Lecture 13 (Ch 3 and 5.5)

Simple linear regression:
\[ y = \alpha + \beta x \]

Regression on transformed data (e.g.):
\[ y = \alpha + \beta \log x \]

Polynomial regression:
\[ y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots \]

Multiple (linear) regression:
\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \]

For all of these,
- Don't worry about overfitting!
- ANOVA decomposition:
  \[ \frac{SS_{exp}}{SS_T} = R^2 \]
  \[ s_e = \sqrt{\frac{SSE}{n-2}} \]
  Goodness-of-fit, typical error in prediction

In multiple regression:
- \( (n-2) \) [e.g. in \( s_e \)] \( \rightarrow \) \( n-(k+1) \) \( \approx \) df
- There are 2 "new" issues: Interaction, collinearity, e.g. \((x_1 \times x_2)\) (below)

Both of these affect the interpretation of \( \beta \)'s.

E.g. if \[ y = 1.2 + 2.3 x_1 - 3.4 x_2 \text{ wealth, health} \]

Then, on average, e.g. life span health

- Does \( y \) change +2.3 units when \( x_1 \) increases by 1 unit?
- Does \( y \) change -3.4 units when \( x_2 \) increases by 1 unit?

Yes, iff there is no interaction and no collinearity. (see below).
What does it look like?

\[ y = x_1 \times x_2 \]

The effect of one predictor on \( y \) depends on other predictor(s)!

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 = \alpha + \beta_1 x_1 + (\beta_2 + \beta_3 x_1) x_2 \]

For this reason, the \( \beta \)'s are uninterpretable, i.e. They don't tell you how much \( y \) changes when \( x \) is changed by 1 unit.

Guidance

Q: How do we know when to include an interaction term?

A: 1) Look at scatterplots for signs of a saddle surface.
2) Include it, and see if it significantly improves \( R^2 \).
3) Ultimately, if it improves predictions, then include it! But always be mindful of overfitting.

4) Look at residual plots for signs of a saddle

5) In class we will learn another way of deciding whether or not an interaction exists.
2) In multiple regression, in addition to interaction, there is one more thing to worry about: **collinearity.**

Let's return to the first (important) step: *Look at data!*

Multiple predictors $\rightarrow$ matrix of scatterplots:

\[
\begin{align*}
\mathbf{X} &= \begin{pmatrix}
\mathbf{x}_1 \\
\mathbf{x}_2
\end{pmatrix}, \\
\mathbf{Y} &= \begin{pmatrix}
\mathbf{y}
\end{pmatrix}
\end{align*}
\]

A linear association between $x_1$ and $x_2$ is called "**collinearity.**"

It's a "bad" disease. (See below, for why.) One cure is to simply exclude one of the predictors $x_1, x_2$.

![Diagram showing scatterplots](image)

Almost anything is OK here. Don't look for a linear relation. In fact, if you get a high-correlation linear pattern in one of them, then, you probably don't need the other one at all!

**A consequence** of collinearity is that it renders the $\beta$'s **un-interpretable** (as the avg. rate of change of $y$ ...):

Ordinarily, in $y = \alpha + \beta_1 x_1 + \beta_2 x_2$,

- $\beta_1$ = avg. rate of change in $y$, for 1 unit change in $x_1$; **if** $x_2$ **is held constant.**

But if $x_1$ and $x_2$ are correlated, then one cannot hold one of them fixed.

In fact, in an example, lifespan $\sim \alpha + \beta_1$ (health) + $\beta_2$ (wealth).

I once got a value of $\beta_1$ that was negative, in spite of the positive association displayed in the scatterplot of life span vs. health. The culprit was collinearity.
Guidance: When you suspect extreme collinearity, then just drop one of the predictors from the model. In less severe cases, ask a statistician! Or look-up "principal component regression."

In summary, even though both interaction and collinearity make the β's un-interpretable, they are very different concepts.

\[ \text{collinearity} \neq \text{interaction} \]

Again, don’t forget the main purpose of regression: To make predictions with reduced uncertainty.

<table>
<thead>
<tr>
<th>best prediction</th>
<th>± uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before regression: ( \bar{y} )</td>
<td>( s_y )</td>
</tr>
<tr>
<td>After regression: ( \hat{y} )</td>
<td>( s_e = \sqrt{\frac{\text{SSE}}{n-(k+1)}} )</td>
</tr>
</tbody>
</table>

Recall that multiple regression, even with polynomial terms, is still an example of linear regression (1) because the model \( y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon \) is linear in the parameters \( \alpha, \beta_1, \ldots \). So, when we do derivatives of SSE, we get a bunch of linear equations that can be solved exactly, giving a unique solution for \( \hat{\alpha}, \hat{\beta}_1, \ldots \). This is the advantage of linear models. Meanwhile, the models can be as non-linear (in \( x \)'s) as you would like them to be. So, multiple regression is a win-win.
Geometrically, the reason why the $\beta$'s become uncertain and uninterpretable is that we are then trying to fit a plane through a cigar-shaped cloud in 3D, as opposed to a planar cloud.

