Chapter 3: Linear Regression

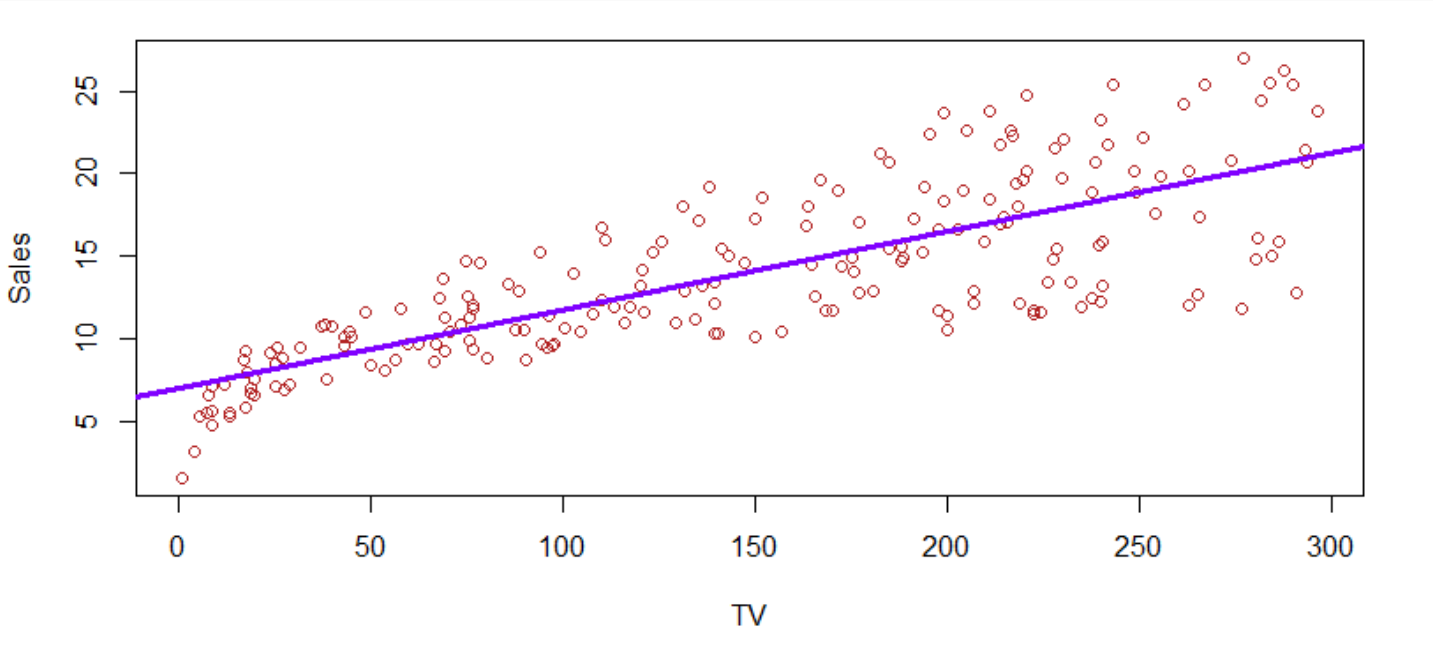
Yonghyun Kwon

Department of Mathematics, Korea Military Academy

Simple Linear Regression

The Advertising Data Set (I)

- Data from 200 regional markets, each with:
  - Ad budgets for **TV** ($1,000s)
  - Product **sales** (1,000s of units)
- Goal: Understand how ad budgets influence sales
  - *Predictors:* TV (X)
  - *Response:* Sales (Y)



Simple Linear Regression

Simple linear regression

Simple linear regression predicts a quantitative response Y from a single predictor X.

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 and β_1 are the model parameters, called *regression coefficients*, and ε is the *error term*.

- Parameters, β_0 and β_1, are estimated from data.
  - The pointe estimates are written as $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Usually, we use X to predict Y.
  - X: *Predictor*(explanatory) variable
  - Y: *Response* variable

Prediction

Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we *predict* the response Y using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$.

Estimating the Coefficients

- The coefficients β_0 and β_1 are unknown.
- Suppose we observe n data points $(x_1, y_1), \ldots, (x_n, y_n)$.
- We fit a line so that the *fitted values* are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{for } i = 1, \ldots, n$$

- *Residuals* are defined as:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- Goal: minimize the *Residual Sum of Squares (RSS)*:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Least Squares Estimators

- The values minimizing RSS are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Here, $\bar{x}$ and $\bar{y}$ are the *sample means*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- These give the *least squares estimates* of the coefficients.

Note: To minimize $S(\beta_0, \beta_1) := \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, *consider normal equations:*

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{1}$$

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \tag{2}$$

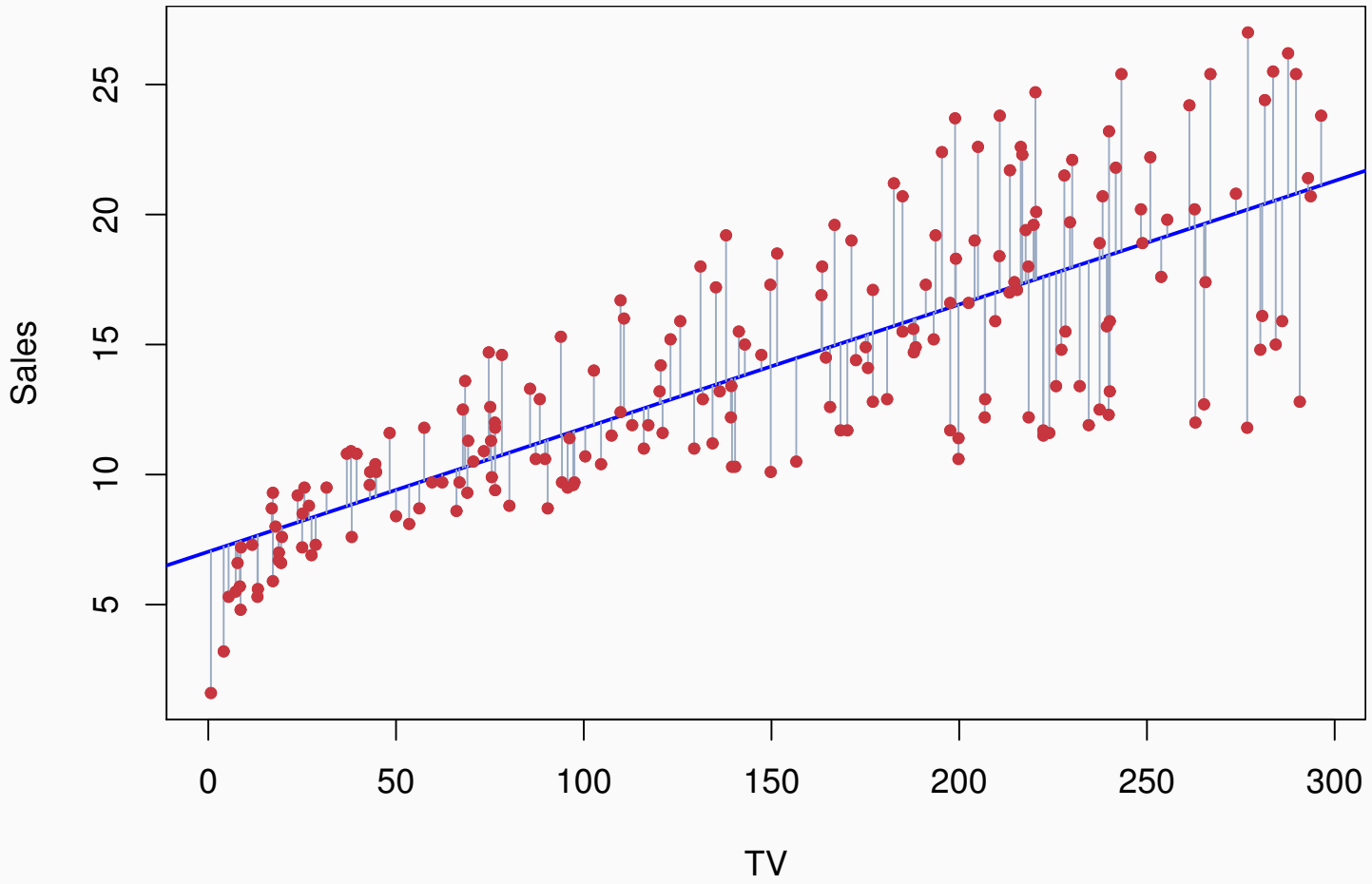
(1) *gives* $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. *Plugging* $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ *into* (2),

