DAY 2, Q4

Consider the moment generating function of Y_i,

$$\mathrm{E}(e^{tY_i}) = \mathrm{E}(e^{t(T_i+U_i)}) = \mathrm{E}(e^{tT_i})\mathrm{E}(e^{tU_i}) = (1-t/\lambda)^{-2}$$

which is the moment generating function of a $Ga(2, \lambda)$ random variable.

2. *More Bang!*

Investors are concerned about the return of their money. Suppose you invested \$1000 in the US stock market last year. After one year, your investment is worth \$1080. The return is then $1080/1000 = 1.08$, and the rate of the return is $\log(1080/1000) = 0.077$.

Now, we consider a model for investment strategy. Label the initial value of of your investment as W_0,(e.g. $W_0 = \$1000$) and the annual return of the investment as R_t (e.g. $R_1 = 1.08$) during year t. The value at the end of the first year is $W_1 = W_0R_1$, and by the end of year T the value is

$$W_T = W_{T-1}R_T = \cdots = W_0R_1R_2\cdots R_T.$$

We suppose $\{R_t\}_{t=0}^{T}$ are i.i.d. random variables.

Based upon the historical data, we have summary statistics of the return R_t for different assets as in Table 1.

	Stocks	T-bills
Mean	1.10	1.05
Std Dev	0.20	0.04

Table 1: *Mean, standard deviations of annual returns, R_t, on US stocks and Treasury Bills*

If we start with \$1000 in each of the stock and the T-bills, we would expect to have \$1,100 in stock and \$1,050 in T-bills after one year. Because the expected value of a product of independent random variables is the product of expectations, we can find the expectations for each investment over a longer horizon given this assumption. Over 20 years, the initial investment of \$1000 in stock grows in expectation to $\$1000 \times (1.1)^{20} = \$6727$. By comparison, the initial investment in T-bills grows to \$2653.

At first glance, the above calculation of expected values seems quite reasonable. However, it uses only the mean of the returns and has no appreciation of the standard deviation - the risks!

To have a deeper understanding of impact of the variance on the long term returns, we convert the product to a sum:

$$\log(W_T) = \log(W_0) + \sum_{t=1}^{T} \log(R_t) = \log(W_0) + \sum_{t=1}^{T} r_t$$

where $r_t = \log(R_t)$ is the *continuously compounded rate of return* for $t = 1, \cdots, T$. Now for large T and by law of large numbers, we have

$$\log(W_T) \approx \log(W_0) + T\mathrm{E}(r_t)$$

(a) Let μ_r be the expectation of the rate of return (i.e. $\mu_r \equiv \mathrm{E}(r_t)$). Approximate μ_r for stocks and T-bills, respectively, using the mean and variance for R_t in Table 1 and the second-order Talor series expansion $\log(1+x) \approx x - x^2/2$

The expected value $\mu_r = \mathrm{E}\log R_t$ is called the expected log return in finance, yet another name for the long-run growth rate. To obtain this, let $\mu_R = \mathrm{E}(R_t), \sigma_R^2 = \mathrm{Var}(R_t)$, write $R_t = \mu_R(1 + (R_t/\mu_R - 1))$, then use the Taylor expansion

$$\log(1+x) \approx x - x^2/2,$$

for $\log[1 + (R_t/\mu_R - 1)]$:

$$\begin{aligned} \mu_r &= \mathrm{E}(\log R_t) \\ &= \mathrm{E}\log\Big\{\mu_R(1 + (R_t/\mu_R - 1))\Big\} \\ &\approx \log\mu_R + \mathrm{E}\Big\{(R_t/\mu_R - 1) - \frac{(R_t/\mu_R - 1)^2}{2}\Big\} \\ &= \log(\mu_R) - \frac{\sigma_R^2}{2\mu_R^2}. \end{aligned} \tag{1}$$

Substituting values of μ_R and σ_R^2 in Table 1 into (1) yields

$$\mu_{r_S} = \log(1.10) - 0.2^2/(2 * 1.1^2) = 0.079$$

for stock and

$$\mu_{r_T} = \log(1.05) - 0.04^2/(2 * 1.05^2) = 0.048$$

for T-bill.

N.B. There are more than one reasonable ways to make use of the Taylor series expansion. For example, note that R_t is close to 1. We can write $R_t = 1 + (R_t - 1)$ and apply the Taylor series expansion to $\log[1 + (R_t - 1)]$. In any event, the values of the μ_r should be no greater than μ_R and be around 0.8 for stock and 0.5 for T-bills.

(b) Let σ_r^2 be the variance of the rate of return (i.e. $\sigma_r^2 \equiv \mathrm{Var}(r_t)$). Approximate σ_r^2 for stocks and T-bills, respectively, using the mean and variance for R_t in Table 1 and the first order Taylor series expansion $\log(1+x) \approx x$?

$$\begin{aligned} \sigma_r^2 &= \mathrm{Var}(r_t) \\ &= \mathrm{Var}\ \log\Big\{\mu_R(1 + (R_t/\mu_R - 1))\Big\} \\ &= \mathrm{Var}\ \log\Big\{(1 + (R_t/\mu_R - 1))\Big\} \\ &\approx \mathrm{Var}\Big\{(R_t/\mu_R - 1)\Big\} \\ &\approx \mathrm{Var}(R_t/\mu_R) \\ &= \sigma_R^2/\mu_R^2 \end{aligned} \tag{2}$$

Thus,

$$\sigma_{r_s}^2 = 0.2/1.1^2 = 0.17$$

for stock and

$$\sigma_{r_T}^2 = 0.04/1.05^2 = 0.036$$

for T-bills.

