Business Analytics

Linear Model: Multiple Linear Regression

Dr. Jiun-Yu Yu
BA, NTU

25 Sep 2018

Outlines

- Multiple linear regression model
- Least squares estimators, goodness of fit
- Assumptions about residual standard errors
- Model building and statistical inference
  - Global usefulness test
  - Nested model test / Extra-sums-of-squares F-test
  - Individual test
- Model diagnosis
- Prediction

Multiple Linear Regression

- Y is a quantitative response variable (a.k.a. dependent, outcome, or output variable).
- $(X_1, X_2, \ldots)$ are quantitative explanatory variables (a.k.a. predictor, independent/input variables, or covariates).
- Important to identify variables and define them carefully, e.g.:
  - Y is final exam score, out of 100;
  - X_1 is time spent partying during last week of term, in hours;
  - X_2 is average time spent studying during term, in hours per week
- How much do we expect Y to change by when we change the values of X_1 and/or X_2?
- What do we expect the value of Y to be when $X_1 = 7.5$ and $X_2 = 1.3$?

Multiple Linear Regression Model

- Model: $E(Y|X_1, X_2, \ldots) = b_0 + b_1X_1 + b_2X_2 + \ldots$.
- Interpretation:
  - b_0: expected Y-value when $X_1 = X_2 = \ldots = 0$;
  - b_1: "slope in the X_1 -direction" (i.e., when $X_2, X_3, \ldots$ are held constant);
  - b_2: "slope in the X_2-direction" (i.e., when $X_1, X_3, \ldots$ are held constant).
- Sample: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \ldots\ldots$.
  - How can we estimate $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots\ldots$?

Estimating the Model

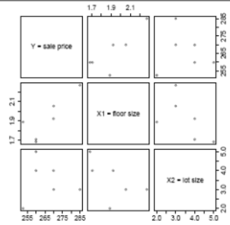
- Model: $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + \varepsilon_i$, $i = 1, \ldots, n$.
- Estimate: $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \ldots + \hat{b}_k X_{ki}$.
- Obtain $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots \hat{b}_k$ by finding best fit "hyperplane" (using least squares).
- Mathematically, *minimize* sum of squared errors (*SSE*):

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i} - \ldots - \hat{b}_k X_{ki})^2$$

Example: HOMES3 Data

- X_1 : floor-size, X_2 : lot size

Y	252.5	259.9	259.9	269.9	270.0	285.0
X_1	1.888	1.683	1.708	1.922	2.053	2.269
X_2	2	5	4	4	3	3



Scatterplot Matrix

- A matrix of scatterplots showing all bivariate relationships in a multivariate dataset (e.g., previous slide).
- However, patterns cannot tell us whether a multiple linear regression model can provide a useful mathematical approximation to these bivariate relationships.
- Primarily useful for identifying any strange patterns or odd-looking values that might warrant further investigation before we start modeling.
- Home price–floor size example:
  - No odd values to worry about.

Multiple Linear Regression Model

- Propose this multiple linear regression model:
  $Y = E(Y) + \varepsilon$
  $= b_0 + b_1X_1 + b_2X_2 + \varepsilon.$
- Random errors, ε, represent variation in Y due to factors other than X_1 and X_2 that we haven't measured, e.g., numbers of bedrooms/bathrooms, property age, garage size, or nearby schools.
- Use least squares to estimate the deterministic part of the model, $E(Y)$, as $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2$
  - i.e., use statistical software to find the values of $\hat{b}_0, \hat{b}_1, \hat{b}_2$ that minimize

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i})^2$$

R Output: HOMES3 Data

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.357     14.786   8.275  0.00370 **
X1            61.976      6.113  10.139  0.00204 **
X2             7.091      1.281   5.535  0.01162 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Fitted model: $\hat{Y} = 122.36 + 61.98X_1 + 7.09X_2$.
- Expect Y to change by $\hat{b}_1$ when X_1 increases by one and X_2 stays constant, i.e., expect sale price to increase \$6200 when floor size increases 100 sq. feet and lot size stays constant.
- Expect Y to change by $\hat{b}_2$ when X_2 increases by one and X_1 stays constant, i.e., expect sale price to increase \$7090 when lot size increases one category and floor size stays constant.

