STAC67H3F Assignment 2: Solutions

Fall 2026

Question 1

The simulation uses an example student number as its seed. Your numbers will differ from these, but the pattern should be the same.

```
set.seed(1004123456)
x <- 1:40
```

(a)

```
bf_test <- function(e, x) {
  g1 <- x < median(x)                       # group 1: x below the median
  d1 <- abs(e[g1] - median(e[g1]))          # absolute deviations from each group's median
  d2 <- abs(e[!g1] - median(e[!g1]))
  n1 <- length(d1); n2 <- length(d2)
  s2 <- ((n1 - 1) * var(d1) + (n2 - 1) * var(d2)) / (n1 + n2 - 2)   # pooled variance
  t_bf <- (mean(d1) - mean(d2)) / sqrt(s2 * (1/n1 + 1/n2))
  df <- n1 + n2 - 2
  c(t_BF = t_bf, df = df, p_value = 2 * pt(-abs(t_bf), df))
}

y <- 10 + 2 * x + rnorm(40, mean = 0, sd = 2)
fit <- lm(y ~ x)
bf_test(resid(fit), x)
```

```
##       t_BF         df    p_value
##  0.2221195 38.0000000  0.8254103
```

Each group has 20 observations, so the statistic has $40 - 2 = 38$ degrees of freedom. Here the variance is constant and the test does not reject.

(b) For each sample we record two logical values: does each test reject at 5%?

```
one_sample <- function(errors) {
  y <- 10 + 2 * x + errors
  e <- resid(lm(y ~ x))
  g1 <- x < median(x)
  c(BF = bf_test(e, x)[["p_value"]] < 0.05,
    F  = var.test(e[g1], e[!g1])$p.value < 0.05)
}
K <- 2000
rate <- function(make_errors) rowMeans(replicate(K, one_sample(make_errors())))

tab <- rbind(
  "Normal(0, 4)"      = rate(function() rnorm(40, 0, 2)),
  "Student t, 3 df"   = rate(function() rt(40, 3)),
  "Exponential(1) - 1" = rate(function() rexp(40) - 1))
```

```
knitr::kable(tab, digits = 3, col.names = c("Brown-Forsythe", "F test"))
```

	Brown-Forsythe	F test
Normal(0, 4)	0.034	0.046
Student t, 3 df	0.040	0.274
Exponential(1) - 1	0.050	0.262

With normal errors both tests reject about 5% of the time (up to Monte Carlo error, about ± 0.01 with $K = 2000$), as they should. With t_3 or skewed errors the Brown-Forsythe test still rejects about 5% of the time, while the F test rejects 27% and 26% of the time, although the variance is constant. **Only the Brown-Forsythe test keeps its level.** The F test would report non-constant variance on roughly one data set in four where there is none.

(c)

```
power <- rate(function() rnorm(40, 0, 0.1 * x))
power
```

```
##     BF      F
## 0.9335 0.9585
```

Both tests detect this funnel most of the time: power 0.93 for Brown-Forsythe and 0.96 for the F test. The F test is slightly more powerful when the errors are normal, but (b) shows that its rejections cannot be trusted otherwise. In practice we do not know that the errors are normal (checking it is part of the same analysis), so **the Brown-Forsythe test is the one to use.**

Question 2

```
mu <- c(2, -1, 4)
S  <- matrix(c(4, 2, 1,
               2, 9, -3,
               1, -3, 4), nrow = 3, byrow = TRUE)
A  <- rbind(c(1, 1, 1),
            c(1, 0, -2))
EW <- A %*% mu
VW <- A %*% S %*% t(A)
rW <- VW[1, 2] / sqrt(VW[1, 1] * VW[2, 2])
```

(a)

