    0.023 *
## Gleason8       1.755      0.269    6.51  3.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.965 on 94 degrees of freedom
## Multiple R-squared:  0.314,  Adjusted R-squared:   0.3
## F-statistic: 21.6 on 2 and 94 DF,  p-value: 1.97e-08
```

```
# change reference level
?relevel()
prca_df$Gleason
```

```
##  [1] 6 7 7 6 6 6 6 6 7 6 6 6 7 7 7 6 7 6 6 7 6 7 6 7 6 6 7 7 7 6 6 6 6 6 6 7 8 7
## [39] 7 7 8 7 6 7 7 7 8 7 6 6 7 6 7 7 8 7 7 6 8 7 6 7 8 7 6 7 7 7 6 7 7 7 7 8 8 6
## [77] 8 8 7 8 7 7 7 8 8 8 6 7 8 8 6 7 8 8 8 8 8
## Levels: 6 7 8
```

```
prca_df$Gleason <- relevel(prca_df$Gleason, ref = "7")
prca_df$Gleason
```

```
##  [1] 6 7 7 6 6 6 6 6 7 6 6 6 7 7 7 6 7 6 6 7 6 7 6 7 6 6 7 7 7 6 6 6 6 6 6 7 8 7
## [39] 7 7 8 7 6 7 7 7 8 7 6 6 7 6 7 7 8 7 7 6 8 7 6 7 8 7 6 7 7 7 6 7 7 7 7 8 8 6
## [77] 8 8 7 8 7 7 7 8 8 8 6 7 8 8 6 7 8 8 8 8 8
## Levels: 7 6 8
```

```
fit_gleason2 <- lm(log_PSA~Gleason, data = prca_df)
summary(fit_gleason2)
```

```
##
## Call:
## lm(formula = log_PSA ~ Gleason, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5462 -0.6000  0.0052  0.5840  2.1600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.386      0.147   16.21  < 2e-16 ***
## Gleason6      -0.516      0.223   -2.31    0.023 *
## Gleason8       1.239      0.257    4.82  5.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.965 on 94 degrees of freedom
## Multiple R-squared:  0.314,  Adjusted R-squared:   0.3
## F-statistic: 21.6 on 2 and 94 DF,  p-value: 1.97e-08
```

```
prca_df$Gleason <- relevel(prca_df$Gleason, ref = "6")
```

Let's adjust for age:

```
fit_gleason3 <- lm(log_PSA~Gleason + age, data = prca_df)
summary(fit_gleason3)
```

```
##
## Call:
```

```
## lm(formula = log_PSA ~ Gleason + age, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6336 -0.6132  0.0181  0.5386  2.1002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2972     0.8569    1.51    0.133
## Gleason7      0.4878     0.2278    2.14    0.035 *
## Gleason8      1.7140     0.2767    6.19  1.6e-08 ***
## age           0.0093     0.0137    0.68    0.497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.968 on 93 degrees of freedom
## Multiple R-squared:  0.318,  Adjusted R-squared:  0.296
## F-statistic: 14.4 on 3 and 93 DF,  p-value: 8.42e-08
```

```
fit_gleason4 <- lm(log_PSA~Gleason + I(age - min(age)), data = prca_df)
summary(fit_gleason4)
```

```
##
## Call:
## lm(formula = log_PSA ~ Gleason + I(age - min(age)), data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6336 -0.6132  0.0181  0.5386  2.1002
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.6786     0.3275    5.13  1.6e-06 ***
## Gleason7            0.4878     0.2278    2.14    0.035 *
## Gleason8            1.7140     0.2767    6.19  1.6e-08 ***
## I(age - min(age))   0.0093     0.0137    0.68    0.497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.968 on 93 degrees of freedom
## Multiple R-squared:  0.318,  Adjusted R-squared:  0.296
## F-statistic: 14.4 on 3 and 93 DF,  p-value: 8.42e-08
```

```
fit_gleason5 <- lm(log_PSA~Gleason * age, data = prca_df)
summary(fit_gleason5)
```

```
##
## Call:
## lm(formula = log_PSA ~ Gleason * age, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7096 -0.6077  0.0198  0.4919  2.0275
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.2895     1.3455   -0.22   0.8301
## Gleason7       1.5521     2.0264    0.77   0.4457
## Gleason8       6.2421     2.1540    2.90   0.0047 **
## age            0.0351     0.0217    1.62   0.1093
## Gleason7:age  -0.0177     0.0319   -0.55   0.5803
## Gleason8:age  -0.0704     0.0333   -2.11   0.0375 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.955 on 91 degrees of freedom
## Multiple R-squared:  0.351,  Adjusted R-squared:  0.315
## F-statistic: 9.84 on 5 and 91 DF,  p-value: 1.53e-07
```

- the effect of age when (Gleason score = 6) = 0.0351
- the effect of age when (Gleason score = 7) = 0.0351 + (-0.0177)
- the effect of age when (Gleason score = 8) = 0.0351 + (-0.0704)

What are important factors that have an effect on PSA levels?