$$\sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})x_i = \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})(x_i - \bar{x}) = 0$$

gives

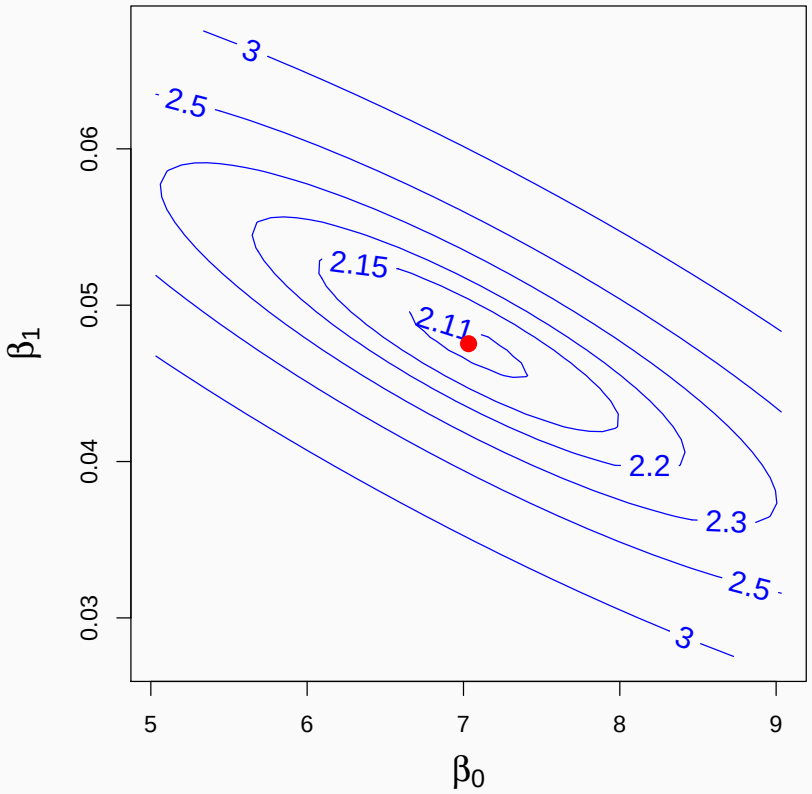
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

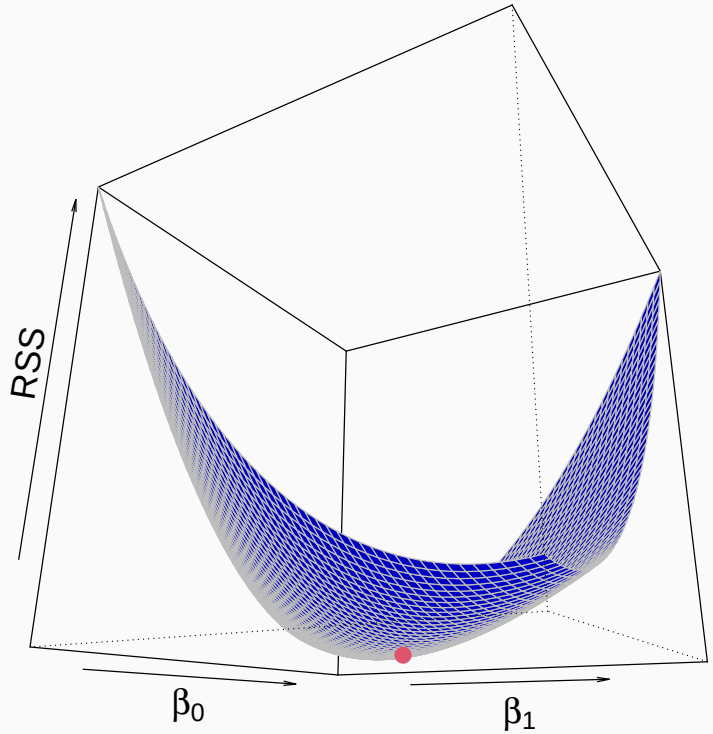
Example: advertising data



- Each vertical line segment = residual.
- The regression line minimizes the total squared residuals.

Contour and 3D RSS Visualization





- RSS surface shows a clear minimum at $(\hat{\beta}_0, \hat{\beta}_1)$.
- Advertising example: $\hat{\beta}_0 = 7.03$, $\hat{\beta}_1 = 0.0475$

Assessing the Accuracy of the Coefficient Estimates

Recall that simple linear regression takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n.$$

We now assume that $\varepsilon_i \sim N(0, \sigma^2)$ independently. That is, ε_i is normally distributed with mean 0 and variance σ^2 independently.

- For the simple linear regression model,

$$E(\hat{\beta}_1) = \beta_1, \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$E(\hat{\beta}_0) = \beta_0, \quad \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right),$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$, given that $x_1, \cdots, x_n$ are fixed.