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \end{bmatrix}, \qquad E(\boldsymbol{W}) = \boldsymbol{A}\, E(\boldsymbol{Z}) = \begin{bmatrix} 2 - 1 + 4 \\ 2 - 8 \end{bmatrix} = \begin{bmatrix} 5 \\ -6 \end{bmatrix},$$

$$\boldsymbol{\sigma^2}\{\boldsymbol{W}\} = \boldsymbol{A}\, \boldsymbol{\sigma^2}\{\boldsymbol{Z}\}\, \boldsymbol{A}' = \begin{bmatrix} 17 & 3 \\ 3 & 16 \end{bmatrix}.$$

As a check by hand: $\mathrm{Var}(W_1)$ is the sum of all entries of $\boldsymbol{\sigma^2}\{\boldsymbol{Z}\}$, $4 + 9 + 4 + 2(2 + 1 - 3) = 17$, $\mathrm{Var}(W_2) = 4 + 4(4) - 4(1) = 16$, and $\mathrm{Cov}(W_1, W_2) = 4 - 2(1) + 2 - 2(-3) + 1 - 2(4) = 3$.

(b) $\mathrm{Corr}(W_1, W_2) = 3/\sqrt{17 \times 16} = 0.1819$.

(c) $\mathrm{Cov}(U, Z_1) = \mathrm{Cov}(Z_2, Z_1) - c\,\mathrm{Var}(Z_1) = 2 - 4c$, which is zero for $c = 1/2$. Then

$$\mathrm{Var}(U) = \mathrm{Var}(Z_2) - 2c\,\mathrm{Cov}(Z_1, Z_2) + c^2\,\mathrm{Var}(Z_1) = 9 - 2 + 1 = 8.$$

Note: In general, write $\sigma_1^2 = \mathrm{Var}(Z_1)$, $\sigma_2^2 = \mathrm{Var}(Z_2)$ and $\sigma_{12} = \mathrm{Cov}(Z_1, Z_2)$, so that $\rho = \sigma_{12}/(\sigma_1\sigma_2)$.

Covariance is linear in each argument, so

$$\mathrm{Cov}(U, Z_1) = \mathrm{Cov}(Z_2, Z_1) - c\,\mathrm{Cov}(Z_1, Z_1) = \sigma_{12} - c\,\sigma_1^2,$$

which is zero exactly when $c = \sigma_{12}/\sigma_1^2$. By $\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$ with $a = 1$ and $b = -c$,

$$\mathrm{Var}(U) = \sigma_2^2 - 2c\,\sigma_{12} + c^2\sigma_1^2.$$

Substituting $c = \sigma_{12}/\sigma_1^2$, the middle term is $2\sigma_{12}^2/\sigma_1^2$ and the last one is $(\sigma_{12}^2/\sigma_1^4)\,\sigma_1^2 = \sigma_{12}^2/\sigma_1^2$, so

$$\mathrm{Var}(U) = \sigma_2^2 - 2\frac{\sigma_{12}^2}{\sigma_1^2} + \frac{\sigma_{12}^2}{\sigma_1^2} = \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2} = \sigma_2^2\left(1 - \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}\right) = (1 - \rho^2)\,\sigma_2^2,$$

where the last step recognizes $\sigma_{12}^2/(\sigma_1^2\sigma_2^2) = \rho^2$. Here $\rho = 2/\sqrt{4 \times 9} = 1/3$, and $(1 - 1/9) \times 9 = 8$, as found.

Write $Z_2 = cZ_1 + U$. Since U and Z_1 are uncorrelated, the variances of the two pieces simply add:

$$\mathrm{Var}(Z_2) = c^2\sigma_1^2 + \mathrm{Var}(U) = \rho^2\sigma_2^2 + (1 - \rho^2)\,\sigma_2^2.$$

The variance of Z_2 splits into a part carried by cZ_1, which moves with Z_1, and a part carried by U, which has no linear relation with Z_1 at all. The first part is the fraction ρ^2 of the total. Here $9 = 1 + 8$: Z_1 accounts for one ninth of the variance of Z_2, and the remaining eight ninths cannot be predicted linearly from Z_1.

