Josh Rhoades
Biostatistics
Take Home Final

'1.

```
> anova(lm((COST~COPAY + GS + RI), data=drugcost))

> anova(lm((COST~RXPM + AGE + F + MM +COPAY + GS + RI), data=drugcost))
Analysis of Variance Table

Response: COST
          Df   Sum Sq  Mean Sq F value    Pr(>F)
RXPM       1 0.000089 0.000089  0.0130 0.9102085
AGE        1 0.045344 0.045344  6.6204 0.0177322 *
F          1 0.024645 0.024645  3.5983 0.0716769 .
MM         1 0.001618 0.001618  0.2363 0.6319360
COPAY      1 0.012735 0.012735  1.8593 0.1871420
GS         1 0.110616 0.110616 16.1503 0.0006213 ***
RI         1 0.000175 0.000175  0.0256 0.8744682
Residuals 21 0.143832 0.006849
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We see that the F values of GS, F, COPAY and AGE are higher than the rest and worth looking into...

```
> pf(6.6204, 1, 21, lower.tail=F)
[1] 0.01773185
> pf(3.5983, 1, 21, lower.tail=F)
[1] 0.0716776
> pf(1.8593, 1, 21, lower.tail=F)
[1] 0.1871457
> pf(16.1503, 1, 21, lower.tail=F)
[1] 0.0006212593
```

The only signigicant (to .05) factors are Age and GS.

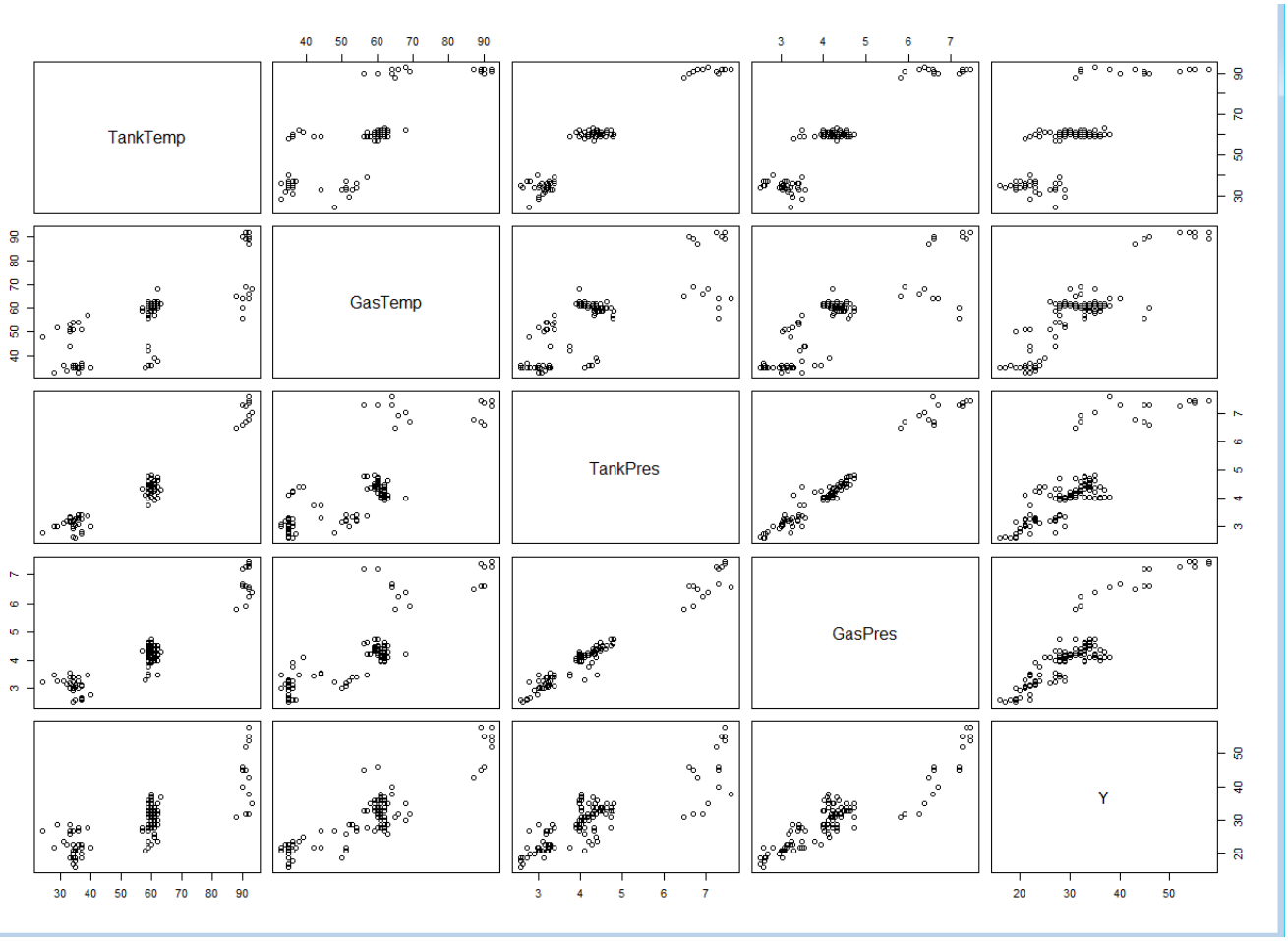
```
Analysis of Variance Table

