# Workspace 'Workspace' in 'Lec4'
Page 1 (row 1, column 1)

# Module 3: Linear Regression

Lecture 4
Jan 19th, 2023

Oracle $f(x) = E(Y \mid X = x)$

# Recap

$\hat{f}_1(x)\ Y = 0.8 + 0.6x$

$\hat{f}_2(x) = 0.78 + 0.5x$

random

random

train $\rightarrow \hat{f}(x)$

knn

$Y = \beta_0 + \beta_1 X$

Estimating $f$: non-parametric vs parametric

Assessing Model Accuracy: Training MSE and $\boxed{\text{Test}}$ MSE

$Var(X) = E((X - EX)^2)$

train $\rightarrow \hat{f} \rightarrow$ Test $\rightarrow$ test MSE

Bias-Variance Trade-off

$X \sim N(0, 1)$

$EX = 0$

$E(Y \mid X = x)$

$E(|\hat{f}(x) - f(x)|) = \text{Bias}$

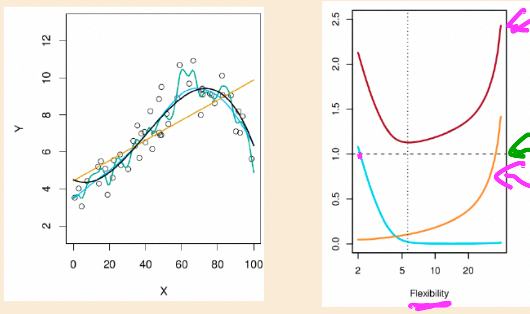


$Var(\hat{f}(x)) = E([\hat{f}(x) - E(\hat{f}(x))]^2)$

Estimating *f*: non-parametric vs parametric

Assessing Model Accuracy: Training MSE and Test MSE

Bia




Test MSE

$Y_0 = f(x_0) + \varepsilon$

Label the line corresponding to each of the following:

- MSE
- Bias
- Variance of $\hat{f}(x_0)$
- Variance of $\varepsilon$

Irreducible error

$$E(y_0 - \hat{f}(x_0))^2 = \boxed{\mathrm{Var}(\hat{f}(x_0))} + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \boxed{\mathrm{Var}(\varepsilon)}$$

# Simple Linear Regression

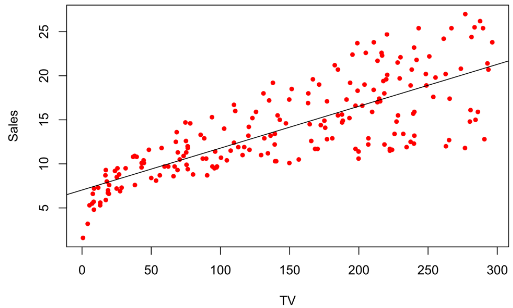


- We have only one feature

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

Where $\beta_0$ and $\beta_1$ are two unknown constants also known as coefficient or parameters.

- Example:




$$\text{Sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

# Topics

MICHIGAN STATE UNIVERSITY

- Least squares coefficient estimates for linear regression
- Residual sum of squares (RSS)
- Confidence interval, hypothesis test, and p-value for coefficient estimates
- Residual standard error (RSE)
- R squared

# Simple Linear Regression



- We have only one feature

$$\boxed{Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon}$$

Where $\beta_0$ and $\beta_1$ are two unknown constants also known as coefficient or parameters.

- Given the training data, we can estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients and predict future Y using

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

Where $\hat{y}$ indicates a prediction of $Y$ on the basis of $X = x$. The hat symbol denotes an estimated value.