Among all constants c, this one also makes $\mathrm{Var}(Z_2 - cZ_1)$ as small as possible: the derivative of $\sigma_2^2 - 2c\sigma_{12} + c^2\sigma_1^2$ is $-2\sigma_{12} + 2c\sigma_1^2$, which vanishes at the same $c = \sigma_{12}/\sigma_1^2$. So "leave nothing correlated with Z_1" and "leave as little variance as possible" are the same requirement. Least squares works the same way: the residuals minimize the sum of squares, and they are uncorrelated with x in the sample, since $\sum e_i = 0$ and $\sum x_i e_i = 0$ (Question 3(b)).

Question 3

(a) Since $\boldsymbol{X'X}$ is symmetric, so is its inverse, and $\boldsymbol{H'} = \big(\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\big)' = \boldsymbol{X}\big((\boldsymbol{X'X})^{-1}\big)'\boldsymbol{X'} = \boldsymbol{H}$. Next,

$$\boldsymbol{HH} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\underbrace{\boldsymbol{X'X}(\boldsymbol{X'X})^{-1}}_{\boldsymbol{I}}\boldsymbol{X'} = \boldsymbol{H}, \qquad \boldsymbol{HX} = \boldsymbol{X}\underbrace{(\boldsymbol{X'X})^{-1}\boldsymbol{X'X}}_{\boldsymbol{I}} = \boldsymbol{X}.$$

Then $(\boldsymbol{I} - \boldsymbol{H})' = \boldsymbol{I} - \boldsymbol{H'} = \boldsymbol{I} - \boldsymbol{H}$, $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{HH} = \boldsymbol{I} - \boldsymbol{H}$, and $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}$.

(b) $\boldsymbol{X'e} = \boldsymbol{X'}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \big((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\big)'\boldsymbol{Y} = \boldsymbol{0'Y} = \boldsymbol{0}$, using the symmetry of $\boldsymbol{I} - \boldsymbol{H}$. The rows of $\boldsymbol{X'}$ are $(1, \ldots, 1)$ and $(x_1, \ldots, x_n)$, so the two equations are

$$\sum_{i=1}^{n} e_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_i e_i = 0.$$

The residuals sum to zero (they average out around the fitted line), and they are uncorrelated with x in the sample (no linear trend in x is left in them). These are the normal equations.

(c) Since $\boldsymbol{\sigma^2\{Y\}} = \boldsymbol{\sigma^2\{\varepsilon\}} = \sigma^2\boldsymbol{I}$,

$$\boldsymbol{\sigma^2\{e\}} = (\boldsymbol{I} - \boldsymbol{H})\,\sigma^2\boldsymbol{I}\,(\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}),$$

$$\boldsymbol{\sigma\{e, \hat{Y}\}} = (\boldsymbol{I} - \boldsymbol{H})\,\sigma^2\boldsymbol{I}\,\boldsymbol{H'} = \sigma^2(\boldsymbol{H} - \boldsymbol{HH}) = \boldsymbol{0}.$$

If the errors are normal, $\boldsymbol{e}$ and $\hat{\boldsymbol{Y}}$ are linear functions of the same normal vector $\boldsymbol{Y}$, hence jointly normal, and zero covariance then means **independence**. Under the model, the residuals carry no information about the fitted values: neither their mean nor their spread should change with $\hat{Y}$. So any pattern in a plot of e_i against $\hat{Y}_i$, such as a curve or a funnel, is evidence against the model.

(d) The ith diagonal entry of $\boldsymbol{H}$ is $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$ with $\boldsymbol{x}_i = (1, x_i)'$, the ith row of $\boldsymbol{X}$. So

$$h_{ii} = \frac{1}{S_{xx}}\left(\frac{\sum_j x_j^2}{n} - 2\bar{x}x_i + x_i^2\right).$$

Since $\sum_j x_j^2 = S_{xx} + n\bar{x}^2$, we have $\sum_j x_j^2/n = S_{xx}/n + \bar{x}^2$, and

$$h_{ii} = \frac{1}{S_{xx}}\left(\frac{S_{xx}}{n} + \bar{x}^2 - 2\bar{x}x_i + x_i^2\right) = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

Summing, $\sum_i h_{ii} = n \cdot \frac{1}{n} + \frac{S_{xx}}{S_{xx}} = 2$, the number of parameters. From (c), $\mathrm{Var}(e_i)$ is σ^2 times the ith diagonal entry of $\boldsymbol{I} - \boldsymbol{H}$, that is $\sigma^2(1 - h_{ii})$.

