REGRESSION

- We have now a way to determine whether there is association between variables (remember, correlation ≠ causation).
- However, in more general settings we are interested in prediction: given values of the predictor, we want to know the outcome.
REGRESSION

- Researchers have studies bears by anesthetizing them in order to obtain vital measurements, such as age, gender, length, and weight. Because most bears are quite heavy and difficult to lift, researchers and hunters experience considerable difficulty actually weighing a bear in the wild. Can we determine the weight of a bear from other measurements that are easier to obtain (for example, length)?
**Simple Linear Regression**

- We will concentrate in linear regression. Therefore, we think that the true relationship is
  \[ y = \beta_0 + \beta_1 x + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \]
- Given pairs \((x, y)\), we want to get estimate
  \[ \hat{y} = b_0 + b_1 x \]

- \(\hat{y}\) Predicted value
- \(b_0\) Estimate of intercept
- \(b_1\) Estimate of slope
\[ Zx: z = 53^2 + (67.5)^2 + 72^2 + \ldots + 37^2 \]
\[ z = 80 \times 34 + 67.5 \times 34 + \ldots + 37 \times 34 \]
\[ = 151829 \]

We need to use the formula for \( b_2 \)

\[ b_2 = \frac{\sum zx}{\sum z} \]

\[ b_2 = \frac{151829}{315} \]

Given the following lengths and weights of male bears, what is the estimate of the regression line:

<table>
<thead>
<tr>
<th>Length</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.5</td>
<td>99.9</td>
</tr>
<tr>
<td>73.6</td>
<td>101.2</td>
</tr>
<tr>
<td>80.0</td>
<td>114.8</td>
</tr>
<tr>
<td>83.0</td>
<td>127.8</td>
</tr>
</tbody>
</table>

Given the following lengths and weights of male bears, what is the estimate of the regression line:

<table>
<thead>
<tr>
<th>Length</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.5</td>
<td>99.9</td>
</tr>
<tr>
<td>73.6</td>
<td>101.2</td>
</tr>
<tr>
<td>80.0</td>
<td>114.8</td>
</tr>
<tr>
<td>83.0</td>
<td>127.8</td>
</tr>
</tbody>
</table>
\[ b_1 = \frac{8 \times 151879 - 516.5 \times 2176}{8 \times 34525.75 - (516.5)^2} \]

\[ b_1 = 9.66 \]

\[ b_0 = \frac{2176}{8} - 9.66 \frac{516.5}{8} = -351.66 \]
**Simple Linear Regression**

- Computing regression estimates:

\[
b_1 = \frac{n \sum y_i x_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}
\]

\[
b_0 = \bar{y} - b_1 \bar{x}
\]

- These estimates provide the line that “best fits” the data, in the sense that the vertical distance between the points and the line is minimized.
Simple linear regression

- The slope $b_1$ represents the marginal change in the outcome for each unit of change in the predictor.
- Residuals are the difference between the actually observed values and the ones predicted by the regression line.

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1x_i)$$

- If a linear regression is a reasonable model, then the residuals should have no pattern.

This is what I called di on 297!!
PREDICTION

○ Once the regression line has been obtained, it can be used to get point predictions for new values by simply plugging in the value of the new predictors.
○ If there is no significant correlation, do not use the regression to make predictions.
○ Do not extrapolate!!! Keep within the scope of the data you used to fit the model.
○ Do not make predictions about a population that is different from that where the original data was collected.
○ Regressions can change in time!!!
PREDICTION

- Good predictions include prediction intervals in addition to point predictions.
- To get prediction intervals, we need to estimate the variance $\sigma^2$.
- We do that based on the residuals:

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$
**Another Look at the Correlation Coefficient**

- A useful way to think about prediction is as splitting the total variation in the observation into an explain component (corresponding to the regression) and an unexplained one (corresponding to the residuals).

\[
\text{Total variation} = \text{Explained variation} + \text{Unexplained variation}
\]

\[
\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2
\]
ANOTHER LOOK AT THE CORRELATION COEFFICIENT

- The coefficient of determination is the amount of the variation in $y$ that is explained by the regression line.
- In simple linear regressions (one regressor), it can be computed by squaring the correlation coefficient $r$.

$$r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$
PREDICTION INTERVAL FOR A NEW OBSERVATION

- Given the new predictor \( x_0 \), a prediction interval for the outcome is given by

\[
\hat{y}_0 - E < y_0 < \hat{y}_0 + E
\]

where

\[
\hat{y}_0 = b_0 + b_1 x_0
\]

\[
E = t_{1-\alpha/2} \sqrt{\frac{s^2}{n} \left( 1 + \frac{n(x_0 - \bar{x})^2}{\sum x_i^2 - (\sum x_i)^2} \right)}
\]
QUESTION

- In the bear weight example, what would be the point prediction for the weight of a bear with length of 71.0 in?
  