```r
fit0 <- lm(log_PSA~vol + wt + age + BPH + invasion + penetration + Gleason, data = prca_df)
summary(fit0)
```

```
##
## Call:
## lm(formula = log_PSA ~ vol + wt + age + BPH + invasion + penetration +
##     Gleason, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8512 -0.4541  0.0702  0.4555  1.5093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.49646    0.72260    2.07   0.0413 *
## vol          0.06745    0.01537    4.39  3.1e-05 ***
## wt           0.00127    0.00185    0.69   0.4942
## age         -0.00280    0.01178   -0.24   0.8127
## BPH          0.08911    0.02997    2.97   0.0038 **
## invasion1    0.79318    0.27061    2.93   0.0043 **
## penetration -0.02653    0.03301   -0.80   0.4238
## Gleason7     0.29677    0.18891    1.57   0.1198
## Gleason8     0.74661    0.26604    2.81   0.0062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.771 on 88 degrees of freedom
## Multiple R-squared:  0.59,   Adjusted R-squared:  0.553
## F-statistic: 15.8 on 8 and 88 DF,  p-value: 3.14e-14
```

```r
# equivalently
fit0 <- lm(log_PSA~. - PSA, data = prca_df)
summary(fit0)
```

```
##
## Call:
```

```
## lm(formula = log_PSA ~ . - PSA, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8512 -0.4541  0.0702  0.4555  1.5093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.49646    0.72260    2.07   0.0413 *
## vol          0.06745    0.01537    4.39  3.1e-05 ***
## wt           0.00127    0.00185    0.69   0.4942
## age         -0.00280    0.01178   -0.24   0.8127
## BPH          0.08911    0.02997    2.97   0.0038 **
## invasion1    0.79318    0.27061    2.93   0.0043 **
## penetration -0.02653    0.03301   -0.80   0.4238
## Gleason7     0.29677    0.18891    1.57   0.1198
## Gleason8     0.74661    0.26604    2.81   0.0062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.771 on 88 degrees of freedom
## Multiple R-squared:  0.59,   Adjusted R-squared:  0.553
## F-statistic: 15.8 on 8 and 88 DF,  p-value: 3.14e-14
```

```
## keeping only significant variables, fit final model
fit1 <- lm(log_PSA~vol + BPH + invasion + Gleason, data = prca_df)
summary(fit1)
```

```
##
## Call:
## lm(formula = log_PSA ~ vol + BPH + invasion + Gleason, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8524 -0.4578  0.0674  0.5165  1.5320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3882     0.1561    8.89  5.3e-14 ***
## vol           0.0624     0.0137    4.57  1.6e-05 ***
## BPH           0.0927     0.0263    3.53  0.00066 ***
## invasion1     0.6965     0.2384    2.92  0.00439 **
## Gleason7      0.2603     0.1828    1.42  0.15790
## Gleason8      0.7055     0.2571    2.74  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.764 on 91 degrees of freedom
## Multiple R-squared:  0.585,  Adjusted R-squared:  0.562
## F-statistic: 25.6 on 5 and 91 DF,  p-value: 4.72e-16
```

```
# prediction
new_data <- data.frame(vol = 0.42, BPH = 1.8, invasion = "1", Gleason = "6")
predict(fit1, new_data)
```

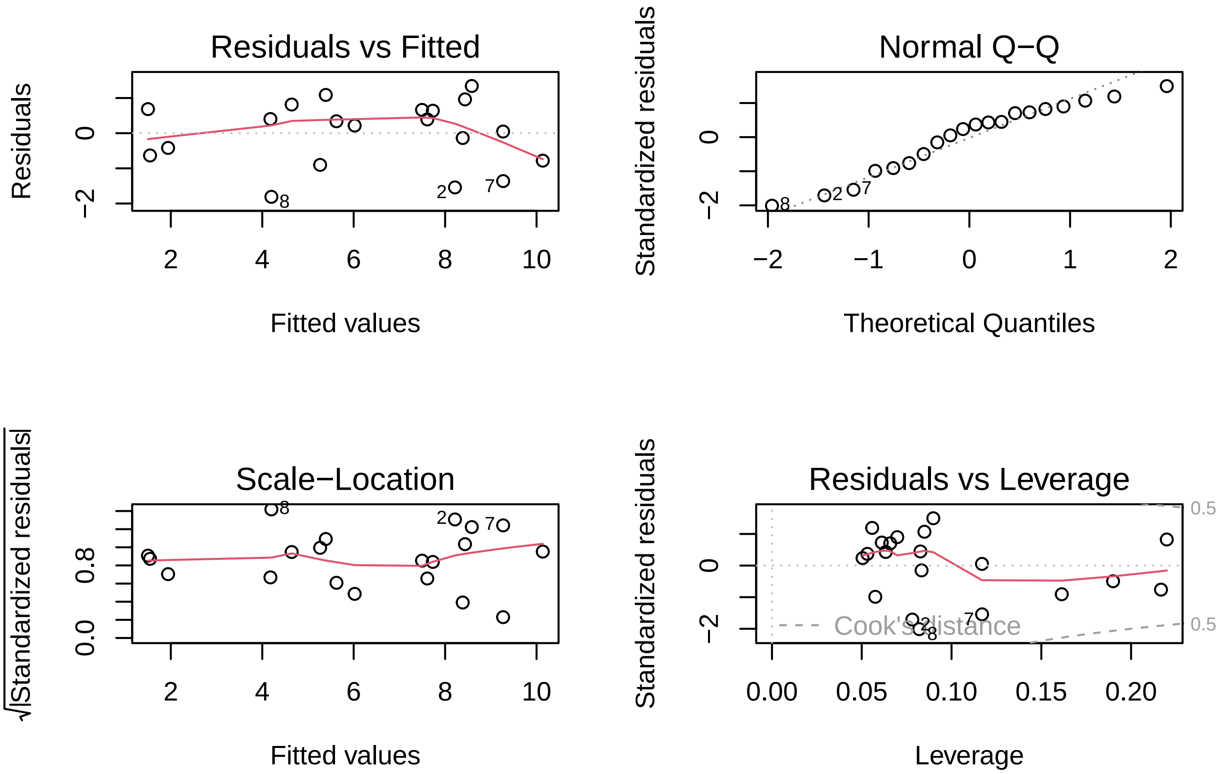
```
##      1
## 2.2776
```

Model Diagnostics

For detailed information, read this article on STHDA: http://sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

Let’s evaluate the diagnostic plots for the initial simple linear regression model `fit_simple`:

