Facultad de Ciencias
UNAM

Estadística II

Tarea 2B

Regresión Lineal Múltiple

Enríquez Hernández Leobardo
Huitrón Zambrano Victor Manuel
Suárez López David

21 de mayo de 2024

Índice

Ejercicio 1

Considere el modelo de regresión

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

y los estimadores obtenidos por mínimos cuadrados en forma matricial $\hat{\beta} = (X^t X)^{-1} X^t y$

Usando la matriz proyección H y sus propiedades indique:

I) A que es igual: $e^t X$

$$\begin{aligned} e^t X &= (y - \hat{y})^t X \\ &= (y - Hy)^t X \\ &= (y^t - y^t H^t) X \\ &= y^t X - y^t H X \\ &= y^t X - y^t (X(X^t X)^{-1} X^t) X \\ &= y^t X - y^t X[(X^t X)^{-1} X^t X] \\ &= y^t X - y^t X = 0 \end{aligned}$$

$\therefore e^t X = 0$

II) A que es igual: $Cov(e, \hat{y})$

Primero notemos lo siguiente:

$$\begin{aligned} e &= (y - \hat{y}) \\ &= (y - Hy) \\ &= (I - H)y \end{aligned}$$

Donde I es la matriz identidad, entonces tenemos:

$$\begin{aligned} Cov(e, \hat{y}) &= Cov((I - H)y, \hat{y}) \\ &= Cov((I - H)y, Hy) \\ &= (I - H)Cov(y, y)H^t \\ &= (I - H)Var(y)H \\ &= (I - H)\sigma^2 H \\ &= \sigma^2(H - HH) \\ &= \sigma^2(H - H) \\ &= \sigma^2(0) = 0 \end{aligned}$$

$\therefore Cov(e, \hat{y}) = 0$

Ejercicio 2

Considere el modelo de regresion

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(3x_i^2 - 2\right) + \xi_i \quad i = 1, 2, 3$$

donde

$$x_1 = -1, x_2 = 0, x_3 = 1$$

I) Matriz de diseño

$$X = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} \begin{matrix} 3x_1^2 - 2 = 1 \\ \text{pues } 3x_2^2 - 2 = -2 \\ 3x_3^2 - 2 = 1 \end{matrix}$$

Asi

$$X^t X = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{pmatrix}$$

y

$$\left(X^t X\right)^{-1} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/6 \end{pmatrix}$$

II)

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

asi:

$$\hat{\beta} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/6 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ -1/2 & 0 & 1/2 \\ 1/6 & -1/3 & 1/6 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \frac{1}{3}(y_1 + y_2 + y_3) \\ \frac{1}{2}(y_3 - y_1) \\ \frac{1}{6}(y_1 - 2y_2 + y_3) \end{pmatrix}$$

Por lo tanto

$$\hat{\beta}_0 = \frac{1}{3}(y_1 + y_2 + y_3) \quad \hat{\beta}_1 = \frac{1}{2}(y_3 - y_1) \quad \hat{\beta}_2 = \frac{1}{6}(y_1 - 2y_2 + y_3)$$

III) Obtenemos los estimadores del modelo reducido:

$$y_i = \beta_0^* + \beta_1^* x_i + \xi_i^* \quad i = 1, 2, 3$$

Obtenemos

$$\left(X^t X\right)^{-1} = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$\rightarrow \hat{\beta}^* = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ -1/2 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$\hat{\beta}^* = \begin{pmatrix} \frac{1}{3}(y_1 + y_2 + y_3) \\ \frac{1}{2}(y_3 - y_1) \end{pmatrix}$$

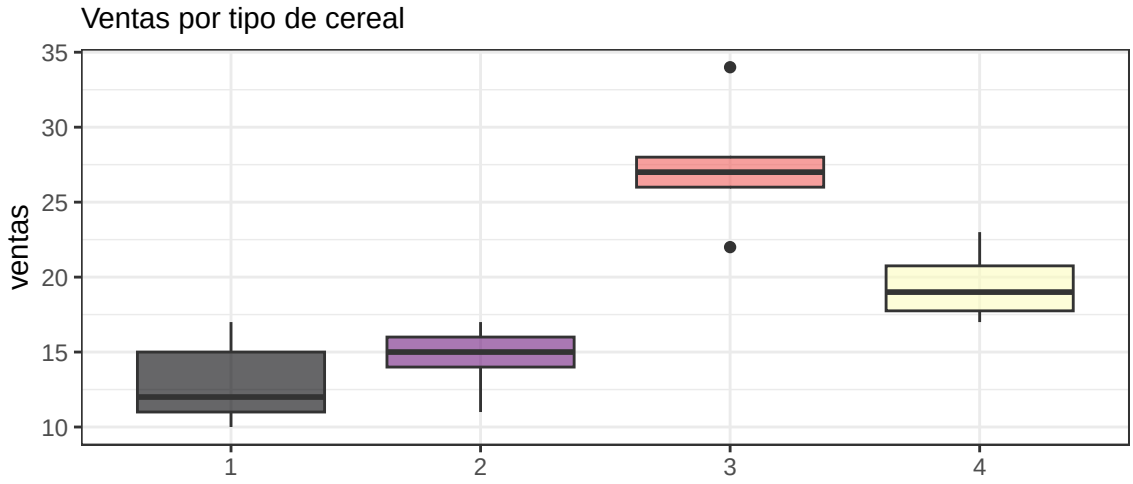
Asi $\hat{\beta}_0^* = \frac{1}{3}(y_1 + y_2 + y_3)$ y $\hat{\beta}_1^* = \frac{1}{2}(y_3 - y_1)$

$\therefore \hat{\beta}_0^* = \hat{\beta}_0$ y $\hat{\beta}_1^* = \hat{\beta}_1$

Ejercicio 3

Compararemos 4 distintos diseños de empaque de un nuevo cereal, asignados aleatoriamente a 5 tiendas como unidades muestrales, y las ventas en un periodo de 2 semanas.

