Week 3: Multiple Linear Regression

Polynomial regression, categorical variables, interactions & main effects, $R^2$

Max H. Farrell
The University of Chicago Booth School of Business
Multiple vs simple linear regression

Fundamental model is the same.

Basic concepts and techniques translate directly from SLR.

- Individual parameter inference and estimation is the same, conditional on the rest of the model.
- We still use \texttt{lm}, \texttt{summary}, \texttt{predict}, etc.

The hardest part would be moving to matrix algebra to translate all of our equations. Luckily, R does all that for you.
Polynomial regression

A nice bridge between SLR and MLR is polynomial regression.

Still only one $X$ variable, but we add powers of $X$:

$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m$$

You can fit any mean function if $m$ is big enough.

- Usually, $m = 2$ does the trick.

This is our first “multiple linear regression”!
Example: telemarketing/call-center data.

- How does length of employment (months) relate to productivity (number of calls placed per day)?

```r
> attach(telemkt <- read.csv("telemarketing.csv"))
> tele1 <- lm(calls~months)
> xgrid <- data.frame(months = 10:30)
> par(mfrow=c(1,2))
> plot(months, calls, pch=20, col=4)
> lines(xgrid$months, predict(tele1, newdata=xgrid))
> plot(months, tele1$residuals, pch=20, col=4)
> abline(h=0, lty=2)
```
It looks like there is a polynomial shape to the residuals.

- We are leaving some predictability on the table
  ... just not linear predictability.
Testing for nonlinearity

To see if you need more nonlinearity, try the regression which includes the next polynomial term, and see if it is significant.

For example, to see if you need a quadratic term,

- fit the model then run the regression
  \[ E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2. \]
  - If your test implies \( \beta_2 \neq 0 \), you need \( X^2 \) in your model.

Note: \( p \)-values are calculated “given the other \( \beta \)’s are nonzero”; i.e., conditional on \( X \) being in the model.
Test for a quadratic term:

```r
> months2 <- months^2
> tele2 <- lm(calls ~ months + months2)
> summary(tele2) # abbreviated output

Coefficients:

|                | Estimate  | Std. Err | t value | Pr(>|t|)  |
|----------------|-----------|----------|---------|-----------|
| (Intercept)    | -0.140471 | 2.32263  | -0.060  | 0.952     |
| months         | 2.310202  | 0.25012  | 9.236   | 4.90e-08  *** |
| months2        | -0.040118 | 0.00633  | -6.335  | 7.47e-06  *** |

The quadratic $months^2$ term has a very significant $t$-value, so a better model is $calls = \beta_0 + \beta_1 months + \beta_2 months^2 + \varepsilon$. 
Everything looks much better with the quadratic mean model.

