Notes on STAT-331: Applied Linear Models

Unversity of Waterloo

David Duan

Last Updated: February 3, 2021 (Draft)

Contents

Chapter 1. Simple Linear Regression

Section 1. Overview

1.1. Suppose we are given a set of data points $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

- How do we characterize the relationship between x and y?
- How do we predict y given x?
- How does the mean of y change when x increases by a?

We can answer questions like these with **simple linear regression** (SLR):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Intuitively, we are assuming that there exists some underlying linear relationship between the **covariate** x_i and the **outcome** y_i, where the **regression coefficients** β_0 and β_1 are unknown. The **error term** ε_i captures the difference between the actual value of y_i and our prediction $\beta_0 + \beta_1 x_i$.

1.2. The model above is "simple" because there is only one explanatory variable x. Suppose now each sample x_i has three covariates x_{i1}, x_{i2}, and x_{i3}. We generalize SLR to **multiple linear regression** (MLR), where each covariate x_{ij} has a corresponding β_j parameter:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

The meaning of y_i and ε_i remain the same; we just have more covariates to work with.

1.3. This course will focus on developing regression models in the following aspects:

- theoretically/mathematically: derive estimators;
- practically: how to fit these models in R;
- how to choose and compare a model, i.e., which covariates to include;
- how to evaluate the appropriateness of the model and assumptions.

Section 2. Simple Linear Regression

1.4. Remark: We make the following assumptions (acronym: LINE):

- **L**inearity: there exists a linear relationship between x and y.
- **I**ndependence: the error terms $\varepsilon_1, \dots, \varepsilon_n$ are independent.
- **N**ormality: the error terms have mean 0.
- **E**qual variance (aka **homoskedasticity**): all error terms share the same variance σ^2.

1.5. Definition: The general form of simple linear regression is given by

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).}$$

- $\beta_0, \beta_1, \sigma^2$: fixed, *unknown* parameters.
- ε_i: *unobserved* random error term.
- y_i, x_i are observed data (we treat x_i as fixed in this course).

Equivalently, we can write

$$\boxed{y_i \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2).}$$

Note here y_i's are independent but no longer have the same distribution because they have different means (depending on x_i).

1.6. Example: How to interpret β_0 and β_1? We make the following observations:

1. $\mathbb{E}[y_i \mid x_i] = \beta_0 + \beta_1 x_i$.
2. $\mathbb{E}[y_i \mid x_i = 0] = \beta_0$.
3. $\mathbb{E}[y_i \mid x_i = x^*] = \beta_0 + \beta_1 x^*$.
4. $\mathbb{E}[y_i \mid x_i = x^* + 1] = \beta_0 + \beta_1(x^* + 1) = \beta_0 + \beta_1 x^* + \beta_1$.
5. $\mathbb{E}[y_i \mid x_i = x^* + 1] - \mathbb{E}[y_i \mid x_i = x^*] = \beta_1$.

Therefore,

- By observation 2, β_0 is the average outcome when $x_0 = 0$.
- By observation 5, β_1 is the expected/average change in y when x moves by 1 unit.

Section 3. SLR: Estimation

1.7. Theorem: *The LS estimators for β_0 and β_1 are given by*

$$\boxed{\begin{aligned}&\hat{\beta}_0^{LS} = \bar{y} - \hat{\beta}_1^{LS}\bar{x}\\ &\hat{\beta}_1^{LS} = \frac{(\sum_i x_i y_i) - n\bar{x}\bar{y}}{(\sum_i x_i^2) - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}\end{aligned}}$$

Proof. The goal is to choose β_0 and β_1 that minimizes the sum of squared errors given by

$$S(\beta_0, \beta_1) := \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Differentiate, set the partial derivatives to 0, and solve for β_0 and β_1:

$$\begin{aligned}\frac{\partial S(\delta_0, \delta_1)}{\partial \beta_0} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)\\ \frac{\partial S(\delta_0, \delta_1)}{\partial \beta_1} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)\end{aligned}$$

$$\begin{aligned}\text{(Set) } 0 &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\\ &= \left(\sum_{i=1}^{n} y_i\right) - n\beta_0 - \left(\beta_1 \sum_{i=1}^{n} x_i\right)\\ \implies \beta_0 &= \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - \beta_1\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \bar{y} - \beta_1\bar{x}\end{aligned}$$

$$\begin{aligned}\text{(Set) } 0 &= \sum_{i=1}^{n} (y_i x_i - \beta_0 x_i - \beta_1 x_i^2)\\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - \left(\beta_0 \sum_{i=1}^{n} x_i\right) - \left(\beta_1 \sum_{i=1}^{n} x_i^2\right)\\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - (\bar{y} - \beta_1\bar{x})n\bar{x} - \left(\beta_1 \sum_{i=1}^{n} x_i^2\right) && \text{plug in previous result}\\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - n\bar{y}\bar{x} + \beta_1 n\bar{x}^2 - \left(\beta_1 \sum_{i=1}^{n} x_i^2\right)\\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - n\bar{y}\bar{x} + \beta_1\left(n\bar{x}^2 - \sum_{i=1}^{n} x_i^2\right)\\ \implies \beta_1 &= \frac{(\sum_{i=1}^{n} y_i x_i) - n\bar{y}\bar{x}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} && \text{See Proposition 1.24}\end{aligned}$$

□

1.8. Theorem: *The ML estimators for β_0 and β_1 coincide with the LS estimators.*

Proof. The **joint likelihood function** of $Y_1, \dots, Y_n$ with $Y_i \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$ is

$$\begin{aligned} L(\beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} f(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right). \end{aligned}$$

The **log-likelihood function** is given by

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Maximizing the log-likelihood is equivalent to solving the following system of equations:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right) = 0 \\ \frac{\partial \ell}{\partial \beta_1} &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right) x_i = 0 \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right)^2 = 0 \end{aligned}$$

Observe solving the first two equations is equivalent to minimizing the sum of squares! In other words, the ML estimators $\hat{\beta}_0^{\text{ML}}, \hat{\beta}_1^{\text{ML}}$ coincide with the LS estimators $\hat{\beta}_0^{\text{LS}}, \hat{\beta}_1^{\text{LS}}$. Therefore, we will remove the superscripts and simply call them $\hat{\beta}_0$ and $\hat{\beta}_1$. □

1.9. Definition: The **fitted values** $\hat{y}_i$ and the **residuals** e_i are given by

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

Note the residuals e_i and the errors $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ are not the same thing.

1.10. Remark: Solving the third equation, we obtain the ML estimator for σ^2:

$$\hat{\sigma}^2_{\text{ML}} = \frac{\sum_{i=1}^{n} e_i^2}{n}.$$

This is slightly different from the unbiased estimator for σ^2 (notice the $n-2$ in the denominator):

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}.}$$

This difference often doesn't matter when $n \geq 50$.

