Handout 2

STAT3530 Applied Linear Models

This handout follows the lecture slides. Part 1 derives the least squares estimator two ways, Part 2 works out its properties, Part 3 builds up to the variance estimate, and Part 4 applies all of it. The simulation section uses a script from the course website.

Part 1: Deriving the least squares estimator

The model in matrix form

The simple linear model for n observations is $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, or in matrix form

$$Y = X\beta + \epsilon\,, \qquad \mathbb{E}[\epsilon] = 0\,, \qquad \mathrm{Var}[\epsilon] = \sigma^2 I$$

1. Fill in the dimensions.

   Y is _____×_____ , X is _____×_____ , β is _____×_____ , ϵ is _____×_____

2. Suppose $n = 4$ with predictor values $x = (2, 4, 6, 8)$. Write out X explicitly, then compute $X^T X$.

By calculus

The least squares criterion is $SSE(\beta) = (Y - X\beta)^T(Y - X\beta)$. Recall the two gradient rules, valid for a fixed vector a and a fixed *symmetric* matrix A:

$$\nabla\, a^T x = a \qquad \text{and} \qquad \nabla\, x^T A x = 2Ax$$

3. Multiply out $SSE(\beta) = (Y^T - \beta^T X^T)(Y - X\beta)$ into four terms.

4. Two of those terms are equal. Which two, and why? *(Hint: what are their dimensions?)*

5. Identify the vector a and the matrix A needed to apply the two rules, then fill in the gradient step by step.

   $a =$ _____ , $A =$ _____

$$\nabla SSE(\beta) = \nabla\Big[\; Y^TY - \; 2\_\_\_\_\_\beta \; + \; \beta^T\_\_\_\_\_\beta \;\Big]$$

$$= \quad 0 \; - \; 2\_\_\_\_\_ \; + \; 2\_\_\_\_\_\beta$$

6. Set the gradient to zero to obtain the **normal equations**, then solve for β.

By geometry

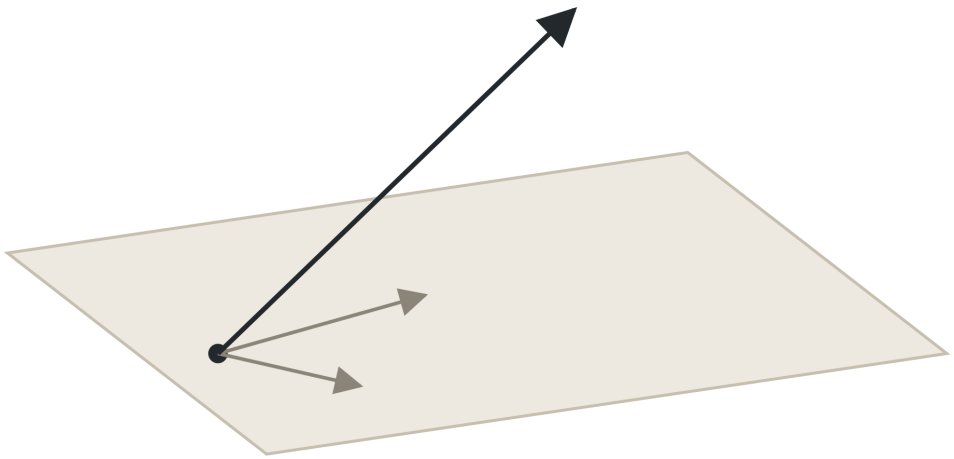
First, two definitions to have straight.

7. Complete each sentence.

   Two vectors u and v in $\mathbb{R}^n$ are **orthogonal** if _____ .

   A vector v is **orthogonal to a subspace** S if _____ .

8. The figure shows the same problem geometrically. Label the shaded subspace and the vector Y. Then **draw in**: a point $X\beta$ where you think the projection lands, the residual vector joining it to Y, and a right-angle mark showing the two are perpendicular.



9. Least squares looks for the point of $\text{col}(X)$ closest to Y. Using your second definition from question 7, explain why the residual at that point must be orthogonal to $\text{col}(X)$.

10. "The residual is orthogonal to every column of X" can be written as a single matrix equation. Write it, and then rearrange it into the normal equations from question 6.

Part 2: Properties of the least squares estimator

Mean and variance

Write $\hat{\beta} = (X^T X)^{-1} X^T Y$, and note that $(X^T X)^{-1} X^T$ is a *fixed* matrix while Y is random.

11. First the mean. Using $\mathbb{E}[\epsilon] = 0$, show that $\mathbb{E}[Y] = X\beta$.

12. Now apply $\mathbb{E}[AY] = A\,\mathbb{E}[Y]$ with $A = (X^T X)^{-1} X^T$ to show that $\mathbb{E}[\hat{\beta}] = \beta$. What is this property called?