Note: *Let*

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \; S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i.$$

Since $\hat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})y_i/S_{xx}$ *and* $E(y_i) = \beta_0 + \beta_1 x_i$,

$$\begin{aligned}\mathrm{E}(\hat{\beta}_1) &= \sum_{i=1}^{n}(x_i - \bar{x})E(y_i)/S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)/S_{xx} \\ &= \frac{\beta_0}{S_{xx}}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} + \frac{\beta_1}{S_{xx}}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})x_i}_{=S_{xx}} = \beta_1.\end{aligned}$$

Also, since y_i*'s are independent with* $\mathrm{Var}(y_i) = \sigma^2$,

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_{xx}}y_i\right) = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{S_{xx}^2}\mathrm{Var}\,(y_i) = \frac{S_{xx}}{S_{xx}^2}\sigma^2 = \frac{\sigma^2}{S_{xx}}.$$

Note: *Note that*

$$\mathrm{E}(\bar{y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\beta_0 + \beta_1 x_i + \varepsilon_i\right) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i) = \beta_0+\beta_1\bar{x}.$$

Since $\mathrm{E}(\hat{\beta}_1) = \beta_1$ *from the previous slide, we have*

$$\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\bar{y} - \hat{\beta}_1\bar{x}) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0.$$

Furthermore, $\hat{\beta}_0$ *can be represented as*

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}\frac{(x_i-\bar{x})}{S_{xx}}\right)y_i.$$

Again, since y_i*'s are independent with* $\mathrm{Var}(y_i) = \sigma^2$,

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}\left(\sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}\frac{(x_i-\bar{x})}{S_{xx}}\right)y_i\right) = \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}\frac{(x_i-\bar{x})}{S_{xx}}\right)^2$$

$$= \sigma^2\left(\underbrace{\sum_{i=1}^{n}\frac{1}{n^2}}_{=1/n} + \underbrace{\bar{x}^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{S_{xx}^2}}_{=\bar{x}^2/S_{xx}} - 2\frac{\bar{x}}{nS_{xx}}\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})}_{=0}\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right).$$

Standard error

- *Standard error (SE)* reflects how an estimate varies under repeated sampling.
  - It is the square root of the variance estimator.
- For the simple linear regression model:

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \quad \mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)}$$

where $\hat{\sigma}^2$ is an estimator of the error variance σ^2.

Confidence Intervals

- SEs can be used to construct *confidence intervals (CIs)*.
- A 95% CI for β_1 takes the form:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1) = \left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

- **Interpretation:** there is approximately a 95% chance that the interval will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample).
- Advertising data example:

$$\text{CI for } \beta_1 : 0.0475 \pm 2 \times 0.0027 = [0.0421, 0.0529]$$

Hypothesis Testing

- SEs also allow us to test hypotheses:

$$H_0 : \text{There is no relationship between } X \text{ and } Y,$$
$$H_A : \text{There is some relationship between } X \text{ and } Y,$$

  where H_0 is the *null hypothesis* and H_A is the *alternatvie hypothesis*.

- Mathematically:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

- If $\beta_1 = 0$, then $Y = \beta_0 + \varepsilon$ and X is unrelated to Y

Hypothesis Testing: t-statistic and p-value

- To test the null hypothesis H_0, compute a *t-statistic*:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

  This has a *t-distribution* with $n - 2$ degrees of freedom under H_0.
- The *p-value* is the probability of observing a value as extreme or more extreme than the computed t.
- If the p-value is small (e.g., < 0.05), reject H_0.

	Coefficient	**Std. Error**	**t-statistic**	**p-value**
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Note: Note that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

It follows that $Z := \dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$. Furthermore, it is known that

$$V := \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2),$$

and V is independent of Z. Therefore,

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \underbrace{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}_{Z} \left(\underbrace{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}}_{V} \cdot \frac{1}{n-2} \right)^{-1/2} = \frac{Z}{\sqrt{V/(n-2)}} \sim t(n-2).$$

Assessing the Overall Accuracy of the Model

- The *residual standard error* is an estimate of σ.

$$\hat{\sigma} = \text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the *residual sum of squares*.

- *R-squared* is the proportion of variability in Y that can be explained using X.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

  where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*. It can also be shown that $TSS = ESS + RSS$, where $\text{ESS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the *explained sum of squares*.

R^2 and Correlation

In simple linear regression, it can be shown that:

$$R^2 = r^2,$$

where r is the *correlation* between X and Y:

$$r = \mathrm{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

For the advertising data, we have the following results:

Quantity	**Value**
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Note: *Let*

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2,\ S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

From the definition of correlation r,

$$\begin{aligned} r^2 &= \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{S_{xx}}{S_{yy}}\frac{S_{xy}^2}{S_{xx}^2} = \frac{S_{xx}}{S_{yy}}\hat{\beta}_1^2 \\ &= \frac{\sum_{i=1}^{n}(\hat{\beta}_1 x_i - \hat{\beta}_1\bar{x})^2}{S_{yy}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{ESS}{TSS}. \end{aligned}$$

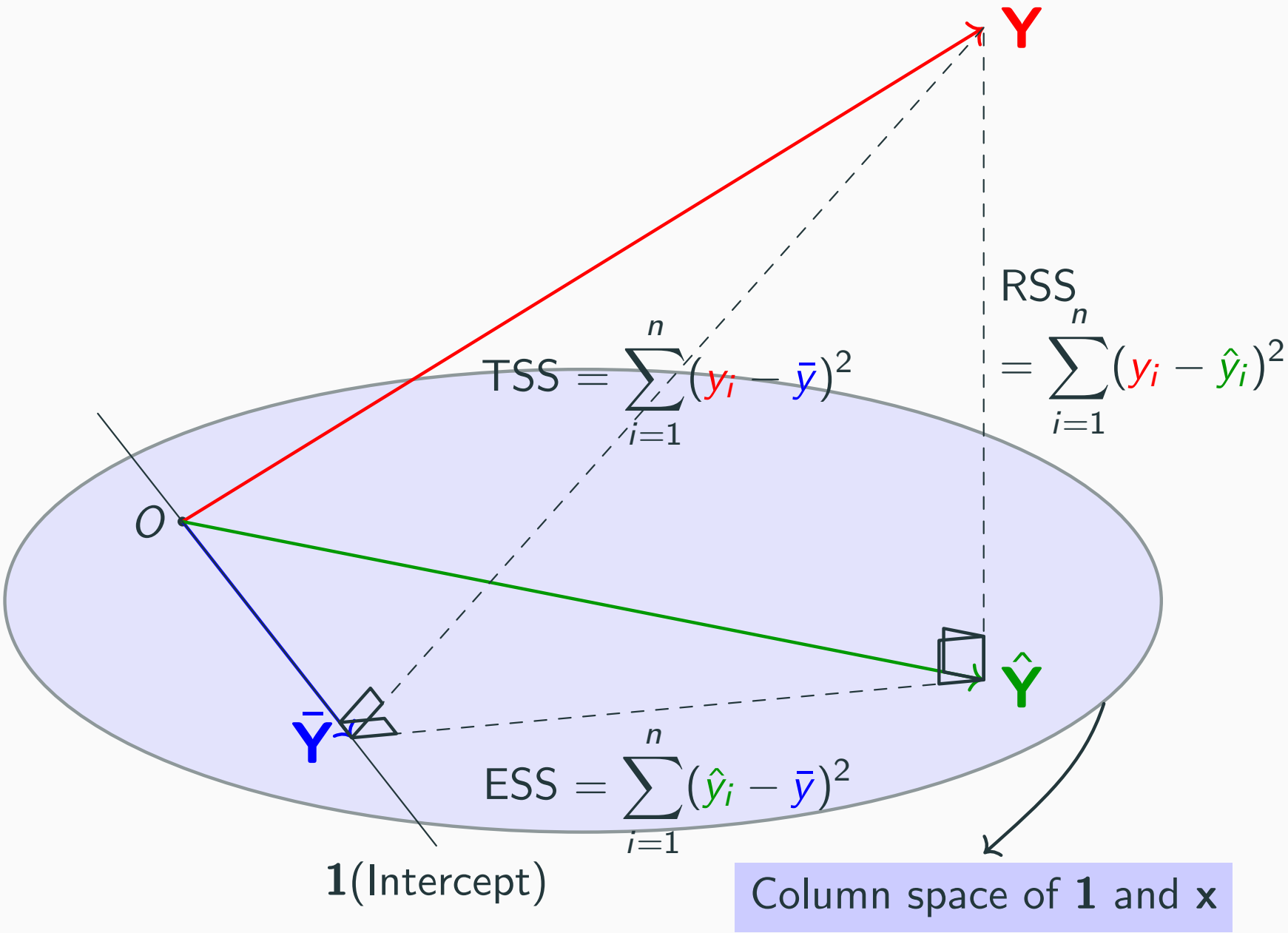
Furthermore, TSS *can be decomposed as the sum of* ESS *and* RSS *because*

