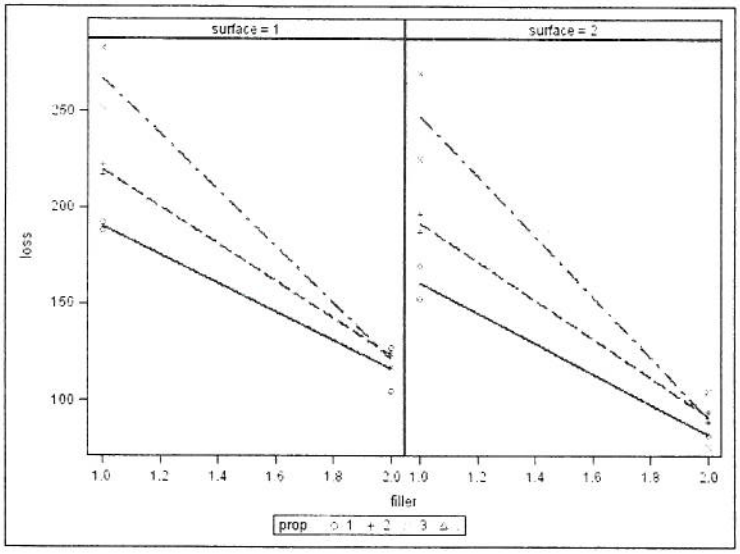
May 2013 PhD Qual Solutions

Problem 1

(a) Here is an interaction plot:



One plot looks like the other shifted vertically, so surface will likely be included only as an additive effect. However, the effect of proportion is quite different among the two levels of filler, so there will likely be a filler*prop interaction term. Broadly, surface=2 tends to reduce fabric loss. Overall, filler=2 yields considerably less loss than filler=1. When filler=2 is used, there appears to be little to no difference among the proportions in terms of loss (for a fixed surface). However, for filler=1, smaller proportion yields less loss, and it appears to be significantly so.

(b) A fit of the full interaction model yields:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	90092.12500	8190.19318	40.86	<.0001
Error	12	2405.50000	200.45833		
Corrected Total	23	92497.62500			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
surface	1	5017.04167	5017.04167	25.03	0.0003
filler	1	70959.37500	70959.37500	353.99	<.0001
surface*filler	1	57.04167	57.04167	0.28	0.6035
prop	2	7969.00000	3984.50000	19.88	0.0002
surface*prop	2	44.33333	22.16667	0.11	0.8962
filler*prop	2	6031.00000	3015.50000	15.04	0.0005
surface*filler*prop	2	14.33333	7.16667	0.04	0.9650

Here's a fit of the model with the three-way interaction, prop*surface, and filler*surface dropped:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	89976.41667	14996.06944	101.12	<.0001
Error	17	2521.20833	148.30637		
Corrected Total	23	92497.62500			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
surface	1	5017.04167	5017.04167	33.83	<.0001

filler	1	70959.37500	70959.37500	478.46	<.0001
prop	2	7969.00000	3984.50000	26.87	<.0001
filler*prop	2	6031.00000	3015.50000	20.33	<.0001

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		91.2916667 B	6.57693130	13.88	<.0001
surface	1	28.9166667 B	4.97169275	5.82	<.0001
surface	2	0.0000000 B	.	.	.
filler	1	151.7500000 B	8.61122444	17.62	<.0001
filler	2	0.0000000 B	.	.	.
prop	1	-6.7500000 B	8.61122444	-0.78	0.4439
prop	2	1.5000000 B	8.61122444	0.17	0.8638
prop	3	0.0000000 B	.	.	.
filler*prop	1 1	-75.5000000 B	12.17811038	-6.20	<.0001
filler*prop	1 2	-53.5000000 B	12.17811038	-4.39	0.0004
filler*prop	1 3	0.0000000 B	.	.	.
filler*prop	2 1	0.0000000 B	.	.	.
filler*prop	2 2	0.0000000 B	.	.	.
filler*prop	2 3	0.0000000 B	.	.	.

