Stat 135, Fall 2006 A. Adhikari

HOMEWORK 10 SOLUTIONS

1a) The model is $cw_i = \beta_0 + \beta_1 el_i + \epsilon_i$, where cw_i is the weight of the ith chick, el_i the length of the egg from which it hatched, and ϵ_i the normal error. The index i ranges from 1 to n and the ϵ are i.i.d. normal with mean 0 and variance σ^2.

The plot looks linear and roughly (though not perfectly) homoscedastic. I'm willing to believe that the model is OK.

The mean chick weight is 6.145455 grams with an SD of 0.4105892 grams. The mean egg length is 31.38955 mm with an SD of 1.100892 mm. The correlation between egg length and chick weight is 0.6761419. So by our old formulas the slope of the regression line is $rs_y/s_x = 0.2521742$ gm/mm, and the intercept is $\bar{y} - \text{slope}\bar{x} = -1.770180$ gm. The equation of the regression line is:
estimated chick weight = 0.2521742gm/mm (egg length) - 1.770180 gm

b) According to R the slope of the regression line is 0.2522 and the intercept is -1.7702, both of which agree with the values computed in **a**. Assuming that the linear model holds, the t-test for the intercept is testing whether or not the intercept of the true line is 0; the p-value is large (19%) which supports the null hypothesis that the intercept is 0.

In this case (simple regression) the t-test for the slope and the F-test are both testing the same thing: whether or not the slope is 0. Both reject the hypothesis that the slope is 0. I hope you noticed that the F-statistic of 35.37 is the square of the t-statistics of 5.947.

c) Not surprisingly, it's egg weight. The correlation between the weights of the eggs and the chicks is 0.847225. The plot is nice and linear but heteroscedasticity is an issue at the edges. The assumptions of the linear model are not terrible for the main bulk of the data. The residual plot shows the same thing. The R^2 is 0.72, not bad, and consistent with the correlation from the correlation matrix.

d) Use the estimated slope and intercept from **c**: the estimate of the mean weight is $0.71852 \times 8.5 - 0.05827 = 6.04915$. By the formula in the class handout, the standard error is estimated as

$$0.2207 \times \sqrt{\frac{1}{44} + \frac{(8.5 - \text{mean(ew)})^2}{43 \times \text{var(ew)}}} = 0.03455294 \; gm.$$

Use the t_{42} distribution to see that the confidence interval is $6.04915 \pm 2.018082 \times .03455294$ which is about $(5.98, 6.12) grams$.

e) The estimate is the same as in **d** and so is the value of t, but now the standard error is estimated as

$$0.2207 \times \sqrt{\frac{1}{44} + \frac{(8.5 - \text{mean(ew)})^2}{43 \times \text{var(ew)}} + 1} = 0.2233884 \; gm.$$

f) The given egg weight is well outside the range of the data. I don't know whether the model holds at those weights or not. So, because we are dealing with such an outlier, I will declare this one to be "not possible".

In general, beware of "extrapolation", that is, making estimates outside the range of your data. Unless you really have reason to believe that the model continues to hold even in where you have no observations, don't do it.

2a) The two regressions are very similar. Both show the same problems with the homoscedasticity assumptions. Both have an R^2 of about 0.71. The normal quantile plot is worse for the

multiple regression. But overall, not much to choose between them.

b) The R^2 is an astonishing 0.95, and the rest of the regression diagnostics are not bad. So you can think of egg weight as essentially a linear function of egg length and egg breadth. That explains why the two regressions in **a** are so similar: linear functions of egg weight are essentially linear functions of egg length and egg breadth.

c) Given the discussion in **b)**, this is a really silly thing to do. And the silliness shows up in the apparently contradictory results of the F-test and the three t-tests for the slopes. Each t-test concludes that the corresponding slope is 0, but the F-test concludes that not all are 0. It's all correct! The predictor variables are so highly correlated with each other (see e.g. **b**) that the individual slopes have no meaning.

It's true that the R^2 is just a shade higher than those of the earlier regressions, but the adjusted R^2 is a shade lower than that of the regression on egg weight alone.

To summarize: if you use predictors that are highly correlated with each other, don't try to interpret results for individual slopes. Better still, don't use predictor variables that are highly correlated with each other!

d) The two in **a** are the best. I don't find anything else that compares well. If you use a combination of egg weight and either of the other two variables, the only significant slope is that of egg weight and the R^2 are all around 0.7. So, for comprehensibility of the model and minimal correlation between predictors, I'd go with the two in **a**.

