STAT 230 Homework 5

Owen Forman

1. The following is a sequential ANOVA table (with some entries removed) for the regression of Y on x_1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	3	2309	??	??	0.01
Residual	12	1650	137.5		

(1a) Fill in the question marks.

(1b) What is the estimate of σ for the regression model?

(1c) The degrees of freedom for x is 3. Explain why this means that x is a categorical variable and say how many categories the variable can take.

(1d) Interpret the p-value in the table.

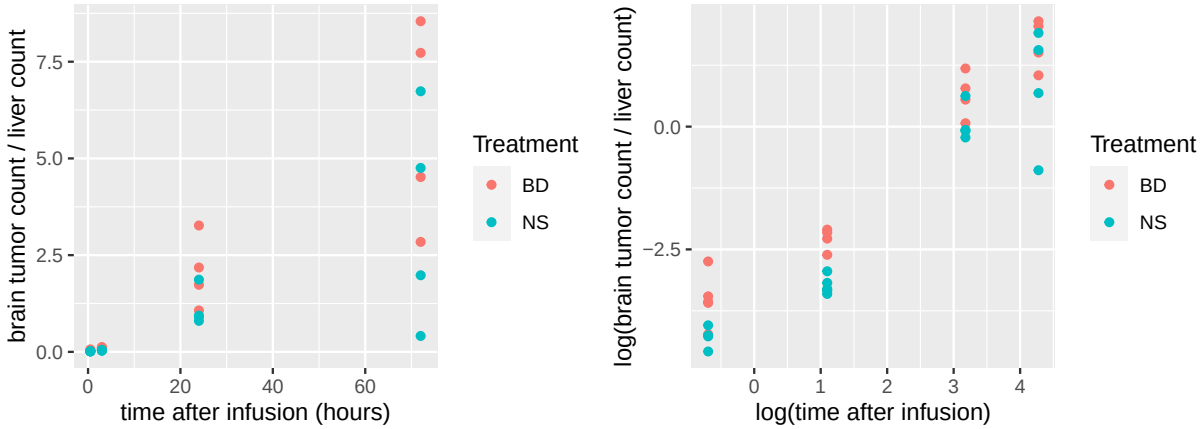
2. (Modification of Exercises 12.14 and 12.15). Reread Section 11.1.2 and then answer the questions below. The response variable is the ratio of the brain tumor antibody count to the liver antibody count, so first calculate that as a new variable:

```
data(case1102, package = "Sleuth3")
case1102$Y <- case1102$Brain / case1102$Liver
```

(2a) Plot the response variable, `Y`, against the length of time after infusion (`Time`) and color-code by Treatment. Make the same plot but taking the logarithm of `Y` and `Time`. First, explain why the book suggests taking logarithms of these variables. Then describe whether there appears to be an interaction between the treatment and and length of time after infusion.

```
origPlot <- ggplot(case1102, aes(y = Y, x = Time, color = Treatment)) + geom_point() +
  labs(y = "brain tumor count / liver count", x = "time after infusion (hours)")
loglogPlot <- ggplot(case1102, aes(y = log(Y), x = log(Time), color = Treatment)) + geom_point() +
  labs(y = "log(brain tumor count / liver count)", x = "log(time after infusion)")

origPlot | loglogPlot
```



(2b) The book asks the following on p. 315: "Was the antibody concentration in the tumor increased by the use of the blood-brain barrier disruption infusion? Is so, by how much? Do the answers to these two questions depend on the length of time after the infusion (from 1/2 to 72 hours)? What is the effect of treatment on antibody concentration after weight loss, total weight, and other covariates are accounted for?"

