ECON-UB 251 Econometrics I Assignment 2

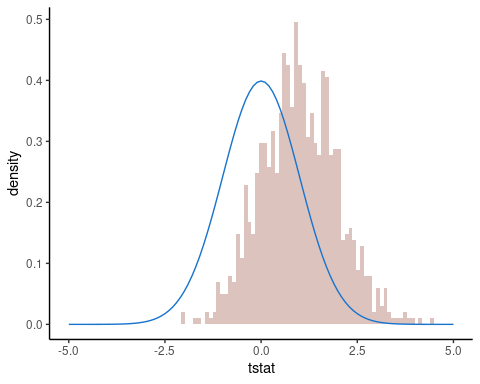
Kevin Song

# Theory

## 1.1

A simulation exercise means generating artificial data to evaluate the behavior of an estimator in that specific controlled environment.

library(ggplot2)  
  
set.seed(123)   
N = 500   
B = 1000   
beta1 <- rep(NA, B)  
tstat <- beta1  
rho <- beta1  
  
for (b in 1:B)   
{  
 X1 = rnorm(N)   
 X2 = 0.1 \* X1 + rnorm(N)   
 Y = 1.1 \* X1 + 0.5 \* X2 + rnorm(N)   
 rho[b] <- cor(X1, X2)   
 fit <- lm(Y ~ X1)   
 beta1[b] <- summary(fit)$coefficients[2,1]   
 tstat[b] <- (beta1[b] - 1.1) / summary(fit)$coefficients[2,2]  
}  
  
ggplot(data.frame(tstat = tstat), aes(x = tstat)) +  
 geom\_histogram(aes(y = ..density..), bins = 100, fill = "tomato4", alpha=0.3) +  
 stat\_function(fun = dnorm, args=list(mean = 0, sd = 1), color="dodgerblue3") +  
 theme\_classic() +  
 xlim(min(-5, min(tstat)), max(5, max(tstat)))



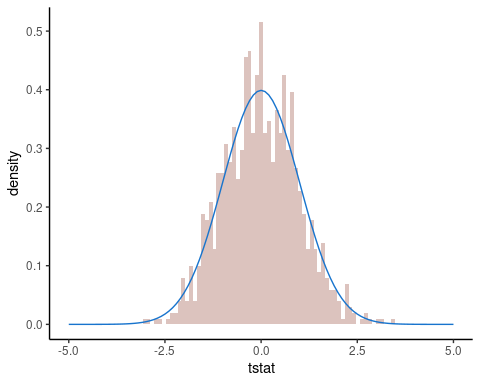
Under what conditions do we have OV bias? Did we design the simulation in a way to produce biased estimates? Why?

OV bias occurs when the regressor, x1, is correlated with an omitted regressor, x2, from the dependent variable y1. Knowing that, this simulation is designed to be bias because the formula to simulate y1 includes x2, but x2 is not included in the OLS.

The simulation results in terms of the distribution of t1 and whether you find evidence of OV bias. There is evidence of OV bias in the simulation because the entire histogram shifted right away from the normal distribution, despite all the variables having been generated as copies of the normal distribution. That must mean there is OV bias.

What would happen to the plot if we would generate X2 as X2i = ni instead of using X2i = 0.1X1i + ni? Since X2i would no longer be correlated with X1i, the OLS regression for Y should be normal as there will be no OV bias.

library(ggplot2)  
  
set.seed(123)   
N = 500   
B = 1000   
beta1 <- rep(NA, B)  
tstat <- beta1  
rho <- beta1  
  
for (b in 1:B)   
{  
 X1 = rnorm(N)   
 X2 = rnorm(N)   
 Y = 1.1 \* X1 + 0.5 \* X2 + rnorm(N)   
 rho[b] <- cor(X1, X2)   
 fit <- lm(Y ~ X1)   
 beta1[b] <- summary(fit)$coefficients[2,1]   
 tstat[b] <- (beta1[b] - 1.1) / summary(fit)$coefficients[2,2]  
}  
  
ggplot(data.frame(tstat = tstat), aes(x = tstat)) +  
 geom\_histogram(aes(y = ..density..), bins = 100, fill = "tomato4", alpha=0.3) +  
 stat\_function(fun = dnorm, args=list(mean = 0, sd = 1), color="dodgerblue3") +  
 theme\_classic() +  
 xlim(min(-5, min(tstat)), max(5, max(tstat)))



This histogram looks much more normal than the previous one.

