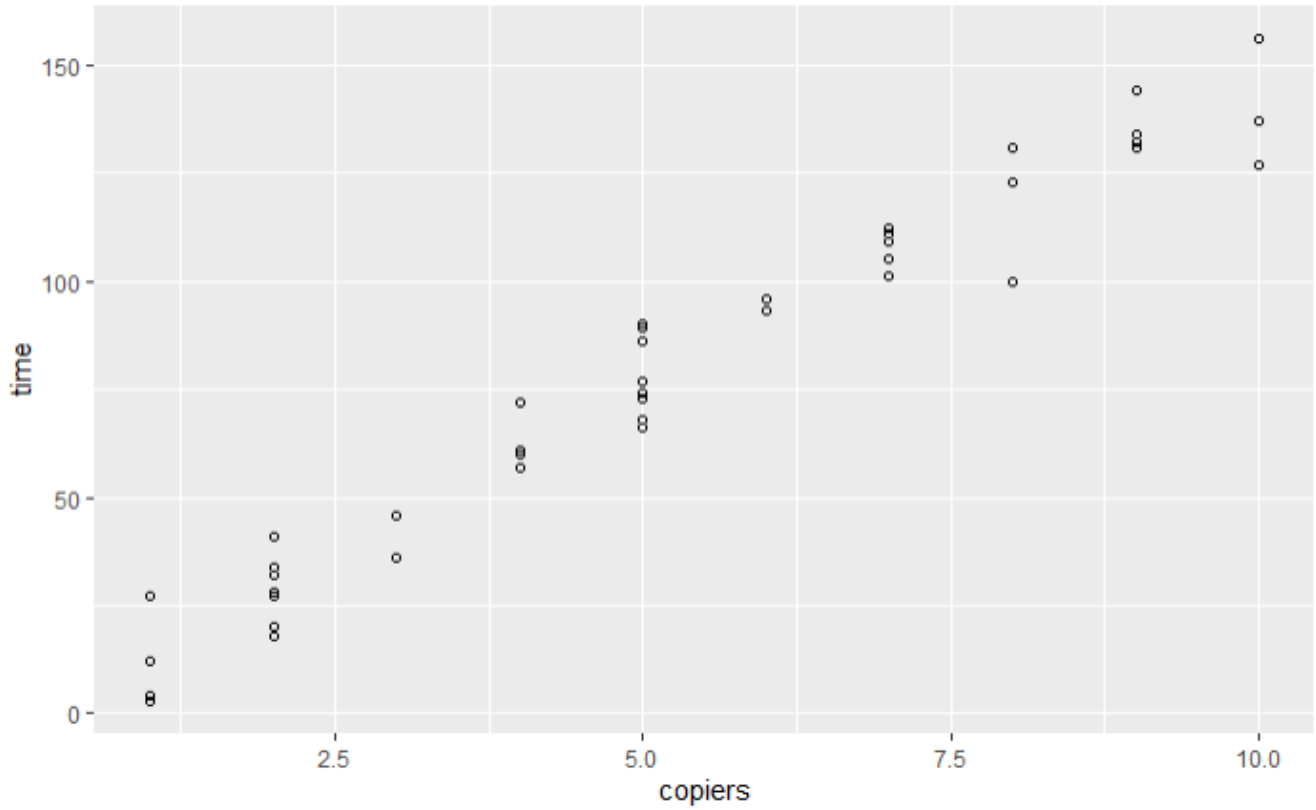
Zadanie 1

For this assignment you will use the data from the file ch01pr20.txt. Second column contains the number of copiers and the first column contains the time (in hours) needed to maintain these copiers.

Plot the data. Is the relationship approximately linear?



Możemy zauważyć wyraźną liniową relację pomiędzy liczbą kopiarek, a czasem.

Zadanie 2

Run the linear regression with y = service time and x = number of machines serviced.

(a) Give the estimated regression equation.

