Section 5: Regression

Ricky Truong (rickytruong@college.harvard.edu),
Emily Xing (exing@college.harvard.edu)

1 Introduction

1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

1.2 Office Hours

- Mondays, 7:30 - 9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM - 12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30 – 11:30 AM in Cabot D-Hall (Emily).

2 Big Picture

Thus far, we've been working with mainly one variable Y, but in *regression*, the data for each observation comes as a pair $(\vec{X}, Y)$, where Y is our outcome variable and $\vec{X}$ is our predictor variable(s). In *predictive regression*, we focus on the *conditional* distribution of Y given $\vec{X}$—i.e., we use $\vec{X}$ to predict Y. In *descriptive regression*, we focus on the *joint* distribution of $(\vec{X}, Y)$—i.e., we describe and study the extent to which they vary together linearly. Regardless of our approach, our recurring estimator will be $\hat{\theta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$.

Importantly, the course uses slightly unconventional notation for linear regression, denoting the regression coefficients as θ_j and the number of predictors as K. In other courses, these are sometimes denoted as β_j and p or J, respectively. Related, we distinguish between regression error, denoted in the course as $U(\vec{x})$, and the error term in a model, often denoted elsewhere as ε_i. These two are closely related, so we often return to the model $Y_i = \theta X_i + \varepsilon_i$ to illustrate our ideas.

3 Predictive Regression

Idea. In *predictive regression*, we're interested in modeling Y as a function of $\vec{X}$. E.g., Y can be annual salary, and X can be age. This can take many different forms, but in the course, we focus on *linear regression* (used when Y can take on any real number) and briefly touch on *logistic regression* (used when Y is binary).

We begin with *simple linear regression*, where we have only one predictor. With *no intercept*, our model is $Y_i = \theta X_i + \varepsilon_i$ for observation i. I.e., there is some underlying θ that captures the relationship between X and Y for every observation. We want to estimate this! The ε_i term is the *random noise* surrounding each observation. Without this, Y would be perfectly

deterministic given $X = x$. The estimation of θ is like a line of best fit. We observe n noisy outcomes $y_1, \ldots, y_n$ with their corresponding predictors $\vec{x}_1, \ldots, \vec{x}_n$. With this, we can estimate θ with three different approaches: MOM, MLE, and OLS. Incredibly, they arrive at the same estimator! We can extend linear regression to a model *with an intercept*—i.e., $Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i$, where θ_0 is the intercept and θ_1 is the slope—and to a model with K *predictors*—i.e., $Y_i = \theta_0 + \theta_1 X_{i,1} + \ldots + \theta_K X_{i,K} + \varepsilon_i$, where θ_j is the slope for predictor X_j. In matrix form, our full model is $\vec{Y}_{n\times 1} = \mathbf{X}_{n\times(K+1)}\vec{\theta}_{(K+1)\times 1} + \vec{\varepsilon}_{n\times 1}$. Equivalently,

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & \ldots & X_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \ldots & X_{n,K} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_K \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

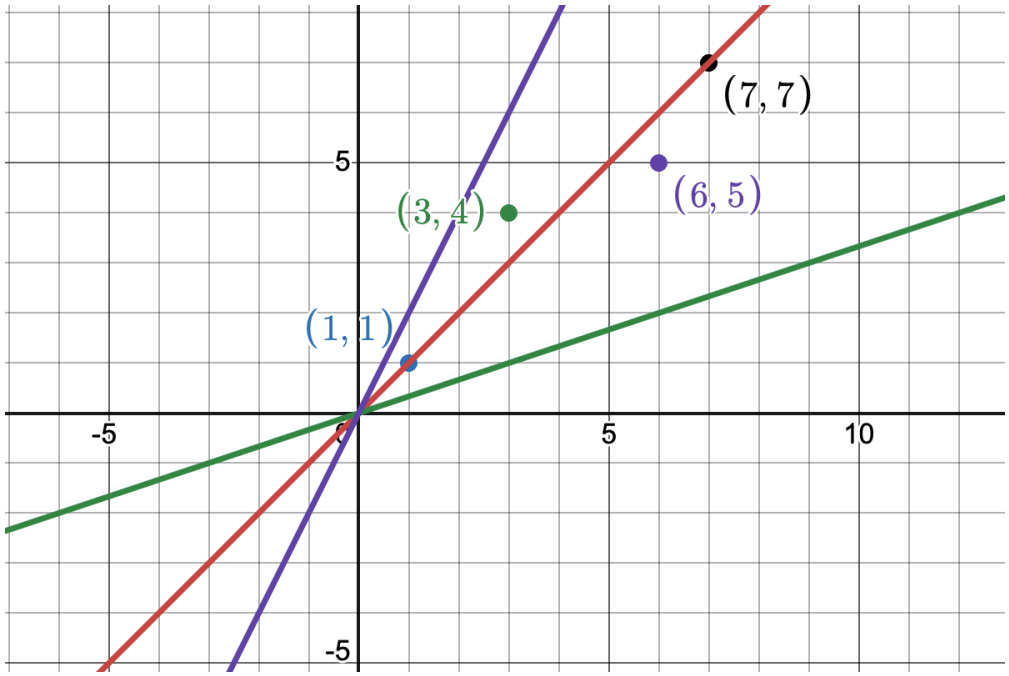

FIGURE 1: Suppose we observe $n = 4$ pairs in a no-intercept simple linear regression model. Each line is a result of a different estimate for θ: $\hat{\theta} = 1$ (in red), $\hat{\theta} = 2$ (in purple), and $\hat{\theta} = \frac{1}{3}$ (in green).