```r
> xgrid <- data.frame(months=10:30, months2=(10:30)^2)
> par(mfrow=c(1,2))
> plot(months, calls, pch=20, col=4)
> lines(xgrid$months, predict(tele2, newdata=xgrid))
> plot(months, tele2$residuals, pch=20, col=4)
> abline(h=0, lty=2)
```
A few words of caution

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

- If you add a higher order term, the lower order term is kept *regardless* of its individual \( t \)-stat.
  
  (see handout on website)

Be very careful about predicting outside the data range;

- the curve may do unintended things beyond the data.

Watch out for over-fitting.

- You can get a “perfect” fit with enough polynomial terms,
- but that doesn’t mean it will be any good for prediction or understanding.
Beyond SLR

Many problems involve more than one independent variable or factor which affects the dependent or response variable.

- Multi-factor asset pricing models (beyond CAPM).
- Demand for a product given prices of competing brands, advertising, household attributes, etc.
- More than size to predict house price!

In SLR, the conditional mean of $Y$ depends on $X$. The multiple linear regression (MLR) model extends this idea to include more than one independent variable.
The MLR Model

The MLR model is same as always, but with more covariates.

\[ Y|X_1, \ldots, X_d \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d, \sigma^2) \]

Recall the key assumptions of our linear regression model:

(i) The conditional mean of \( Y \) is linear in the \( X_j \) variables.
(ii) The additive errors (deviations from line)
   - are Normally distributed
   - independent from each other and all the \( X_j \)
   - identically distributed (i.e., they have constant variance)
Our interpretation of regression coefficients can be extended from the simple single covariate regression case:

$$\beta_j = \frac{\partial \mathbb{E}[Y|X_1, \ldots, X_d]}{\partial X_j}$$

- Holding all other variables constant, $\beta_j$ is the average change in $Y$ per unit change in $X_j$.

∂ is from calculus and means “change in”
If $d = 2$, we can plot the regression surface in 3D.

Consider sales of a product as predicted by price of this product ($P_1$) and the price of a competing product ($P_2$).

$\text{Sales} = 1 - 1.0P_1 + 1.1P_2$

- hold $P_2$ fixed and vary $P_1$
- hold $P_1$ fixed and vary $P_2$

Everything measured on log scale (next week)
How do we estimate the MLR model parameters?

The principle of least squares is unchanged; define:

- fitted values \( \hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_d X_{di} \)
- residuals \( e_i = Y_i - \hat{Y}_i \)
- standard error \( s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p}} \), where \( p = d + 1 \).

Then find the best fitting plane, i.e., coefs \( b_0, b_1, b_2, \ldots, b_d \), by minimizing the sum of squared residuals, \( s^2 \).
Obtaining these estimates in R is very easy:

```r
> salesdata <- read.csv("sales.csv")
> attach(salesdata)
> salesMLR <- lm(Sales ~ P1 + P2)
> salesMLR

Call:
lm(formula = Sales ~ P1 + P2)

Coefficients:
(Intercept)         P1         P2
   1.003     -1.006      1.098
```
Forecasting in MLR

Prediction follows exactly the same methodology as in SLR.

For new data $\mathbf{x}_f = [X_{1,f} \cdots X_{d,f}]'$,

- $\hat{Y}_f = b_0 + b_1 X_{1f} + \cdots + b_d X_{df}$
- $\text{var} [Y_f | \mathbf{x}_f] = \text{var}(\hat{Y}_f) + \text{var}(\varepsilon_f) = s^2_{\text{fit}} + s^2 = s^2_{\text{pred}}$.
- $(1 - \alpha)$ level prediction interval is still $\hat{Y}_f \pm z_{\alpha/2} s_{\text{pred}}$. 
The syntax in R is also exactly the same as before:

```r
> predict(salesMLR, data.frame(P1=1, P2=1),
+   interval="prediction", level=0.95)

  fit     lwr     upr
1 1.094661 1.064015 1.125306
```

```r
> predict(salesMLR, data.frame(P1=1, P2=1),
+   se.fit=TRUE)$se.fit

[1] 0.005227347
```
Residuals in MLR

As in the SLR model, the residuals in multiple regression are purged of any relationship to the independent variables.

We decompose $Y$ into the part predicted by $X$ and the part due to idiosyncratic error.

$$Y = \hat{Y} + e$$

$$\text{corr}(X_j, e) = 0 \quad \text{corr}(\hat{Y}, e) = 0$$
Inference for coefficients

As before in SLR, the LS linear coefficients are random (different for each sample) and correlated with each other.

The LS estimators are unbiased:

$$\mathbb{E}[b_j] = \beta_j \quad \text{for } j = 0, \ldots, d.$$ 

In particular, the sampling distribution for $b$ is a multivariate normal, with mean $\beta = [\beta_0 \cdots \beta_d]'$ and covariance matrix $S_b$.

$$b \sim \mathcal{N}_p(\beta, S_b)$$
Coefficient covariance matrix

$$S_b = \text{var}(b):$$ the $p \times p$ covariance matrix for random vector $b$

$$S_b = \begin{bmatrix}
\text{var}(b_0) & \text{cov}(b_0, b_1) \\
\text{cov}(b_1, b_0) & \text{var}(b_1) \\
& \ddots \\
& & \text{var}(b_{d-1}) & \text{cov}(b_{d-1}, b_d) \\
& & \text{cov}(b_d, b_{d-1}) & \text{var}(b_d)
\end{bmatrix}$$

- Variance decreases with $n$ and $\text{var}(X)$; increases with $s^2$.

⇒ Standard errors are the square root of the diagonal of $S_b$. 
## Standard errors

Conveniently, R’s `summary` gives you all the standard errors.

(or do it manually, see `week3-Rcode.R`)