En el siguiente Boxplot, podemos observar que el empaque que más se vendió es el 3, aunque con una mayor variabilidad entre las tiendas que las venden y el empaque que menos se vendió es el empaque 1.



Ajustaremos un modelo de regresión lineal múltiple del número de ventas promedio por cada tipo de empaque.

```
##
## Call:
## lm(formula = ventas ~ cereal, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.40  -1.75  -0.40   1.70   6.60
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   13.000      1.421   9.150 0.000000159 ***
## cereal2        1.600      2.009   0.796      0.4383
## cereal3       14.400      2.009   7.167 0.000003247 ***
## cereal4        6.500      2.131   3.050      0.0081 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.177 on 15 degrees of freedom
## Multiple R-squared:  0.8056, Adjusted R-squared:  0.7667
## F-statistic: 20.71 on 3 and 15 DF,  p-value: 0.00001367
```

Las expresiones del número de ventas promedio por cada tipo de empaque son.

$E(ventas; cereal1) = \hat{\beta}_0 = 13$

$E(ventas; cereal2) = \hat{\beta}_0 + \hat{\beta}_1 = 13 + 1.6 = 14.6$

$E(ventas; cereal3) = \hat{\beta}_0 + \hat{\beta}_2 = 13 + 14.4 = 27.4$

$E(ventas; cereal4) = \hat{\beta}_0 + \hat{\beta}_3 = 13 + 6.5 = 19.5$

Las hipótesis que se contrastan con la prueba F asociada a la tabla ANOVA son, la hipotesis nula $H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$ contra la alternativa de $H_a : \hat{\beta}_i \neq 0$ para al menos un $i = 1, 2, 3$. Esta prueba se presenta en la salida o `summary` anterior, en donde podemos observar un $p - value : 1{,}367e - 05$ de la prueba `F-statistic` con valor de 20,71 con 3 y 15 grados de libertad. Como el valor del $p - value < 0{,}05$, i.e., considerando un nivel de significancia estadística $\alpha = 0{,}05$, podemos concluir que se rechaza la hipótesis nula H_0, por lo que al menos un β_i es distinto de cero en el modelo planteado.

Para ver si el diseño del empaque afecta las ventas promedio, plantearemos algunas pruebas de hipótesis, usando un nivel de confianza del 95 %. Nos preguntamos si $E(ventas; cereal1) \neq E(ventas; cereal2)$, $E(ventas; cereal1) \neq E(ventas; cereal3)$, $E(ventas; cereal1) \neq E(ventas; cereal4)$, $E(ventas; cereal2) \neq E(ventas; cereal3)$, $E(ventas; cereal2) \neq E(ventas; cereal4)$, $E(ventas; cereal3) \neq E(ventas; cereal4)$. Entonces, planteamos las siguientes pruebas.

Planteamiento de la hipótesis nula:

$\hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_1 \rightarrow \hat{\beta}_1 = 0$

$\hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_2 \rightarrow \hat{\beta}_2 = 0$

$\hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_3 \rightarrow \hat{\beta}_3 = 0$

$\hat{\beta}_0 + \hat{\beta}_1 = \hat{\beta}_0 + \hat{\beta}_2 \rightarrow \hat{\beta}_1 = \hat{\beta}_2$

$\hat{\beta}_0 + \hat{\beta}_1 = \hat{\beta}_0 + \hat{\beta}_3 \rightarrow \hat{\beta}_1 = \hat{\beta}_3$

$\hat{\beta}_0 + \hat{\beta}_2 = \hat{\beta}_0 + \hat{\beta}_3 \rightarrow \hat{\beta}_2 = \hat{\beta}_3$

Tenemos términos redundantes, por lo que nos quedaría la siguiente prueba de hipótesis.

$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$ VS $H_a : \hat{\beta}_i \neq 0$ p.a. $i = 1, 2, 3$.

En la prueba global F asociada a la tabla ANOVA descrita anteriormente, se rechzó H_0. Por lo que podemos concluir que al menos un diseño de empaque afecta las ventas promedio, sin embargo no nos dice explícitamente cuál o cuáles en un análisis simultáneo.

Para esto, podemos plantear una prueba de hipótesis simultánea asociada a la igualdad de las ventas promedio entre todos los posibles pares de diferentes empaques, que puede resolverse con la prueba lineal general simultánea. A continuación, se muestra la salida de la prueba.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = ventas ~ cereal, data = data)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0    1.600      2.009   0.796   0.8549
## 2 == 0   14.400      2.009   7.167   <0.001 ***
## 3 == 0    6.500      2.131   3.050   0.0366 *
## 4 == 0  -12.800      2.009  -6.370   <0.001 ***
## 5 == 0   -4.900      2.131  -2.299   0.1423
## 6 == 0    7.900      2.131   3.707   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Finalmente realizaremos una prueba de hipótesis para argumentar en favor o en contra de la hipótesis de que el diseño de empaque 3 es el que más aumenta las ventas en comparación con el resto de empaques. Esto es, que $E(ventas; cereal3) > E(ventas; cereal1)$, $E(ventas; cereal3) > E(ventas; cereal2)$, $E(ventas; cereal3) > E(ventas; cereal4)$. Entonces planteamos, las siguientes hipótesis.