3.1 Fundamentals

Definition 1 (Predictive regression)**.** The task of estimating the conditional expectation of the outcome as function of the predictors: $\mu(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}]$, where $\mu(\vec{x})$ is the (theoretical) prediction function.

- We "regress" Y on $\vec{X}$.

Definition 2 (Homoskedastic)**.** Let $\sigma^2(\vec{x}) = \text{Var}[Y \mid \vec{X} = \vec{x}]$. If $\sigma^2(\vec{x})$ does not vary with $\vec{x}$ (i.e., if the conditional variance $\sigma^2(\vec{x})$ is some constant σ^2), then the regression error is homoskedastic. Otherwise, it is heteroskedastic.

- We often assume homoskedasticity as it simplifies notation and allows for more inferential tasks.

Definition 3 (Regression error)**.** The difference between the random outcome and the theoretical prediction as a function of the predictors: $U(\vec{x}) = Y - \mathbb{E}[Y \mid \vec{X} = \vec{x}]$.

• This implies $\sigma^2(\vec{x}) = \text{Var}[U(\vec{x}) \mid \vec{X} = \vec{x}]$ since $\text{Var}[U(\vec{x}) \mid \vec{X} = \vec{x}] = \text{Var}[Y - \mathbb{E}[Y \mid \vec{X} = \vec{x}] \mid \vec{X} = \vec{x}] = \text{Var}[Y \mid \vec{X} = \vec{x}] = \sigma^2(\vec{x})$.

• Suppose we assume $Y_i = \theta X_i + \varepsilon_i$ and $\mathbb{E}[\varepsilon_i \mid \vec{X} = \vec{x}] = 0$ (so that $\mathbb{E}[Y_i \mid X_i = x_i] = \theta x_i$). By rearranging, we have $\varepsilon_i = Y_i - \theta X_i = Y_i - \mathbb{E}[Y_i \mid X_i]$, which we can think of as a function of X_i to allow for heteroskedasticity. In the model, the conditional variance of Y_i is from the conditional variance of ε_i since, conditionally, θX_i is a constant.

• ☣: If the model is misspecified (e.g., if the conditional expectation of the outcome isn't actually linear), then regression error still exists as $Y - \mathbb{E}[Y \mid \vec{X} = \vec{x}]$, but this isn't equal to the (assumed) error term in the model: $Y_i - \mathbb{E}[Y_i \mid X_i] \neq Y_i - \theta X_i$. These are the same when we assume our model is correctly specified (which we often do), but don't mix up the two!

• ☣: The notation is a bit misleading since the regression error $U(\vec{X})$ isn't perfectly deterministic given $\vec{X} = \vec{x}$, so we can't "take out what's known." This is best illustrated by the story behind the simple model. Even after $X_i = x_i$, there is still random noise (from ε_i) surrounding Y_i that prevents it from being perfectly deterministic.

Definition 4 (Signal-noise decomposition)**.** By rearranging, we can decompose Y into the signal (i.e., the theoretical prediction $\mu(\vec{x})$) and the noise (i.e., the regression error $U(\vec{x})$): $Y = \mu(\vec{x}) + U(\vec{x})$.

• Again, recall the model $Y_i = \theta X_i + \varepsilon_i$, where θX_i is the signal and ε_i is the noise.

Definition 5 (Properties of regression error)**.** As a random variable, regression error has an expectation of 0 (conditionally and unconditionally) and a covariance of 0 with each predictor: $\mathbb{E}[U(\vec{X}) \mid \vec{X} = \vec{x}] = 0$, $\mathbb{E}[U(\vec{X})] = 0$, and $\text{Cov}[U(\vec{X}), X_j] = 0 \ \forall \ X_j$.

Proof.

$$\begin{aligned}
\mathbb{E}[U(\vec{X}) \mid \vec{X} = \vec{x}] &= \mathbb{E}[U(\vec{x}) \mid \vec{X} = \vec{x}] && \text{by conditioning} \\
&= \mathbb{E}[Y - \mathbb{E}[Y \mid \vec{X} = \vec{x}] \mid \vec{X} = \vec{x}] && \text{by definition of } U(\vec{x}) \\
&= \mathbb{E}[Y \mid \vec{X} = \vec{x}] - \mathbb{E}[Y \mid \vec{X} = \vec{x}] && \text{by linearity} \\
&= 0 && \text{by simplifying}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[U(\vec{X})] &= \mathbb{E}[\mathbb{E}[U(\vec{X}) \mid \vec{X}]] && \text{by Adam's Law} \\
&= \mathbb{E}[0] && \text{since } \mathbb{E}[U(\vec{X}) \mid \vec{X} = \vec{x}] = 0 \\
&= 0 && \text{by linearity}
\end{aligned}$$

$$\begin{aligned}
\text{Cov}[U(\vec{X}), X_j] &= \mathbb{E}[U(\vec{X})X_j] - \mathbb{E}[U(\vec{X})]\mathbb{E}[X_j] && \text{by definition} \\
&= \mathbb{E}[U(\vec{X})X_j] && \text{since } \mathbb{E}[U(\vec{X})] = 0 \\
&= \mathbb{E}[\mathbb{E}[U(\vec{X})X_j \mid \vec{X}]] && \text{by Adam's Law} \\
&= \mathbb{E}[X_j\mathbb{E}[U(\vec{X}) \mid \vec{X}]] && \text{by taking out what's known} \\
&= \mathbb{E}[X_j 0] && \text{since } \mathbb{E}[U(\vec{X}) \mid \vec{X} = \vec{x}] = 0 \\
&= 0 && \text{by simplifying}
\end{aligned}$$

□

Definition 6 (Variance of outcome)**.** Conditionally, we already defined $\text{Var}[Y \mid \vec{X} = \vec{x}] = \sigma^2(\vec{x})$. Unconditionally, $\text{Var}[Y] = \text{Var}[\mu(\vec{X})] + \text{Var}[U(\vec{X})]$.