# Example

MICHIGAN STATE UNIVERSITY

| | TV | Radio | Newspaper | Sales |
|---|---|---|---|---|
| 1 | 230.1 | 37.8 | 69.2 | 22.1 |
| 2 | 44.5 | 39.3 | 45.1 | 10.4 |
| 3 | 17.2 | 45.9 | 69.3 | 9.3 |
| 4 | 151.5 | 41.3 | 58.5 | 18.5 |
| 5 | 180.8 | 10.8 | 58.4 | 12.9 |
| 6 | 8.7 | 48.9 | 75 | 7.2 |
| 7 | 57.5 | 32.8 | 23.5 | 11.8 |
| 8 | 120.2 | 19.6 | 11.6 | 13.2 |
| 9 | 8.6 | 2.1 | 1 | 4.8 |
| 10 | 199.8 | 2.6 | 21.2 | 10.6 |
| 11 | 66.1 | 5.8 | 24.2 | 8.6 |

$\leftarrow \hat{y}_1$

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

- $$\text{sales} \approx \beta_0 + \beta_1 \text{TV}$$
- $\beta_0$ intercept; $\beta_1$ slope
- Coefficients or parameters : $\{\beta_0, \beta_1\}$
- Once we have good guesses for $\hat{\beta}_i$, model is

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

# How to estimate?

MICHIGAN STATE UNIVERSITY

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction of Y based on the $i$th value of X.

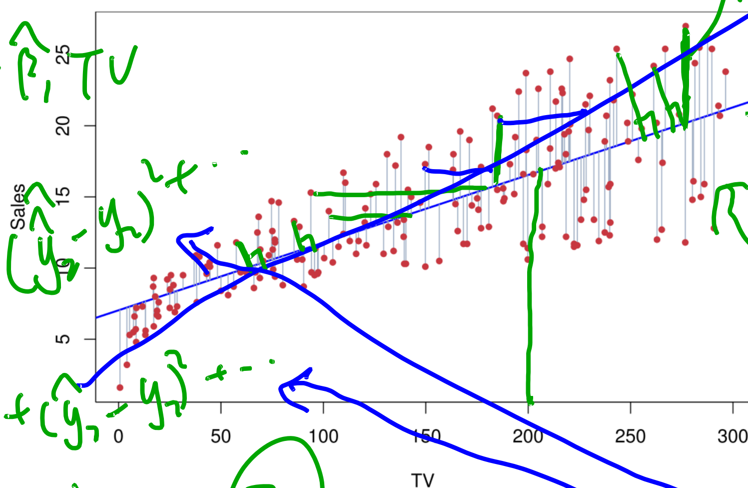
$\hat{TV} = \hat{\alpha}_0 + \hat{\alpha}_1 \, Sale$

- Errors: $e_i = y_i - \hat{y}_i$

$Sale = \hat{\gamma}_0 + \hat{\gamma}_1 \, TV \Leftarrow$ min ⊥ distance

$\alpha_0 \neq \beta_0$

$Sale = \hat{\beta}_0 + \hat{\beta}_1 \, TV$




$e$

$\hat{\beta}_0 + \hat{\beta}_1 x$

$A = (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \cdots$

$B = (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \cdots$

$A > B$

$\hat{Y} = \gamma_0 + \gamma_1 x$

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

# How to estimate?

MICHIGAN STATE UNIVERSITY

- Residual Sum of Squares

$$\text{RSS} = \underline{e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2} = \sum_{i=1}^{n} e_i^2$$

- Equivale

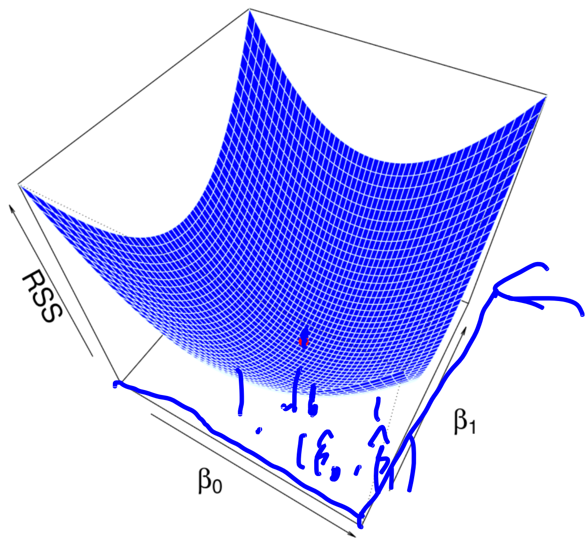
$$\min \ \underline{\text{RSS}} = \boxed{\sum_{i=1}^{n}(y_i - \hat\beta_0 - \hat\beta_1 x_i)^2}$$