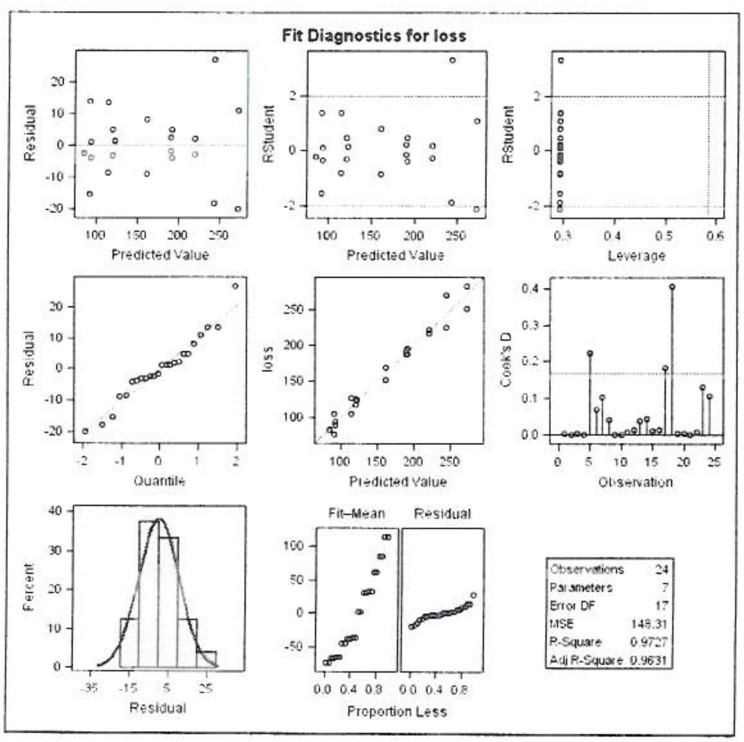
Here's the F-statistic for dropping the three interactions at the same time along with the p-value:

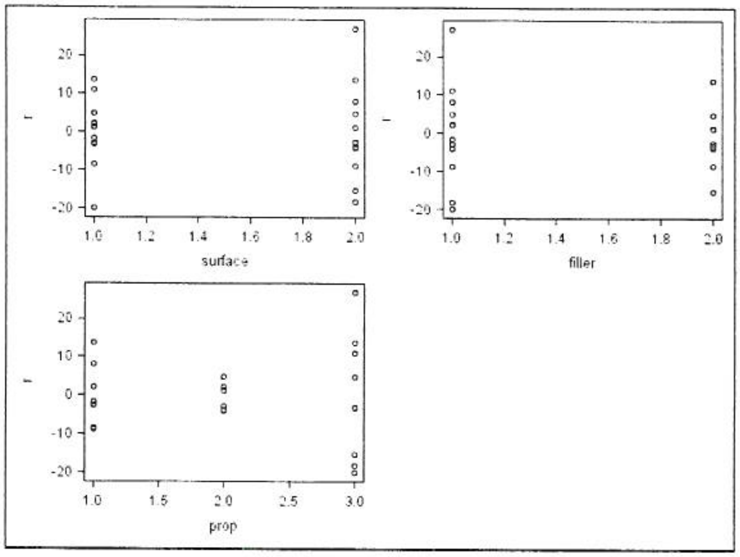
Obs	_TYPE_	_FREQ_	minss	mindf	maxss	maxdf	fstar	pvalue
1	0	2	2405.5	12	2521.21	17	0.11544	0.98654

With p=0.987, we can easily drop the three interactions at the 5% level. The fitted model is

$$\begin{aligned} E(L) &= 91.3 + 28.9\ I\{S=1\} + 151.8\ I\{F=1\} - 6.8\ I\{P=1\} + 1.5\ I\{P=2\} \\ &\quad -75.5\ I\{F=1\}I\{P=1\} - 53.4\ I\{F=1\}I\{P=2\} \end{aligned}$$

(c) Here are the requested diagnostic plots





Constant variance seems reasonable overall, although the variability seems to be quite a bit smaller for prop=2 than for prop=1 or prop=3. With a deleted residual (2nd plot in top row) over 3, there is one quite ill-fit point. There are also three points with fairly large Cook's distances. It might be worth taking a closer look at these points. The normal probability plot looks reasonably straight, so the normality assumption seems satisfied.