(c) Now suppose r_t are i.i.d. $N(\mu_r, \sigma_r^2)$ so that R_t follows a *log-normal distribution*. Use the calculated μ_r and σ_r from (a) and (b). Simulate the path of $\{W_t\}_{t=1}^{40}$ 10,000 times for $W_0 = \$1000$ and for stock. What is the largest W_{40} among these 10,000 simulations ?

The largest W_{40} among 10000 simulations is likely to be within $[1 \times 10^8, 3 \times 10^9]$.
Note that $r_t \sim N(\mu_r, \sigma_r^2)$. Thus $\sum_{r=1}^{40} r_t \sim N(40\mu_r, 40\sigma_r^2)$. For stock, $\mu_r = 0.079$ and $\sigma_r^2 = 0.17$ and then $\sum_{r=1}^{40} r_t \sim N(3.16, 6.8)$. The distribution of the maximum value among 10000 i.i.d. draws from $N(3, 8)$ can be obtained by simulation or by EVT. We do this by simulation. The 2.5% and 97.5% empirical quantiles of $\max\{\sum_{r=1}^{40} r_t^{(1)}, \cdots, \sum_{r=1}^{40} r_t^{(10000)}\}$ from 20000 replication are 12 and 15 respectively. Consequently, $\max\{W_{40}^{(1)}, \cdots, W_{40}^{(10000)}\}$ are likely to be within $[\$1000e^{12} = \$1 \times 10^8, \$1000e^{15} = \$3 \times 10^9]$.

(d) There are millions of investors seeking profits in the US stock market. A few of them, such as Warren Buffet or Peter Lynch, are famous for consistently generating huge positive returns over time. Their performances are often attributed to their knowledgable investment strategy. Criticize these statements based on your statistical thinking.

This is an open question with no definite answer. One important perspective from a statistical point of view should be related to the issue of identifying signals (here, investment skills) in the context of multiple testing (here, millions of investors).
We can argue that Warren Buffett can be simply "got lucky". For example, in the simulation of part(c), all of the paths are from the same distribution. They all start with \$1000, but there will be a lucky few that can generate huge amount of return with millions of money in the end. It may be simply luck that produced the Warren Buffetts out of the millions of investors. Of course, we are not concluding that that Warren Buffett became successful by sheer luck. We simply point out the difficulty in separating skill from luck, a problem that bedevils fund investors.

DAY2, Q5

BIRD DATA PROBLEM

(a) The scatterplot strongly suggests the following linear model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \cdot x_{2i} + \epsilon_i,$$

Where, for $i = 1, 2, \ldots, 32$,

Y_i = amount of energy

x_{1i} = species
($0 = A$; $1 = B$)

x_{2i} = temperature

As usual, assume $\epsilon_i \sim \text{iid } N(0, \sigma^2)$

Note that the proposed model is simple parallel lines regression. This is a parsimonious model; however, we should check for a lack of parellelism.

The fitted model is

$$\widehat{Y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_{2i},$$

Where

2

least squares estimates. $\begin{cases} \hat{\beta}_0 = 36.47408 \\ \hat{\beta}_1 = 4.48125 \\ \hat{\beta}_2 = -0.44577 \end{cases}$