Chcę wyznaczyć β_1 oraz β_0 z równania regresji $Y = \beta_1 * X + \beta_0$

```
B1 = cov(data$copiers, data$time) / var(data$copiers)
B0 = mean(data$time) - mean(data$copiers) * B1
```

Poprawność powyższych obliczeń mogę sprawdzić, korzystając z funkcji *lm*:

```
regression = lm(time~copiers, data=data)
coef(regression)
```

Otrzymane parametry to: $\beta_1 = 15.0352480, \beta_0 = -0.5801567$

(b) Give a 95% confidence interval for the slope.

```
s2 = 1 / 43 * sum((data$time - B1 * data$copiers - B0)^2)
s2_B1 = s2 / sum((data$copiers - mean(data$copiers))^2)
upper_B1 = B1 + qt(1-0.05/2, 43) * sqrt(s2_B1)
lower_B1 = B1 - qt(1-0.05/2, 43) * sqrt(s2_B1)
sprintf("Lower: %f, Upper: %f", lower_B1, upper_B1)
```

```
[1] "Lower: 14.061010, Upper: 16.009486"
```

Z prawdopodobieństwem 95%, estymowany parametr β_1 znajduje się w przedziale $[14.061, 16.009]$.

(c) Describe the results of the significance test for the slope. Give the hypothesis being tested, the test statistic with degrees of freedom, the P-value, and your conclusion in a brief sentence.

Testuję: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
Liczba stopni swobody wynosi $45 - 2 = 43$.
Obliczę statystykę T:

```
T = B1 / sqrt(s2_B1)
```

Wynosi ona 31.12326.

```
p_value = pt(q=T, df=43, lower.tail=FALSE) + pt(q=-T, df=43, lower.tail=TRUE)
```

$Wartość\ p$ wynosi $4.009032e-31$, co jest mniejsze od 0.05, więc możemy odrzucić hipotezę H_0. Pokazuje to, że istnieje relacja pomiędzy liczbą kopiarek a czasem.

Powyższe obliczenia mogę zweryfikować, korzystając z funkcji *summary*:

```
summary(regression)
```

```
Call:
lm(formula = time ~ copiers, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.7723  -3.7371   0.3334   6.3334  15.4039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5802     2.8039  -0.207    0.837
copiers      15.0352     0.4831  31.123   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.914 on 43 degrees of freedom
Multiple R-squared:  0.9575,   Adjusted R-squared:  0.9565
F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Zadanie 3

Give an estimate of the mean service time that you would expect if 11 machines were serviced; and a 95% condence interval for this estimate.

Estymuję wartość $E(Y_{11})$:

```
x = 11
prediction = B0 + B1 * x

s2_x = s2 *
  (1/45 + (x - mean(data$copiers))^2/sum((data$copiers - mean(data$copiers))^2))
diff = qt(1-0.05/2, 43) * sqrt(s2_x)
lower = prediction - diff
upper = prediction + diff
sprintf("Prediction: %f, Lower: %f, Upper: %f", prediction, lower, upper)
```

```
[1] "Prediction: 164.807572, Lower: 158.475440, Upper: 171.139704"
```

Otrzymałem, że estymowania wartość $E(Y_{11})$ wynosi 164.80, natomiast rzeczywista wartość z prawdopodobieństwem 0.95 znajduje się w przedziale $[158.47, 171.13]$.

Zadanie 4

Give a prediction for the actual service time that you would expect if 11 machines were serviced; and 95% prediction interval for this time

Estymuję wartość Y_{11}.

```
s2_x = s2 *
  (1 + 1/45 + (x - mean(data$copiers))^2/sum((data$copiers - mean(data$copiers))^2))
diff = qt(1-0.05/2, 43) * sqrt(s2_x)
lower = prediction - diff
upper = prediction + diff
sprintf("Prediction: %f, Lower: %f, Upper: %f", prediction, lower, upper)
```