  a) 327
  
  b) 361
  
  c) 334
  
  d) 390

<table>
<thead>
<tr>
<th>Length</th>
<th>53.0</th>
<th>67.5</th>
<th>72.0</th>
<th>72.0</th>
<th>73.0</th>
<th>73.0</th>
<th>69.5</th>
<th>73.0</th>
<th>37.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>80</td>
<td>344</td>
<td>416</td>
<td>348</td>
<td>252</td>
<td>360</td>
<td>332</td>
<td>34</td>
<td></td>
</tr>
</tbody>
</table>

\[
b_0 = -351.66 \quad x_0 = 71.0
\]

\[
b_1 = 9.66
\]

\[
\hat{y}_0 = -351.66 + 9.66 \times 71 = 334.2
\]
**QUESTION**

- In the bear weight example, what would be the 95% prediction interval for the weight of a bear with length of 71.0 in?
  - a) [267, 401]
  - b) [158, 510]
  - c) [240, 428]
  - d) [133, 535]

<table>
<thead>
<tr>
<th>Length</th>
<th>53.0</th>
<th>57.5</th>
<th>62.0</th>
<th>72.0</th>
<th>73.0</th>
<th>68.5</th>
<th>73.0</th>
<th>37.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>80</td>
<td>344</td>
<td>416</td>
<td>348</td>
<td>262</td>
<td>360</td>
<td>332</td>
<td>34</td>
</tr>
</tbody>
</table>

\[
\hat{Y}_0 = 334.2 \quad n = 8
\]

\[
t_{1-\alpha/2}(6) = 2.446
\]

\[
s = \sqrt{\frac{80}{6} \left( \frac{344 - 160.32}{2} + \frac{344 - 300.39}{2} \right) + \ldots}
\]

\[
n = 66.6
\]

\[
\hat{Y}_1 = -351.66 + 9.66 \times 53 = 160.32
\]

\[
\hat{Y}_2 = -351.66 + 9.66 \times 67.5 = 300.39
\]

\[
\hat{Y}_3 = -351.66 + 9.66 \times 72 = 343.86
\]

\[
\hat{Y}_4 = -351.66 + 9.66 \times 72 = 343.86
\]
We still need to compute

\[
\frac{n(x_0 - \bar{x})^2}{n \sum x_i^2 - (\sum x_i)^2} = \frac{8(71.0 - 64.56)^2}{8 \times 34525.75 - (516.5)^2}
\]

= 0.0351

\[
E = 7.446 \times 66.6 \times \sqrt{1 + \frac{1}{8} + 0.0351}
\]

= 175.46

\[
\hat{\pi}(\hat{y}_0) = [334.2 - 175.46, 334.2 + 175.46]
\]

= [158, 510]
MULTIPLE LINEAR REGRESSION

- A multiple regression expresses a linear association between a dependent variable \( y \) and two or more independent variables \( (x_1, x_2, \ldots, x_k) \).
- In a multiple linear regression, we assume that the relationship among variables is:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)
\]
MULTIPLE LINEAR REGRESSION

- The estimated regression is
  \[ y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \]

- Formulas to compute the estimates are beyond the scope of this class, but we can use computer software to do the calculations.

- The estimate of the variance is
  \[ s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - k - 1} \]
Iris setosa. This is a famous dataset by Fisher.
EXAMPLE

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 2.30374  | 0.36529    | 6.312   | < 2e-16  |
| Sepal_width| 0.66742  | 0.05036    | 13.266  | < 2e-16  |
| Petal_length| 0.28342 | 0.19722    | 1.437   | 0.157    |

Signif. codes:  < 0.001 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2359 on 47 degrees of freedom
Multiple R-squared: 0.8703, Adjusted R-squared: 0.853
F-statistic: 31.16 on 2 and 47 DF, p-value: 2.40e-09

\[ b_0 = 2.30374 \]
\[ b_1 = 0.66742 \]
\[ b_2 = 0.28342 \]

\[ H_0: \beta_1 = 0 \] vs \[ H_1: \beta_1 \neq 0 \]
EXAMPLE

- Note that we can use the previous table to decide if any given coefficient is significant or not (but we can only conclude about one variable at a time).

\[ H_0: \beta_j = 0 \quad \text{Ha: } \beta_j \neq 0 \]

- The interpretation for \( R^2 \) is the same as before (percentage of explained variability). In multiple regression, it is called the \textit{multiple coefficient of determination}.

- We want \( R^2 \) as close to 1 as possible!!
ADJUSTED $R^2$

- $R^2$ always grows as more variables are included in the model, even if those variables really explain very little.
- The adjusted $R^2$ corrects for the "complexity" of the model. Occam's razor is a concept that shows up frequently.

$$\text{adjusted } R^2 = 1 - \frac{(n - 1)}{n - k - 1}(1 - R^2)$$
TEST OF SIGNIFICANCE OF REGRESSION

- Is there at least one variable that explains the variability in the outcome? In other words, are we gaining anything with the regression? The null hypothesis is that the regression is worthless.
- In this case we reject the null hypothesis, so there is evidence that at least one variable explains the outcome.
Residuals vs. Fits

Good.

Bad

You are missing some variable

= Variability is not constant
QQ PLOTS

- Quantile-quantile plots: quantiles of the empirical distribution vs. the quantiles of the standard normal distribution. If it looks like a straight line, then the empirical distribution is approximately normal.