The residuals with the smallest variance are those with the largest h_{ii}, the points whose x_i is **farthest from** $\bar{x}$.

The fitted line is pulled towards such points, because they have the most influence on the slope, so their residuals are systematically small. Even when the errors have constant variance, the residuals do not: residuals at the ends of the x range are less spread out, which is worth remembering when reading a residual plot.

Question 4

```
n5  <- 16
XtX <- matrix(c(16, 967, 967, 63921), 2)
XtY <- c(734, 49968)
YtY <- 39842
detX <- XtX[1, 1] * XtX[2, 2] - XtX[1, 2]^2
XtXi <- matrix(c(XtX[2, 2], -XtX[1, 2], -XtX[1, 2], XtX[1, 1]), 2) / detX
bhat <- drop(XtXi %*% XtY)
SSTO <- YtY - XtY[1]^2 / n5
SSE  <- YtY - sum(bhat * XtY)
SSR  <- SSTO - SSE
MSE  <- SSE / (n5 - 2)
Fst  <- SSR / MSE
R2   <- SSR / SSTO
Vb   <- MSE * XtXi
xbar5 <- XtX[1, 2] / n5
Sxx5  <- XtX[2, 2] - XtX[1, 2]^2 / n5
Xh   <- c(1, 60)
Yh   <- sum(Xh * bhat)
s2Yh <- drop(t(Xh) %*% Vb %*% Xh)
tq   <- qt(0.975, n5 - 2)
ci   <- Yh + c(-1, 1) * tq * sqrt(s2Yh)
pi5  <- Yh + c(-1, 1) * tq * sqrt(MSE + s2Yh)
```

(a) $\det(\boldsymbol{X}'\boldsymbol{X}) = 16 \times 63921 - 967^2 = 87{,}647$, so

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{87647}\begin{bmatrix} 63921 & -967 \\ -967 & 16 \end{bmatrix} = \begin{bmatrix} 0.7293 & -0.011033 \\ -0.011033 & 0.00018255 \end{bmatrix},$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \frac{1}{87647}\begin{bmatrix} 63921 \times 734 - 967 \times 49968 \\ -967 \times 734 + 16 \times 49968 \end{bmatrix} = \begin{bmatrix} -15.9851 \\ 1.02354 \end{bmatrix}.$$

The fitted line is $\hat{Y} = -15.99 + 1.0235\,x$. Each additional centimetre of trunk circumference is associated with an increase of about 1.02 kg in the mean seasonal yield of a tree (about 10.2 kg per 10 cm).

(b) $\sum Y_i = 734$ is the first entry of $\boldsymbol{X}'\boldsymbol{Y}$. $\text{SST} = 39842 - 734^2/16 = 6169.75$ and $\text{SSE} = 39842 - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} = 430.9$, so $\text{SSR} = 5738.85$.

Source	df	SS	MS	F^*
Regression	1	5738.85	5738.85	186.5
Error	14	430.9	30.779	
Total	15	6169.75		

$R^2 = \text{SSR}/\text{SST} = 0.9302$: trunk circumference accounts for about 93% of the variation in yield. Since $F^* = 186.5 > F(0.95; 1, 14) = 4.6$ ($p = 1.8 \times 10^{-9}$), we reject $H_0 : \beta_1 = 0$ at the 5% level: yield is linearly related to circumference.