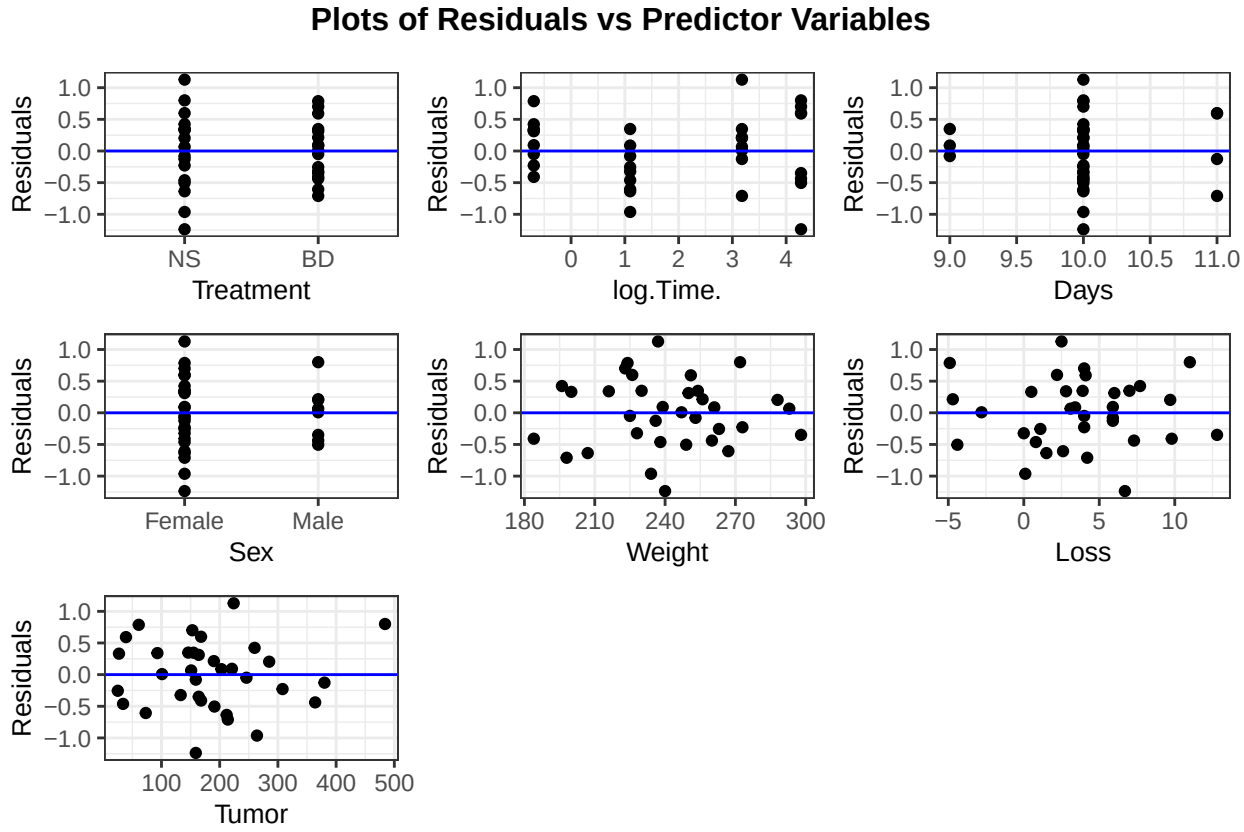
Describe the goals of the analysis. Is the primary goal one of prediction or of understanding?

(2c) Fit the model implied by the questions quoted in part (b) and check the residuals from this model. (In the code below, we set the normal saline solution as the reference level since it is the control treatment.) Is there any evidence of serious nonlinearity or heteroskedasticity?

Notes:
1. In order to save you time, I am not asking you to make plots of the response vs. all predictor variables before fitting the model below. In a realistic setting, you should do that.
2. You may see differences in variability in the plots against the `Days` and `Sex` variables, but this is due to the small sample sizes in some of the groups for those variables.

```
case1102$Treatment <- relevel(case1102$Treatment,
                              ref = "NS")
model1 <- lm(log(Y) ~ Treatment * log(Time) + Days + Sex +  Weight + Loss + Tumor,
                 data = case1102)

library(ggResidpanel)
resid_xpanel(model1)
```



(2d) Answer the questions posed in part (b). Be sure to interpret your model on the original data scale, not on the log scale. (If the answer to the question "Do the answers to these two questions depend on the length of time after the infusion (from 1/2 to 72 hours)?" is "no", refit the model without the interaction before answering the other questions).

