### Multiple Regression Model Building

- Transformation of Variables
- Collinearity
- Dropping Regressors

### Simple Linear Regression Model Review

The simple linear regression model states that the response variable $y$ and the explanatory variable $x$ have a linear relationship of the form

$$ y = \beta_0 + \beta_1 x + \epsilon $$

where

- $\beta_0$ and $\beta_1$ are the $y$-intercept and the slope of the true population regression line.
- $\epsilon \sim N(0, \sigma)$

- The $\epsilon_i$ corresponding to the pairs $(x_i, y_i)$ are independent of each other
- Given $x$, $y$ has mean $\beta_0 + \beta_1 x$ and variance $\sigma^2$

$$ E(y|x) = \mu_y|x = \beta_0 + \beta_1 x $$

is called the **population regression line**.

$$ \text{Var}(y|x) = \sigma^2_{y|x} = \sigma^2 $$

### Example

The following (hypothetical) data give stopping distances $Y$ (ft) for $n = 62$ trials of various automobiles traveling at speed $X$ (mph). (Ezekiel and Fox, 1959)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>20</td>
<td>48</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>2.4,8.8</td>
<td>21</td>
<td>39,42,55</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>7.7</td>
<td>24</td>
<td>56</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>8.9,11.13</td>
<td>25</td>
<td>33,48,56,59</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>5.5,13</td>
<td>26</td>
<td>39,41</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>8.14,17</td>
<td>27</td>
<td>57,78</td>
<td>27</td>
</tr>
<tr>
<td>12</td>
<td>11.19,21</td>
<td>28</td>
<td>64,84</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>15.18,27</td>
<td>29</td>
<td>54,68</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>14.16</td>
<td>30</td>
<td>60,67,101</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>31</td>
<td>77</td>
<td>31</td>
</tr>
<tr>
<td>16</td>
<td>14.19,34</td>
<td>35</td>
<td>85,107</td>
<td>35</td>
</tr>
<tr>
<td>17</td>
<td>22,29</td>
<td>36</td>
<td>79</td>
<td>36</td>
</tr>
<tr>
<td>18</td>
<td>29.34,47</td>
<td>37</td>
<td>138</td>
<td>37</td>
</tr>
<tr>
<td>19</td>
<td>30</td>
<td>40</td>
<td>110,134</td>
<td>40</td>
</tr>
</tbody>
</table>

![Stopping Distance vs. Speed](image-url)

```r
lm(formula = dist ~ speed)
```

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -20.1644 | 3.3362 | -6.044 | 1.04e-07 *** |
| speed | 3.1488 | 0.1568 | 20.087 < 2e-16 *** |

---

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 12.11 on 60 degrees of freedom
Multiple R-Squared: 0.8705, Adjusted R-squared: 0.8684
F-statistic: 403.5 on 1 and 60 DF, p-value: < 2.2e-16
The curvature indicates that the relationship between the response variable $Y$ and the predictor $X$ is non-linear. There is also an increasing of the variance of the residuals as the fitted values increase.

Both are indications of a departure from the model assumptions. How might we remedy this?

One way is to transform the response variable $Y$.

Maybe the true relationship is exponential. That is,

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

It looks like we overcompensated for the curvature by taking $\log(Y)$. Perhaps we should take $\log(X)$ as well. That is,

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

Two predictors, $X_1$ and $X_2$, are exactly collinear if there is a linear equation such as,

$$c_1 X_1 + c_2 X_2 = c_0$$

For example

- height in feet and meters
- temperature in Celsius and Fahrenheit

It’s rare that a regression would include two exactly collinear predictors, but there may be approximately collinear predictors, that is, predictors whose correlation is close to one in absolute value.

For more than two predictors, the set $X_1, X_2, \ldots, X_p$ is collinear if, for some constants $c_0, c_1, \ldots, c_p$

$$c_1 X_1 + c_2 X_2 + \cdots + c_p X_p = c_0$$

In the case of collinearity, the estimating procedure is very unstable; it becomes very sensitive to random errors. Collinear predictors will typically lead to large variances for estimated coefficients.
Example

Suppose in the stopping distance example, we create a second predictor

\[ W = -X + \delta \]

where \( \delta \sim N(0,1) \).

The exact correlation between \( X \) and \( W \) depends on the realization of \( \delta \), but for one simulation it came out to \( r^2 = 0.9959765 \).

What happens if we fit the model

\[ Y = \beta_0 + \beta_1 X + \beta_2 W + \epsilon \]

\[ \text{lm(formula = dist \sim speed} + W) \]

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -20.1252 | 3.3963 | -5.926 | 1.72e-07 *** |
| speed | 3.2919 | 1.7082 | 1.927 | 0.0588 . |
| W | 0.1455 | 1.7288 | 0.084 | 0.9332 |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.21 on 59 degrees of freedom
Multiple R-Squared: 0.8706, Adjusted R-squared: 0.8662
F-statistic: 198.4 on 2 and 59 DF, p-value: < 2.2e-16

Dropping a Regressor

Suppose we regress \( y \) on \( x_1, x_2, \ldots, x_p \). If the \( t \) value for \( x_2 \), say, is small (or equivalently if the CI for \( \beta_2 \) contains 0), do we leave \( x_2 \) out of the regression?

If our prior belief is that the regressor has an effect on the response ...

If our prior belief is that the regressor has little/no effect on the response ...