$\hat{\beta}_0 + \hat{\beta}_2 > \hat{\beta}_0 \rightarrow \hat{\beta}_2 > 0$

$\hat{\beta}_0 + \hat{\beta}_2 > \hat{\beta}_0 + \hat{\beta}_1 \rightarrow \hat{\beta}_2 > \hat{\beta}_1$

$\hat{\beta}_0 + \hat{\beta}_2 > \hat{\beta}_0 + \hat{\beta}_3 \rightarrow \hat{\beta}_2 > \hat{\beta}_3$

Hipótesis nula: $H_0 : \hat{\beta}_2 \leq 0$, $\hat{\beta}_2 \leq \hat{\beta}_1$, $\hat{\beta}_2 \leq \hat{\beta}_3$

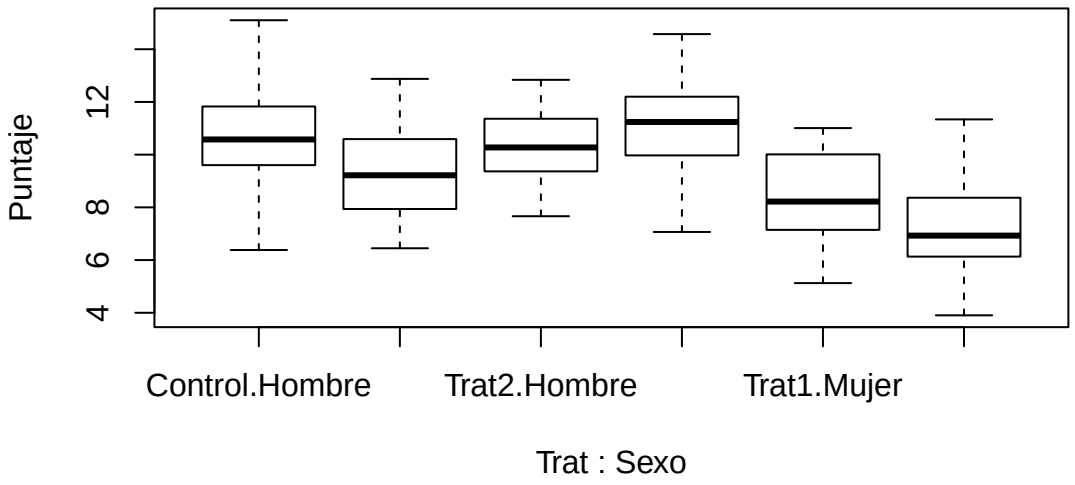
Hipótesis alternativa: $H_a : \hat{\beta}_2 > 0$, $\hat{\beta}_2 > \hat{\beta}_1$, $\hat{\beta}_2 > \hat{\beta}_3$

A continuación se muestra la salida, donde podemos observar que se rechaza H_0, por lo que podemos afirmar que el diseño de empaque 3 es el que más aumenta las ventas en comparación con el resto de empaques.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = ventas ~ cereal, data = data)
##
## Linear Hypotheses:
##        Estimate Std. Error t value  Pr(>t)
## 1 <= 0   14.400      2.009   7.167 < 0.001 ***
## 2 <= 0   12.800      2.009   6.370 < 0.001 ***
## 3 <= 0    7.900      2.131   3.707 0.00304 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Ejercicio 4

I)



En estos boxplots podemos observar que para el caso Control se observa que de puntaje que presentan ansiedad, tamto en hombres como mujeres, se encuentra aproximadamente entre 10 y 11.

En el caso de aplicar el tratamiento actual (Trat1) podemos notar que hay una disminucion en el nivel de ansiedad que presentan tanto hombres como mujeres, aunque parece ser que el tratamiento actual reduce más el nivel de ansiedad en mujeres que en hombres.

Por otro lado, al aplicar el nuevo tratamiento (Trat2) podemos notar una ligera disminucion en el nivel de ansiedad de los hombres respecto al caso Control, pero los niveles de ansiedad son mayores en comparacion al aplicar el tratamiento actual. Mientras que en el caso de las mujeres el nuevo tratamiento reduce la ansiedad a un nivel muchisimo mas bajo en comparación al caso control e incluso a un nivel mas bajo comparandolo al aplicar el tratamiento actual.

II)

El modelo general es el siguiente:

$$E(Puntaje; Trat, Sexo) = \beta_0 + \beta_1 * Trat1 + \beta_2 * Trat2 + \beta_3 * Mujer + \beta_4(Trat1 * Mujer) + \beta_5(Trat2 * Mujer)$$

A partir del modelo general podemos obtener los modelos individuales:

$E(Puntaje; Control, Hombre) = \beta_0$

$E(Puntaje; Trat1, Hombre) = \beta_0 + \beta_1$

$E(Puntaje; Trat2, Hombre) = \beta_0 + \beta_2$

$E(Puntaje; Control, Mujer) = \beta_0 + \beta_3$

$E(Puntaje; Trat1, Mujer) = \beta_0 + \beta_1 + \beta_3 + \beta_4$

$E(Puntaje; Trat2, Mujer) = \beta_0 + \beta_2 + \beta_3 + \beta_5$

Ajunstamos nuestro modelo:

```
##
## Call:
## lm(formula = Puntaje ~ Trat * Sexo, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3785 -1.1800 -0.0518  1.2159  4.3400
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           10.7602     0.3948  27.252 < 0.0000000000000002 ***
## TratTrat1             -1.5100     0.5584  -2.704               0.0079 **
## TratTrat2             -0.4789     0.5584  -0.858               0.3929
## SexoMujer              0.5231     0.5584   0.937               0.3509
## TratTrat1:SexoMujer   -1.3758     0.7897  -1.742               0.0842 .
## TratTrat2:SexoMujer   -3.5914     0.7897  -4.548            0.0000136 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.766 on 114 degrees of freedom
## Multiple R-squared:  0.4007, Adjusted R-squared:  0.3744
## F-statistic: 15.24 on 5 and 114 DF,  p-value: 0.00000000001873
```

