Section 10.2 - More Detail About Simple Linear Regression

Statistics 104

Autumn 2004

Analysis of Variance for Regression

```
. regress cityfuel Weight

  Source |       SS       df       MS              Number of obs =      93
---------+------------------------------           F(  1,    91) =  374.31
   Model |  68.8245208     1  68.8245208           Prob > F      =  0.0000
Residual |  16.7322059    91  .183870394           R-squared     =  0.8044
---------+------------------------------           Adj R-squared =  0.8023
   Total |  85.5567267    92  .929964421           Root MSE      =   .4288

------------------------------------------------------------------------------
cityfuel |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |   .0014662   .0000758     19.35    0.000       .0013157    .0016168
   _cons |   .1936668   .2370884      0.82    0.416      -.2772802    .6646138
------------------------------------------------------------------------------
```

The Analysis of Variance (ANOVA) Table is an alternative approach to examining a regression model.

The idea behind it is based on

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

The variance in the data y is expressed by the deviations

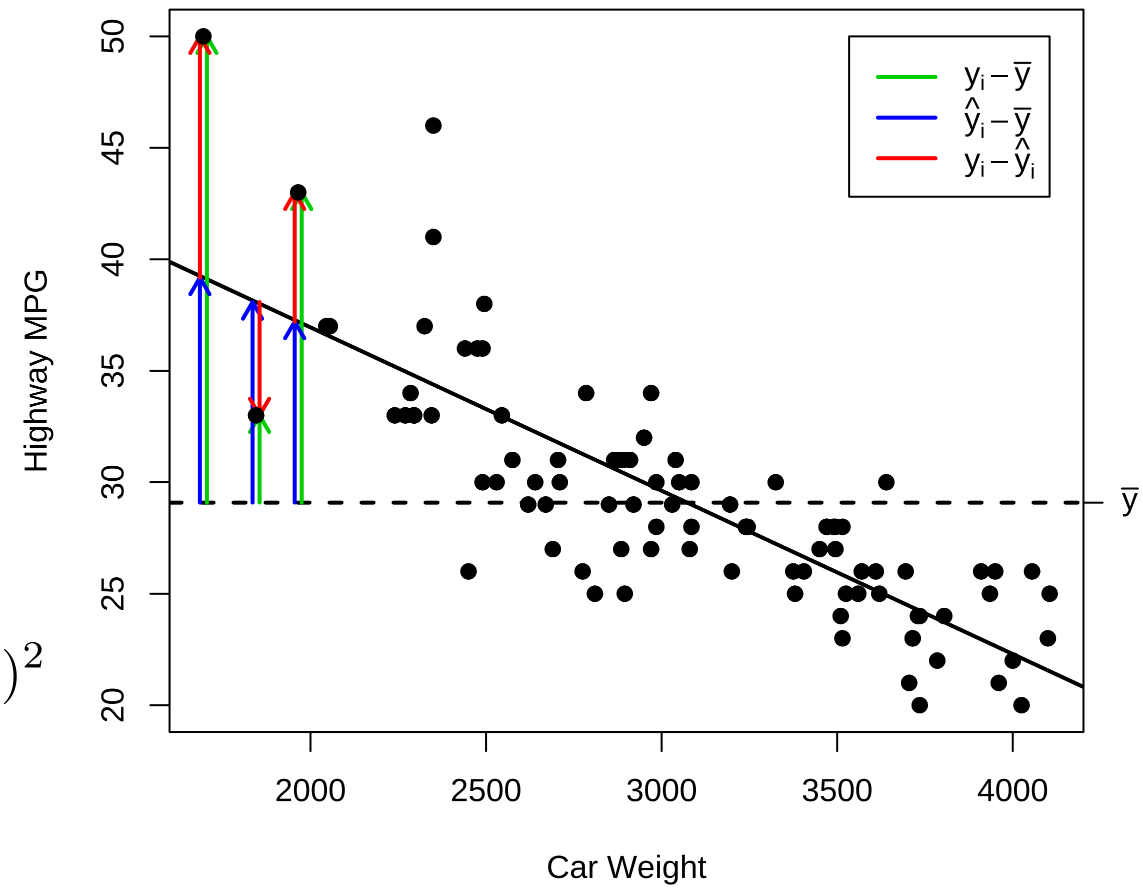
$$y_i - \bar{y}$$

This can be broken down as

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

It is possible to show that

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$



We can rewrite this formula as

$$SST = SSM + SSE$$

where

$$\begin{aligned} SST &= \sum (y_i - \bar{y})^2 && \text{(Total sums of squares)} \\ SSM &= \sum (y_i - \bar{y})^2 && \text{(Model SS)} \\ SSE &= \sum (y_i - \bar{y})^2 && \text{(Error or Residual SS)} \end{aligned}$$