The error variance estimate is

$$\hat{\sigma}^2 = MSE = 0.35080.$$

To see if an interaction term

$$\text{Species} * \text{temp} \leftrightarrow x_{1i} * x_{2i}$$

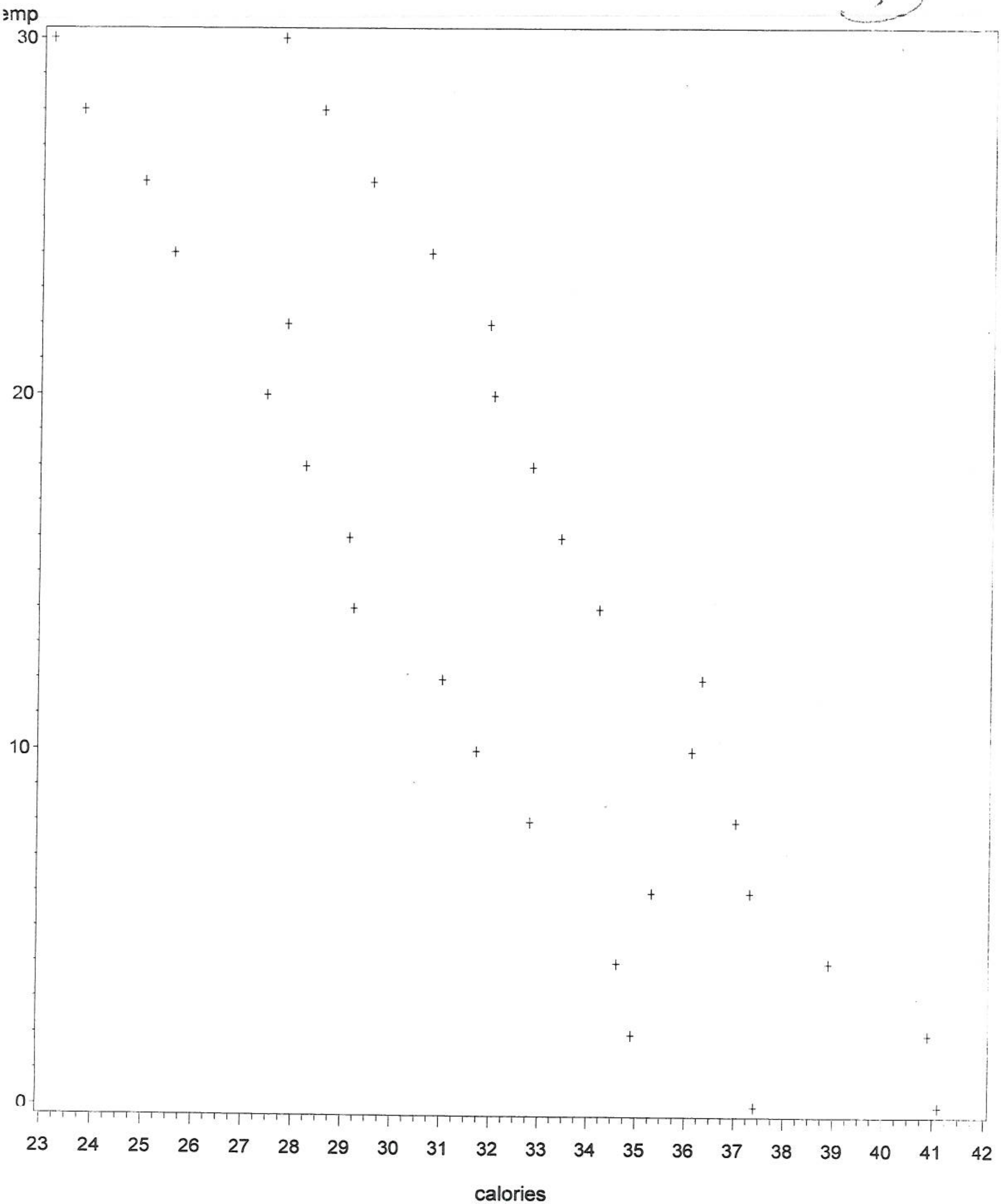
is needed, I used proc glm to fit

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

The p-value for

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

is large ($p = 0.3826$). There is insufficient evidence against H_0 (stick w/ parallel lines model).



emp
30
20
10
0
23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
calories

Diagnostics → for parallel lines model fit

Residual vs. Fitted plot → looks ok; maybe a slight amount of increasing variability as temperature increases.

Qq-plot → slight departure from normality, but perhaps still ok.

The diagnostic conclusions are not "ideal," but reasonable.

No evidence of severe outliers/ influential observations.

We settle on the parallel lines model as stated earlier.

(b) A 95% prediction interval is

(28.5389, 31.0361).

For a new bird of species A subjected to 15 deg C, we are 95% confident this bird's energy response will be between 28.5389 and 31.0361 calories.

(c) With temperature fixed, we would like to test

$$H_0: \beta_1 = 0 \qquad \text{(no difference)}$$
$$H_1: \beta_1 > 0 \qquad \text{(species B > species A)}$$

The test statistic is

$$t = 21.40$$

Therefore, Ho would be rejected at any reasonable level of significance.

There is strong evidence that birds from species B burn more calories than birds from species A when the temperature is fixed.

9

```
data birds;
input species temp calories;
cards;
0 0 37.4
0 2 34.9
0 4 34.6
0 6 35.3
0 8 32.8
0 10 31.7
0 12 31.0
0 14 29.2
0 16 29.1
0 18 28.2
0 20 27.4
0 22 27.8
0 24 25.5
0 26 24.9
0 28 23.7
0 30 23.1
1 0 41.1
1 2 40.9
1 4 38.9
1 6 37.3
1 8 37.0
1 10 36.1
1 12 36.3
1 14 34.2
1 16 33.4
1 18 32.8
1 20 32.0
1 22 31.9
1 24 30.7
1 26 29.5
1 28 28.5
1 30 27.7
;

proc gplot;
plot temp*calories;
run;

proc reg data = birds;
model calories = species temp/all;
output out=diag p=yhat r=resid h=hat;
plot r.*p.;
plot nqq.*student. /haxis=-3 to 3 by 1 vaxis=-3 to 3 by 1;
run;


data new;
input species temp calories;
lines;
0 15 .
;
run;

data both;
set birds new; /* merging two datasets */
run;

proc reg data=both;
model calories = species temp/clm cli;
run;

proc glm data = birds;
model calories = species temp species*temp;
run;
```

7

The REG Procedure
Model: MODEL1

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	species	temp	calories
Intercept	32	16	480	1024.9
species	16	16	240	548.3
temp	480	240	9920	14161
calories	1024.9	548.3	14161	33536.95

Disregard

The REG Procedure
Model: MODEL1
Dependent Variable: calories

Number of Observations Read	32
Number of Observations Used	32

X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	species	temp	calories
Intercept	0.1452205882	-0.0625	-0.005514706	36.474080882
species	-0.0625	0.125	0	4.48125
temp	-0.005514706	0	0.0003676471	-0.445772059
calories	36.474080882	4.48125	-0.445772059	10.173253676