$(\hat\beta_0, \hat\beta_1)$

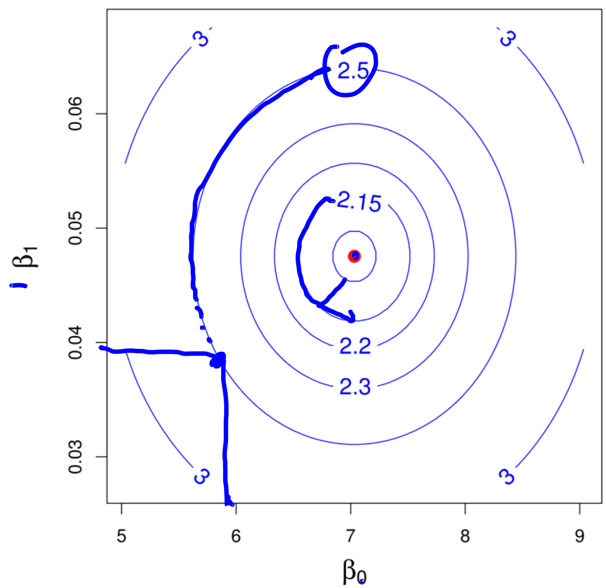
# How to estimate?

- Minimizing RSS to find the best estimates.

$$\underset{\beta_0,\beta_1}{\min}\ \mathrm{RSS} = \underset{\beta_0,\beta_1}{\min} \sum_{i=1}^{n} e_i^2 = \underset{\beta_0,\beta_1}{\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



$$\min_{\beta_0,\beta_1} \text{RSS} = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



# Solving for minimal RSS

MICHIGAN STATE UNIVERSITY

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- RSS is a **convex** function of $\beta_0, \beta_1$
- Minimum achieved when (recall the chain rule):

$$\frac{\partial\,\text{RSS}}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial\,\text{RSS}}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

# Linear Regression Coefficients

MICHIGAN STATE UNIVERSITY

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Solution:

$$\boxed{\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}}$$

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

where

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

# Linear Regression Coefficients



$Y = X\beta + \varepsilon$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_{11}, & \cdots & x_{1p} \\ \vdots & & \\ x_{n1}, & \cdots & x_{np} \end{pmatrix} . \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

$$RSS = \underbrace{(Y - X\beta)}_{n\times 1}^T (Y - X\beta) = Y^T Y - \beta^T X^T X\beta - 2Y^T X\beta$$

$$\frac{\partial RSS}{\partial \beta} = 0 + 2X^T X\beta - 2X^T Y = 0$$

$$\Rightarrow (X^T X)\beta = X^T Y \Rightarrow \hat{\beta} = (\underset{p\times n}{X^T} \underset{n\times p}{X})^{-1} X^T Y$$

$$= (X^T X)^{-1} X^T (X\beta + \varepsilon)$$

$$= \cancel{(X^T X)^{-1}} \cancel{X^T X} \beta + (X^T X)^{-1} X^T \varepsilon$$

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon$$

$$E(\hat{\beta}) = \beta + E((X^T X)^{-1} X^T \varepsilon)$$

$$= \beta + (X^T X)^{-1} X^T \underbrace{E(\varepsilon)}_{=0}$$

$$= \beta$$

# Why Minimizing RSS?

MICHIGAN STATE UNIVERSITY

- It has closed form!!
- Maximize likelihood when $Y = \beta_0 + \beta_1 X + \epsilon$ , when $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Best Linear Unbiased Estimator (BLUE): Gauss-Markov Theorem.