Section 4. SLR: Inference

1.11. Theorem: *The estimator $\hat\beta_1$ follows the Normal distribution with parameters*

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

Proof. Recall $y_i \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Let us rewrite $\hat\beta_1$ as

$$\begin{aligned}\hat\beta_1 = \frac{S_{xy}}{S_{xx}} &= \frac{\sum_i (y_i - \bar y)(x_i - \bar x)}{\sum_i (x_i - \bar x)^2} \\ &= \frac{\sum_i y_i (x_i - \bar x)}{\sum_i (x_i - \bar x)^2} =: \sum_{i=1}^n w_i y_i, \qquad\qquad w_i := \frac{(x_i - \bar x)}{\sum_i (x_i - \bar x)}.\end{aligned}$$

Since we assumed that x_i's are fixed, the variables w_i's are fixed wrt the y_i's. Thus, $\hat\beta_1$ is a linear combination independent Normal random variables $y_1, \dots, y_n$. Moreover, all y_i's share the same variance (homoskedasticity). By the Fact above, β_i follows the normal distribution with parameters

$$\hat\beta_1 \sim N\left(\sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i), \sigma^2 \sum_{i=1}^n w_i^2\right).$$

It remains to simplify the parameters.

$$\begin{aligned}\mathbb{E}[\hat\beta_1] &= \sum_{i=1}^n w_i \left(\beta_0 + \beta_1 x_i\right) \\ &= \sum_{i=1}^n \frac{(x_i - \bar x)}{\sum (x_i - \bar x)^2} \left(\beta_0 + \beta_1 x_i\right) \\ &= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar x)}{\sum (x_i - \bar x)^2} + \beta_1 \frac{\sum_{i=1}^n x_i (x_i - \bar x)}{\sum (x_i - \bar x)^2} \\ &= 0 + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar x)(x_i - \bar x)}{\sum (x_i - \bar x)^2} = \beta_1 \qquad\qquad \sum_{i=1}^n (x_i - \bar x) = 0\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[\hat\beta_1] &= \sigma^2 \sum_{i=1}^n w_i^2 \\ &= \sigma^2 \sum_{i=1}^n \left[\frac{(x_i - \bar x)}{\sum_{i=1}^n (x_j - \bar x)^2}\right]^2 \\ &= \sigma^2 \sum_{i=1}^n \frac{(x_i - \bar x)^2}{\left[\sum_{j=1}^n (x_j - \bar x)^2\right]^2} \\ &= \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar x)^2}{\left[\sum_{j=1}^n (x_j - \bar x)^2\right]^2} = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar x)^2} = \frac{\sigma^2}{S_{xx}}\end{aligned}$$

□

1.12. Theorem: *The estimator $\hat{\beta}_0$ follows the Normal distribution with parameters*

$$\boxed{\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]\right).}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0\right] &= \mathbb{E}\left[\bar{y}-\hat{\beta}_1\bar{x}\right] = \mathbb{E}[\bar{y}] - \mathbb{E}[\hat{\beta}_1\bar{x}] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_i^n y_i\right] - \bar{x}\mathbb{E}\left[\hat{\beta}_1\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^n \mathbb{E}[y_i]\right) - \bar{x}\beta_1 && \mathbb{E}[\hat{\beta}_1] = \beta_1 \\
&= \frac{1}{n}\left(\sum_{i=1}^n (\beta_0+\beta_1 x_i)\right) - \bar{x}\beta_1 && \mathbb{E}[y_1] = \beta_0+\beta_1 x_i \\
&= \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\hat{\beta}_0 &= \mathrm{Var}\left(\bar{y}-\hat{\beta}_1\bar{x}\right) \\
&= \mathrm{Var}(\bar{y}) - 2\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\bar{x}\right) + \mathrm{Var}\left(\hat{\beta}_1\bar{x}\right) \\
&= \frac{\sigma^2}{n} - 2\bar{x}\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right) + \bar{x}^2\mathrm{Var}\hat{\beta}_1 && \text{See (1.25)} \\
&= \frac{\sigma^2}{n} - 2\bar{x}\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right) + \bar{x}^2\frac{\sigma^2}{S_{xx}} \\
&= \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right] - 2\bar{x}\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right).
\end{aligned}$$

It remains to show that $\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0$.

$$\begin{aligned}
\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^n y_i, \frac{\sum_i (x_i-\bar{x})\,y_i}{\sum_i (x_i-\bar{x})^2}\right) \\
&= \frac{1}{n\sum_i (x_i-\bar{x})^2}\mathrm{Cov}\left(\sum_i Y_i, \sum_i (x_i-\bar{x})\,Y_i\right) \\
&= \frac{1}{n\sum_i (x_i-\bar{x})^2}\sum_{i,j}\mathrm{Cov}\left(y_i, (x_i-\bar{x})\,y_j\right) && \mathrm{Cov}\left(y_i, (x_i-\bar{x})\,y_j\right) \propto \delta_{i,j} \\
&= \frac{1}{n\sum_i (x_i-\bar{x})^2}\sum_i (x_i-\bar{x})\,\mathrm{Var}(y_i) && \mathrm{Cov}(y_i, y_i) = \mathrm{Var}(y_i) \\
&= \frac{\sigma^2}{n\sum_i (x_i-\bar{x})^2}\sum_i (x_i-\bar{x}) && \mathrm{Var}(y_i) = \sigma^2 \\
&= 0 && \sum_i (x_i-\bar{x}) = 0
\end{aligned}$$

□

Section 5. SLR: Confidence Interval

1.13. Let us derive a 95% confidence interval for β_1. Recall that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \implies Z := \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1). \tag{1.1}$$

Suppose σ is known. Then

$$\begin{aligned} 0.95 &= P(-1.96 \le Z \le 1.96) \\ &= P\left(-1.96 \le \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \le 1.96\right) \\ &= P\left(-1.96\frac{\sigma}{\sqrt{S_{xx}}} \le \hat{\beta}_1 - \beta_1 \le 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right) \\ &= P\left(-1.96\frac{\sigma}{\sqrt{S_{xx}}} \le \beta_1 - \hat{\beta}_1 \le 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right) \\ &= P\left(\hat{\beta}_1 - 1.96\frac{\sigma}{\sqrt{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right) \end{aligned}$$

Thus, a 95% CI for β_1 is

$$\hat{\beta}_1 \pm 1.96\frac{\sigma}{\sqrt{S_{xx}}}.$$

In practice, σ^2 is often unknown. We can estimate it using the unbiased estimator $\hat{\sigma}^2$.

1.14. Definition: The **standard error** $\mathrm{SE}(\hat{\beta}_1)$ is an estimator of $\hat{\beta}_1$'s standard deviation:

$$\boxed{\mathrm{SE}(\hat{\beta}_1) := \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}.}$$

1.15. Theorem: *The confidence interval of $\hat{\beta}_1$ is given by*

$$\boxed{\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\mathrm{SE}(\hat{\beta}_1).}$$

Proof. Replacing σ^2 by $\hat{\sigma}^2$ in (1.1) gives the t-distributed pivotal quantity

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{(n-2)}.$$

A $100(1-\alpha)\%$ confidence interval is given by

$$1-\alpha = \Pr\left(-q \le \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \le q\right) = \Pr\left(\hat{\beta}_1 - q\frac{\hat{\sigma}}{\mathrm{SE}(\hat{\beta}_1)} \le \beta_1 \le \hat{\beta}_1 + q\frac{\hat{\sigma}}{\mathrm{SE}(\hat{\beta}_1)}\right)$$

Thus, a 95% CI for β_1 is

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

where $t_{1-\alpha/2,n-2}$ is can be found with `qt(p = alpha/2, df = n-2)` in R. □

Section 6. SLR: Hypotheses Testing

1.16. Suppose we want to test a null hypothesis $H_0 : \beta_1 = \theta_0$ against some alternative hypothesis $H_1 : \beta_1 \neq \theta_0$. For SLR, we often set

- $H_0 : \beta_1 = 0$: no linear relationship;
- $H_1 : \beta_1 \neq 0$: two-sided alternative.