• Again, recall the model $Y_i = \theta X_i + \varepsilon_i$, where θX_i is the signal and ε_i is the noise. We often "don't care" about the distribution of X, but it is still a random variable at the end of the day (with its own variance). Thus, unconditionally, $\text{Var}[Y_i] = \text{Var}[\theta X_i + \varepsilon_i] = \text{Var}[\theta X_i] + \text{Var}[\varepsilon_i] + 2\text{Cov}[\theta X_i, \varepsilon_i] = \text{Var}[\theta X_i] + \text{Var}[\varepsilon_i]$ since the signal and the noise have covariance of 0.

Proof.

$$\begin{aligned}
\text{Var}[U(\vec{X})] &= \mathbb{E}[\text{Var}[U(\vec{X}) \mid \vec{X}]] + \text{Var}[\mathbb{E}[U(\vec{X}) \mid \vec{X}]] && \text{by Eve's Law} \\
&= \mathbb{E}[\text{Var}[U(\vec{X}) \mid \vec{X}]] + \text{Var}[0] && \text{since } \mathbb{E}[U(\vec{X}) \mid \vec{X} = \vec{x}] = 0 \\
&= \mathbb{E}[\text{Var}[U(\vec{X}) \mid \vec{X}]] && \text{by bilinearity} \\
&= \mathbb{E}[\sigma^2(\vec{X})] && \text{since } \sigma^2(\vec{x}) = \text{Var}[U(\vec{x}) \mid \vec{X} = \vec{x}]
\end{aligned}$$

$$\begin{aligned}
\text{Var}[Y] &= \mathbb{E}[\text{Var}[Y \mid \vec{X}]] + \text{Var}[\mathbb{E}[Y \mid \vec{X}]] && \text{by Eve's Law} \\
&= \mathbb{E}[\sigma^2(\vec{X})] + \text{Var}[\mu(\vec{X})] && \text{by definition of } \sigma^2(\vec{X}) \text{ and } \mu(\vec{X})
\end{aligned}$$

By substituting, we have $\text{Var}[Y] = \text{Var}[\mu(\vec{X})] + \text{Var}[U(\vec{X})]$. For some intuition, we've decomposed the variance in the outcome into the variance in the signal plus the variance in the noise. □

Definition 7 (R^2 statistic)**.** The share of the variation in Y accounted for by the variation in the theoretical prediction: $R^2 = \frac{\text{Var}[\mu(\vec{X})]}{\text{Var}[Y]} = \frac{\text{Var}[Y] - \text{Var}[U(\vec{X})]}{\text{Var}[Y]} = 1 - \frac{\text{Var}[U(\vec{X})]}{\text{Var}[Y]}$.

• If R^2 is close to 1, then very little variation in the outcome is due to the regression error (i.e., random noise), so the model explains the data well.

Concept Checker 1. Suppose we correctly specify the model as $Y_i = \theta X_i + \varepsilon_i$, where $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Assume homoskedasticity (i.e., $\sigma_i^2(x_i) = \sigma^2 \ \forall \ x_i \in \mathbb{R}$ and $i \in \{1, \ldots, n\}$). What is R^2?

Solution

By definition, $R^2 = \frac{\text{Var}[\mu_i(X_i)]}{\text{Var}[Y_i]}$. For the numerator, $\mu(X_i) = \mathbb{E}[Y_i \mid X_i] = \mathbb{E}[\theta X_i + \varepsilon_i \mid X_i] = \theta X_i + \mathbb{E}[\varepsilon_i \mid X_i] = \theta X_i$, so $\text{Var}[\mu_i(X_i)] = \text{Var}[\theta X_i] = \theta^2$. For the denominator, $\text{Var}[Y_i] = \text{Var}[\mu_i(X_i)] + \text{Var}[U_i(X_i)] = \theta^2 + \sigma^2$. Thus, $R^2 = \frac{\theta^2}{\theta^2+\sigma^2}$.

Definition 8 (Linear regression model)**.** A model where the conditional expectation of the outcome is a linear function of the parameters (i.e., "linear in the parameters," not necessarily in the predictors): $\mu(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}] = \theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K$, where $\vec{\theta} = (\theta_0, \ldots, \theta_K)^\top$ is the vector of parameters/regression coefficients.

• In order to predict Y given $\vec{X} = \vec{x}$, we must estimate $\vec{\theta}$ with $\hat{\vec{\theta}}$. Once we do, our estimator (before $\vec{X}$ crystallizes) is $\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K$.

Concept Checker 2. Which of the following models are linear in their parameters?

1. $Y_i = \theta X_i + \varepsilon_i$
2. $Y_i = \theta_0 + \theta_1 X_i^{\theta_2} + \varepsilon_i$
3. $Y_i = \theta_0 + \theta_1 X_{i,1} + \theta_2 X_{i,2} + \theta_3 X_{i,1} X_{i,2} + \varepsilon_i$
4. $Y_i = \theta \sin(X_i) + \varepsilon_i$

Solution

All models except 2 are linear in their parameters. Even though $X_{i,1}X_{i,2}$ and $\sin(X_i)$ are not linear functions, those are for the predictors. We are concerned about whether the parameters are linear.

Definition 9 (Residual)**.** The difference between the true outcome and the predicted outcome: $\hat{U}(\vec{X}) = Y - \hat{Y} = Y - (\hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K)$.

• Residuals are orthogonal to the predictors, so in the no-intercept simple linear regression model, $\sum_{i=1}^n X_i \hat{U}_i(X_i) = 0$.