OLS $\hat{\beta}$

$E(\hat{\beta}) = \beta$

# Bias in Estimation

MICHIGAN STATE UNIVERSITY

- Assume a true value $\mu^*$
- An estimate from training data $\hat{\mu}$
- The estimate is unbiased if E( $\hat{\mu}$ ) = $\mu^*$

$E X = \mu^*$

$X_1, X_2, \ldots, X_n$

- Sample mean is unbiased for population mean

$$E(\hat{\mu}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mu$$

$S = \frac{X_1 + X_2 + \cdots + X_n}{n}$

$E(S) = E\left(\frac{X_1 + \cdots + X_n}{n}\right)$

$= \frac{EX_1 + EX_2 + \cdots + EX_n}{n}$

$= \frac{n\mu^*}{n} = \mu^*$

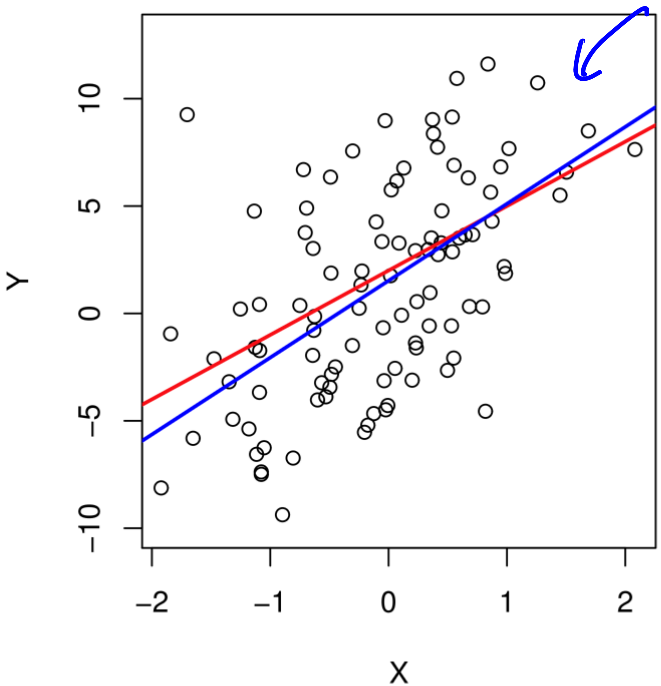
- Standard variance estimate is biased

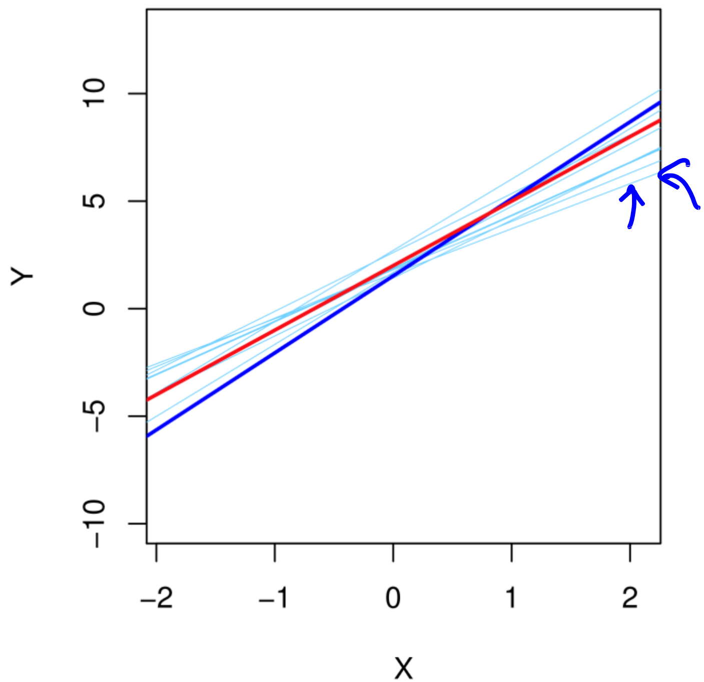
$$E(\hat{\sigma}^2) = E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2] \neq \sigma^2$$