(2e) Describe the process that you would use to further refine the model from part (c). (You do not need to do any further refinement here, just describe the process that you would use.)

(2f) Consider selecting a model via backward elimination, starting with a model with all pairwise interactions. (You can use the code below.) Briefly discuss the relative merits of the selected model for answering the questions posed in part (b).

```
upper_model <- lm(log(Y) ~ (log(Time) + Treatment + Sex + Days +  Weight + Loss + Tumor)^2,
                  data = case1102)
lower_model <- lm(Y ~ 1, data = case1102)

library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:patchwork':
##
##     area

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
backwardSelectModel <- stepAIC(upper_model, scope = list(lower = lower_model,
                                                         upper = upper_model),
                               direction = "backward")
```

```
## Start:  AIC=-50.19
## log(Y) ~ (log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor)^2
##
##
## Step:  AIC=-50.19
## log(Y) ~ log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Days +
##     log(Time):Weight + log(Time):Loss + log(Time):Tumor + Treatment:Sex +
##     Treatment:Days + Treatment:Weight + Treatment:Loss + Treatment:Tumor +
##     Sex:Weight + Sex:Loss + Sex:Tumor + Days:Weight + Days:Loss +
##     Days:Tumor + Weight:Loss + Weight:Tumor + Loss:Tumor
##
##                        Df Sum of Sq    RSS     AIC
## - log(Time):Weight      1   0.00186 1.4985 -52.145
## - Treatment:Sex         1   0.01354 1.5101 -51.881
## - log(Time):Days        1   0.05636 1.5530 -50.931
## - Weight:Tumor          1   0.08263 1.5792 -50.360
## <none>                              1.4966 -50.188
## - Sex:Loss              1   0.12510 1.6217 -49.458
## - Treatment:Days        1   0.13538 1.6320 -49.243
## - Days:Loss             1   0.16164 1.6582 -48.700
## - Treatment:Weight      1   0.23705 1.7337 -47.188
## - log(Time):Sex         1   0.32732 1.8239 -45.463
## - Treatment:Tumor       1   0.37237 1.8690 -44.633
## - Treatment:Loss        1   0.51756 2.0142 -42.089
## - Weight:Loss           1   0.53379 2.0304 -41.816
## - log(Time):Treatment   1   0.66032 2.1569 -39.761
## - Sex:Weight            1   0.70803 2.2046 -39.017
## - Loss:Tumor            1   0.88735 2.3840 -36.358
## - Sex:Tumor             1   1.28330 2.7799 -31.134
## - log(Time):Loss        1   1.39115 2.8878 -29.840
## - Days:Weight           1   1.47796 2.9746 -28.833
## - Days:Tumor            1   1.74047 3.2371 -25.958
## - log(Time):Tumor       1   1.78208 3.2787 -25.523
##
## Step:  AIC=-52.15
## log(Y) ~ log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Days +
##     log(Time):Loss + log(Time):Tumor + Treatment:Sex + Treatment:Days +
##     Treatment:Weight + Treatment:Loss + Treatment:Tumor + Sex:Weight +
##     Sex:Loss + Sex:Tumor + Days:Weight + Days:Loss + Days:Tumor +
##     Weight:Loss + Weight:Tumor + Loss:Tumor
##
##                        Df Sum of Sq    RSS     AIC
## - Treatment:Sex         1   0.01368 1.5122 -53.836
## - log(Time):Days        1   0.07724 1.5757 -52.436
## - Weight:Tumor          1   0.08132 1.5798 -52.348
## <none>                              1.4985 -52.145
## - Sex:Loss              1   0.12443 1.6229 -51.433
```