(c)

$$\boldsymbol{s^2}\{\hat{\boldsymbol{\beta}}\} = \text{MSE}\,(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 22.447 & -0.3396 \\ -0.3396 & 0.005619 \end{bmatrix},$$

so $\text{se}(\hat{\beta}_0) = \sqrt{22.447} = 4.738$ and $\text{se}(\hat{\beta}_1) = \sqrt{0.005619} = 0.075$.

(d) At $x_h = 60$: $\hat{Y}_h = -15.9851 + 1.02354 \times 60 = 45.43$ kg, and

$$s^2\{\hat{Y}_h\} = \boldsymbol{X}_h'\,\boldsymbol{s^2}\{\hat{\boldsymbol{\beta}}\}\,\boldsymbol{X}_h = 1.9247, \qquad s\{\hat{Y}_h\} = 1.3874.$$

(The same value comes from $\text{MSE}\,(1/16 + (60 - 60.4375)^2/5477.9375)$.) With $t(0.975; 14) = 2.145$:

- 95% CI for the mean yield of trees of 60 cm: $45.43 \pm 2.145 \times 1.387 = (42.45,\ 48.4)$ kg.
- 95% PI for the yield of one tree of 60 cm, with $s\{\text{pred}\} = \sqrt{\text{MSE} + s^2\{\hat{Y}_h\}} = 5.719$: $(33.16,\ 57.69)$ kg.

The prediction interval is much wider: it must also account for the variation of a single tree around the mean, σ^2, which does not shrink with more data. Since 60 cm is very close to $\bar{x}$, the confidence interval is almost as narrow as it can be.

Question 5

(a) At a given x, $\log Y \sim N(\mu, \sigma^2)$ with $\mu = \beta_0 + \beta_1 x$. The median of a normal distribution is its mean μ, so $P(\log Y \le \mu) = 1/2$. Since exp is increasing, $P(Y \le e^{\mu}) = P(\log Y \le \mu) = 1/2$: the median of Y is $e^{\mu} = \exp(\beta_0 + \beta_1 x)$. By the hint with $V = \log Y$ and $\tau^2 = \sigma^2$, $E(Y) = E(e^{\log Y}) = \exp(\beta_0 + \beta_1 x + \sigma^2/2)$.

Replacing x by $x + 1$ multiplies both by $\exp(\beta_1)$, because $\exp(\beta_0 + \beta_1(x + 1)) = e^{\beta_1}\exp(\beta_0 + \beta_1 x)$, and the factor $e^{\sigma^2/2}$ in the mean does not depend on x. So on the original scale β_1 acts multiplicatively: one more unit of x changes the median and the mean of Y by $100(e^{\beta_1} - 1)\%$.

(b) Since exp is increasing, the events $\{L < \log Y_{h(\text{new})} < U\}$ and $\{e^L < Y_{h(\text{new})} < e^U\}$ are the same event, so they have the same probability, 0.95. Prediction intervals are statements about a random variable's own values, and they survive any increasing transformation.