Por tanto las estimaciones puntuales son:

$E(Puntaje; Control, Hombre) = 10{,}7602$

$E(Puntaje; Trat1, Hombre) = 10{,}7602 + (-1{,}51) = 9{,}2502$

$E(Puntaje; Trat2, Hombre) = 10{,}7602 + (-0{,}4798) = 10{,}2804$

$E(Puntaje; Control, Mujer) = 10{,}7602 + 0{,}5231 = 11{,}2833$

$E(Puntaje; Trat1, Mujer) = 10{,}7602 + (-1{,}5100) + 0{,}5231 + (-1{,}3758) = 8{,}3975$

$E(Puntaje; Trat2, Mujer) = 10{,}7602 + (-0{,}4789) + 0{,}5231 + (-3{,}5914) = 7{,}213$

III)

Las hipótesis que se contrastan con la tabla ANOVA son:

$H_0 : \beta_0 = 0,\ \beta_1 = 0,\ \beta_2 = 0,\ \beta_3 = 0,\ \beta_4 = 0,\ \beta_5 = 0$ vs $H_\alpha : \beta_0 \neq 0$ ó $\beta_1 \neq 0\ \beta_2 \neq 0,\ \beta_3 \neq 0,\ \beta_4 \neq 0,\ \beta_5 \neq 0$

Con el comando $summary(fit)$ vemos que el p-value asociado a la tabla ANOVA es de 1.873e-11 que es menor a nuestra significancia de .05, por lo que rechazamos H_0.

IV)

Para determinar si el sexo tiene un efecto en el puntaje realizaremos una prueba de hipotesis, la cual tiene como hipotesis nula:

$$H_0 = \begin{cases} E(Puntaje; Control, Hombre) &=& E(Puntaje; Control, Mujer) \\ E(Puntaje; Trat1, Hombre) &=& E(Puntaje; Trat1, Mujer) \\ E(Puntaje; Trat2, Hombre) &=& E(Puntaje; Trat2, Mujer) \end{cases} \iff$$

$$\iff H_0 = \begin{cases} \beta_0 &=& \beta_0 + \beta_3 \\ \beta_0 + \beta_1 &=& \beta_0 + \beta_1 + \beta_3 + \beta_4 \\ \beta_0 + \beta_2 &=& \beta_0 + \beta_2 + \beta_3 + \beta_5 \end{cases} \iff H_0 = \begin{cases} 0 &=& \beta_3 \\ 0 &=& \beta_3 + \beta_4 \\ 0 &=& \beta_3 + \beta_5 \end{cases}$$

Notemos que al comparar dos a dos las igualdades de H_0 podemos obtener una hipotesis nula que es equivalente, la cual es:

$$H_0 = \begin{cases} 0 &= \beta_3 \\ 0 &= \beta_4 \\ 0 &= \beta_5 \end{cases}$$

Por lo tanto, la prueba de hipotesis a realizar es:

$H_0 : \beta_3 = 0,\ \beta_4 = 0,\ \beta_5 = 0$ vs $H_\alpha : \beta_3 \neq 0,\ \beta_4 \neq 0,\ \beta_5 \neq 0$

```
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##        Estimate
## 1 == 0   0.5231
## 2 == 0  -1.3758
## 3 == 0  -3.5914
##
## Global Test:
##       F DF1 DF2      Pr(>F)
## 1 11.13   3 114 0.000001828
```

Vemos que el p-valor es de 1.827792e-06, por lo que rechazamos H_0. Ahora realizaremos una prueba simultanea para ver si podemos decir que el sexo tiene algun efecto en el tratamiento

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Puntaje ~ Trat * Sexo, data = datos)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0   0.5231     0.5584   0.937    0.623
## 2 == 0  -1.3758     0.7897  -1.742    0.182
## 3 == 0  -3.5914     0.7897  -4.548   <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Con la prueba simultanea podemos ver que es plausible quitar a β_3 y β_4 del modelo, pues el p-value en ambos es mayor a .025

Por tanto nuestro modelo reducido seria:

$$E(Puntaje; Trat, Sexo) = \beta_0 + \beta_1 * Trat1 + \beta_2 * Trat2 + \beta_3(Trat2 * Mujer)$$

V)

Ajustamos nuestro modelo reducido:

```
##
## Call:
## lm(formula = Puntaje ~ I(Trat == "Trat1") + I(Trat == "Trat2") +
##     I((Sexo == "Mujer") * (Trat == "Trat2")), data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6400 -1.0701  0.0033  1.0982  4.1249
```

```
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                               11.0217     0.2806  39.272
## I(Trat == "Trat1")TRUE                    -2.1979     0.3969  -5.538
## I(Trat == "Trat2")TRUE                    -0.7405     0.4861  -1.523
## I((Sexo == "Mujer") * (Trat == "Trat2"))  -3.0683     0.5613  -5.466
##                                                    Pr(>|t|)
## (Intercept)                              < 0.0000000000000002 ***
## I(Trat == "Trat1")TRUE                            0.000000194 ***
## I(Trat == "Trat2")TRUE                                  0.13
## I((Sexo == "Mujer") * (Trat == "Trat2"))          0.000000267 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.775 on 116 degrees of freedom
## Multiple R-squared:  0.3838, Adjusted R-squared:  0.3678
## F-statistic: 24.08 on 3 and 116 DF,  p-value: 0.000000000003461
```

Entonces las expresiones del puntaje promedio de cada uno de los valores en las variables categoricas, junton con su estimacion puntual, son:

$E(Puntaje; Control, Hombre) = \beta_0 = 11{,}0217$

$E(Puntaje; Trat1, Hombre) = \beta_0 + \beta_1 = 11{,}0217 + (-2{,}1979) = 8{,}8238$