That is ambiguous! There are lots of planes one can fit through a cigar-shaped cloud in 3D. Of course, these different fits differ in their $\hat{\beta}_1, \hat{\beta}_2$. That's why they become meaningless. You can also see that the predictions, $\hat{y}$, are affected by collinearity; however, note that the effect is mostly in their uncertainty. (Move, in Ch. II).

Another bad consequence of coll. is that it effectively reduces the amount of information in the data, which, in turn, leads to more uncertain estimates of the $\beta$'s and predictions. We'll see that in Ch. II.

Another bad consequence of coll. is that it can lead to overfitting. This is because the various predictors come with pams to be estimated from data, but the various predictors are essentially carrying the same information, i.e. there is effectively more params. Than data, hence overfitting can happen.
For different levels of collinearity, the problem of uncertain \( \beta \)'s and predictions can be qualitatively different.

For very little collinearity, there is a reasonably unique plain one can fit the black dots. For mild collinearity (red), there is no unique surface to fit the "cigar." For extreme collinearity (blue), the "fit" is a "vertical" surface.

Think about what this does to the predictions.

Perhaps, you will see these in the lab.
One last thing in regression (until Ch. 11): The visual assessment of fit quality.

Plot \( \hat{y} \) vs. \( y \):
- Random, symmetric about diagonal → Good
- Diagonal → Bad
- Under/over forecasting small/large \( y \).

Residual Plot: (Mona Lisa)
- Errors (last column of "table" before)
- Random, symmetric about error = 0 line → Good Model/fit
  - So nothing more to do
- Not recommended → Bad Model/fit
  - Things to do:
    1) Transform data, or
    2) Fit different model (e.g., polynomial or interaction)
n = 100
x = rnorm(n, 0, 1)
y = 1 + x + x^2 + rnorm(n, 0, 0.1)
lm.1 = lm(y ~ x)
yhat1 = predict(lm.1)
lm.2 = lm(y ~ x + I(x^2))
yhat2 = predict(lm.2)
par(mfrow=c(2, 2))
plot(x, y)
abline(lm.1)
plot(yhat1, y - yhat1); abline(h=0)
plot(yhat2, y - yhat2); abline(h=0)

The residual plot for $y = \alpha + \beta x$ clearly shows that $y = \alpha + \beta x$ is not a good fit to the data; because it shows a non-linear pattern, we try $y = \alpha + \beta x + \beta x^2$ and then we see that the residual plot looks good (i.e. shows no pattern). This may seem trivial in this case, because you can see that you need $y = \alpha + \beta x + \beta x^2$ from the scatterplot of the data, to begin with! But even in cases where you cannot make a scatterplot of data (e.g. in multiple regression) you can still make a residual plot.
Why \((n-1), (n-2), \ldots, \chi = \frac{1}{n} \sum_{i=1}^{n} Y_i \), why \(n\)?

A. \([Y_1, Y_2, \ldots, Y_n]\) are all independent \(\Rightarrow\) \(df(\text{of } \sum_{i=1}^{n} Y_i) = n\)

Q. \(S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2\), why \((n-1)\)?

A. \([Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y}]\) are not all independent.

There is 1 constraint on them \(\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0\)

I.e. There are \((n-1)\) independent terms \(\Rightarrow\) \(df(\text{of } S^2) = n-1\)

[There are other reasons for \((n-1)\), too].

Q. \(s^2_0 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2\), why \((n-2)\)?

\([Y_i - \hat{Y}_i]\) satisfy 2 constraints (below) \(\Rightarrow\) \(df(\text{of } SSE) = n-2\)

\(^1\text{st constraint:}\ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i) = \bar{Y} - \hat{\alpha} - \hat{\beta} \bar{x} = 0\)

\(\bar{Y} - \hat{\beta} \bar{x}\) (See \(\hat{\alpha}\) eqn.)

\(^2\text{nd constraint:}\ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) x_i = \frac{1}{n} \sum_{i=1}^{n} (Y_i x_i - (\hat{\alpha} + \hat{\beta} x_i) x_i) = \bar{Y}x - \hat{\alpha} \bar{x} - \hat{\beta} \bar{x}^2\)

\(= \bar{Y}x - (\bar{Y} - \hat{\beta} \bar{x}) \bar{x} + \hat{\beta} \bar{x}^2\)

\(= (\bar{Y}x - \bar{Y} \bar{x}) - \hat{\beta} (\bar{x}^2 - \bar{x}^2) = 0\)

\(\frac{\bar{Y}x - \bar{Y} \bar{x}}{\bar{x}^2 - \bar{x}} = 0\) (See \(\hat{\beta}\) eqn.)
Ch 5 (5.5, 5.6) "Bridge between 1st & 2nd half of course"

Sampling Distribution: Extremely Important!!

Population, \( x \)

Sample 1

\( \overline{x}, s, p \)

Sample 2 \( \rightarrow \) \( \overline{x}, s, p \) \( \rightarrow \) \( \overline{x}, s, p \)

Sample prop.

