Universidad de los Andes
(2023)

ARQUITECTURA DE SOFTWARE

Grupo 37

Camilo Duque - 202024289
Natalia Quintero - 201720156
Esteban Guerra - 201912735
Melissa Bayona - 201618981

CONTEXTUALIZACIÓN

Una pequeña empresa enfrenta problemas de disponibilidad en su aplicación, pero carece del presupuesto necesario para replicar el sistema. Se propone abordar la situación mediante modelos predictivos, específicamente la regresión lineal.

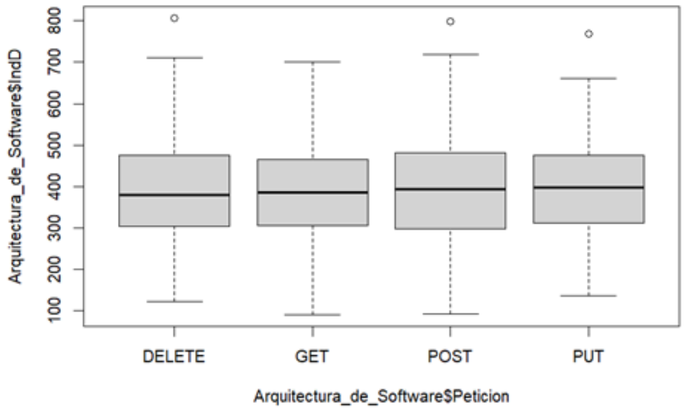
OBJETIVO GENERAL:

Mejorar la disponibilidad de la aplicación de software utilizando un modelo de regresión lineal para prever el índice de indisponibilidad.

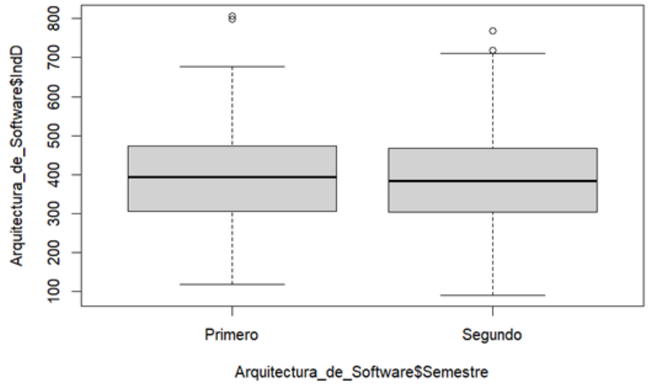
OBJETIVOS ESPECÍFICOS:

- Evaluar la relevancia de los datos en la bitácora digital.
- Desarrollar un modelo de regresión lineal que identifique las variables más significativas y sus contribuciones al problema de indisponibilidad.
- Validar el modelo y proporcionar recomendaciones para mejorar la disponibilidad.

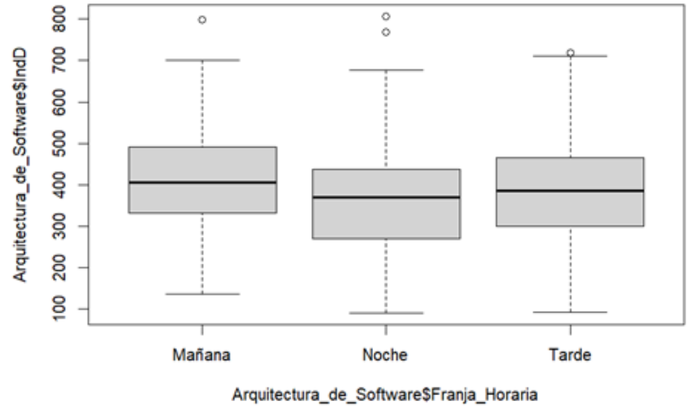
EXPLORACIÓN



Petición	Count of IndD	Average of IndD	StdDev of IndD
DELETE	108	395.037037	130.6074182
GET	450	386.2733333	116.5843069
POST	342	394.3888889	126.0839835
PUT	227	395.7180617	118.8211632
Total general	**1127**	**391.4782609**	**121.2730256**



Semestre	Count of IndD	Average of IndD	StdDev of IndD
Primero	570	393.4473684	118.7153773
Segundo	557	389.4631957	123.910101
(blank)			
Grand Total	**1127**	**391.4782609**	**121.2730256**

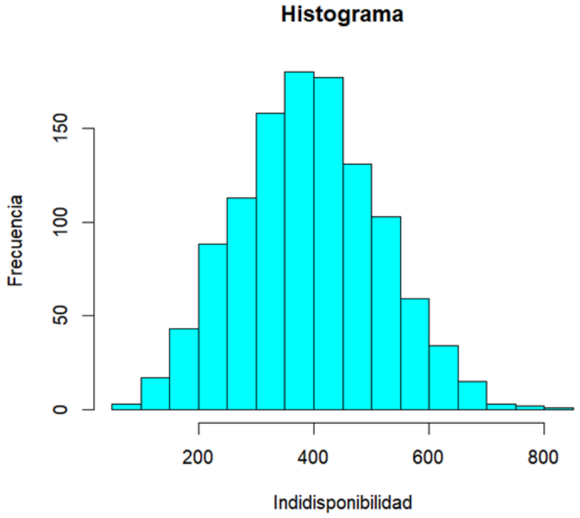


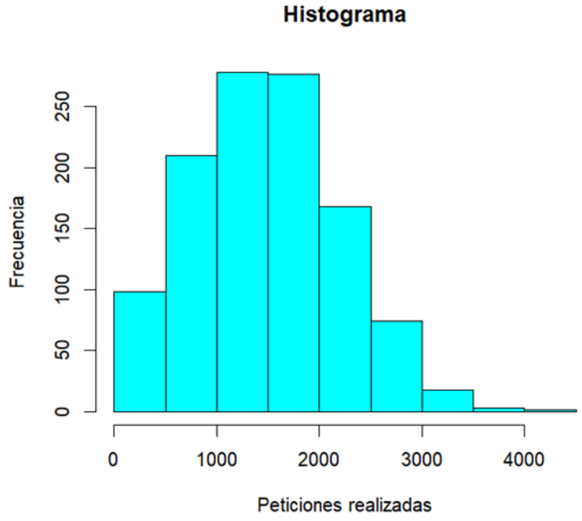
Franja Horaria	Count of IndD	Average of IndD	StdDev of IndD
Mañana	447	410.1946309	115.0003711
Noche	224	365.3973214	129.5387754
Tarde	456	385.9429825	120.4107304
(blank)			
Grand Total	**1127**	**391.4782609**	**121.2730256**

