IEOR 165 – Lecture 4
Diagnostics

1 Some Linear Regression Examples

Linear regression is an important method, and so we discuss a few additional examples. First, recall that a linear model given by

$$y = m \cdot x + b + \epsilon,$$

where $x \in \mathbb{R}$ is a single predictor, $y \in \mathbb{R}$ is the response variable, $m, b \in \mathbb{R}$ are the coefficients of the linear model, and ϵ is zero-mean noise with finite variance that is also assumed to be independent of x.

For this linear model, the method of least squares can be used to estimate m, x. If we let

$$\begin{aligned}
\overline{x} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i \\
\overline{y} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n y_i \\
\overline{xy} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i y_i \\
\overline{x^2} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^2,
\end{aligned}$$

then the least squares estimates of m, x are given by

$$\begin{aligned}
\hat{m} &= \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \\
\hat{b} &= \overline{y} - \hat{m} \cdot \overline{x}.
\end{aligned}$$

These equations look different from the ones we derived in the previous lecture, but they can be shown to be equivalent after performing some algebraic manipulation.

1.1 Example: Linear Model of Demand

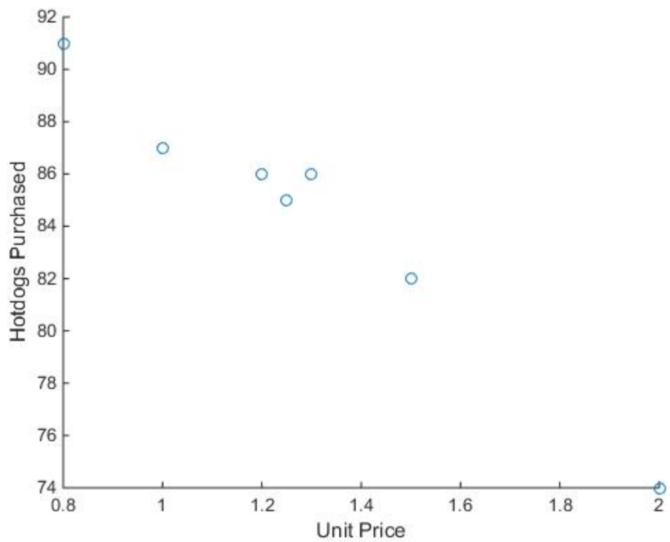
Imagine we are running a several hot dog stands, and for $n = 7$ different stands we have different prices for a hotdog. For these different stands, we record the number of hotdogs purchased (in a single day) at the i-th stand and the price of a single hotdog at the i-th stand. Suppose we would like to build a linear model that predicts demand of hotdogs as a function price, and assume the (paired) data is $H = \{91, 86, 74, 85, 86, 87, 82\}$ and $P = \{0.80, 1.30, 2.00, 1.25, 1.20, 1.00, 1.50\}$.

Consider the following questions and answers:

1. Q: What is the predictor? What is the response?
   A: The predictor is the price P of a hotdog, and the response is the number H of hotdogs purchased.

2. Q: What is the linear model?
   A: The linear model is $H = m \cdot P + b$.

3. Q: Construct a scatter plot of the raw data.
   A:



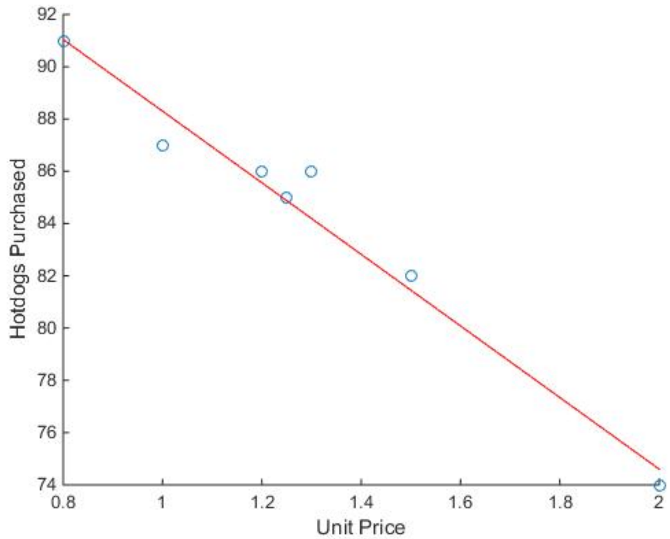
4. Q: Estimate the parameters of the linear model.
   A: We first compute the sample averages:

$$
\begin{aligned}
\overline{x} &= \tfrac{1}{7} \cdot (0.80 + 1.30 + 2.00 + 1.25 + 1.20 + 1.00 + 1.50) = 1.2929 \\
\overline{y} &= \tfrac{1}{7} \cdot (91 + 86 + 74 + 85 + 86 + 87 + 82) = 84.4286 \\
\overline{xy} &= \tfrac{1}{7} \cdot (91 \cdot 0.80 + 86 \cdot 1.30 + 74 \cdot 2.00 + 85 \cdot 1.25 + 86 \cdot 1.20 + 87 \cdot 1.00 + 82 \cdot 1.50) \\
&= 107.4357 \\
\overline{x^2} &= \tfrac{1}{7} \cdot (0.80^2 + 1.30^2 + 2.00^2 + 1.25^2 + 1.20^2 + 1.00^2 + 1.50^2) = 1.7975.
\end{aligned}
$$

Inserting these into the equations for estimating the model parameters gives:

$$
\begin{aligned}
\hat{m} &= \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{107.4357 - 1.2929 \cdot 84.4286}{1.7975 - (1.2929)^2} = -13.7 \\
\hat{b} &= \overline{y} - \hat{m} \cdot \overline{x} = 84.4286 - (-13.7) \cdot 1.2929 = 102.
\end{aligned}
$$

5. Q: Draw the estimated linear model on the scatter plot.
   A:



6. Q: What is the predicted demand if the price was 1.25?
   A: The predicted demand is

$$\hat{H}(1.25) = \hat{m} \cdot 1.25 + \hat{b} = -13.7 \cdot 1.25 + 102 = 85.$$

1.2 Example: Linear Model of Vehicle Miles

Imagine we conduct a survey in which we ask a random subset of the population to provide the following information:

- Annual salary S
- Vehicle miles driven last month M
- County of residence, and we only survey people from the counties

$$C \in \{\text{Alameda}, \text{San Francisco}, \text{San Mateo}, \text{Santa Clara}\}$$

Suppose we would like to build a linear model that predicts vehicle miles driven last month based on an person's annual salary and what county they live in.

Consider the following questions and answers:

1. Q: What is the predictor? What is the response?
   A: The response is vehicle miles M. The predictor variables are more complicated because C is a categorical variable. The predictor variables are: S, $C_1 = \mathbf{1}(C = \text{Alameda})$, $C_2 = \mathbf{1}(C = \text{San Francisco})$, and $C_3 = \mathbf{1}(C = \text{San Mateo})$.

In general, if C is a categorical variable with d possibilities, then we must define $d-1$ binary variables to represent the d possibilities. The $d-1$ binary variables represent d possible combinations because setting the $d-1$ binary variables to zero represents the d-th category. Note that we do *not* define d binary variables.

1.3 Example: Ball Trajectory

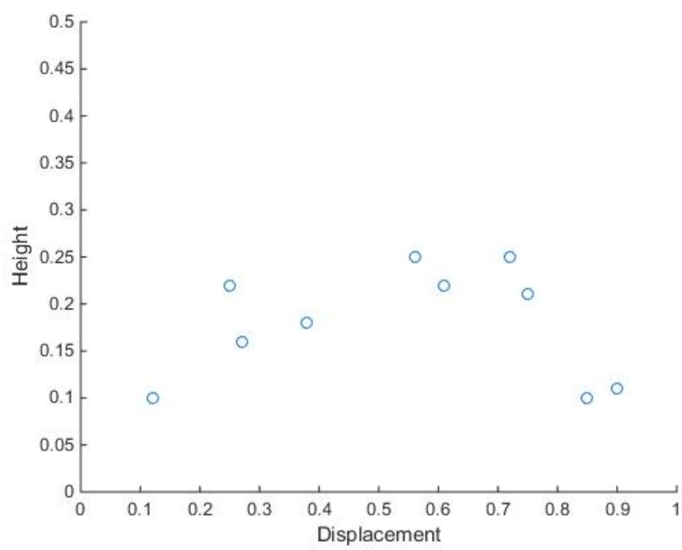
Imagine we are conducting a physics experiment for our class, and the experiment is that we through a small ball and measure its displacement x and height y. Suppose we would like to build a linear model that predicts height of the ball as a function of displacement, and assume the (paired) data is $x = \{0.56, 0.61, 0.12, 0.25, 0.72, 0.85, 0.38, 0.90, 0.75, 0.27\}$ and $y = \{0.25, 0.22, 0.10, 0.22, 0.25, 0.10, 0.18, 0.11, 0.21, 0.16\}$.

