Regression Model Assumptions

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon \]

Recall the key assumptions of our linear regression model:

(i) The mean of \( Y \) is linear in \( X \)'s.

(ii) The additive errors (deviations from line)

- are normally distributed
- independent from each other
- identically distributed (i.e., they have constant variance)

\[ Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \]
Regression Model Assumptions

Inference and prediction relies on this model being “true”!

If the model assumptions do not hold, then all bets are off:

▶ prediction can be systematically biased
▶ standard errors, intervals, and t-tests are wrong

We will focus on using graphical methods (plots!) to detect violations of the model assumptions.
Here we have two datasets... Which one looks compatible with our modeling assumptions?
Example

### Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.816420516</td>
</tr>
<tr>
<td>R Square</td>
<td>0.66654246</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.629491622</td>
</tr>
<tr>
<td>Standard Error</td>
<td>1.236603323</td>
</tr>
<tr>
<td>Observations</td>
<td>11</td>
</tr>
</tbody>
</table>

#### ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>27.51000091</td>
<td>27.51000091</td>
<td>17.99994297</td>
<td>0.002169629</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>13.76269</td>
<td>1.529187778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>41.27269091</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.000090909</td>
<td>1.124746791</td>
<td>2.867347828</td>
<td>0.025734051</td>
<td>0.455768905</td>
<td>5.544444913</td>
</tr>
<tr>
<td>X1</td>
<td>0.500090909</td>
<td>0.117905501</td>
<td>4.241455289</td>
<td>0.002169629</td>
<td>0.233370137</td>
<td>0.768811681</td>
</tr>
</tbody>
</table>

### Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.816236508</td>
</tr>
<tr>
<td>R Square</td>
<td>0.666242034</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.629157815</td>
</tr>
<tr>
<td>Standard Error</td>
<td>1.237214205</td>
</tr>
<tr>
<td>Observations</td>
<td>11</td>
</tr>
</tbody>
</table>

#### ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>27.5</td>
<td>27.5</td>
<td>17.96564849</td>
<td>0.002178816</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>13.77629091</td>
<td>1.53069899</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>41.27629091</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.000090909</td>
<td>1.125302416</td>
<td>2.66757884</td>
<td>0.025758941</td>
<td>0.455280175</td>
<td>5.546520007</td>
</tr>
<tr>
<td>X2</td>
<td>0.5</td>
<td>0.117963746</td>
<td>4.23859039</td>
<td>0.002178816</td>
<td>0.233147468</td>
<td>0.768852532</td>
</tr>
</tbody>
</table>
Example

The regression output values are exactly the same...

Thus, whatever decision or action we might take based on the output would be the same in both cases!
Example

...but the residuals (plotted against $\hat{Y}$) look totally different!!

Plotting $e$ vs $\hat{Y}$ is your #1 tool for finding fit problems.
Residual Plots

We use residual plots to “diagnose” potential problems with the model.

From the model assumptions, the error term ($\epsilon$) should have a few properties... we use the residuals ($e$) as a proxy for the errors as:

$$
\epsilon_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) 
\approx y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi}) 
= e_i
$$
Residual Plots

What kind of properties should the residuals have??

\[ e_i \approx N(0, \sigma^2) \text{ iid and independent from the X’s} \]

- We should see no pattern between \( e \) and each of the \( X \)’s
- This can be summarized by looking at the plot between \( \hat{Y} \) and \( e \)
- Remember that \( \hat{Y} \) is “pure \( X \)”, i.e., a linear function of the \( X \)’s.

If the model is good, the regression should have pulled out of \( Y \) all of its “\( x \) ness”... what is left over (the residuals) should have nothing to do with \( X \).
Example – Mid City (Housing)

Left: $\hat{y}$ vs. $y$
Right: $\hat{y}$ vs $e$

![Graph showing $y=price$ vs $\hat{y}$](image1)

![Graph showing $e=resid$ vs $\hat{y}$](image2)
Example – Mid City (Housing)

Size vs. e

![Scatter plot of size vs. residuals for midcity multiple regression.](image)
In the Mid City housing example, the residuals plots (both $X$ vs. $e$ and $\hat{Y}$ vs. $e$) showed no obvious problem...

This is what we want!!

Although these plots don’t guarantee that all is well it is a very good sign that the model is doing a good job.
Non Linearity

Example: *Telemarketing*

- How does length of employment affect productivity (number of calls per day)?
Non Linearity

Example: Telemarketing

- Residual plot highlights the non-linearity!
Non Linearity

What can we do to fix this?? We can use multiple regression and transform our $X$ to create a no linear model...