$$\begin{aligned} TSS &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{RSS} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{ESS} + 2\underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{0}, \end{aligned}$$

and the last term vanishes since

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) = 0$$

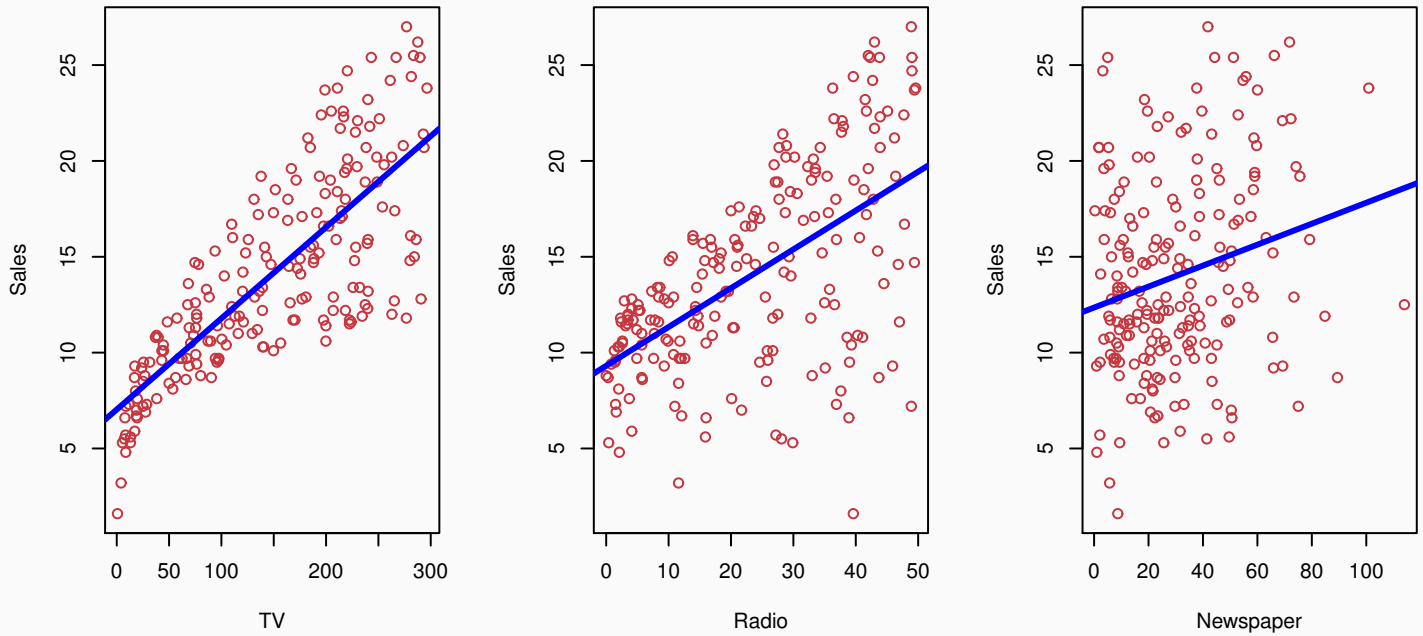
from the normal equations.


Y
RSS
$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
O
$\hat{\mathbf{Y}}$
$\bar{\mathbf{Y}}$
$\text{ESS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
$\mathbf{1}$(Intercept)
Column space of $\mathbf{1}$ and $\mathbf{x}$

Multiple Linear Regression

The Advertising Data Set (II)

- Data from 200 regional markets, each with:
  - Ad budgets for **TV**, **Radio**, **Newspaper** ($1,000s)
  - Product **sales** (1,000s of units)
- Goal: Understand how ad budgets influence sales
- *Predictors:* TV, Radio, Newspaper (X_1, X_2, X_3)
- *Response:* Sales (Y)



Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

- We interpret β_j as the *average* effect on Y of a one unit increase in X_j, *holding all other predictors fixed*.
- In the advertising example, the model becomes:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \varepsilon$$