```
[1] "Prediction: 164.807572, Lower: 145.749099, Upper: 183.866044"
```

Ponownie, estymowana wartość wynosi 164.80. Przedział ufności jest tym razem większy, wynosi $[145.74, 183.86]$, co wynika z faktu, że wariancja błędu predykcji jest większa niż wariancja estymatora $E(Y_{11})$.

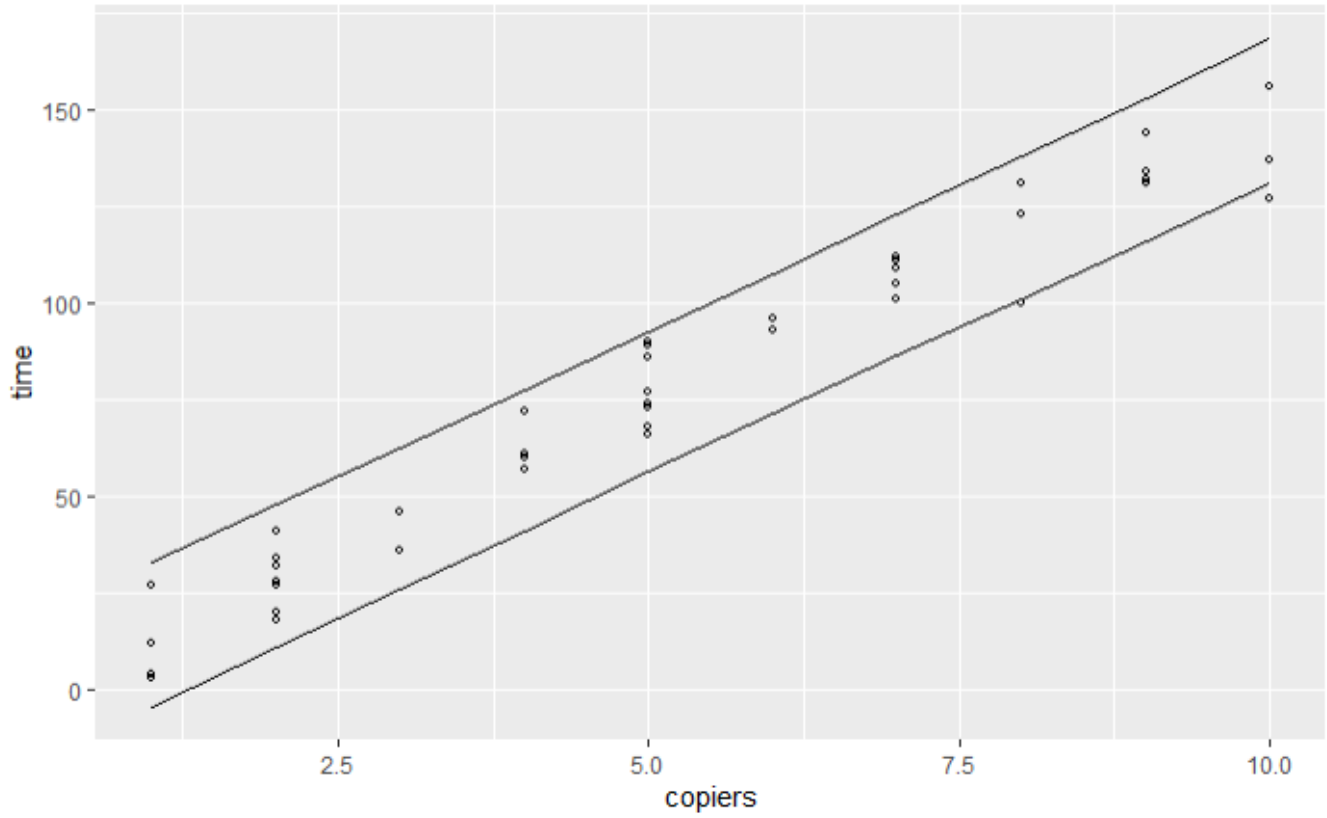
Zadanie 5

Plot the data with the 95% prediction bounds for individual observations.