Let’s try

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

The data...

<table>
<thead>
<tr>
<th>months</th>
<th>months2</th>
<th>calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>121</td>
<td>22</td>
</tr>
<tr>
<td>14</td>
<td>196</td>
<td>23</td>
</tr>
<tr>
<td>15</td>
<td>225</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Linear Model

Regression Statistics

<table>
<thead>
<tr>
<th>Regression</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>1</td>
<td>397.445862</td>
<td>397.445862</td>
<td>124.408882</td>
<td>1.62235E-09</td>
</tr>
<tr>
<td>R Square</td>
<td>1</td>
<td>397.445862</td>
<td>397.445862</td>
<td>124.408882</td>
<td>1.62235E-09</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>1</td>
<td>397.445862</td>
<td>397.445862</td>
<td>124.408882</td>
<td>1.62235E-09</td>
</tr>
<tr>
<td>Standard Error</td>
<td>18</td>
<td>57.50413798</td>
<td>3.194674332</td>
<td>1.62235E-09</td>
<td>1.62235E-09</td>
</tr>
<tr>
<td>Observations</td>
<td>19</td>
<td>454.95</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANOVA Table

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.426971138</td>
<td>9.580270766</td>
<td>1.7206E-08</td>
<td>10.67281476</td>
<td>16.66872498</td>
</tr>
<tr>
<td>months</td>
<td>0.743514848</td>
<td>0.066659792</td>
<td>11.1538296</td>
<td>1.62235E-09</td>
<td>0.603467823</td>
</tr>
</tbody>
</table>
Telemarketing

With $X^2$

**Regression Statistics**

- Multiple R: 0.981014716
- R Square: 0.962389873
- Adjusted R Square: 0.957965152
- Standard Error: 1.003251396
- Observations: 20

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>437.8392728</td>
<td>218.9196364</td>
<td>217.5029608</td>
<td>7.76E-09</td>
</tr>
<tr>
<td>Residual</td>
<td>17</td>
<td>17.11072717</td>
<td>1.006513363</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>454.95</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.140471176</td>
<td>-0.060479351</td>
<td>0.95247918</td>
<td>-5.040792846</td>
<td>4.759850493</td>
</tr>
<tr>
<td>months</td>
<td>2.310202389</td>
<td>9.236313153</td>
<td>4.89632E-08</td>
<td>1.782491725</td>
<td>2.837913052</td>
</tr>
<tr>
<td>Months2</td>
<td>-0.04011825</td>
<td>-6.334983539</td>
<td>7.46662E-06</td>
<td>-0.053479312</td>
<td>-0.026757188</td>
</tr>
</tbody>
</table>
Telemarketing

\[ \hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2 \]
Telemarketing

What is the marginal effect of $X$ on $Y$?

$$\frac{\partial E[Y|X]}{\partial X} = \beta_1 + 2\beta_2 X$$

- To better understand the impact of changes in $X$ on $Y$ you should evaluate different scenarios.
- Moving from 10 to 11 months of employment raises productivity by 1.47 calls
- Going from 25 to 26 months only raises the number of calls by 0.27.
Polynomial Regression

Even though we are limited to a linear mean, it is possible to get nonlinear regression by transforming the $X$ variable.

In general, we can add powers of $X$ to get polynomial regression:
\[
Y = \beta_0 + \beta_1 X + \beta_2 X^2 \ldots + \beta_m X^m
\]

You can fit any mean function if $m$ is big enough. Usually, $m = 2$ does the trick.
Closing Comments on Polynomials

We can always add higher powers (cubic, etc) if necessary.

Be very careful about predicting outside the data range. The curve may do unintended things beyond the observed data.

Watch out for over-fitting... remember, simple models are “better”.
Be careful when extrapolating...
...and, be careful when adding more polynomial terms!
Variable Interaction

So far we have considered the impact of each independent variable in a additive way.