Beta Coefficients

- These explanatory variables are measured in different units, thus, to see which one has larger impact on Y, it is not sensible to compare their regression coefficients.
- Instead, we can use standardized regression model, in which all the explanatory and response variables are standardized (mean = 0, standard deviation = 1).
- It can be shown that beta coefficients, estimated from above, are:

$$\hat{b}_j^* = \frac{S_{X_j}}{S_Y} \hat{b}_j, \quad j = 1, \ldots, k$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}, \qquad S_{X_j} = \sqrt{\frac{\sum_{i=1}^{n} (X_{j,i} - \bar{X}_j)^2}{n-1}}$$

= sample standard deviation of Y = sample standard deviation of X_j

Beta Coefficients

$$Y_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \cdots + \hat{b}_k X_{ki} + \hat{\varepsilon}_i$$

$$E(Y) = \bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{X}_1 + \cdots + \hat{b}_k \bar{X}_k$$

$$\therefore Y_i - \bar{Y} = \hat{b}_1 (X_{1i} - \bar{X}_1) + \cdots + \hat{b}_k (X_{ki} - \bar{X}_k) + \hat{\varepsilon}_i$$

$$\therefore \frac{Y_i - \bar{Y}}{S_Y} = \hat{b}_1 \frac{S_{X_1}}{S_Y} \left(\frac{X_{1i} - \bar{X}_1}{S_{X_1}} \right) + \cdots + \hat{b}_k \frac{S_{X_k}}{S_Y} \left(\frac{X_{ki} - \bar{X}_k}{S_{X_k}} \right) + \frac{\hat{\varepsilon}_i}{S_Y}$$

$$\therefore Y_i^* = \hat{b}_1^* X_{1i}^* + \cdots + \hat{b}_k^* X_{ki}^* + \hat{\varepsilon}_i^*$$

```
> beta.X1 <- hm3.lm$coef["X1"]*sd(X1)/sd(Y)
> beta.X2 <- hm3.lm$coef["X2"]*sd(X2)/sd(Y)
> beta.X1
      X1
1.196173
> beta.X2
     X2
0.65297
```


Calculating R^2

- Without model, estimate Y with sample mean $\bar{Y}$.
- With model, estimate Y using fitted $\hat{Y}$-value.
- How much do we reduce our error when we do this?
- Total error without model (variation in Y about $\bar{Y}$):
  $SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- Remaining error with model (unexplained variation):
  $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- Proportional reduction in error: $R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$
- Home price–floor size example: $R^2 = 0.972$. Thus, 97.2% of the variation in sale price (about its mean) can be explained by a multiple linear regression relationship between sale price and (floor size, lot size).

Disadvantage of R^2 for Model Building

- Model building: what is the best way to model the relationship between Y and $(X_1, X_2, \ldots, X_k)$?
  - e.g., should we use all k predictors, or just a subset?
- Consider a sequence of *nested models*, with each model in the sequence adding explanatory variable to the previous model.
- Which model would R^2 say is the "best" model? The final model with k explanatory variables.
- Geometrical argument: start with a regression line on a 2D-scatterplot, then add a second explanatory variable to make the line a plane in a 3D-scatterplot.
- In other words, R^2 always increases (or stays the same) as you add explanatory variables to a model.

Adjusted R^2

- R^2 has a clear interpretation since it represents the proportion of variation in Y (about its mean) explained by a multiple linear regression relationship between Y and $(X_1, X_2, \ldots)$.
- But, R^2 is not appropriate for finding a model that captures the major, important population relationships without overfitting every slight twist and turn in the sample relationships.
- We need an alternate criterion, which penalizes models that contain too many unimportant predictor variables:

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right)$$

- In practice, we can obtain the value for adjusted R^2 directly from statistical software.

Using Adjusted R^2

```
lm(formula = Y ~ X1)
Residual standard error: 7.178 on 4 degrees of freedom
Multiple R-squared: 0.6823,     Adjusted R-squared: 0.6029
F-statistic: 8.591 on 1 and 4 DF,  p-value: 0.04277

lm(formula = Y ~ X1 + X2)
Residual standard error: 2.475 on 3 degrees of freedom
Multiple R-squared: 0.9717,     Adjusted R-squared: 0.9528
F-statistic: 51.43 on 2 and 3 DF,  p-value: 0.00477
```

- Since adjusted R^2 is 0.603 for the single-predictor model, but 0.953 for the two-predictor model, the two-predictor model is better than the single-predictor model (according to this criterion).
- In other words, there is no indication that adding X_2 = lot size to the model causes overfitting.
- What happens to R^2 and $\hat{\sigma}$?