```
options(warn=-1)
preds = predict(regression, interval='prediction')
xs = v[,1]
lower_xs = v[,2]
upper_xs = v[,3]

confidence_data = data.frame(data$copiers, data$time, lower_xs, xs, upper_xs)

ggplot(confidence_data) +
  geom_point(aes(data.copiers, data.time), shape = 1, size=1) +
  geom_line(aes(data.copiers, upper_xs)) +
  geom_line(aes(data.copiers, lower_xs)) +
  xlab('copiers') + ylab('time')
```



Możemy zauważyć, że 43 spośród 45 punktów znajduje się w obrębie przedziału predykcji.

Zadanie 6

Assume $n = 40$, $\sigma^2 = 120$, $SSX = \sum(Xi - \bar{X})^2 = 1000$.

```
n = 40
sigma2 = 120
ssx = 1000
```

(a) Find the power for rejecting the null hypothesis that the regression slope is zero using a $\alpha = 0.05$ *significance test when the true slope is* $\beta_1 = 1$.

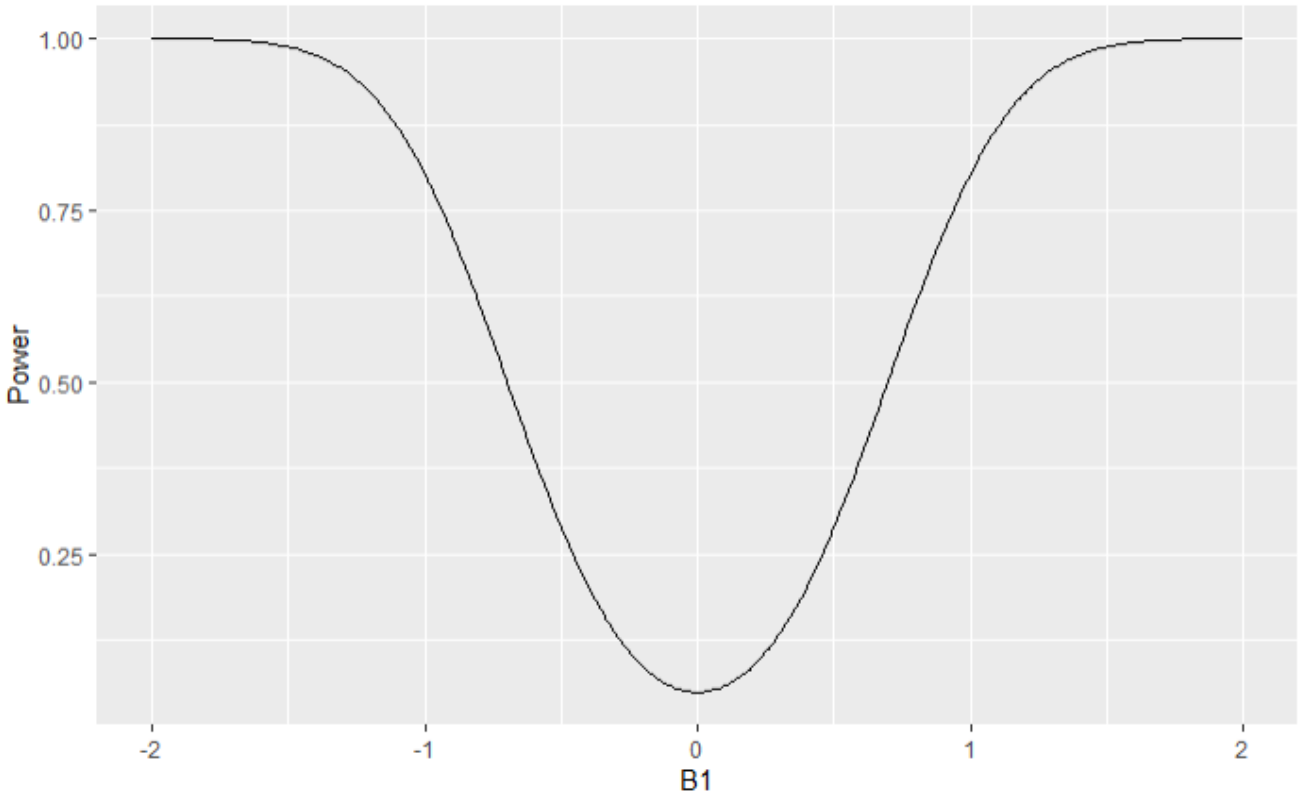
Testuję następujący problem: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 = 1$.
Moc jest równa :
$\pi(1) = P_{\beta_1=1}(\text{test odrzucił hipotezę } H_0) = P_{\beta_1=1}(|T| > t_c) = F_{\beta_1=1}(-t_c) + 1 - F_{\beta_1=1}(t_c)$,
gdzie $t_c = t^*(1 - \alpha/2, n - 2)$

Obliczam więc $\sigma^2(\hat{\beta}_1) = \sigma^2/SSX$ oraz parametr niecentralności $\delta = \beta_1/\sigma(\hat{\beta}_1)$

