Applied Regression Modeling: A Business Approach
Chapter 3: Multiple Linear Regression
Sections 3.3.3–3.3.5

by Iain Pardoe

Evaluating fit numerically

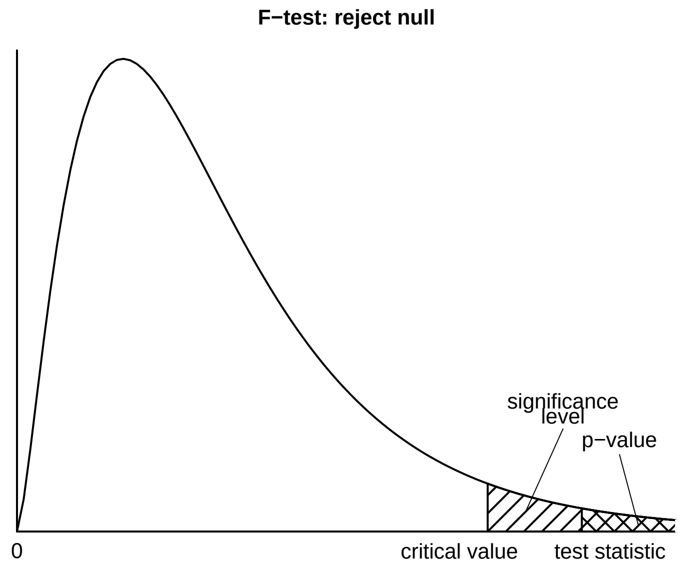
Three methods:

- How close are the actual observed Y-values to the model-based fitted values, $\hat{Y}$?
  - Calculate *regression standard error*, s (3.3.1).
- How much of the variability in Y have we been able to explain with our model?
  - Calculate *coefficient of determination*, R^2 (3.3.2).
- How strong is the evidence of our modeled relationship between Y and $(X_1, X_2, \ldots)$?
  - Estimate/test *regression parameters*, $b_1, b_2, \ldots$
  - Globally (3.3.3), in subsets (3.3.4), individually (3.3.5).

Global usefulness test

- Model: $\mathrm{E}(Y) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$.
  Could all k population regression parameters be 0?
- NH: $b_1 = b_2 = \cdots = b_k = 0$
  AH: at least one of b_1, b_2, ..., b_k is not equal to 0.
- Global F-stat $= \frac{(\text{TSS}-\text{SSE})/k}{\text{SSE}/(n-k-1)} = \frac{\text{R}^2/k}{(1-\text{R}^2)/(n-k-1)}$.
- Significance level $= 5\%$.
- Critical value is 95^{th} percentile of the F-distribution with k numerator df and $n-k-1$ denominator df.
- The p-value is the area to the right of the global F-statistic for the F-distribution with k numerator df and $n-k-1$ denominator df.
- If the global F-statistic falls in the rejection region, or the p-value is less than the significance level, then we reject NH in favor of AH.

Density curve for an F-distribution



Global usefulness test for HOMES3 data

ANOVA [a]

Model		Sum of Squares	df	Mean Square	Global F-stat	Pr(>F)
1	Regression	630.259	2	315.130	51.434	0.005 [b]
	Residual	18.381	3	6.127		
	Total	648.640	5			

[a] Response variable: Y.
[b] Predictors: (Intercept), X1, X2.

- $$\text{Global F-stat} = \frac{(\text{TSS}-\text{SSE})/k}{\text{SSE}/(n-k-1)} = \frac{(648.640-18.381)/2}{18.381/(6-2-1)}$$
  $$= \frac{\text{R}^2/k}{(1-\text{R}^2)/(n-k-1)} = \frac{0.97166/2}{(1-0.97166)/(6-2-1)}$$
  $$= 51.4.$$
- Critical value, `FINV(0.05,2,3)`, is 9.55.
- p-value, `FDIST(51.4,2,3)`, is 0.005.
- Reject NH in favor of AH; at least one of the predictors, (X_1, X_2), is linearly related to Y.

Global usefulness test for SHIPDEPT data

ANOVA [a]

Model		Sum of Squares	df	Mean Square	Global F-stat	Pr(>F)
1	Regression	5646.052	4	1411.513	?	? [b]
	Residual	1242.898	15	82.860		
	Total	6888.950	19			

[a] Response variable: Y.

[b] Predictors: (Intercept), X1, X2, X3, X4.