$E(Puntaje; Trat2, Hombre) = \beta_0 + \beta_2 = 11{,}0217 + (-0{,}7405) = 10{,}2812$

$E(Puntaje; Control, Mujer) = \beta_0 = 11{,}0217$

$E(Puntaje; Trat1, Mujer) = \beta_0 + \beta_1 = 11{,}0217 + (-2{,}1979) = 8{,}8238$

$E(Puntaje; Trat2, Mujer) = \beta_0 + \beta_2 + \beta_3 = 11{,}0217 + (-0{,}7405) + (-3{,}0683) = 7{,}2129$

VI)

Queremos realizar una prueba de hipotesis para ver si el nuevo tratamiento tiene mejor desempeño, por lo que nuestra hipotesis alternativa seria:

$$H_\alpha = \begin{cases} E(puntaje; Trat2, Hombre) & < & E(puntaje; Control, Hombre) \\ E(puntaje; Trat2, Hombre) & < & E(puntaje; Trat1, Hombre) \\ E(puntaje; Trat2, Mujer) & < & E(puntaje; Control, Mujer) \\ E(puntaje; Trat2, Mujer) & < & E(puntaje; Trat1, Mujer) \end{cases} \iff$$

$$\iff H_\alpha = \begin{cases} \beta_0 + \beta_2 & < & \beta_0 \\ \beta_0 + \beta_2 & < & \beta_0 + \beta_1 \\ \beta_0 + \beta_2 + \beta_3 & < & \beta_0 \\ \beta_0 + \beta_2 + \beta_3 & < & \beta_0 + \beta_1 \end{cases} \iff H_\alpha = \begin{cases} 0 & < & -\beta_2 \\ 0 & < & \beta_1 - \beta_2 \\ 0 & < & -\beta_2 - \beta_3 \\ 0 & < & \beta_1 - \beta_2 - \beta_3 \end{cases}$$

Entonces nuestra prueba de hipotesis a realizar es:

$$H_0 = \begin{cases} 0 & \geq & -\beta_2 \\ 0 & \geq & \beta_1 - \beta_2 \\ 0 & \geq & -\beta_2 - \beta_3 \\ 0 & \geq & \beta_1 - \beta_2 - \beta_3 \end{cases} \quad \text{vs} \quad H_\alpha = \begin{cases} 0 & < & -\beta_2 \\ 0 & < & \beta_1 - \beta_2 \\ 0 & < & -\beta_2 - \beta_3 \\ 0 & < & \beta_1 - \beta_2 - \beta_3 \end{cases}$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
```

```
## Fit: lm(formula = Puntaje ~ I(Trat == "Trat1") + I(Trat == "Trat2") +
##     I((Sexo == "Mujer") * (Trat == "Trat2")), data = datos)
##
## Linear Hypotheses:
##        Estimate Std. Error t value  Pr(>t)
## 1 <= 0   0.7405     0.4861   1.523 0.18152
## 2 <= 0  -1.4574     0.4861  -2.998 1.00000
## 3 <= 0   2.9383     0.7425   3.957 < 0.001 ***
## 4 <= 0   1.6109     0.4861   3.314 0.00206 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Con esta prueba podemos ver que hay evidencia para no rechazar H0, pues tenemos que en los hombres el nuevo tratamiento no resulta ser mejor. Por lo que podriamos decir que el nuevo tratamiento no tiene mejor desempeño

VII)

Queremos realizar una prueba de hipotesis para ver que el tratamiento nuevo tiene mejor desempeño en mujeres mientras que el tratamiento actual lo tiene en hombres, por lo que nuestra hipotesis alternativa seria:

$$H_\alpha = \begin{cases} E(puntaje; Trat1, Hombre) & < & E(puntaje; Control, Hombre) \\ E(puntaje; Trat1, Hombre) & < & E(puntaje; Trat2, Hombre) \\ E(puntaje; Trat2, Mujer) & < & E(puntaje; Control, Mujer) \\ E(puntaje; Trat2, Mujer) & < & E(puntaje; Trat1, Mujer) \end{cases}$$

$$\iff$$

$$H_\alpha = \begin{cases} \beta_0 + \beta_1 & < & \beta_0 \\ \beta_0 + \beta_1 & < & \beta_0 + \beta_2 \\ \beta_0 + \beta_2 + \beta_3 & < & \beta_0 \\ \beta_0 + \beta_2 + \beta_3 & < & \beta_0 + \beta_1 \end{cases} \iff H_\alpha = \begin{cases} 0 & < & -\beta_1 \\ 0 & < & \beta_2 - \beta_1 \\ 0 & < & -\beta_2 - \beta_3 \\ 0 & < & \beta_1 - \beta_2 - \beta_3 \end{cases}$$

Entonces nuestra prueba de hipotesis a realizar es:

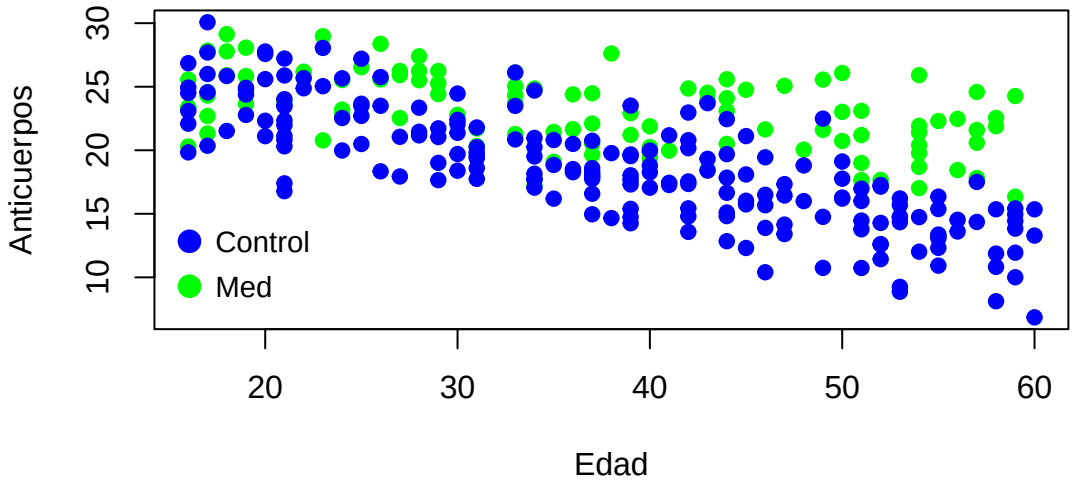
$$H_0 = \begin{cases} 0 & \geq & -\beta_1 \\ 0 & \geq & \beta_2 - \beta_1 \\ 0 & \geq & -\beta_2 - \beta_3 \\ 0 & \geq & \beta_1 - \beta_2 - \beta_3 \end{cases} \quad \text{vs} \quad H_\alpha = \begin{cases} 0 & < & -\beta_1 \\ 0 & < & \beta_2 - \beta_1 \\ 0 & < & -\beta_2 - \beta_3 \\ 0 & < & \beta_1 - \beta_2 - \beta_3 \end{cases}$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Puntaje ~ I(Trat == "Trat1") + I(Trat == "Trat2") +
##     I((Sexo == "Mujer") * (Trat == "Trat2")), data = datos)
##
## Linear Hypotheses:
##        Estimate Std. Error t value  Pr(>t)
## 1 <= 0   2.1979     0.3969   5.538 < 0.001 ***
## 2 <= 0   1.4574     0.4861   2.998 0.00617 **
## 3 <= 0   3.8087     0.4861   7.835 < 0.001 ***
## 4 <= 0   1.6109     0.4861   3.314 0.00238 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Con esta prueba podemos ver que todos los p-value son menores a .05, por lo que podemos decir que el medicamento actual es mejor en los hombres mientras que el nuevo medicamento tiene mejor desempeño en las mujeres.

Ejercicio 5

I)



Observamos un aumento en los anticuerpos de la poblacion con medicamento apartir de cierta edad, lo cual podria significar una diferencia en la pendiente con respecto a la poblacion control

II)

Tenemos el modelo con interacciones:

$$E(y;x) = \beta_0 + \beta_1 + \beta_2 TratMed + \beta_3(Edad * TratMed)$$

```
##
## Call:
## lm(formula = Ant ~ Edad * Trat, data = Datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6211 -1.9539  0.0277  1.6018  7.0063
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  29.34298    0.57573  50.966 < 0.0000000000000002 ***
## Edad         -0.28290    0.01440 -19.645 < 0.0000000000000002 ***
## TratMed      -2.25730    0.96763  -2.333               0.0203 *
## Edad:TratMed  0.17307    0.02437   7.101     0.00000000000921 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 296 degrees of freedom
## Multiple R-squared:  0.6709, Adjusted R-squared:  0.6676
## F-statistic: 201.2 on 3 and 296 DF,  p-value: < 0.00000000000000022
```

Se rechaza la hipotesis de que todas las β´s sean cero.

III)

a) $E(y; Trat : Contol; Edad) = \beta_0 + \beta_1 Edad$

$E(y; Trat : Contol; Edad) = 29{,}34298 + (-0{,}2829) Edad$

b) $E(y; Trat : Med; Edad) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Edad$

$E(y; Trat : Med; Edad) = (29{,}34298 - 2{,}2573) + (-0{,}2829 + 0{,}17307) Edad$

IV)

Para corroborar si la edad afecta por igual a ambos grupos, se requiere la siguiente pueba de hipotesis:

$H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$

es decir se busca una diferencia en las pendientes

```
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##        Estimate
## 1 == 0   0.1731
##
## Global Test:
##       F DF1 DF2           Pr(>F)
## 1 50.43   1 296 0.00000000000921
```

De esta forma se rechaza H_0, por lo que se puede decir que la edad no afecta de la misma forma al grupo control y al grupo que se aplico el medicamento.

U)

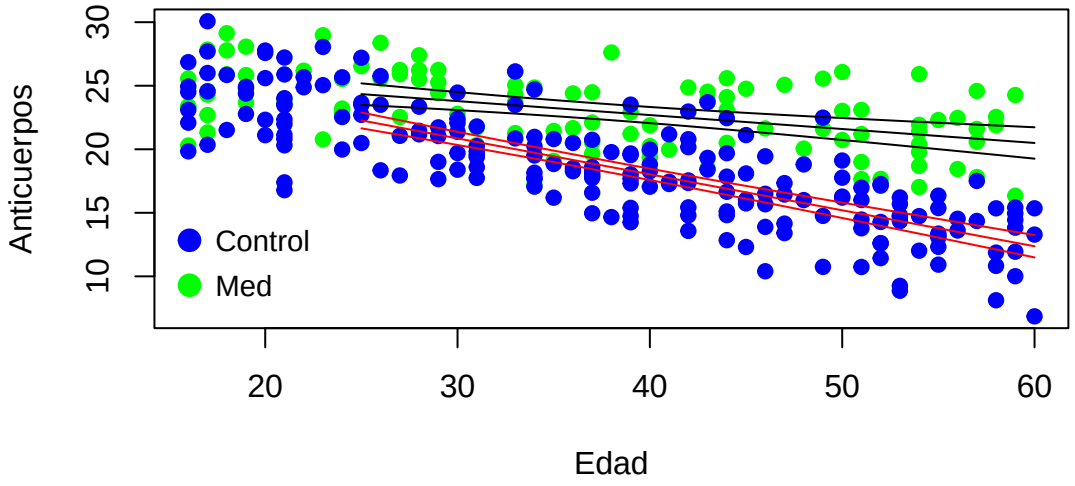
Este modelo parece indicar que el medicamento aumenta la produccion de anticuerpos pero dicho tratamiento es mas efectivo en edades avanzadas.