# 2.1

library(readr)  
library(dplyr)  
sample\_orig\_2012 <- read\_delim("/cloud/project/sample\_orig\_2012.txt","|", escape\_double = FALSE, col\_names = FALSE, trim\_ws = TRUE, show\_col\_types = FALSE)  
  
orig.names <- c("Credit\_Score","First\_Payment \_ate","First\_Time\_Homebuyer", "Maturity\_Date",  
"MSA","Mortgage\_Insurance\_Percentage","Number\_Units","Occupancy\_Status","CLTV",  
"DTI","UPB","LTV","Interest\_Rate","Channel","Prepayment\_Penalty",  
"Amortization\_Type","State","Property\_Type","Postal\_Code","Sequence\_Number",  
"Purpose","Loan\_Term","Number\_Borrowers","Seller\_Name","Servicer\_Name",  
"Super\_Conforming","Pre-HARP\_Loan","Program\_Indicator","HARP\_Indicator",  
"Valuation\_Method","Interest\_Only")  
colnames(sample\_orig\_2012) <- orig.names  
  
sample\_orig\_2012 <- filter(sample\_orig\_2012, sample\_orig\_2012$Interest\_Rate != 999, sample\_orig\_2012$Credit\_Score != 999,  
 sample\_orig\_2012$DTI != 999, sample\_orig\_2012$UPB != 999, sample\_orig\_2012$LTV != 999)  
Interest\_Rate <- sample\_orig\_2012$Interest\_Rate  
Credit\_Score <- sample\_orig\_2012$Credit\_Score  
DTI <- sample\_orig\_2012$DTI  
UPB <- sample\_orig\_2012$UPB  
LTV <- sample\_orig\_2012$LTV  
  
sample\_orig\_2012 %>%  
 group\_by(State) %>%  
 summarize(N = n(),  
 IR = mean(Interest\_Rate),  
 CS = mean(Credit\_Score),  
 DTI = mean(DTI),  
 UPB = mean(UPB/1000),  
 LTV = mean(LTV)) %>%  
 arrange(desc(N)) %>%  
head(20) %>% knitr::kable(digits=3)

| State | N | IR | CS | DTI | UPB | LTV |
| --- | --- | --- | --- | --- | --- | --- |
| CA | 5005 | 3.696 | 771.159 | 33.126 | 317.598 | 60.673 |
| IL | 1801 | 3.550 | 765.418 | 30.672 | 199.917 | 68.444 |
| OH | 1481 | 3.422 | 770.156 | 27.486 | 158.125 | 69.355 |
| TX | 1454 | 3.654 | 757.961 | 31.298 | 198.444 | 74.355 |
| MA | 1348 | 3.612 | 767.099 | 31.056 | 268.606 | 64.030 |
| VA | 1208 | 3.649 | 768.041 | 30.429 | 278.159 | 66.358 |
| WI | 1097 | 3.483 | 767.392 | 29.814 | 171.104 | 69.500 |
| NC | 1062 | 3.628 | 767.006 | 29.443 | 192.472 | 69.692 |
| PA | 1050 | 3.559 | 764.774 | 31.078 | 193.212 | 69.850 |
| NY | 1046 | 3.624 | 760.969 | 33.061 | 233.612 | 66.142 |
| MI | 960 | 3.497 | 766.174 | 29.794 | 149.418 | 71.130 |
| IN | 933 | 3.508 | 761.987 | 27.771 | 146.989 | 70.398 |
| FL | 926 | 3.702 | 765.786 | 32.371 | 181.593 | 70.366 |
| MN | 910 | 3.513 | 769.433 | 30.025 | 196.470 | 70.934 |
| MO | 888 | 3.600 | 767.245 | 29.973 | 174.082 | 70.865 |
| NJ | 842 | 3.641 | 761.343 | 33.112 | 276.939 | 65.350 |
| WA | 804 | 3.653 | 771.261 | 31.113 | 238.463 | 65.973 |
| MD | 788 | 3.624 | 767.251 | 30.204 | 282.867 | 65.753 |
| CO | 748 | 3.674 | 767.207 | 30.841 | 239.523 | 68.789 |
| KY | 660 | 3.487 | 762.697 | 28.053 | 156.662 | 69.883 |

Interest Rate and Credit Score both seem to be very similar across different states. Interest rates fluctuate between 3.6 and 3.9 mainly, and credit scores mainly between 750 and 770. Debt to income is also pretty consistently between 28 and 33. Unpaid principal balance is the most varying, especially since it is divided by 1000, meaning the actual values are 1000 times larger. Ohio has a UPB/1000 of 158.125 while California’s is 317.598 . signalling that mortgage prices vary highly between states. LTV seems to be relatively consistent with most states in the seventies, though some go to above 90 or below 70.