• ☣: Regression error is unobservable (as a theoretical quantity) while residual is observable (as a statistic). Don't mix up the two!

Concept Checker 3. Which of the following are equivalent?

1. $U_i(\vec{x}_i)$
2. $\mu_i(\vec{x}_i)$
3. $\mathbb{E}[Y_i \mid \vec{X}_i = \vec{x}_i]$
4. $Y_i - \mu_i(\vec{x}_i)$
5. $\hat{Y}_i$
6. $\hat{\theta}_0 + \hat{\theta}_1 X_{i,1} + \ldots + \hat{\theta}_K X_{i,K}$
7. $\theta_0 + \theta_1 X_{i,1} + \ldots + \theta_K X_{i,K}$
8. $\vec{X}_i \vec{\theta}$, where $\vec{X}_i$ is $1 \times (K+1)$ and $\vec{\theta}$ is $(K+1) \times 1$.

9. $\sigma_i^2(\vec{x}_i)$
10. $\text{Var}[Y_i]$
11. $\text{Var}[Y_i \mid \vec{X}_i = \vec{x}_i]$
12. $\text{Var}[U_i(\vec{x}_i) \mid \vec{X}_i = \vec{x}_i]$
13. $\text{Var}[\mu_i(\vec{X}_i)] + \text{Var}[U_i(\vec{X}_i)]$.

Solution

1 and 4: $U_i(\vec{x}_i) = Y_i - \mu_i(\vec{x}_i)$.
2 and 3: $\mu_i(\vec{x}_i) = \mathbb{E}[Y_i \mid \vec{X}_i = \vec{x}_i]$.
5 and 6: $\hat{Y}_i = \hat{\theta}_0 + \hat{\theta}_1 X_{i,1} + \ldots + \hat{\theta}_K X_{i,K}$.
7 and 8: $\theta_0 + \theta_1 X_{i,1} + \ldots + \theta_K X_{i,K} = \vec{X}_i \vec{\theta}$, where $\vec{X}_i$ is $1 \times (K+1)$ and $\vec{\theta}$ is $(K+1) \times 1$.
9, 11, and 12: $\sigma_i^2(\vec{x}_i) = \text{Var}[Y_i \mid \vec{X}_i = \vec{x}_i] = \text{Var}[U_i(\vec{x}_i) \mid \vec{X}_i = \vec{x}_i]$.
10 and 13: $\text{Var}[Y_i] = \text{Var}[\mu_i(\vec{X}_i)] + \text{Var}[U_i(\vec{X}_i)]$.

3.2 No-Intercept Simple Linear Regression Model

Definition 10 (Model)**.** $Y_i = \theta X_i + \varepsilon_i$. We will show three different approaches to estimate θ, which all result in the same estimator!

Definition 11 (Method of moments estimator)**.** The estimator where the theoretical moments are replaced with their sample analogues: $\hat{\theta}_{\text{MOM}} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$.

Proof.

$$\begin{aligned}
\mathbb{E}[X_i Y_i] &= \mathbb{E}[\mathbb{E}[X_i Y_i \mid X_i]] && \text{by Adam's Law} \\
&= \mathbb{E}[X_i \mathbb{E}[Y_i \mid X_i]] && \text{by taking out what's known} \\
&= \mathbb{E}[X_i(\theta X_i)] && \text{since } \mathbb{E}[Y_i \mid X_i] = \theta X_i \\
&= \theta \mathbb{E}[X_i^2] && \text{by linearity}
\end{aligned}$$

This implies $\theta = \frac{\mathbb{E}[X_i Y_i]}{\mathbb{E}[X_i^2]}$, which implies $\hat{\theta}_{\text{MOM}} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\frac{1}{n}\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$. □

Definition 12 (Ordinary least squares estimator)**.** The estimator that minimizes the sum of the squared errors: $\hat{\theta}_{\text{OLS}} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$.

Proof. By definition, $\hat{\theta}_{\text{OLS}} = \arg\min_{\theta \in \mathbb{R}} \left(\sum_{i=1}^n \varepsilon_i^2\right) = \arg\min_{\theta \in \mathbb{R}} \left(\sum_{i=1}^n (Y_i - \theta X_i)^2\right)$.

$$\begin{aligned}
\frac{\partial}{\partial\theta}\left(\sum_{i=1}^{n}(Y_i-\theta X_i)^2\right) &= \sum_{i=1}^{n}\frac{\partial}{\partial\theta}(Y_i-\theta X_i)^2 && \text{by linearity}\\
&= \sum_{i=1}^{n}2(Y_i-\theta X_i)(-X_i) && \text{by chain rule}\\
&= 2\left(-\sum_{i=1}^{n}X_iY_i+\theta\sum_{i=1}^{n}X_i^2\right) && \text{by simplifying}\\
\frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^{n}(Y_i-\theta X_i)^2\right) &= \frac{\partial}{\partial\theta}\left(2\left(-\sum_{i=1}^{n}X_iY_i+\theta\sum_{i=1}^{n}X_i^2\right)\right) && \text{by differentiating}\\
&= \sum_{i=1}^{n}X_i^2>0 && \text{by differentiating}
\end{aligned}$$

Thus, the value for θ at which the first derivative is equal to 0 is a minimum. By setting it equal to 0 and rearranging, we get $\theta=\frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2}$, which implies $\hat{\theta}_{\text{OLS}}=\frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2}$. □

Definition 13 (Maximum likelihood estimator)**.** Assume $Y_i \mid X_i \overset{\text{ind.}}{\sim} \mathcal{N}(\theta X_i,\sigma^2)$, with θ and σ^2 unknown. The estimator that maximizes the likelihood (i.e., the most likely value for θ, given the data): $\hat{\theta}_{\text{MLE}}=\frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2}$.