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	701.15143	350.57572	999.36	<.0001
Error	29	10.17325	0.35080		
Corrected Total	31	711.32469			

Root MSE	0.59229	R-Square	0.9857
Dependent Mean	32.02813	Adj R-Sq	0.9847
Coeff Var	1.84927		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	--Heteroscedasticity Consistent- Standard Error	t Value	Pr > |t|
Intercept	1	36.47408	0.22571	161.60	<.0001	0.26523	137.52	<.0001
species	1	4.48125	0.20940	21.40	<.0001	0.19935	22.48	<.0001
temp	1	-0.44577	0.01136	-39.25	<.0001	0.01052	-42.36	<.0001

Parameter Estimates

Variable	DF	Type I SS	Type II SS	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Partial Corr Type I	Squared Semi-partial Corr Type II
Intercept	1	32826	9160.95020	0	.	.	.
species	1	160.65281	160.65281	0.47524	0.22585	0.22585	0.22585
temp	1	540.49862	540.49862	0.87169	0.75985	0.98153	0.75985

8

The REG Procedure
Model: MODEL1
Dependent Variable: calories

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	37.4000	36.4741	0.2257	36.0125	36.9357	35.1777	37.7704	0.9259
2	34.9000	35.5825	0.2091	35.1549	36.0102	34.2979	36.8672	-0.6825
3	34.6000	34.6910	0.1937	34.2948	35.0872	33.4165	35.9655	-0.0910
4	35.3000	33.7994	0.1799	33.4315	34.1674	32.5334	35.0655	1.5006
5	32.8000	32.9079	0.1681	32.5642	33.2516	31.6487	34.1671	-0.1079
6	31.7000	32.0164	0.1586	31.6920	32.3407	30.7623	33.2704	-0.3164
7	31.0000	31.1248	0.1519	30.8141	31.4356	29.8742	32.3754	-0.1248
8	29.2000	30.2333	0.1485	29.9295	30.5370	28.9844	31.4821	-1.0333
9	29.1000	29.3417	0.1485	29.0380	29.6455	28.0929	30.5906	-0.2417
10	28.2000	28.4502	0.1519	28.1394	28.7609	27.1996	29.7008	-0.2502
11	27.4000	27.5586	0.1586	27.2343	27.8830	26.3046	28.8127	-0.1586
12	27.8000	26.6671	0.1681	26.3234	27.0108	25.4079	27.9263	1.1329
13	25.5000	25.7756	0.1799	25.4076	26.1435	24.5095	27.0416	-0.2756
14	24.9000	24.8840	0.1937	24.4878	25.2802	23.6095	26.1585	0.0160
15	23.7000	23.9925	0.2091	23.5648	24.4201	22.7078	25.2771	-0.2925
16	23.1000	23.1009	0.2257	22.6393	23.5625	21.8046	24.3973	-0.000919
17	41.1000	40.9553	0.2257	40.4937	41.4170	39.6590	42.2517	0.1447
18	40.9000	40.0638	0.2091	39.6361	40.4914	38.7792	41.3484	0.8362
19	38.9000	39.1722	0.1937	38.7760	39.5685	37.8977	40.4468	-0.2722
20	37.3000	38.2807	0.1799	37.9127	38.6487	37.0147	39.5467	-0.9807
21	37.0000	37.3892	0.1681	37.0454	37.7329	36.1300	38.6483	-0.3892
22	36.1000	36.4976	0.1586	36.1733	36.8220	35.2436	37.7516	-0.3976
23	36.3000	35.6061	0.1519	35.2953	35.9168	34.3555	36.8566	0.6939
24	34.2000	34.7145	0.1485	34.4108	35.0183	33.4657	35.9634	-0.5145
25	33.4000	33.8230	0.1485	33.5192	34.1267	32.5741	35.0718	-0.4230
26	32.8000	32.9314	0.1519	32.6207	33.2422	31.6809	34.1820	-0.1314
27	32.0000	32.0399	0.1586	31.7155	32.3642	30.7859	33.2939	-0.0399
28	31.9000	31.1483	0.1681	30.8046	31.4921	29.8892	32.4075	0.7517
29	30.7000	30.2568	0.1799	29.8888	30.6248	28.9908	31.5228	0.4432
30	29.5000	29.3653	0.1937	28.9690	29.7615	28.0907	30.6398	0.1347
31	28.5000	28.4737	0.2091	28.0461	28.9014	27.1891	29.7583	0.0263
32	27.7000	27.5822	0.2257	27.1205	28.0438	26.2858	28.8785	0.1178

Output Statistics

<table>
<tr><th>Obs</th><th>Std Error Residual</th><th>Student Residual</th><th>-2-1 0 1 2</th><th>Cook's D</th></tr>
<tr><td>1</td><td>0.548</td><td>1.691</td><td>| |*** |</td><td>0.162</td></tr>
<tr><td>2</td><td>0.554</td><td>-1.232</td><td>| **| |</td><td>0.072</td></tr>
<tr><td>3</td><td>0.560</td><td>-0.163</td><td>| | |</td><td>0.001</td></tr>
<tr><td>4</td><td>0.564</td><td>2.659</td><td>| |***** |</td><td>0.240</td></tr>
<tr><td>5</td><td>0.568</td><td>0.190</td><td>| | |</td><td>0.001</td></tr>
</table>