Residual Standard Error, $\hat{\sigma}$

```
Call:
lm(formula = Y ~ X1 + X2)
Residual standard error: 2.475 on 3 degrees of freedom
Multiple R-squared: 0.9717,     Adjusted R-squared: 0.9528
F-statistic: 51.43 on 2 and 3 DF,  p-value: 0.00477
```

- *Residual/Regression standard error*, $\hat{\sigma}$, estimates the standard deviation of the multiple linear regression random errors:

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-k-1}}$$

- Unit of measurement for $\hat{\sigma}$ is the same as unit of measurement for Y.

Example 2: SHIPDEPT Data

Y (labor hours)	X_1 (weight shipped)	X_2 (truck proportion)	X_3 (average weight)	X_4 (week)
100	5.1	90	20	1
85	3.8	99	22	2
...	...	...	...	...
85	4.8	58	25	20

- Y = weekly labor hours
- X_1 = total weight shipped in thousands of pounds
- X_2 = proportion shipped by truck
- X_3 = average shipment weight in pounds
- X_4 = week
- Compare two models:
  - $E(Y) = b_0 + b_1 X_1 + b_3 X_3$;
  - $E(Y) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$.

Adjusted R^2 for SHIPDEPT data

```
lm(formula = Y ~ X1 + X3, data = SHIPDEPT)
Residual standard error: 8.815 on 17 degrees of freedom
Multiple R-squared: 0.8082,     Adjusted R-squared: 0.7857
F-statistic: 35.83 on 2 and 17 DF,  p-value: 8.008e-07

lm(formula = Y ~ X1 + X2 + X3 + X4, data = SHIPDEPT)
Residual standard error: 9.103 on 15 degrees of freedom
Multiple R-squared: 0.8196,     Adjusted R-squared: 0.7715
F-statistic: 17.03 on 4 and 15 DF,  p-value: 1.889e-05
```

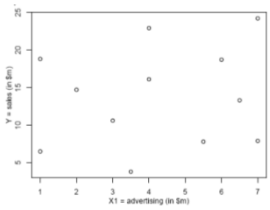
- Since adjusted R^2 is 0.786 for the two-predictor model, but 0.772 for the four-predictor model, the two-predictor model is better than the four-predictor model (according to this criterion).
- In other words, there is a suggestion that adding X_2 = truck proportion and X_4 = week to the model causes overfitting.
- What happens to R^2 and $\hat{\sigma}$?

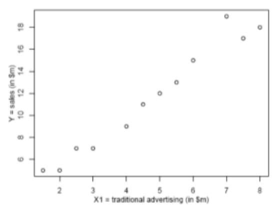
Multiple Correlation Coefficient

```
sqrt(summary(fit3)$r.squared)
0.9857275
```

- The multiple correlation coefficient, <u>multiple *R*</u>, measures the strength and direction of linear association between the observed *Y*-values and the fitted -values from the model.
- Multiple linear regression: multiple $R = +\sqrt{R^2}$.
  - e.g., $0.986 = \sqrt{0.972}$ for the home price–floor size example above.
- Beware: intuition about correlation can be seriously misleading when it comes to multiple linear regression.

Correlation between Y and X_1

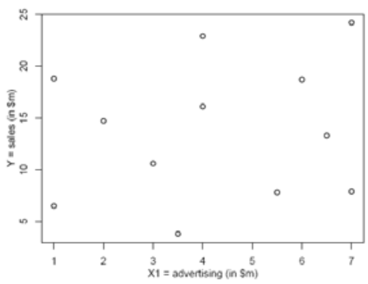




- X_1 on the left might still be a useful predictor of Y in a Multiple LR model.
- X_1 on the right might still be a poor predictor of Y in a Multiple LR model.

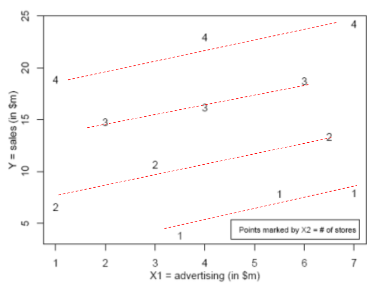
Correlation: Y and X_1 Uncorrelated

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



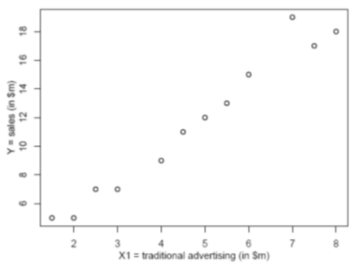
But Y associated with (X_1, X_2) together

- Linear association between Y and X_1 for fixed X_2



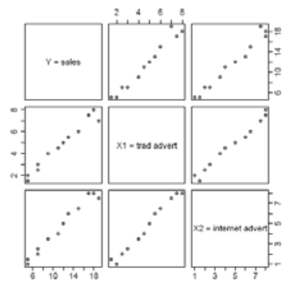
Correlation: Y and X_1 Correlated

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



But X_1 and X_2 Highly Correlated

- Unstable estimates when both X_1 and X_2 in model.