The goal is to characterize how much evidence we have against H_0, or how "extreme" our data are relative to H_0. We can test the null hypothesis with the t-statistic

$$T := \frac{\hat{\beta}_1 - \theta_0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}.$$

Assuming H_0 is true, what's the probability to have some as extreme or more than what we observe?

$$\Pr(|T| \geq |t_{\text{obs}}|) = 2\Pr(T \geq |t_{\text{obs}}|) = 2[1 - \Pr(T \leq -|t_{\text{obs}}|)].$$

We typically reject the null hypothesis at the 5% level, i.e., reject H_0 if $p < 0.05$. Would we accept H_0 if $p > 0.05$? **No, we simply would not have enough evidence to reject.**

1.17. Remark: Does this mean $\Pr(\beta_1 = 0) = p$? No. Instead, it means under the null hypothesis, i.e., assuming $\beta_1 = 0$, the probability of a test statistic as extreme as the one observed is equal to p. That's why a small p-value is evidence against the null, since it would be particularly "rare" under the null.

1.18. Remark: Note that a $100(1-\alpha)\%$ CI (e.g., 0.95) corresponds with a hypothesis test with a $100\alpha\%$ significance level (e.g., 0.05), i.e., we will derive a similar conclusion. In particular, if we reject H_0 at the 0.05-level (i.e., when the p-value is less than 0.05), then the 95% CI will not contain the value of 0.

Section 7. SLR: Estimation of Mean Response

1.19. Theorem: *Given new x_0, the estimated mean response is given by*

$$\hat{\mu}_0 = \beta_0 + \beta_1 x_0 \sim N\left(\mu_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

Proof. The mean response for an arbitrary x_0 is given by

$$\hat{\mu}_0 = \mathbb{E}[y \mid x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0 = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}).$$

The estimate of mean response is unbiased:

$$\mathbb{E}[\hat{\mu}_0] = \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \mathbb{E}[\hat{\beta}_0] + \mathbb{E}[\hat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0 =: \mu_0.$$

The variance is given by

$$\begin{aligned}
\operatorname{Var}[\hat{\mu}_0] &= \operatorname{Var}\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right] \\
&= \operatorname{Var}\left[\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 x_0\right] \\
&= \operatorname{Var}\left[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})\right] \\
&= \operatorname{Var}\left[\left(\sum_{i=1}^n \frac{1}{n} y_i\right) + \left(\sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} y_i\right)(x_0 - \bar{x})\right] \\
&= \operatorname{Var}\left[\sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right) y_i\right] && \star \\
&= \sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right)^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} + \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2\frac{1}{n}\frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right) \\
&= \sigma^2 \left(\sum_{i=1}^n \frac{1}{n^2} + \sum_{i=1}^n \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2\sum_{i=1}^n \frac{1}{n}\frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2} S_{xx} + 2\frac{1}{n}\frac{(x_0 - \bar{x})}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}$$

Note in the 5th line of the derivation of variance (labeled $\star$), we see that $\hat{\mu}_0$ is a linear combination of Normal random variables y_i, so μ_0 is also Normal:

$$\hat{\mu}_0 \sim N\left(\mu_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

□

1.20. Note: From above, we know that

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma\sqrt{\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}} \sim N(0,1) \quad \text{and} \quad \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}} \sim t_{n-2}.$$

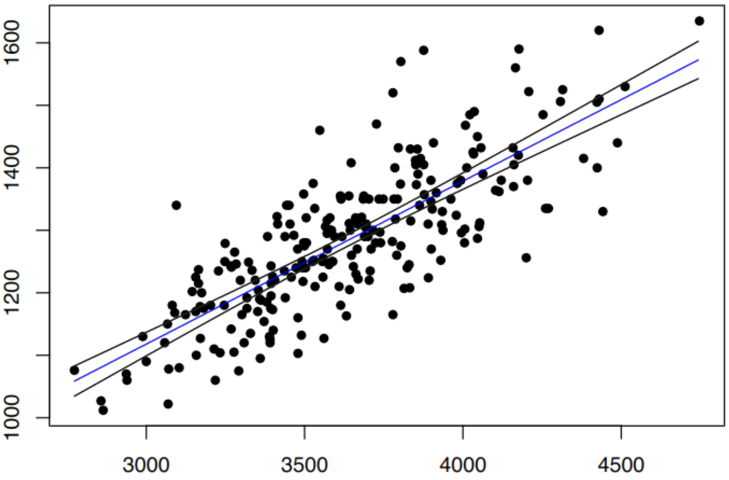
Thus, a 95% CI is given by

$$0.95 = P\left(-t_{n-2,1-\frac{\alpha}{2}} \leq \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)^{1/2}} \leq t_{n-2,1-\frac{\alpha}{2}}\right)$$

In general, a $100(1-\alpha)\%$ CI is given by

$$\boxed{\hat{\mu}_0 \pm t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}.}$$

Note that the CIs get bigger as $x \to \infty$ and $x \to -\infty$ as we have fewer data points there.



Note that many points fall outside of the CI. What if we don't just care about the mean, but also the predictions? That is, even if we got the mean absolutely perfect, the new points wouldn't fall directly on the line!

Section 8. SLR: Prediction of a Single Response

1.21. Note: Suppose we want to predict the response for a new covariate value:

$$y_{\text{new}} = \beta_0 + \beta_1 x_{\text{new}} + \varepsilon_{\text{new}}.$$

Define the predicted value $\hat{y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$ and prediction error $\hat{y}_{\text{new}} - y_{\text{new}}$. Let's quantify the prediction error.

$$\begin{aligned} E\left[\hat{y}_{\text{new}} - y_{\text{new}}\right] &= E\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}\right) - \left(\beta_0 + \beta_1 x_{\text{new}} + \varepsilon_{\text{new}}\right)\right] \\ &= \beta_0 + \beta_1 x_{\text{new}} - \left(\beta_0 + \beta_1 x_{\text{new}}\right) = 0 \end{aligned}$$

Note that $\hat{y}_{\text{new}}$ and y_{new} are independent, because the former is a linear combination of the known y_i's while the latter has nothing to do with those. Moreover, $\hat{y}_{\text{new}}$ is Normal as y_i's are Normal.

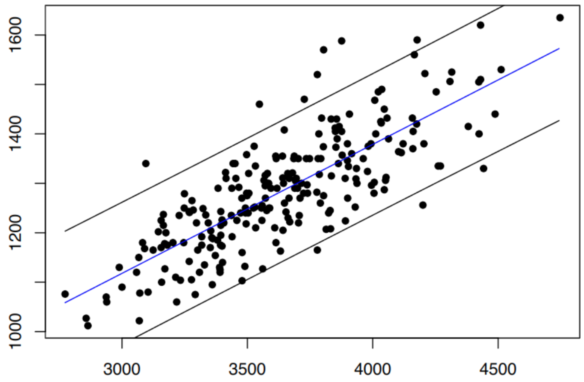
$$\begin{aligned} \operatorname{Var}\left[\hat{y}_{\text{new}} - y_{\text{new}}\right] &= \operatorname{Var}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}\right) - y_{\text{new}}\right] \\ &= \operatorname{Var}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}\right)\right] + \operatorname{Var}\left[y_{\text{new}}\right] \\ &= \left[\sigma^2\left(\frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\right)\right] + \left[\sigma^2\right] \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\right) \end{aligned}$$

Using the same approach as above, we have

$$\frac{\hat{y}_{\text{new}} - y_{\text{new}}}{\sigma\sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}} \sim N(0,1) \quad \text{and} \quad \frac{\hat{y}_{\text{new}} - y_{\text{new}}}{\hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}} \sim t_{n-2}.$$

Thus, a $100(1-\alpha)\%$ **prediction interval** is given by

$$\boxed{\hat{y}_{\text{new}} \pm t_{n-2,\left(1-\frac{\alpha}{2}\right)} \hat{\sigma} \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\right)}.}$$



Note the margin of error of PI is much wider compared to the previous CI.