We can extend this notion by the inclusion of multiplicative effects through interaction terms. This provides another way to model non-linearities

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i}X_{2i}) + \varepsilon \]

\[ \frac{\partial E[Y|X_1, X_2]}{\partial X_1} = \beta_1 + \beta_3 X_2 \]

What does that mean?
Example: College GPA and Age

Consider the connection between college and MBA grades:
A model to predict McCombs GPA from college GPA could be

$$GPA^{MBA} = \beta_0 + \beta_1 GPA^{Bach} + \varepsilon$$

| Estimate  | Std.Error | t value | Pr(>|t|)    |
|-----------|-----------|---------|-------------|
| BachGPA   | 0.26269   | 0.09244 | 2.842       |
|           |           |         | 0.00607 **  |

For every 1 point increase in college GPA, your expected GPA at McCombs increases by about .26 points.
College GPA and Age

However, this model assumes that the marginal effect of College GPA is the same for any age.

It seems that how you did in college should have less effect on your MBA GPA as you get older (farther from college).

We can account for this intuition with an interaction term:

\[
GPA_{MBA} = \beta_0 + \beta_1 GPA^{Bach} + \beta_2 (Age \times GPA^{Bach}) + \varepsilon
\]

Now, the college effect is

\[
\frac{\partial E[GPA_{MBA} | GPA^{Bach}, Age]}{\partial GPA^{Bach}} = \beta_1 + \beta_2 Age.
\]

Depends on Age!
College GPA and Age

\[ GPA^{\text{MBA}} = \beta_0 + \beta_1 GPA^{\text{Bach}} + \beta_2 (\text{Age} \times GPA^{\text{Bach}}) + \varepsilon \]

Here, we have the interaction term but do not the main effect of age... what are we assuming?

|              | Estimate  | Std.Error | t value | Pr(>|t|)   |
|--------------|-----------|-----------|---------|------------|
| BachGPA     | 0.455750  | 0.103026  | 4.424   | 4.07e-05 *** |
| BachGPA:Age | -0.009377 | 0.002786  | -3.366  | 0.00132 **  |
College GPA and Age

Without the interaction term

- Marginal effect of College GPA is $b_1 = 0.26$.

With the interaction term:

- Marginal effect is $b_1 + b_2 \cdot \text{Age} = 0.46 - 0.0094 \cdot \text{Age}$.

<table>
<thead>
<tr>
<th>Age</th>
<th>Marginal Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.22</td>
</tr>
<tr>
<td>30</td>
<td>0.17</td>
</tr>
<tr>
<td>35</td>
<td>0.13</td>
</tr>
<tr>
<td>40</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Non-constant Variance

Example...

This violates our assumption that all $\varepsilon_i$ have the same $\sigma^2$. 
Non-constant Variance

Consider the following relationship between $Y$ and $X$:

$$Y = \gamma_0 X^{\beta_1} (1 + R)$$

where we think about $R$ as a random percentage error.

- On average we assume $R$ is 0...
- but when it turns out to be 0.1, $Y$ goes up by 10%!
- Often we see this, the errors are multiplicative and the variation is something like $\pm 10\%$ and not $\pm 10$.
- This leads to non-constant variance (or heteroskedasticity)
The Log-Log Model

We have data on $Y$ and $X$ and we still want to use a linear regression model to understand their relationship... what if we take the log (natural log) of $Y$?

$$
\log(Y) = \log \left[ \gamma_0 X^{\beta_1}(1 + R) \right]
$$

$$
\log(Y) = \log(\gamma_0) + \beta_1 \log(X) + \log(1 + R)
$$

Now, if we call $\beta_0 = \log(\gamma_0)$ and $\epsilon = \log(1 + R)$ the above leads to

$$
\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon
$$

a linear regression of $\log(Y)$ on $\log(X)$!
The Log-Log Model

Consider a country’s GDP as a function of IMPORTS:

- Since trade multiplies, we might expect to see %GDP to increase with %IMPORTS.
Elasticity and the log-log Model

In a log-log model, the slope $\beta_1$ is sometimes called **elasticity**.

In English, a 1% increase in $X$ gives a beta % increase in $Y$.

$$\beta_1 \approx \frac{d\%Y}{d\%X} \quad (Why?)$$

For example, economists often assume that GDP has import elasticity of 1. Indeed,

$$\log(GDP) = \beta_0 + \beta_1 \log(IMPORTS)$$

Coefficients:

(Intercept) log(IMPORTS)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.8915</td>
</tr>
<tr>
<td>log(IMPORTS)</td>
<td>0.9693</td>
</tr>
</tbody>
</table>
Price Elasticity

In economics, the slope coefficient $\beta_1$ in the regression
\[ \log(sales) = \beta_0 + \beta_1 \log(price) + \varepsilon \]
is called price elasticity.