Proof. By definition, $\hat{\theta}_{\text{MLE}}=\arg\max_{\theta\in\mathbb{R}}\mathcal{L}(\theta;\vec{Y})=\arg\max_{\theta\in\mathbb{R}}\ell(\theta;\vec{Y})$.

$$\begin{aligned}
\mathcal{L}(\theta,\sigma^2;\vec{Y}) &= f_{\vec{Y}}(\vec{Y};\theta,\sigma^2) && \text{by definition of likelihood}\\
&= \prod_{i=1}^{n}f_{Y_i}(Y_i;\theta,\sigma^2) && \text{by conditional independence}\\
&= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(Y_i-\theta X_i)^2\right) && \text{by Normal PDF}\\
&= \prod_{i=1}^{n}\frac{1}{\sqrt{\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(Y_i-\theta X_i)^2\right) && \text{by removing multiplicative constants}\\
&= \frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\theta X_i)^2\right) && \text{by product}\\
\ell(\theta,\sigma^2;\vec{Y}) &= \log\left(\frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\theta X_i)^2\right)\right) && \text{by definition of log-likelihood}\\
&= -\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\theta X_i)^2 && \text{by log}
\end{aligned}$$

As a function of θ, we're subtracting a constant by $\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\theta X_i)^2$. To maximize, we want to subtract by as little as possible, so maximizing the log-likelihood with respect to θ

is equivalent to minimizing $\sum_{i=1}^n (Y_i - \theta X_i)^2$, which is what the OLS estimator does! Thus, $\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$. □

Definition 14 (Properties of $\hat{\theta}$)**.** Assume the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ have conditionally independent outcomes—i.e., $(Y_i \perp\!\!\!\perp Y_j) \mid \vec{X}\ \forall\ i \neq j$. Recall $\mu_i(x_i) = \mathbb{E}[Y_i \mid X_i = x_i]$ and $\sigma_i^2(x_i) = \text{Var}[Y_i \mid X_i = x_i]$. We have $\mathbb{E}[\hat{\theta} \mid X_i = x_i] = \frac{\sum_{i=1}^n x_i \mu_i(x_i)}{\sum_{i=1}^n x_i^2}$ and $\text{Var}[\hat{\theta} \mid X_i = x_i] = \frac{\sum_{i=1}^n x_i^2 \sigma_i^2(x_i)}{\left(\sum_{i=1}^n x_i^2\right)^2}$.

• If $\mu_i(x_i) = \theta x_i$ (which is the case if $Y_i = \theta X_i + \varepsilon_i$ is correctly specified), then $\mathbb{E}[\hat{\theta} \mid X_i = x_i] = \theta$, so $\hat{\theta}$ is conditionally unbiased.

• If $\sigma_i^2(x_i) = \sigma^2$ (which is the case if $Y_i = \theta X_i + \varepsilon_i$ is correctly specified with ε_i homoskedastic), then $\text{Var}[\hat{\theta} \mid X_i = x_i] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$.

• If $Y_i \mid X_i \overset{\text{ind.}}{\sim} \mathcal{N}(\theta X_i, \sigma^2)$ (which is the setup for the MLE), then $\hat{\theta} \mid X_i = x_i \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$ and $\text{Var}[\hat{\theta} \mid X_i = x_i] = \frac{1}{\mathcal{I}_{\vec{Y}}(\theta)}$, so $\hat{\theta}$ conditionally achieves the Cramér-Rao lower bound.

Concept Checker 4. Suppose we correctly specify the model as $Y_i = \theta X_i + \varepsilon_i$. Show $\sum_{i=1}^n X_i \hat{U}_i(X_i) = 0$.

Solution

First, notice the following.

$$\begin{aligned}
\sum_{i=1}^n \hat{\theta} X_i^2 &= \hat{\theta} \sum_{i=1}^n X_i^2 && \text{by linearity} \\
&= \left(\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\right) \sum_{i=1}^n X_i^2 && \text{since } \hat{\theta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} \\
&= \sum_{i=1}^n X_i Y_i && \text{by simplifying}
\end{aligned}$$

Next, notice the following.

$$\begin{aligned}
\sum_{i=1}^{n} X_i \hat{U}_i(X_i) &= \sum_{i=1}^{n} X_i(Y_i - \hat{\theta} X_i) && \text{since } \hat{U}_i(X_i) = Y_i - \hat{\theta} X_i \\
&= \sum_{i=1}^{n} X_i Y_i - \hat{\theta} X_i^2 && \text{by distributing} \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \hat{\theta} X_i^2 && \text{by linearity} \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i Y_i && \text{since } \sum_{i=1}^{n} \hat{\theta} X_i^2 = \sum_{i=1}^{n} X_i Y_i \\
&= 0 && \text{by simplifying}
\end{aligned}$$

For some intuition, if the residuals were not orthogonal to the predictors, then this would imply we haven't used all of the information the predictors can tell us about the outcome.

3.3 Extensions to Other Setups

Setup	Model	Estimator
No-intercept simple linear regression	$Y_i = \theta X_i + \varepsilon_i$	$\hat{\theta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$
Simple linear regression	$Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i$	$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$, $\hat{\theta}_0 = \bar{Y} - \hat{\theta}_1 \bar{X}$
Matrix form with K predictors	$\vec{Y} = \mathbf{X}\vec{\theta} + \vec{\varepsilon}$	$\hat{\vec{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \vec{Y}$

3.4 Logistic Regression

Definition 15 (Logit function)**.** $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ for $0 < p < 1$, where $\frac{p}{1-p}$ is the odds.

- $\text{logit} : (0, 1) \to \mathbb{R}$.

Definition 16 (Logistic function)**.** $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$ for $x \in \mathbb{R}$.