Section 9. Appendix

1.22. Definition: Let $\bar{x}, \bar{y}$ denote the mean of x's and y's. Define

$$\boxed{\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \end{aligned}}$$

1.23. Lemma: *Let $\bar{x}$ be the mean of $\{x_1, \dots, x_n\}$. Then*

$$\boxed{\sum_{i=1}^{n} (x_i - \bar{x}) = 0.}$$

Proof. Observe that

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x}) &= \left[\sum_{i=1}^{n} x_i\right] - n\bar{x} \\ &= \left[\sum_{i=1}^{n} x_i\right] - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \left[\sum_{i=1}^{n} x_i\right] - \left[\sum_{i=1}^{n} x_i\right] = 0. \end{aligned}$$

□

1.24. Proposition: *We have the following equalities for S_{xx} and S_{xy}:*

$$\boxed{\begin{aligned} S_{xx} &= \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2 \\ S_{xy} &= \left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}. \end{aligned}}$$

Proof. Observe that

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) \\ &= \left[\sum_{i=1}^{n} x_i (x_i - \bar{x})\right] - \left[\bar{x} \sum_{i=1}^{n} (x_i - \bar{x})\right] && \bar{x} \text{ does not depend on } i \\ &= \sum_{i=1}^{n} x_i (x_i - \bar{x}) && \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \end{aligned}$$

$$
\begin{aligned}
&= \left[\sum_{i=1}^{n} x_i^2\right] - \left[\bar{x}\sum_{i=1}^{n} x_i\right] && \bar{x} \text{ does not depend on } i \\
&= \left[\sum_{i=1}^{n} x_i^2\right] - \bar{x}(n\bar{x}) && \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \implies \sum_{i=1}^{n} x_i = n\bar{x} \\
&= \left[\sum_{i=1}^{n} x_i^2\right] - n\bar{x}^2
\end{aligned}
$$

The second property can be derived using a similar approach (Exercise). □

1.25. Lemma:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Chapter 2. Multiple Linear Regression

Section 1. Review: Linear Algebra and Calculus

2.1. Remark: It's often a lot easier to understand formulas intuitively in higher-dimensional spaces once you know their sizes/dimensions (sanity check!). I will try to label the dimensions of vectors and spaces as much as possible. **Warning**: There will be abuse of notations for random variables, e.g., I will label a random vector $\mathbf{x}$ with three elements as $\mathbf{x} \in \mathbb{R}^3$.

2.2. Note: We briefly review some facts about matrices. Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ be matrices.

- $[\mathbf{C}^T]_{ij} = [\mathbf{C}]_{ij}$.
- $\mathbf{C}$ is **symmetric** if $\mathbf{C}^T = \mathbf{C}$.
- $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.
- If a square matrix $\mathbf{B}$ is non-singular, then $\mathbf{B}\mathbf{B}^{-1} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$.
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ is both are non-singular square matrices.
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$.
- $\operatorname{tr}(\mathbf{A}) = \sum_j^n a_{jj}$ for square matrix $\mathbf{A}$.
- $\operatorname{tr}(c\mathbf{A} + \mathbf{B}) = c \cdot \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$.
- $\operatorname{tr}(\mathbf{A}^T) = \operatorname{tr}(\mathbf{A})$.
- $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$.

2.3. Note: We briefly review some matrix calculus.

- Let $\mathbf{y} = (y_1, \ldots, y_k) \in \mathbb{R}^k$ and $f : \mathbb{R}^k \to \mathbb{R}$ be a function of $\mathbf{y}$. Then
$$\frac{\partial f}{\partial \mathbf{y}} = \begin{bmatrix} \dfrac{\partial f}{\partial y_1} \\ \vdots \\ \dfrac{\partial f}{\partial y_k} \end{bmatrix} \in \mathbb{R}^{k \times 1}$$
- If $z = \mathbf{a}^T\mathbf{y} \in \mathbb{R}$ where $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^k$ is a column vector, then
$$\frac{\partial z}{\partial \mathbf{y}} = \mathbf{a} \in \mathbb{R}^{k \times 1}.$$
- If $z = \mathbf{y}^T A \mathbf{y} \in \mathbb{R}$ where $A \in \mathbb{R}^{k \times k}$, then
$$\frac{\partial z}{\partial \mathbf{y}} = (A + A^T)\mathbf{y} \in \mathbb{R}^{k \times 1}.$$
In particular, if A is symmetric, then
$$\frac{\partial z}{\partial \mathbf{y}} = 2A\mathbf{y} \in \mathbb{R}^{k \times 1}.$$

Section 2. Random Vectors

2.4. Definition: A **random vector** is a vector of random variables.
Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a random vector. The **mean** of $\mathbf{y}$ is

$$\boxed{\mathbb{E}[\mathbf{y}] = \begin{bmatrix} \mathbb{E}[y_1] \\ \vdots \\ \mathbb{E}[y_n] \end{bmatrix} \in \mathbb{R}^{n \times 1}.}$$

The **variance** of $\mathbf{y}$ is given by the **covariance matrix**:

$$\begin{aligned} \mathrm{Var}(\mathbf{y}) = \mathbf{V} &= \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T] \\ &= \begin{bmatrix} \mathrm{Var}(y_1) & \mathrm{Cov}(y_1, y_2) & \cdots & \mathrm{Cov}(y_1, y_n) \\ \mathrm{Cov}(y_2, y_1) & \mathrm{Var}(y_2) & \cdots & \mathrm{Cov}(y_1, y_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(y_n, y_1) & \mathrm{Cov}(y_n, y_2) & \cdots & \mathrm{Var}(y_n) \end{bmatrix} \in \mathbb{R}^{n \times n} \end{aligned}$$

In particular,

$$\boxed{\mathbf{V}_{ij} = \mathrm{Cov}(y_i, y_j).}$$

2.5. Proposition: *Let $V = \mathrm{Var}(\mathbf{y})$ be the covariance matrix of $\mathbf{y}$.*

- $\mathbf{V}$ *is* ***symmetric****, i.e.,* $\mathbf{V}_{ij} = \mathbf{V}_{ji}$.
- $\mathbf{V}$ *is* ***positive semidefinite****, i.e.,* $\forall \mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^T \mathbf{V} \mathbf{a} \geq 0$.