- $$\begin{aligned}\text{Global F-stat} &= \frac{(\text{TSS}-\text{SSE})/k}{\text{SSE}/(n-k-1)} \\ &= \frac{\text{R}^2/k}{(1-\text{R}^2)/(n-k-1)} \\ &= ?\end{aligned}$$
- Critical value, `FINV(0.05,4,15)`, is 3.06.
- p-value, `FDIST(?,4,15)`, is ?
- Reject NH in favor of AH; at least one of the predictors, (X_1, X_2, X_3, X_4), is linearly related to Y.

3.3.4 Nested model test

Do some predictors overfit the data?

- Suppose a global usefulness test suggests at least one of $(X_1, X_2, \ldots, X_k)$ is linearly related to Y.
- Can a *reduced* model with less than k predictor variables be better than a *complete* k-predictor model?
  - If a subset of the X's provides no useful information about Y *beyond* the information provided by the other X's.
- Complete k-predictor model: SSE_C.
- Reduced r-predictor model: SSE_R.
- Which is larger? (recall geometric argument)
- Which model is favored if it is a lot larger?
- Which model is favored if it is just a little larger?

Nested model test

- Reduced model: $\mathrm{E}(Y) = b_0 + b_1X_1 + \cdots + b_rX_r$.
- Complete model: $\mathrm{E}(Y) = b_0 + b_1X_1 + \cdots + b_rX_r + b_{r+1}X_{r+1} + \cdots + b_kX_k$.
- NH: $b_{r+1} = \cdots = b_k = 0$
  AH: at least one of $b_{r+1}, \ldots, b_k$ is not equal to 0.
- Nested F-stat $= \frac{(\mathrm{SSE_R} - \mathrm{SSE_C})/(k-r)}{\mathrm{SSE_C}/(n-k-1)}$.
- Significance level $= 5\%$.
- Critical value is 95^{th} percentile of the F-distribution with $k-r$ numerator df and $n-k-1$ denominator df.
- The p-value is the area to the right of the nested F-statistic for the F-distribution with $k-r$ numerator df and $n-k-1$ denominator df.
- If the nested F-statistic falls in the rejection region, or the p-value is less than the significance level, then we reject NH in favor of AH.

Nested F-statistic for SHIPDEPT data

ANOVA [a]

Model		Sum of Squares	df	Mean Square	Global F-stat	Pr(>F)
C	Regression	5646.052	4	1411.513	17.035	0.000 [b]
	Residual	1242.898	15	82.860		
	Total	6888.950	19			

[a] Response variable: Y.
[b] Predictors: (Intercept), X1, X2, X3, X4.

Model		Sum of Squares	df	Mean Square	Global F-stat	Pr(>F)
R	Regression	5567.889	2	2783.945	35.825	0.000 [b]
	Residual	1321.061	17	77.709		
	Total	6888.950	19			

[a] Response variable: Y.
[b] Predictors: (Intercept), X1, X3.

- $$\begin{aligned} \text{Nested F-stat} &= \frac{(\mathrm{SSE_R} - \mathrm{SSE_C})/(k-r)}{\mathrm{SSE_C}/(n-k-1)} \\ &= \frac{(1321.061 - 1242.898)/(4-2)}{1242.898/(20-4-1)} \\ &= 0.472. \end{aligned}$$

Nested model test results

- Reduced model: $\mathrm{E}(Y) = b_0 + b_1X_1 + b_3X_3$.
- Complete model: $\mathrm{E}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$.
- NH: $b_2 = b_4 = 0$
  AH: at least one of b_2 or b_4 is not equal to 0.
- Nested F-stat $= 0.472$.
- Significance level $= 5\%$.
- Critical value, `FINV(0.05,2,15)`, is 3.68.
- p-value, `FDIST(0.472,2,15)`, is 0.633.
- Cannot reject NH in favor of AH.
- Neither X_2 nor X_4 appears to provide useful information about Y beyond the information provided by X_1 and X_3.
- Reduced model is favored.

Compare reduced and complete models

Model Summary

Model	R Squared	Adjusted R Squared	Regression Std. Error	Change Statistics F-stat	df1	df2	Pr(>F)
R	0.808 [a]	0.786	8.815				
C	0.820 [b]	0.771	9.103	0.472	2	15	0.633

[a] Predictors: (Intercept), X1, X3.
[b] Predictors: (Intercept), X1, X2, X3, X4.

- There is a suggestion that adding $X_2 =$ truck proportion and $X_4 =$ week to the model causes overfitting. Why?
  - Adjusted R^2 is higher for the reduced model.
  - The regression standard error, s, is lower for the reduced model.
  - The nested F-stat is not significant (high p-value), so the reduced model is favored.

