Linear models list 2

Klaudia Weigel

1 Introduction

The objective of this report is performing a linear regression analysis on a dataset containing a response variable and a predictor variable.

Let's assume that we have n observations. A linear regression model with a response vector Y and one predictor variable X is defined as follows

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where Y_i and is the response in the ith trial, β_0 and β_1 are the regression parameters we want to approximate, X_i is a known value of the predictor variable in the ith trial, ϵ_i is a random error from the normal distribution $N(0, \sigma^2)$ (errors are independent). It follows that $E(Y_i) = \beta_0 + \beta_1 X_i$ and $Var(Y_i) = \sigma^2$.

We can approximate the values of β_0 and β_1 from the equation

$$(\hat{\beta}_0, \hat{\beta}_1)^T = \underset{(b_0,b_1)^T \in \mathbb{R}^2}{argmin} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2.$$

The expression under the sum is a squared distance between the real value of the ith observation and the value estimated by our linear model. Solving the equation by taking partial derivatives and equating them to zero gives the following estimators

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Using this method we cannot obtain the estimator for σ^2, to do that we will calculate a maximum likelihood estimator. The likelihood function for the response vector is

$$L(Y; b_0, b_1, z^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi z^2}} exp(-\frac{1}{2z^2}(Y_i - b_0 - b_1 X_i)^2).$$

By maximizing the log-likelihood function, we can obtain the estimators for β_0, β_1 and σ^2. It turns out that the estimators for β_0 and β_1 are the same as the ones obtained using the ols method. The unbiased estimator for σ^2 is

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Properties of the estimators:

- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$
- $\hat{\beta}_1 \sim N(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right))$

1.1 Hypothesis testing

1.1.1 T-test for β_0

Let us consider the following testing problem

$$H_0 : \beta_0 = a \quad \text{against} \quad H_1 : \beta_0 \neq a.$$

The test statistic has the form

$$T = \frac{\hat{\beta}_0 - a}{s(\hat{\beta}_0)} \quad \text{where} \quad s^2(\hat{\beta}_0) = s^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}).$$

T has the t-distribution with n-2 degrees of freedom. The critical region for this test is $|T| > t_c$, where t_c is obtained from $P_{H_0}(|T| > c) = 2 * (1 - P_{H_0}(T < t_c)) = \alpha$, with α being the chosen significance level. Therefore t_c is the $(1 - \alpha/2)$-th quantile from the $t(n-2)$ distribution.
The power of a statistical test is the probability of rejecting the null hypothesis when it is not true. In our case the test statistic under the alternative hypothesis has the noncentral t-distribution with $n-2$ degrees of freedom and the noncentrality parameter $\delta = \beta_0/\sigma(\hat{\beta}_0)$. Therefore to obtain the power of the test, we need to calculate