Estimation in Multiple Linear Regression

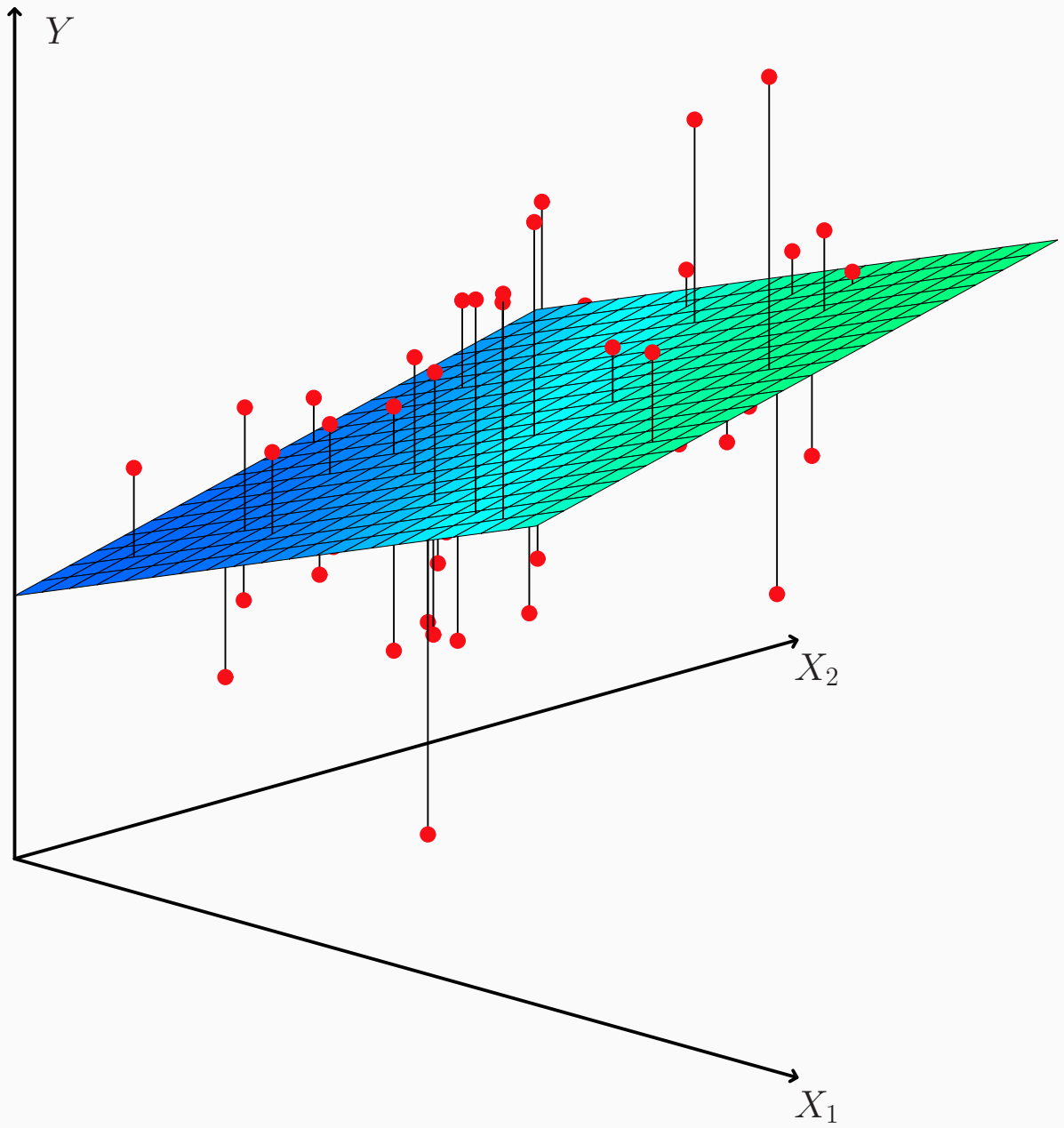
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we predict the response as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

- The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.


Y
X_2
X_1

Results for advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
Radio	0.189	0.0086	21.89	< 0.0001
Newspaper	-0.001	0.0059	-0.18	0.8599

- `TV` and `radio` are statistically significant.
- `Newspaper` has a large p-value $\Rightarrow$ not significant.

Results for advertising data

	TV	Radio	Newspaper	Sales
TV	1.0000	0.0548	0.0567	0.7822
Radio		1.0000	0.3541	0.5762
Newspaper			1.0000	0.2283
Sales				1.0000

Table 1: Correlation matrix for `TV`, `radio`, `newspaper`, and `sales`.

- `TV` and `sales` are highly correlated.
- `Newspaper` and `sales` show weak correlation.

Model and matrix notation

Let

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then the multiple linear regression model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

MVN denotes the multivariate normal distribution, and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Assume $\boldsymbol{X}$ is full rank: $\text{rank}(\boldsymbol{X}) = p + 1$.

Least squares estimator and distributional results

1. The *ordinary least squares estimator(OLS)* of $\boldsymbol{\beta}$, which minimizes

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

   is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

2. The estimator $\hat{\boldsymbol{\beta}}$ has distribution

$$\hat{\boldsymbol{\beta}} \sim MVN\left(\boldsymbol{\beta},\ \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right).$$

3. The estimators $\hat{\boldsymbol{\beta}}$ and

$$\hat{\sigma}^2 := \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - p - 1}$$

   are independent.

4. The scaled variance estimator satisfies

$$\frac{(n - p - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - p - 1).$$

Note: We seek the least-squares estimator $\hat{\boldsymbol{\beta}}$ that minimizes

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The first-order condition yields the normal equations

$$\boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y}.$$

If $\boldsymbol{X}$ is full rank, then $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible and

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Since $\hat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$,

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathrm{E}(\boldsymbol{y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathrm{E}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}, \end{aligned}$$

where we used $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$. Hence, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. Also, since $\boldsymbol{y}$ has covariance matrix $\mathrm{Var}(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$,

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right] \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathrm{Var}(\boldsymbol{y}) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \end{aligned}$$

It can be shown that the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is the UMVUE(Uniformly Minimum Variance Unbiased Estimator) of β.

Gauss-Markov Theorem: BLUE

Suppose $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n..$

The ordinary least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is the *best among all linear unbiased estimators(BLUE)* of $\boldsymbol{\beta}$.

Gauss-Markov Theorem

For all estimators of the form $\tilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y}$ with an $(p+1) \times n$ matrix $\boldsymbol{C}$ such that $\mathrm{E}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, we have

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}) \text{ is nonnegative definite,}$$

or equivalently, $\mathrm{Var}(\boldsymbol{\ell}^\top \tilde{\boldsymbol{\beta}}) \geq \mathrm{Var}(\boldsymbol{\ell}^\top \hat{\boldsymbol{\beta}})$ for any $(p+1)$-vector $\boldsymbol{\ell}$.

The theorem can be proved by letting $\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top + \boldsymbol{B}$ for a $(p+1) \times n$ matrix $\boldsymbol{B}$, and note that the unbiasedness is equivalent to $\boldsymbol{B}\boldsymbol{X} = \boldsymbol{O}$.

Some Important Questions

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help explain Y, or is only a subset useful?
- How well does the model fit the data?
- Given a set of predictor values, what response should we predict, and how accurate is our prediction?

Is At Least One Predictor Useful?

We test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,$$

versus the alternative

$$H_A : \text{at least one } \beta_j \text{ is non-zero.}$$

We may use the *F-statistic*:

$$F = \frac{\text{ESS}/p}{\text{RSS}/(n-p-1)} \overset{H_0}{\sim} F_{p,n-p-1}$$

Quantity	**Value**
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Note: Derivation of F-statistic: Suppose the multiple linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n),$$

and consider

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0, \; H_A : \text{not } H_0.$$

From the result of slide 27,

$$V_2 := RSS/\sigma^2 \sim \chi^2(n-p-1), \; RSS = (n-p-1)\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

If H_0 is true, it can be also shown that

$$V_1 := ESS/\sigma^2 \sim \chi^2(p), \; ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

Independence of V_1 and V_2 gives

$$F = \frac{V_1/p}{V_2/(n-p-1)} = \frac{ESS/p}{RSS/(n-p-1)} \overset{H_0}{\sim} F(p, n-p-1).$$

Best Subset Selection

- Compute the least squares fit for *all possible subsets* of predictors.
- Choose model using a criterion that balances training error with model size.
- Problem: Number of models $= 2^p$ becomes infeasible for large p.
- For example, when $p = 40$, there are over a billion models.

Forward Selection

- Start with the *null model* (intercept only).
- Add the variable that minimizes RSS among all one-variable models.
- Then add the next variable that minimizes RSS among all two-variable models, and so on.
- Stop when no remaining variables have a p-value below a chosen threshold.

