STAC67H3F: Midterm Test, solutions

Saturday October 24, 2026

Exercise 1 (14 points in total)

(a) TRUE or FALSE

(i) FALSE. From the lectures, $\boldsymbol{\sigma^2\{e\}} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$. So $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$, which is smaller than σ^2 as soon as $h_{ii} > 0$, and $\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$, which is not zero in general. The residuals are not even free to vary independently: with an intercept, the normal equations give $\sum_i e_i = 0$, so each residual is determined by the $n-1$ others. The *errors* ε_i are independent with variance σ^2; the residuals are not.

(ii) TRUE. In simple linear regression,

$$F^* = \frac{MSR}{MSE} = \frac{\hat\beta_1^2 S_{xx}}{MSE} = \left(\frac{\hat\beta_1}{s\{\hat\beta_1\}}\right)^2 = (t^*)^2,$$

and the critical values match: $F(1-\alpha;\ 1,\ n-2) = \left[t(1-\alpha/2;\ n-2)\right]^2$, since the square of a $t(n-2)$ variable has an $F(1, n-2)$ distribution. Hence $F^* > F(1-\alpha; 1, n-2)$ exactly when $|t^*| > t(1-\alpha/2; n-2)$: the two tests always reach the same conclusion.

(b) A linear function of the response vector

(i) Each row of $\boldsymbol{A}$ holds the coefficients of Y_1, Y_2, Y_3 in one W: $W_1 = 1\,Y_1 - 1\,Y_2 + 0\,Y_3$ and $W_2 = 1\,Y_1 + 1\,Y_2 - 2\,Y_3$, so

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & -2 \end{bmatrix}.$$

The Y_i are independent with variance $3^2 = 9$, so $\boldsymbol{\sigma^2\{Y\}} = 9\boldsymbol{I}$ and

$$\boldsymbol{\sigma^2\{W\}} = \boldsymbol{A}\,(9\boldsymbol{I})\,\boldsymbol{A}' = 9\,\boldsymbol{A}\boldsymbol{A}' = 9\begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix} = \begin{bmatrix} 18 & 0 \\ 0 & 54 \end{bmatrix}.$$

(ii) Yes. $\boldsymbol{Y}$ is a normal vector, so $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{Y}$, a linear function of it, is a normal vector too. Its two components are uncorrelated, since $\mathrm{Cov}(W_1, W_2) = 0$ by (i). For a normal vector, uncorrelated components are independent: W_1 and W_2 are independent. (Without normality, zero covariance alone would not be enough.)

Exercise 2 (22 points in total)

(a) Predicted overall rating

$$\hat{Y}_h = 14.376 + 0.7546 \times 80 = 74.74.$$

(b) Confidence interval for the mean overall rating at $x_h = 80$

With $\boldsymbol{X}_h = (1, 80)'$,

$$s^2\{\hat{Y}_h\} = \boldsymbol{X}_h' \, \boldsymbol{s^2}\{\hat{\boldsymbol{\beta}}\} \, \boldsymbol{X}_h = s^2\{\hat{\beta}_0\} + 80^2 \, s^2\{\hat{\beta}_1\} + 2 \times 80 \times s\{\hat{\beta}_0, \hat{\beta}_1\}$$

$$= 43.824 + 6400 \times 0.0095127 + 160 \times (-0.63354) = 3.339, \qquad s\{\hat{Y}_h\} = 1.827.$$

With $n - 2 = 28$ degrees of freedom, $t(0.975;\ 28) = 2.048$:

$$74.74 \pm 2.048 \times 1.827 = 74.74 \pm 3.74 = (71.00,\ 78.49).$$

(c) Test of $H_0 : \beta_1 = 1$ against $H_a : \beta_1 \neq 1$

$$t^* = \frac{\hat{\beta}_1 - 1}{s\{\hat{\beta}_1\}} = \frac{0.7546 - 1}{\sqrt{0.0095127}} = -2.52,$$

on $n - 2 = 28$ degrees of freedom under H_0. The test is two-sided: reject H_0 when $|t^*| > t(0.975;\ 28) = 2.048$. Since $|t^*| = 2.52 > 2.048$, we reject H_0 at the 5% level. (From the table, $t(0.99; 28) = 2.467 < 2.52 < t(0.995; 28) = 2.763$, so the p-value is between 0.01 and 0.02.) The manager's claim is not supported: since $t^* < 0$, the data indicate that, on average, a one-point increase in the complaints rating comes with an increase of less than one point in the overall rating.