$$\pi(\alpha) = P_{H_1}(|T| > t_c) = P_{H_1}(T < -t_c) + P_{H_1}(T > t_c).$$

1.1.2 T-test for β_1

Now let us consider the testing problem

$$H_0 : \beta_1 = a \quad \text{against} \quad H_1 : \beta_1 \neq a.$$

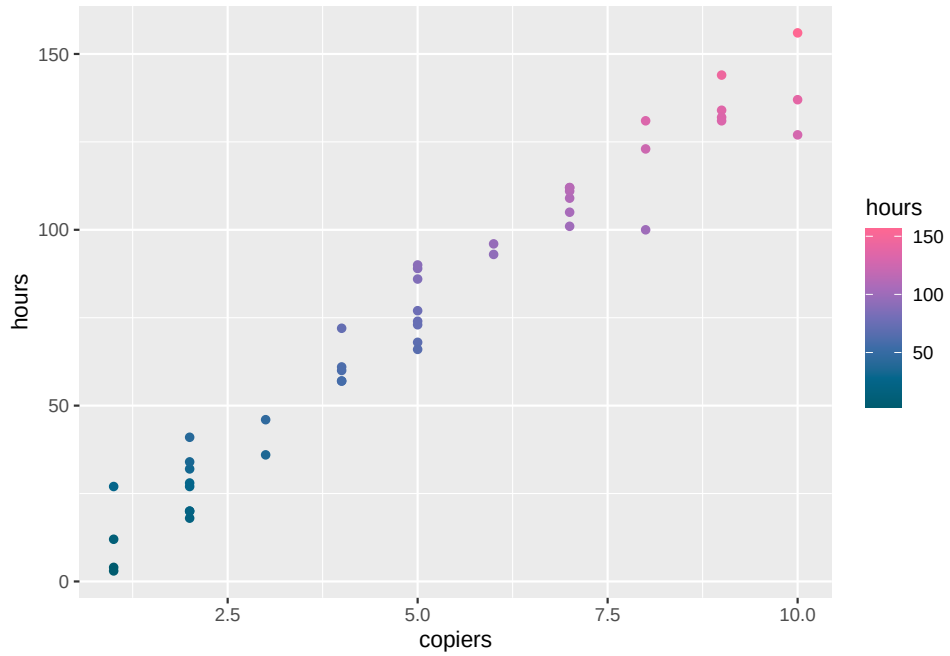
The test statistic is

$$T = \frac{\hat{\beta}_1 - a}{s(\hat{\beta}_1)} \sim t(n-2) \quad \text{where} \quad s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Similarly as in the case of testing β_0 we reject the null hypothesis when $|T| > t_c = t(1 - \frac{\alpha}{2}, n-2)$. The power of the test is computed the same way as for β_0.

2 Exercise 1

```
df = read.table("CH01PR20.txt", col.names = c("hours", "copiers"))
df = data.frame(df)
ggplot(data = df, aes(x=copiers, y=hours, color = hours)) +
  geom_point() +
  scale_color_gradientn(colors = space_palette)
```



We see that the relationship between the number of machines and service time is approximately linear.

3 Exercise 2

We want to find a linear regression equation to predict the time needed to maintain the copiers, based on the number of these copiers. Our theoretical model is

$$Time_i = \beta_0 + \beta_1 Number_i \ + \epsilon_i \qquad i = 1, ..., 45 = n$$

a)

We can approximate the regression coefficients β_0 and β_1 from

$$\hat{\beta}_0 = \overline{Time} - \hat{\beta}_1 \overline{Number}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{45}(Number_i - \overline{Number})(Time_i - \overline{Time})}{\sum_{i=1}^{45}(Number_i - \overline{Number})^2}.$$

To obtain the parameters of the model we will use a built in function `lm`.

```
reg = lm(hours~copiers, data = df)
reg$coefficients
```

```
## (Intercept)     copiers
##  -0.5801567  15.0352480
```

```
sigma(reg)
```

```
## [1] 8.913508
```

We can compare the results with self-written formulas.

```
# Y = Time, X = Number
Y = df[,1]; X = df[,2]; n = length(Y)
beta_1 = sum((X-mean(X)) %*% (Y - mean(Y)))/sum((X-mean(X))^2)
```

```
beta_0 = mean(Y) - beta_1*mean(X)
sigma_squared = (1/(n-2))*sum((Y - beta_0 - beta_1*X)^2)
beta_0; beta_1; sqrt(sigma_squared)
```

```
## [1] -0.5801567
```

```
## [1] 15.03525
```

```
## [1] 8.913508
```

We see that in both cases we obtained the same values. Finally our regression function is

$$E(Time) = -0.58 + 15.035 Number.$$

b)
To construct the confidence interval for β_1 we use $T = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t(n-2)$. The $1-\alpha$ confidence limits for β_1 are

$$\hat{\beta}_1 \pm t(1 - \frac{\alpha}{2}, n-2) s(\hat{\beta}_1)$$

where $t(1-\frac{\alpha}{2}, n-2)$ is $1-\frac{\alpha}{2}$ -th quantile from the $t(n-2)$ distribution.
In R we can compute the confidence interval using the built in function `confint`.

```
confint(reg)[2,]
```

```
##    2.5 %   97.5 %
## 14.06101 16.00949
```

Let's compare the results with a self-written formulas.

```
alpha = 0.05
s_2_beta_1 = sigma_squared/sum((X-mean(X))^2)
conf_beta_1 = beta_1 + c(-qt(1-alpha/2, df = n-2), qt(1-alpha/2, df = n-2))*sqrt(s_2_beta_1)
conf_beta_1
```

```
## [1] 14.06101 16.00949
```

c) We test the following hypothesis

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0.$$

The test statistic is

$$T = \frac{\hat{\beta}_1 - 0}{s(\hat{\beta}_1)} \sim t(n-2) = t(43) \quad \text{where} \quad s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n} (Number_i - \overline{Number})^2}.$$

The critical region is of the form $C = \{T : |T| > t_c\}$. The p-value is $P(|T| > |t|)$, where t is the observed value of T.

```
T = beta_1/sqrt(sigma_squared/sum((X - mean(X))^2))
T; 2*(1 - pt(abs(T), df=n-2))
```

```
## [1] 31.12326
```

```
## [1] 0
```

We can compare obtained values with the output of `summary` in R.

```
summary(reg)
```

```
##
## Call:
## lm(formula = hours ~ copiers, data = df)
```

```
##
## 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
```

The value of the T statistic is 31.123 and the p-value is equal to zero. This means that we reject the null hypothesis and conclude that the service time depends on the number of machines.