β_0 se podria interpreta como una aproximacion de anticuerpos que la personas cercanas a los 20 años tienen sin que se les haya aplicado el tratamiento, de esta forma al ser β_2 pequeña nos dice que hay poca diferencia con el numero de anticuerpos de a quienes si se les aplico y que igualmente tienen edad cercana a 20 años.

β_1 mide el como la edad afecta el numero de anticuerpos en el grupo sin tratamiento, al se β_1 negativa se podria decir que afecta negativamente, al ser β_3 positiva el tratamiento reduce los efectos negativos de la edad respecto al numero de anticuerpos.

VI)

Bajamos la confianza al 90 %



Observamos los intervalos de confianza no se intersectan en el rango de edades de 25 a 60 por lo que podemos decir que en estas edades el medicamento funciona.

Ejercicio 6

Con datos de `Ex6.csv` se considera un modelo de regresión lineal con las covariables X_1 a X_6 sin interacción. Se muestra el resultado en la primera columna del Cuadro de `MODELOS`, en donde se observa la prueba global F asociado a la tabla ANOVA cuyo p-value es menor a 0,05, por lo que se rechaza la hipótesis nula $H_0 : \hat{\beta}_i = 0, \forall i = 1, 2, ..., p$, a favor de la alternativa de que al menos un coeficiente estimado $\hat{\beta}_i$ es distinto de cero. Se observa además que dado que están las otras variables, la variable X3 no agrega información adicional al modelado, lo mismo para el caso de X5. (Chunk modelo1, línea de código 34).

Si hacemos una prueba visual de los supuestos del modelo, tales como la linealidad (Residuals vs Fitted), homocedasticidad (Scale-Location), normalidad (Q-Q Residuals) y presencia de outliers influyentes (Residuals vs Leverage), se observa que no se cumple la linealidad y normalidad, aunque parece no haber problemas con la homocedasticidad y la presencia de outliers influyentes (que se salgan de la distancia de Cook). (Chunk plotsmodelo1, línea de código 47)

De acuerdo con la prueba `studentized Breusch-Pagan` se tiene un p-value de 0.1576004 por lo que no se rechaza la hipótesis nula de homocesaticidad, mientras que las pruebas de normalidad Jarque-Bera, Shapiro-Wilk y Kolmogorov-Smirnov rechazan la hipótesis nula de normalidad, con p-value de 0, 0.0000005 y 0.0284986, respectivamente. (Chunk pruebasmodelo1, línea de código 56).

Para el caso de la linealidad, la prueba de Tukey rechaza la hipótesis nula de linealidad, particularmente con X4 y X6, el p-value asociado es menor a 0.05. (Chunk residualplotsmodelo1, línea de código 82).

```
##                                       X3                                       X4
## 0.77754089573502105370295112152234651 0.00122725412693692201640538463891517
##                                       X5                                       X6
## 0.09239394887177095061492337890740600 0.00000616869402705057755503109673256
##                                       X1                                       X2
##                                       NA                                       NA
##                                 Tukey test
## 0.00000000000000000000000000001185248
```

Esto se refleja en las gráficas individuales, para X4 y X6 es evidente la no linealidad. (Chunk residualPlots-modelo1, línea de código 96).

En las gráficas crPlots, se muestra que es posible ajustar tal vez un polinomio para X4 y X6. (Chunk crPlotsmodelo1, linea de código 109)

Haciendo la prueba `powerTransform` se rechaza la hipótesis nula de $\lambda = 1$ por lo que hay que hacer una transformación a la variable dependiente Y, en la misma prueba se rechaza la hipótesis nula de la transformación logarítimica $\lambda = 0$, y se presenta un valor sugerido de 2 como exponente. (Chunk powerTransform1, linea de código 119).

Luego, al hacer la prueba de `boxTidwell` para ver si se requiere transformar X4 y X6, tenemos que para el caso de X6 no se rechaza la hipótesis nula de una $\lambda = 1{,}032$, por lo que no requiere transformación, sin embargo, para el caso de X4, se muestra una $\lambda = 0{,}38193$ que redondearemos a $1/2$, por lo que tomaremos la raíz cuadrada de la variable X4 en el modelado. (Chunk boxTidwell1, linea de codigo 125).

Entonces, planteamos un segundo modelo que considera Y^2 en función de las covariables $X1$ a $X6$ considerando $\sqrt{X4}$ y los demás covariables sin cambios. Los resultados se muestran en la segunda columna del Cuadro de `MODELOS`, en donde se observa la prueba global F asociado a la tabla ANOVA cuyo p-value es menor a 0,05, por lo que se rechaza la hipótesis nula $H_0 : \hat{\beta}_i = 0, \forall i = 1, 2, ..., p$, a favor de la alternativa de que al menos un coeficiente estimado $\hat{\beta}_i$ es distinto de cero. Se observa además que dado que están las otras variables, la variable X3 no agrega información adicional al modelado, lo mismo para el caso de X5. (Chunk modelo2, linea de código 137).

Con este segundo modelo se cumple linealidad con la prueba Tukey, al igual que para todas las variables individuales. (Chunk residualplotsmodelo2, linea de código 144).