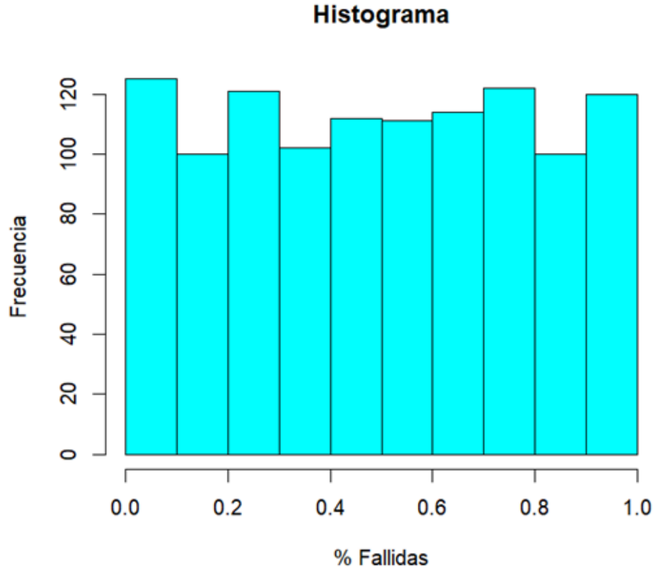
EXPLORACIÓN

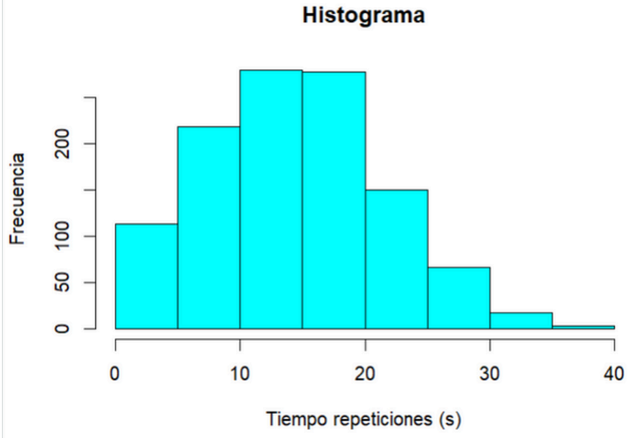
	CPR	*RampUp (seg)*	*%Error*	*Media tiempo respuesta*	*Desvest del tiempo*	*Latencia Maxima*	*Utilizacion promedio CPU*	*Indisponibilidad*
Media	1477.262	14.774	0.499	199.792	115.100	598.426	0.010	391.478
Error estandar	21.379	0.214	0.009	0.706	2.358	3.682	0.000	3.612
Mediana	1459.000	15.000	0.504	199.944	105.930	591.027	0.010	390.000
Desviacion Estandar	717.699	7.188	0.291	23.696	79.150	123.623	0.002	121.273
Minimo	11.000	0.000	0.001	115.159	0.333	252.055	0.002	91.000
Maximo	4021.000	40.000	0.999	309.797	433.857	1075.651	0.016	806.000
Suma	1664874.000	16650.000	562.465	225165.247	129718.057	674425.697	11.278	441196.000