# 2.2

library(dplyr)  
prime<-subset(sample\_orig\_2012,Credit\_Score >670)  
subprime <- subset(sample\_orig\_2012, Credit\_Score < 670)  
  
prime %>%  
 group\_by(State) %>%  
 summarize(N = n(),  
 IR = mean(Interest\_Rate),  
 CS = mean(Credit\_Score),  
 DTI = mean(DTI),  
 UPB = mean(UPB/1000),  
 LTV = mean(LTV)) %>%  
 arrange(desc(N)) %>%  
head(20) %>% knitr::kable(digits=3)

| State | N | IR | CS | DTI | UPB | LTV |
| --- | --- | --- | --- | --- | --- | --- |
| CA | 4928 | 3.692 | 773.013 | 33.061 | 318.714 | 60.636 |
| IL | 1755 | 3.547 | 768.422 | 30.615 | 201.129 | 68.311 |
| OH | 1455 | 3.418 | 772.215 | 27.474 | 158.775 | 69.373 |
| TX | 1398 | 3.650 | 762.147 | 31.141 | 199.677 | 74.494 |
| MA | 1320 | 3.606 | 769.476 | 30.936 | 269.422 | 64.105 |
| VA | 1185 | 3.646 | 770.232 | 30.323 | 279.815 | 66.493 |
| WI | 1073 | 3.476 | 769.881 | 29.731 | 172.315 | 69.419 |
| NC | 1040 | 3.620 | 769.495 | 29.357 | 193.856 | 69.839 |
| PA | 1026 | 3.553 | 767.371 | 31.054 | 194.135 | 69.818 |
| NY | 1012 | 3.619 | 764.765 | 32.914 | 236.232 | 66.255 |
| MI | 940 | 3.495 | 768.526 | 29.760 | 149.537 | 71.095 |
| IN | 907 | 3.504 | 765.083 | 27.699 | 148.418 | 70.473 |
| FL | 901 | 3.692 | 768.958 | 32.297 | 182.345 | 70.415 |
| MN | 890 | 3.509 | 771.994 | 29.955 | 196.409 | 71.113 |
| MO | 877 | 3.598 | 768.708 | 29.867 | 174.572 | 70.855 |
| NJ | 823 | 3.637 | 763.808 | 32.984 | 277.705 | 65.329 |
| WA | 791 | 3.650 | 773.319 | 31.058 | 239.295 | 66.024 |
| MD | 773 | 3.620 | 769.498 | 30.163 | 284.543 | 65.690 |
| CO | 733 | 3.674 | 769.596 | 30.880 | 239.889 | 68.768 |
| KY | 640 | 3.486 | 766.167 | 28.006 | 156.967 | 70.000 |

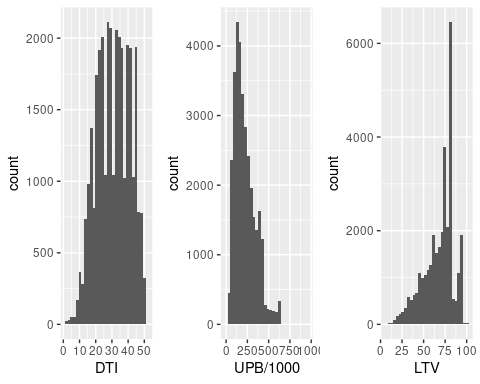
subprime %>%  
 group\_by(State) %>%  
 summarize(N = n(),  
 IR = mean(Interest\_Rate),  
 CS = mean(Credit\_Score),  
 DTI = mean(DTI),  
 UPB = mean(UPB/1000),  
 LTV = mean(LTV)) %>%  
 arrange(desc(N)) %>%  
head(20) %>% knitr::kable(digits=3)