Response: COST
          Df   Sum Sq  Mean Sq F value   Pr(>F)
COPAY      1 0.000032 0.000032  0.0035 0.953511
GS         1 0.110411 0.110411 12.1001 0.001863 **
RI         1 0.000492 0.000492  0.0540 0.818216
Residuals 25 0.228120 0.009125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> pf(0.0035, 1, 25, lower.tail=F)
[1] 0.9532944
> pf(12.1001, 1, 25, lower.tail=F)
[1] 0.001863141
> pf(0.0540, 1, 25, lower.tail=F)
[1] 0.8181355

We see that GS is highly related to COST, and that COPAY and RI have little effect on COST.

For some reason the eastcoast/midatlantic region have the highest cost, perhaps there is another factor missing, such as population density, distribution distance or health risk in that particular area.

2A.



I would eliminate Tank Temperature, as it has the loosest correlation with Y.

2B.

model<- lm(Y~TankPres + GasPres + GasTemp, data=sniffer)
> summary(model)

```
Call:
lm(formula = Y ~ TankPres + GasPres + GasTemp, data = sniffer)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6373 -1.5419  0.0535  1.4436  6.6534

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45089    1.02806   0.439    0.662
TankPres    -5.73444    1.24613  -4.602 1.04e-05 ***
GasPres     10.83605    1.53201   7.073 1.06e-10 ***
GasTemp      0.15456    0.03591   4.303 3.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.779 on 121 degrees of freedom
Multiple R-squared:  0.8907,    Adjusted R-squared:  0.888
F-statistic: 328.7 on 3 and 121 DF,  p-value: < 2.2e-16
```

Without including TankTwmp, all three other vatiables have a significant effect on Y.

```
> model1<- lm(Y~TankPres + TankTemp + GasPres + GasTemp, data=sniffer)
> summary(model1)

Call:
lm(formula = Y ~ TankPres + TankTemp + GasPres + GasTemp, data = sniffer)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5425 -1.2938  0.0495  1.2259  7.0413

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.15391    1.03489   0.149   0.8820
TankPres    -4.05962    1.58000  -2.569   0.0114 *
TankTemp    -0.08269    0.04857  -1.703   0.0912 .
GasPres      9.85744    1.62515   6.066 1.57e-08 ***
GasTemp      0.18971    0.04118   4.606 1.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.758 on 120 degrees of freedom
Multiple R-squared:  0.8933,    Adjusted R-squared:  0.8897
F-statistic: 251.1 on 4 and 120 DF,  p-value: < 2.2e-16
```

When I put Tank Temp into the model, not only is tank temp not significantly related to Y, but TankPres now is not as significant as it was before.

Y is a measure of escaping hydrocarbons in the air, so if pressure is greater in the tank more vapors will escape as the tank is filled. Also, as gas temperature increases, the amount of vapors also increases, therefore all three of these variables are important predictors of Y. TankTemp is loosely tied to Y only because of the effect it may have on the Tank Pressure, but it is better to just use TankPres. The best model uses TankPres, GasPres and GasTemp as predictors of Y.

3A.

```
> fishmodelno3<-glm(period~ age + length, family=gaussian, data=walleyeno3)
> summary(fishmodelno3)

Call:
glm(formula = period ~ age + length, family = gaussian, data = walleyeno3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0795  -0.5777   0.3153   0.3711   0.6256

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.0448609  0.0600708  34.041  < 2e-16 ***
age          0.0799103  0.0103494   7.721 1.65e-14 ***
length      -0.0022916  0.0002972  -7.711 1.78e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2282264)

    Null deviance: 587.12  on 2512  degrees of freedom
Residual deviance: 572.85  on 2510  degrees of freedom
AIC: 3423.8

Number of Fisher Scoring iterations: 2

> 1-pchisq(587.12-572.85, 2512-2510)
[1] 0.0007967258
```

There is no significant difference between periods. Also age and length are strongly correlated.

