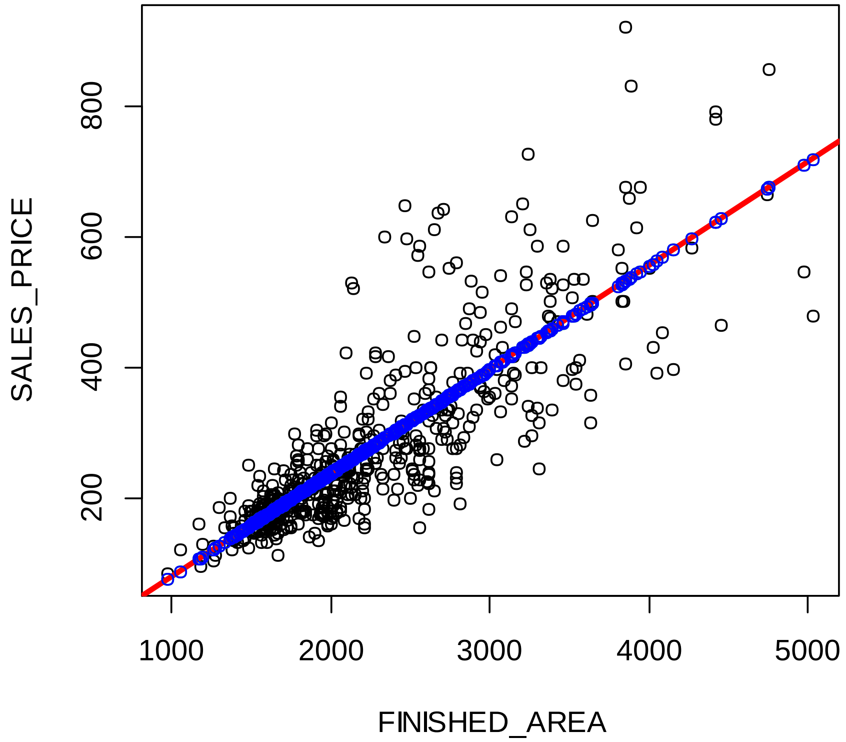
R Lab 3. Univariate Linear Regression

```
> H = read.csv("HOME_SALES.csv")
> attach(H)
> names(H)
 [1] "ID"               "SALES_PRICE"      "FINISHED_AREA"    "BEDROOMS"
 [5] "BATHROOMS"        "GARAGE_SIZE"      "YEAR_BUILT"       "STYLE"
 [9] "LOT_SIZE"         "AIR_CONDITIONER" "POOL"             "QUALITY"
[13] "HIGHWAY"
> plot(FINISHED_AREA, SALES_PRICE)
```

Familiar stuff so far. Now, we are fitting a regression model that we can use to predict the house sales price based on its area. So, X = area, Y = price.
reg – conducts regression analysis, estimates regression slope and intercept
abline – graphs the sample regression line in red
Yhat – computes predicted values based on the obtained regression equation
points – plots these predicted values in blue

```
> reg = lm( SALES_PRICE ~ FINISHED_AREA )
> abline(reg,col="red",lwd=3)
> Yhat = predict(reg, x=FINISHED_AREA)
> points(FINISHED_AREA, Yhat, col="blue")
```



Prediction. Predict the price for three houses that have the finished area of 2500, 4000, and 6000 square feet.

```
> predict(reg,data.frame(FINISHED_AREA=c(2500,4000,6000)))
       1        2        3
315.9426 554.3680 872.2684
```

Inference. Use “summary” to see results of the regression analysis.

```
> summary(reg)

Call:
lm(formula = SALES_PRICE ~ FINISHED_AREA)

Residuals:
    Min      1Q  Median      3Q     Max
-239.40  -39.84   -7.64   23.52  388.36

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -81.432946  11.551846  -7.049 5.74e-12 ***
FINISHED_AREA   0.158950   0.004875  32.605  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.12 on 520 degrees of freedom
Multiple R-squared:  0.6715,    Adjusted R-squared:  0.6709
F-statistic:  1063 on 1 and 520 DF,  p-value: < 2.2e-16
```

Conclusion: the sample regression equation is Price = -81.4 + 0.159(area). The slope and the intercept are both significant. The area can actually be used as an important factor to predict the sales price. This variable alone explains 67.15% of the total variation of house sales prices.

Analysis of Variance

```
> anova(reg)
Analysis of Variance Table

Response: SALES_PRICE
               Df  Sum Sq Mean Sq F value    Pr(>F)
FINISHED_AREA   1 6655486 6655486  1063.1 < 2.2e-16 ***
Residuals     520 3255426    6260
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```