This is the % change in sales per 1% change in price.

The model implies that $E[sales] = A \cdot price^{\beta_1}
where $A = \exp(\beta_0)$.
Price Elasticity of OJ

A chain of gas station convenience stores was interested in the dependency between price of and Sales for orange juice...

They decided to run an experiment and change prices randomly at different locations. With the data in hands, let’s first run an regression of Sales on Price:

\[
Sales = \beta_0 + \beta_1 Price + \epsilon
\]

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.719</td>
</tr>
<tr>
<td>R Square</td>
<td>0.517</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.507</td>
</tr>
<tr>
<td>Standard Error</td>
<td>20.112</td>
</tr>
<tr>
<td>Observations</td>
<td>50.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>1.000</td>
</tr>
<tr>
<td>SS</td>
<td>20803.071</td>
</tr>
<tr>
<td>MS</td>
<td>20803.071</td>
</tr>
<tr>
<td>F</td>
<td>51.428</td>
</tr>
<tr>
<td>Significance F</td>
<td>0.000</td>
</tr>
<tr>
<td>Regression</td>
<td>1.000</td>
</tr>
<tr>
<td>Residual</td>
<td>48.000</td>
</tr>
<tr>
<td>Total</td>
<td>49.000</td>
</tr>
<tr>
<td>SS</td>
<td>40219.520</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>89.642</td>
<td>8.610</td>
<td>10.411</td>
<td>72.330</td>
<td>106.955</td>
</tr>
</tbody>
</table>
Price Elasticity of OJ

No good!!
Price Elasticity of OJ

But... would you really think this relationship would be linear? Moving a price from $1 to $2 is the same as changing it form $10 to $11?? We should probably be thinking about the price elasticity of OJ...

\[
\log(Sales) = \gamma_0 + \gamma_1 \log(Price) + \epsilon
\]

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.812</td>
<td>0.148</td>
<td>32.504</td>
<td>0.000</td>
<td>4.514</td>
</tr>
<tr>
<td>LogPrice</td>
<td>-1.752</td>
<td>0.144</td>
<td>-12.173</td>
<td>0.000</td>
<td>-2.042</td>
</tr>
</tbody>
</table>

How do we interpret \( \hat{\gamma}_1 = -1.75 \)?

(When prices go up 1%, sales go down by 1.75%)
Price Elasticity of OJ

Fitted Model

Residual Plot

Much better!!
Making Predictions

What if the gas station store wants to predict their sales of OJ if they decide to price it at $1.8?

The predicted $\log(Sales) = 4.812 + (-1.752) \times \log(1.8) = 3.78$

So, the predicted $Sales = \exp(3.78) = 43.82$.

How about the plug-in prediction interval?

In the log scale, our predicted interval in $[\log(Sales) - 2s; \log(Sales) + 2s] =$

$[3.78 - 2(0.38); 3.78 + 2(0.38)] = [3.02; 4.54]$.

In terms of actual $Sales$ the interval is $[\exp(3.02), \exp(4.54)] = [20.5; 93.7]$
Making Predictions

In the log scale (right) we have $[\hat{Y} - 2s; \hat{Y} + 2s]$

In the original scale (left) we have $[\exp(\hat{Y}) \times \exp(-2s); \exp(\hat{Y}) \exp(2s)]$
Some additional comments...

- Another useful transformation to deal with non-constant variance is to take only the \( \log(Y) \) and keep \( X \) the same. Clearly the “elasticity” interpretation no longer holds.
- Always be careful in interpreting the models after a transformation
- Also, be careful in using the transformed model to make predictions
Summary of Transformations

Coming up with a good regression model is usually an iterative procedure. Use plots of residuals vs $X$ or $\hat{Y}$ to determine the next step.

Log transform is your best friend when dealing with non-constant variance ($\log(X)$, $\log(Y)$, or both).

Add polynomial terms (e.g. $X^2$) to get nonlinear regression.

The bottom line: you should combine what the plots and the regression output are telling you with your common sense and knowledge about the problem. Keep playing around with it until you get something that makes sense and has nothing obviously wrong with it.
Outliers

Body weight vs. brain weight...

\[ X = \text{body weight of a mammal in kilograms} \]

\[ Y = \text{brain weight of a mammal in grams} \]

Do additive errors make sense here??