Statistical Inference for MLR (I)

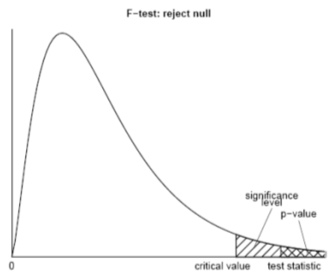
- Model building and statistical inference
  - Model evaluation
  - Variable transformation and interactions
  - Influential points and outliers
  - Variable selection
- Model diagnosis

- Model evaluation: How strong is the evidence of our modeled relationship between Y and $(X_1, X_2, \ldots)$?
  - Global usefulness test
  - Nested model test
  - Individual test

Global Usefulness Test

- Model: $E(Y) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$.
  Could all k population regression parameters be 0?
- H_0: $b_1 = b_2 = \cdots = b_k = 0$
  H_A: at least one of $b_1, b_2, \ldots, b_k$ is not equal to 0.
- Global F-stat = $\dfrac{(SST - SSE)/k}{SSE/(n-k-1)} = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$
- 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 H_0 in favor of H_A.

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 is 9.55. `(qf(0.95, 2, 3))`
- p-value is 0.005. `(1-pf(51.434, 2, 3))`
- Reject H_0 in favor of H_A; 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.

$$\text{Global F-stat} = \frac{(\text{TSS}-\text{SSE})/k}{\text{SSE}/(n-k-1)}$$
$$= \frac{\text{R}^2/k}{(1-\text{R}^2)/(n-k-1)}$$
$$= ?$$

- Critical value is 3.06. `(qf(0.95, 4, 15))`
- p-value is ?
- Reject H_0 in favor of H_A; at least one of the predictors, (X_1, X_2, X_3, X_4), is linearly related to Y.

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 (Full) k-predictor model: SSE_C.
- Reduced r-predictor model: SSE_R.
- Which is larger?
- 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: $E(Y) = b_0+b_1X_1+\cdots+b_rX_r$.
- Complete (Full) model: $E(Y) = b_0+b_1X_1+\cdots+b_rX_r+b_{r+1}X_{r+1}+\cdots+b_kX_k$.
- $H_0 : b_{r+1} =\cdots= b_k = 0$
  H_A : at least one of $b_{r+1}, \ldots, b_k$ is not equal to 0.
- Nested F-stat = $\dfrac{(SSE_R - SSE_C)/(k-r)}{SSE_C/(n-k-1)}$
- 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 H_0 in favor of H_A.

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.

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

Nested Model Test Results

- Reduced model: $E(Y) = b_0 + b_1X_1 + b_3X_3$.
- Complete model: $E(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$.
- $H_0 : b_2 = b_4 = 0$
  H_A : at least one of b_2 or b_4 is not equal to 0.
- Nested F-stat = 0.472.
- Significance level = 5%.
- Critical value is 3.68. `(qf(0.95,2,15))`
- p-value is 0.633. `(1-pf(0.472,2,15))`
- Cannot reject H_0 in favor of H_A.
- 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	Residual 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 residual standard error, $\hat{\sigma}$, is lower for the reduced model.
  - The nested F-stat is not significant (high p-value), so the reduced model is favored.
- `anova(Model.R, Model.C)`