```r
> summary(salesMLR) ## abbreviated output

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1.002688 | 0.007445   | 134.7   | <2e-16   *** |
| P1             | -1.005900| 0.009385   | -107.2  | <2e-16   *** |
| P2             | 1.097872 | 0.006425   | 170.9   | <2e-16   *** |

Residual standard error: 0.01453 on 97 degrees of freedom
Multiple R-squared:  0.998, Adjusted R-squared:  0.9979
F-statistic: 2.392e+04 on 2 and 97 DF,  p-value: < 2.2e-16
Inference for individual coefficients

Intervals and \( t \)-statistics are exactly the same as in SLR.

\>
\>
A \((1 - \alpha)100\%\) C.I. for \( \beta_j \) is \( b_j \pm z_{\alpha/2} s_{b_j} \).

\>
\>
\( z_{b_j} = (b_j - \beta_j^0) / s_{b_j} \sim \mathcal{N}(0, 1) \) is number of standard errors between the LS estimate and the null value.

Intervals/testing via \( b_j \) & \( s_{b_j} \) are one-at-a-time procedures:

\>
\>
You are evaluating the \( j^{th} \) coefficient conditional on the other \( X \)'s being in the model, but regardless of the values you’ve estimated for the other \( b \)'s.
Categorical effects/dummy variables

To represent qualitative factors in multiple regression, we use dummy, binary, or indicator variables.

- temporal effects (1 if Holiday season, 0 if not)
- spatial (1 if in Midwest, 0 if not)

If a factor $X$ takes $R$ possible levels, we use $R - 1$ dummies

- Allow the intercept to shift by taking on the value 0 or 1
- $1_{[X=r]} = 1$ if $X = r$, 0 if $X \neq r$.

$$E[Y|X] = \beta_0 + \beta_1 1_{[X=2]} + \beta_2 1_{[X=3]} + \cdots + \beta_{R-1} 1_{[X=R]}$$

What is $E[Y|X = 1]$?
Example: back to the pickup truck data.

Does price vary by make?

> attach(pickup <- read.csv("pickup.csv"))
> c(mean(price[make=="Dodge"]),
    mean(price[make=="Ford"]),
    mean(price[make=="GMC"]))

[1] 6554.200 8867.917 7996.208

▶ GMC seems lower on average, but lots of overlap.
▶ Not much of a pattern.
Now fit with linear regression:

$$E[\text{price}|\text{make}] = \beta_0 + \beta_1 \mathbb{1}_{\text{make=Ford}} + \beta_2 \mathbb{1}_{\text{make=GMC}}$$

Easy in R (if \text{make} is a \textit{factor} variable)

\begin{verbatim}
> summary(trucklm1 <- lm(price ~ make, data=pickup))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(In)\text{tercept} 6554      1787   3.667  0.000671 ***
\text{make}\text{Ford}  2314      2420   0.956  0.344386
\text{make}\text{GMC}  1442      2127   0.678  0.501502

The coefficient values correspond to our dummy variables.

What are the \textit{p}-values?
What if you also want to include mileage?