Backward Selection

- Start with the *full model* (all p predictors).
- Remove the predictor with the largest (least significant) p-value.
- Repeat until all remaining variables have p-values below a significance threshold.

Other Model Selection Criteria

- More systematic criteria for choosing optimal models:
  - Mallow's C_p
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
  - Adjusted R^2
  - Cross-validation (CV)

Effect of Deleting Regressors: Bias and Variance

Note: *Assume the true model is*

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}, \qquad \mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_n,$$

but we fit the subset model using only $\boldsymbol{X}_p$*:*

$$\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top \boldsymbol{y}.$$

Taking expectation under the full model,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p + (\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top \boldsymbol{X}_r\,\boldsymbol{\beta}_r =: \boldsymbol{\beta}_p + \boldsymbol{A}\boldsymbol{\beta}_r.$$

So omitting regressors generally introduces bias. Also,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_p) = \sigma^2(\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}, \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}_p^{\mathrm{full}}) = \sigma^2(\boldsymbol{X}_p^\top \left(\boldsymbol{I} - \boldsymbol{X}_r(\boldsymbol{X}_r^\top \boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^\top\right) \boldsymbol{X}_p)^{-1}$$

where $\hat{\boldsymbol{\beta}}_p^{\mathrm{full}}$ *denotes the coefficients on* $\boldsymbol{X}_p$ *from the full model.*

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_p) \preceq \mathrm{Var}(\hat{\boldsymbol{\beta}}_p^{\mathrm{full}}).$$

Thus, deleting regressors can reduce variance, but at the cost of bias.

Mean Square Error: Bias–Variance Trade-off

Note: *For any estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$,*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathrm{Var}(\hat{\boldsymbol{\theta}}) + \left[\mathrm{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right]\left[\mathrm{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right]^{\top}.$$

Applying this to $\hat{\boldsymbol{\beta}}_p$ and $\hat{\boldsymbol{\beta}}_p^{\mathrm{full}}$,

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p) = \sigma^2(\boldsymbol{X}_p^{\top}\boldsymbol{X}_p)^{-1} + \boldsymbol{A}\boldsymbol{\beta}_r\boldsymbol{\beta}_r^{\top}\boldsymbol{A}^{\top}$$

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p^{\mathrm{full}}) = \sigma^2(\boldsymbol{X}_p^{\top}\left(I - \boldsymbol{X}_r(\boldsymbol{X}_r^{\top}\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^{\top}\right)\boldsymbol{X}_p)^{-1}$$

The full-model estimator $\hat{\boldsymbol{\beta}}_p^{\mathrm{full}}$ is unbiased but has larger variance. Hence, if the omitted coefficients $\boldsymbol{\beta}_r = \mathbf{0}$,

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p) \preceq \mathrm{MSE}(\hat{\boldsymbol{\beta}}_p^{\mathrm{full}}).$$

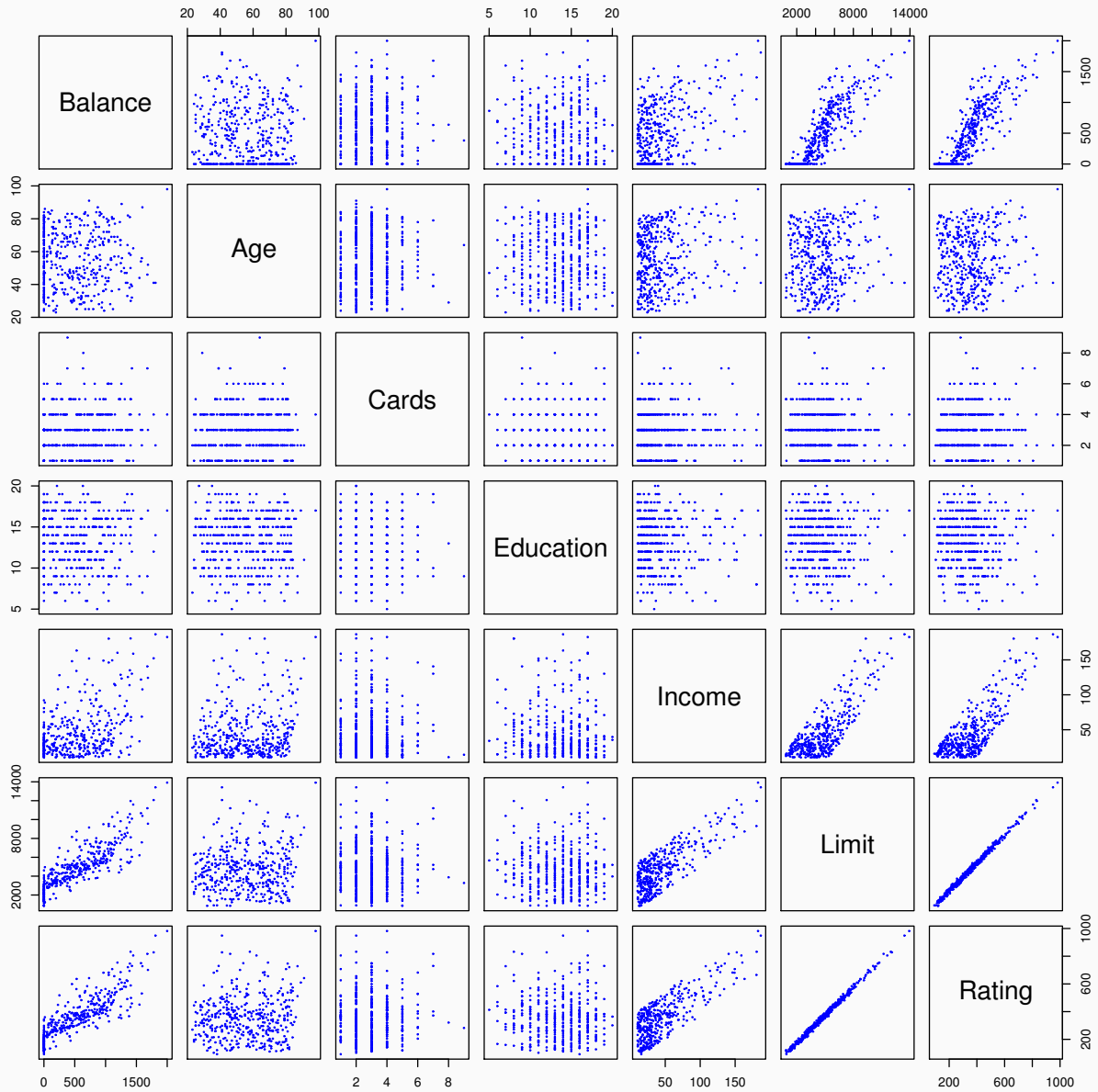
Conclusion: deleting variables can improve estimation accuracy by reducing variance more than the bias it introduces.