$Var(X) = \sigma^2$

$\hat{\sigma}^2 = \frac{1}{(n-1)}\sum (X_i - \bar{X})^2$

# Linear Regression is Unbiased



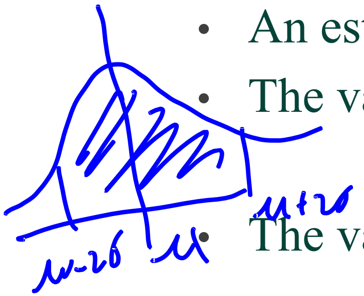
# Variance of the estimates



$E(\hat{\beta}_1) = \beta_1$

$\hat{\beta}$

- Assume a true value $\mu^*$
- An estimate from training data $\hat{\mu}$
- The variance of sample mean is: $\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$




$2\sigma$

- The variances of the linear regression estimates are

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

where $\sigma^2 = \mathrm{Var}(\epsilon)$

$\mathrm{Var}(\epsilon) = \sigma^2$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$$

d.f.

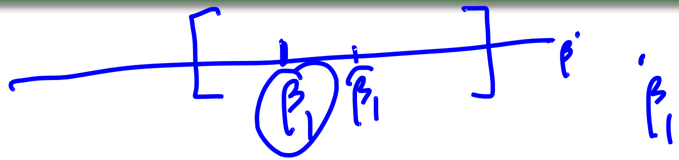
$RSS = \sum_{i=1}^{n} e_i$

$e_i^2 = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2$

- We then can calculate the confidence intervals. The 95% confidence interval for $\beta_1$ approximately takes the form
- $$\hat{\beta}_1 \pm \underset{1.96}{\cancel{2}} \cdot \mathrm{SE}(\hat{\beta}_1)$$

$\hat{\beta}_1 \pm 2\hat{\sigma} \Leftarrow$ 95% CI

# Confidence Interval



That is, there is approximately a 95% chance that the interval

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

will contain the true value of $\beta_1$ (under a scenario where we got repeated samples like the present sample)

For the advertising data, the 95% confidence interval for $\beta_1$ is $[0.042, 0.053]$

# Hypothesis Testing

MICHIGAN STATE UNIVERSITY

- Standard errors can also be used to perform **hypothesis tests** on the coefficients. The most common hypothesis test involves testing the **null hypothesis** of

$H_0$ : There is no relationship between $X$ and $Y$

versus the *alternative hypothesis*

$H_A$ : There is some relationship between $X$ and $Y$.

- Mathematically, it is

$$\boxed{H_0 : \beta_1 = 0}$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and $X$ is not associated with $Y$.

$Y = \beta_0 + \beta_1 X$

$\beta_0$

$\beta_1$

# Hypothesis Testing

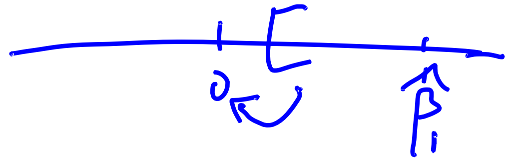
- Mathematically, it is

$$H_0 : \beta_1 = 0$$

versus

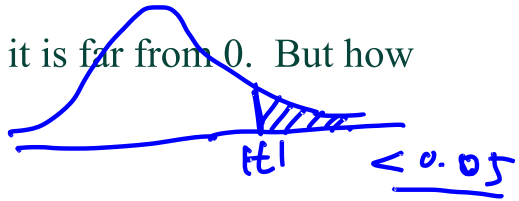
$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and $X$ is not associated with $Y$.




- We have $\hat{\beta}_1$ from data and want to test whether it is far from 0. But how far?

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




- This will have a t-distribution with *n-2* degrees of freedom, assuming $\beta_1 = 0$

- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value.

- We will specify an alpha value (0.05) before Hypothesis testing. If p-value less than the alpha value, we will reject the null hypothesis.

# Results for the Advertising Data

MICHIGAN STATE UNIVERSITY

| | Coefficient | Std. Error | t-statistic | p-value |
|---|---|---|---|---|
| $\hat{\beta}_0$ Intercept | 7.0325 | 0.4578 | 15.36 | < 0.0001 |
| $\hat{\beta}_1$ TV | $\hat{}$0.0475 | 0.0027 | 17.67 | < 0.0001 |

$<< 0.05$

- Since p-value < 0.05, we reject the null hypothesis and conclude that TV is related to sale.