Proof. The first claim follows from the fact that the Cov operator is symmetric. For the second claim, observe that

$$\mathbf{a}^T \mathbf{V} \mathbf{a} = \mathbf{a}^T \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T] \mathbf{a} = \mathbb{E}[\mathbf{a}^T (\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{a}] \overset{\star}{=} \mathbb{E}[((\mathbf{y} - \boldsymbol{\mu})^T \mathbf{a})^2] \geq 0$$

where $\star$ follows from the fact that $\mathbf{a}^T(\mathbf{y} - \boldsymbol{\mu})$ and $(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{a}$ are scalars. □

2.6. Note: Recall the following facts. Let $a_i, b_i, c, d \in \mathbb{R}$ be constants, y_i be random variables, and $z = \sum_{i=1}^n a_i y_i + c$, $u = \sum_{i=1}^n b_i y_i + d$ be linear combinations of y_i's; $z, u \in \mathbb{R}$. Then

$$\boxed{\begin{aligned} \mathbb{E}[z] &= \sum_{i=1}^n a_i \mathbb{E}[y_i] + c \in \mathbb{R} \\ \mathrm{Cov}(z, u) &= \sum_{i=1}^n \sum_{j=1}^n a_i b_j \mathrm{Cov}(y_i, y_j) \in \mathbb{R}. \end{aligned}}$$

Equivalently in matrix notation, $z = \mathbf{a}^T \mathbf{y} + c \in \mathbb{R}$, $u = \mathbf{b}^T \mathbf{y} + d \in \mathbb{R}$, then

$$\boxed{\begin{aligned} \mathbb{E}[z] &= a^T \boldsymbol{\mu} + c \in \mathbb{R} \\ \mathrm{Cov}(z, u) &= \mathbf{a}^T \mathbf{V} \mathbf{b} \in \mathbb{R} \end{aligned}}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}]$ and $\mathbf{V} = \mathrm{Var}(\mathbf{y})$. We now consider their multivariate counterparts.

2.7. Note: Consider a random vector $\mathbf{z} = (z_1, \ldots, z_k)^T$ of k linear combinations of random $\mathbf{y}$:

$$\begin{aligned} z_1 &= a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n \\ z_2 &= a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n \\ &\vdots \\ z_k &= a_{k1}y_1 + a_{k2}y_2 + \cdots + a_{kn}y_n \end{aligned}$$

We can equivalently write $\mathbf{z} = \mathbf{A}\mathbf{y} \in \mathbb{R}^k$ for $\mathbf{A} \in \mathbb{R}^{k\times n}$, $[\mathbf{A}]_{ij} = a_{ij}$. Then

$$\begin{aligned} \mathbb{E}[\mathbf{A}\mathbf{y}] &= \mathbf{A}\mathbb{E}[\mathbf{y}] = \mathbf{A}\boldsymbol{\mu} \in \mathbb{R}^k \\ \mathrm{Var}(\mathbf{A}\mathbf{y}) &= \mathbb{E}[(\mathbf{A}\mathbf{y} - \mathbb{E}[\mathbf{A}\mathbf{y}])(\mathbf{A}\mathbf{y} - \mathbb{E}[\mathbf{A}\mathbf{y}])^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{A}(\mathbf{y} - \mathbb{E}[\mathbf{y}]))^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T A^T] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T]A^T \\ &= \mathbf{A}\mathrm{Var}(\mathbf{y})\mathbf{A}^T \\ &= \mathbf{A}\mathbf{V}\mathbf{A}^T \in \mathbb{R}^{k\times k} \end{aligned}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}]$ and $\mathbf{V} = \mathrm{Var}(\mathbf{y})$. In other words, you can pull out a matrix of constants from the expectation and the variance operator much like what you do with vectors. We summarize this result into the following proposition (with an extra bias term $\mathbf{b}$).

2.8. Theorem: *Let $y \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{k\times n}$. Then*

$$\boxed{\begin{aligned} \mathbb{E}[\mathbf{A}\mathbf{y} + \mathbf{b}] &= \mathbf{A}\mathbb{E}[\mathbf{y}] + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} \in \mathbb{R}^k \\ \mathrm{Var}(\mathbf{A}\mathbf{y} + \mathbf{b}) &= \mathbf{A}\mathrm{Var}(\mathbf{y})\mathbf{A}^T = \mathbf{A}\mathbf{V}\mathbf{A}^T \in \mathbb{R}^{k\times k}. \end{aligned}}$$

Section 3. Multivariate Normal Distribution

2.9. Definition: A vector $\mathbf{y}$ has a **multivariate normal distribution** $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its density function has the form

$$\boxed{f(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\}}$$

where $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{y}) = \boldsymbol{\Sigma}$.

2.10. Example: Let $\mathbf{z} = (z_1, \ldots, z_n) \in \mathbb{R}^n$ be a random vector of iid standard normal random variables, i.e., $z_i \overset{\text{iid}}{\sim} N(0, 1)$ for all i's. Then for any $\mathbf{A} \in \mathbb{R}^{k \times n}$,

$$\mathbf{y} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu} \in \mathbb{R}^k \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{y}) = \boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$.

2.11. Proposition: *Some nice properties of MVN:*

- *Linearity: If* $\mathbf{u} = \mathbf{C}\mathbf{y} + \mathbf{d}$, *then*

$$\mathbf{u} \sim MVN(\mathbf{C}\boldsymbol{\mu} + \mathbf{d}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T).$$

- *Marginal distribution: If* $\tilde{\mathbf{y}} = (y_1, \ldots, y_m)^T \subseteq \mathbf{y}$ *is a vector subset of* $\mathbf{y}$, *then* $\tilde{\mathbf{y}}$ *is MVN-distributed. In particular, every* $y_j \in \mathbf{y} \sim N(\mu_j, \Sigma_{jj})$ *is normally distributed.*
- *Conditional distribution: If* $\mathbf{u} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T \sim MVN$ *(i.e., breaking a column vector* $\mathbf{u}$ *into two pieces), then* $\mathbf{y}_1^T \mid \mathbf{y}_2^T$ *is MVN-distributed.*
- *Independence: If* $\Sigma_{ij} = 0$, *then* y_i *and* y_j *are independent.*
  - *Note this only holds for Normal variables: independence* $\implies$ Cov $= 0$ *always holds, but the other direction is generally false (but true for MVN).*

Section 4. Multiple Linear Regression

2.12. Definition: The **multiple linear regression** (MLR) model is given by

$$\boxed{\begin{array}{l} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{iP} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \\ \iff \\ y_i \mid x_i \overset{\text{indep}}{\sim} N\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2\right) \end{array}}$$

- (x_i, y_i): the ith observation, but now we have P covariates instead of just 1.
- The meaning of other symbols remain the same.
- Assume $p < n$, or we have more variates than observations.

2.13. (Cont'd): Equivalently, we can write

$$\boxed{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1P} \\ 1 & x_{21} & x_{22} & \dots & x_{2P} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nP} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},}$$

or more compactly,

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I}) \iff \mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I),}$$

where

- $\mathbf{X}$ is the **design matrix**,
- $\boldsymbol{\beta}$ is the **parameter vector**,
- $\boldsymbol{\varepsilon}$ is the **error vector**, and
- $\mathbf{y}$ is the **response vector**.

Note $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$, where each row represents a sample and each column correspond to a covariate.

2.14. Note: How to interpret the regression coefficients:

- β_0 is the mean outcome when all variates are set to 0.
- β_j represents the difference in mean outcome for a 1-unit change in the jth variate x_j, *holding other covariates* fixed.