- No problem.

```r
> pickup$miles <- pickup$miles/10000
> trucklm2 <- lm(price ~ make + miles, data=pickup)
> summary(trucklm2)
```

Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|)   |
|---------------------|----------|------------|---------|------------|
| (Intercept)         | 12761.8  | 1746.6     | 7.307   | 5.31e-09 *** |
| makeFord            | 2185.7   | 1842.9     | 1.186   | 0.242      |
| makeGMC             | 2298.8   | 1627.0     | 1.413   | 0.165      |
| miles               | -654.1   | 115.3      | -5.671  | 1.18e-06 *** |

All three brands expect to lose $\$654$ per 10k miles.
Different intercepts, same slope!

```r
> plot(miles, price, pch=20, col=make,
    xlab="miles (10k)", ylab="price ($)"
> abline(a=coef(trucklm2)[1],b=coef(trucklm2)[4],col=1)
> abline(a=(coef(trucklm2)[1]+coef(trucklm2)[2]),
    b=coef(trucklm2)[4],col=2)
```

▶ Dodge trucks affect all slopes!
Variable interaction

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

We can extend this notion and include interaction effects through multiplicative terms.

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

\[ \frac{\partial \mathbb{E}[Y|X_1, X_2]}{\partial X_1} = \beta_1 + \beta_3 X_2 \]
Interactions with dummy variables

Dummy variables separate out categories
  ▶ Different intercept for each category

Interactions with dummies separate out trends
  ▶ Different slope for each category

\[ Y_i = \beta_0 + \beta_1 \mathbb{1}_{\{X_{1i}=1\}} + \beta_2 X_{2i} + \beta_3 (\mathbb{1}_{\{X_{1i}=1\}} X_{2i}) + \cdots + \varepsilon_i \]

\[ \frac{\partial \mathbb{E}[Y|X_1 = 0, X_2]}{\partial X_2} = \beta_2 \]
\[ \frac{\partial \mathbb{E}[Y|X_1 = 1, X_2]}{\partial X_2} = \beta_2 + \beta_3 \]
Same slope, different intercept

▶ Price difference does not depend on mileage!

> trucklm2 <- lm(price ~ make + mile, data=pickup)

▶ Dodge trucks affect all slopes!
Now add individual slopes!

- Price difference *varies* with miles!

> `trucklm3 <- lm(price ~ make*miles, data=pickup)`

- Dodge doesn’t effect Ford, GMC
  
  $b_0, b_1$
What do the numbers show?

```r
> summary(trucklm3)

     Estimate  Std. Error   t value Pr(>|t|)  
(Intercept)  8862.1987    2508.048    3.532  0.00122 **  
makeFord     5216.0507    3707.966    1.408  0.16692    
makeGMC      8360.7481    3080.603    2.727  0.01004 **  
miles        -243.1789    225.3194   -1.081  0.28721    
makeFord:miles-317.0086    346.9587   -0.915  0.36594    
makeGMC:miles-611.1195    267.5601   -2.295  0.02623 *  
```

```r
> c(coef(trucklm3)[1], coef(trucklm3)[4]) # (b_0, b_1) Dodge
  (Intercept) miles
     8862.1987   -243.1789
```

```r
> c(coef(trucklm3)[1]+coef(trucklm3)[2], coef(trucklm3)[4]+coef(trucklm3)[5]) # (b_0, b_1) Ford
  (Intercept) miles
    14078.6715  -560.5871
```
What do the numbers show?

```r
> summary(trucklm3)

  Estimate  Std. Error t value  Pr(>|t|)  
(Intercept)  8862.0 2508.3  3.528  0.001 **
makeFord  5216.0 3707.4  1.404  0.167
makeGMC  8360.0 3080.1  2.716  0.010 **
miles -243.0 225.1 -1.082  0.287
makeFord:miles -317.0 347.5 -0.916  0.366
makeGMC:miles -611.0 268.3 -2.272  0.028 *
```

```r
> price.Ford <- price[make=="Ford"]
> miles.Ford <- miles[make=="Ford"]
> summary(lm(price.Ford ~miles.Ford))

  Estimate  Std. Error t value  Pr(>|t|)  
(Intercept)  14078.8 3094.6  4.549  0.00106 **
miles.Ford  -560.6 299.3 -1.873  0.09054 .
```
Interactions with continuous variables