Also, what are the standardized residuals plotted above?
Standardized Residuals

In our model $\epsilon \sim N(0, \sigma^2)$

The residuals $e$ are a proxy for $\epsilon$ and the standard error $s$ is an estimate for $\sigma$

Call $z = e/s$, the standardized residuals... We should expect

$$z \approx N(0, 1)$$

(How often should we see an observation of $|z| > 3$?)
Outliers

Let’s try logs...

Great, a lot better!! But we see a large and positive potential outlier... the Chinchilla!
Outliers

It turns out that the data had the brain of a Chinchilla weighing 64 grams!! In reality, it is 6.4 grams... after correcting it:
How to Deal with Outliers

When should you delete outliers?

Only when you have a really good reason!

There is nothing wrong with running regression with and without potential outliers to see whether results are significantly impacted.

Any time outliers are dropped the reasons for removing observations should be clearly noted.
Other non-Linear Models

- We can always try to add more polynomial terms...
Other non-Linear Models

- We can always try to add more polynomial terms...
Other non-Linear Models

- We can always try to add more polynomial terms...
Other non-Linear Models

- We can always try to add more polynomial terms...
Other non-Linear Models

- Or use a different set of "Basis Function"

\[ Y_i = \beta_0 + \beta_1 b_1(X_i) + \beta_2 b_2(X_i) + \cdots + \beta_p b_p(X_i) + \epsilon_i \]

- Notice that this has a form of a linear model and can, therefore, be estimated via LS.

- There are a million different choices of bases... we’ll focus on the most popular one: Splines
Regression Splines

- A spline works by fitting different low dimensional polynomials over different regions of the predictor space.
- The points where the coefficients change are called knots.
- The more knots you use, the more flexibility you have (but also more complexity!)
Regression Splines
Regression Splines
Smoothing Splines

- This approach tries to battle complexity by imposing “smoothness” constraints
- It looks for a function $g$ that minimizes

$$\sum_{i=1}^{N} (Y_i - g(X_i))^2 + \lambda \int g''(t)^2 dt$$

- $\lambda$ is the tuning parameter (greater than 0)... equivalent to choosing the number of knots
- $\lambda = \infty$ leads to linear regression... larger $\lambda$ leads to complex (wiggly) functions
Smoothing Splines

![Graph showing a scatter plot with a smoothing spline.]
Smoothing Splines
Smoothing Splines

![Graph showing different smoothing splines with varying sparsities](image-url)
Choosing the Model

How do we evaluate a forecasting model? Make predictions!

Basic Idea: We want to use the model to forecast outcomes for observations we have not seen before.

▶ Use the data to create a prediction problem.
▶ See how our candidate models perform.

We’ll use most of the data for training the model, and the left over part for validating the model.
Cross-Validation

In a cross-validation scheme, you fit a bunch of models to most of the data (training sample) and choose the model that performed the best on the rest (left-out sample).

- Use the model to obtain fitted $\hat{Y}_j = x'_j b$ values for all of the $N_{LO}$ left-out data points.
- Calculate the Mean Square Error for these predictions.

$$MSE = \frac{1}{N_{LO}} \sum_{j=1}^{N_{LO}} (Y_j - \hat{Y}_j)^2$$
Generalized Additive Models (GAMs)

- We use “Gams” when we have more than one predictor variable.
- It tries to find a different non-linear function to connect each $X$ to $Y$... and adds them all together (additive model).

$$Y_i = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_p(X_{ip}) + \epsilon$$
Generalized Additive Models (GAMs)
Generalized Additive Models (GAMs)
Generalized Additive Models (GAMs)
Out-of-Sample Comparison

- MSE for a randomly chosen 100 observations...

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{Y}$</td>
<td>83.78 (13.70)</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>29.59 (6.16)</td>
</tr>
<tr>
<td>GAM</td>
<td>18.94 (5.88)</td>
</tr>
</tbody>
</table>
Model Building Process

When building a regression model remember that simplicity is your friend... smaller models are easier to interpret and have fewer unknown parameters to be estimated.

Keep in mind that every additional parameter represents a cost!!

The first step of every model building exercise is the selection of the universe of variables to be potentially used. This task is entirely solved through you experience and context specific knowledge...

- Think carefully about the problem
- Consult subject matter research and experts
- Avoid the mistake of selecting too many variables
Model Building Process

With a universe of variables in hand, the goal now is to select the model. Why not include all the variables in?