4 Exercise 3

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

Let the estimate of mean be $\hat{\mu_h}$, in our case $h = 11$. Then $\hat{\mu_h} \sim N(\mu_h, \sigma^2(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}))$. We can define a T statistic similarly as for β_0 and β_1:

$$T = \frac{\hat{\mu_h} - E(\hat{\mu_h})}{s(\hat{\mu_h})} \sim t(n-2) \quad \text{where } s^2(\hat{\mu_h}) = s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).$$

The $1 - \alpha$ confidence interval for the mean is

$$\hat{\mu_h} \pm t(1 - \frac{\alpha}{2}, n-2)s(\hat{\mu_h})$$

We will calculate the mean estimate and the 95% confidence interval using a built in function `predict` and compare the results.

```
predict(reg, data.frame(copiers = c(11)), interval='confidence')
```

```
##        fit      lwr      upr
## 1 164.8076 158.4754 171.1397
```

```
mu_11_pred = beta_0 + 11*beta_1
X_h = 11
s_2_mu_11 = sigma_squared*(1/n + (X_h -mean(X))^2/sum((X-mean(X))^2))
quantile = qt(1-alpha/2, df = n-2)
conf_mu_11 = mu_11_pred + c(-quantile, quantile)*sqrt(s_2_mu_11)
mu_11_pred; conf_mu_11
```

```
## [1] 164.8076
```

```
## [1] 158.4754 171.1397
```

5 Exercise 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.

The actual value for Y_h is equal to $\beta_0 + \beta_1 X_h + \epsilon_h$. The prediction will remain the same as in the previous exercise ($\hat{Y_h} = \hat{\mu_h} = \hat{\beta_0} + \hat{\beta_1} X_h$), what will change however are the prediction bounds, as we have a different variance.

$$Var(Y_h - \hat{Y_h}) = Var(Y_h - \hat{\mu_h}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

```
predict(reg, data.frame(copiers = c(11)), interval='prediction')
```

```
##        fit      lwr     upr
## 1 164.8076 145.7491 183.866
```

```
s_2_pred = sigma_squared*(1 + 1/n + (X_h -mean(X))^2/sum((X-mean(X))^2))
conf_pred = mu_11_pred + c(-quantile, quantile)*sqrt(s_2_pred)
conf_pred
```

```
## [1] 145.7491 183.8660
```

We can observe that the bounds for the prediction interval are wider than the ones for the confidence interval obtained in the previous exercise. This is an immediate result of the fact that the variance of the prediction error is larger than the variance of $\hat{\mu_h}$.

6 Exercise 5

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

Let us consider a problem of simultaneous estimation of several $E(Y_h)$ at several levels of X_h. A confidence band for the entire regression line at a confidence level $1 - \alpha$ has the following bounds:

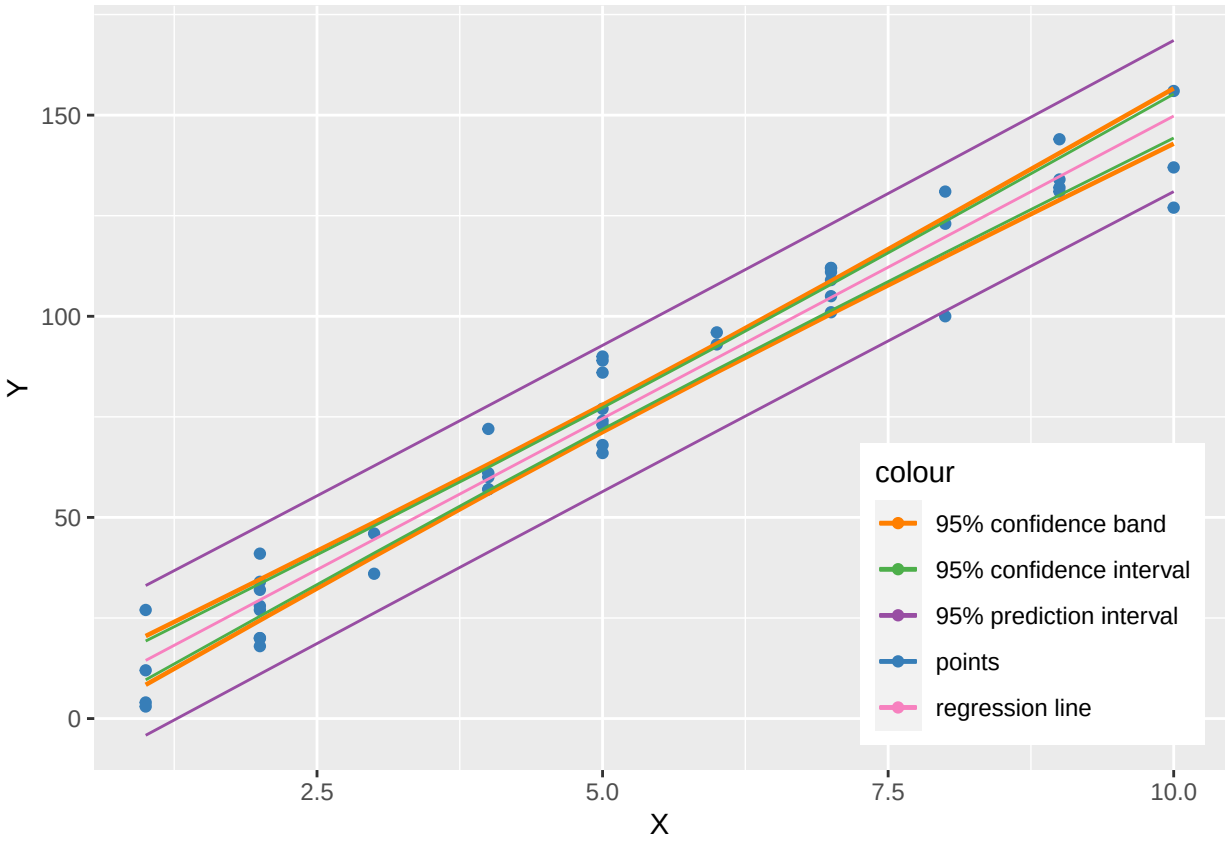
$$\hat{Y_h} \pm W s(\hat{\mu_h}),$$

where $W^2 = 2F(1 - \alpha, 2, n - 2)$ is a double quantile from the Fisher-Snedecor distribution with 2 and n-2 degrees of freedom. We will also plot the confidence and prediction intervals obtained from pointwise calculation.

```
m2 = X*beta_1 + beta_0 # = predict(reg)
# = predict(reg, se.fit = TRUE)$se.fit
s2 = sqrt(sigma_squared*(1/n + (X - mean(X))^2/sum((X - mean(X))^2)))
w = sqrt(2*qf(1-alpha,2,n-2))
up = m2 + w*s2; down = m2 - w*s2
dat = data.frame(X, Y, m2, up, down)
pred = predict(reg, interval = "prediction")
conf = predict(reg, interval = "confidence")
dat = cbind(dat, pred[,2:3], conf[,2:3])
colnames(dat) = c("X", "Y", "fit", "up", "down", "lwr_pred", "upr_pred", "lwr_conf", "upr_conf")
```

```
colors = c(palette("Set1"))
colors = c(palette("Set1"))
ggplot(data = dat, aes(x = X, y = Y)) +
  geom_point(aes(color = "points")) +
  geom_line(aes(y=up, color = "95% confidence band"), size = 0.8) +
  geom_line(aes(y=down, color = "95% confidence band"), size = 0.8) +
```

```
  geom_line(aes(y=lwr_pred, color = "95% prediction interval"), size = 0.5) +
  geom_line(aes(y=upr_pred, color = "95% prediction interval"), size = 0.5) +
  geom_line(aes(y=lwr_conf, color = "95% confidence interval"), size = 0.5) +
  geom_line(aes(y=upr_conf, color = "95% confidence interval"),  size = 0.5) +
  geom_line(aes(y = fit, color = "regression line")) +
  scale_color_manual(values = c("points" = colors[2], "95% confidence band" = colors[5],
                                "95% prediction interval" = colors[4],
                                "95% confidence interval" = colors[3],
                                "regression line" = colors[8])) +
  theme(legend.position= c(0.83, 0.23))
```



The boundary points of all the intervals are wider apart the further we get from the sample mean. The W term will be larger than the t term used for calculating pointwise intervals in two previous exercises. This follows from the fact that the confidence band must encompass the entire regression line, whereas the confidence limits for $E(Y_h)$ at X_h apply only at the single level X_h.