```
## - Treatment:Days       1   0.13500 1.6335 -51.212
## - Days:Loss            1   0.18853 1.6870 -50.116
## - Treatment:Weight     1   0.24904 1.7475 -48.918
## - log(Time):Sex        1   0.36974 1.8682 -46.647
## - Treatment:Tumor      1   0.37121 1.8697 -46.620
## - Treatment:Loss       1   0.53473 2.0332 -43.770
## - Weight:Loss          1   0.58085 2.0793 -43.007
## - log(Time):Treatment  1   0.70091 2.1994 -41.098
## - Loss:Tumor           1   0.88590 2.3844 -38.353
## - Sex:Weight           1   0.93614 2.4346 -37.644
## - Sex:Tumor            1   1.35140 2.8499 -32.289
## - log(Time):Loss       1   1.39471 2.8932 -31.776
## - Days:Tumor           1   1.75298 3.2515 -27.807
## - log(Time):Tumor      1   1.78276 3.2812 -27.497
## - Days:Weight          1   1.90077 3.3992 -26.295
##
## Step:  AIC=-53.84
## log(Y) ~ log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Days +
##     log(Time):Loss + log(Time):Tumor + Treatment:Days + Treatment:Weight +
##     Treatment:Loss + Treatment:Tumor + Sex:Weight + Sex:Loss +
##     Sex:Tumor + Days:Weight + Days:Loss + Days:Tumor + Weight:Loss +
##     Weight:Tumor + Loss:Tumor
##
##                       Df Sum of Sq    RSS     AIC
## - log(Time):Days       1   0.07157 1.5837 -54.264
## - Weight:Tumor         1   0.07207 1.5842 -54.253
## <none>                             1.5122 -53.836
## - Sex:Loss             1   0.11151 1.6237 -53.417
## - Treatment:Days       1   0.13013 1.6423 -53.029
## - Treatment:Weight     1   0.23985 1.7520 -50.830
## - Days:Loss            1   0.24367 1.7558 -50.756
## - log(Time):Sex        1   0.35796 1.8701 -48.612
## - Treatment:Tumor      1   0.42344 1.9356 -47.442
## - Treatment:Loss       1   0.55153 2.0637 -45.264
## - Weight:Loss          1   0.59972 2.1119 -44.479
## - log(Time):Treatment  1   0.72241 2.2346 -42.559
## - Loss:Tumor           1   0.90228 2.4144 -39.927
## - Sex:Weight           1   1.23492 2.7471 -35.538
## - log(Time):Loss       1   1.54367 3.0558 -31.917
## - Sex:Tumor            1   1.65181 3.1640 -30.734
## - Days:Tumor           1   1.73972 3.2519 -29.802
## - log(Time):Tumor      1   1.78245 3.2946 -29.359
## - Days:Weight          1   1.89059 3.4027 -28.260
##
## Step:  AIC=-54.26
## log(Y) ~ log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Loss +
##     log(Time):Tumor + Treatment:Days + Treatment:Weight + Treatment:Loss +
##     Treatment:Tumor + Sex:Weight + Sex:Loss + Sex:Tumor + Days:Weight +
##     Days:Loss + Days:Tumor + Weight:Loss + Weight:Tumor + Loss:Tumor
##
##                       Df Sum of Sq    RSS     AIC
## - Weight:Tumor         1   0.08382 1.6675 -54.510
```

```