Consider the following questions and answers:

1. Q: What is the predictor? What is the response?
   A: The predictor x, and the response y.

2. Q: What is the linear model?
   A: The linear model is $y = m \cdot x + b$.

3. Q: Construct a scatter plot of the raw data.
   A:



4. Q: Estimate the parameters of the linear model.

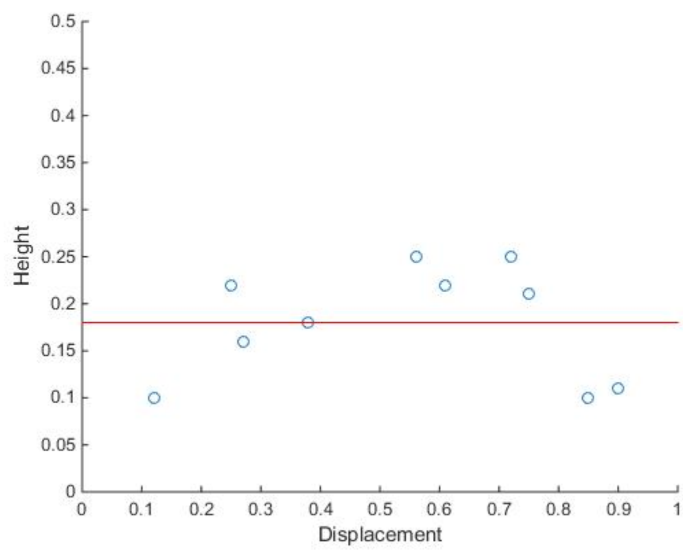
A: We first compute the sample averages:

$$\begin{aligned}\overline{x} &= 0.5410\\ \overline{y} &= 0.1800\\ \overline{xy} &= 0.0974\\ \overline{x^2} &= 0.3593.\end{aligned}$$

Inserting these into the equations for estimating the model parameters gives:

$$\begin{aligned}\hat{m} &= \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{0.0974 - 0.5410 \cdot 0.1800}{0.3593 - (0.5410)^2} = 0\\ \hat{b} &= \overline{y} - \hat{m}\cdot\overline{x} = 0.1800 - 0 \cdot 0.5410 = 0.18.\end{aligned}$$

5. Q: Draw the estimated linear model on the scatter plot.
   A:



6. Q: What is the predicted height if the displacement was 0.2?
   A: The predicted height is

$$\hat{y}(0.2) = \hat{m}\cdot 0.2 + \hat{b} = 0 \cdot 0.2 + 0.18 = 0.18.$$

2 Coefficient of Determination

From a practical standpoint, it can be useful to evaluate the accuracy of a linear model. Given the ubiquity of linear models, a large number of approaches have been developed. The simplest approach is to visually compare a scatter plot of the data to the plot of the estimated linear model;

however, this comparison can be misleading or difficult to evaluate. Another simple approach is known as the coefficient of determination, which is defined as the quantity

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\overline{(y - \hat{y})^2}}{\overline{(y - \overline{y})^2}}$$

This is a common approach, and it is popular because it is easy to compute.

The quantity R^2 ranges in value from 0 to 1. The intuition for this range is that if we chose estimates of $\hat{m} = 0$ and $\hat{b} = \hat{y}$, then the least squares objective would be $\sum_{i=1}^n (y_i - \overline{y})^2$. And since $\hat{y}_i$ correspond to the $\hat{m}, \hat{b}$ estimates that minimize the least squares objective, we must have that $0 \leq \overline{(y - \hat{y})^2} \leq \overline{(y - \overline{y})^2}$. Thus, it holds that

$$0 \leq \frac{\overline{(y - \hat{y})^2}}{\overline{(y - \overline{y})^2}} \leq 1,$$

which means that $0 \leq R^2 \leq 1$.