13. Next the variance. Using $\mathrm{Var}[\epsilon] = \sigma^2 I$, explain why $\mathrm{Var}[Y] = \sigma^2 I$ as well.

14. Now apply $\mathrm{Var}[AY] = A\,\mathrm{Var}[Y]\,A^T$ with the same A. Fill in the missing matrices at each step.

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}] &= \mathrm{Var}\Big[\underline{\qquad\qquad\qquad} \; Y \Big] \\
&= \underline{\qquad\qquad\qquad} \;\; \mathrm{Var}[Y] \;\; \underline{\qquad\qquad\qquad} \\
&= \underline{\qquad\qquad\qquad} \;\; (\sigma^2 I) \;\; \underline{\qquad\qquad\qquad} \\
&= \sigma^2 \; \underline{\qquad\qquad\qquad\qquad\qquad\qquad} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

Remark. The collapse in the last step happens because $X^T X$ meets one of its own inverses and cancels, leaving a single $(X^T X)^{-1}$.

Seeing it in simulation

Download and run `02-sampling-variability.R` from the course website. It simulates 2000 datasets from a known model and stores the estimates from each.

15. Write down the model the script simulates from, filling in the numbers it uses. Then say which pieces are **fixed** and which are **random**, and which are **known to you as the analyst** and which are not.

16. Compare `colMeans(B)` with `beta`. Which property from question 12 does this illustrate? Why don't the numbers agree exactly?

17. Compare `apply(B, 2, var)` with `diag(V)`. Which property does *this* illustrate?

18. Look at the scatterplot of the simulated pairs. Is the correlation positive or negative? Explain. *Hint: Picture a line pivoting about* $(\bar{x}, \bar{Y})$.

19. Re-run the script with `x` replaced by `x - mean(x)`. What happens to the correlation, and which entry of $(X^TX)^{-1}$ explains it?

Optimality

20. Complete the statement.

    Gauss-Markov theorem. Under the model $Y = X\beta + \epsilon$ with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2 I$, among all estimators of β that are

    _____ and _____ ,

    the least squares estimator $\widehat{\beta}$ has the smallest _____ . We abbreviate this by saying $\widehat{\beta}$ is **BLUE**.

Part 3: Estimating the variance

The hat matrix

The fitted values are $\hat{Y} = X\hat{\beta}$. Substituting $\hat{\beta} = (X^TX)^{-1}X^TY$ gives $\hat{Y} = HY$ for a single matrix H, and the residuals are then $e = Y - \hat{Y} = (I - H)Y$.

21. Write down H, and give its dimensions. Why is it called the "hat matrix"?

22. Show that H is idempotent by filling in the cancellation.

$$HH = \Big[X(X^TX)^{-1}X^T\Big]\Big[X(X^TX)^{-1}X^T\Big]$$

$$= X(X^TX)^{-1} \_\_\_\_\_ (X^TX)^{-1}X^T$$

$$= X \_\_\_\_\_ X^T = H$$

In words, why should $HH = H$ hold for *any* projection?

23. Show that $HX = X$. What does that say about vectors already lying in $\text{col}(X)$?

24. Now show $I - H$ is *also* symmetric and idempotent, so it is a projection too.

$$(I - H)^T = I^T - \_\_\_\_\_ = I - H$$

$$(I - H)(I - H) = I - H - H + \_\_\_\_\_ = I - 2H + \_\_\_\_\_ = I - H$$

25. Show in one line that $H(I - H) = 0$. This establishes that H and $I - H$ project onto orthogonal subspaces.

26. Use question 23 to fill in the following.

$$e \;=\; (I-H)Y = (I-H)(X\beta+\epsilon)$$

$$= \_\_\_\_\_\;\beta \;+\; (I-H)\,\epsilon \;=\; (I-H)\,\epsilon$$

So the residual depends only on ϵ, never on β.

The error variance σ^2

Since $\mathbb{E}[\epsilon_i] = 0$, the error variance is $\sigma^2 = \mathrm{Var}[\epsilon_i] = \mathbb{E}[\epsilon_i^2]$ — it is the average squared error. So if we could see the errors, we would estimate σ^2 by averaging their squares.

27. Use that fact to fill in the expected squared length of the whole error vector.

$$\mathbb{E}[\|\epsilon\|^2] \;=\; \mathbb{E}\Big[\textstyle\sum_{i=1}^{n}\epsilon_i^2\Big] \;=\; \_\_\_\_\_$$

Each of the n coordinates of ϵ contributes σ^2 to the total.

The catch is that ϵ cannot be observed. The residual e can be, and question 26 says exactly how the two are related: $e = (I-H)\epsilon$, the error vector with the part lying in $\mathrm{col}(X)$ projected away. Question 25 says what survives lies in $\mathrm{col}(X)^{\perp}$, which has dimension $n-2$.

