1. Miscellaneous.
2. Review list.
3. Practice problems.

1) Miscellaneous.
a) The midterm is Wednesday, May 12, 3:00pm to 3:50pm, in Boelter 2444.
b) 10 multiple choice worth 8.5 points each and 1 short answer worth 15 points.
c) If something goes wrong in Boelter 2444, then the midterm will be in BH 9413.
d) The final exam is Thur June 10, 3-6pm, in Math-Science 6229.

2) Review list.
A) Regression basics and problems in regression.
A1) Ordinary linear regression assumptions.
A2) Confounding factors, observational studies, experiments, and causation.
A3) Statistical and practical significance.
A4) Extrapolation and prediction.
A5) Curvature.
A6) Leverage.
A7) Dependance and nonidentifiability.
A8) Overfitting.
B) Variable selection.
B1) Maximum likelihood.
B2) AIC and Corrected AIC.
B3) BIC.
B4) Comparing Information Criteria.
B5) Forward and backward selection.

B6) Inference after variable selection (artificially small p-values, testing and training).

B7) LASSO.

C) Logistic regression.

C1) Bernoulli random variables.

C2) Binomial random variables.

C3) The logistic function.

C4) The logistic regression model.

C5) Odds.

C6) Likelihood and estimation for logistic regression.

C7) Testing and deviance for logistic regression.

C8) Residuals for logistic regression.

C9) Residuals and overfitting for binary data.

C10) Interpreting logistic regression.

D) Poisson regression.

D1) Poisson random variables.

D2) The Poisson regression model.

D3) Likelihood and estimation for Poisson regression.

E) Kernel regression.

E1) The purpose of kernel regression.

E2) Kernel density estimates.

E3) The Nadaraya-Watson estimator.

E4) Choosing the bandwidth.

F) Robust regression.

F1) m-estimation, LAD regression, and Huber's method.

F2) LTS.

F3) B-spline regression.

F4) GAM.

G) GLM.

G1) The purpose of GLM.

G2) The exponential family, including the normal, binomial, and Poisson.

G3) Properties of GLMs.

G4) Link functions.

G5) Probit and complementary log-log regression.

G6) GLM residuals, partial regression and partial residual plots.

H) Serial correlation.

H1) Time series and serial correlation.

H2) Autocovariance function, ACF and correlogram.

H3) Properties of correlograms.

3) Example problems.

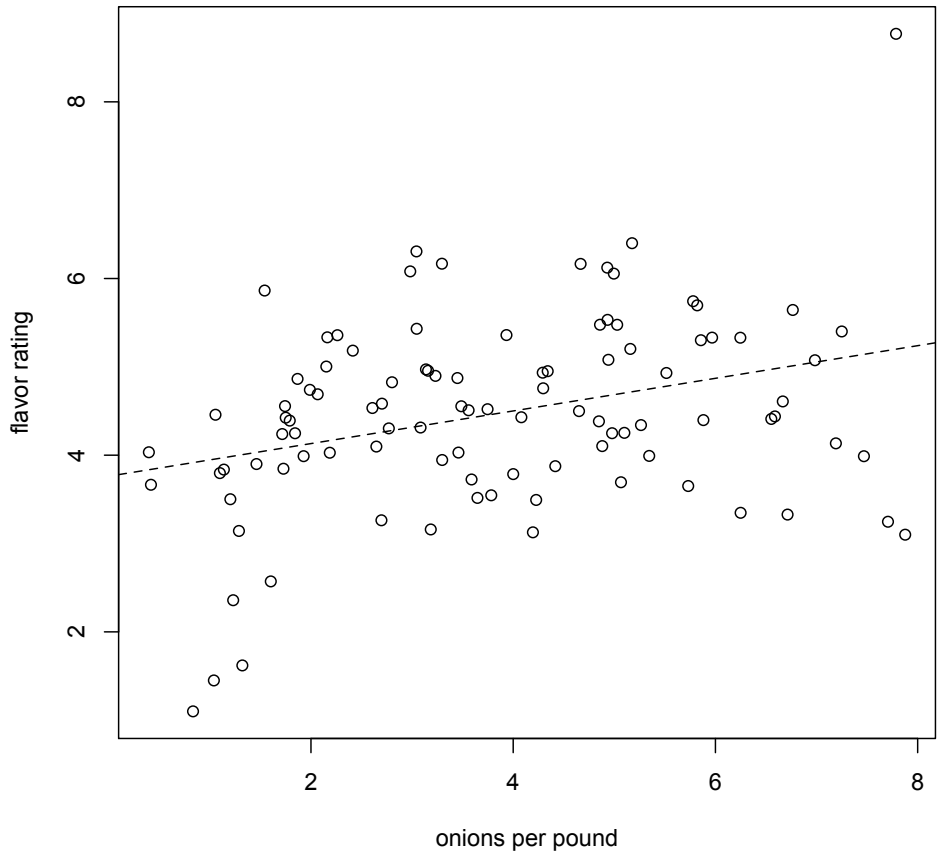
a) According to the Poisson regression model with log link function, with $\beta_0=1$ and $\beta_1=2$, if $X_{100}=3$, then what is the standard deviation of Y_{100}?

(i) exp(5) (iii) exp(7) (iii) log(5) (iv) log(7)
(v) $\sqrt{}$exp(5) (vi) $\sqrt{}$exp(7) (vii) $\sqrt{}$log(5) (viii) NTA.

b) Suppose that you do a survey on students' art knowledge and happiness. You have a list of 100 paintings and you ask students how many of them they recognize. In addition, you ask the students to rate their general happiness from 1 to 10. You consider happiness the response variable, and find that, using ordinary least squares linear regression, when X = 80, your prediction $\hat{}Y = 8.2$, and when X = 70, $\hat{}Y = 8.1$. The residual sum of squares for OLS is 10,420 and $\hat{}\sigma = 0.2$. Using kernel regression, when X = 80, $\hat{}Y = 8.2$, and when X = 70, $\hat{}Y = 8.0$. The residual sum of squares for kernel regression is 8,010 and $\hat{}\sigma = 0.2$. What can you conclude?

(i) Art makes people happier.
(ii) The least squares residual estimator has higher correlation bias due to logistic extrapolation heteroskedasticity.

(iii) Increasing your art knowledge from 70% to 80% would cause an increase of about 0.1 to 0.2 in your happiness score.
(iv) For students with 80% art knowledge, their happiness scores tend to range between 8 and 8.4.
(v) Since kernel regression has a smaller residual sum of squares, predictions based on kernel regression should be favored over linear regression.
(vi) NTA.



c) Lasagnes at 100 restaurants are tested for onion content and rated from 0 to 10 for flavor by food critics. A plot is shown above. Using linear regression, the estimated intercept is around 4 and the estimated slope is nearly 0.2. A researcher concludes that a recipe

would therefore need 30 onions per pound in order to get a flavor rating of 10. What are the main problems with this conclusion?