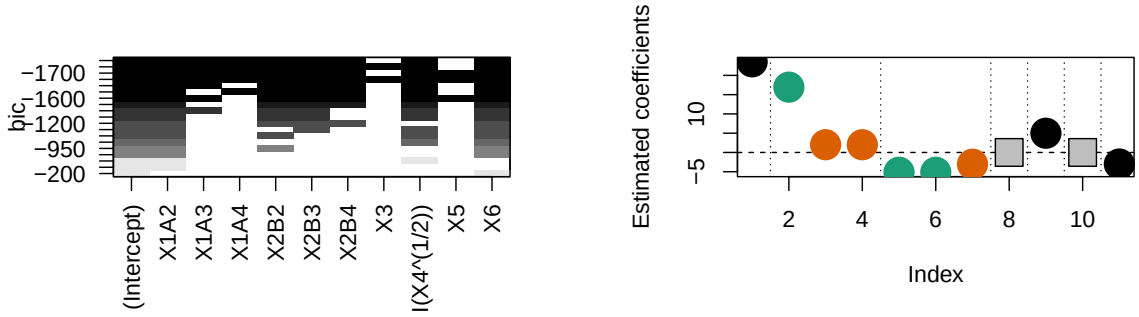
```
##          X1          X2          X3 I(X4^(1/2))          X5          X6
```

```
##           NA           NA    0.4045252    0.6665737    0.1516197    0.2226187
##  Tukey test
##   0.6844529
```

Además, de acuerdo con la prueba `studentized Breusch-Pagan` se tiene un p-value de 0.9656083 por lo que no se rechaza la hipótesis nula de homocesaticidad, mientras que las pruebas de normalidad Jarque-Bera, Shapiro-Wilk y Kolmogorov-Smirnov no rechazan la hipótesis nula de normalidad, con p-value de 0.5337811, 0.3336509 y 0.109361, respectivamente. (Chunk pruebasmodelo2, línea de código 158).

A partir de este segundo modelo, haremos una selección de variables.

Con la función `regsubsets` de la biblioteca `leaps` se analizaron los mejores subconjuntos, se presenta la siguiente gráfica izquierda mostrando los resultados, con esto decidimos tomar las covariables X1, X2, X4 y X6, omitiendo las variables X3 y X5, con lo que se podría tener un tercer modelo que muestra un BIC de 1236.7351928 y que se muestra en la columna 3 del Cuadro de MODELOS. (Chunk subconjuntosleaps, linea de código 187). Este resultado se refuerza con el análisis hecho con los métodos por pasos forward y backward, con la función `step` y el criterio BIC. (Chunk stepBIC, linea de código 207). Adicionalmente se realizó con la biblioteca `smurf` un modelo lasso de seleccion via penalizacion en la logverosimilitud, considerando la familia gausiana, pesos glm.stand, y lambda is.bic, cuyo BIC es de 1226.1559243 y resultado se muestra se presenta en la gráfica de la derecha (Chunk glm_lambdabic, linea de código 221). Como se puede observar en la gráfica derecha, las variables X3 y X5 tienen un valor de cero, además podemos notar que X1A3 y X1A4 pueden combinarse en una sola variable, al igual que X2B2 y X2B3 en otra, así se realizó el ajuste con el modelo lasso final cuyos resultados se presentan en la cuarta y última columna del Cuadro de MODELOS y cuyo BIC es de 1226.0633129. (Chunk modsellasso, linea de código 238).



Finalmente, se presentan los resultados de los modelos referidos anteriormente en el Cuadro de MODELOS. El modelo final elegido es el de la última columna donde la variable dependiente es $V1 = I((Y)\hat{}2)$, podemos observar que la prueba F tiene un p-value menor a 0,05, los tres asteriscos indican que incluso menor que 0,01, por lo que al menos un coeficiente estimado es distinto de cero. Además, cada uno de los coeficientes, en una análisis individual, dado que están las otras variables, agregan información al modelo.

MODELOS

	Dependent variable:			
	Y	I((Y)^2)		V1
	(1)	(2)	(3)	(4)
X3	0.002	0.036		
	(0.003)	(0.028)		
X4	0.066***			
	(0.003)			
I(X4^(1/2))		5.014***	5.018***	5.006***
		(0.167)	(0.167)	(0.167)
X5	−0.002	−0.016		
	(0.002)	(0.018)		
X6	−0.238***	−2.999***	−3.005***	−3.002***
	(0.003)	(0.028)	(0.028)	(0.028)
X1A2	1.295***	16.979***	16.967***	16.969***
	(0.014)	(0.152)	(0.151)	(0.151)
X1A3	0.177***	2.032***	2.030***	
	(0.014)	(0.151)	(0.151)	
X1A4	0.177***	2.019***	2.014***	
	(0.014)	(0.152)	(0.151)	
X2B2	−0.410***	−5.078***	−5.073***	
	(0.014)	(0.151)	(0.151)	
X2B3	−0.410***	−5.248***	−5.244***	
	(0.014)	(0.151)	(0.151)	
X1A3.4				2.023***
				(0.131)
X2B2.3				−5.159***
				(0.131)
X2B4	−0.235***	−3.035***	−3.047***	−3.048***
	(0.014)	(0.152)	(0.152)	(0.151)
Constant	5.550***	23.262***	23.410***	23.448***
	(0.030)	(0.548)	(0.541)	(0.540)
Observations	400	400	400	400
R^2	0.982	0.988	0.988	0.999
Adjusted R^2	0.981	0.988	0.988	0.999
Residual Std. Error	0.102 (df = 389)	1.065 (df = 389)	1.066 (df = 391)	1.065 (df = 393)
F Statistic	2,101.510*** (df = 10; 389)	3,204.967*** (df = 10; 389)	3,999.207*** (df = 8; 391)	89,411.300*** (df = 7; 393)
Note:				*p<0.1; **p<0.05; ***p<0.01