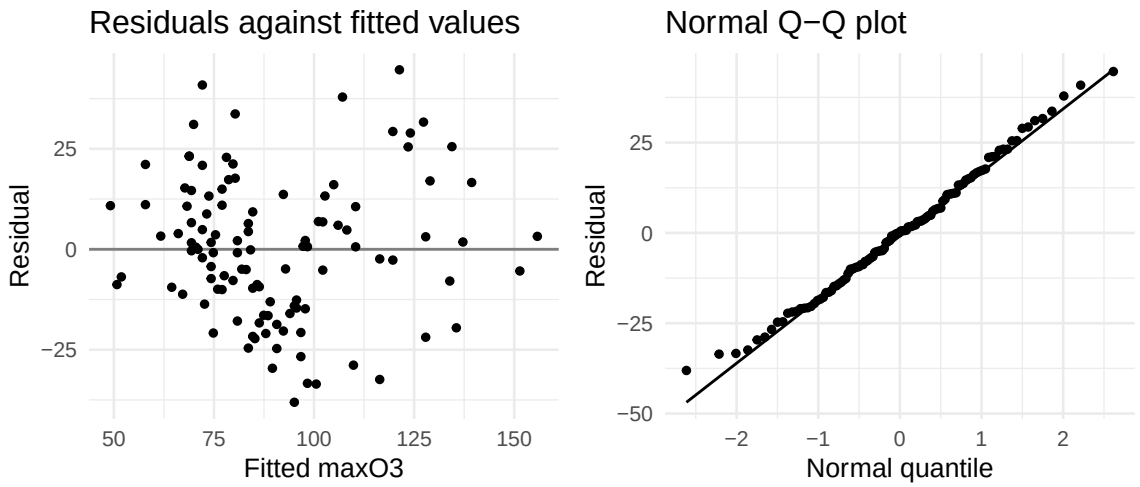
The same argument shows that (e^{L^*}, e^{U^*}) covers $e^{E(\log Y_h)} = \exp(\beta_0 + \beta_1 x_h)$ with probability 0.95. By (a), that is the **median** of Y_h, not its mean: $E(Y_h) = e^{\sigma^2/2}\exp(\beta_0 + \beta_1 x_h)$ is larger. The exponentiated interval is a 95% confidence interval for the median of Y_h. In general $E(g(Y)) \neq g(E(Y))$ for a non-linear g, so a confidence interval for a mean does not transform into one for the mean of the transformed variable.

Question 6

```
oz <- read.table("ozone.txt", header = TRUE)
fit1 <- lm(maxO3 ~ T12, data = oz)
r1 <- data.frame(fitted = fitted(fit1), resid = resid(fit1))
```

(a)

```
p_res <- ggplot(r1, aes(fitted, resid)) +
  geom_hline(yintercept = 0, colour = "grey50") + geom_point(size = 1) +
  labs(x = "Fitted maxO3", y = "Residual", title = "Residuals against fitted values")
p_qq <- ggplot(r1, aes(sample = resid)) + stat_qq(size = 1) + stat_qq_line() +
  labs(x = "Normal quantile", y = "Residual", title = "Normal Q-Q plot")
gridExtra::grid.arrange(p_res, p_qq, nrow = 1)
```




```
bf1 <- bf_test(resid(fit1), oz$T12); bf1
```

```
##         t_BF           df      p_value
##  -2.50982288 110.00000000   0.01353503
```

```
sw1 <- shapiro.test(resid(fit1)); sw1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit1)
## W = 0.99235, p-value = 0.792
```

- **Normality** looks reasonable: the Q-Q plot is close to a line, and the Shapiro-Wilk test does not reject ($p = 0.792$).
- **Constant variance** does not: the residuals spread out more on hot days (large fitted values) than on cool ones, and the Brown-Forsythe test rejects at 5% ($t_{BF} = -2.51$ on 110 df, $p = 0.0135$). The negative sign means the mean absolute deviation is larger in group 2, the warmer days.

The following two assumptions are more difficult to assess, and would need more evidence. * **Linearity** is roughly acceptable, although a run of negative residuals at fitted values between about 85 and 100 suggests the line misses something for mid-range temperatures (see (d)). * **Independence** cannot be judged from these plots; the data are 112 days of one summer, so it should be questioned (see (d)).

(b) A value λ is rejected at the 5% level when $2[\ell(\hat{\lambda}) - \ell(\lambda)] > \chi^2_1(0.95) = 3.841$. So the interval keeps the values of λ whose log-likelihood is within $\chi^2_1(0.95)/2 = 1.921$ of the maximum:

$$\{\lambda : \ \ell(\lambda) \geq \ell(\hat{\lambda}) - \chi^2_1(0.95)/2\}.$$

With `plotit = FALSE`, `boxcox` returns the grid of λ in `x` and the values of $\ell(\lambda)$ in `y`.