The REG Procedure
Model: MODEL1
Dependent Variable: calories

Output Statistics

<table>
<tr><th>Obs</th><th>Std Error Residual</th><th>Student Residual</th><th>-2-1 0 1 2</th><th>Cook's D</th></tr>
<tr><td>6</td><td>0.571</td><td>-0.554</td><td>| *| |</td><td>0.008</td></tr>
<tr><td>7</td><td>0.572</td><td>-0.218</td><td>| | |</td><td>0.001</td></tr>
<tr><td>8</td><td>0.573</td><td>-1.802</td><td>| ***| |</td><td>0.073</td></tr>
<tr><td>9</td><td>0.573</td><td>-0.422</td><td>| | |</td><td>0.004</td></tr>
<tr><td>10</td><td>0.572</td><td>-0.437</td><td>| | |</td><td>0.004</td></tr>
<tr><td>11</td><td>0.571</td><td>-0.278</td><td>| | |</td><td>0.002</td></tr>
<tr><td>12</td><td>0.568</td><td>1.995</td><td>| |*** |</td><td>0.116</td></tr>
<tr><td>13</td><td>0.564</td><td>-0.488</td><td>| | |</td><td>0.008</td></tr>
<tr><td>14</td><td>0.560</td><td>0.0286</td><td>| | |</td><td>0.000</td></tr>
<tr><td>15</td><td>0.554</td><td>-0.528</td><td>| *| |</td><td>0.013</td></tr>
<tr><td>16</td><td>0.548</td><td>-0.0017</td><td>| | |</td><td>0.000</td></tr>
<tr><td>17</td><td>0.548</td><td>0.264</td><td>| | |</td><td>0.004</td></tr>
<tr><td>18</td><td>0.554</td><td>1.509</td><td>| |*** |</td><td>0.108</td></tr>
<tr><td>19</td><td>0.560</td><td>-0.486</td><td>| | |</td><td>0.009</td></tr>
<tr><td>20</td><td>0.564</td><td>-1.738</td><td>| ***| |</td><td>0.102</td></tr>
<tr><td>21</td><td>0.568</td><td>-0.685</td><td>| *| |</td><td>0.014</td></tr>
<tr><td>22</td><td>0.571</td><td>-0.697</td><td>| *| |</td><td>0.012</td></tr>
<tr><td>23</td><td>0.572</td><td>1.212</td><td>| |** |</td><td>0.035</td></tr>
<tr><td>24</td><td>0.573</td><td>-0.897</td><td>| *| |</td><td>0.018</td></tr>
<tr><td>25</td><td>0.573</td><td>-0.738</td><td>| *| |</td><td>0.012</td></tr>
<tr><td>26</td><td>0.572</td><td>-0.230</td><td>| | |</td><td>0.001</td></tr>
<tr><td>27</td><td>0.571</td><td>-0.0699</td><td>| | |</td><td>0.000</td></tr>
<tr><td>28</td><td>0.568</td><td>1.323</td><td>| |** |</td><td>0.051</td></tr>
<tr><td>29</td><td>0.564</td><td>0.785</td><td>| |* |</td><td>0.021</td></tr>
<tr><td>30</td><td>0.560</td><td>0.241</td><td>| | |</td><td>0.002</td></tr>
<tr><td>31</td><td>0.554</td><td>0.0474</td><td>| | |</td><td>0.000</td></tr>
<tr><td>32</td><td>0.548</td><td>0.215</td><td>| | |</td><td>0.003</td></tr>
</table>

Sum of Residuals	0
Sum of Squared Residuals	10.17325
Predicted Residual SS (PRESS)	12.36943

7

The REG Procedure
Model: MODEL1
Dependent Variable: calories

Number of Observations Read	33
Number of Observations Used	32
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	701.15143	350.57572	999.36	<.0001
Error	29	10.17325	0.35080		
Corrected Total	31	711.32469			

Root MSE	0.59229	R-Square	0.9857
Dependent Mean	32.02813	Adj R-Sq	0.9847
Coeff Var	1.84927		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	36.47408	0.22571	161.60	<.0001
species	1	4.48125	0.20940	21.40	<.0001
temp	1	-0.44577	0.01136	-39.25	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: calories