Other Considerations in the Regression Model

Qualitative Predictors

- Some predictors are not quantitative but *qualitative*, taking a discrete set of values.
- These are also called *categorical* or *factor variables*.
- In the credit card data set, in addition to the 7 quantitative variables, there are 4 qualitative variables:
  - `gender`
  - `student` (student status)
  - `status` (marital status)
  - `ethnicity` (Caucasian, African American, Asian)

Quantitative Variables for Credit Card Data



Qualitative Predictors — Binary Example

- Example: Compare credit card `balance` between males and females, ignoring other variables.
- Define a *dummy variable* (or *indicator variable*):

$$x_i = \begin{cases} 1 & \text{if person } i \text{ is female} \\ 0 & \text{if person } i \text{ is male} \end{cases}$$

- Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person is female.} \\ \beta_0 + \varepsilon_i & \text{if } i\text{-th person is male.} \end{cases}$$

Credit Card Data — Gender Model

Coefficient	Std. Error	t-statistic	p-value	
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

- The coefficient for `gender` is not statistically significant.

Qualitative Predictors with More Than Two Levels

- For categorical variables with more than two levels (e.g., `ethnicity`), use multiple dummy variables.
- Example: `ethnicity` levels = Asian, Caucasian, African American (baseline)

$$x_{i1} = \begin{cases} 1 & \text{if person } i \text{ is Asian} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if person } i \text{ is Caucasian} \\ 0 & \text{otherwise} \end{cases}$$

Ethnicity Regression Model

- Resulting model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if person } i \text{ is Asian,} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if person } i \text{ is Caucasian,} \\ \beta_0 + \varepsilon_i & \text{if person } i \text{ is African American.} \end{cases}$$

- Always use one fewer dummy variable than the number of levels.

Credit Card Data — Ethnicity Model

Coefficient	Std. Error	t-statistic	p-value	
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

- Neither group shows significant differences from the baseline (African American).

Removing the Additive Assumption: Interactions

- Linear model (without interaction):

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{radio}$$

- Assumes the effect of one medium is independent of the level of the others (the slope for TV is always β_1).
- In practice, advertising media can reinforce each other (*synergy* = *interaction*).
- With a fixed budget, splitting between TV and radio may outperform spending all on one medium.

Interaction Model and Estimates

$$\texttt{sales} = \beta_0 + \beta_1\,\texttt{TV} + \beta_2\,\texttt{radio} + \beta_3\,(\texttt{TV} \times \texttt{radio}) + \varepsilon$$
$$= \beta_0 + (\beta_1 + \beta_3\,\texttt{radio})\,\texttt{TV} + \beta_2\,\texttt{radio} + \varepsilon.$$

Coefficient	Estimate	Std. Error	t-stat	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV$\times$radio	0.0011	0.000	20.73	< 0.0001

Strong evidence for *interaction*: $H_A : \beta_3 \neq 0$.

Interpreting the Coefficients

- *Effect of TV* (per \$1,000) at radio level:

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio} \quad \text{(units of sales)}.$$

- *Effect of radio* (per \$1,000) at TV level:

$$(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV} \quad \text{(units of sales)}.$$

Marginal effects increase with the other medium.

The Hierarchy Principle

- Sometimes an interaction has a tiny p-value, but one or both *main effects* do not.
- *Hierarchy principle:* If we include an *interaction* term, we should also include the corresponding *main effects*, even if their p-values are not significant.
- Reason: Interactions are hard to interpret without main effects; their meaning changes if main effects are omitted.

Quantitative × Qualitative Interactions (Credit Data)

Consider the Credit data set, and suppose that we wish to predict `balance` using `income` (quantitative) and `student`(qualitative).

No interaction:

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} 0, & \text{(non-student)}, \\ \beta_2, & \text{(student)}. \end{cases}$$

- Two parallel lines: same slope β_1, different intercepts.

With interaction:

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} 0, & \text{(non-student)}, \\ \beta_2 + \beta_3 \times \texttt{income}_i, & \text{(student)}. \end{cases}$$

- Different slopes, different intercepts.

