COMPARING COEFFICIENTS FOR MODELS WITH SAME and DIFFERENT X VARIABLES

```
> HouseFit1=lm(log(Salesprice)~ SquareFt + AC + Quality, data = HouseData)
> summary (HouseFit1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.216e+01  8.310e-02 146.327   <2e-16 ***
SquareFt     3.445e-04  1.715e-05  20.086   <2e-16 ***
AC           5.361e-02  2.568e-02   2.087   0.0374 *
Quality     -2.512e-01  2.012e-02 -12.485   <2e-16 ***

> HouseFit2=lm(log(Salesprice)~ AC + Quality + SquareFt, data = HouseData)
> summary(HouseFit2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.216e+01  8.310e-02 146.327   <2e-16 ***
AC           5.361e-02  2.568e-02   2.087   0.0374 *
Quality     -2.512e-01  2.012e-02 -12.485   <2e-16 ***
SquareFt     3.445e-04  1.715e-05  20.086   <2e-16 ***
---

> HouseFit4=lm(log(Salesprice)~ Bedrooms + Bathrooms + LotSize, data = HouseData)
> summary (HouseFit4)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.148e+01  4.900e-02 234.276  < 2e-16 ***
Bedrooms    2.904e-02  1.504e-02   1.931    0.054 .
Bathrooms   2.777e-01  1.438e-02  19.314  < 2e-16 ***
LotSize     4.913e-06  1.072e-06   4.583 5.74e-06 ***


> HouseFit5=lm(log(Salesprice)~ SquareFt + LotSize, data = HouseData)
> summary(HouseFit5)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.121e+01  3.726e-02 300.704  < 2e-16 ***
SquareFt    4.989e-04  1.435e-05  34.774  < 2e-16 ***
LotSize     4.167e-06  8.731e-07   4.772 2.37e-06 ***
```

LESSON:
When the **same X variables** are in the model, the **coefficients, standard errors, t*, p-value** all remain the same. The t-test is testing that each coefficient = 0, *given* that the other X variables are in the model.

When **different X variables** are in the model, **coefficients, standard errors, t*, p-value** all change. So, the coefficient for one X variable depends on what *other* X variables are in the model.

COMPARING Residual SS and df, Multiple R^2, Adjusted R^2, F Statistic, df ACROSS MODELS

```
> HouseFit1=lm(log(Salesprice)~ SquareFt + AC + Quality, data = HouseData)
> summary (HouseFit1)

Residual standard error: 0.1999 on 518 degrees of freedom
Multiple R-Squared: 0.7868,     Adjusted R-squared: 0.7855
F-statistic:   637 on 3 and 518 DF,  p-value: < 2.2e-16

> HouseFit2=lm(log(Salesprice)~ AC + Quality + SquareFt, data = HouseData)
> summary(HouseFit2)

Residual standard error: 0.1999 on 518 degrees of freedom
Multiple R-Squared: 0.7868,     Adjusted R-squared: 0.7855
F-statistic:   637 on 3 and 518 DF,  p-value: < 2.2e-16



> HouseFit4=lm(log(Salesprice)~ Bedrooms + Bathrooms + LotSize, data = HouseData)
> summary (HouseFit4)

Residual standard error: 0.2824 on 518 degrees of freedom
Multiple R-Squared: 0.5744,     Adjusted R-squared: 0.572
F-statistic: 233.1 on 3 and 518 DF,  p-value: < 2.2e-16



> HouseFit5=lm(log(Salesprice)~ SquareFt + LotSize, data = HouseData)
> summary(HouseFit5)

Residual standard error: 0.23 on 519 degrees of freedom
Multiple R-Squared: 0.7173,     Adjusted R-squared: 0.7162
F-statistic: 658.5 on 2 and 519 DF,  p-value: < 2.2e-16
```

LESSON:
When the **same X variables** are in the model, what stays the same:

- **Residual standard error** (what the book calls **s**)
- **Df for residual standard error = n – p**
- **Multiple R-Squared** and **Adjusted R-squared**
- **F-Statistic (F*), degrees of freedom (p – 1** and **n – p)**

When **different X variables** are in the model:
None of the above stay the same, *except* **degrees of freedom** stay the same, *if* **p** is the same. In other words, if two models have the same *number* of explanatory variables.