(d) A qualitative assessment is provided in part (a). Surface=2 significantly reduces fabric loss by an estimated 29 grams: a confidence interval for this difference is (18, 39) grams.

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha
surface	surf2-surf1	-28.9167	4.9717	17	-5.82	<.0001	<.0001	0.05

The effect of proportion depends on the filler used. Here is a table of pairwise differences in proportion at the two levels of filler; multiple comparisons are adjusted using Tukey's method, and capped at FER≤ 0.05. Bonferroni might work better here.

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
filler*prop	prop1-prop2 @ filler=1	30.2500	8.6112	17	3.51	0.0027	0.0027
filler*prop	prop3-prop1 @ filler=1	82.2500	8.6112	17	9.55	<.0001	<.0001
filler*prop	prop3-prop2 @ filler=1	52.0000	8.6112	17	6.04	<.0001	<.0001
filler*prop	prop1-prop2 @ filler=2	8.2500	8.6112	17	0.96	0.3515	0.3515
filler*prop	prop3-prop1 @ filler=2	6.7500	8.6112	17	0.78	0.4439	0.4439
filler*prop	prop3-prop2 @ filler=2	-1.5000	8.6112	17	-0.17	0.8638	0.8638

For filler=2, all three adjusted p=values are larger than 0.05, and we find no significant differences among the three proportions. For filler=1 proportion does make a significant difference, with increased proportion yielding significantly more loss. This confirms the interaction plot.

To see whether the effect of filler is significant, there are several approaches one might take. Let's simply look at whether there's differences in filler at the three proportions, again using Tukey:

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
filler*prop	filler1-filler2 @ prop=1	76.2500	8.6112	17	8.85	<.0001	<.0001
filler*prop	filler1-filler2 @ prop=2	98.2500	8.6112	17	11.41	<.0001	<.0001
filler*prop	filler1-filler2 @ prop=3	151.75	8.6112	17	17.62	<.0001	<.0001

Filler 2 results in significantly less loss at all three proportions.

Prob 2

$Y_i \sim Exp(\beta x_i)$, $i=1\cdots,n$.

(a) $Q(\beta) = \sum_{i=1}^{n} (Y_i - \beta x_i)^2$

$\frac{dQ}{d\beta} = -\sum_i (Y_i - \beta x_i)\cdot x_i = 0$. [set the first derivative to be 0].

$\Rightarrow \tilde{\beta} = \left(\sum x_i^2\right)^{-1} \sum_i x_i Y_i$ is the least squares estimator of β.

$E(\tilde{\beta}) = \left(\sum x_i^2\right)^{-1} \sum x_i \cdot E(Y_i) = \left(\sum x_i^2\right)^{-1} \cdot \sum x_i \cdot \beta x_i$ $\quad$ Since $E(Y_i) = \beta x_i$

$= \left(\sum x_i^2\right)^{-1} \cdot \left(\sum x_i^2\right) \cdot \beta = \beta$.

$Var(\tilde{\beta}) = Var\left(\left(\sum x_i^2\right)^{-1} \sum_i x_i Y_i \right)$

$= \left(\sum x_i^2\right)^{-2} \cdot Var\left(\sum_i x_i Y_i\right)$

$= \left(\sum x_i^2\right)^{-2} \cdot \sum_i x_i^2 Var(Y_i)$ $\quad$ $\boxed{\text{due to the independence among } Y_i\text{'s}}$

$= \left(\sum x_i^2\right)^{-2} \cdot \sum_i x_i^2 \cdot (x_i\beta)^2$

$= \left(\sum_i x_i^2\right)^{-2} \cdot \left(\sum_i x_i^4\right) \beta^2$

(b). The observed likelihood is

$$L(\beta) = \prod_i \frac{1}{\beta x_i} \exp\left(-\frac{y_i}{\beta x_i}\right)$$

The log-likelihood is

$$l(\beta) = -n\log\beta - \sum_i \log x_i - \frac{1}{\beta}\sum_i \frac{y_i}{x_i}$$

Consider the first derivative w.r.t. β.

$$\frac{dl}{d\beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_i \frac{y_i}{x_i}.$$

Second derivative

$$\frac{d^2 l}{\partial\beta^2} = +\frac{n}{\beta^2} - \frac{2}{\beta^3}\sum_i \frac{y_i}{x_i}.$$