| State | N | IR | CS | DTI | UPB | LTV |
| --- | --- | --- | --- | --- | --- | --- |
| CA | 75 | 3.971 | 652.067 | 37.093 | 241.987 | 63.173 |
| TX | 50 | 3.787 | 651.460 | 34.620 | 165.560 | 70.040 |
| IL | 45 | 3.658 | 650.356 | 32.867 | 149.311 | 73.178 |
| NY | 31 | 3.798 | 645.871 | 38.097 | 150.935 | 61.806 |
| MA | 27 | 3.931 | 654.481 | 36.630 | 223.222 | 59.815 |
| FL | 25 | 4.048 | 651.480 | 35.040 | 154.480 | 68.600 |
| IN | 25 | 3.685 | 653.360 | 30.400 | 97.760 | 67.880 |
| OH | 25 | 3.659 | 654.320 | 28.440 | 121.400 | 69.640 |
| PA | 23 | 3.804 | 653.043 | 31.609 | 153.391 | 72.174 |
| VA | 22 | 3.824 | 654.500 | 35.955 | 182.636 | 59.682 |
| MI | 20 | 3.592 | 655.650 | 31.400 | 143.800 | 72.800 |
| MN | 20 | 3.731 | 655.450 | 33.150 | 199.200 | 62.950 |
| NC | 20 | 4.006 | 647.250 | 33.150 | 125.650 | 64.100 |
| KY | 19 | 3.555 | 650.684 | 30.158 | 153.053 | 67.000 |
| TN | 19 | 3.774 | 650.895 | 33.526 | 195.579 | 68.789 |
| NJ | 18 | 3.847 | 653.722 | 39.333 | 246.278 | 67.833 |
| WI | 18 | 3.771 | 651.500 | 35.278 | 105.667 | 73.056 |
| GA | 16 | 4.093 | 652.062 | 32.062 | 161.438 | 75.500 |
| CO | 15 | 3.667 | 650.467 | 28.933 | 221.600 | 69.800 |
| MD | 14 | 3.907 | 650.143 | 32.500 | 189.714 | 69.857 |

From the data, it can be observed that prime borrowers have lower interest rates, <4, than subprime borrowers, where in many states have rates > 4. Additionally, debt to income is on average higher for prime borrowers, because their higher credit score means they’re more likely to pay off debt, even if the ratio of debt to their income is high. UPB/1000 also seems to be on average smaller for subprime clients, which possibly signals that lower credit score clients usually mortgage for lower value homes. LTV seems to be relatively equal between prime and subprime borrowers, which makes sense as it is a proportion between the mortgage and value, thus not affected by literal changes in price.

# 3.1

library(dplyr)  
library(ggplot2)  
  
p1 <- ggplot(sample\_orig\_2012, aes(DTI)) + geom\_histogram( )  
p2 <- ggplot(sample\_orig\_2012, aes(UPB/1000)) + geom\_histogram()  
p3 <- ggplot(sample\_orig\_2012, aes(LTV)) + geom\_histogram()  
gridExtra::grid.arrange(p1,p2,p3,ncol = 3)



All 3 variables have skew. DTI has a slight left skew, UPB/1000 has right skew, and LTV has a large left skew, with many data points gradually deviating from the mode. All 3 variables have very high modes, with thousands of data points at 28 for DTI, 200 for UPB/1000, and 80 for LTV. The only graph with outliers seems to be UPB/1000, where there are a few data points above the 750 point, though it may be a bit hard to see them.

These variables should be pricing factors of the interest rate because they all relate to the value or risk of a loan. A high DTI is riskier than a lower one, as it would take more years of income to pay off a high ratio. UPB is the literal amount one still has to pay, which directly relates to a potential loss if the client defaults. LTV corresponds with one’s down payment, so a higher LTV means a larger/longer mortgage which is riskier for the bank.

# 3.2

cor(sample\_orig\_2012[, c('Credit\_Score', 'DTI', 'UPB', 'LTV')])

Credit\_Score DTI UPB LTV  
Credit\_Score 1.00000000 -0.17244094 0.03170205 -0.13921964  
DTI -0.17244094 1.00000000 0.09954079 0.09888268  
UPB 0.03170205 0.09954079 1.00000000 0.05481868  
LTV -0.13921964 0.09888268 0.05481868 1.00000000

Based on these correlations, none of them are above |0.2|, meaning weak correlations so I would not be particularly worried about multicollinearlity.

