Classification & Regression

Model Inference
Multiple Regression
Multicollinearity
Last Lecture

\[ \hat{y} = 3.6 + 1.7x_i \]

*Interpret this Equation*

- \( \hat{y} \) = weight
- \( x_i \) = age
So you've fit a model, and ensured each of the assumptions of linear models were met...

\[ \hat{y} = 3.6 + 1.7x_i \]

- \( \hat{y} \) = weight
- \( x_i \) = age

What's Next

We can begin by asking three broad questions:

- Do our predictors add any value? Is so, how “good” is the model?
- Is age significantly related to weight, or does it vary with it by chance?
- How confident that the relation of age and weight has a slope of 1.7
Q1: Do Our Predictors Add Value

\[ \hat{y} = 3.6 + 1.7x_i \]

- \( \hat{y} \) = weight
- \( x_i \) = age

Does the equation capture a significant relation to the target variable?
Test of the Regression

\[ \hat{y} = 3.6 + 1.7x_i \]

**Does the equation capture a significant relation to the target variable**

To assess the overall ability of the set of independent variables to explain differences in the dependent variable within the regression model we again utilize hypothesis testing.
Test of the Regression

\[ \hat{y} = 3.6 + 1.7x_i \]

Does the equation capture a significant relation to the target variable

To assess the overall ability of the set of independent variables to explain differences in the dependent variable within the regression model we again utilize hypothesis testing.

What is the null hypothesis?

\[ \hat{y} = \text{weight} \]
\[ x_i = \text{age} \]
Test of the Regression

• Null hypothesis
  – The independent variables do not provide any added information in determining the value of the target variable
  – Equivalent to saying $\beta_1 = 0$ in $y = \beta_0 + \beta_1 x_i$

• This hypothesis can be tested using a F-Test
  – Known as a “Global Test” of the model
Remember:

\[ y = \beta_0 + \beta_1 x_i \]

Goal: Find the parameters \( \beta_0 \) and \( \beta_1 \), which minimize the SSE

\[ \text{SSE} = \sum_{i=1}^{N} [y_i - f(x_i)]^2 = \sum_{i=1}^{N} [y_i - \beta_0 - \beta_1 x_i]^2 \]

\[ \beta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = r_{xy} \frac{\sigma_y}{\sigma_x} \]
How Does this Become the Null Hypothesis

\[ \hat{y} = 3.6 + 1.7x_i \]

\[ \hat{y} = 5.9 + 0.001x_i \]
How Does this Become the Null Hypothesis

\[ \hat{y} = 3.6 + 1.7x_i \]

\[ \hat{y} = 5.9 + 0.001x_i \]

\[ \beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r_{xy} \frac{\sigma_y}{\sigma_x} \]
Model Summary Statistics

[Diagram showing regression analysis with terms like Actual y, Estimated y, Total Deviation, Unexplained Deviation (Residual), Explained Deviation (Regression), and Mean of y.]
Model Summary Statistics

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$</td>
</tr>
<tr>
<td>ERROR</td>
<td>$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$</td>
</tr>
<tr>
<td>TOTAL</td>
<td>$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$</td>
</tr>
</tbody>
</table>

- **SSR**: “sum of squares regression”
  - Quantifies how far the estimated sloped regression line, is from the horizontal “no relationship line,” i.e. the sample mean

- **SSE**: “sum of squares error”
  - Quantifies how much the data points, vary around the estimated regression line

- **SST**: “sum of squares total”
  - Quantifies how much the data points, vary around their mean

![Diagram showing regression line and summary statistics](image-url)
Global F-Test

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>( SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 )</td>
<td>( p )</td>
<td>( MSR = \frac{SSR}{p} )</td>
<td>( F = \frac{MSR}{MSE} )</td>
</tr>
<tr>
<td>ERROR</td>
<td>( SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 )</td>
<td>( n - p - 1 )</td>
<td>( MSE = \frac{SSE}{n-p-1} )</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>( S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 )</td>
<td>( n - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( p = \) number independent variables, \( n = \) total instances.

\[
F = \frac{MSR}{MSE}
\]
How “good” is the model?
**Goodness of Fit ($R^2$)**

\[ R^2 = \frac{SSR}{SST} = \frac{\sum_i [f(x_i) - \bar{y}]^2}{\sum_i [y - \bar{y}]^2} \]