So e is what is left of ϵ after a projection, and projecting can only shorten a vector. Of the n contributions of σ^2 above, the two spanned by $\mathrm{col}(X)$ are removed and $n-2$ remain:

$$\mathbb{E}[\|e\|^2] \;=\; (n-2)\,\sigma^2$$

This is why $\|e\|^2$ is the right thing to look at: it is the observable stand-in for $\|\epsilon\|^2$, just systematically two dimensions short.

28. Fill in the divisor that makes the estimator unbiased.

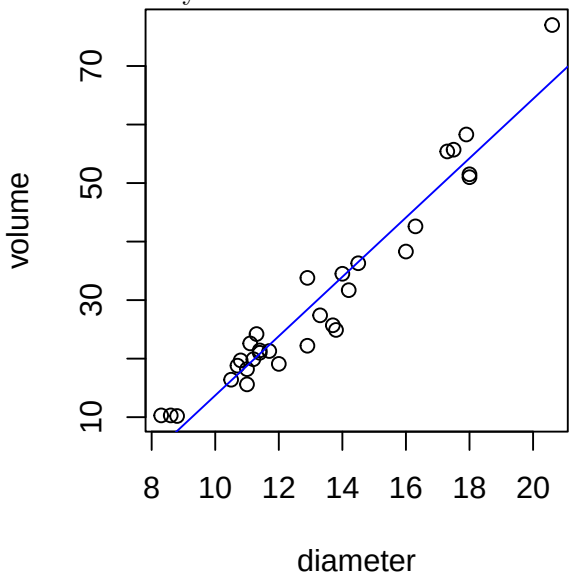
$$\hat{\sigma}^2 \;=\; \_\_\_\_\_ \qquad \text{so that} \qquad \mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

29. Recall that you divide by $n-1$ when computing a sample variance. Consider possible connections with the present context. Can you explain why we use $n-2$ for the denominator in $\hat{\sigma}^2$?

Part 4: Examples

Black cherry trees

Timber volume (cubic ft) vs. diameter (in) for 31 felled cherry trees.




```
# fit the model
fit <- lm(volume ~ diameter, data = cherry)

# coefficient estimates
coef(fit)
```

```
(Intercept)    diameter
 -36.943459    5.065856
```

```
# variance parameter estimate
sigma(fit)
```

```
[1] 4.251988
```

```
# estimated variance-covariance matrix
vcov(fit)
```

```
             (Intercept)    diameter
(Intercept)  11.3242005 -0.81073976
diameter     -0.8107398  0.06119536
```

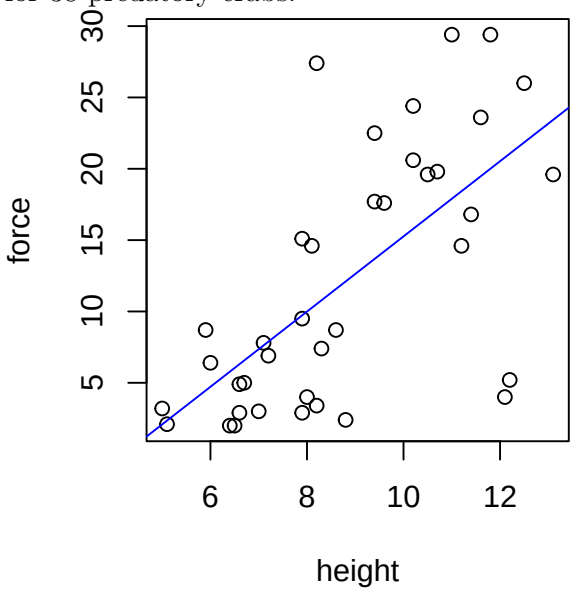
30. Read off $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}$, and give the units of each.

31. Take square roots of the diagonal of `vcov(fit)` to get the standard errors.

    $SE(\hat{\beta}_0) =$ _____ , $SE(\hat{\beta}_1) =$ _____

Crab claws — your turn

Claw closing force (N) and claw height (mm) for 38 predatory crabs.




```
# fit the model
fit <- lm(force ~ height, data = crab)

# coefficient estimates
coef(fit)
```

```
(Intercept)      height
 -11.086902    2.634823
```

```
# variance parameter estimate
sigma(fit)
```

```
[1] 6.89171
```

```
# estimated variance-covariance matrix
vcov(fit)
```

```
             (Intercept)     height
(Intercept)   21.366599 -2.2825772
height        -2.282577  0.2589965
```

Interpret each of the estimates above.

32. The coefficients $\hat{\beta}$. What are the units of $\hat{\beta}_1$?

33. The error variance estimate $\hat{\sigma}$. What does it measure, and in what units?

34. The coefficient variances $\hat{\sigma}^2(X^TX)^{-1}$. Compute and interpret $SE(\hat{\beta}_1)$. How does it differ from $\hat{\sigma}$?