# 4.1

library(ggplot2)  
library(dplyr)  
regression <- lm(Interest\_Rate~Credit\_Score+DTI+I(UPB/1000)+LTV, data = sample\_orig\_2012)  
  
vcov<- sandwich::vcovHC(regression, type = "HC1")  
fit.robus <- lmtest::coeftest(regression, vcov. = vcov)  
summary(regression)

Call:  
lm(formula = Interest\_Rate ~ Credit\_Score + DTI + I(UPB/1000) +   
 LTV, data = sample\_orig\_2012)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1.58594 -0.28321 0.03213 0.29793 2.15599   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.978e+00 5.186e-02 76.71 <2e-16 \*\*\*  
Credit\_Score -1.194e-03 6.299e-05 -18.95 <2e-16 \*\*\*  
DTI 4.643e-03 2.412e-04 19.25 <2e-16 \*\*\*  
I(UPB/1000) 2.470e-04 1.941e-05 12.72 <2e-16 \*\*\*  
LTV 5.024e-03 1.389e-04 36.17 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4266 on 32539 degrees of freedom  
Multiple R-squared: 0.08014, Adjusted R-squared: 0.08003   
F-statistic: 708.7 on 4 and 32539 DF, p-value: < 2.2e-16

The coefficient of Credit\_Score means that or every increase in 1 credit score, interest rate decreases by 0.001194.

They are positive, meaning higher values mean higher interest rates. This makes sense as high debt to income is more risk, high unpaid balance is more risk, and high loan to value corresponds to higher unpaid balance which is more risk.

All the variables have p values of less than 2e^-16, meaning they are all statistically significant at the 5% level. All 4 variables have an impact on interest rate.

My r-squared for Interest\_Rate~Credit\_Score was 0.04685. This r-squared with 4 variables is 0.08014. That is nearly double, but still very small and thus does not explain much of the variance in interest rates still.

The coefficient estimate of Credit\_Score went from -2.338e-03 to -1.194e-03. The standard error went from 0.00004717 to 0.00006299. Using:

results in Z = -14.54, meaning p-value < 0.001 This is a significant change meaning DTI, UPB, and LTV were previously omitted variables.

# 4.2

library(lmtest)  
library(sandwich)  
model <- lm(Interest\_Rate~Credit\_Score, data = sample\_orig\_2012)  
model\_unres <- lm(Interest\_Rate~Credit\_Score+DTI+I(UPB/1000)+LTV, data = sample\_orig\_2012)  
  
waldtest(model, model\_unres, vcov = vcovHC(model\_unres, type = "HC0"))

Wald test  
  
Model 1: Interest\_Rate ~ Credit\_Score  
Model 2: Interest\_Rate ~ Credit\_Score + DTI + I(UPB/1000) + LTV  
 Res.Df Df F Pr(>F)   
1 32542   
2 32539 3 733.39 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The heteroskedastic robust F-statistic for beta\_DTI = beta\_UDP = beta\_LTV = 0 is 733.39. This is smaller than the 0.05 significance level. This means the added variables DTI, UPB/1000, and LTV are statistically significant and improve the model.

# 4.3

regression <- lm(Interest\_Rate~Credit\_Score+DTI+I(UPB/1000)+LTV+State, data = sample\_orig\_2012)  
summary(regression)