```
B1 = 1
alpha = 0.05
sigma2_B1 = sigma2 / ssx
delta = B1 / sqrt(sigma2_B1)
tc = qt(1-alpha/2, n-2)
hypothesis_power = pt(-tc, n-2, delta) + 1 - pt(tc, n-2, delta)
hypothesis_power
```

Otrzymuję moc równą 0.8032105.

(b) Plot the power as a function of β_1 for values of β_1 between −2 and 2.



Zgodnie z oczekiwaniami, możemy zauważyć, że prawdopodobieństwo odrzucenia hipotezy $H_0 : \beta_1 = 0$ maleje, gdy prawdziwa wartość nachylenia zbliża się do 0.

Zadanie 7

Generate the vector $X = (X1, \ldots, X_{200})^T$ from the multivariate normal distribution $N(0, \frac{1}{200} I)$ Then generate 1000 vectors Y from the model $Y = 5 + \beta 1X + \epsilon$.
For each replication of the experiment test the hypothesis that $\beta_1 = 0$ and estimate the probability of rejection by the frequency of rejections in your sample (separately for each of the points (a)-(d)). Compare these estimated probabilities to the theoretical probability of the type I error (a and b) and the theoretical power (c and d) calculated under the assumption that the noise has the normal distribution. Summarize the results.

Na początku wygeneruję wektor X:

```
X = rnorm(200, 0, sqrt(1/200))
```

Skorzystam z funkcji estimate_p, którą wywołam dla każdego z podpunktów a)-d). Wykonuje ona 2000 iteracji, gdzie każda iteracja to wygenerowanie wektora Y, a następnie sprawdzenie, z prawdopodobieństwem 95%, czy parametr β_1 nie jest równy 0, co umożliwi oszacowanie prawdopodobieństwa odrzucenia hipotezy, że $\beta_1 = 0$.
Funkcja ta, na potrzeby podpunktów c) i d), oblicza również moc testu dla odpowiednich β_1.

```
estimate_p = function(fun, X, B1=0, iters=2000){
  counter = 0
  powers_sum = 0
  for (i in 1:iters){
    Y = fun(X)
    data = data.frame(X, Y)
    model = lm(Y~X, data)
    confidence = confint(model, "X")
    if (confidence[1] > 0 || confidence[2] < 0){
      counter = counter + 1
    }

    # calculating power for c) and d)
    s = sd(model$residuals) * sqrt(199/198)
    sigma_B1 = s/sum((X - mean(X))^2)
    delta = B1/sigma_B1
    tc = qt(1-0.05/2, 198)
    powers_sum = powers_sum + pt(-tc, 198, delta) + 1 - pt(tc, 198, delta)
  }
  return (c(counter/iters, powers_sum/iters))
}
```

a) $\beta_1 = 0, \epsilon \sim N(0, I)$

```
case_a = function(X){
  return (sapply(X, function(x) 5 + rnorm(1, 0, 1)))
}

estimate_p(case_a, X, 1)[1]
```

Otrzymujemy, że szacowane prawdopodobieństwo odrzucenia hipotezy wynosi:

```
[1] 0.047
```

co jest bardzo bliskie teoretycznej wartości błędu pierwszego rodzaju, czyli 5%.

b) $\beta_1 = 0, \epsilon_1, \dots \epsilon_{200} - \ iid\ from\ the\ exponential\ distribution\ with\ \lambda = 1$

```
case_b = function(X){
  return (sapply(X, function(x) 5 + rexp(1, 1)))
}

estimate_p(case_b, X)[1]
```

Otrzymujemy, że szacowane prawdopodobieństwo odrzucenia hipotezy wynosi:

```
[1] 0.0395
```

Ponownie, otrzymaliśmy wartość bliską 5%.

c) $\beta_1 = 1.5, \epsilon \sim N(0, I)$

```
case_c = function(X){
  return (sapply(X, function(x) 5 + 1.5 * x + rnorm(1, 0, 1)))
}

estimate_p(case_c, X, B1=1.5)
```

Otrzymujemy dwie wartości:

```
[1] 0.3485000 0.3335263
```

0.3485000 to szacowane prawdopodobieństwo odrzucenia hipotezy.
0.3335263 to teoretyczna wartość mocy testu dla $\beta_1 = 1.5$. Widać, że obie wartości są podobne.

d) $\beta_1 = 1.5, \epsilon_1, \dots \epsilon_{200} - \ iid\ from\ the\ exponential\ distribution\ with\ \lambda = 1$

```
case_c = function(X){
  return (sapply(X, function(x) 5 + 1.5 * x + rexp(1, 1)))
}

estimate_p(case_c, X, B1=1.5)
```

Otrzymujemy:

```
[1] 0.3305000 0.3431282
```

Ponownie zauważamy, że szacowane prawdopodobieństwo odrzucenia hipotezy (0.3305) jest bliskie teoretycznej wartości mocy testu (0.3431282).

Zadanie 8

You use $n = 20$ observations to t the linear model $Y = \beta 0 + \beta 1X + \epsilon$
Your estimators are $b_0 = 1$, $b_1 = 3$ and $s = 4.0$.
a) The estimated standard deviation of $b1$, $s(b_1)$, is equal to 1. Construct the 95 % confidence interval for β_1.
b) Do you have statistical evidence to believe that Y depends on X?
c) The 95% condence interval for $E(Y)$ when $X = 5$ is $[13, 19]$. Find the corresponding prediction interval.

a)
Skonstruujmy 95% procentowy przedział ufności dla β_1:
$b_1 + t_c s(b_1) = 3 + 2.1 * 1 = 5.1$
$b_1 - t_c s(b_1) = 3 - 2.1 * 1 = 0.9$

b)
Aby sprawdzić, czy Y zależy od X, przetestuję hipotezę:

$$H_0 : \beta_1 = 0$$

$$T = b_1 / s(b_1) = 3.$$
$$|T| > t_c = 2.1$$

więc mogę odrzucić hipotezę H_0, co wskazuje, że istnieje zależność pomiędzy X a Y.

c)
Przedział ufności jest dany wzorem: $\hat{\mu}_5 \pm t_c s(\hat{\mu}_5)$
Przedział predykcji jest dany wzorem: $\hat{\mu}_5 \pm t_c s(pred)$
Wiemy, że 95% procentowy przedział ufności dla $E(Y)$, gdy $X = 5$, wynosi $[13, 19]$.
Możemy zauważyć, że $\hat{\mu}_5 = 16$. Z tego wynika, że $s(\hat{\mu}_5) = (19 - 16)/t_c$, co w przybliżeniu wynosi 1.428.
Wiedząc, że $s^2(pred) = s^2(\hat{\mu}_5) + s^2$, możemy wyznaczyć:

$$s^2(pred) \approx 2.04 + 16 = 18.04$$
$$lower \approx 16 - 2.1 * \sqrt{18.04} = 7.08056$$
$$upper \approx 16 + 2.1 * \sqrt{18.04} = 24.9194$$

Otrzymujemy, że przedział predykcji w przybliżeniu wynosi $[7.08, 24.92]$