STATS 415: Assessing Model Accuracy Part 1 + linear regression

Prof. Liza Levina

Department of Statistics, University of Michigan

Assessing model accuracy

- Supervised learning: predicting an outcome
- Regression: predicting a continuous outcome
- Classification: predicting a categorical outcome
- Before we develop methods, we need to decide how to evaluate them
- Will start from simple methods as examples to build up assessment tools

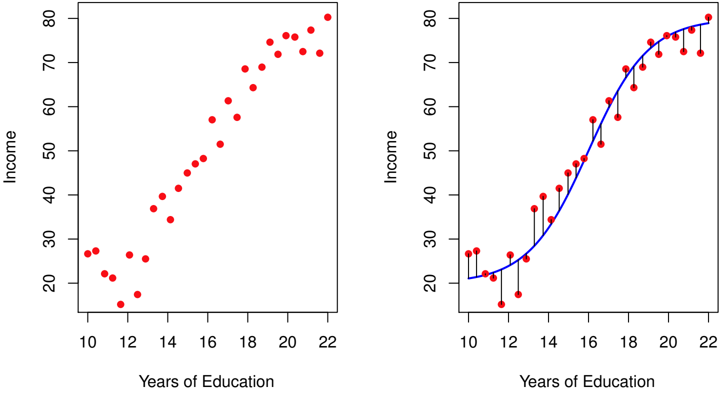
Regression

- Suppose we observe a quantitative y_i and $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, \ldots, n$
- We believe that there is a relationship between y (response) and at least one of the x's (predictors).
- We can model the relationship as

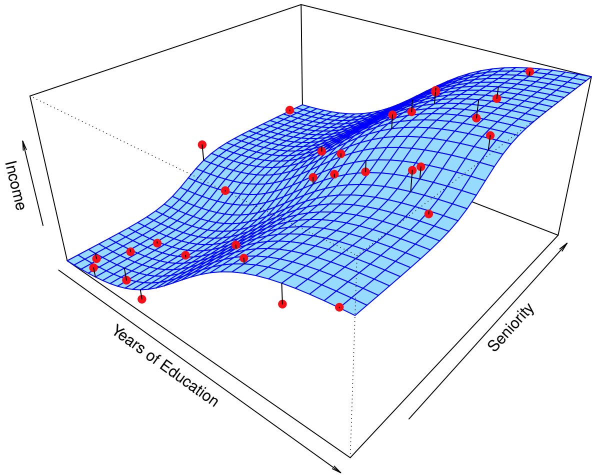
$$y_i = f(x_i) + \varepsilon_i$$

where $f(\cdot)$ is an unknown function and ε_i is an unobserved random error with mean zero, $E(\varepsilon_i) = 0$.

Example: Income vs Education



Example: Income vs Education and Seniority



Estimating $f(\cdot)$

- "Statistical learning" refers to using the data to "learn" $f(\cdot)$.
- Why do we want to learn $f(\cdot)$?
  - Prediction: If we can produce a good estimate for $f(\cdot)$ (and the variance of ε is not too large) we can make accurate predictions for the response, y, for a new value of x.
  - Inference: if we are interested in the type of relationship between y and the x's.

How Do We Estimate $f(\cdot)$?

- We need a set of training data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- We then use the training data and a statistical method to estimate $f(\cdot)$.
- Statistical learning methods:
  - Parametric methods
  - Non-parametric methods

Parametric Methods

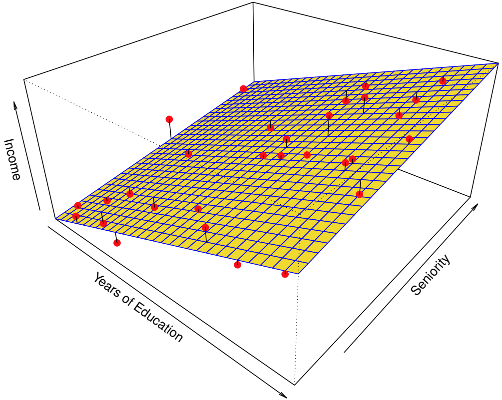
- Reduce the problem of estimating $f(\cdot)$ to estimating a set of parameters (numbers).
- Step 1: Make an assumption about the functional form of $f(\cdot)$, for example a linear model:

$$f(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- Step 2: Use the training data to fit the model, i.e.,estimate the unknown parameters such as β_0, β_1, ..., β_p; for example, with ordinary least squares (OLS).
- This course will cover many more complex and flexible models than linear regression, and methods superior to OLS for fitting them.