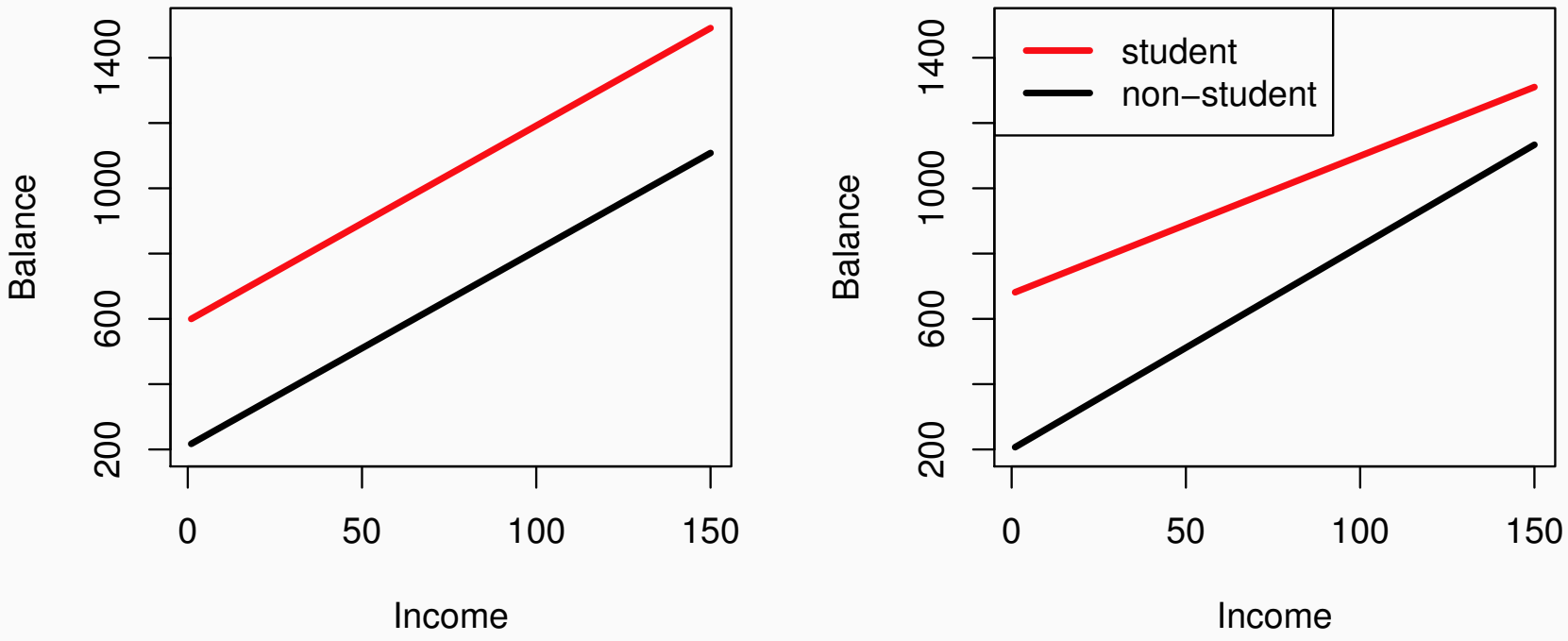
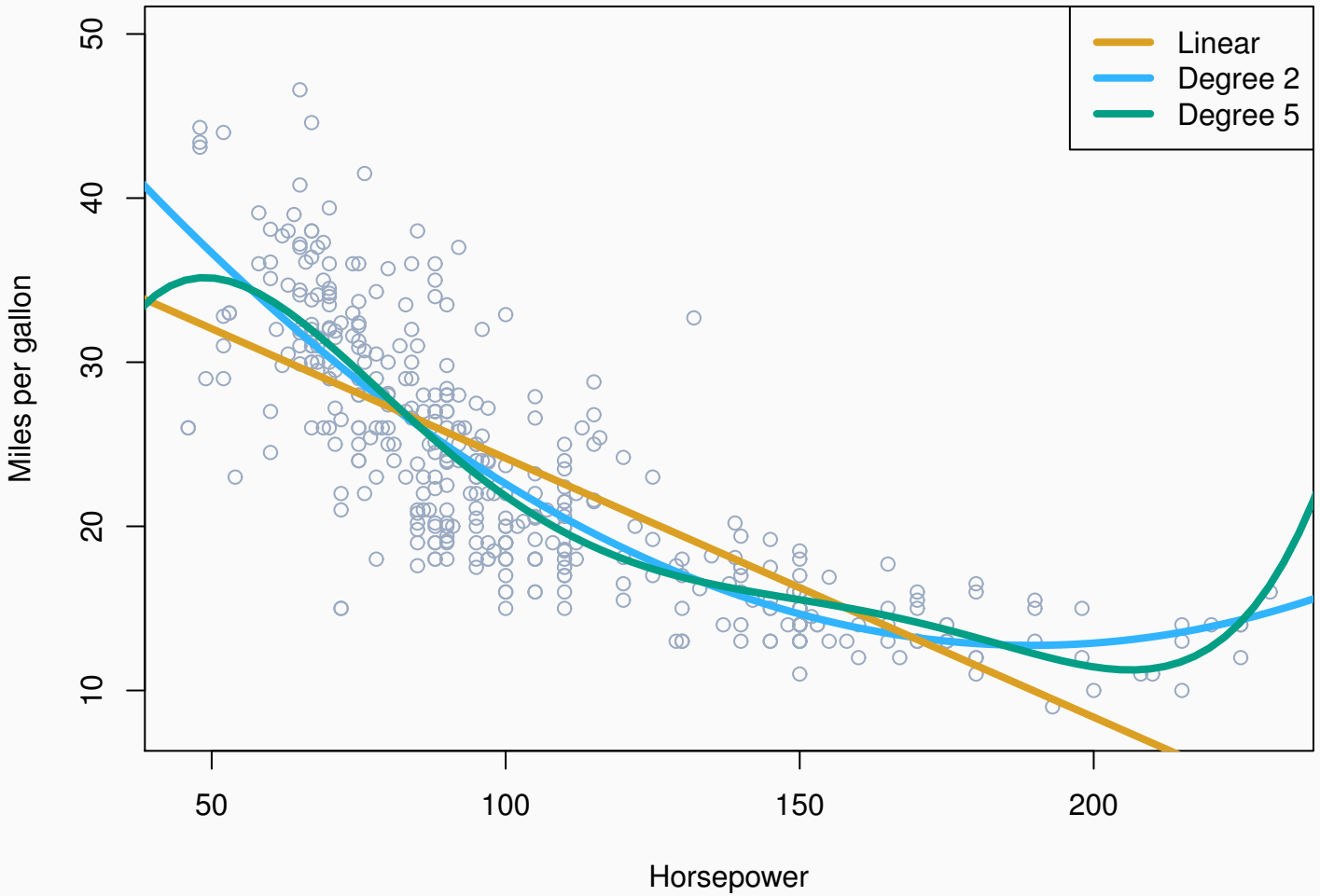


Figure 1: Left: no interaction between `income` and `student`. Right: with an interaction term between `income` and `student`.

Non-linear Effects of Predictors



polynomial regression on `Auto` data

The figure suggests that

$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{horsepower} + \beta_2 \texttt{horsepower}^2 + \varepsilon.$$

may provide a better fit.

Coefficient	Estimate	Std. Error	t-stat	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

What We Did Not Cover (Here)

- Outliers
- Non-constant error variance (heteroskedasticity)
- High leverage points
- Collinearity

Exercises

1. Consider the simple linear regression without an intercept:

$$y_i = \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ independently.

A. Find the least square estimate of β_1, $\hat{\beta}_1$, that minimizes

$$S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2.$$

B. Let the ith fitted value be $\hat{y}_i = \hat{\beta}_1 x_i$. Using A, show that

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} \hat{y}_i^2.$$

Exercises(continued)

1. Consider the simple linear regression without an intercept:

$$y_i = \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ independently.

C. Find $\mathrm{E}(\hat{\beta}_1)$. Also, show that $\mathrm{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^n x_i^2}$.

D. Find $\mathrm{E}(\hat{y}_i)$. Also, show that $\mathrm{Var}(\hat{y}_i) = \dfrac{\sigma^2 x_i^2}{\sum_{i=1}^n x_i^2}$.

E. Find $\mathrm{E}(e_i)$. Also, show that $\mathrm{Var}(e_i) = \sigma^2 \left(1 - \dfrac{x_i^2}{\sum_{i=1}^n x_i^2}\right)$,

where $e_i = y_i - \hat{y}_i$. (Hint: use
$\mathrm{Var}(y_i) = \mathrm{Var}(y_i - \hat{y}_i) + \mathrm{Var}(\hat{y}_i)$)

Exercises

2. Suppose that we have the two multiple linear regression models.

$$\text{Model A: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$
$$\text{Model B: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

for $i = 1, \cdots, n$, where $\varepsilon_i \sim N(0, \sigma^2)$ independently for both models. Let $\hat{y}_i^{(\mathrm{A})}$ and $\hat{y}_i^{(\mathrm{B})}$ be fitted values for Model A and B, respectively. Compare each of the following quantities. (Hint: use projection)

- $\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(\mathrm{A})} \right)^2$ and $\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(\mathrm{B})} \right)^2$
- $\sum_{i=1}^{n} \left(\hat{y}_i^{(\mathrm{A})} - \bar{y} \right)^2$ and $\sum_{i=1}^{n} \left(\hat{y}_i^{(\mathrm{B})} - \bar{y} \right)^2$
- R^2 for model A and R^2 for model B.

Exercises

3. In slide 44, recall the regression of `ethnicity`(with three levels Asian, Caucasian, African American) on `balance` in the Credit data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

where

$$x_{i1} = \begin{cases} 1 & \text{if person } i \text{ is Asian} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if person } i \text{ is Caucasian} \\ 0 & \text{otherwise} \end{cases}$$

Express the mean difference in `balance` between Asian and Caucasian in terms of regression coefficients β_0, β_1, and β_2.