Individual regression parameter test

- Which predictors to test in a nested model test?
- One possible approach is to consider the regression parameters individually.
- What do the estimated sample estimates, $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$, tell us about likely values for the population parameters, $b_1, b_2, \ldots, b_k$?
- An individual t-test for b_p considers whether there is evidence that X_p provides useful information about Y beyond the information provided by the other $k-1$ predictors. In other words:
  - should we retain X_p in the model with the other $k-1$ predictors (evidence suggests $b_p \neq 0$);
  - or, should we consider removing X_p from the model and retain only the other $k-1$ predictors (evidence cannot rule out $b_p = 0$)?

Hypothesis test for b_p

- Recall slope t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- Here, t-statistic for $b_p = \frac{\hat{b}_p - b_p}{s_{\hat{b}_p}} \sim t_{n-k-1}$.
- Example: NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{6.074 - 0}{2.662} = 2.28$.
- Significance level $= 5\%$.
- Critical value, `TINV(0.05,15)`, is 2.13.
- p-value, `TDIST(2.28,15,2)`, is 0.038.
- Since t-statistic $(2.28) >$ critical value (2.13) and p-value $<$ signif. level, reject NH in favor of AH.
- Sample data favor $b_1 \neq 0$ (at a 5% signif. level).
- There appears to be a linear relationship between Y and X_1, once X_2, X_3, and X_4 have been accounted for (or holding X_2, X_3, and X_4 constant).

Individual t-test computer output

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	95.415	30.036	3.177	0.006
	X1	6.074	2.662	2.281	0.038
	X2	0.084	0.089	0.951	0.357
	X3	−1.746	0.760	−2.297	0.036
	X4	−0.124	0.380	−0.328	0.748

[a] Response variable: Y.

- Last two cols: individual t-stats and two tail p-values.
- Low p-values indicate potentially useful predictors that should be retained (i.e., X_1 and X_3 here).
- High p-values indicate possible candidates for removal from the model (i.e., X_2 and X_4 here).
- However, high p-value for X_2 means we can remove X_2, but only if we retain X_1, X_3, and X_4.
- Similarly, high p-value for X_4 means we can remove X_4, but only if we retain X_1, X_2, and X_3.

Individual t-tests and nested F-tests

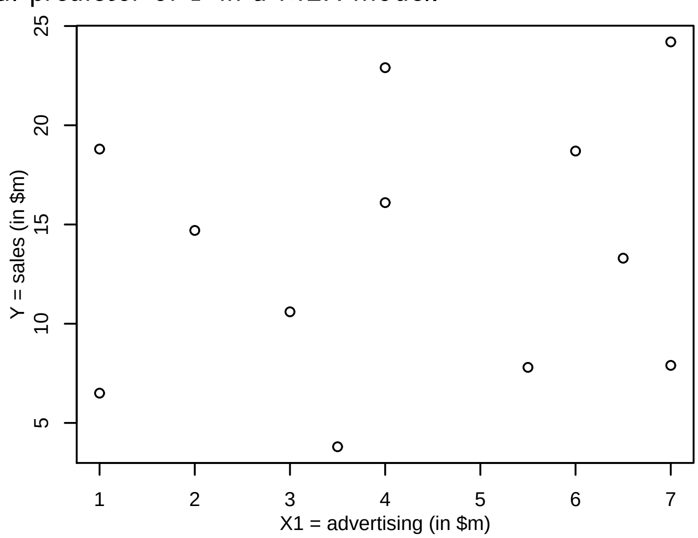
- Can do individual regression parameter t-tests to:
  - remove just one redundant predictor at a time;
  - or to identify which predictors to investigate with a nested model F-test.
- Need to do a nested model F-test to remove more than one predictor at a time.
- Using nested model F-tests allows us to use fewer hypothesis tests overall to help identify redundant predictors (so that the remaining predictors appear to explain Y adequately).
  - This also lessens the chance of making any hypothesis test errors.

Regression parameter confidence intervals

- Calculate a 95% confidence interval for b_1.
- 97.5^{th} percentile of t_{15} is `TINV(0.05,15)`$=2.131$.
- $\hat{b}_1 \pm 97.5^{\text{th}}$ percentile $(s_{\hat{b}_1}) = 6.074 \pm 2.131 \times 2.662 = 6.074 \pm 5.673 = (0.40,\, 11.75)$.
- Loosely speaking: based on this dataset, we are 95% confident that the the population regression parameter, b_1, is between 0.40 and 11.75 in the model $\mathrm{E}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$.
- More precisely: if we were to take a large number of random samples of size 20 from our population of shipping numbers and calculate a 95% confidence interval for b_1 in each, then 95% of those confidence intervals would contain the true (unknown) population regression parameter.
- What happens to this interval in the model $\mathrm{E}(Y) = b_0 + b_1X_1 + b_3X_3$?