## <none>                              1.5837 -54.264
## - Sex:Loss             1   0.15972 1.7434 -52.997
## - Treatment:Days       1   0.16312 1.7468 -52.931
## - Treatment:Weight     1   0.25221 1.8359 -51.240
## - log(Time):Sex        1   0.34963 1.9334 -49.482
## - Treatment:Tumor      1   0.39253 1.9763 -48.735
## - Days:Loss            1   0.51277 2.0965 -46.727
## - Treatment:Loss       1   0.54136 2.1251 -46.267
## - Weight:Loss          1   0.64786 2.2316 -44.604
## - log(Time):Treatment  1   0.79890 2.3826 -42.377
## - Loss:Tumor           1   0.94852 2.5322 -40.307
## - Sex:Weight           1   1.25969 2.8434 -36.366
## - log(Time):Loss       1   1.50373 3.0875 -33.566
## - Sex:Tumor            1   1.71686 3.3006 -31.297
## - Days:Weight          1   1.82822 3.4119 -30.169
## - log(Time):Tumor      1   1.92354 3.5073 -29.232
## - Days:Tumor           1   1.97457 3.5583 -28.741
##
## Step:  AIC=-54.51
## log(Y) ~ log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Loss +
##     log(Time):Tumor + Treatment:Days + Treatment:Weight + Treatment:Loss +
##     Treatment:Tumor + Sex:Weight + Sex:Loss + Sex:Tumor + Days:Weight +
##     Days:Loss + Days:Tumor + Weight:Loss + Loss:Tumor
##
##                        Df Sum of Sq    RSS     AIC
## - Sex:Loss              1   0.07596 1.7435 -54.996
## - Treatment:Days        1   0.08562 1.7532 -54.808
## <none>                              1.6675 -54.510
## - Treatment:Weight      1   0.28173 1.9493 -51.203
## - log(Time):Sex         1   0.39859 2.0661 -49.223
## - Treatment:Tumor       1   0.42029 2.0878 -48.868
## - Days:Loss             1   0.43227 2.0998 -48.673
## - Treatment:Loss        1   0.59772 2.2653 -46.095
## - Loss:Tumor            1   0.95766 2.6252 -41.081
## - Weight:Loss           1   1.19486 2.8624 -38.140
## - log(Time):Treatment   1   1.33280 3.0003 -36.539
## - Sex:Weight            1   1.64704 3.3146 -33.153
## - Sex:Tumor             1   1.83697 3.5045 -31.259
## - log(Time):Loss        1   1.85596 3.5235 -31.075
## - Days:Weight           1   1.92159 3.5891 -30.447
## - log(Time):Tumor       1   2.33748 4.0050 -26.720
## - Days:Tumor            1   2.55246 4.2200 -24.942
##
## Step:  AIC=-55
## log(Y) ~ log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Loss +
##     log(Time):Tumor + Treatment:Days + Treatment:Weight + Treatment:Loss +
##     Treatment:Tumor + Sex:Weight + Sex:Tumor + Days:Weight +
##     Days:Loss + Days:Tumor + Weight:Loss + Loss:Tumor
##
##                        Df Sum of Sq    RSS     AIC
## - Treatment:Days        1    0.0497 1.7932 -56.040
## <none>                              1.7435 -54.996
```

```
## - Treatment:Weight     1    0.3123 2.0558 -51.393
## - Days:Loss            1    0.3568 2.1003 -50.666
## - log(Time):Sex        1    0.5743 2.3178 -47.316
## - Treatment:Tumor      1    0.6369 2.3804 -46.409
## - Loss:Tumor           1    0.8818 2.6253 -43.080
## - Treatment:Loss       1    1.1088 2.8523 -40.260
## - Weight:Loss          1    1.1309 2.8744 -39.997
## - log(Time):Treatment  1    1.2671 3.0106 -38.424
## - Sex:Weight           1    1.7359 3.4794 -33.503
## - Days:Weight          1    2.1172 3.8607 -29.967
## - Sex:Tumor            1    2.5581 4.3016 -26.291
## - log(Time):Tumor      1    2.6331 4.3766 -25.703
## - Days:Tumor           1    2.6745 4.4180 -25.383
## - log(Time):Loss       1    3.3771 5.1206 -20.365
##
## Step:  AIC=-56.04
## log(Y) ~ log(Time) + Treatment + Sex + Days + Weight + Loss +
##     Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Loss +
##     log(Time):Tumor + Treatment:Weight + Treatment:Loss + Treatment:Tumor +
##     Sex:Weight + Sex:Tumor + Days:Weight + Days:Loss + Days:Tumor +
##     Weight:Loss + Loss:Tumor
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              1.7932 -56.040