\( \Rightarrow 10^6 \overline{x} \)'s \( \Rightarrow \) histogram

or \( 10^6 s \)'s

frec. \( \sim N(\mu, \sigma^2) \)

The sampling dist. (of the sample mean) is a distribution, ie. a \( p(x) \) or an \( f(x) \) that can be derived mathematically, or simply assumed as a description of the population of all \( \overline{x} \)'s.

The only reason I talk about a histogram is to make the concept of the sampling dist. more intuitive. The histogram is sometimes called the "empirical sampling dist."
```r
ntrial = 64
xbar = numeric(ntrial)
par(mfrow=c(8,8))
for( trial in 1:ntrial ){
  x = rnorm(50, 0, 1)
  hist(x, breaks=10)
  xbar[trial] = mean(x)
}
hist(xbar, main="")
```

Q: What's $\bar{x}$ in each hist above? What's the mean of the $\bar{x}$'s?

Q: What's $s$ in each hist above? What's $s$ of the $\bar{x}$'s?
Consider fitting the model $y_i = \beta x_i x_2 + \epsilon_i$, $i = 1, 2, \ldots, n$ to data on $y$, $x_1$, $x_2$. Show that the OLS estimate of $\beta$ is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i x_2 y_i}{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} x_2^2)}$$

Hint: find the critical value of $\beta$ that minimizes

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i x_2)^2$$

The article "The Undrained Strength of Some Thawed Permafrost Soils" (Canadian Geotech. J., 1979: 420-427) contained the accompanying data on $y$ shear strength of sandy soil (kPa), $x_1$ depth (m), and $x_2$ water content (%). Obs Depth Content Strength

<table>
<thead>
<tr>
<th>Obs</th>
<th>Depth</th>
<th>Content</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.9</td>
<td>31.5</td>
<td>14.7</td>
</tr>
<tr>
<td>2</td>
<td>36.6</td>
<td>27.0</td>
<td>48.0</td>
</tr>
<tr>
<td>3</td>
<td>36.8</td>
<td>25.9</td>
<td>25.6</td>
</tr>
<tr>
<td>4</td>
<td>6.1</td>
<td>39.1</td>
<td>10.0</td>
</tr>
<tr>
<td>5</td>
<td>6.9</td>
<td>39.2</td>
<td>16.0</td>
</tr>
<tr>
<td>6</td>
<td>6.9</td>
<td>38.3</td>
<td>16.8</td>
</tr>
<tr>
<td>7</td>
<td>7.3</td>
<td>33.9</td>
<td>20.7</td>
</tr>
<tr>
<td>8</td>
<td>8.4</td>
<td>33.8</td>
<td>38.8</td>
</tr>
<tr>
<td>9</td>
<td>6.5</td>
<td>27.9</td>
<td>16.9</td>
</tr>
<tr>
<td>10</td>
<td>8.0</td>
<td>33.1</td>
<td>27.0</td>
</tr>
<tr>
<td>11</td>
<td>4.5</td>
<td>26.3</td>
<td>16.0</td>
</tr>
<tr>
<td>12</td>
<td>9.9</td>
<td>37.0</td>
<td>24.9</td>
</tr>
<tr>
<td>13</td>
<td>2.9</td>
<td>34.6</td>
<td>7.3</td>
</tr>
<tr>
<td>14</td>
<td>2.0</td>
<td>36.4</td>
<td>12.8</td>
</tr>
</tbody>
</table>

a) Perform regression to predict $y$ from $x_1$, $x_2$, $x_3 = x_1^2$, $x_4 = x_2^2$, and $x_5 = x_1 x_2$; and write down the coefficients of the various terms.

b) Can you interpret the regression coefficients? Explain.

c) Compute $R^2$ and explain what it says about goodness-of-fit ("in English").

d) Compute $s_\text{e}$, and interpret ("in English").

e) Produce the residual plot (residuals vs. *predicted* $y$), and explain what it suggests, if any.

f) Now perform regression to predict $y$ from $x_1$ and $x_2$ only.

g) Compute $R^2$ and explain what it says about goodness-of-fit.

h) Compare the above two $R^2$ values. Does the comparison suggest that at least one of the higher-order terms in the regression eqn provides useful information about strength?

i) Compute $s_\text{e}$ for the model in part f, and compare it to that in part d. What do you conclude?
For each of the data sets a) `hw_3_dat1.txt` and b) `hw_3_dat2.txt`, find the "best" (OLS) fit, and report R-squared and the standard deviation of the errors. Do not use some ad hoc criterion (like maximum R2) to determine what is the "best" model. Instead, use your knowledge of regression to find the best model, and explain in words why you think you have the best model. Specifically, make sure you address 1) collinearity, 2) interaction, and 3) nonlinearity.

Suppose we had to come up with a measure of spread that was based on $\sum |x_i - \bar{x}|$. What should it be divided by? Explain.

a) write R code to produce the sampling distribution of the sample maximum, for samples of size 50 taken from a standard Normal. Use 5000 trials,
b) Repeat for the sample minimum.

Turn-in your code, and the resulting 2 histograms.
FYI, these distributions arise naturally when one tries to model extreme events, e.g. the biggest storms, the strongest earthquakes, the brightest stars, the smallest forms of life, etc.