Example: A Linear Regression Estimate

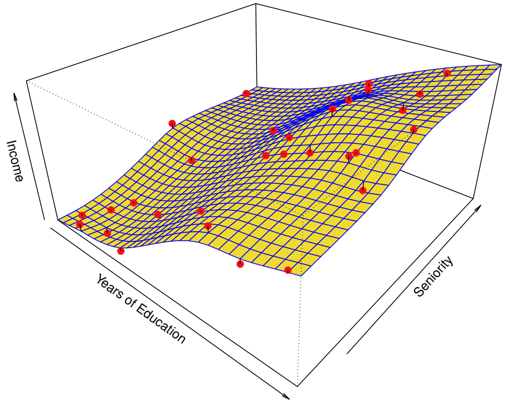
$$f = \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$



Non-parametric Methods

- They do not make explicit assumptions about the functional form of $f(\cdot)$.
- Advantage: They accurately fit a wider range of possible shapes of $f(\cdot)$ (more flexible), therefore likely more accurate
- Disadvantage: A very large number of observations is required to obtain an accurate estimate of $f(\cdot)$.

Example: A Thin-Plate Spline Estimate

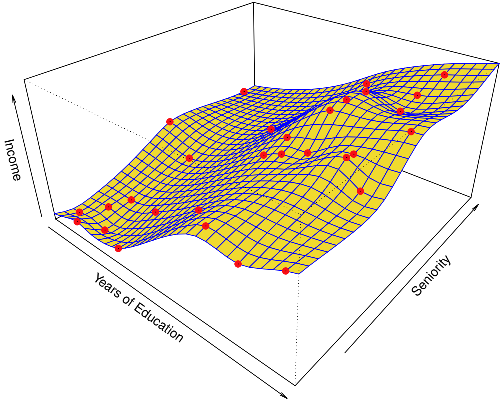


Prediction Accuracy vs Model Interpretability

Why not just always use a more flexible method?

- A simple method such as linear regression is much easier to interpret (inference). For example, in linear regression β_j is the average increase in y for a one unit increase in x_j holding all other variables constant.
- Even for prediction purposes, a simple model can be more accurate if there are not enough data points to fit a more flexible model
- Overfitting: too much flexibility follows the noise too closely

Example of overfitting by a flexible method



Need to design model assessment tools that take all this into account

The Linear Regression Model

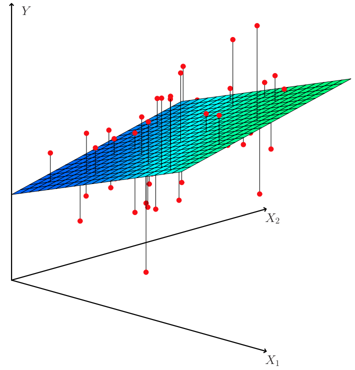
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

- Start with a simple concrete example
- A parametric, linear model
- Parameters are very easy to interpret
- β_0 is the intercept (i.e. the predicted value for y if all the x's are zero) centered data:^beta0=0
- β_j is the slope for the jth variable x_j (i.e. the predicted change in y for one unit increase in x_j if all other variables are held constant).