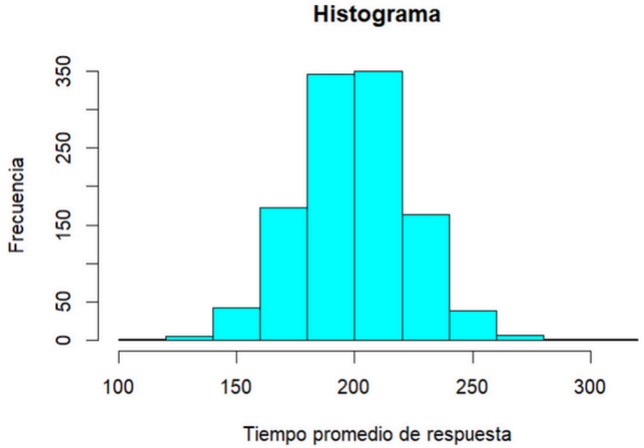
EXPLORACIÓN

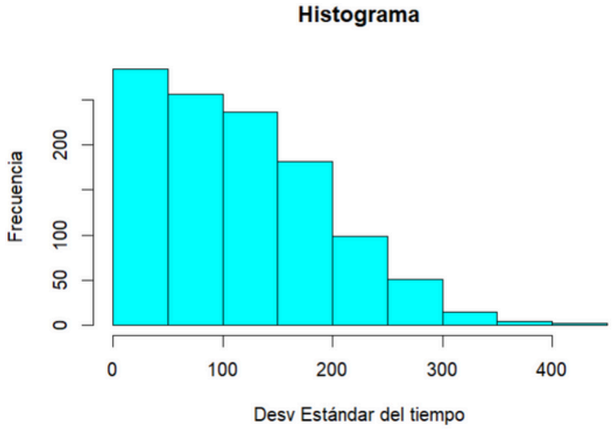




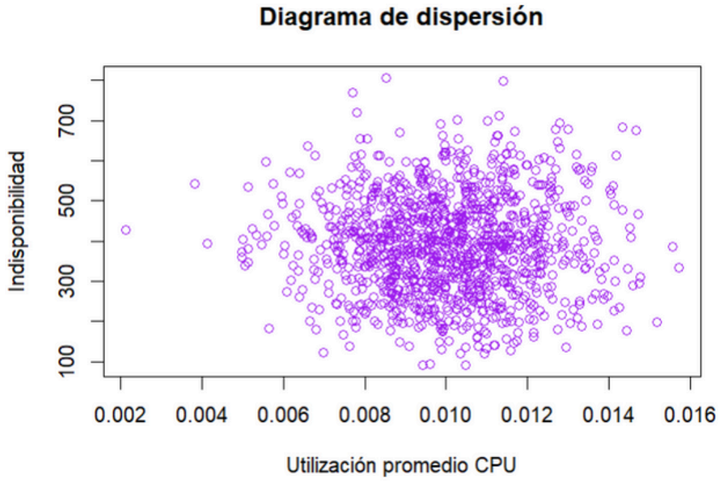


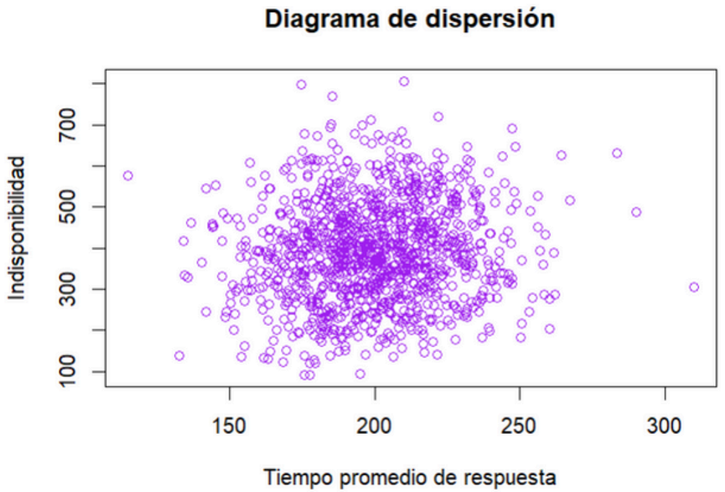


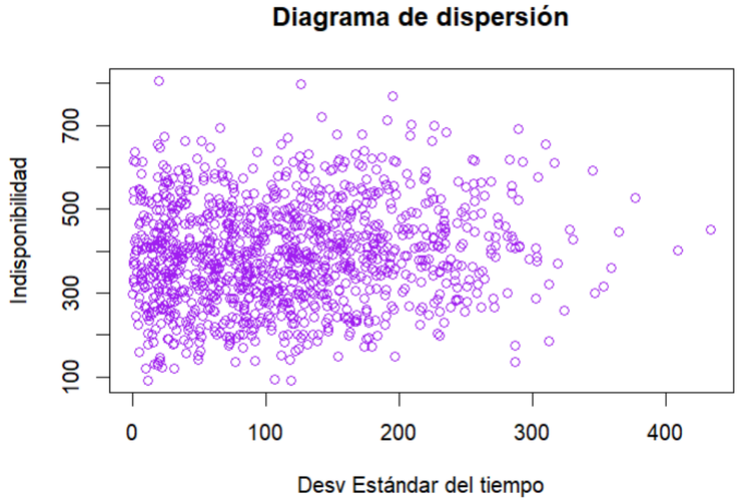


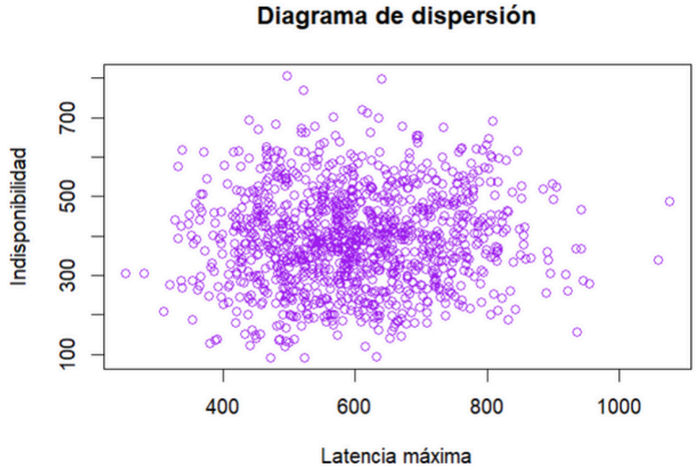


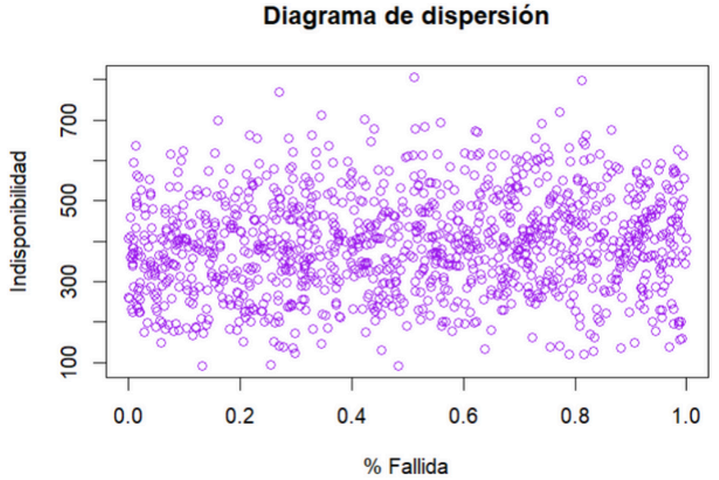
EXPLORACIÓN



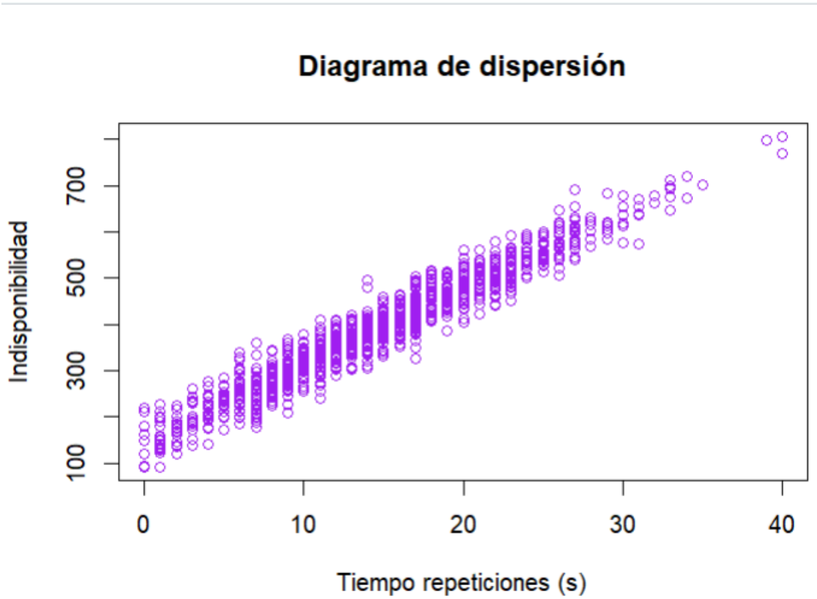


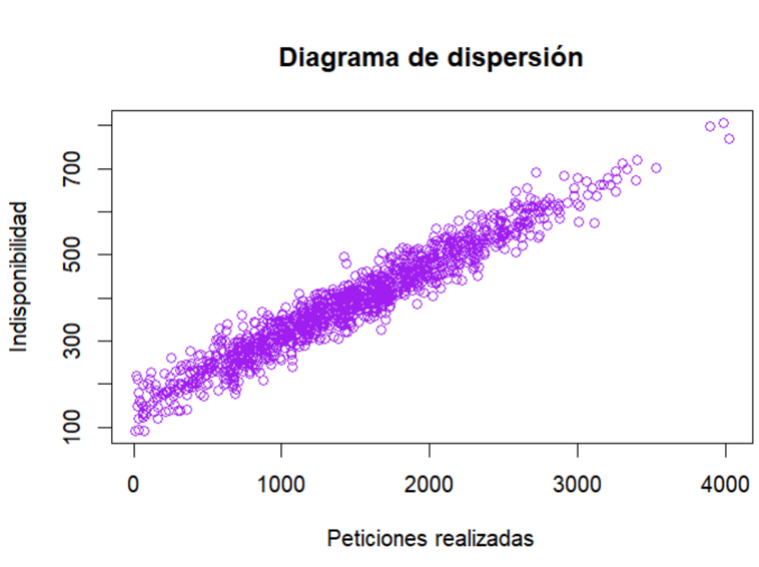






EXPLORACIÓN





Es probable que el modelo tenga problemas de multicolinealidad debido a la relación de dependencia entre estas dos variables

STEP WISE

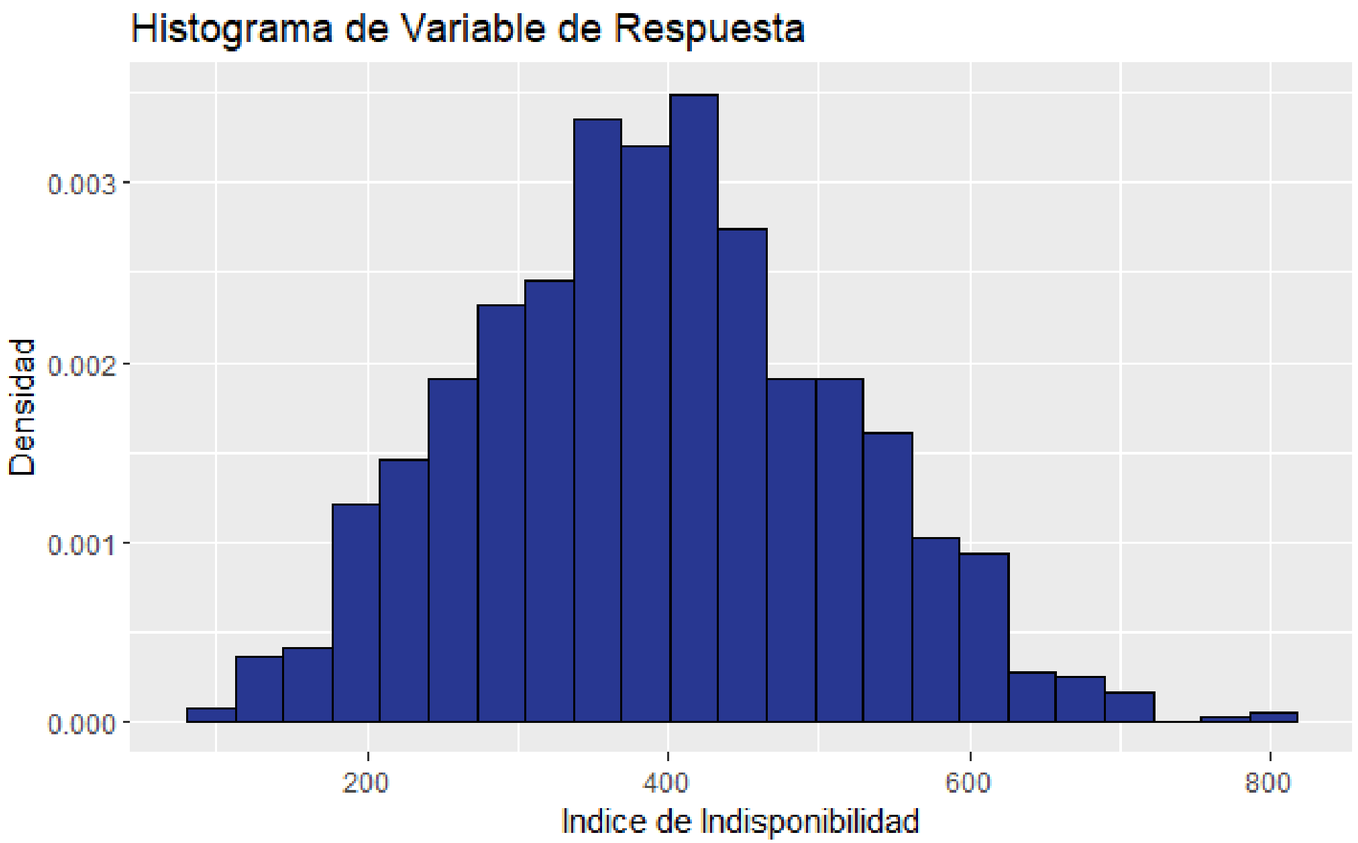
Minimizar el AIC, es decir, minimizar la cantidad de información perdida en el modelo.

```
Step:  AIC=6893.72
IndD ~ CPR + `%Error` + Media + Desvest + LatenciaMax + CPU +
    Peticion + Semestre + Franja_Horaria + LatenciaMax:Peticion +
    CPU:Peticion + CPU:Semestre + Desvest:Franja_Horaria

                        Df Sum of Sq      RSS     AIC
<none>                                  491467  6893.7
- CPU:Semestre           1       934   492401  6893.9
- LatenciaMax:Peticion   3      2704   494171  6893.9
- Desvest:Franja_Horaria 2      1910   493377  6894.1
- CPU:Peticion           3      3660   495127  6896.1
- Media                  1     11839   503305  6918.5
- `%Error`               1     45048   536515  6990.6
- CPR                    1  14840399 15331866 10768.9

Call:
lm(formula = IndD ~ CPR + `%Error` + Media + Desvest + LatenciaMax +
    CPU + Peticion + Semestre + Franja_Horaria + LatenciaMax:Peticion +
    CPU:Peticion + CPU:Semestre + Desvest:Franja_Horaria, data = Arquitectura_de_Software)

Coefficients:
                (Intercept)                      CPR                  `%Error`                    Media                  Desvest
                  7.380e+01                1.618e-01                  2.198e+01                1.709e-01                1.067e-01
               LatenciaMax                      CPU                PeticionGET               PeticionPOST                PeticionPUT
                  6.145e-02                1.279e+03                 -3.158e+00                -5.813e+00                 4.328e+01
           SemestreSegundo      Franja_HorariaNoche     Franja_HorariaTarde  LatenciaMax:PeticionGET  LatenciaMax:PeticionPOST
                  1.038e+01                -4.124e+01                 -2.135e+01                -1.525e-02                -2.392e-03
   LatenciaMax:PeticionPUT           CPU:PeticionGET          CPU:PeticionPOST           CPU:PeticionPUT       CPU:SemestreSegundo
                 -3.691e-02                -6.910e+02                 -7.008e+01                -2.515e+03                -9.784e+02
Desvest:Franja_HorariaNoche  Desvest:Franja_HorariaTarde
                 -4.091e-02                 1.792e-03
```