Setting $\frac{dl}{\partial\beta} = 0$, we get

$$\hat\beta = \frac{1}{n}\sum_i \frac{y_i}{x_i} \quad \text{as the unique solution.}$$

$$\left.\frac{\partial^2 l}{\partial\beta^2}\right|_{\hat\beta} = \frac{n}{\hat\beta^2} - \frac{2}{\hat\beta^3}\sum_i \frac{y_i}{x_i}$$

$$= \frac{1}{\hat\beta^2}\left[n - \frac{2}{\hat\beta}\sum_i \frac{y_i}{x_i}\right]$$

$$= \frac{1}{\hat\beta^2}\left[n - 2n\right]$$

$$= -\frac{n}{\hat\beta^2} < 0$$

So, $\hat\beta = \frac{1}{n}\sum_i \frac{y_i}{x_i}$ is the MLE.

(c) Find the exact sampling distribution of $\hat{\beta} = \frac{1}{n}\sum_i \frac{Y_i}{X_i}$.

Since $Y_i \sim Exp(\beta x_i)$ with mean equal to βx_i,

$$\frac{Y_i}{X_i} \sim Exp(\beta), \text{ i.e., } Ga(1, \beta), \text{ for } i = 1, \ldots, n.$$

Further $\sum_{i=1}^{n} \frac{Y_i}{X_i} \sim Ga(n, \beta)$ using the fact $\frac{Y_i}{X_i}$'s are independent.

So, $\hat{\beta} = \frac{1}{n}\sum_i \frac{Y_i}{X_i} \sim Ga(n, \frac{\beta}{n})$

(d) $Var(\hat{\beta}) = n \times \left(\frac{\beta}{n}\right)^2 = \frac{1}{n}\beta^2$ using the result in (c).

$$Var(\tilde{\beta}) = \frac{\sum_i X_i^4}{(\sum X_i^2)} \beta^2.$$

$$\text{So, } Var(\tilde{\beta}) - Var(\hat{\beta}) = \frac{(\sum_i X_i^4)}{(\sum X_i^2)^2}\beta^2 - \frac{1}{n}\beta^2$$

$$= \frac{\beta^2}{n(\sum_i X_i^2)^2} \cdot \left[n \cdot (\sum_i X_i^4) - (\sum_i X_i^2)^2 \right] \geq 0$$

The second item is nonnegative by using the Cauchy-Schwarz inequality.

$$(\sum_i a_i^2)(\sum_i b_i^2) \geq (\sum a_i b_i)^2, \text{ with } a_i = X_i^2 \text{ and } b_i = 1, \forall i.$$

The equality holds when $X_1^2 = X_2^2 = \cdots = X_n^2$ only.

So, in general, $\hat{\beta}$ has a smaller variance than $\tilde{\beta}$.

$\boxed{\text{Alternative solution:}}$ one can argue that $\sum_i \frac{y_i}{x_i}$ is the sufficient and complete statistic of β using the result for Exponential family. Since $\hat{\beta}$ is unbiased for β and is a function of $\sum \frac{y_i}{x_i}$, $\hat{\beta}$ is the UMVUE of β. This will lead to $Var(\hat{\beta}) \leq Var(\tilde{\beta})$ since $\tilde{\beta}$ is also unbiased for β.

Prob 3

(a) Since $U = F_Z(Z)$ and $F(\cdot)$ is a CDF function, the domain of U is $[0,1]$.

For $u \in [0,1]$,

$$F_U(u) = P(U \le u) = P(F_Z(Z) \le u)$$

$$= P(Z \le F_Z^{-1}(u))$$ The inverse function exists and is increasing since $F_Z(\ )$ is strictly increasing.

$$= \int_{\infty}^{F_Z^{-1}(u)} 1 \, dF_Z(z)$$

$$= F_Z(z) \Big|_{-\infty}^{F_Z^{-1}(u)}$$

$$= F_Z(F_Z^{-1}(u)) - \underbrace{F_Z(-\infty)}_{=0}$$

$$= u - 0$$

$$= u$$

For $u < 0$, $F_U(u) = P(U \le u) = P(F_Z(Z) \le u) \le P(F_Z(Z) < 0) = 0$

For $u > 1$, $F_U(u) = P(F_Z(Z) \le u) \ge P(F_Z(Z) \le 1) = 1$.