Ordinary Least Squares (OLS)

We estimate the parameters using least squares, i.e.

$$\arg\min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$



Solving the OLS problem*

* *optional material*

- The problem is quadratic in the unknown variables (β's)

$$\arg\min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} \cdots - \beta_p x_{ip})^2$$

- To minimize, take derivative with respect to each β, set to 0, and solve the system of linear equations: for example,

$$\frac{d}{d\beta_1}\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} \cdots - \beta_p x_{ip})^2 = \frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} \cdots - \beta_p x_{ip})x_{i1}$$

The OLS problem in matrix form *

* *optional material*

- To convert everything to matrix form, add a column of "1"s in front of the data matrix, and write

$$X_{n\times(p+1)} = [\mathbf{1}, X]\ ,\ \ \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$$

- The model is then

$$Y = X\beta + \varepsilon$$

  where Y is n-vector of responses and ε is n-vector of errors
- The OLS problem is

$$\min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

- The solution is

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T = (X^T X)^{-1} X^T Y$$

- If errors are i.i.d. (independent identically distributed) normals $N(0, \sigma^2)$, $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{n}(X^T X)^{-1}$ (unbiased, normally distributed)

Fitted Model

- Fitted model: $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$
- Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$
- Residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS): $\sum_{i=1}^n \hat{\varepsilon}_i^2$

Goodness of Fit: R^2

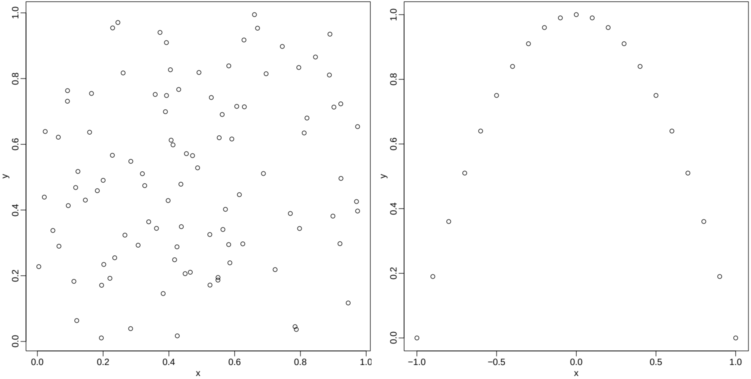
- Some of the variation in y can be explained by changes in x's and some cannot.
- Total variation (Total Sum of Squares): $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Unexplained variation (Residual Sum of Squares): $\sum_{i=1}^{n}\hat{\varepsilon}_i^2$
- R^2: the fraction of variance "explained" by x.

$$R^2 = 1 - \frac{\text{RSS}}{\sum(y_i - \bar{y})^2}$$

- R^2 is always between 0 and 1.
- $R^2 = 0$ means no variance in y is explained by x. $R^2 = 1$ means perfect fit to the data ($\hat{y}_i = y + i$ for all i).

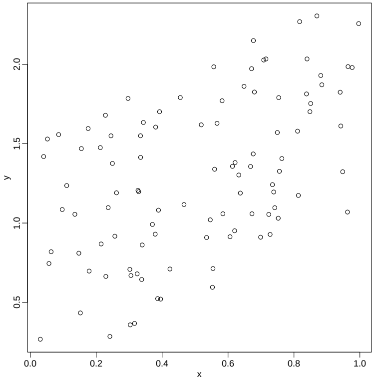
Remarks on R^2

- R^2 near 0 could be ...



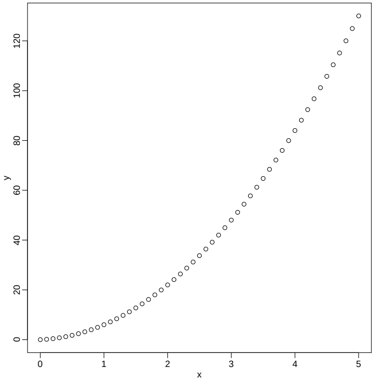
Remarks on R^2

- R^2 close to 0 does not mean that y and x are not linearly related. It could also mean a high error variance.



Remarks on R^2

- Likewise, R^2 close to 1 does not mean the linear model is correct.



