Linear Regression

Data Set

h2o

```
ames <- AmesHousing::make_ames()
ames.h2o <- as.h2o(ames)
```

stratified (*Sale_Price*) training sample

```
set.seed(123)

split <- initial_split(ames, prop = 0.7,
                       strata = "Sale_Price")

ames_train <- training(split)
ames_test <- testing(split)
```

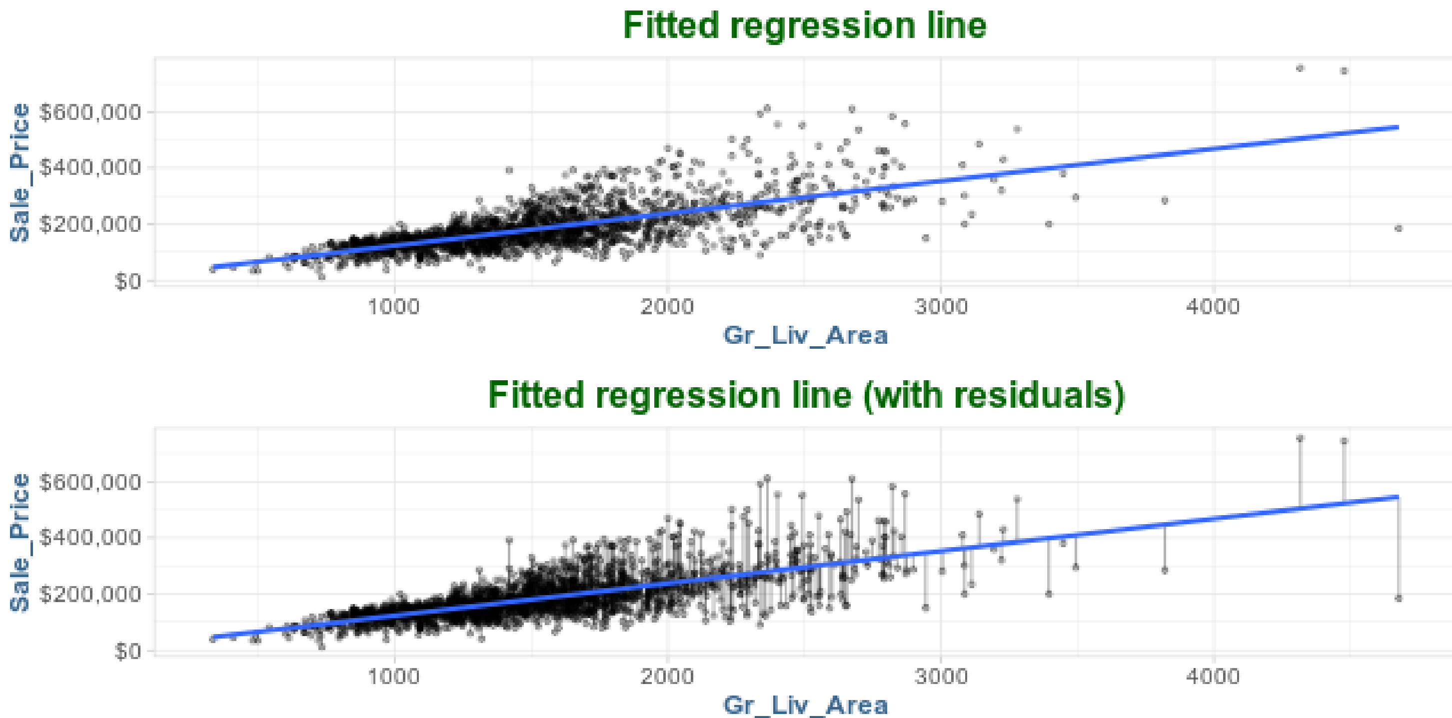
Simple Linear Model

```
model1 <- lm(Sale_Price ~ Gr_Liv_Area, data = ames_train)
```

```
# Fitted regression line (full training)
p1 <- model1 %>%
   broom::augment() %>%
   ggplot(aes(Gr_Liv_Area, Sale_Price)) +
   geom_point(size = 1, alpha = 0.3) +
   geom_smooth(se = F, method = "lm") +
   scale_y_continuous(labels = scales::dollar) +
   ggtitle("Fitted regression line")

# Fitted regression line (restricted range)
p2 <- model1 %>%
   broom::augment() %>%
   ggplot(aes(Gr_Liv_Area, Sale_Price)) +
   geom_segment(aes(x = Gr_Liv_Area, y = Sale_Price,
                    xend = Gr_Liv_Area, yend = .fitted),
                alpha = .3) +
   geom_point(size = 1, alpha = 0.3) +
   geom_smooth(se = F, method = "lm") +
   scale_y_continuous(labels = scales::dollar) +
   ggtitle("Fitted regression line (with residuals)")
```

```r
grid.arrange(p1, p2, nrow = 2)
```




```r
summary(model1)
```

```

Call:
lm(formula = Sale_Price ~ Gr_Liv_Area, data = ames_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-361143  -30668   -2449   22838  331357 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8732.938   3996.613   2.185    0.029 *  
Gr_Liv_Area  114.876      2.531  45.385   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56700 on 2051 degrees of freedom
Multiple R-squared:  0.5011,    Adjusted R-squared:  0.5008 
F-statistic:  2060 on 1 and 2051 DF,  p-value: < 2.2e-16
```

```
[1] 56704.78
```

```
[1] 3215432370
```

Inference

The variability of an estimate is its *standard error (SE)*, the square root of its variance.

t-test for the coefficents are simply the estimated coefficent divided by the standard error (t value = Estimate / Std. Error)

t-test measure the number of standard deviations each coefficent is away from zero (basically abs(T) > 2 is significant at 95% conf)

The confidence interval for coefficents is:

$\hat{\beta}_j \pm t_{1-\alpha/2,n-p} \hat{SE}(\hat{\beta}_j)$

```
confint(model1, level = .95)
```

```
                2.5 %     97.5 %
(Intercept) 895.0961 16570.7805
Gr_Liv_Area 109.9121   119.8399
```