## - Treatment:Weight     1    0.2715 2.0647 -53.247
## - Days:Loss            1    0.3171 2.1103 -52.505
## - log(Time):Sex        1    0.5254 2.3186 -49.303
## - Treatment:Tumor      1    0.7063 2.4995 -46.749
## - Treatment:Loss       1    1.0625 2.8557 -42.220
## - Weight:Loss          1    1.0886 2.8818 -41.910
## - Loss:Tumor           1    1.2580 3.0512 -39.969
## - log(Time):Treatment  1    1.3232 3.1164 -39.249
## - Sex:Weight           1    1.9379 3.7311 -33.128
## - Days:Weight          1    2.1200 3.9133 -31.508
## - log(Time):Tumor      1    2.6409 4.4341 -27.259
## - Days:Tumor           1    2.6692 4.4624 -27.043
## - Sex:Tumor            1    2.6742 4.4674 -27.005
## - log(Time):Loss       1    3.4066 5.1998 -21.843
```

```
summary(backwardSelectModel)
```

```
##
## Call:
## lm(formula = log(Y) ~ log(Time) + Treatment + Sex + Days + Weight +
##     Loss + Tumor + log(Time):Treatment + log(Time):Sex + log(Time):Loss +
##     log(Time):Tumor + Treatment:Weight + Treatment:Loss + Treatment:Tumor +
##     Sex:Weight + Sex:Tumor + Days:Weight + Days:Loss + Days:Tumor +
##     Weight:Loss + Loss:Tumor, data = case1102)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88289 -0.13639  0.02933  0.14263  0.39588
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)            3.990e+01  2.433e+01   1.640 0.126960
## log(Time)              3.943e-01  2.669e-01   1.477 0.165307
## TreatmentBD            3.381e+00  1.876e+00   1.803 0.096601 .
## SexMale               -1.915e+01  6.262e+00  -3.058 0.009930 **
## Days                  -4.346e+00  2.408e+00  -1.805 0.096280 .
## Weight                -3.871e-01  1.020e-01  -3.796 0.002548 **
## Loss                   1.178e+00  1.381e+00   0.853 0.410574
## Tumor                  1.860e-01  4.484e-02   4.147 0.001354 **
## log(Time):TreatmentBD  2.824e-01  9.492e-02   2.976 0.011578 *
## log(Time):SexMale      1.279e+00  6.819e-01   1.875 0.085309 .
## log(Time):Loss        -1.330e-01  2.786e-02  -4.775 0.000453 ***
## log(Time):Tumor        6.320e-03  1.504e-03   4.204 0.001223 **
## TreatmentBD:Weight    -1.049e-02  7.783e-03  -1.348 0.202615
## TreatmentBD:Loss      -2.649e-01  9.935e-02  -2.666 0.020544 *
## TreatmentBD:Tumor      5.309e-03  2.442e-03   2.174 0.050422 .
## SexMale:Weight         6.640e-02  1.844e-02   3.601 0.003638 **
## SexMale:Tumor         -2.669e-02  6.309e-03  -4.230 0.001167 **
## Days:Weight            3.865e-02  1.026e-02   3.767 0.002688 **
## Days:Loss             -1.927e-01  1.323e-01  -1.457 0.170882
## Days:Tumor            -1.926e-02  4.557e-03  -4.226 0.001176 **
## Weight:Loss            3.451e-03  1.279e-03   2.699 0.019345 *
## Loss:Tumor             1.304e-03  4.494e-04   2.901 0.013292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3866 on 12 degrees of freedom
## Multiple R-squared:  0.9892, Adjusted R-squared:  0.9704
## F-statistic: 52.58 on 21 and 12 DF,  p-value: 8.992e-09
```