7 Exercise 6

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

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

We want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The test statistic is

$$T = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t(n-2).$$

The critical region has the form $|T| > t_c$, where t_c is $(1 - \frac{\alpha}{2})$th quantile from the t-distribution with n-2 degrees of freedom.
The power of a statistical test is the probability of rejecting the null hypothesis when it is not true, namely

$$\pi(\alpha) = P_{H_1}(|T| > t_c)$$

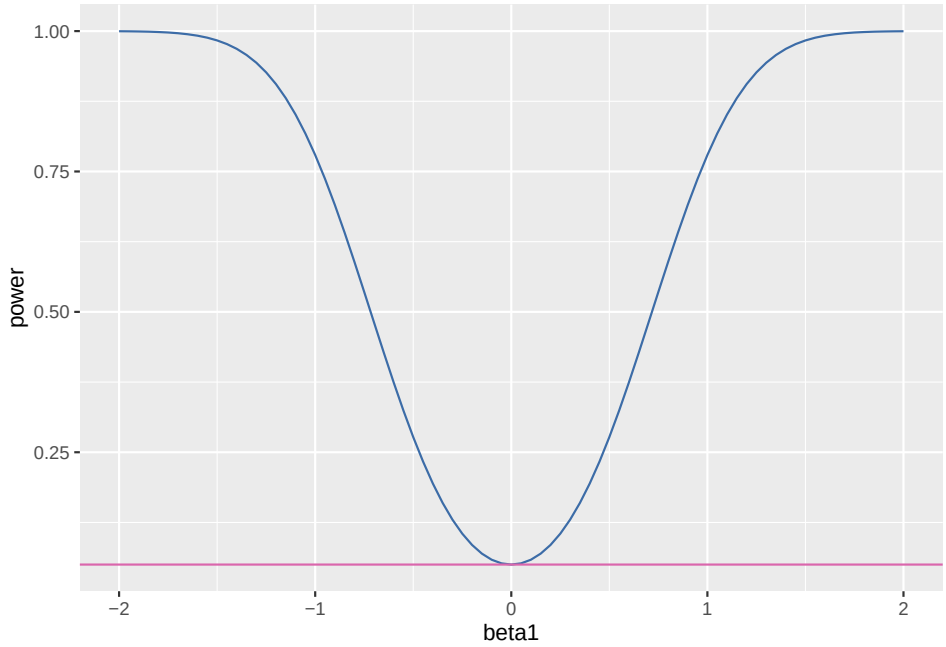
Under H_1 the test statistic has the noncentral t-distribution with a noncentrality parameter being equal to $\delta = \beta_1/\sigma(\hat{\beta}_1)$.

```
n = 20; sig2 = 120; ssx = 1000
sig2b1<-sig2/ssx
df=n-2
tc<-qt(1-alpha/2,df)
beta1 = 1
delta = beta1/sqrt(sig2b1)
prob1 = function(delta){pt(tc,df,delta)}
prob2 = function(delta){pt(-tc,df,delta)}
power = 1-prob1(delta)+prob2(delta)
power
```

```
## [1] 0.7793586
```

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

```
beta1 = seq(from=-2.0, to= 2.0, by= .05)
delta = beta1/sqrt(sig2b1)
power = 1-prob1(delta)+prob2(delta)
ggplot(data = data.frame(beta1, power), aes(x=beta1, y=power)) +
  geom_line(color = space_palette[3]) +
  geom_hline(yintercept = 0.05, color = space_palette[6])
```



As the true value of β_1 gets closer to zero, the power gets smaller and bounded from below by the significance level. That is the expected result, since when the true value is close to zero we don't want to reject the null hypothesis.