Furthermore, the closer the quantity R^2 is to the value 1, then the better the estimated linear model fits the measured data. The reason is that the better the model fits the data, the closer the $\hat{y}_i$ are to the y_i. Thus, $\overline{(y - \hat{y})^2}$ will be close to the value 0 when the model fits the data very well. And so $\frac{\overline{(y-\hat{y})^2}}{\overline{(y-\overline{y})^2}}$ will be close to 0, and R^2 will be close to 1.

There is a subtlety to this definition, however. In particular, it is the case that $\hat{y}_i$ can only get closer to y_i as the number of predictors increases. So if we have a large number of predictors (even if the predictors are completely irrelevant to the real system), it is typically the case that $(y_i - \hat{y}_i)^2$ is small. Hence, R^2 will go closer to 1 as the number of predictors increases. As a result, sometimes the adjusted R^2 value is used instead. The adjusted R^2 is defined as

$$R^2_{\text{adj}} = R^2 - (1 - R^2) \cdot \frac{d}{n - d - 1},$$

where d is the total number of predictors (not including the constant/intercept term), and n is the number of data points. The adjusted R^2 is only of interest when we have more than one predictor variable.

2.1 Example: Linear Model of Demand

We can compute R^2 for the hotdog example. First, we compute $\hat{y}_i$:

$$\begin{aligned}
&\hat{y}_1 = \hat{m} \cdot x_1 + \hat{b} = -13.7 \cdot 0.8 + 102 = 91.0400 && \hat{y}_5 = 85.5600 \\
&\hat{y}_2 = 84.1900 && \hat{y}_6 = 88.3000 \\
&\hat{y}_3 = 74.6000 && \hat{y}_7 = 81.4500 \\
&\hat{y}_4 = 84.8750 &&
\end{aligned}$$

Next, we compute $(y_i - \hat{y}_i)^2$:

$$(y_1 - \hat{y}_1)^2 = (91 - 91.0400)^2 = 0.0016 \qquad (y_5 - \hat{y}_5)^2 = 0.1936$$
$$(y_2 - \hat{y}_2)^2 = 3.2761 \qquad (y_6 - \hat{y}_6)^2 = 1.6900$$
$$(y_3 - \hat{y}_3)^2 = 0.3600 \qquad (y_7 - \hat{y}_7)^2 = 0.3025$$
$$(y_4 - \hat{y}_4)^2 = 0.0156$$

We also compute $(y_i - \overline{y})^2$:

$$(y_1 - \overline{y})^2 = (91 - 84.4286)^2 = 43.1833 \qquad (y_5 - \overline{y})^2 = 2.4694$$
$$(y_2 - \overline{y})^2 = 2.4694 \qquad (y_6 - \overline{y})^2 = 6.6122$$
$$(y_3 - \overline{y})^2 = 108.7551 \qquad (y_7 - \overline{y})^2 = 5.8980$$
$$(y_4 - \overline{y})^2 = 0.3265$$

Finally, we can compute R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} =$$
$$1 - \frac{0.0016+3.2761+0.3600+0.0156+0.1936+1.69+0.3025}{43.1833+2.4694+108.7551+0.3265+2.4694+6.6122+5.8980} = 0.96$$

2.2 Example: Ball Trajectory

We can compute R^2 for the physics example. First, we compute $\hat{y}_i$. Since $\hat{m} = 0$, we have that $\hat{y}_i = \hat{b} = 0.18$. Next, we compute $(y_i - \hat{y}_i)^2$:

$$(y_i - \hat{y}_i)^2 = \{8248.3, 7365.1, 5449.4, 7194.4, 7365.1, 7537.7, 6694.5\}.$$

We also compute $(y_i - \overline{y})^2$:

$$(y_i - \overline{y})^2 = \{8248.3, 7365.1, 5449.4, 7194.4, 7365.1, 7537.7, 6694.5\}.$$

Since $(y_i - \hat{y}_i) = (y_i - \overline{y})$, we have that

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - 1 = 0.$$