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	37.4000	36.4741	0.2257	36.0125	36.9357	35.1777	37.7704	0.9259
2	34.9000	35.5825	0.2091	35.1549	36.0102	34.2979	36.8672	-0.6825
3	34.6000	34.6910	0.1937	34.2948	35.0872	33.4165	35.9655	-0.0910
4	35.3000	33.7994	0.1799	33.4315	34.1674	32.5334	35.0655	1.5006
5	32.8000	32.9079	0.1681	32.5642	33.2516	31.6487	34.1671	-0.1079
6	31.7000	32.0164	0.1586	31.6920	32.3407	30.7623	33.2704	-0.3164
7	31.0000	31.1248	0.1519	30.8141	31.4356	29.8742	32.3754	-0.1248
8	29.2000	30.2333	0.1485	29.9295	30.5370	28.9844	31.4821	-1.0333
9	29.1000	29.3417	0.1485	29.0380	29.6455	28.0929	30.5906	-0.2417
10	28.2000	28.4502	0.1519	28.1394	28.7609	27.1996	29.7008	-0.2502
11	27.4000	27.5586	0.1586	27.2343	27.8830	26.3046	28.8127	-0.1586
12	27.8000	26.6671	0.1681	26.3234	27.0108	25.4079	27.9263	1.1329
13	25.5000	25.7756	0.1799	25.4076	26.1435	24.5095	27.0416	-0.2756
14	24.9000	24.8840	0.1937	24.4878	25.2802	23.6095	26.1585	0.0160
15	23.7000	23.9925	0.2091	23.5648	24.4201	22.7078	25.2771	-0.2925
16	23.1000	23.1009	0.2257	22.6393	23.5625	21.8046	24.3973	-0.000919
17	41.1000	40.9553	0.2257	40.4937	41.4170	39.6590	42.2517	0.1447
18	40.9000	40.0638	0.2091	39.6361	40.4914	38.7792	41.3484	0.8362
19	38.9000	39.1722	0.1937	38.7760	39.5685	37.8977	40.4468	-0.2722
20	37.3000	38.2807	0.1799	37.9127	38.6487	37.0147	39.5467	-0.9807
21	37.0000	37.3892	0.1681	37.0454	37.7329	36.1300	38.6483	-0.3892
22	36.1000	36.4976	0.1586	36.1733	36.8220	35.2436	37.7516	-0.3976
23	36.3000	35.6061	0.1519	35.2953	35.9168	34.3555	36.8566	0.6939
24	34.2000	34.7145	0.1485	34.4108	35.0183	33.4657	35.9634	-0.5145
25	33.4000	33.8230	0.1485	33.5192	34.1267	32.5741	35.0718	-0.4230
26	32.8000	32.9314	0.1519	32.6207	33.2422	31.6809	34.1820	-0.1314
27	32.0000	32.0399	0.1586	31.7155	32.3642	30.7859	33.2939	-0.0399
28	31.9000	31.1483	0.1681	30.8046	31.4921	29.8892	32.4075	0.7517
29	30.7000	30.2568	0.1799	29.8888	30.6248	28.9908	31.5228	0.4432
30	29.5000	29.3653	0.1937	28.9690	29.7615	28.0907	30.6398	0.1347
31	28.5000	28.4737	0.2091	28.0461	28.9014	27.1891	29.7583	0.0263
32	27.7000	27.5822	0.2257	27.1205	28.0438	26.2858	28.8785	0.1178
33	.	29.7875	0.1481	29.4847	30.0903	28.5389	31.0361	.

Sum of Residuals	0
Sum of Squared Residuals	10.17325
Predicted Residual SS (PRESS)	12.36943

95% PI

10

The GLM Procedure

Number of Observations Read	32
Number of Observations Used	32

The GLM Procedure

Dependent Variable: calories

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	701.4294669	233.8098223	661.60	<.0001
Error	28	9.8952206	0.3534007		
Corrected Total	31	711.3246875			

R-Square	Coeff Var	Root MSE	calories Mean
0.986089	1.856104	0.594475	32.02813

Source	DF	Type I SS	Mean Square	F Value	Pr > F
species	1	160.6528125	160.6528125	454.59	<.0001
temp	1	540.4986213	540.4986213	1529.42	<.0001
species*temp	1	0.2780331	0.2780331	0.79	0.3826