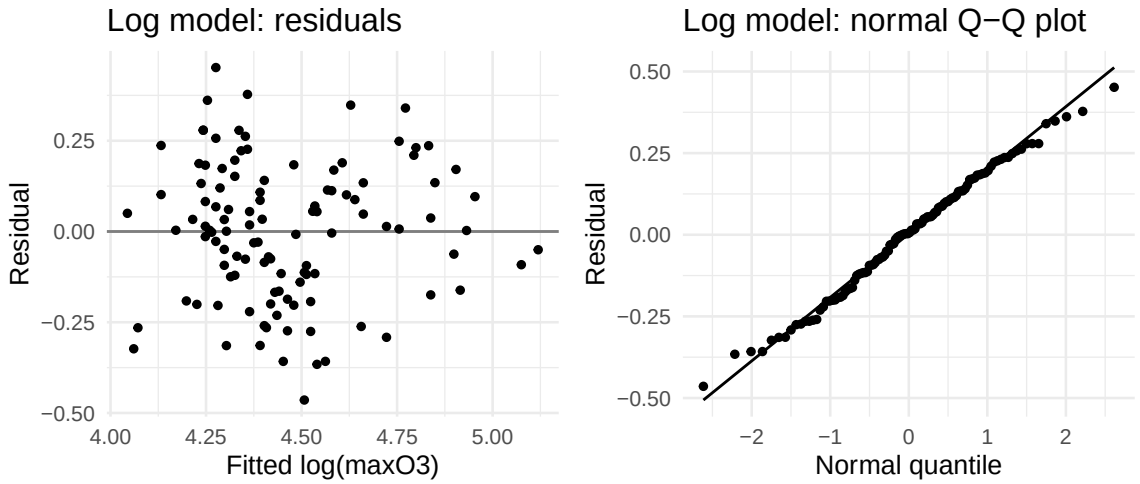
```
bc <- MASS::boxcox(fit1, lambda = seq(-2, 2, 0.01), plotit = FALSE)
lam <- bc$x[which.max(bc$y)]                          # lambda-hat maximizes l(lambda)
keep <- 2 * (max(bc$y) - bc$y) <= qchisq(0.95, 1)     # values not rejected at 5%
```

```r
lam_ci <- range(bc$x[keep])
c(lambda_hat = lam, lower = lam_ci[1], upper = lam_ci[2])
```

```
## lambda_hat      lower      upper
##       0.06      -0.43       0.55
```

$\hat{\lambda} = 0.06$, with 95% interval $(-0.43, 0.55)$. The interval excludes $\lambda = 1$ (no transformation) and contains $\lambda = 0$, which suggests the **log** transformation.

```r
fit2 <- lm(log(maxO3) ~ T12, data = oz)
r2 <- data.frame(fitted = fitted(fit2), resid = resid(fit2))
gridExtra::grid.arrange(
  p_res %+% r2 + labs(x = "Fitted log(maxO3)", title = "Log model: residuals"),
  p_qq %+% r2 + labs(title = "Log model: normal Q-Q plot"), nrow = 1)
```




```r
bf2 <- bf_test(resid(fit2), oz$T12); bf2
```

```
##        t_BF          df     p_value
##  -0.7487372 110.0000000   0.4556136
```

```r
sw2 <- shapiro.test(resid(fit2)); sw2
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit2)
## W = 0.99191, p-value = 0.7549
```

On the log scale the spread is about the same on cool and hot days: Brown-Forsythe no longer rejects ($p = 0.456$), and the residuals are still consistent with normality ($p = 0.755$). The run of negative residuals in the middle is still there.

(c)

```r
b1 <- coef(fit2)[["T12"]]
c(slope = b1, multiplier = exp(b1), percent = 100 * (exp(b1) - 1))
```

```
##      slope multiplier    percent
## 0.05512588 1.05667362 5.66736180
```

```r
new <- data.frame(T12 = c(15, 30))
pi_raw <- predict(fit1, new, interval = "prediction")
pi_log <- exp(predict(fit2, new, interval = "prediction"))
```

```r
out <- data.frame(T12 = new$T12,
                  raw_lwr = pi_raw[, "lwr"], raw_upr = pi_raw[, "upr"],
                  raw_width = pi_raw[, "upr"] - pi_raw[, "lwr"],
                  log_lwr = pi_log[, "lwr"], log_upr = pi_log[, "upr"],
                  log_width = pi_log[, "upr"] - pi_log[, "lwr"])
knitr::kable(out, digits = 1)
```