Example: connection between college & MBA grades. A model to predict Booth GPA from college GPA could be

\[
\text{GPA}^{\text{MBA}} = \beta_0 + \beta_1 \text{GPA}^{\text{Bach}} + \varepsilon.
\]

> grades <- read.csv("grades.csv")
> summary(grades) #output not shown
> attach(grades)
> summary(lm(MBAGPA ~ BachGPA)) ## severly abbrev.

| Estimate (Intercept)         | Std. Error | t value | Pr(>|t|) |
|------------------------------|------------|---------|----------|
| 2.58985                      | 0.31206    | 8.299   | 1.2e-11  *** |
| 0.26269                      | 0.09244    | 2.842   | 0.00607  ** |

- For every 1 point increase in college GPA, your expected GPA at Booth increases by about 0.26 points.
However, this model assumes that the marginal effect of College GPA is the same for any age.

But I’d guess that how you did in college has less effect on your MBA GPA as you get older (farther from college).

We can account for this intuition with an interaction term:

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

Now, the college effect is

\[
\frac{\partial \mathbb{E}[GPA^{\text{MBA}} \mid GPA^{\text{Bach}}, \text{Age}]}{\partial GPA^{\text{Bach}}} = \beta_1 + \beta_2 \text{Age}.
\]

⇒ Depends on Age!
Fitting interactions in R is easy:

\[
\text{lm}(Y \sim X_1 \times X_2) \text{ fits } \mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.
\]

Here, we want the interaction but do not want to include the main effect of age (should age matter individually?).

\[
> \text{summary(lm(MBAGPA} \sim \text{BachGPA}\times\text{Age \sim \text{- Age}}))
\]

Coefficients: \# output abbreviated

|                | Estimate  | Std. Error | t value | Pr(>|t|)   |
|----------------|-----------|------------|---------|------------|
| (Intercept)    | 2.820494  | 0.296928   | 9.499   | 1.23e-13 *** |
| BachGPA        | 0.455750  | 0.103026   | 4.424   | 4.07e-05 *** |
| BachGPA:Age    | -0.009377 | 0.002786   | -3.366  | 0.00132 **  |
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>
Case study in interaction

Use census data to explore the relationship between log wage rate (log(income/hours)) and age—a proxy for experience.

We look at people earning >$5000, working >500 hrs, and <60 years old.
A discrepancy between mean $\log(WR)$ for men and women.

- Female wages flatten at about 30, while men’s keep rising.

```r
> men <- sex=="M"
> malemean <- tapply(log.WR[men], age[men], mean)
> femalemean <- tapply(log.WR[!men], age[!men], mean)
```

![Graph showing the mean log wage rate for men and women across different ages.]
The most simple model has

\[ \mathbb{E}[\log(WR)] = 2 + 0.016 \cdot \text{age}. \]

> wagereg1 <- lm(log.WR ~ age)

You get one line for both men and women.
Add a sex effect with

\[ \mathbb{E}[\log(WR)] = 1.9 + 0.016 \cdot \text{age} + 0.2 \cdot 1_{[\text{sex}=\text{M}]} \].

```r
> wagereg2 <- lm(log.WR ~ age + sex)
```

The male wage line is shifted up from the female line.
With interactions
\[ E[\log(WR)] = 2.1 + 0.011 \cdot \text{age} + (-0.13 + 0.009 \cdot \text{age}) I[\text{sex}=M]. \]

\[
\begin{align*}
> \text{wagereg3} & \leftarrow \text{lm}(\log(WR) \sim \text{age*sex})
\end{align*}
\]

The interaction term gives us different slopes for each sex.
& quadratics ...

\[ E[\log(\text{WR})] = 0.9 + 0.077 \cdot \text{age} - 0.0008 \cdot \text{age}^2 + (-0.13 + 0.009 \cdot \text{age})I[\text{sex}=M]. \]

\[ \text{wagereg4 <- lm(log.WR ~ age*sex + age2)} \]

\[ \text{predicted log wagerate} \]

\[ \text{age}^2 \text{ allows us to capture a nonlinear wage curve.} \]
Finally, add an interaction term on the curvature ($age^2$)

$$E[\log(WR)] = 1 + .07 \cdot age - .0008 \cdot age^2 + (.02 \cdot age - .00015 \cdot age^2 - .34)I_{[\text{sex}=M]}.$$

> `wagereg5 <- lm(log.WR ~ age*sex + age2*sex)`

This full model provides a generally decent looking fit.
We could also consider a model that has an interaction between age and edu.