- Values range from 0 to 1
- Represents the percentage of variation explained by the regression model
  - Interpretation: "$r^2 \times 100$: percent of the variation in $y$ is 'explained by' the variation in predictor $x$."
- Equivalent to the sample correlation squared.
Caution: Goodness of Fit ($R^2$)

$$R^2 = \frac{SSM}{SST} = \frac{\sum_i [f(x_i) - \bar{y}]^2}{\sum_i [y - \bar{y}]^2}$$

- Does not guarantee model has a good fit to the data
  - Must check the assumptions hold first
- Example:
  - Regression line consistently under and over-predicts the data along the curve, which is bias.
  - The Residuals versus Fits plot emphasizes this unwanted pattern.
  - This type of specification bias occurs when your linear model is underspecified.
    - In other words, it is missing significant independent variables, polynomial terms, and interaction terms. (We will cover these over the next few weeks)
Interpret this Equation: Q2

\[ \hat{y} = 3.6 + 1.7x_i \]

Is age significantly related to weight, or does it vary with it by chance?

- \( \hat{y} \) = weight
- \( x_i \) = age
Regression Coefficient Significance

We can utilize the standard t-test to identify significance of a feature

\[
t_{n-2} = \frac{\hat{\beta}_i - c}{\sqrt{SE(\beta_i)}}
\]

The reason why is that we lose two degrees of freedom for estimating the intercept and slope of the line.

\[
SE(\beta_i) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}
\]

\[
MSE = \frac{\sum_{i=1}^{n}(e_i)^2}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-k-1}
\]
Interpret this Equation: Q3

\[ \hat{y} = 3.6 + 1.7x_i \]

Is the 1.7 exact?

- \( \hat{y} \) = weight
- \( x_i \) = age
Digging Deeper – Inference on Coefficients

• Is the 1.7 exact?
  – **No.** Beta values (weights, slopes) are in fact estimates, and with such estimates comes some degree of uncertainty
  – Similar to other point estimates, can generate confidence intervals for the slope of each variable, giving a more representative range for its relation to the target variable
  – Utilizes the t-distribution
    • Note the importance of the linearity assumption! (and large sample sizes 😊)
Digging Deeper – Inference on Coefficients

Just as we did to test the coefficient significance, we can apply t-distribution to estimate a confidence interval for $\beta_i$

$$
\beta_i \pm t_{(1-\frac{\alpha}{2}), df=(n-2)} \times SE(\beta_i)
$$

Again, we lose two degrees of freedom for estimating the intercept and slope of the line.

$$
SE(\beta_i) = \sqrt{\frac{MSE}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
$$

$$
MSE = \frac{\sum_{i=1}^{n} (e_i)^2}{n-k+1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-k+1}
$$
Bonus: Target Value Inference

• As the regression model is predicing (estimating) the value $\hat{y}$, it is also possible to construct a confidence interval for the predicted value as follows:

$$\hat{y} \pm t_{(n-2)} \times \sqrt{MSE} \times \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})}{(x_i - \bar{x})^2}}$$

$$MSE = \frac{\sum_{i=1}^{n}(e_i)^2}{n-k+1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-k+1}$$
Target: Confidence vs. Prediction Interval

- **Confidence Interval**
  - X% confident of the average y for \( x_p \) falls between...

- **Prediction Interval**
  - X% confident that y for an individual with \( x_p \) falls between...
More than 1 Variable (Multiple Regression)
Multiple Regression

- More commonly the modeling of health (and most real-world) data requires the combination of attributes, to accurately predict the target variable.