Call:  
lm(formula = Interest\_Rate ~ Credit\_Score + DTI + I(UPB/1000) +   
 LTV + State, data = sample\_orig\_2012)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1.56529 -0.28011 0.02471 0.29339 2.05283   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.947e+00 6.861e-02 57.529 < 2e-16 \*\*\*  
Credit\_Score -1.237e-03 6.206e-05 -19.926 < 2e-16 \*\*\*  
DTI 3.592e-03 2.396e-04 14.990 < 2e-16 \*\*\*  
I(UPB/1000) -4.869e-05 2.167e-05 -2.247 0.024670 \*   
LTV 6.056e-03 1.421e-04 42.626 < 2e-16 \*\*\*  
StateAL 5.020e-02 5.366e-02 0.936 0.349527   
StateAR -9.419e-03 5.508e-02 -0.171 0.864221   
StateAZ 1.734e-01 4.941e-02 3.509 0.000451 \*\*\*  
StateCA 2.318e-01 4.691e-02 4.941 7.81e-07 \*\*\*  
StateCO 1.600e-01 4.892e-02 3.271 0.001074 \*\*   
StateCT 1.008e-01 5.023e-02 2.006 0.044835 \*   
StateDC 2.715e-01 6.074e-02 4.469 7.88e-06 \*\*\*  
StateDE 5.834e-02 6.036e-02 0.967 0.333777   
StateFL 1.680e-01 4.849e-02 3.466 0.000530 \*\*\*  
StateGA 5.430e-02 4.963e-02 1.094 0.273895   
StateGU -8.957e-02 1.470e-01 -0.609 0.542209   
StateHI 1.766e-01 5.820e-02 3.035 0.002408 \*\*   
StateIA -1.417e-01 5.030e-02 -2.816 0.004859 \*\*   
StateID 7.471e-02 5.769e-02 1.295 0.195286   
StateIL 3.469e-02 4.751e-02 0.730 0.465293   
StateIN -1.513e-02 4.850e-02 -0.312 0.754999   
StateKS -1.806e-02 5.063e-02 -0.357 0.721221   
StateKY -3.278e-02 4.928e-02 -0.665 0.505867   
StateLA 1.074e-01 5.434e-02 1.977 0.048019 \*   
StateMA 1.277e-01 4.786e-02 2.669 0.007608 \*\*   
StateMD 1.333e-01 4.881e-02 2.730 0.006337 \*\*   
StateME 1.342e-01 5.534e-02 2.425 0.015322 \*   
StateMI -3.323e-02 4.844e-02 -0.686 0.492775   
StateMN -9.883e-03 4.851e-02 -0.204 0.838567   
StateMO 7.318e-02 4.857e-02 1.507 0.131902   
StateMS 1.065e-02 6.966e-02 0.153 0.878432   
StateMT 7.748e-02 6.126e-02 1.265 0.205918   
StateNC 1.110e-01 4.823e-02 2.302 0.021349 \*   
StateND -8.765e-02 6.638e-02 -1.320 0.186706   
StateNE 6.370e-02 5.464e-02 1.166 0.243721   
StateNH 4.088e-02 5.338e-02 0.766 0.443767   
StateNJ 1.345e-01 4.867e-02 2.764 0.005709 \*\*   
StateNM 1.484e-01 5.904e-02 2.514 0.011952 \*   
StateNV 2.121e-01 5.969e-02 3.553 0.000381 \*\*\*  
StateNY 1.101e-01 4.824e-02 2.282 0.022467 \*   
StateOH -8.333e-02 4.777e-02 -1.744 0.081100 .   
StateOK 1.125e-01 5.504e-02 2.044 0.040950 \*   
StateOR 1.561e-01 5.026e-02 3.105 0.001904 \*\*   
StatePA 3.211e-02 4.824e-02 0.665 0.505748   
StatePR -8.077e-04 2.459e-01 -0.003 0.997379   
StateRI 1.519e-01 6.026e-02 2.520 0.011727 \*   
StateSC 1.218e-01 5.073e-02 2.401 0.016369 \*   
StateSD -1.985e-02 8.301e-02 -0.239 0.811024   
StateTN 6.044e-02 5.050e-02 1.197 0.231401   
StateTX 9.107e-02 4.776e-02 1.907 0.056573 .   
StateUT 2.505e-02 5.043e-02 0.497 0.619347   
StateVA 1.542e-01 4.801e-02 3.212 0.001320 \*\*   
StateVT 7.427e-02 5.711e-02 1.300 0.193443   
StateWA 1.606e-01 4.876e-02 3.293 0.000992 \*\*\*  
StateWI -3.468e-02 4.818e-02 -0.720 0.471720   
StateWV 1.320e-01 6.327e-02 2.086 0.036944 \*   
StateWY 3.968e-02 7.390e-02 0.537 0.591267   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4182 on 32487 degrees of freedom  
Multiple R-squared: 0.1175, Adjusted R-squared: 0.1159   
F-statistic: 77.21 on 56 and 32487 DF, p-value: < 2.2e-16

Maine has a p value of 0.015322 with positive t stat, and Iowa has a p value of 0.004859 with negative t stat. The coefficient of of Maine is 0.01342, meaning a client from Maine changes their predicted interest rate by 0.01342. The coefficient of Iowa is -0.01417, meaning a client from Iowa changes their predicted interest rate by -0.01417.

Overall, the inclusion of the state variable does significantly change the estimate of Credit\_Score, as well as DTI, UDP, and LTV. Using:

beta1 is -0.001237 and beta2 is -0.00194. SEbeta1 is then 0.00006206 and SEbeta2 is 0.00006299. Plugging these values into the formula gives a z score of -35.92, giving a p value less than 0.0001. This means that despite the small literal change in value of Credit\_Score, the standard error and large sample size actually shows this change is still significant. We can also then conclude the other variables DTI, UDP, and LTV are significantly different too because they all have similar standard errors and result in similar p values < 0.0001.