Section 5. MLR: Least Squares Estimation

2.15. Theorem: *The LS estimators for $\boldsymbol{\beta}$ is given by*

$$\boxed{\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.}$$

Proof. We wish to minimize the sum of squares:

$$\begin{aligned}
S(\boldsymbol{\beta}) &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} && \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}, \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} \in \mathbb{R}
\end{aligned}$$

Taking its derivative with respect to the vector $\boldsymbol{\beta}$, we get

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + (\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X})\boldsymbol{\beta}$$

Note the last term comes from the derivative of the quadratic form

$$\frac{\partial}{\partial \mathbf{y}}(\mathbf{y}^T\mathbf{A}\mathbf{y}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{y}.$$

Now set the derivative to 0,

$$\begin{aligned}
-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= 0 \\
\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{y} \\
\implies \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.
\end{aligned}$$

Note the inverse exists iff $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ $\mathbf{X}$ has full column rank (i.e., the columns of $\mathbf{X}$ are linearly independent). Thus, we require $n \geq p+1$. □

2.16. Remark: Maximum likelihood gives the same estimators. We omit the derivation.

2.17. Theorem: *The LS estimator $\hat{\boldsymbol{\beta}}$ has the following properties:*

$$\boxed{\begin{aligned}
&\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \\
&\hat{\beta}_j \sim N(\beta_j, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj})
\end{aligned}}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{y}] && \text{Linearity of } \mathbb{E} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

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

Finally, since $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is a linear combination of $\mathbf{y} \sim$ MVN, $\hat{\boldsymbol{\beta}}$ is also MVN. The second statement follows from the *marginal distribution* property of MVN. □

2.18. Theorem: *The unbiased estimator of σ^2 is given by*

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-(p+1)}\mathbf{e}^T\mathbf{e}.}$$

Section 6. MLR: Fitted Values and Residuals

2.19. Definition: Let $\hat{\boldsymbol{\beta}}$ be the LS estimator of $\boldsymbol{\beta}$. The **fitted values** is defined as

$$\begin{aligned}\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \left[\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{y} =: \mathbf{H}\mathbf{y}.\end{aligned}$$

The matrix $\boxed{\mathbf{H} := \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}$ is called the **Hat matrix**, as applying $\mathbf{H}$ to $\mathbf{y}$ yields $\hat{\mathbf{y}}$ ("adding a hat to $\mathbf{y}$"). You should be familiar with the property $\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}.}$

2.20. Proposition: *The Hat matrix* $\mathbf{H}$ *is symmetric and idempotent (i.e., a projection matrix).*

Proof. $\mathbf{H}\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H}.$ □

2.21. Corollary: $\mathbf{I} - \mathbf{H}$ *is symmetric and idempotent (i.e., a projection matrix).*

Proof. $(\mathbf{I} - \mathbf{H}) = \mathbf{I}^T - \mathbf{H}^T = (\mathbf{I} - \mathbf{H})$. Also, $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I}\mathbf{I} - 2\mathbf{H} + \mathbf{H}\mathbf{H} = \mathbf{I} - \mathbf{H}$. □

2.22. Proposition: $\boxed{\mathbb{E}[\hat{\mathbf{y}}] = \mathbf{X}\boldsymbol{\beta}, \operatorname{Var}[\hat{\mathbf{y}}] = \sigma^2\mathbf{H}.}$

Proof.

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{y}}] &= \mathbb{E}[\mathbf{H}\mathbf{y}] \\ &= \mathbf{H}\mathbb{E}[\mathbf{y}] \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{X}\boldsymbol{\beta}\end{aligned} \qquad \begin{aligned}\operatorname{Var}[\hat{\mathbf{y}}] &= \operatorname{Var}[\mathbf{H}\mathbf{y}] \\ &= \mathbf{H}\operatorname{Var}[\mathbf{y}]\mathbf{H}^T \\ &= \mathbf{H}\sigma^2\mathbf{I}\mathbf{H} \\ &= \sigma^2\mathbf{H}\end{aligned}$$

□

2.23. Definition: Define **residuals** as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = \boxed{(\mathbf{I} - \mathbf{H})\mathbf{y}}.$

2.24. Remark: Note that the sum of residuals is zero:

$$\begin{bmatrix}\sum_{i=1}^n e_i \cdot 1 \\ \sum_{i=1}^n x_{i1}e_i \\ \vdots \\ \sum_{i=1}^n x_{ip}e_i\end{bmatrix} = \mathbf{X}^T\mathbf{e} = \mathbf{X}^T(\mathbf{y} - \mathbf{H}\mathbf{y}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} = \mathbf{0}.$$

2.25. Proposition: $\boxed{\mathbb{E}[\mathbf{e}] = \mathbf{0}, \operatorname{Var}[\mathbf{e}] = \sigma^2(\mathbf{I} - \mathbf{H}).}$

Proof.

$$\begin{aligned}\mathbb{E}[\mathbf{e}] &= \mathbb{E}[(\mathbf{I} - \mathbf{H})\mathbf{y}] \\ &= (\mathbf{I} - \mathbf{H})\mathbb{E}[\mathbf{y}] \\ &= \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)(\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}\end{aligned} \qquad \begin{aligned}\operatorname{Var}[\mathbf{e}] &= \operatorname{Var}[(\mathbf{I} - \mathbf{H})\mathbf{y}] \\ &= (\mathbf{I} - \mathbf{H})\operatorname{Var}[\mathbf{y}](\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2(\mathbf{I} - \mathbf{H})\end{aligned}$$

□

2.26. Note: Recall $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ are both linear combinations of $\mathbf{y}$. Since $\mathbf{y}$ is MVN-distributed, the vector obtained by stacking rows of $\hat{\boldsymbol{\beta}}$ on top of the rows of $\mathbf{e}$,

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ (\mathbf{I} - \mathbf{H})\mathbf{y} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ \mathbf{I} - \mathbf{H} \end{pmatrix} \mathbf{y}$$

is also MVN-distributed. We now explore the relationship between $\hat{\boldsymbol{\beta}}$ and $\mathbf{e}$.

2.27. Theorem:

$$\boxed{\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{e} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I} - \mathbf{H}) \end{bmatrix}\right).}$$

Moreover,

1. $\hat{\boldsymbol{\beta}} \sim \text{MVN}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$
2. $\mathbf{e} \sim \text{MVN}\left(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})\right)$, *and*
3. $\hat{\boldsymbol{\beta}}$ *and* $\mathbf{e}$ *are independent.*

Proof. We already proved Claim 1. For Claim 2 and 3, it suffices to prove that the vector has the claim distribution, as $\boldsymbol{\Sigma}_{22} = \text{Var}[\mathbf{e}]$ and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = 0$ indicates variables $\hat{\boldsymbol{\beta}}$ and $\mathbf{e}$ are independent.