(d) Prediction interval for one department at $x_h = 80$

$MSE = 6.993^2 = 48.902$, so

$$s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} = 48.902 + 3.339 = 52.241, \qquad s\{\text{pred}\} = 7.228,$$

$$74.74 \pm 2.048 \times 7.228 = 74.74 \pm 14.80 = (59.94,\ 89.55).$$

Exercise 3 (28 points in total)

(a) The missing values

In simple linear regression $F^* = (t^*)^2$ and $F^* = (n-2)R^2/(1-R^2)$, so

$$R^2 = \frac{(t^*)^2}{(t^*)^2 + n - 2} = \frac{91.374}{91.374 + 30} = 0.7528 \quad \text{(a)}, \qquad s\{\hat{\beta}_1\} = \frac{\hat{\beta}_1}{t^*} = \frac{-5.3445}{-9.559} = 0.5591 \quad \text{(b)}.$$

Interpretation. About 75% of the variation in fuel efficiency among these 32 cars is explained by its linear relationship with weight.

(b) A 95% prediction interval

The model has $p = 3$ parameters, so the error degrees of freedom are $n - p = 32 - 3 = 29$, and $t(0.975;\ 29) = 2.045$. From the printed interval,

$$s\{\hat{Y}_h\} = \frac{21.82 - 20.83}{2.045} = 0.4841.$$

With $MSE = 2.593^2 = 6.7236$,

$$s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} = 6.7236 + 0.2344 = 6.9580, \qquad s\{\text{pred}\} = 2.6378,$$

$$20.83 \pm 2.045 \times 2.6378 = 20.83 \pm 5.39 = (15.44,\ 26.22) \text{ miles per gallon.}$$

(c) The missing values of the multiple regression output, and the F test

(i) The error degrees of freedom are $n - p = 29$, and $SSE = 29 \times MSE = 29 \times 6.7236 = 194.99$. Then

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{194.99}{1126.05} = 0.8268, \qquad R^2_a = 1 - \frac{n-1}{n-p}\frac{SSE}{SST} = 1 - \frac{31}{29} \times \frac{194.99}{1126.05} = 0.8149,$$

$$F^* = \frac{SSR/(p-1)}{MSE} = \frac{(1126.05 - 194.99)/2}{6.7236} = 69.2, \quad \text{on 2 and 29 degrees of freedom.}$$

(ii) The global F test: $H_0 : \beta_1 = \beta_2 = 0$ against H_a: at least one of β_1, β_2 is not zero. The critical value is $F(0.95;\ 2,\ 29) = 3.328$. Since $F^* = 69.2 > 3.328$, we reject H_0: weight and engine power, taken together, are related to fuel efficiency.

(d) Checking the assumptions

- **Linearity: violated.** The residuals show a clear U shape, followed by the lowess curve: they are positive for the smallest and the largest fitted values, negative in between. The mean of mpg is not a linear function of weight and horsepower.
- **Constant variance.** No funnel in the residual plot, and the Brown-Forsythe test does not reject equal variances ($p = 0.57$).
- **Normality: in doubt.** The Q-Q plot bends upwards at the right end (three large positive residuals), and the Shapiro-Wilk test rejects normality at the 5% level ($p = 0.034$). This is partly a consequence of the curvature: the largest residuals sit at the two ends of the U.
- **Independence** cannot be assessed from these two graphs.

Remedy. Change the mean function, for example by transforming the response, such as log(mpg) or 1/mpg (gallons per mile), possibly chosen with Box-Cox, or by transforming the predictors; then fit again and check the diagnostics again. For these data, regressing log(mpg) or 1/mpg on the same predictors removes most of the curvature, and the Shapiro-Wilk test no longer rejects for log(mpg).

Exercise 4 (16 points in total)

(a) Confidence interval for β_2

$MSE = SSE/(n-p) = 904.20/22 = 41.100$, and

$$s^2\{\hat{\beta}_2\} = MSE \times \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{33} = 41.100 \times 0.002419 = 0.09942, \qquad s\{\hat{\beta}_2\} = 0.3153.$$

With $t(0.975;\ 22) = 2.074$:

$$-1.345 \pm 2.074 \times 0.3153 = -1.345 \pm 0.654 = (-2.00,\ -0.69).$$

Interpretation. With 95% confidence, at a given maximum temperature, each additional millimetre of rain is associated with a decrease in the mean water consumption of the city of between 0.69 and 2.00 megalitres.

Test. The interval does not contain 0, so $H_0 : \beta_2 = 0$ is rejected against $H_a : \beta_2 \neq 0$ at the 5% level: rainfall is related to water consumption, once temperature is in the model.