So, $$F_U(u) = \begin{cases} 0 & \text{if } u < 0 \\ u & \text{if } u \in [0,1] \\ 1 & \text{if } u > 1 \end{cases}$$

i.e., $U = F_Z(Z) \sim U_{(0,1)}$.

(b). $E(F_Z(X)) = E[\Phi(X)]$ here I use $\Phi()$ as the CDF of $Z \sim N(0,1)$.

$$= \int_{-\infty}^{\infty} \Phi(x) \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx. \quad \text{Since } X \sim N(\mu, \sigma^2).$$

$$= \int_{-\infty}^{\infty} \Phi(\mu+\sigma y) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy. \quad \underline{\text{let } \frac{x-\mu}{\sigma} = y.}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t+\mu+\sigma y)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dt\, dy$$

use fact
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t+x)^2} dt$$

$$= \int_{-\infty}^{0} \int_{-\infty}^{\infty} \frac{1}{2\pi} \cdot e^{-\frac{1}{2}\left[(t+\mu+\sigma y)^2 + y^2\right]} dy\, dt$$

by Fubini's Theorem. The integrand is positive.

$$= \int_{-\infty}^{0} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}\left[(1+\sigma^2)y^2 + 2(t+\mu)\sigma y + (t+\mu)^2\right]} dy\, dt$$

$$= \int_{-\infty}^{0} \int_{-\infty}^{\infty} \frac{\sqrt{1+\sigma^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}(1+\sigma^2)\left(y + \frac{t+\mu}{1+\sigma^2}\right)^2} dy \cdot \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1+\sigma^2}} e^{-\frac{1}{2}\left[(t+\mu)^2 - \frac{(t+\mu)^2\sigma^2}{1+\sigma^2}\right]} dt$$

$$= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi} \cdot \sqrt{1+\sigma^2}} e^{-\frac{1}{2}\frac{(t+\mu)^2}{1+\sigma^2}} dt$$

$$= \int_{-\infty}^{\frac{\mu}{\sqrt{1+\sigma^2}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \quad \text{let } \frac{t+\mu}{\sqrt{1+\sigma^2}} = z$$

$$= \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) = F_Z\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right)$$

(c). $X_i \overset{iid}{\sim} N(\mu, 1)$.

$$\tau(\mu) = P_\mu(X_1 < c) = P\left(Z < \frac{c-\mu}{1}\right) = F_Z(c-\mu) = \Phi(c-\mu).$$

Define $Y = \sqrt{\frac{n}{n-1}}\,(c - \bar{X})$. | We know $\bar{X} \sim N(\mu, \frac{1}{n})$

Then $Y \sim N\left(\sqrt{\frac{n}{n-1}} \cdot (c-\mu),\ \underbrace{\frac{n}{n-1} \cdot \frac{1}{n} = \frac{1}{n-1}}_{\sigma_Y^2}\right)$

Using the result in (b),

$$E[\Phi(Y)] = \Phi\left(\frac{\mu_Y}{\sqrt{1+\sigma_Y^2}}\right)$$

$$= \Phi\left(\frac{\sqrt{\frac{n}{n-1}} \cdot (c-\mu)}{\sqrt{1 + \frac{1}{n-1}}}\right)$$

$$= \Phi(c-\mu) = \tau(\mu)$$

This suggests that $\Phi(Y)$, i.e., $F_Z\left(\sqrt{\frac{n}{n-1}}\,(c-\bar{X})\right)$, is an unbiased estimator of $\tau(\mu)$.

It is known that $\bar{X}$ is a sufficient and complete statistic of μ since $N(\mu, 1)$ belongs to the exponential family.

Since $F_Z\left(\sqrt{\frac{n}{n-1}}\,(c-\bar{X})\right)$ is a function of $\bar{X}$ and is unbiased for $\tau(\mu)$,

$F_Z\left(\sqrt{\frac{n}{n-1}}\,(c-\bar{X})\right)$ is the UMVUE of $\tau(\mu)$.