Standard output from linear regression model fit:

Section 1: Model description

```
1)   Number of obs =      51
2)   F(  2,    48) =   75.55
3)   Prob > F      =  0.0000
4)   R-squared     =  0.7589
5)   Adj R-squared =  0.7489
6)   Root MSE      =  302.65
```

1) Number of observations, always make sure you are using the correct data (this is one place to check)

2) Overall model F-test. First you will see the F-statistic for testing:

   H_o: all beta's in the model are equal to 0
   H_A: at least one beta in the model is not equal to 0

3) P-value corresponding to overall model F-test

4) R-squared :

   fraction of variability in the response that can be explained using the linear model compared to just using the mean.

   r^2= SS(MODEL) / SS(TOTAL)

   The R-squared value will increase with each additional explanatory variable, even if the additional variable is not scientifically meaningful.

5) Adjusted R-squared:

   R-square plus penalty for number of explanatory variables in the model

   Adj. R-square = $1-\left(1-r^2\right)*\left(\frac{n-1}{n-p-1}\right)$

   where n = number of observations, p = number of explanatory variables.

6) Root MSE:

   Recall, that in a simple/multiple linear regression, you are assuming that your residuals are independent and identically distributed from a normal distribution with mean 0 and variance σ^2. The MSE (see below) is the estimate of this common variance. Therefore, the ROOT MSE is the estimate of the standard deviation for the distribution of the residuals.

In a residual plot, apply the general rule that approximately 95% of normal data will fall within 2 standard deviations of the mean and draw 95% bounds on the graph to identify observations, which may be labelled as “extreme”.

Section 2: Analysis of Variance table

Source	SS	df	MS
Model	13840987.8	2	6920493.90
Residual	4396616.24	48	91596.1716
Total	18237604.0	50	364752.081

In the analysis of variance table, we decompose the variability in the response, Y, into the portion that we can explain with our linear prediction and error. Specifically, we have that

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

SS(Total) = SS(Model) + SS(Error)

Degrees of Freedom Total = DF(Total) = number of observations – 1 = n-1
Degrees of Freedom Model = DF(Model) = number of explanatory variables in the model = p
Degrees of Freedom Error = DF(Error) = DF Total – DF Model = n – 1 – p

The Mean Squares are simply the Sums of Squares divided by the degrees of freedom (MS(total) = Variance for Y).

The overall F-test from Section 1 is F = MS(Model) / MS(Error) which under the null hypothesis stated above will follow an F distribution with numerator degrees of freedom = DF(Model) and denominator degrees of freedom = DF(Error)

Section 3: Regression Coefficients

expadm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
los	213.7967	42.20769	5.065	0.000	128.9325	298.661
salary	.248994	.0217992	11.422	0.000	.2051638	.2928241
_cons	-2582.736	464.77	-5.557	0.000	-3517.219	-1648.254

This table will contain a row for every explanatory variable in the model plus a row for the intercept.

For each row you will be given:

a) the estimated regression coefficient ($\hat{\beta}$)

b) the estimated standard error of the regression coefficient (se($\hat{\beta}$))

c) the t-statistic for testing H_o: $\beta = 0$ vs. H_A: $\beta \neq 0$

$$t = \frac{\hat{\beta} - 0}{se(\hat{\beta})}$$

Under the null hypothesis, this statistic will follow the t distribution with degrees of freedom equal to DF(Error).

d) the corresponding p-value for t

e) the 95% confidence interval for the regression coefficient

Additional Regression Diagnostics:

We discussed using added variables plots (aka adjusted variables plots) and residual plots to look for curvature and outlying/influencial data values. We then discussed DFBETAS to assess the influence of individual data values in the analysis. There are other common regression diagnostics used to look for influence of data values.

First some quick definitions:

Leverage point: data value that is "extreme" to the distribution of an explanatory variable

Outlier: data value that is "extreme" to the distribution of the response variable or away from the general pattern of the response.

Influencial data point: high leverage and an outlier

Now some more regression diagnostics (See a reference for more details, these are additional diagnostics that you may have heard of or that are commonly used):

1) L-R plots: plots the leverage verses the squared residuals. Here you can identify points with high leverage and large residuals. However, these don't have to go hand-in-hand to warrant special attention. You may have data points with high leverage and small residual that influence the regression estimates and vice-versa.

2) Standardized and Studentized residuals: take your residuals and "standardize" them by dividing by some estimate of the error associated with the residuals. Then identify standardized or studentized residuals with values more extreme than +/- 2.

3) DFITS, Cook's Distance and Welsch Distance: attempts to summarize information from the L-R plot.

4) Variance inflation factors (VIFs): statistic calculated to measure presence of multi-collinearity. General rules of thumb exist for deleting explanatory variables based on this measure. Best ways to address multi-collinearity:

    - science: let the science inform you of possible strong associations between explanatory variables.
    - Use scatter plot matrix to look for associations across explanatory variables: note only identifies pairwise associations.
    - Use added variable plots in conjunction with scatter plots: how will this work?