$$\begin{aligned} \mathbb{E}[\mathbf{e}] &= \mathbb{E}[(\mathbf{I} - \mathbf{H})\mathbf{y}] \\ &= (\mathbf{I} - \mathbf{H})E[\mathbf{y}] \\ &= (\mathbf{I} - \mathbf{H})\mathbf{X}\hat{\beta} \\ &= \mathbf{X}\hat{\beta} - \mathbf{H}\mathbf{X}\hat{\beta} \\ &= \mathbf{X}\hat{\beta} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\hat{\beta} \\ &= \mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta} \\ &= \mathbf{0} \end{aligned}$$

$$\begin{aligned} \text{Var}\begin{bmatrix} \hat{\beta} \\ \mathbf{e} \end{bmatrix} &= \text{Var}\left[\begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix}\mathbf{y}\right] \\ &= \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix} \text{Var}[\mathbf{y}] \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} X^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix}^T \\ &= \sigma^2 \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix} \begin{pmatrix} \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T & (\mathbf{I} - \mathbf{H})^T \end{pmatrix} & \text{Var}[\mathbf{y}] = \sigma^2\mathbf{I} \\ &= \sigma^2 \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix} \begin{pmatrix} \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} & (\mathbf{I} - \mathbf{H}) \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} & \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{H}) \\ (\mathbf{I} - \mathbf{H})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} & (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \end{aligned}$$

Now

$$\mathbf{A} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

$$\begin{aligned}
\mathbf{B} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{H}) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{0} = \mathbf{C}^T
\end{aligned}$$

$$\begin{aligned}
\mathbf{D} &= (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^T \\
&= \left(\mathbf{I}\mathbf{I}^T - \mathbf{I}\mathbf{H}^T - \mathbf{H}\mathbf{I}^T + \mathbf{H}\mathbf{H}^T\right) \\
&= (\mathbf{I} - 2\mathbf{H} - \mathbf{H}) = (\mathbf{I}-\mathbf{H})
\end{aligned}$$

□

Section 7. MLR: Deriving t-Statistic*

2.28. Remark (Review on eigen-decomposition)**:** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ with n linearly independent eigenvectors q_i, $1 \le i \le n$. Then $\mathbf{A}$ can be factorized as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ where $\mathbf{Q} \in \mathbb{R}^{n\times n}$, whose ith column is the eigenvector q_i of $\mathbf{A}$, and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ii} = \lambda_i$. Only diagonalizable matrices can be factorized in this way.

2.29. Note: So far, we have proved that

$$\hat{\beta} \sim N\left(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right) \Longrightarrow \hat{\beta}_j \sim N\left(\beta_j, \sigma^2 V_{jj}\right).$$

If we can show that

1. $\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e} \sim \chi^2_{n-(p+1)}$, and
2. it is independent of $\hat{\boldsymbol{\beta}}$,

then we obtain the following t-statistic, which can be used for constructing confidence intervals and hypothesis testing. Note we did something similar for SLR but we didn't give a mathematical proof back then.

$$\frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma^2 V_{jj}}}}{\sqrt{\left(\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e}\right)/(n-(p+1))}} = \frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \sim t_{n-(p+1)}.$$

Intuitively, we have $n-(p+1)$ degrees of freedom because we have n data points and we are trying to estimate $p+1$ regression parameters. We now show the math behind this.

2.30. (Cont'd): The second condition is easy. Since $\mathbf{e}$ is independent of $\hat{\boldsymbol{\beta}}$, $\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e}$ as a function of $\mathbf{e}$ is also independent of $\hat{\boldsymbol{\beta}}$. Now for the first condition, recall that $\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{y}$. Consider the eigen-decomposition $\mathbf{I}-\mathbf{H} = \mathbf{\Gamma}^T\mathbf{D}\mathbf{\Gamma}$ where $\mathbf{\Gamma}^{-1} = \mathbf{\Gamma}^T$ and

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}$$

is the diagonal matrix whose diagonal contains the eigenvalues of $(\mathbf{I}-\mathbf{H})$. Define $\tilde{\mathbf{e}} = \mathbf{\Gamma}\mathbf{e}$. Then

$$\begin{aligned} \mathbb{E}[\tilde{\mathbf{e}}] &= \mathbb{E}[\mathbf{\Gamma}\mathbf{e}] = \mathbf{\Gamma}\mathbb{E}[\mathbf{e}] = \mathbf{0} \\ \operatorname{Var}[\tilde{\mathbf{e}}] &= \operatorname{Var}[\mathbf{\Gamma}\mathbf{e}] \\ &= \mathbf{\Gamma}\operatorname{Var}[\mathbf{e}]\mathbf{\Gamma}^T && \operatorname{Var}[\mathbf{A}\mathbf{e}] = \mathbf{A}\operatorname{Var}[\mathbf{e}]\mathbf{A}^T \\ &= \sigma^2\mathbf{\Gamma}(\mathbf{I}-\mathbf{H})\mathbf{\Gamma}^T && \operatorname{Var}[\mathbf{e}] = \mathbf{I}-\mathbf{H} \\ &= \sigma^2\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{D}\mathbf{\Gamma}\right)\mathbf{\Gamma}^T \\ &= \sigma^2\mathbf{D} \end{aligned}$$

Thus, $\tilde{\mathbf{e}} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{D})$ as $\mathbf{e} \sim \text{MVN}$, which implies

$$\tilde{e}_i \overset{\text{indep}}{\sim} N(0, \sigma^2[\mathbf{D}]_{ii}) = N(0, \sigma^2\lambda_i^2).$$

2.31. Remark (Review on χ^2 Distributions)**:** For standard normal rvs $Z_i \overset{\text{iid}}{\sim} N(0,1)$,

$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2.$$

2.32. (Cont'd): Next,

$$\tilde{\mathbf{e}}^T\tilde{\mathbf{e}} = (\boldsymbol{\Gamma}\mathbf{e})^T(\boldsymbol{\Gamma}\mathbf{e}) = \mathbf{e}^T\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\mathbf{e} = \mathbf{e}^T\mathbf{e},$$

so we can write

$$\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e} = \frac{1}{\sigma^2}\tilde{\mathbf{e}}^T\tilde{\mathbf{e}} = \sum_{i=1}^{n}\left(\frac{\tilde{e}_i}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2, \qquad Z_i \overset{\text{indep}}{\sim} N(0,\lambda_i^2).$$

Thus, $\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e}$ is a sum of squared independent normally distributed rvs. To show

$$\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e} \sim \chi^2_{(n-(p+1))},$$

we need to show that $n-(p+1)$ of the eigenvalues λ_j's are equal to 1, and all others are equal to 0. (Indeed, if $\lambda_j = 0$, then $Z_j \sim N(0,0)$ becomes a constant.) We know that $(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \mathbf{I}-\mathbf{H}$. This gives

$$\begin{aligned}(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) &= \mathbf{I}-\mathbf{H}\\ (\boldsymbol{\Gamma}^T\mathbf{D}\boldsymbol{\Gamma})(\boldsymbol{\Gamma}^T\mathbf{D}\boldsymbol{\Gamma}) &= (\boldsymbol{\Gamma}^T\mathbf{D}\boldsymbol{\Gamma})\\ \boldsymbol{\Gamma}^T\mathbf{D}\mathbf{D}\boldsymbol{\Gamma} &= \boldsymbol{\Gamma}^T\mathbf{D}\boldsymbol{\Gamma},\end{aligned}$$

i.e., $\mathbf{D}\mathbf{D} = \mathbf{D}$ and thus $\lambda_j^2 = \lambda_j$. Thus all λ_j are either 0 or 1. Next,

$$\begin{aligned}\sum_j \lambda_j = \operatorname{tr}(\mathbf{D}) &= \operatorname{tr}\left(\mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\right) && \text{trace is similarity-invariant}\\ &= \operatorname{tr}\left(\boldsymbol{\Gamma}^T\mathbf{D}\boldsymbol{\Gamma}\right) && \text{invariant under cyclic permutation}\\ &= \operatorname{tr}(\mathbf{I}-\mathbf{H})\\ &= \operatorname{tr}(\mathbf{I}) - \operatorname{tr}(\mathbf{H}) && \text{trace is linear}\\ &= n - \operatorname{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\\ &= n - \operatorname{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) && \text{invariant under cyclic permutation}\\ &= n - \operatorname{tr}(\mathbf{I}_{p+1}) && \mathbf{X}\in\mathbb{R}^{n\times(p+1)} \implies \mathbf{X}^T\mathbf{X}\in\mathbb{R}^{(p+1)(p+1)}\\ &= n-(p+1)\end{aligned}$$

This concludes our proof.