```
par(mfrow = c(2, 2))
plot(fit_simple)
```



```
par(mfrow = c(1, 1))
```

Residuals vs Fitted

Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship. In this case, there is a slight deviation, which is an issue.

Normal Q-Q

Used to examine whether the residuals are normally distributed.

Scale-Location (or Spread-Location)

Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. In this case, there seems to be a heteroscedasticity issue.

Residuals vs. Leverage

Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis

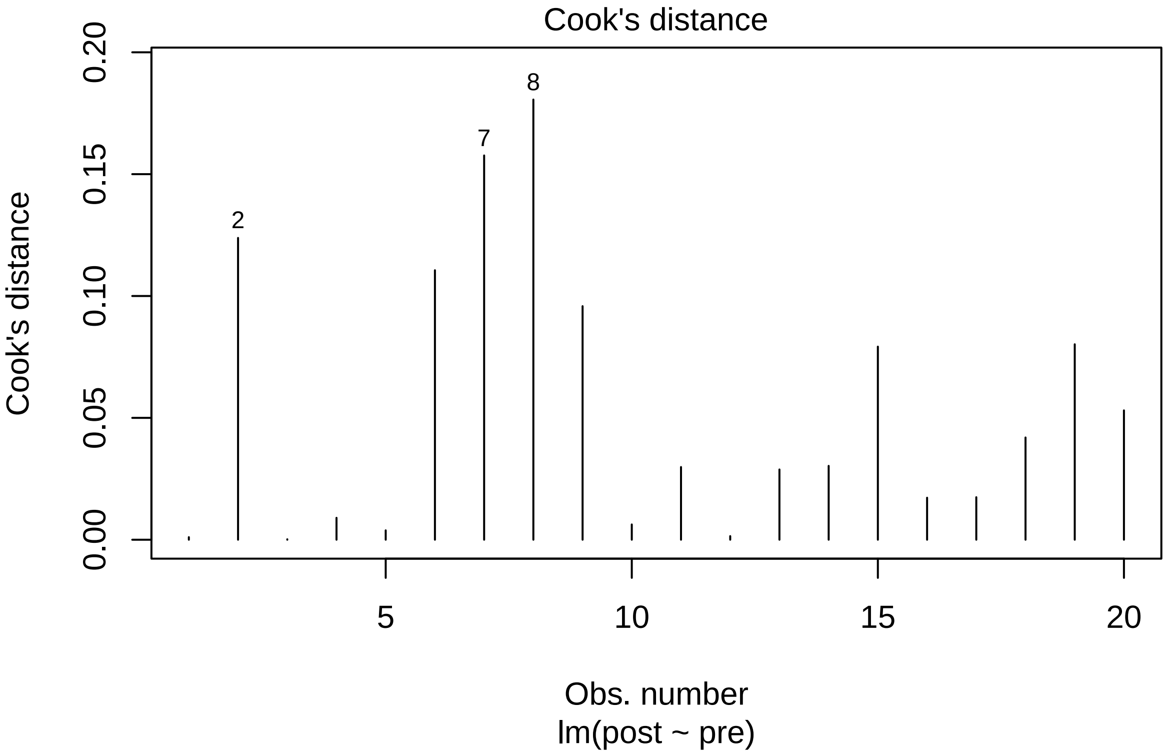
- Standardized residual: the residual divided by its estimated standard error. Standardized residuals can be interpreted as the number of standard errors away from the regression line. Observations whose standardized residuals are greater than 3 in absolute value are possible outliers
- Leverage: A data point has high leverage, if it has extreme predictor x values

Outlying values are generally located at the upper right corner or at the lower right corner. Those spots are the places where data points can be influential against a regression line.

Cook's Distance

This metric defines influence as a combination of leverage and residual size .

```
plot(fit_simple, 4)
```


Cook's distance
Cook's distance
Obs. number
lm(post ~ pre)