- $\text{logit}^{-1} : \mathbb{R} \to (0, 1)$.

Definition 17 (Logistic regression model)**.** Let Y be binary and $\mu(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}] = P(Y = 1 \mid \vec{X} = \vec{x})$. Logistic regression models the logit of the conditional expectation of the outcome as a linear function of the parameters: $\text{logit}(\mu(\vec{x})) = \log\left(\frac{\mu(\vec{x})}{1-\mu(\vec{x})}\right) = \theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K$.

- By applying logit^{-1} to both sides, we have $\mu(\vec{x}) = \text{logit}^{-1}(\theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K) = \frac{\exp(\theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K)}{1 + \exp(\theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K)}$, so our estimator (before $\vec{X}$ crystallizes) is $\hat{\mu}(\vec{X}) = \text{logit}^{-1}(\hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K) = \frac{\exp\left(\hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K\right)}{1 + \exp\left(\hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K\right)}$.

Concept Checker 5. With a binary outcome, an alternative to the logistic model is the probit model: $\Phi^{-1}(\mu(\vec{x})) = \theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K$. First, why is this valid? Next, if we find estimators $\hat{\theta}_0, \ldots, \hat{\theta}_K$, how can we estimate $\mu(\vec{x})$?

Solution

First, this is valid because, like the logit function, $\Phi^{-1}:(0,1)\to\mathbb{R}$ and $\Phi:\mathbb{R}\to(0,1)$. Next, by applying Φ to both sides, we have $\mu(\vec{x})=\Phi(\theta_0+\theta_1x_1+\ldots+\theta_Kx_K)$, so our estimator (before $\vec{X}$ crystallizes) is $\hat{\mu}(\vec{X})=\Phi(\hat{\theta}_0+\hat{\theta}_1X_1+\ldots+\hat{\theta}_KX_K)$.

4 Descriptive Regression

Idea. In *descriptive regression*, we no longer condition on X (one-dimensional for simplicity). Instead, we assume (X,Y) follows some joint distribution $F_{X,Y}$, which we wish to learn about. Incredibly, despite a different way of thinking, we arrive at the same estimator as before: $\hat{\theta}=\frac{\sum_{i=1}^n X_iY_i}{\sum_{i=1}^n X_i^2}$.

Definition 18 (Descriptive regression)**.** $\beta_{Y\sim X}=\frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}$. One way to think about this summary measure to find θ where $(\alpha,\theta)=\arg\min_{(a,b)\in\mathbb{R}^2}\mathbb{E}\left[(Y-(a+bX))^2\right]$, which results in $\alpha=\mathbb{E}[Y]-\theta\mathbb{E}[X]$ and $\theta=\frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}=\beta_{Y\sim X}$.

• Thus, $\alpha+\theta X$ best "mimics" Y in the sense that any other linear mimic $a+bX$ will have a larger expected square error.

• By adding θX to both sides of $\alpha=\mathbb{E}[Y]-\theta\mathbb{E}[X]$, we have $\alpha+\theta X=\mathbb{E}[Y]+\beta_{Y\sim X}(X-\mathbb{E}[X])$, where the right-hand side is the linear projection.

Definition 19 (Linear projection)**.** Assume X, Y have finite variance. The linear projection of Y on X at $X=x$ is $\mathrm{LP}(Y\mid X=x)=\mathbb{E}[Y]+\beta_{Y\sim X}(x-\mathbb{E}[X])$.

• $\mathrm{LP}(Y\mid X)$ is not the conditional expectation in general. It is the best linear function of x for approximating Y.

Definition 20 (Estimator for θ)**.** Like before, let $\mu(x)=\mathbb{E}[Y\mid X=x]$. It can be shown $\beta_{Y\sim X}=\frac{\mathrm{Cov}[\mu(X),X]}{\mathrm{Var}[X]}=\beta_{\mu(X)\sim X}$. Let $(X_i,Y_i)\overset{\text{i.i.d.}}{\sim}F_{X,Y}$. If $\mathbb{E}[Y]=\mathbb{E}[X]=0$, then our estimand is $\theta=\frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}=\frac{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2]-(\mathbb{E}[X])^2}=\frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}$. With a method of moments approach, $\hat{\theta}=\frac{\frac{1}{n}\sum_{i=1}^n X_iY_i}{\frac{1}{n}\sum_{i=1}^n X_i^2}=\frac{\sum_{i=1}^n X_iY_i}{\sum_{i=1}^n X_i^2}$.

5 Practice Problems

Problem 1. Jerry loves predicting things: stock returns, basketball scores, and whether his friends will like his K-pop dances. He approaches these problems via the conditional mean given some predictors, $\mu(x)=\mathbb{E}[Y\mid X=x]$, which he approximates with a parametric model $\mu(x\mid\theta)$ for scalar θ. Suppose Jerry observes $(X_1,Y_1),\ldots,(X_n,Y_n)$ and uses the model

$$Y_j\mid(X_1=x_1,\ldots,X_n=x_n)\overset{\text{ind.}}{\sim}\mathcal{N}(\mu(x_j\mid\theta),\sigma^2),\quad j=1,\ldots,n,$$

where σ^2 is **known**. Let $\hat{\theta}$ denote the MLE for θ.

(a) Why does it make sense to use a *conditional* likelihood here rather than the joint likelihood over (X_j,Y_j)?

(b) Show that the conditional log-likelihood simplifies to

$$\ell(\theta) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}\{Y_j - \mu(x_j \mid \theta)\}^2.$$

(c) Find the score $s(\theta) = \ell'(\theta)$.

(d) Find the Fisher information $\mathcal{I}(\theta)$ for θ in the sample. What is the approximate variance of $\hat{\theta}$?