PREGUNTA 1

Verifique todos los supuestos de su modelo de regresión. Si se encuentran problemas de multicolinealidad proponga y ejecute un método para solucionarlo (diferente al de eliminar las variables que generan el problema). Analice el modelo resultante, compare este método con el de eliminar variables ¿es este mejor?

NORMALIDAD



Prueba Bondad de Ajuste

Chi-Cuadrado de Pearson

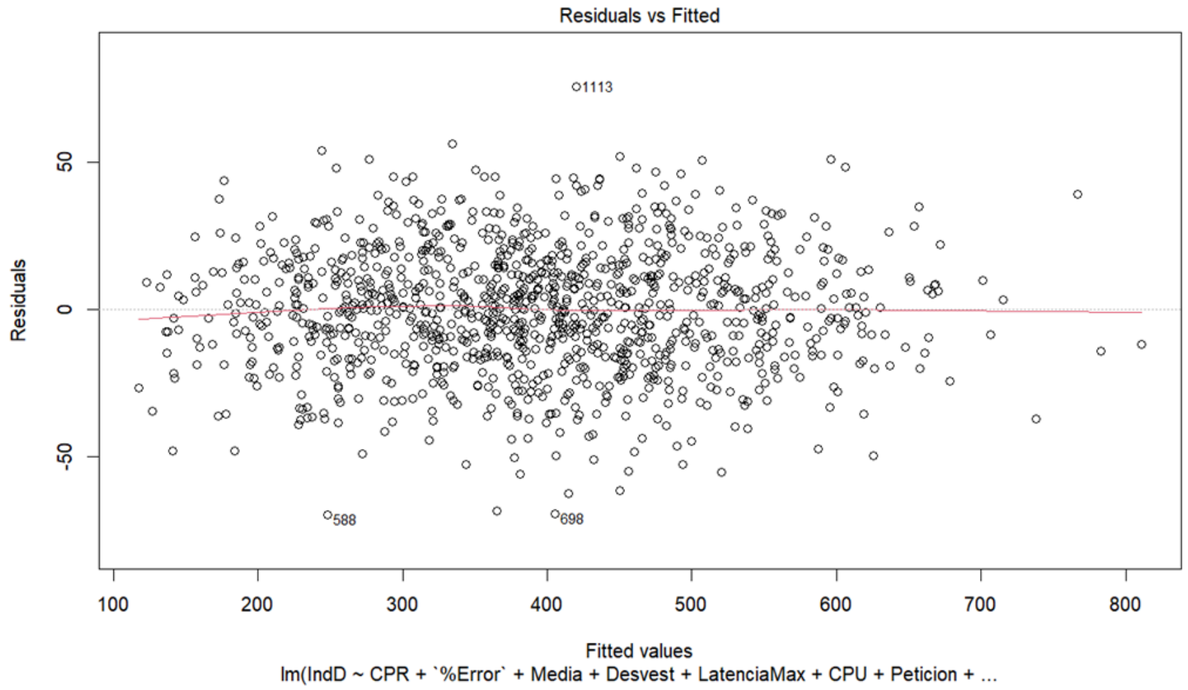
$$H_0 : Ajusta\ a\ una\ distribucion\ normal$$

$$H_A : \neg H_0$$

p-value = 70.74%

NO se rechaza hipótesis nula, es decir con un nivel de significancia del 5%, se cumple el supuesto de normalidad.