8 Exercise 7

Generate a vector $X = (X_1, ..., X_{200})^T$ *from the multivariate normal distribution* $N(0, \frac{1}{200}I)$. *Then generate 1000 vectors* Y *from the model* $Y = 5 + \beta_1 X + \epsilon$, *where*

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

```
library(mvtnorm)
X = rmvnorm(1, mean = rep(0, 200), sigma = 1/200 * diag(200)); num_vectors = 1000

make_y = function(x, beta) {
  epsilon = rmvnorm(1, mean = rep(0, 200), sigma = diag(200))
  Y = 5 + beta*X + epsilon
  Y
}

make_y_b = function(x, beta) {
  epsilon = rexp(200, 1)
  Y = 5 + beta*X + epsilon
  Y
}

test = function(Y, X) {
  alpha = 0.05; n = length(X)
  s_squared = sum((Y - mean(Y))^2)/(n-2)
  beta_1_est = sum((X-mean(X)) %*% t((Y - mean(Y))))/sum((X-mean(X))^2)
  t = beta_1_est/sqrt((s_squared/sum((X-mean(X))^2))) # test statistic
  # calculate p-value, return 1 if we reject the null hypothesis
  return(2*(1 - pt(abs(t), df = n-2)) < alpha)
}

Y_a = lapply(1:num_vectors, make_y, beta = 0)
test_a = lapply(Y_a, test, X)
sum(unlist(test_a))/length(test_a)
```

```
## [1] 0.043
```

The empirical probabilities of type I error from multiple simulations are 0.052, 0.043. 0.05, 0.056, 0.053, 0.046, 0.062, 0.059, 0.042. All the values are close to the theoretical value equal to 0.05.

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

```
Y_b = lapply(1:num_vectors, make_y_b, beta = 0)
test_b = lapply(Y_b, test, X)
sum(unlist(test_b))/length(test_b)
```

```
## [1] 0.044
```

Values obtained from multiple simulations are 0.04, 0.049, 0.045, 0.054, 0.039, 0.048, 0.044, 0.056. Again the results are close to the theoretical values.

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

```
power = function(beta1) {
  n = 1000; sig2 = 1; ssx = sum((X-mean(X))^2); sig2b1<-sig2/ssx; df=n-2
  tc = qt(1-alpha/2,df)
  df = n-2
  tc = qt(1-alpha/2,df)
```

```
  delta = beta1/sqrt(sig2b1)
  prob1 = function(delta){pt(tc,df,delta)}
  prob2 = function(delta){pt(-tc,df,delta)}
  power = 1-prob1(delta)+prob2(delta)
  power
}


Y_c = lapply(1:num_vectors, make_y, beta = 1.5)
test_c = lapply(Y_c, test, X)
sum(unlist(test_c))/length(test_c)
```

```
## [1] 0.269
```

```
power(1.5)
```

```
## [1] 0.2907105
```

The empirically calculated power is close to the theoretical value.

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

```
Y_d = lapply(1:num_vectors, make_y_b, beta = 1.5)
test_c = lapply(Y_c, test, X)
sum(unlist(test_c))/length(test_c)
```

```
## [1] 0.269
```

```
power(1.5)
```

```
## [1] 0.2907105
```

In this case the two values are also close.

9 Exercise 8

You use n=20 observations to fit the linear model $Y = \beta_0 + \beta_1 X + \epsilon$. *Your estimators are* $\hat{\beta}_0 = 1, \hat{\beta}_1 = 3$ *and* $s = 4$.

a) *The estimated standard deviation of* $\hat{\beta}_1, s(\hat{\beta}_1)$, *is equal to 1. Construct the 95% confidence interval for* β_1.

The confidence interval is $\hat{\beta}_1 \pm t_c s(\hat{\beta}_1)$, where t_c is the $(1 - \frac{\alpha}{2})$-th ($\alpha = 0.05$) quantile from the t(n-2) distribution.

```
alpha = 0.05
tc = qt(1-alpha/2, df = 18)
tc
```

```
## [1] 2.100922
```

```
3 + c(-tc, tc)
```

```
## [1] 0.899078 5.100922
```

b) *Do you have statistical evidence to believe that Y depends on X?*

We will test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The test statistic $T = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{3}{1} > tc = 2.1$. We reject the null hypothesis and therefore have a reason to think that Y depends on X.

c) *The 95% confidence interval for E(Y) when X = 5 is [13,19]. Find the corresponding prediction interval.*

The 95% prediction interval is $\hat{\mu_h} \pm t_c s(pred)$, $\hat{\mu_h} = 1 + 3 * 5 = 16$. We know that $\hat{\mu_h} \pm t_c s(\hat{\mu_h}) = [13, 19]$, then $s(\hat{\mu_h}) = 3/2.1 = 1.428$. Now $s^2(pred) = s^2 + s^2(\hat{\mu_h}) = 16 + 2.039 = 18.039$, then $s(pred) = 4.247$. Finally the prediction interval is $16 \pm 2.1 * 4.247 = [7, 24.92]$.