COMPARING ANOVA TABLE RESULTS ACROSS MODELS

```
> HouseFit1=lm(log(Salesprice)~ SquareFt + AC + Quality, data = HouseData)
> anova (HouseFit1)

Response: log(Salesprice)
           Df Sum Sq Mean Sq  F value    Pr(>F)
SquareFt    1 68.435  68.435 1712.293 < 2.2e-16 ***
AC          1  1.716   1.716   42.929 1.372e-10 ***
Quality     1  6.229   6.229  155.863 < 2.2e-16 ***
Residuals 518 20.703   0.040
```

NOTE: The SS here sum to SSR = 68.435 + 1.716 + 6.229 = 76.38

NOTE: SSTO = SSR+SSE = 76.38 + 20.703 = 97.083

```
> HouseFit2=lm(log(Salesprice)~ AC + Quality + SquareFt, data = HouseData)

           Df Sum Sq Mean Sq F value    Pr(>F)
AC          1 12.101  12.101  302.77 < 2.2e-16 ***
Quality     1 48.155  48.155 1204.89 < 2.2e-16 ***
SquareFt    1 16.124  16.124  403.43 < 2.2e-16 ***
Residuals 518 20.703   0.040
```

NOTE: The SS here sum to SSR = 12.101 + 48.155 + 16.124 = 76.38

NOTE: SSTO = SSR+SSE = 76.38 + 20.703 = 97.083

```
> HouseFit3=lm(log(Salesprice)~ Quality + SquareFt + AC, data = HouseData)

           Df Sum Sq Mean Sq  F value  Pr(>F)
Quality     1 60.169  60.169 1505.465 < 2e-16 ***
SquareFt    1 16.037  16.037  401.264 < 2e-16 ***
AC          1  0.174   0.174    4.356 0.03737 *
Residuals 518 20.703   0.040
```

NOTE: The SS here sum to SSR = 60.169 + 16.037 + 0.174 = 76.38

NOTE: SSTO = SSR+SSE = 76.38 + 20.703 = 97.083

```
> HouseFit4=lm(log(Salesprice)~ Bedrooms + Bathrooms + LotSize, data = HouseData)

           Df Sum Sq Mean Sq F value    Pr(>F)
Bedrooms    1 22.782  22.782 285.633 < 2.2e-16 ***
Bathrooms   1 31.309  31.309 392.528 < 2.2e-16 ***
LotSize     1  1.676   1.676  21.007 5.742e-06 ***
Residuals 518 41.316   0.080
```

NOTE: SSR = 55.767

NOTE: SSTO = SSR+SSE = 55.767 + 41.316 = 97.083

```
> HouseFit5=lm(log(Salesprice)~ SquareFt + LotSize, data = HouseData)

           Df Sum Sq Mean Sq  F value    Pr(>F)
SquareFt    1 68.435  68.435 1294.211 < 2.2e-16 ***
LotSize     1  1.204   1.204   22.776 2.369e-06 ***
Residuals 519 27.444   0.053
```

NOTE: SSR=68.435 + 1.204 = 69.639

NOTE: SSTO = SSR+SSE = 69.639 + 27.444 = 97.083

LESSON:
When the **same X variables** are used, **SSE and MSE (**listed in the **Residuals** row) stay the same.
The **SSR and MSR combined** also remain the same. Add up the individual SS for the variables.
But the **SSR for each X variable changes.** It measures the *addition* to SSR when that variable is added, *given* that the ones *listed before it* are in the model already, *not the ones listed after it.*

When **different X variables** are in the model, even if **p** remains the same, **SSE and MSE change.**

ONLY SSTO remains the same, across all models with the same Y variable. Remember that SSTO does not depend on the X variables at all. But SSTO = SSR + SSE.