Stat 306: Output of lm() in R

```
b=read.table("burnabycondo.txt",header=T,skip=2)  # sample size n=69
b=b[,2:9]
b$askprice=b$askprice/10000; b$ffarea=b$ffarea/100;
b$mfee=b$mfee/10;  b$sqfl=sqrt(b$floor)
attach(b)
bur=lm(askprice~ ffarea+beds+baths+sqfl+view+age+mfee)
print(summary(bur))

Residuals:  (check if there appears to be an extreme residual)
     Min       1Q   Median       3Q      Max
-11.4276  -3.3001  -0.5997   2.7321  13.6735

Coefficients Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.20635    3.35351   2.447 0.017297 *    (ignore t-test for beta0)
ffarea       2.95164    0.78018   3.783 0.000355 ***
beds         4.43059    2.39921   1.847 0.069645 .
baths        0.52776    2.67639   0.197 0.844333
sqfl         2.70950    0.46065   5.882 1.84e-07 ***
view        -0.96760    1.55042  -0.624 0.534894
age         -0.53610    0.06648  -8.065 3.40e-11 ***
mfee        -0.13907    0.18297  -0.760 0.450133
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.928 on 61 degrees of freedom  (residual SD)
Multiple R-squared:  0.8187,    Adjusted R-squared:  0.7979
F-statistic: 39.35 on 7 and 61 DF,  p-value: < 2.2e-16  (ignore F-statistic)
```

df=$\nu = n - k$, k =number of estimated betas (here, $k = 8$, $\nu = 69 - 8 = 61$).
Least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_7$ withy $p = 7$.
Assume model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_7 x_{i7} + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$.
For the random variables, (to be shown later) $\mathrm{Var}\,(\hat{\beta}_j) = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ where this denotes the entry in row and column $j+1$ of $(\mathbf{X}^T\mathbf{X})^{-1}$ with row/column indices $0, \ldots, 7$.
$SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$, $\hat{\sigma} = [\sum_{i=1}^n r_i^2/(n-k)]^{1/2}$, $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_{i1}x_{i1} - \cdots - \hat{\beta}_{i7}x_{i7}$.
90% confidence interval for $\hat{\beta}_j$ is

$$\hat{\beta}_j \pm t_{\nu,0.95} SE(\hat{\beta}_j).$$

95%, 99% and 99.5% confidence intervals for $\hat{\beta}_j$ are

$$\hat{\beta}_j \pm t_{\nu,0.975} SE(\hat{\beta}_j).$$

$$\hat{\beta}_j \pm t_{\nu,0.995} SE(\hat{\beta}_j).$$

$$\hat{\beta}_j \pm t_{\nu,0.9995} SE(\hat{\beta}_j).$$

For $\nu = 61$, the critical values are $1.670219, 1.999624, 2.658857, 3.457294$ for the 0.95, 0.975, 0.995, 0.9995 quantiles of t_{61}.

Comment on usage of "standard error" versus "standard deviation".

SE=standard error is the SD of the sampling distribution of an estimator.

Refer to page of Appendix, Section 7, page 13, for explanations/definitions of other quantities that are in the output of lm() in R.

Important explanatory variables

For a single regression equation, look at the t-ratios or confidence intervals for $\hat{\beta}$'s to determine the most important explanatory variables.

From output, large absolute t-ratios and small P-values indicate the most important explanatory variables. This is same as explanatory variables x_j where the 95% confidence interval for β_j does not include 0 (interval is completely above 0 or it is completely below 0).

Explanation of P-value from hypothesis testing: Section 2.6 of coursepack.

2-sided P-value $= 2[1 - \texttt{pt}(\texttt{tratio}, \nu)]$ (draw picture).

P-value $< 0.10 \iff$ 90% confidence interval for β_j doesn't contain 0.
P-value $< 0.05 \iff$ 95% confidence interval for β_j doesn't contain 0.
P-value $< 0.01 \iff$ 99% confidence interval for β_j doesn't contain 0.
P-value $< 0.001 \iff$ 99.9% confidence interval for β_j doesn't contain 0.

Example, for flarea, $\hat{\beta}_1 = 2.9516$, $SE_1 = SE(\hat{\beta}_1) = 0.7802$, tratio=3.78, 99.9% confidence interval is

$$2.9516 \pm 3.457 \times 0.7802 = 2.9516 \pm 2.6974 = (0.254, 5.649),$$

so P-value $= 0.0004 < 0.001$.

The best prediction model will typically not include all of the explanatory variables in a data set. Generally different explanatory variables are correlated so that they provide related information, so not all of them are needed.

After residual diagnostics, quadratic or interaction terms might be added to the model. Also categorical variables can be introduced via binary dummy variables.

The number of β's in the final model can be more $p + 1$ where p is the number of explanatory variables. The number of β's is denoted as k.

Comparisons of prediction/regression equations with different subsets of explanatory variables

p	R^2	adj R^2	residSD $\hat{\sigma}$	signs of $\hat{\beta}$s
2	0.450	0.434	8.25	as expected
3	0.798	0.789	5.04	as expected
4	0.816	0.804	4.85	as expected
7	0.819	0.798	4.93	view unexpected, maybe mfee

R^2 and adj R^2: larger is better; $0 \le R^2 \le 1$ (see coursepack for interpretation and definition). adj $R^2 \le R^2 \le 1$.
$\hat{\sigma}$: smaller is better.

R^2 increases for nested models as more explanatory variables are added;
adj R^2 need not be increasing, if added variables have little additional explanatory power then it decreases;
$\hat{\sigma}$ decreases if added variables have good additional explanatory power, otherwise it can increase.

Later, another comparison is cross-validated root mean square prediction error.

More for R^2

$R^2 = 1 - \frac{SS(Res)}{SS(Total)}$,

$SS(Res) = S(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n r_i^2 = (n-k)\hat{\sigma}^2$, $k = p + 1$ if no quadratic terms etc.

$SS(Total) = \sum_{i=1}^n (y_i - \overline{y})^2 = (n-1)s_y^2$.

Boundary cases:
1. No explanatory variable is useful: $\hat{\beta}_1 = \cdots = \hat{\beta}_p = 0$ and $\hat{\beta}_0 = \overline{y}$: $R^2 = 0$.
2. The p explanatory variables provide a perfect fit: residuals $r_i = 0$ for all i and $R^2 = 1$.

Residual plots: (i) check for normality (normal quantile plot of residuals); (ii) check for homoscedasticity versus heteroscedasticity and possible structural deviations from model (plot of residuals versus predicted values, plots of residuals versus each explanatory variable).

Examples:
(a) Burnaby condominium data set;
(b) y versus x with scatter increasing with x;
(c) y versus x with curvilinear relation.