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 in SLR, slope t-statistic = $\dfrac{\hat{b}_1 - b_1}{S_{\hat{b}_1}} \sim t_{n-2}$.
- Here in MLR, t-statistic for $b_p = \dfrac{\hat{b}_p - b_p}{S_{\hat{b}_p}} \sim t_{n-k-1}$.
- Example: H_0: $b_1 = 0$ versus H_A: $b_1 \neq 0$, and
- t-statistic = $\dfrac{\hat{b}_1 - b_1}{S_{\hat{b}_1}} = \dfrac{6.074 - 0}{2.662} = 2.28.$ (SHIPDEPT data)
- With signif. level = 5%, critical value = 2.13, p-value = 0.038.
- Since t-statistic (2.28) > critical value (2.13) and p-value < signif. level, reject H_0 in favor of H_A.
- 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 R Output: SHIPDEPT

```
Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 95.41495   30.03577   3.177  0.00626 **
X1           6.07391    2.66245   2.281  0.03755 *
X2           0.08435    0.08870   0.951  0.35673
X3          -1.74600    0.76018  -2.297  0.03645 *
X4          -0.12450    0.37993  -0.328  0.74768
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 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 2.131.
- $\hat{b}_1 \pm 97.5^{th}$ 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 population regression parameter, b_1, is between 0.40 and 11.75 in the model $E(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$.
- 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.

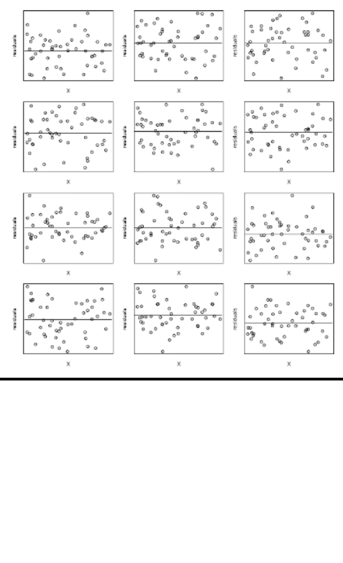
Regression Model Assumptions

- Four assumptions about random errors, $\varepsilon = Y - E(Y) = Y - b_0 - b_1X_1 - \ldots - b_kX_k$:
  - Probability distribution of ε at each set of values $(X_1, X_2, \ldots, X_k)$ has a mean of **zero**;
  - Probability distribution of ε at each set of values $(X_1, X_2, \ldots, X_k)$ has **constant variance**;
  - Value of ε for one observation is **independent** of the value of ε for any other observation;
  - Probability distribution of ε at each set of values $(X_1, X_2, \ldots, X_k)$ is **normal**.

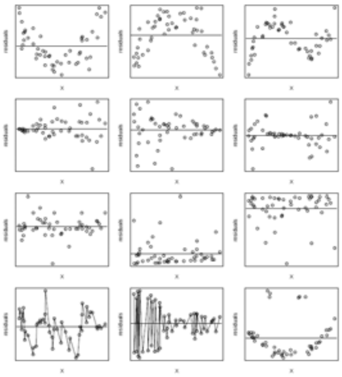
Checking Model Assumptions

- Calculate residuals, $\hat{\varepsilon} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1X_1 - \cdots - \hat{b}_kX_k$
- Draw a residual plot with $\hat{\varepsilon}$ along the vertical axis and a function of $(X_1, X_2, \ldots, X_k)$ along the horizontal axis (e.g., $\hat{Y}$ or one of the X's).
  - Assess zero mean assumption – do the residuals average out to zero as we move across the plot from left to right?
  - Assess constant variance assumption – is the (vertical) variation of the residuals similar as we move across the plot from left to right?
  - Assess independence assumption – do residuals look "random" with no systematic patterns?
- Draw a histogram and QQ-plot of the residuals.
  - Assess normality assumption – does histogram look approximately bell-shaped and symmetric and do QQ-plot points lie close to line?

Residual Plots which Pass

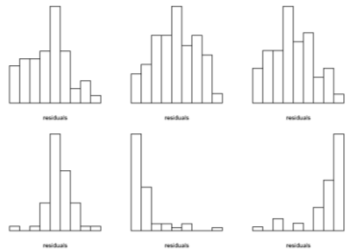


Residual Plots which Fail



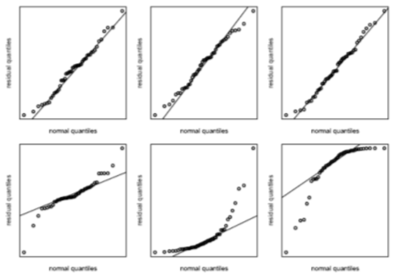
Histograms of Residuals

- Upper three pass, lower three fail



QQ-plots of Residuals

- Upper three pass, lower three fail



Model Interpretation: SHIPDEPT Data

Model Summary

Model	R Squared	Adjusted R Squared	Regression Std. Error	Change Statistics F-stat	df1	df2	Pr(>F)
1	0.808[a]	0.786	8.815				
2	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 no evidence at the 5% significance level that X_2 (proportion shipped by truck) or X_4 (week) provide useful information about Y (weekly labor hours) beyond the information provided by X_1 (total weight shipped in thousands of pounds) and X_3 (average shipment weight in pounds).

R Results: SHIPDEPT