If the slope $\beta_1 = 0$, the observations can be viewed as coming from a single population with mean μ_y with the variance described by the sample variance

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{SST}{n-1}$$

You can think of SST as the total error variability of the 0 slope model.

As we have seen before, $n-1$ is the degrees of freedom for the single population model and $n-1$ is the degrees of freedom for error in the simple linear regression model.

We can breakdown the degrees of freedom like we did the sums of squares

$$DFT = DFM + DFE$$

For simple linear regression

$$\begin{aligned} DFT &= n-1 \quad && \text{(Total degrees of freedom)} \\ DFM &= 1 && \text{(Model df)} \\ DFE &= n-2 && \text{(Error or Residual df)} \end{aligned}$$

Instead of looking at the sums of squares, we can also look at the mean squares (variability per degree of freedom).

$$MS = \frac{\text{sums of squares}}{\text{degrees of freedom}}$$

So

$$\begin{aligned} MSE &= \frac{SSE}{n-2} = s^2 \\ MSM &= \frac{SSM}{1} \end{aligned}$$

We can also fit correlation into this approach to the simple linear regression model. It is possible to show that

$$r^2 = \frac{SSM}{SST}$$

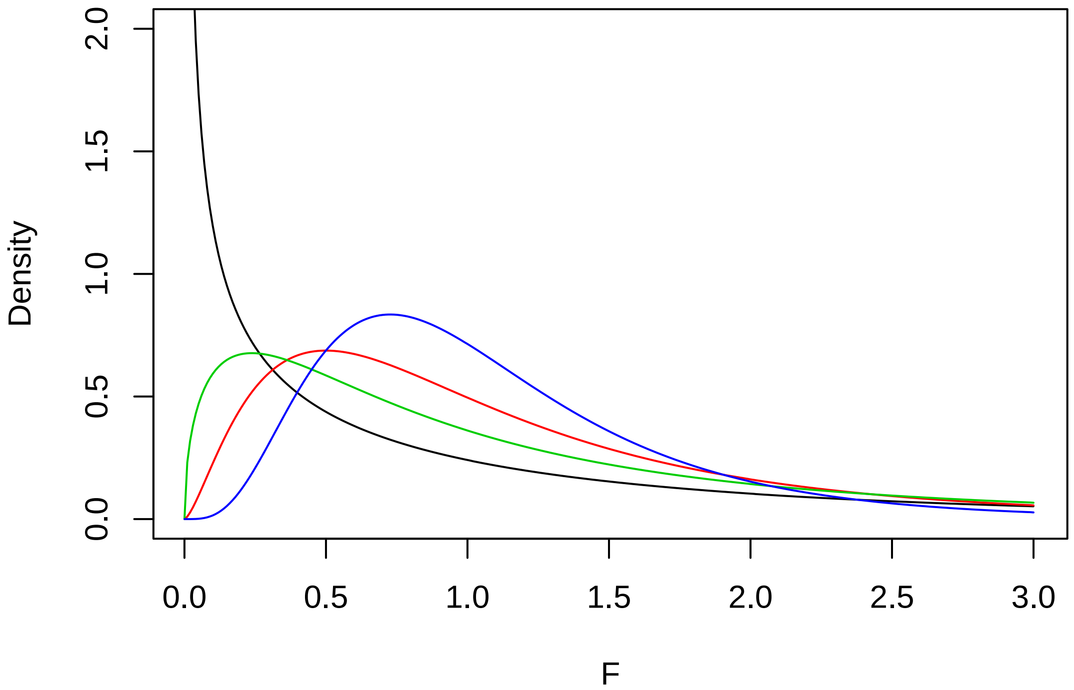
ANOVA F Test

Instead of using the t test to investigate the hypotheses $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$, we can look at the ratio

$$F = \frac{MSM}{MSE}$$

If the null hypothesis is false, MSE should be small and MSM should be large (leading to $F > 1$). If H_0 is true, $MSE \approx MSM (F \approx 1)$.