(e) State the asymptotic distribution of $\hat{\theta}$.

(f) Using the delta method, find the asymptotic distribution of $\mu(x_j \mid \hat{\theta})$ as an estimator of $\mu(x_j \mid \theta)$.

(g) Jerry wants to predict Y_{n+1} for a new individual with $X_{n+1} = x_{n+1}$. His prediction is $\mu(x_{n+1} \mid \hat{\theta})$ and his prediction error is $Y_{n+1} - \mu(x_{n+1} \mid \hat{\theta})$. Find the approximate distribution of his prediction error. *Hint*: Add and subtract $\mu(x_{n+1} \mid \theta)$.

Solution

(a) In predictive regression, we are interested in modeling the conditional distribution of Y given X, that is, we want to use X to predict Y but not model how X itself is distributed. Using the conditional likelihood avoids having to specify a model for X, which may be complicated. This focuses inference exactly on the relationship between X and Y (what we actually care about).

(b) Given $X_j = x_j$, the conditional density of Y_j is $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(Y_j-\mu(x_j|\theta))^2}{2\sigma^2}\right)$. Taking the log and summing over j:

$$\ell(\theta) = \sum_{j=1}^{n}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y_j - \mu(x_j \mid \theta))^2}{2\sigma^2}\right] = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\{Y_j - \mu(x_j \mid \theta)\}^2$$

Dropping the constant term gives the result.

(c) Differentiating with respect to θ:

$$s(\theta) = \frac{1}{\sigma^2}\sum_{j=1}^{n}\frac{\partial\mu(x_j \mid \theta)}{\partial\theta}\{Y_j - \mu(x_j \mid \theta)\}.$$

(d) Differentiating the score:

$$s'(\theta) = -\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(\frac{\partial\mu(x_j \mid \theta)}{\partial\theta}\right)^2 + \frac{1}{\sigma^2}\sum_{j=1}^{n}\frac{\partial^2\mu(x_j \mid \theta)}{\partial\theta^2}\{Y_j - \mu(x_j \mid \theta)\}.$$

The second term has expectation 0 since $\mathbb{E}[Y_j - \mu(x_j \mid \theta) \mid X_j = x_j] = 0$ by definition of μ. By the information equality, $\mathcal{I}(\theta) = -\mathbb{E}[s'(\theta)] = \frac{1}{\sigma^2}\sum_{j=1}^{n}\left(\frac{\partial\mu(x_j|\theta)}{\partial\theta}\right)^2$. The approximate variance of $\hat{\theta}$ is $\mathcal{I}(\theta)^{-1}$.

(e) By the asymptotic normality of the MLE:

$$\hat{\theta} - \theta \,\dot{\sim}\, \mathcal{N}\big(0,\, \mathcal{I}(\theta)^{-1}\big)\,.$$

(f) Let $\mu'(x_j \mid \theta) = \frac{\partial \mu(x_j|\theta)}{\partial \theta}$. By the delta method applied to $g(\theta) = \mu(x_j \mid \theta)$:

$$\mu(x_j \mid \hat{\theta}) - \mu(x_j \mid \theta) \,\dot{\sim}\, \mathcal{N}\big(0,\, \mathcal{I}(\theta)^{-1} \cdot [\mu'(x_j \mid \theta)]^2\big)\,.$$

(g) Using the hint, decompose:

$$Y_{n+1} - \mu(x_{n+1} \mid \hat{\theta}) = \underbrace{\{Y_{n+1} - \mu(x_{n+1} \mid \theta)\}}_{\text{irreducible noise}} + \underbrace{\{\mu(x_{n+1} \mid \theta) - \mu(x_{n+1} \mid \hat{\theta})\}}_{\text{estimation error}}.$$

These two terms are independent, since the first depends only on Y_{n+1} (our new predicted observation) while the second depends on $\hat{\theta}$ (computed from the training data). The first term is $\mathcal{N}(0, \sigma^2)$ by our model assumption. By part (f), the second is approximately $\mathcal{N}(0,\, \mathcal{I}(\theta)^{-1}[\mu'(x_{n+1} \mid \theta)]^2)$. Combining:

$$Y_{n+1} - \mu(x_{n+1} \mid \hat{\theta}) \,\dot{\sim}\, \mathcal{N}\big(0,\ \sigma^2 + \mathcal{I}(\theta)^{-1}[\mu'(x_{n+1} \mid \theta)]^2\big)\,.$$

Notice the prediction uncertainty has two components: σ^2 from the irreducible noise in Y_{n+1}, and the estimation error in $\hat{\theta}$.

Problem 2. Suppose

$$Y \mid (X = x) \sim \text{Expo}\big(\mu(x)^{-1}\big)\,, \quad \text{where } \mu(x) = \mathbb{E}[Y \mid X = x] = e^{\theta x}.$$

Assume $(X_1, Y_1), \ldots, (X_n, Y_n)$ yields outcomes conditionally independent given predictors. Let (x_j, y_j) be the observed value of (X_j, Y_j). Assume the x_j's are not all 0.

(a) Write down the conditional log-likelihood $\ell(\theta)$.

(b) Find the score $s(\theta)$ and show that the log-likelihood is *strictly concave*. What does this imply about the MLE?

(c) Find the Fisher information $\mathcal{I}(\theta)$ in the sample.

(d) Construct a nominal 95% confidence interval for θ in terms of $\hat{\theta}$ and the data.