HOMOCEDASTICIDAD



Breusch- Pagan Test

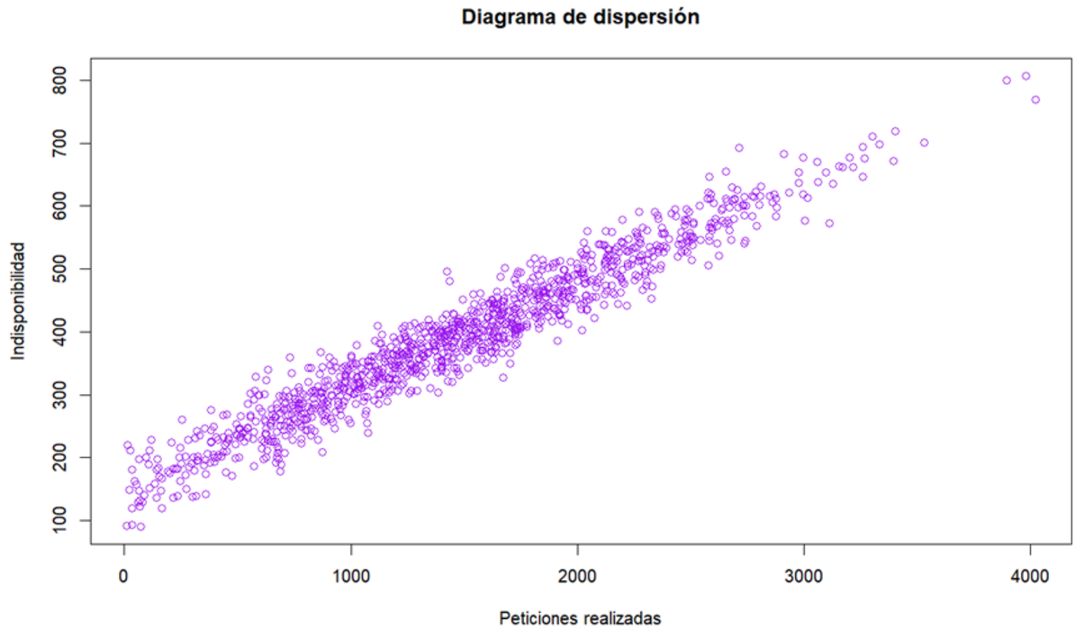
$H_0: Homocedasticidad$

$H_A: Heterocedasticidad$

p-value = 67.35%

NO se rechaza hipótesis nula, es decir con un nivel de significancia del 5%, se cumple el supuesto de homocedasticidad.

INSESGADO



Ramsey-RESET Test

$$H_0: Especificación\ correcta$$
$$H_A: Especificación\ errada$$

p-value = 34.33%

NO se rechaza hipótesis nula, es decir con un nivel de significancia del 5%, no hay problemas de especificación.

INDEPENDENCIA LINEAL

```
> durbinWatsonTest(modelo)
 lag Autocorrelation D-W Statistic p-value
   1     0.008145074      1.980112    0.69
 Alternative hypothesis: rho != 0
```

Durbin-Watson Test

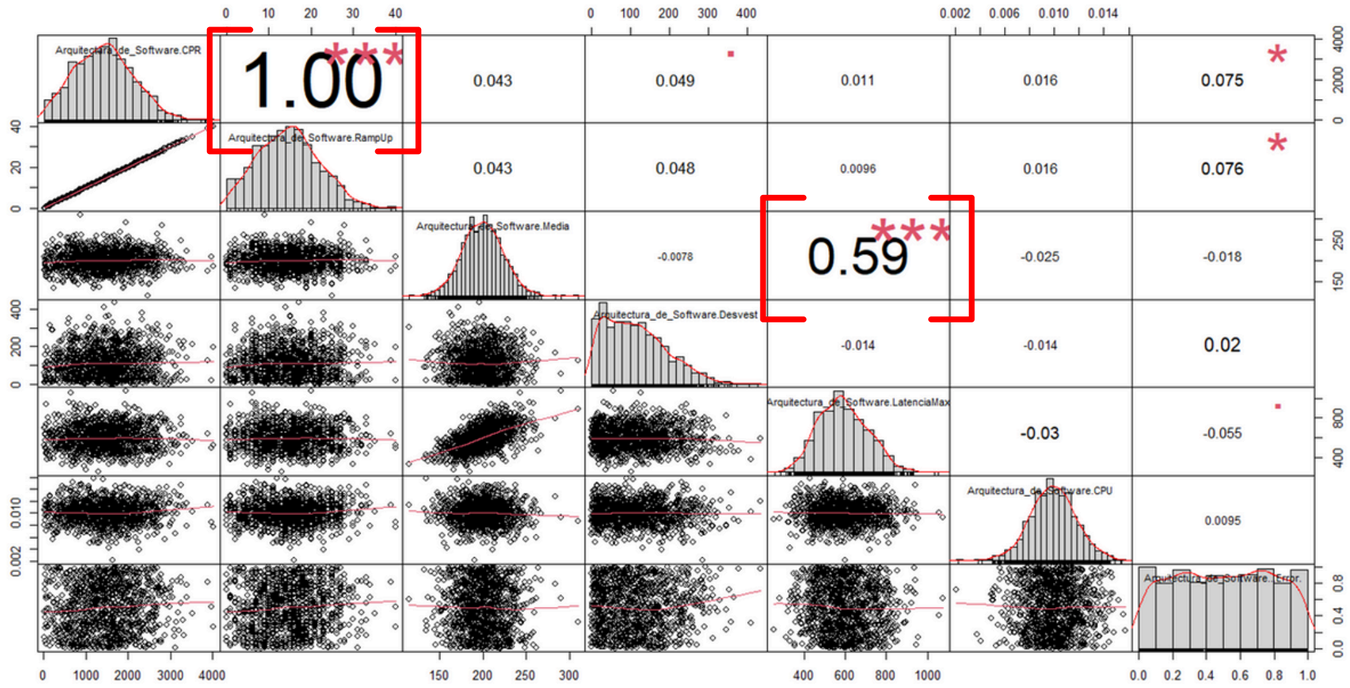
$$H_0 : Independencia\ Lineal\ de\ Residuales$$

$$H_A : \neg H_0$$

p-value = 69%

NO se rechaza hipótesis nula, es decir con un nivel de significancia del 5%, no hay problemas de autocorrelación en los residuales.

INDEPENDENCIA ENTRE REGRESORES



coeficientes de correlacion altos

Problemas de multicolinealidad

Para corregirlo, creamos un índice o eliminamos la variable RampUp o utilizamos Principal Components Analysis (PCA).