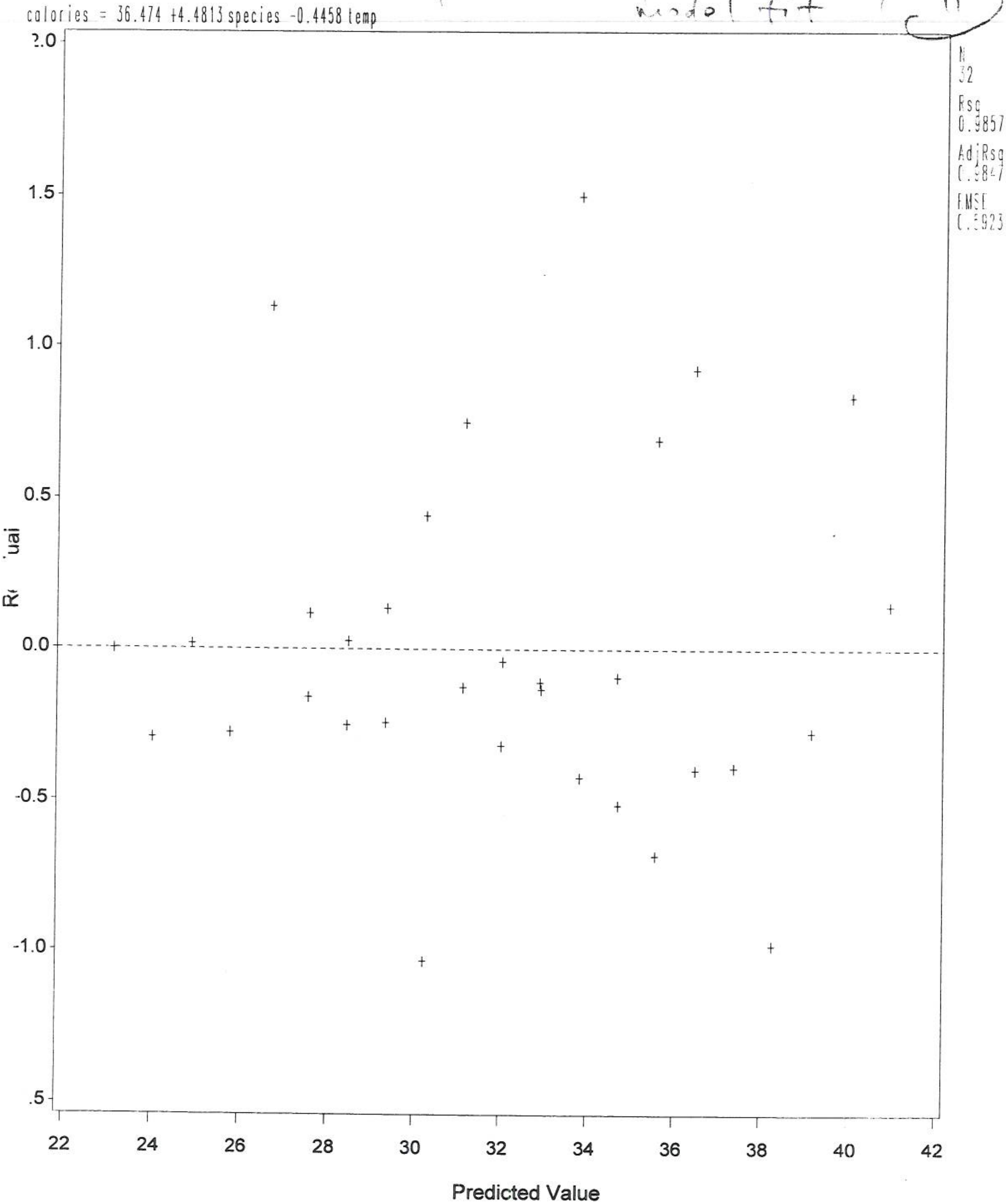
Source	DF	Type III SS	Mean Square	F Value	Pr > F
species	1	38.2888093	38.2888093	108.34	<.0001
temp	1	282.6470588	282.6470588	799.79	<.0001
species*temp	1	0.2780331	0.2780331	0.79	0.3826

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	36.62573529	0.28382139	129.05	<.0001
species	4.17794118	0.40138405	10.41	<.0001
temp	-0.45588235	0.01611997	-28.28	<.0001
species*temp	0.02022059	0.02279708	0.89	0.3826

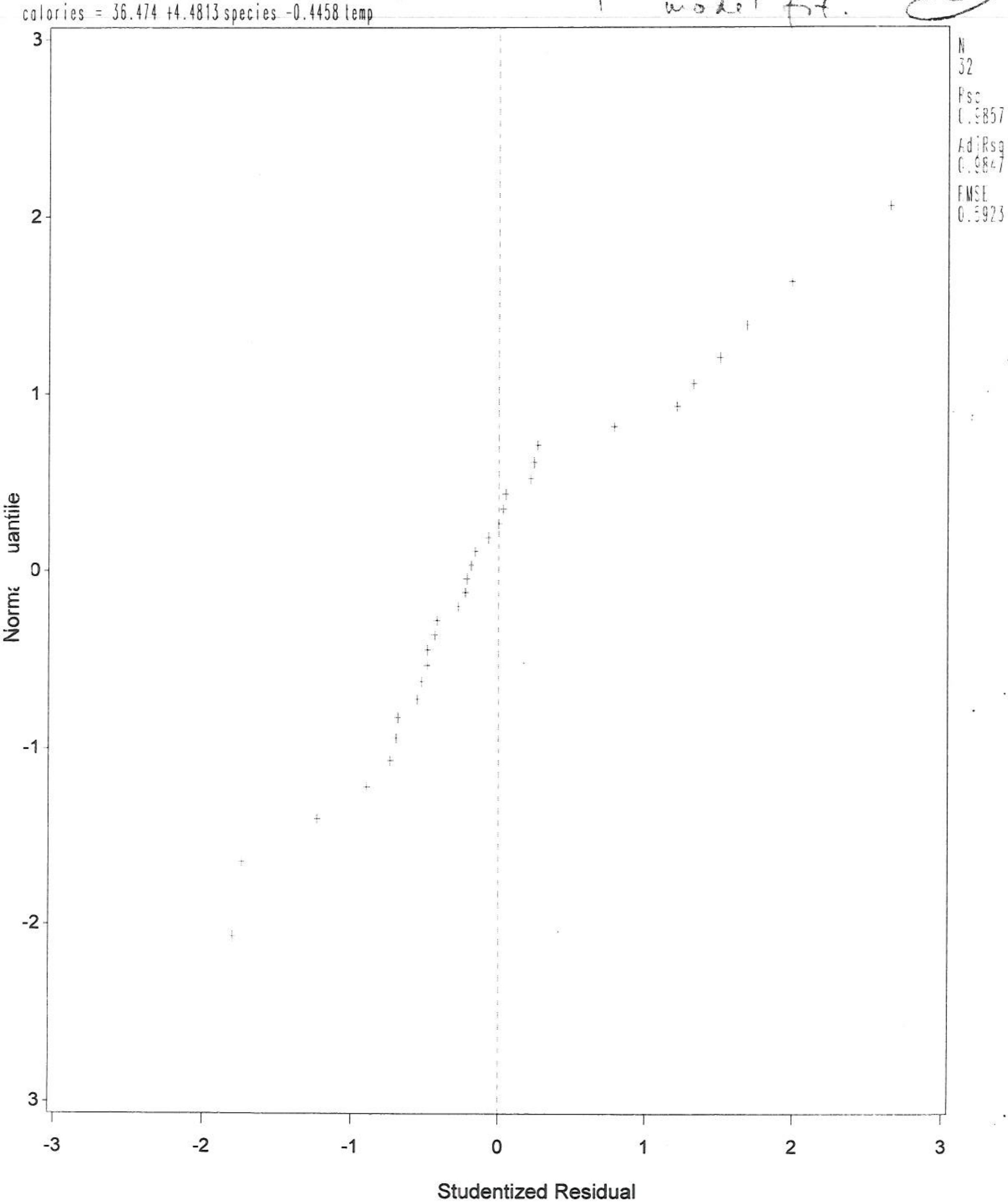
test for interaction.