Follows the general formula where $x_i$ is each independent variable

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + \varepsilon$$
Separation (Plane)

Simple Linear Regression

Multiple Regression (2 features)

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \]
Assessing Fit – Multiple Regression

• Procedurally, very similar to simple linear regression
  – Estimation of parameters in multiple linear regression is done by minimizing the sum of square errors (Ordinary least squares, OLS)

• Same (LINE) assumptions hold as for simple linear regression
  – L – Linear relationship between the dependent variable and each independent variable
  – I – Independence of instances and errors
  – N – Normality of errors
  – E – Equal variance among errors across all X

• Interpretation
  – Regression Coefficients: $\beta_i \equiv$ effect on the dependent variable when increasing the $i^{th}$ independent variable by 1 unit, **holding all other predictors constant**
    • Assume for now all predictors are continuous
Assessing Fit – Multiple Regression (Residuals)

• Residuals should now be plotted against:
  – Target variable ($y$)
  – Each independent variable

• Partial Residual plots can be used to better assess the Linearity assumption
  – Capture relation between of one independent variable on the response variable by taking into account the effects of the other independent variables

• Partial Regression plots
  – Similar to partial residual, but rather than structure is used to identify outliers and high influence / leverage points
Additional Residual Plots

• It is sometimes also helpful to plot the residuals against explanatory variables that are not currently in the model, but which could potentially be included.
  – Any structure in the plot of residuals versus an omitted variable may indicate that incorporation of that variable could improve the model
Assessing Fit – Multiple Regression (Adjusted $R^2$)

$$R^2 = \frac{SSR}{SST} = \frac{\sum_i [f(x_i) - \bar{y}]^2}{\sum_i [y - \bar{y}]^2}$$

- As we add more explanatory variables, $R^2$ increases, so it is typically adjusted as:

$$Adjusted \ R^2 = 1 - \left(\frac{N - 1}{N - d}\right) (1 - R^2),$$

Where $N$ is the number of data points and $d+1$ is the number of parameters of the regression model.
Global F-Test

• The global F-test is exactly the same as in the case of simple linear regression

• Global F-test only evaluates the assumption that at least one variable coefficient is likely non-zero.
  – Does not specify how many, nor which.
  – Inference (significance and CI) on model coefficients can be used to answer these questions, but this is a univariate perspective.

• When looking to assess if a set of coefficients improves a model, we turn to partial F-tests
Partial F-Test

- Begin with the “full” model
  \[ \hat{y} = w_0 + w_i x_i + \varepsilon \]
  - Compute SSE (\( SSE_{full} \))

- Fit the reduced (intercept only) model
  \[ \hat{y} = w_0 + \varepsilon \]
  - Compute SSE (\( SSE_{reduced} \))

\[ F = \frac{SSE_{reduced} - SSE_{full}}{\Delta \text{Number of independent variables}} \]

\[ SSE_{reduced} \text{ can never be smaller than } SSE_{Full}. \]
  - It is always larger than, or possibly the same as.
Partial F-Test

- Begin with the “full” model
  - \( \hat{y} = w_0 + w_i x_i + \epsilon \)
  - Compute SSE (\( SSE_{full} \))

- Fit the reduced (intercept only) model
  - \( \hat{y} = w_0 + \epsilon \)
  - Compute SSE (\( SSE_{reduced} \))

- \( SSE_{Reduced} \) can never be smaller than \( SSE_{Full} \).
  - It is always larger than, or possibly the same as.

\[
F = \frac{SSE_{reduced} - SSE_{full}}{\Delta \text{Number of independent variables}} \div \frac{SSE_{full}}{Df_{full}}
\]
Next Class – Collinearity, Nominal Data, Interactions

<table>
<thead>
<tr>
<th>ADMISSION_TYPE</th>
<th>ADMISSION_LOCATION</th>
<th>DISCHARGE_LOCATION</th>
<th>INSURANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMERGENCY</td>
<td>EMERGENCY ROOM ADMIT</td>
<td>DISC-TRAN CANCER/CHLDRN H</td>
<td>Private</td>
</tr>
<tr>
<td>ELECTIVE</td>
<td>PHYS REFERRAL/NORMAL DELI</td>
<td>HOME HEALTH CARE</td>
<td>Medicare</td>
</tr>
<tr>
<td>EMERGENCY</td>
<td>TRANSFER FROM HOSP/EXTRAM</td>
<td>HOME HEALTH CARE</td>
<td>Medicare</td>
</tr>
<tr>
<td>EMERGENCY</td>
<td>TRANSFER FROM HOSP/EXTRAM</td>
<td>HOME</td>
<td>Private</td>
</tr>
<tr>
<td>EMERGENCY</td>
<td>EMERGENCY ROOM ADMIT</td>
<td>HOME</td>
<td>Private</td>
</tr>
</tbody>
</table>