Solution

(a) For $Y_j \mid X_j = x_j \sim \text{Expo}(\mu(x_j)^{-1})$, the conditional density is $f(y_j \mid x_j; \theta) = \frac{1}{\mu(x_j)} e^{-y_j/\mu(x_j)}$. Plugging in $\mu(x_j) = e^{\theta x_j}$:

$$\ell(\theta) = \sum_{j=1}^{n} \log f(y_j \mid x_j; \theta) = -\sum_{j=1}^{n} \log \mu(x_j) - \sum_{j=1}^{n} \frac{y_j}{\mu(x_j)} = -\theta \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} y_j e^{-\theta x_j}.$$

(b) Differentiating:

$$s(\theta) = -\sum_{j=1}^{n} x_j + \sum_{j=1}^{n} x_j y_j e^{-\theta x_j}.$$

The second derivative is

$$\ell''(\theta) = -\sum_{j=1}^{n} x_j^2 y_j e^{-\theta x_j} < 0,$$

since $y_j > 0$, $x_j^2 \geq 0$ (with strict inequality for at least one j), and the exponential is always positive. Since $\ell''(\theta) < 0$ for all θ, the log-likelihood is strictly concave, which implies the MLE is **unique** if it exists that is.

(c) Using the information equality $\mathcal{I}(\theta) = -\mathbb{E}[\ell''(\theta)]$:

$$\mathcal{I}(\theta) = \sum_{j=1}^{n} x_j^2 e^{-\theta x_j} \mathbb{E}[Y_j] = \sum_{j=1}^{n} x_j^2 e^{-\theta x_j} \cdot e^{\theta x_j} = \sum_{j=1}^{n} x_j^2.$$

Alternatively, using $\mathcal{I}(\theta) = \mathrm{Var}(s(\theta; \vec{Y}))$ and the fact that $\mathrm{Var}(Y_j \mid X_j = x_j) = \mu(x_j)^2 = e^{2\theta x_j}$:

$$\mathcal{I}(\theta) = \sum_{j=1}^{n} x_j^2 e^{-2\theta x_j} \mathrm{Var}(Y_j) = \sum_{j=1}^{n} x_j^2 e^{-2\theta x_j} \cdot e^{2\theta x_j} = \sum_{j=1}^{n} x_j^2.$$

Sanity check! The Fisher information depends only on the predictors, not on θ.

(d) By asymptotic normality of the MLE, $\hat{\theta} \overset{\cdot}{\sim} \mathcal{N}(\theta, \mathcal{I}(\theta)^{-1})$. Plugging in $\hat{\theta}$ for θ (plug-in principle) and $\mathcal{I}(\hat{\theta}) = \sum_{j=1}^{n} x_j^2$:

$$\hat{\theta} \pm \frac{1.96}{\sqrt{\sum_{j=1}^{n} x_j^2}}.$$

Notice the width of the confidence interval shrinks as the x_j's grow in magnitude, so when our predictors are more spread out we have more information about θ.

Problem 3. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown.

(a) Simplify the likelihood and identify a *two-dimensional* sufficient statistic for (μ, σ^2).

(b) We know $\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$, where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Show that $\bar{Y}$ and $\hat{\sigma}^2$ are independent. *Hint*: Consider the MVN vector $(\bar{Y}, Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})$ and compute $\mathrm{Cov}(\bar{Y}, Y_j - \bar{Y})$.

(c) Using the results from (b) and the fact that $\frac{\bar{Y}-\mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$, construct an *exact* 95% confidence interval for μ.

(d) Now suppose σ^2 is known. Propose an exact pivot and construct an exact 95% confidence interval for μ.

(e) Compare the widths of the confidence intervals in (c) and (d). When n is large, do they differ much? Why or why not?

Solution

(a) The joint density is

$$f_{\vec{Y}}(\vec{y};\, \mu, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right).$$

Expanding $\sum (y_i - \mu)^2 = \sum (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$, the density becomes

$$\propto \frac{1}{\sigma^n} \exp\left(-\frac{\sum (y_i - \bar{y})^2}{2\sigma^2} \right) \exp\left(-\frac{n(\bar{y} - \mu)^2}{2\sigma^2} \right).$$

The data enter only through $\bar{y}$ and $\sum (y_i - \bar{y})^2$, so $\left(\bar{Y},\, \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right)$ is a two-dimensional sufficient statistic for (μ, σ^2).

(b) The vector $(\bar{Y}, Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})$ is an affine transformation of $(Y_1, \ldots, Y_n)$, hence MVN. For any j:

$$\mathrm{Cov}(\bar{Y}, Y_j - \bar{Y}) = \mathrm{Cov}(\bar{Y}, Y_j) - \mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.$$

Since uncorrelated jointly Normal random variables are independent, $\bar{Y}$ is independent of every component $Y_j - \bar{Y}$, hence independent of any function of them including $\hat{\sigma}^2$.

(c) Since $\frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$ exactly, we have the pivot

$$1 - \alpha = P\left(Q_{t_{n-1}}\left(\tfrac{\alpha}{2}\right) \le \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \le Q_{t_{n-1}}\left(1 - \tfrac{\alpha}{2}\right) \right).$$

By symmetry of the t-distribution and isolating μ, the exact 95% CI is:

$$\bar{Y} \pm Q_{t_{n-1}}(0.975) \cdot \frac{\hat{\sigma}}{\sqrt{n}}.$$

(d) When σ^2 is known, $\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ exactly. Isolating μ:

$$\bar{Y} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}.$$

(e) The t-interval is wider than the z-interval for any finite n since $Q_{t_{n-1}}(0.975) > 1.96$—the t-distribution has heavier tails and thus is more flexible which reflects the additional uncertainty from estimating σ^2. However, as $n \to \infty$, $t_{n-1} \to \mathcal{N}(0, 1)$ and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, so the two intervals become indistinguishable (and approx. normal!). For very large n, knowing σ^2 provides essentially no practical advantage (Can you think of why?).