VIF:

CPR	RampUp	`%Error`	Media	Desvest	LatenciaMax	CPU
615.033989	615.057727	1.009697	1.537782	1.003902	1.540864	1.001785

INDEPENDENCIA ENTRE REGRESORES

solucion

Creacion de un indice

	CPR	*RampUp (seg)*
Media	1477.262	14.774

Se utilizan las variables estandarizadas ya que la escala de estas es diferente.

$$indice = (\widehat{CPR} + \widehat{RampUp})/2$$

nuevo VIF con indice

indice	`%Error`	Media	Desvest	LatenciaMax	CPU
1.010515	1.009357	1.537573	1.003087	1.539452	1.001561

INDEPENDENCIA ENTRE REGRESORES

```
> # PCA Correccion
>
> dat <- subset(Arquitectura_de_Software, select = c("CPR","RampUp"))

> #PCA
> pca <- prcomp(dat, scale. = TRUE)

> # divide la varianza explicada por cada variable
> # sobre la varinza que explica todas las variable
>
> pca$sdev^2 /sum(pca$sdev^2)
[1] 0.9995923745 0.0004076255

> pc1 <- pca$x[,"PC1"]

> modelo_corr <- lm(IndD ~ pc1 + Media + Desvest + LatenciaMax
+                   + CPU + Peticion + Semestre + Franja_Horaria,
+                   d .... [TRUNCATED]

> VIF(modelo_corr)
                  GVIF Df GVIF^(1/(2*Df))
pc1           1.006552  1        1.003270
Media         1.545184  1        1.243054
Desvest       1.008069  1        1.004027
LatenciaMax   1.540826  1        1.241300
CPU           1.008361  1        1.004172
Peticion      1.017598  3        1.002912
Semestre      1.003465  1        1.001731
Franja_Horaria 1.013737  2        1.003417
```

Solución PCA

Utilizando la función de R **prcomp(),** se realizó el Principal Components Analysis (PCA) entre las dos variables explicativas. Se observó que la Variable PC1 representa el **99.95% de la varianza total**, entonces se escogió esta para dejar en el modelo.

ELIMINAR VS PCA VS INDICE

Solución Eliminar RampUp

```
Residual standard error: 21.09 on 1105 degrees of freedom
Multiple R-squared:  0.9703,    Adjusted R-squared:  0.9698
F-statistic:  1720 on 21 and 1105 DF,  p-value: < 2.2e-16
```

$$R^2_{adj} = 96.98\%$$

Solución PCA

```
Residual standard error: 22.22 on 1115 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9664
F-statistic:  2947 on 11 and 1115 DF,  p-value: < 2.2e-16
```

$$R^2_{adj} = 96.64\%$$

Solución Índice

```
Residual standard error: 21.21 on 1105 degrees of freedom
Multiple R-squared:   0.97,     Adjusted R-squared:  0.9694
F-statistic:  1700 on 21 and 1105 DF,  p-value: < 2.2e-16
```

$$R^2_{adj} = 96.64\%$$

PREGUNTA 2

Identifique, de manera general, cuáles son los aspectos que maximizan el índice de indisponibilidad. A partir de ello, proponga alguna medida de control frente a esos aspectos, de modo que la compañía pueda enfocarse en estos para reducir la indisponibilidad de su servidor.

MAX

VARIABLES MAXIMIZAN

Variables significativas que afectan el índice de indisponibilidad:

Cantidad de peticiones (CPR), el porcentaje de errores (%Error), el tiempo promedio de respuesta (Media), la desviación estándar (Desviación Estándar), la latencia máxima y la franja horaria