Big models tend to over-fit and find features that are specific to the data in hand... ie, not generalizable relationships. The results are bad predictions and bad science!

In addition, bigger models have more parameters and potentially more uncertainty about everything we are trying to learn... (check the beer and weight example!)

We need a strategy to build a model in ways that accounts for the trade-off between fitting the data and the uncertainty associated with the model
Out-of-Sample Prediction

One idea is to focus on the model’s ability to predict... How do we evaluate a forecasting model? Make predictions!

**Basic Idea:** We want to use the model to forecast outcomes for observations we have not seen before.

- Use the data to create a prediction problem.
- See how our candidate models perform.

We’ll use most of the data for *training* the model, and the left over part for *validating* the model.
Out-of-Sample Prediction

In a cross-validation scheme, you fit a bunch of models to most of the data (training sample) and choose the model that performed the best on the rest (left-out sample).

- Fit the model on the training data
- Use the model to predict $\hat{Y}_j$ values for all of the $N_{LO}$ left-out data points
- Calculate the Mean Square Error for these predictions

$$MSE = \frac{1}{N_{LO}} \sum_{j=1}^{N_{LO}} (Y_j - \hat{Y}_j)^2$$
Example

To illustrate the potential problems of “over-fitting” the data, let’s look again at the Telemarketing example... let’s look at multiple polynomial terms...
Example

Let’s evaluate the fit of each model by their $R^2$
(on the training data)
Example

How about the MSE?? (on the left-out data)
Another way to evaluate a model is to use Information Criteria metrics which attempt to quantify how well our model would have predicted the data (regardless of what you’ve estimated for the $\beta_j$’s).

A good alternative is the BIC: Bayes Information Criterion, which is based on a “Bayesian” philosophy of statistics.

$$BIC = n \log(s^2) + p \log(n)$$

You want to choose the model that leads to minimum BIC.
One (very!) nice thing about the BIC is that you can interpret it in terms of model probabilities.

Given a list of possible models \( \{M_1, M_2, \ldots, M_R\} \), the probability that model \( i \) is correct is

\[
P(M_i) \approx \frac{e^{-\frac{1}{2} \text{BIC}(M_i)}}{\sum_{r=1}^{R} e^{-\frac{1}{2} \text{BIC}(M_r)}} = \frac{e^{-\frac{1}{2} [\text{BIC}(M_i) - \text{BIC}_{\text{min}}]}}{\sum_{r=1}^{R} e^{-\frac{1}{2} [\text{BIC}(M_r) - \text{BIC}_{\text{min}}]}}
\]

(Subtract \( \text{BIC}_{\text{min}} = \min\{\text{BIC}(M_1) \ldots \text{BIC}(M_R)\} \) for numerical stability.)
BIC for Model Selection

Thus BIC is an alternative to testing for comparing models.

- It is easy to calculate.
- You are able to evaluate model probabilities.
- There are no “multiple testing” type worries.
- It generally leads to more simple models than $F$-tests.

As with testing, you need to narrow down your options before comparing models. **What if there are too many possibilities?**
Stepwise Regression

One computational approach to build a regression model step-by-step is “stepwise regression.” There are 3 options:

- **Forward**: adds one variable at the time until no remaining variable makes a significant contribution (or meet a certain criteria... could be out of sample prediction)

- **Backwards**: starts will all possible variables and removes one at the time until further deletions would do more harm them good

- **Stepwise**: just like the forward procedure but allows for deletions at each step
LASSO

The LASSO is a shrinkage method that performs automatic selection. Yet another alternative... has similar properties as stepwise regression but it is more automatic... R does it for you!

The LASSO solves the following problem:

$$\arg \min_\beta \left\{ \sum_{i=1}^{N} (Y_i - X_i' \beta)^2 + \lambda |\beta| \right\}$$

- Coefficients can be set exactly to zero (automatic model selection)
- Very efficient computational method
- $\lambda$ is often chosen via CV
One informal but very useful idea to put it all together...

I like to build models from the bottom, up...

- Set aside a set of points to be your validating set (if dataset large enough)
- Working on the training data, add one variable at the time deciding which one to add based on some criteria:
  1. larger increases in $R^2$ while significant
  2. larger reduction in MSE while significant
  3. BIC, etc...
- at every step, carefully analyze the output and check the residuals!
- Stop when no additional variable produces a “significant” improvement
- Always make sure you understand what the model is doing in the specific context of your problem