▶ `reg <- lm(log.WR ~ edu*age)`

Maybe we don’t need the age main effect?

▶ `reg <- lm(log.WR ~ edu*age - age)`

Or perhaps all of the extra edu effects are unnecessary?

▶ `reg <- lm(log.WR ~ edu*age - edu)`
A (bad) goodness of fit measure: $R^2$

How well does the least squares fit explain variation in $Y$?

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

- **Total sum of squares (SST)**
- **Regression sum of squares (SSR)**
- **Error sum of squares (SSE)**

SSR: Variation in $Y$ explained by the regression.
SSE: Variation in $Y$ that is left unexplained.

$$\text{SSR} = \text{SST} \Rightarrow \text{perfect fit.}$$

*Be careful of similar acronyms; e.g. SSR for “residual” SS.*
How does that breakdown look on a scatterplot?

\[(Y_i - \bar{Y})\]
A (bad) goodness of fit measure: $R^2$

The coefficient of determination, denoted by $R^2$, measures goodness-of-fit:

$$R^2 = \frac{SSR}{SST}$$

- SLR or MLR: same formula.
- $R^2 = \text{corr}^2(\hat{Y}, Y) = r_{\hat{y}y}^2$ (\(= r_{xy}^2\) in SLR)
- $0 < R^2 < 1$.
- $R^2$ closer to 1 $\rightarrow$ better fit . . . for these data points
  - No surprise: the higher the sample correlation between $X$ and $Y$, the better you are doing in your regression.
  - So what? What's a “good” $R^2$? For prediction? For understanding?
Adjusted $R^2$

This is the reason some people like to look at adjusted $R^2$

$$R_a^2 = 1 - \frac{s^2}{s_y^2}$$

Since $s^2/s_y^2$ is a ratio of variance estimates, $R_a^2$ will not necessarily increase when new variables are added.

Unfortunately, $R_a^2$ is useless!

- The problem is that there is no theory for inference about $R_a^2$, so we will not be able to tell “how big is big”.
bad $R^2$?
bad model?
bad data?
bad question?
... or just reality?

Pickup regressions:

```r
> summary(trucklm1)$r.square ## make
[1] 0.021
> summary(trucklm2)$r.square ## make + miles
[1] 0.446
> summary(trucklm3)$r.square ## make * miles
[1] 0.511
```

- Is make useless? Is 45% \textit{significantly} better?

  (We’ll formalize in 2 weeks.)
Up next: choosing a regression model

What’s a good regression?

Next week:
  - Problems and diagnostics
  - Some fixes

Week 5:
  - MLR for causation
  - Testing for fit

Week 7
  - BIC, model choice algorithms
  - Big $p$ problems (data mining)
Glossary and equations

\[
\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \quad \text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2, \\
\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]

(Watch out: sometimes SSE is called SSR or RSS!)

\[
R^2 = \frac{\text{SSR}}{\text{SST}} = \text{cor}^2(\hat{Y}, Y) = r_{\hat{y}y}^2
\]

MLR:

- **Model:** \(Y|X_1, \ldots, X_d \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d, \sigma^2)\)
- **Prediction:** \(\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_d X_{di}\)
- **\(b \sim \mathcal{N}_p(\beta, S_b)\)**
- **Interaction:**
  - \(Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i}X_{2i}) + \cdots + \varepsilon\)
  - \(\frac{\partial \mathbb{E}[Y|X_1, X_2]}{\partial X_1} = \beta_1 + \beta_3 X_2\)