The sampling distribution of F is an F distrubution with 1 and $n - 2$ degrees of freedom ($F(1, n - 2)$)



The p-value for the F test is

$$p\text{--value} = P[F(1, n-2) \geq F_{obs}]$$

ANOVA Table

Source	DF	SS	MS	F
Model	1	$SSM = \sum(\hat{y}_i - \bar{y})^2$	$MSM = \frac{SSM}{DFM}$	$F = \frac{MSM}{MSE}$
Error	$n-2$	$SSE = \sum(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{DFE}$	
Total	$n-1$	$SST = \sum(y_i - \bar{y})^2$		

```
  Source |       SS       df       MS              Number of obs =      93
---------+------------------------------           F(  1,    91) =  374.31
   Model |  68.8245208     1  68.8245208           Prob > F      =  0.0000
Residual |  16.7322059    91  .183870394           R-squared     =  0.8044
---------+------------------------------           Adj R-squared =  0.8023
   Total |  85.5567267    92  .929964421           Root MSE      =   .4288

------------------------------------------------------------------------------
cityfuel |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |   .0014662   .0000758     19.35   0.000       .0013157    .0016168
   _cons |   .1936668   .2370884      0.82   0.416      -.2772802    .6646138
------------------------------------------------------------------------------
```

So we have two tests for examining

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0$$

In fact we really one have one, since it's possible to show that $t^2 = F$ and the p-values for the two tests are the same.

In the example $19.35^2 = 374.42$ (within rounding).

The F test is more useful for multiple regression models and in that situation it looks at more complicated hypotheses.

Inference for Correlation

There is a third approach to examining whether the data is better described by a line with slope 0.

If there is no correlation between x and y ($\rho = 0$), the population regression line will have slope $\beta_1 = 0$.

The usual test statistic for examining $H_0 : \rho = 0$ is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

This statistic has a $t(n-2)$ distribution.

The assumption behind this sampling distribution is that x and y are jointly normally distributed.

Getting p-values for this test statistic is similar to other t tests.

$$H_A : \rho < 0 \qquad p\text{−value} = P[T \leq t_{obs}]$$
$$H_A : \rho > 0 \qquad p\text{−value} = P[T \geq t_{obs}]$$
$$H_A : \rho \neq 0 \qquad p\text{−value} = 2 \times P[T \geq |t_{obs}|]$$

Should we be confused by have two different t tests in the linear regression setting? No, as this t test on correlation is exactly the same as the t on the slope.

It is possible to show

$$\frac{b_1}{SE_{b_1}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The following dataset I'll describe in more detail next class, but I want to show how the different approaches all tie in together.

```
  Source |       SS       df       MS              Number of obs =      26
---------+------------------------------           F(  1,    24) =    0.62
   Model |  4510.59756     1  4510.59756           Prob > F      =  0.4382
Residual |    174203.6    24  7258.48334           R-squared     =  0.0252
---------+------------------------------           Adj R-squared = -0.0154
   Total |  178714.198    25   7148.5679           Root MSE      =  85.197

------------------------------------------------------------------------------
    sales |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
----------+-------------------------------------------------------------------
Promotion |   7.332297   9.301349     0.79   0.438     -11.86474    26.52934
    _cons |   130.5569   53.00137     2.46   0.021      21.16744    239.9463
------------------------------------------------------------------------------

. pwcorr sales promotion, sig         (Pairwise Correlations)
             |    sales promotion
-------------+------------------
       sales |   1.0000
             |
   promotion |   0.1589   1.0000
             |   0.4382
```

Theoretical Aside

It is possible to link the parameters from a bivariate normal model to the population regression line model.

This is what motivates the relationship between testing whether a correlation is 0 and whether a slope is 0.

$$\begin{aligned}
\beta_1 &= \rho\frac{\sigma_y}{\sigma_x} \\
\beta_0 &= \mu_y - \beta_1\mu_x \\
\sigma_\epsilon &= \sigma_y\sqrt{1-\rho^2}
\end{aligned}$$