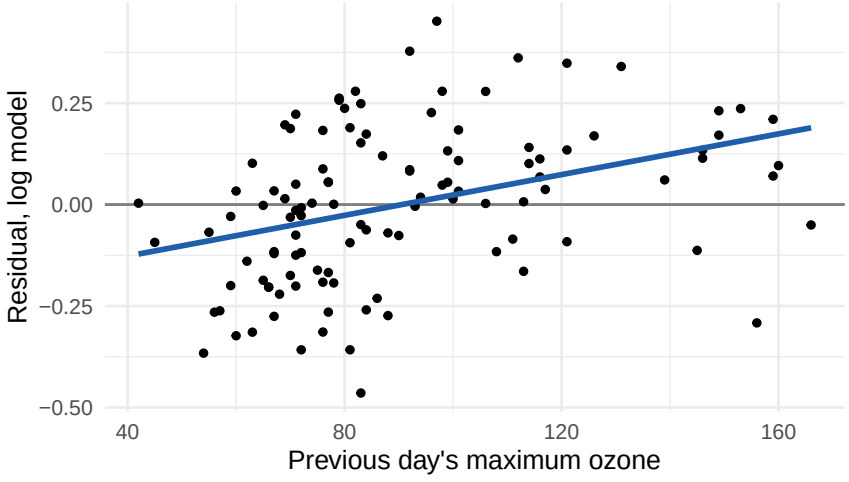
T12	raw_lwr	raw_upr	raw_width	log_lwr	log_upr	log_width
15	19.2	90.0	70.7	41.0	88.6	47.6
30	101.0	172.3	71.3	93.5	203.2	109.7

The slope is $\hat{\beta}_1 = 0.0551$, and $e^{\hat{\beta}_1} = 1.0567$: each additional degree at noon is associated with an increase of about 5.7% in the median (and the mean) daily maximum ozone concentration.

The prediction intervals of the original model have almost the same width, about 71 μg/m^3, at both temperatures: that model assumes the same error variance every day. The intervals from the log model, transformed back by Question 5(b), are narrow on cool days (47.6) and wide on hot days (109.7), and they are not symmetric about the prediction. This matches the funnel in the residual plot of (a): on the original scale the spread grows with the level of ozone, so the original model overstates the uncertainty on cool days and understates it on hot days, which are the days that matter for pollution alerts.

(d)

```r
ggplot(data.frame(max03v = oz$max03v, resid = resid(fit2)), aes(max03v, resid)) +
  geom_hline(yintercept = 0, colour = "grey50") + geom_point(size = 1) +
  geom_smooth(method = "lm", se = FALSE, colour = "#1F5FA9") +
  labs(x = "Previous day's maximum ozone", y = "Residual, log model")
```




```r
cor(oz$max03v, resid(fit2))
```

```
## [1] 0.37239
```

```r
coef(lm(resid(fit2) ~ oz$max03v))
```

```
##  (Intercept)    oz$max03v
## -0.227133287  0.002507781
```

The residuals increase with the previous day's ozone: the least squares line rises clearly (slope 0.00251, correlation 0.37), whereas under the model it should be flat at zero. Since the residuals are on the log scale, each extra 10 μg/m^3 of ozone yesterday goes with about 2.5% more ozone today than the temperature alone

predicts: days that follow a high-ozone day have more ozone than predicted, and days that follow a clean day have less. Two conclusions:

- the errors of consecutive days are probably **not independent**: ozone persists from one day to the next, so knowing yesterday tells us about today's error; and
- the model is **missing a useful predictor**. Adding `maxO3v` to the model, a multiple regression, is the natural remedy, and it would also reduce the unexplained variation and narrow the prediction intervals.