Population and Least Squares Lines

- Population line

how to get population line

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

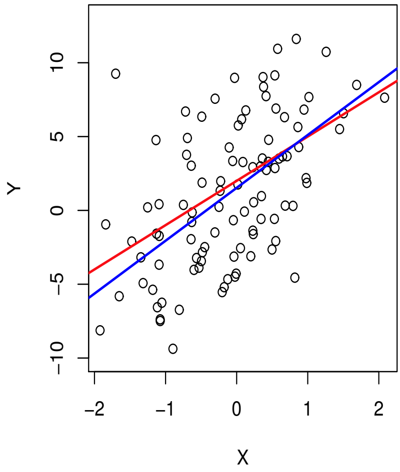
- Least Squares line

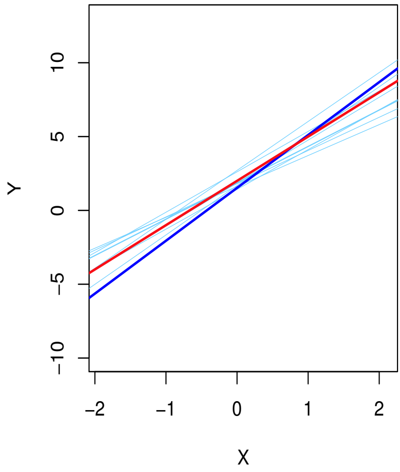
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- We really want to know β_0 through β_p, i.e. the population line. We estimate them with $\hat{\beta}_0$ through $\hat{\beta}_p$, i.e. the least squares line.
- The estimates (our best guesses) $\hat{\beta}_0$ through $\hat{\beta}_p$ are not perfect, just like the sample mean $\bar{x}$ is not a perfect estimate of the population mean $\mu = E(X)$.

Population vs Sample lines

Red: population; Blue: least squares





Some Relevant Inference Questions

- Can we "be sure" x_j is a useful predictor? In other words, are we sure $\beta_j \neq 0$?
- Can we "be sure" that at least one of our variables is a useful predictor?
- These questions can be answered by hypothesis tests

Is x_j a useful predictor?

- Hypothesis testing framework: assume x_j is not useful ($\beta_j = 0$) and see if there is enough evidence to reject this hypothesis.
- $H_0 : \beta_j = 0$ vs $H_a : \beta_j \neq 0$
- Because $\hat{\beta}$ is approximately normal, t-test applies: calculate the t-statistic
$$t = |\hat{\beta}_j| / \mathrm{SD}(\hat{\beta}_j)$$
- If t is large (equivalently p-value is small) we can reject $H_0 : \beta_j = 0$ and conclude x_j is useful in this model.

Marketing example (from the book)

- Response: product sales
- Predictors: money spent on advertising in different media
- Simple regression, model 1: regress Sales on TV ad spending

	Coefficient	Std Err	t-value	p-value
Intercept	7.033	0.458	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001

- Simple regression, model 2: regress Sales on Newpaper ad spending

	Coefficient	Std Err	t-value	p-value
Intercept	12.35	0.621	19.88	<0.0001
Newspaper	0.547	0.0166	3.30	<0.0001

Testing individual variables in multiple regression

- Is there a (statistically detectable) linear relationship between Newspapers and Sales after all the other variables have been accounted for?

	Coefficient	Std Err	t-value	p-value
Intercept	2.939	0.312	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
Radio	0.189	0.0086	21.89	<0.0001
Newspaper	-0.0010	0.0059	-0.18	0.860

- Almost all the explaining that Newspapers could do in simple regression has already been done by TV and Radio in multiple regression!

Is the multiple regression explaining anything?

http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-analysis-of-variance-anova-and-the-f-test

- Need a hypothesis test for
  - H_0: al$\beta_1 = \beta_2 = \cdots = \beta_p = 0$ against
  - H_a: at least one $\beta_j neq 0$
- Tested by the F-test in ANOVA (ANalysis Of VAriance) table.

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

	df	SS	MS	F-value	p-value
Explained	2	4860.2	2430.1	859.6	0.000
Unexplained	197	556.9	2.83		

Categorical (Qualitative) Predictors

- How do you put a categorical variable into a regression equation?
- Code them as indicator variables (dummy variables)
- For example, student status: “not student”=0 and “student”=1.