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.380e+01  1.508e+01   4.894 1.13e-06 ***
CPR                            1.618e-01  8.860e-04 182.666  < 2e-16 ***
`%Error`                       2.198e+01  2.184e+00  10.064  < 2e-16 ***
Media                          1.709e-01  3.313e-02   5.159 2.94e-07 ***
Desvest                        1.067e-01  1.238e-02   8.617  < 2e-16 ***
LatenciaMax                    6.145e-02  1.599e-02   3.842 0.000129 ***
CPU                            1.279e+03  1.095e+03   1.169 0.242834
PeticionGET                   -3.158e+00  1.564e+01  -0.202 0.840013
PeticionPOST                  -5.813e+00  1.639e+01  -0.355 0.722993
PeticionPUT                    4.328e+01  1.723e+01   2.511 0.012170 *
SemestreSegundo                1.038e+01  6.867e+00   1.512 0.130821
Franja_HorariaNoche           -4.124e+01  3.085e+00 -13.369  < 2e-16 ***
Franja_HorariaTarde           -2.135e+01  2.495e+00  -8.558  < 2e-16 ***
LatenciaMax:PeticionGET       -1.525e-02  1.765e-02  -0.864 0.387609
LatenciaMax:PeticionPOST      -2.392e-03  1.838e-02  -0.130 0.896492
LatenciaMax:PeticionPUT       -3.691e-02  1.968e-02  -1.876 0.060958 .
CPU:PeticionGET               -6.910e+02  1.180e+03  -0.586 0.558269
CPU:PeticionPOST              -7.008e+01  1.223e+03  -0.057 0.954322
CPU:PeticionPUT               -2.515e+03  1.243e+03  -2.024 0.043204 *
CPU:SemestreSegundo           -9.784e+02  6.751e+02  -1.449 0.147548
Desvest:Franja_HorariaNoche   -4.091e-02  2.192e-02  -1.866 0.062291 .
Desvest:Franja_HorariaTarde    1.792e-03  1.786e-02   0.100 0.920087
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.09 on 1105 degrees of freedom
Multiple R-squared:  0.9703,    Adjusted R-squared:  0.9698
F-statistic:  1720 on 21 and 1105 DF,  p-value: < 2.2e-16
```

PROPUESTAS DE MEDIDA DE CONTROL

Gestión Dinámica de Peticiones (CPR):

- Ajustar la capacidad del servidor en tiempo real.
- Establecer umbrales para redirigir o rechazar peticiones según la carga.

Control de Calidad y Reducción de Errores (%Error):

- Crear un equipo dedicado para abordar y corregir errores recurrentes.
- Mejorar la validación de peticiones para reducir errores.

Monitoreo y Ajustes Automáticos (Latencia Máxima):

- Implementar alertas tempranas ante latencia crítica.
- Ajustar automáticamente la configuración del servidor para evitar congestiones.

Optimización de Recursos CPU:

- Asignar CPU según perfiles específicos de cada tipo de petición.
- Utilizar algoritmos de equilibrio de carga para distribuir eficientemente la carga.

Gestión de Recursos según Franja Horaria:

- Reservar recursos extras en horas pico.
- Establecer un plan de contingencia para redistribuir recursos ante cambios bruscos.

PREGUNTA 3

$$H_0: \beta_1 = \beta_2 = \cdots . = \beta_{21}$$

H_A: $\neg H_0$

MODELO SIGNIFICATIVO

VARIABLES SIGNIFICATIVAS

H_0: β=0

H_A: $\beta \neq 0$

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.380e+01  1.508e+01   4.894 1.13e-06 ***
CPR                            1.618e-01  8.860e-04 182.666  < 2e-16 ***
`%Error`                       2.198e+01  2.184e+00  10.064  < 2e-16 ***
Media                          1.709e-01  3.313e-02   5.159 2.94e-07 ***
Desvest                        1.067e-01  1.238e-02   8.617  < 2e-16 ***
LatenciaMax                    6.145e-02  1.599e-02   3.842 0.000129 ***
CPU                            1.279e+03  1.095e+03   1.169 0.242834
PeticionGET                   -3.158e+00  1.564e+01  -0.202 0.840013
PeticionPOST                  -5.813e+00  1.639e+01  -0.355 0.722993
PeticionPUT                    4.328e+01  1.723e+01   2.511 0.012170 *
SemestreSegundo                1.038e+01  6.867e+00   1.512 0.130821
Franja_HorariaNoche           -4.124e+01  3.085e+00 -13.369  < 2e-16 ***
Franja_HorariaTarde           -2.135e+01  2.495e+00  -8.558  < 2e-16 ***
LatenciaMax:PeticionGET       -1.525e-02  1.765e-02  -0.864 0.387609
LatenciaMax:PeticionPOST      -2.392e-03  1.838e-02  -0.130 0.896492
LatenciaMax:PeticionPUT       -3.691e-02  1.968e-02  -1.876 0.060958 .
CPU:PeticionGET               -6.910e+02  1.180e+03  -0.586 0.558269
CPU:PeticionPOST              -7.008e+01  1.223e+03  -0.057 0.954322
CPU:PeticionPUT               -2.515e+03  1.243e+03  -2.024 0.043204 *
CPU:SemestreSegundo           -9.784e+02  6.751e+02  -1.449 0.147548
Desvest:Franja_HorariaNoche   -4.091e-02  2.192e-02  -1.866 0.062291 .
Desvest:Franja_HorariaTarde    1.792e-03  1.786e-02   0.100 0.920087
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.09 on 1105 degrees of freedom
Multiple R-squared:  0.9703,    Adjusted R-squared:  0.9698
F-statistic:  1720 on 21 and 1105 DF,  p-value: < 2.2e-16
```

R^2=0.9703

El 97,03% de lal variabilidad del índice de indisponibilidad está siendo explicada por el modelo de regresión.

PREGUNTA 4

Las peticiones de tipo POST, PUT y DELETE siempre realizan un proceso de escritura sobre la base de datos, lo que en general, según el gerente de TI, tiende a sobre cargar el sistema y aumentan el nivel de indisponibilidad. Por otro lado, las peticiones GET solamente realizan proceso de lectura y según el gerente de TI aumentan el nivel de indisponibilidad en menor medida que las otras peticiones. Promedie los efectos de las peticiones POST, PUT y DELETE, y concluya si la afirmación del gerente de TI es correcta

$$\beta_{GET} < \frac{\beta_{PUT} + \beta_{POST} + \beta_{DELETE}}{3} \rightarrow \beta_{PUT} + \beta_{POST} + \beta_{DELETE} - 3\beta_{GET} > 0$$

$$H_0: \beta_{PUT} + \beta_{POST} + \beta_{DELETE} - 3\beta_{GET} = 0$$

$$H_A: \beta_{PUT} + \beta_{POST} + \beta_{DELETE} - 3\beta_{GET} \neq 0$$

$$t = \frac{c^T \beta}{\sqrt{c^T Var(\beta) c}} = 3.145869$$

$$p - value = 2P(t > |3.145869|)) = 0.00170018$$

Rechazamos Ho ya que el p-value < 5%, por lo que las peticiones tipo POST, PUT y DELETE tiene mayor indisponibilidad promedio en comparación a una petición GET.

PREGUNTA 5

El equipo de TI ha concluido que, si el sistema llega a un índice de indisponibilidad mayor a 350 puntos, el sistema no cumple los requerimientos de disponibilidad. Realice un intervalo de predicción para las siguientes condiciones y mencione si bajo estas condiciones el sistema se encuentra indisponible o no.

Condiciones Intervalo

CPR: 10,000 peticiones

RampUp: 10 Segundos

Petición: GET

%Error: 35%

Media: 200ms

DesvEstand: 72ms

Semestre: Primero

Franja Horaria: Tarde

LatenciaMax: 320ms

CPU: 1%

Intervalo de predicción

```
> t(c)%*%B - sqrt(MSE + t(c)%*%V%*%c)
          [,1]
[1,] 260.1878
> t(c)%*%B + sqrt(MSE + t(c)%*%V%*%c)
          [,1]
[1,] 302.7793
```

PREGUNTA 6

El equipo de soporte de TI quiere cambiar la variable de interés de "índice de indisponibilidad" a "probabilidad de indisponibilidad". ¿Qué tipo de modelo propondría su grupo para cumplir con esta petición?

Si la variable de respuesta es la probabilidad de indisponibilidad el grupo le propondría hacer un **Modelo de Regresión Logística**

GRACIAS