3a) You can do the parametric test of the null hypothesis that in the population the mean baseline score is the same as the mean 15-month score (i.e. the treatment did nothing). The data are paired, so you run the t-test on the differences between the scores to see that the value of t is -6.15 (15months - base) with 21 degrees of freedom. The p-value is tiny so you conclude that the means are different. Notice the negative sign of t. This is coming from the fact that the 15-month scores are on average lower than the baseline scores.

To run the nonparametric test of the null hypothesis that the underlying distributions at baseline and at 15 months are the same, do the Wilcoxon signed-rank test. The value of the statistic is 246 and the p-value is tiny, supporting the alternative that the two underlying distributions are different. This is consistent with the result of the parametric test.

b) A glance at the correlation matrix shows that you'll want to include the baseline score (big surprise) and chemo (it's correlated with the response but almost uncorrelated with baseline). You may want to use height, but that's correlated with the baseline measurement so it may not be a good idea. And indeed, it turns out not to be a good idea when you run the regression.

The best one uses baseline and chemo as the predictors; the adjusted R^2 is about 0.43, clearly greater than the values from the other regressions. Both slopes are significantly different from 0. According to the diagnostic plots the model looks OK, though there are clearly some deviations from homoscedasticity and normality.

4a) Looks about as normal as a real dataset gets. The histogram has a fairly symmetric bell shape and the normal q-q plot looks like a straight line.

b) This distribution is skewed to the right. In the q-q plot this is represented by the bow shape of the line. If the skewness was in the other direction, the q-q plot would again be bow-shaped but this time it would be concave instead of convex.

c) The mother's age doesn't seem to matter. The model which includes all the other predictors looks good; even better, you can drop the column of the mother's pregnancy weight without much

loss. In both cases the adjusted R^2 are around 0.25, clearly better than the others, and the diagnostic plots are pretty good.

d) The coefficient is estimated as -8.35 or (or -8.5, depending on which model you used; it doesn't make much difference). It represents the average difference in birthweight between a baby born to a smoker and a baby born to a nonsmoker, provided the other variables included in the model are held constant. The conclusion is that a mother who smokes is expected to have a baby whose birthweight is 8.35 ounces less than a mother who doesn't smoke (*ceterus paribus* - all else assumed to be equal).

5. a) $R = 0.9955$. If you just plot the data, the fit appears to be very good. R^2 is very high and the slope is highly significant. However, the residual plot is u-shaped, showing a strong non-linear pattern. So, regardless of the high R^2, we should not have fit a straight line.

b) Less. Each point in the dataset `women` represents many women – all those of a given height. If you replace the point by the individual points for all the women at that height, the picture will become much more fuzzy. And the correlation will drop.

c) The residual plot of the linear regression suggests a polynomial fit of degree 2. However, once that model is fitted, you can see a clear up-and-down pattern in the residuals. Try a final model, then, of degree 3. R^2 for this model equals 0.9998 which is ridiculously high. The residual plot is better than the previous two, though the normal q-q plot of the residuals is not as good as the one in the quadratic fit.

I won't go to the 4th degree polynomial for several reasons, chief among them being that you don't gain much as far as the fit goes, and fourth powers of inches are beyond most people's comprehensions. Stick with the cubic.

6a) The plots are very similar and neither shows any clear relationship, linear or otherwise. Both appear to be formless blobs.

b) Similar except that the men's heart rates are clearly less variable than the women's.

c) The regression confirms what we saw in **a** – there's nothing much going on in terms of a linear relationship. R^2 is only about 0.04. The residual plot is formless blob, which is good, but then so is the original plot. The estimated slope is 1.645 (not significantly different from 0, big surprise), and the estimated intercept is -87.967.

d) The story for the women is pretty much the same as that for the men except that R^2 is higher (about 0.08) and the slope does come out to be significant. It's positive, so the estimated heart rate increases slightly with increasing temperature. But it's not a very convincing regression because its predictive power is pretty small due to the small R^2.

e) The slope for the men was estimated as 1.645 with an SE of 1.039, and the slope for the women was estimated as 3.128 with an SE of 1.316. So the difference in slopes (women - men) is estimated to be 1.483 with an SE of $\sqrt{1.039^2 + 1.316^2} = 1.68$, since the data for the men and women are independent. The sample sizes are large enough that I'm just going to use the normal approximation and not worry about t distributions. An approximate 95%-confidence interval for the difference between the slopes is $1.483 \pm 2 \times 1.68$ which clearly contains 0. So at the 5% level you can conclude that the slopes are equal.

f) Play the same game as in **e** but with the intercepts. The estimated difference is 145.657 with an SE of 164.767, and the 95%-confidence interval for the difference clearly contains 0. So at the

5% level you can conclude that the intercepts are equal.