Residual plot from parallel lines model fit

11



QQ plot from parallel lines model fit. 12



DAY 2, Q6.

1LD2. A suitable model for the data could be a linear mixed effects model, with Lab as a fixed effect and Inspector as a random effect.

So,

$$Y_{ij} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ij} \quad , i=1,\dots,a \quad j=1,\dots,b$$

$\sum \alpha_i = 0$

$\beta_j \sim N(0, \sigma_\beta^2)$ independently

$\varepsilon_{ij} \sim N(0, \sigma^2)$ independently.

We first check for additivity using Tukey's test. This is where an interaction term is included of the form

$$(\alpha\beta)_{ij} = D\alpha_i\beta_j$$

and the test is $H_0: D = 0$

vs $H_a: D \neq 0$.

Here $F = 1.599$ p-value $= 0.212$, $\hat{D} = 10.192$ and there is no evidence against H_0.

Note this test is strictly for two fixed effects but it can be used as an informal guide in this situation.

The normal Q-Q plot looks straight and the residual plot shows a roughly broad horizontal band of points, and so there is no reason to doubt the equal variance and normality assumptions.

The ANOVA table

	df	SS	MS.
Lab	5	0.0770	0.0154
Inspector	10	0.0436	0.0044
Error	50	0.1674	0.0033

Lab effect.

To test $H_0: \alpha_i = 0$

vs H_a: not all $\alpha_i = 0$

we have $F = \frac{MSA}{MSAB} \sim F_{5,50}$ under H_0.

Here MSAB = MSE as there is no interaction term.

$F = \frac{0.0154}{0.0033} = 4.6017$ $\quad P(F_{5,50} > 4.6017) = 0.0016$.

So, there is strong evidence for a difference in response due to Labs.

Inspector effect $\quad H_0: \sigma_\beta^2 = 0$

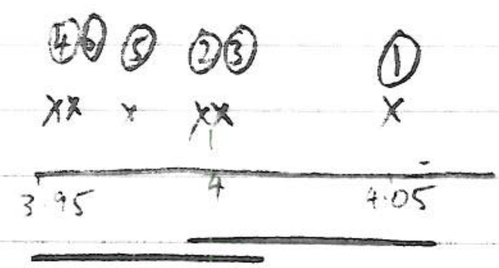
$H_a: \sigma_\beta^2 > 0$.

$F = \frac{MSB}{MSE} = \frac{0.0044}{0.0033} = 1.3034$ $\quad P(F_{10,50} > 1.3034) = 0.2546$.

So, no evidence for variability in inspectors

From the Tukey multiple comparisons procedure we see that the only significant differences are between Labs

	diff	p-value
4 and 1	−0.102	0.001
5 and 1	−0.075	0.039
6 and 1	−0.089	0.0087



Model.

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$\sum \alpha_i = 0$, $\beta_j \sim N(0, \sigma_\beta^2)$ independently, $(\alpha\beta)_{ij} \sim N(0, \sigma_{\alpha\beta}^2)$ independently.

indpt.

Test: $H_0: \sigma_{\alpha\beta}^2 = 0$

$H_a: \sigma_{\alpha\beta}^2 > 0$

$$F = \frac{MSAB}{MSE} \sim F_{(a-1)(b-1),(n-1)ab} \text{ under } H_0.$$

Reject H_0 if $F > F_{1-\alpha, (a-1)(b-1), (n-1)ab}$

Test $H_0: \alpha_i = 0$

vs $H_a: \alpha_i \neq 0$ for at least one i.

$$F = \frac{MSA}{MSAB} \sim F_{a-1,\,(a-1)(b-1)} \quad \text{under } H_0.$$

Reject H_0 if $F > F_{1-\alpha,\,a-1,\,(a-1)(b-1)}$

Test $H_0: \sigma_\beta^2 = 0$

vs $H_a: \sigma_\beta^2 > 0$

$$F = \frac{MSB}{MSE} \sim F_{b-1,\,(n-1)(ab)} \quad \text{under } H_0.$$

Reject H_0 if $F > F_{1-\alpha,\,b-1,\,(n-1)(ab)}$.