Interpretation

- Suppose we want to predict bank balance from income and student status.
- Let the new variable

$$\text{Student} = \begin{cases} 0 & \text{if not student} \\ 1 & \text{if student} \end{cases}$$

- Then the regression model is

$$\begin{aligned} \text{Balance} &= \beta_0 + \beta_1 \times \text{Income} + \beta_2 \times \text{Student} \\ &= \begin{cases} \beta_0 + \beta_1 \times \text{Income} & \text{if not student} \\ \beta_0 + \beta_1 \times \text{Income} + \beta_2 & \text{if student} \end{cases} \end{aligned}$$

- β_2 is the average extra balance (positive or negative) each month that students have for given income level. “Not student” is the “baseline”.

	Coefficient	Std Err	t-value	p-value
Intercept	211.1	32.5	6.51	$<$0.0001
Income	5.984	0.557	10.75	$<$0.0001
StudentYes	382.7	65.3	5.859	$<$0.0001

Interaction

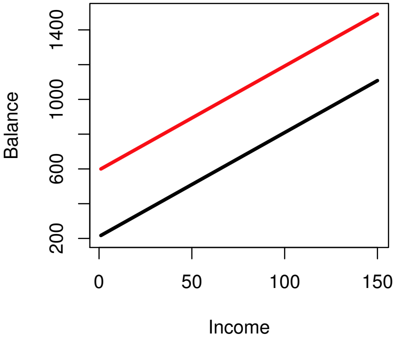
- Our model has forced the line for students and the line for non-students to be parallel.
- Parallel lines say that income has the same effect on balance for students as for non-students.
- If lines aren't parallel then income affects students' and non-students' balances differently.
- Interaction effects: When the effect on y of increasing x_1 depends on another x_2.
  ??

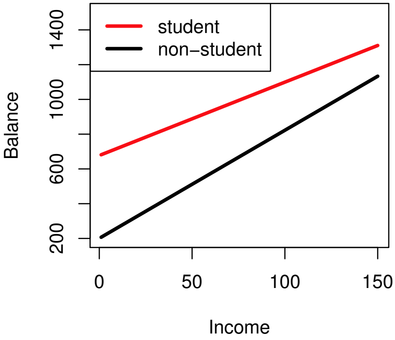
Parallel Regression Lines?

- Regression equation

$$\text{Balance} = \beta_0 + \beta_1 \times \text{Income} + \begin{cases} 0 & \text{if not student} \\ \beta_2 + \beta_3 \times \text{Income} & \text{if student} \end{cases}$$

$$= \begin{cases} \beta_0 + \beta_1 \times \text{Income} & \text{if not student} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{Income} & \text{if student} \end{cases}$$





Should the Lines be Parallel?

	Coefficient	Std Err	t-value	p-value
Intercept	200.6	33.70	5.953	$<$0.0001
Income	6.218	0.592	10.50	$<$0.0001
StudentYes	476.7	104.4	4.568	$<$0.0001
Income*Student	-2.00	1.73	-1.16	0.25

beta3??

whether rejected?

Interaction in Advertising

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{TV} \times \text{Radio}$$

	Coefficient	Std Err	t-value	p-value
Intercept	6.750	0.2479	27.23	$<$0.0001
TV	0.0191	0.0015	12.7	$<$0.0001
Radio	0.0289	0.0089	3.24	0.0014
TV*Radio	0.0011	5.2e-5	20.7	$<$0.0001

- Spending \$1 extra on TV increases average sales by $0.0191 + 0.0011 \times \text{Radio}$

$$\text{Sales} = \beta_0 + (\beta_1 + \beta_3 \times \text{Radio}) \times \text{TV} + \beta_2 \times \text{Radio}$$

- Spending \$1 extra on Radio increases average sales by $0.0289 + 0.0011 \times \text{TV}$

$$\text{Sales} = \beta_0 + (\beta_2 + \beta_3 \times \text{TV}) \times \text{Radio} + \beta_1 \times \text{TV}$$

Potential Fit Problems

There are a number of possible problems that one may encounter when fitting the linear regression model.

- Non-linearity of the data
- Dependence among errors
- Non-constant variance of error terms
- Outliers
- High leverage points
- Collinearity ??

\hat{\beta}=(X^TX)^{-1}X^TY 如果X矩阵的列向量
x1,x2,x3是highly colinear的 那么X^TX很可能rank就
会变小 然后就没有inverse了 或者inverse变得很大