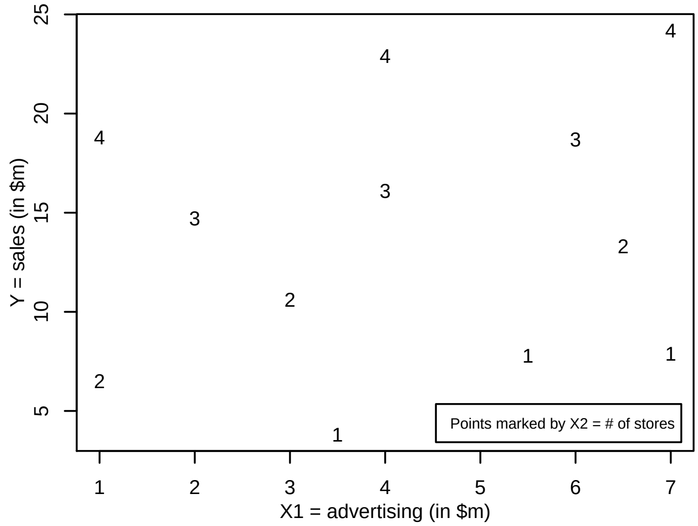
Correlation revisited: Y and X_1 uncorrelated

X_1 can still be a useful predictor of Y in a MLR model.



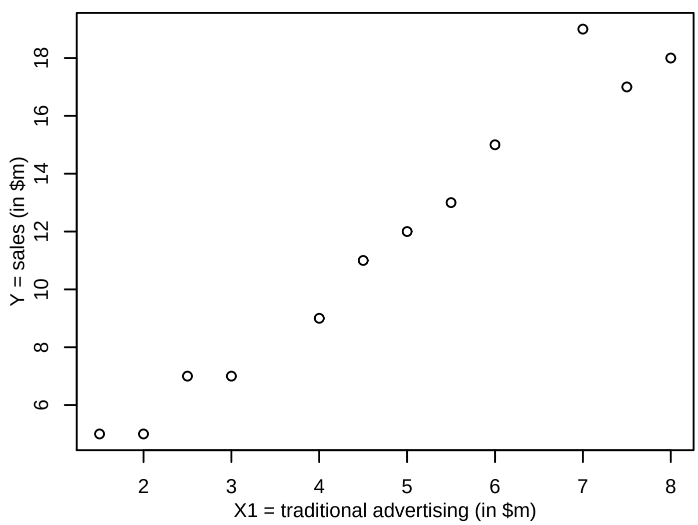
But Y associated with (X_1, X_2) together

Linear association between Y and X_1 for fixed X_2.



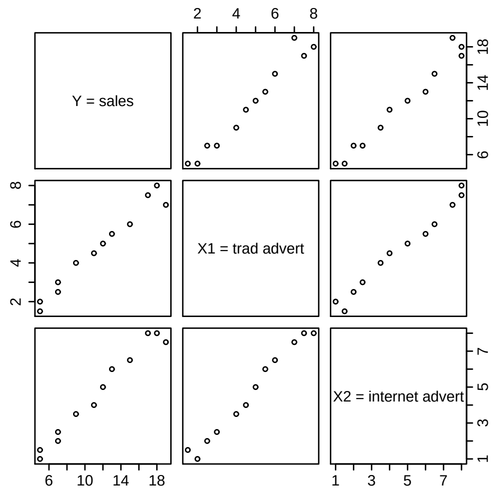
Correlation revisited: Y and X_1 correlated

X_1 may be a poor predictor of Y in a MLR model.



But X_1 and X_2 even more highly correlated

Unstable estimates when both X_1 and X_2 in model.



Predictor selection

- Global usefulness test to determine whether *any* of the potential predictors in a dataset are useful.
- Nested model F-tests and individual parameter t-tests to identify the most important predictors.
- Employ tests judiciously to avoid conducting too many tests and reduce chance of making mistakes.
- If possible, identification of the important predictors should also be guided by practical considerations and background knowledge about the application.
- When k is very large, computer intensive methods can help get things started:
  - *Forward selection:* predictors added sequentially to an initial zero-predictor model;
  - *Backward elimination:* predictors excluded sequentially from the full k-predictor model;
  - Combined *stepwise* method: can proceed forwards or backwards at each stage;
  - Other *machine learning/data mining* methods.