3.: Wages and Race Consider ch.10 exercise 29 to answer the following questions. Review the background info for this exercise and data coding provided by the exercise description.

```
data(ex1029, package = "Sleuth3")
```

(3a) In R, fit the interaction model described below:

$$\begin{aligned}\mu(\log(WeeklyEarnings)) = \beta_0 &+ \beta_1 Educ + \beta_2 Exper + \beta_3 RaceNotBlack + \beta_4 MetStatus + \beta_5 regionNE \\ &+ \beta_6 regionS + \beta_7 regionW + \beta_8 regionNE \times RaceNotBlack \\ &+ \beta_9 regionS \times RaceNotBlack + \beta_{10} regionW \times RaceNotBlack\end{aligned}$$

Use an F test to test whether the effect of race (Black/non-Black) on earnings of males differs by region, after controlling for race, region, education, experience, and metropolitan status. Write down the null and alternative hypotheses in terms of a mean function for log(earnings) (e.g. Null: $\mu(\log(WeeklyEarnings)) =$ ... vs. Alt: $\mu(\log(WeeklyEarnings)) =$ ...), then use R to do the F test of these hypotheses. State your conclusion, in context, for this test.

(3b) Fit the no interaction model (below) and use it to interpret the effect that race (Black vs. non-Black) has on earnings (original scale, not logged scale) after controlling for all other predictors, and give a confidence interval for this effect too.

$$\begin{aligned}\mu(\log(WeeklyEarnings)) = \beta_0 &+ \beta_1 Educ + \beta_2 Exper + \beta_3 RaceNotBlack + \beta_4 MetStatus \\ &+ \beta_5 regionNE + \beta_6 regionS + \beta_7 regionW\end{aligned}$$

(3c) Describe the distribution of the residuals for the model given in part (b) with both a histogram and normal qq plot. Our modeling goal is to explore the effect of race (Black/notBlack) on estimated earnings of males after controlling for region, education, experience, and MetStatus. With this in mind, is the distribution of these residuals concerning? Explain. (Hint: think about when normality is **not** a concern.)

```
library(ggResidpanel)
#resid_panel(wage_lmred, plots = c("hist", "qq"))
```

(3d) Using the `ggResidpanel` package, create the six possible residual plots for this model (fitted + 5 predictors). For each, comment on the linearity and constant variance assumptions made for this model. This is a large data set (with lots of residuals), so add `smoother = TRUE` to add a smoother line to help detect nonlinearity in the overlapping points in your residual plots.

```
#resid_panel(wage_lmred, plots = "resid", smoother = TRUE)
#resid_xpanel(wage_lmred, smoother = TRUE)
```

(3e) One way to "test" for curvature is to add a quadratic term to your model. Since part (d) suggest a nonlinear effect of experience, add a quadratic term for experience to the linear model above and fit this model to the data. Use the t-test results to determine whether the nonlinear effect of experience is significant.

(3f) Using your model from (e), report the case numbers of the cases with the highest leverage, studentized residuals and Cook's distance values. Use the data for these cases and basic EDA to explain why their respective case influence stat is high. Then explain why none of these cases need to be removed from our data to adequately model earnings.

If you'd like to use the `augment`ed data frame to find the row number of these "max" case influence stats, I suggest that you do the following

- Add **row numbers** to your data set, e.g. like this using `dplyr`:

```
library(dplyr)
#my_data <- my_data %>% mutate(case = row_number())
```

- Then `augment` your `lm` and the data set to add the case influence stats to your original data set (otherwise R adds these stats to data that matches your model terms (logged and quadratic terms without case number))

```
library(broom)
#my_data_aug <- augment(my_lm, data = my_data)
```