(b) Prediction interval for one day with $X_1 = 38$ and $X_2 = 5$

With $\boldsymbol{X}_h = (1, 38, 5)'$:

$$\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}} = 56.514 + 2.704 \times 38 - 1.345 \times 5 = 152.541.$$

Expanding the quadratic form $\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h = \sum_j \sum_k X_{hj} X_{hk} c_{jk}$,

$$\begin{aligned}\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h &= 1.401 + 38^2(0.002145) + 5^2(0.002419) \\ &\quad + 2(38)(-0.05277) + 2(5)(0.002155) + 2(38)(5)(-0.0005726) \\ &= 1.401 + 3.0974 + 0.0605 - 4.0105 + 0.0216 - 0.2176 = 0.3523.\end{aligned}$$

Then

$$s^2\{\text{pred}\} = MSE\,(1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h) = 41.100 \times 1.3523 = 55.579, \qquad s\{\text{pred}\} = 7.455,$$

$$152.54 \pm 2.074 \times 7.455 = 152.54 \pm 15.46 = (137.08,\ 168.00) \text{ megalitres.}$$

Can it be interpreted in the usual way? No. A maximum temperature of 38 degrees is outside the range of the data (19.2 to 33.5 degrees): this is an **extrapolation**. The interval is computed as if the linear model held at 38 degrees, but the data contain no day that hot and cannot check it. Water consumption on a very hot day may well depend on temperature differently, for example because of watering bans or a demand that saturates. The width of the interval accounts for the uncertainty in $\hat{\boldsymbol{\beta}}$ and for the day's own error *under the model*, not for the possibility that the model is wrong there. The 95% level is therefore not guaranteed, and the interval should be reported, if at all, with this caveat. (Rainfall, 5 mm, is within its range.)

Exercise 5 (20 points in total)

(a) Matrix form

$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$, where the design matrix $\boldsymbol{X} = \boldsymbol{x}$ is the single column $(x_1, \ldots, x_n)'$ (no column of 1s, since there is no intercept), the parameter vector is the scalar β, and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The model has $p = 1$ parameter.

(b) The estimator and the hat matrix

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^n x_i^2, \qquad \boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{x}'\boldsymbol{Y} = \sum_{i=1}^n x_i Y_i, \qquad \hat{\beta} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{Y} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2},$$

where $\sum_i x_i^2 > 0$ because the x_i are not all zero. Since $(\boldsymbol{x}'\boldsymbol{x})^{-1}$ is the scalar $1/\sum_k x_k^2$,

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \frac{\boldsymbol{x}\boldsymbol{x}'}{\sum_k x_k^2}, \qquad h_{ij} = \frac{x_i x_j}{\sum_k x_k^2}, \qquad h_{ii} = \frac{x_i^2}{\sum_k x_k^2},$$

and $\sum_i h_{ii} = \sum_i x_i^2 / \sum_k x_k^2 = 1$, the number of parameters.

(c) Mean, variance and distribution of $\hat{\beta}$

$E(\boldsymbol{Y}) = \boldsymbol{x}\beta$, so $E(\hat{\beta}) = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{x}\beta = \beta$: $\hat{\beta}$ is unbiased. By the rule for linear functions,

$$\mathrm{Var}(\hat{\beta}) = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\,(\sigma^2\boldsymbol{I})\,\boldsymbol{x}\,(\boldsymbol{x}'\boldsymbol{x})^{-1} = \sigma^2(\boldsymbol{x}'\boldsymbol{x})^{-1} = \frac{\sigma^2}{\sum_i x_i^2}.$$

$\hat{\beta}$ is a linear combination of the independent normal Y_i, so it is normal:

$$\hat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{\sum_i x_i^2}\right).$$

(d) $E(SSE)$ and an unbiased estimator of σ^2

$SSE = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, with $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{H}$ symmetric, $\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{x}\beta$ and $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. Then:

- $\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) = \sigma^2\,\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(n - \sum_i h_{ii}) = \sigma^2(n-1)$, by (b);
- $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{x}'\boldsymbol{x})/\sum_k x_k^2 = \boldsymbol{x}$, so $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} = \beta(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{x}) = \boldsymbol{0}$, and $\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} = 0$.

Hence

$$E(SSE) = (n-1)\,\sigma^2 + 0 = (n-1)\,\sigma^2,$$

and $MSE = SSE/(n-1)$ is an unbiased estimator of σ^2: one degree of freedom is lost for the single estimated parameter, as $n - p$ with $p = 1$.