2.33. Note: This entire section is optional. The only thing you need to remember is that

$$\boxed{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \sim t_{n-(p+1)}.}$$

Moreover, the standard error of $\hat{\beta}_j$ is given by

$$\boxed{\mathrm{SE}(\hat{\beta}_j) = \hat{\sigma}\sqrt{V_{jj}}.}$$

Section 8. MLR: Hypothesis Testing

2.34. Note: Suppose we want to test a null hypothesis $H_0 : \beta_j = \theta_0$ against some alternative hypothesis $H_1 : \beta_j \neq \theta_0$. Our goal is to characterize how much evidence we have against H_0, or more intuitively, how *extreme* are our data relative to H_0. Under H_0 (i.e., if H_0 holds), then

$$T := \frac{\hat{\beta}_j - \theta_0}{\hat{\sigma}/\sqrt{V_{jj}}} \sim t_{n-p-1}.$$

Below we discuss two approaches for hypothesis testing.

2.35. (Cont'd): First, we can compute the p-value and compare it against α.

1. Given observed value

$$T_{\text{obs}} := \frac{\hat{\beta}_j - \theta_0}{\hat{\sigma}/\sqrt{V_{jj}}} \sim t_{n-p-1},$$

2. Compute the p-value $p = \Pr(|T| \geq |T_{\text{obs}}|) = 2\Pr(T \geq T_{\text{obs}})$ given by

```
p <- pq(T_obs, df = n-p-1, lower.tail=FALSE).
```

3. If $p < \alpha$, reject H_0 (at α).

2.36. (Cont'd): Alternatively, we can compute the quantile, known as the **critical value**, of the test statistic T that gives a p-value of α, then compare our observed value with this threshold.

1. Given observed value

$$T_{\text{obs}} := \frac{\hat{\beta}_j - \theta_0}{\hat{\sigma}/\sqrt{V_{jj}}} \sim t_{n-p-1},$$

2. Compute the threshold by

```
q <- qt(p = alpha/2, df=n-p-1).
```

3. If $|T_{\text{obs}}| < t_{n-p-1,1-\alpha/2} =$ `q`, reject H_0 (at α).

2.37. Theorem: *A* $(100-\alpha)\%$ *CI for* β_j *is*

$$\boxed{\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{V_{jj}}.}$$

Proof. Omitted. □

2.38. Note: We can never guarantee that any single CI contains the true value. However, as we repeatedly construct CIs, about $(100-\alpha)\%$ of them will contain the true value.

Section 9. MLR: Estimating Mean Response

2.39. For an arbitrary vector of covariates $\mathbf{x}_0 = [1, x_{01}, x_{02}, \ldots, x_{0p}]$, the mean response is

$$\mu_0 = \mathbb{E}[\mathbf{y}_0 \mid \mathbf{x}_0] = \mathbf{x}_0\boldsymbol{\beta}.$$

We can estimate this as $\hat{\mu}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}}$. We now look at the properties of this estimator.

2.40. Proposition:

$$\boxed{\begin{aligned} \mathbb{E}[\hat{\mu}_0] &= \mathbf{x}_0\boldsymbol{\beta} \\ \operatorname{Var}[\hat{\mu}_0] &= \sigma^2\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T \end{aligned}}$$

Proof.

$$\begin{aligned} \mathbb{E}[\hat{\mu}_0] &= \mathbb{E}[\mathbf{x}_0\hat{\boldsymbol{\beta}}] \\ &= \mathbf{x}_0\mathbb{E}[\hat{\boldsymbol{\beta}}] \\ &= \mathbf{x}_0\boldsymbol{\beta} \end{aligned}$$

$$\begin{aligned} \operatorname{Var}[\hat{\mu}_0] &= \operatorname{Var}\left(\mathbf{x}_0\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{x}_0 \operatorname{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0^T \\ &= \mathbf{x}_0\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T \\ &= \sigma^2\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T \end{aligned}$$

□

2.41. Note: By the same logic as before,

$$\begin{aligned} \frac{\hat{\mu}_0 - \mu_0}{\sigma\sqrt{\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T}} &\sim N(0,1), \\ \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\sqrt{\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T}} &\sim t_{n-p-1}, \end{aligned}$$

and a $100(1-\alpha)\%$ CI is given by

$$\boxed{\hat{\mu}_0 \pm t_{n-p-1,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T}.}$$

Section 10. MLR: Prediction

2.42. Note: For a new response

$$y_{\text{new}} = \mathbf{x}_{\text{new}}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\text{new}} ,$$

our prediction is

$$\hat{y}_{\text{new}} = \mathbf{x}_{\text{new}}\,\hat{\boldsymbol{\beta}}.$$

2.43. Proposition:

$$\boxed{\begin{aligned} \mathbb{E}[\hat{y}_{\text{new}}] &= \mathbf{x}_{\text{new}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new} \\ \text{Var}[\hat{y}_{\text{new}}] &= \sigma^2 \mathbf{x}_{\text{new}} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_{\text{new}}^T \end{aligned}}$$

Proof.

$$\begin{aligned} \mathbb{E}\left[\hat{y}_{\text{new}}\right] &= \mathbb{E}\left[\mathbf{x}_{\text{new}}\,\hat{\boldsymbol{\beta}}\right] \\ &= \mathbf{x}_{\text{new}}\,\mathbb{E}[\hat{\boldsymbol{\beta}}] \\ &= \mathbf{x}_{\text{new}}\,\boldsymbol{\beta} \end{aligned}$$

$$\begin{aligned} \text{Var}\left[\hat{y}_{\text{new}}\right] &= \text{Var}\left(\mathbf{x}_{\text{new}}\,\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{x}_{\text{new}}\ \text{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_{\text{new}}^T \\ &= \mathbf{x}_{\text{new}}\,\sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_{\text{new}}^T \\ &= \sigma^2 \mathbf{x}_{\text{new}} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_{\text{new}}^T \end{aligned}$$

□

2.44. Note: Since y_{new} and $\hat{y}_{\text{new}}$ are independent and normally-distributed, we have

$$\frac{y_{\text{new}} - \hat{y}_{\text{new}}}{\sigma\sqrt{1 + \mathbf{x}_{\text{new}} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_{\text{new}}^T}} \sim N(0,1),$$

$$\frac{y_{\text{new}} - \hat{y}_{\text{new}}}{\hat{\sigma}\sqrt{1 + \mathbf{x}_{\text{new}} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_{\text{new}}^T}} \sim t_{n-p-1}$$

Thus, a $100(1-\alpha)\%$ prediction interval for y_{new} is

$$\boxed{\hat{y}_{\text{new}} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + \mathbf{x}_{\text{new}} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_{\text{new}}^T} .}$$