```
Call:
lm(formula = Y ~ X1 + X3, data = SHIPDEPT)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.4311    24.8556   4.443 0.000357 ***
X1            5.0007     2.2607   2.212 0.040948 *
X3           -2.0122     0.6675  -3.014 0.007810 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.815 on 17 degrees of freedom
Multiple R-squared: 0.8082,     Adjusted R-squared: 0.7857
F-statistic: 35.83 on 2 and 17 DF,  p-value: 8.008e-07
```

95% Confidence Interval

Model	Lower Bound	Upper Bound
X1	0.231	9.770
X3	−3.420	−0.604

[a] Response variable: Y.

```
confint(Model)
```


Interpreting Model Results (1): SHIPDEPT

- We found a statistically significant straight-line relationship (at a 5% significance level) between Y and X_1 (holding X_3 constant) and between Y and X_3 (holding X_1 constant).
- Estimated equation: $\hat{Y} = 110.43 + 5.00\,X_1 - 2.01\,X_3$.
- $X_1 = X_3 = 0$ makes no sense for this application, nor do we have data close to $X_1 = X_3 = 0$, so cannot meaningfully interpret $\hat{b}_0 = 110.43$.
- Expect increase of 5 weekly labor hours when total weight increases 1000 pounds and average shipment weight remains constant, for total weights of 2000–10,000 pounds and average weights of 10–30 pounds (95% confident increase is 0.23–9.77).

Interpreting Model Results (2): SHIPDEPT

- Expect decrease of 2.01 weekly labor hours when average weight increases 1 pound and total weight remains constant, for total weights of 2000–10,000 pounds and average weights of 10–30 pounds (95% confident decrease is 0.60–3.42).
- Can expect a prediction of unobserved weekly labor hours from particular values of total weight shipped and average shipment weight to be accurate to within approximately ±17.6 (with 95% confidence).
- 80.8% of the variation in weekly labor hours (about its mean) can be explained by a multiple linear regression relationship between labor hours and (total weight shipped, average shipment weight).

Confidence Interval for $E(Y)$

- Estimate the mean (or expected) value of Y at particular values of $(X_1, X_2, \ldots, X_k)$.
- Formula: $\hat{Y} \pm t\text{-percentile}\,(SE_{\hat{Y}})$.
- Interval is narrower:
  - when n is large;
  - when X's are close to their sample means;
  - when the regression standard error, $\hat{\sigma}$, is small;
  - for lower levels of confidence.
- Example: for shipping example two-predictor model, the 95% confidence interval for $E(Y)$ when $X_1 = 6$ and $X_3 = 20$ is (95.4, 105.0).
- Interpretation: we're 95% confident that the expected weekly labor hours is between 95.4 and 105.0 when total weight shipped is 6000 pounds and average shipment weight is 20 pounds.

Prediction Interval for a Y-value

- Predict an individual value of Y at particular values of $(X_{1,new}, X_{2,new}, \ldots, X_{k,new})$.
- Formula: $\hat{Y}_{new} \pm t\text{-percentile}\,(SE_{\hat{Y}_{new}})$.
- Interval is narrower:
  - when n is large;
  - when X's are close to their sample means;
  - when the residual standard error, $\hat{\sigma}$, is small;
  - for lower levels of confidence.
- Since $S_{\hat{Y}_{new}} > S_{\hat{Y}}$, prediction interval is wider than confidence interval.
- Example: for shipping example two-predictor model, the 95% prediction interval for Y when $X_1 = 6$ and $X_3 = 20$ is (81.0, 119.4).
- Interpretation: we're 95% confident that the actual labor hours in a week is between 81.0 and 119.4 when total weight shipped is 6000 pounds and average shipment weight is 20 pounds.

Summary

- Multiple linear regression model
- Least squares estimators, goodness of fit
- Assumptions about residual standard errors
- Model building and statistical inference
  - Model evaluation
    - Global usefulness test
    - Nested model test / Extra-sums-of-squares F-test
    - Individual test
  - Variable transformations and interactions
  - Influential points and outliers
  - Variable selection
- Model diagnosis
- Prediction

Reading & Assignment

- Ramsey & Schafer (2002):
  - 9.1, 9.2, 9.5, 9.7
  - 10.1, 10.2, 10.3, 10.5
- Assignment 1
  - Available on course website
  - Use R
  - Due: *9:00 am, Tue 02-Oct-2018*