3B.

```
 walleyeno1<-read.table("walleyeno1.txt", header=T)
>
> fishmodelno1<-glm(period~ age + length, family=gaussian, data=walleyeno1)
> summary(fishmodelno1)

Call:
glm(formula = period ~ age + length, family = gaussian, data = walleyeno1)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.95246  -0.23516  -0.03521   0.29839   1.42063

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.9986909  0.0800355   37.47   <2e-16 ***
age          0.1765231  0.0120548   14.64   <2e-16 ***
length      -0.0053424  0.0003755  -14.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4307283)

    Null deviance: 1359.0  on 2936  degrees of freedom
Residual deviance: 1263.8  on 2934  degrees of freedom
AIC: 5866.1

Number of Fisher Scoring iterations: 2

> 1-pchisq(1359.0-1263.8,2936-2934)
[1] 0
```

4.

```
mantelhaen.test(cancer)

        Mantel-Haenszel chi-squared test with continuity correction

data:  cancer
Mantel-Haenszel X-squared = 7.2873, df = 1, p-value = 0.006944
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.245332 3.703541
sample estimates:
common odds ratio
         2.147589

cancersum<-apply(cancer,c(1,2),sum)
> cancersum
        Cancer
Exposure Yes No
    High  58 58
    Low   46 84
> (58*84)/(58*46)
[1] 1.826087
```

```
 chisq.test(cancersum)

        Pearson's Chi-squared test with Yates' continuity correction

data:  cancersum
X-squared = 4.7836, df = 1, p-value = 0.02873
```

Common odds ratio of 2.147589, the p-value = 0.006944 suggests that the odds ratio is not equal to 1

The odds ratios of 1.826087 and 2.147589 are similar, cannot determine is location a confounding factor. => The test could not determine if cancer incidence was different at any location.

5A.

```
> wilcox.test(plasma$BETAPLASMA,plasma$SEX, alternative="g")

        Wilcoxon rank sum test with continuity correction

data:  plasma$BETAPLASMA and plasma$SEX
W = 98910, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0

 t.test(plasma$BETAPLASMA,plasma$SEX, alternative="g")

        Welch Two Sample t-test

data:  plasma$BETAPLASMA and plasma$SEX
t = 18.2355, df = 314.002, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 171.0152      Inf
sample estimates:
 mean of x  mean of y
189.892063   1.866667
```

Both tests show that betaplasma is significantly different between sexes.

5B.

```
 wilcox.test(plasma$BETAPLASMA,plasma$SMOKESTAT, alternative="g")

        Wilcoxon signed rank test with continuity correction

data:  plasma$BETAPLASMA
V = 49455, p-value < 2.2e-16
alternative hypothesis: true location is greater than 0

> t.test(plasma$BETAPLASMA,plasma$SMOKESTAT, alternative="g")
```

One Sample t-test

data: plasma$BETAPLASMA
t = 18.4166, df = 314, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 172.8819   Inf
sample estimates:
mean of x
 189.8921

Both tests show that smkoing has a significant effect on betaplasma.

> wilcox.test(plasma$RETPLASMA,plasma$SMOKESTAT, alternative="g")

Wilcoxon signed rank test with continuity correction

data: plasma$RETPLASMA
V = 49770, p-value < 2.2e-16
alternative hypothesis: true location is greater than 0

> t.test(plasma$RETPLASMA,plasma$SMOKESTAT, alternative="g")

One Sample t-test

data: plasma$RETPLASMA
t = 51.2145, df = 314, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 583.3734   Inf
sample estimates:
mean of x
 602.7905

Both tests show that smoking has a significant effect on retplasma.

The investigators were also interested in carotene and retinol, so lets see if smoking effects their levels as well.

5C.
I am unsure what else to investigate, we have seen that smoking effects levels of betaplasma and retplasma and that was the main goal of the study. However, perhaps smokers eat less foods containing these compounds, ie smoker eat unhealthy food. Lets see if there is a correlation between BETADIET and smoking.

```
> wilcox.test(plasma$BETADIET,plasma$SMOKESTAT, alternative="g")

        Wilcoxon signed rank test with continuity correction

data:  plasma$BETADIET
V = 49770, p-value < 2.2e-16
alternative hypothesis: true location is greater than 0

> t.test(plasma$BETADIET,plasma$SMOKESTAT, alternative="g")

        One Sample t-test

data:  plasma$BETADIET
t = 26.3186, df = 314, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 2048.604      Inf
sample estimates:
mean of x
 2185.603
```

Both tests show that smoking has a significant correlation with BETADIET, this is odd because we would expect everyone to consume similar amounts of BETACONSUME. Perhaps smokings is associated with bad diets, and it does not actually cause a physiological reason for the lower levels in BETAPLASMA..
Lets see if this holds true for RETDIET...

```
> wilcox.test(plasma$RETDIET,plasma$SMOKESTAT, alternative="g")

        Wilcoxon signed rank test with continuity correction

data:  plasma$RETDIET
V = 49770, p-value < 2.2e-16
alternative hypothesis: true location is greater than 0
```

Again we see that smoking is likely just associated with other poor health choices, such as diets low in retinol. Therefore the findings of part B are missleading and are confounded by dietary preferences.