7a. There are $n = 4$ observations and 2 parameters w_1 and w_2. The model is $Y = X\beta + \epsilon$ where the observed value of Y is the transpose of $(3, 3, 1, 7)$, β is the transpose of (w_1, w_2), the vector of errors ϵ consists of 4 i.i.d. normal $(0, \sigma^2)$ variables, and the design matrix X is

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}$$

b) The least squares estimates are $(X^TX)^{-1}X^TY$. Now X^TX is a particularly simple matrix:

$$\begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

and so its inverse is equally simple:

$$\begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix}$$

And X^TY is the transpose of $(11, 9)$ and so the estimates are $\hat{w}_1 = 11/3$ and $\hat{w}_2 = 9/3$.

c) $\hat{Y}$ is the transpose of $(11/3, 9/3, 2/3, 20/3)$. So the residual vector is the transpose of $(-2/3, 0, 1/3, 1/3)$. So the estimate of σ^2 is $(6/9)/(4-2) = 1/3$.

d) The covariance matrix of the estimates is $\sigma^2(X^TX)^{-1}$ and so both the estimated variances are equal to $1/9$, and so the estimated standard errors are both $1/3$.

e) The estimate is $11/3 - 9/3 = 2/3$, and its standard error is $\sqrt{1/9 + 1/9} = 0.47$ because the covariance is 0 (the off-diagonal elements of the covariance matrix are 0).

f) The relative sizes of the estimated difference and its SE in part **e** show that the 95%-confidence interval for $w_1 - w_2$ must contain 0, whether you use a t distribution or a normal (you should use a t with 2 d.f.). So you'll accept the null hypothesis.

8. Just like the previous one, except now X is an $n \times 2$ matrix whose first column consists of the values of x and the second column consists of the values of x^2.

a) As in the previous problem, X^TX is a 2×2 matrix whose diagonal entries are the sum of squares of x and the sum of fourth powers of x. The off-diagonal entries are the sum of cubes. After this, it's all as in the previous problem except that X^TY is the transpose of $(\sum x_i y_i, \sum x_i^2 y_i)$. I'm not really interested in whether or not you wrote the algebraic formulas out correctly long-hand.

b) The matrix is $\sigma^2(X^TX)^{-1}$ where X^TX was found in part **a**. That's all you have to do.

9. These results are the probability versions of results you derived in HW 9 for lists of real numbers. The derivations are very similar, using expectations instead of averages.

I won't use the hint – I'll just do the equivalent of what we did in class when we derived the formula for the slope and intercept of the regression line.

a) $E(Y - (\alpha + \beta X))^2 = E(Y - \beta X)^2 - 2\alpha E(Y - \beta X) + \alpha^2$. Fix β, treat this as a function of α, differentiate, and set equal to 0:

$$-2E(Y - \beta X) + 2\hat{\alpha} = 0$$

and so for each fixed β, the value of $\hat{\alpha}$ is $E(Y - \beta X) = \mu_y - \beta\mu_y$, which is one of the results we are asked to prove.

For the other, plug $\hat{\alpha}$ into the expected square error:

$$E[(Y-\mu_y)-\beta(X-\mu_X)]^2 \;=\; Var(Y)-2\beta Cov(X,Y)+\beta^2 Var(X)$$

Minimize this with respect to β:

$$-2Cov(X,Y)+2\hat{\beta}Var(X)=0$$

and therefore $\hat{\beta}=\sigma_{xy}/\sigma_x^2$ as was to be shown.

b)

$$\begin{aligned} Var(Y)=Var[(Y-\hat{Y})+\hat{Y}] &= Var(Y-\hat{Y})+Var(\hat{Y})+2Cov(Y-\hat{Y},\hat{Y}) \\ &= Var(Y-\hat{Y})+Var(\hat{Y})+2Cov(Y,\hat{Y})-2Var(\hat{Y}) \\ &= Var(Y-\hat{Y})-Var(\hat{Y})+2Cov(Y,\hat{Y}) \\ &= Var(Y-\hat{Y})-\hat{\beta}^2Var(X)+2\hat{\beta}Cov(Y,X) \end{aligned}$$

by plugging in $\hat{Y}=\hat{\alpha}+\hat{\beta}X$. Now plug in the value of $\hat{\beta}$ to see that

$$Var(Y)-Var(Y-\hat{Y})=-\frac{(\sigma_{xy})^2}{\sigma_x^2}+2\frac{(\sigma_{xy})^2}{\sigma_x^2}=\frac{(\sigma_{xy})^2}{\sigma_x^2}$$

Divide both sides by $Var(Y)$ and you're done. The right hand side is the square of the correlation, by the definition of correlation as $\sigma_{xy}/\sigma_x\sigma_y$.

10. Freebie.