Test Review

Chapters 1 - 4

Dr. Cathy Poliak, cpoliak@uh.edu

University of Houston

Exam Information

- Tuesday February 27 at 11:30 am in SEC 102 during class.
- Approximately 8 questions.
- 75 minutes.
- May bring one-page notes front/back can be typed if wanted to be turned in with the test for bonus points. Only notes, formulas and R code no worked out examples.
- Bring your calculator.

Three problems will present you with a data example and ask you an array of modeling/interpretation questions about that data. (Short answer questions)

Other problems will just be a mix of single questions on general knowledge of the class material. Will be a mixture of multiple choice and short answer questions.

Topics Covered

- Types of statistical learning
- Simple linear regression
- Multiple linear regression
- Polynomial regression
- Best subsets
- Logistic Regression
- Test/Training data
- Confusion Matrix

Statistical Learning General Approach

- We refer to the response usually as Y.
- Let $X = (X_1, X_2, \ldots, X_p)$ be p different predictors (independent) variables.
- We assume there is some sort of relationship between X and Y, which can be written in the general form thus our model is

$$Y = f(X) + \epsilon$$

- Where ϵ captures the measurement errors and other discrepancies.
- Statistical leaning refers to a set of approaches for estimating f.

Lin Reg: $f(x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \hat{y}$

Log Reg: $f(x) = \dfrac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)} = P(Y=1 \mid X)$

Lin Reg: Least squares method

$\min \sum_{i}^{n} (y_i - \hat{y}_i)^2$

Log Reg: Max. Likelihood estimate

$$P(y) = p^{y}(1-p)^{1-y} \qquad y = 0 \text{ or } 1$$

$$L(P(y_i)) = \prod_{i=1}^{n} p_i^{y_i}(1-p)^{1-y_i}$$

$$\ln[L(P(y_i))] = \sum_{i=1}^{n} y_i \ln(p_i) + \sum_{i=1}^{n} (1-y_i) \ln(1-p_i)$$

minimize log-likelihood.

Reducible and Irreducible Error

There are uncertainty associated with this prediction.

1. The coefficients are only an estimate for the true population model $f(X)$. This is related to the **reducible error**. We use the **confidence interval** for the predicted value to determine how close $\hat{Y}$ will be to $f(X)$.
2. We are assuming a linear model for $f(X)$, so there is an additional source of potentially reducible error which we call *model bias*.
3. Even if we know $f(X)$, the response value cannot be predicted perfectly because of the random error ε. This is the **irreducible error**. How much will Y vary from $\hat{Y}$? We use **prediction intervals** to answer this question.

Bias and Variance

- Accuracy - measured by **bias**

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- Precision - measured by its variance, Var$((\hat{\theta})$. The estimated standard deviation of an estimator θ is referred to as its **standard error (SE)**.
- The **mean squared error (MSE)** combines both measures.

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

Lin Reg: $\text{MSE} = \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-p-1} = \dfrac{RSS}{n-p-1}$

Log Reg: $\text{MSE} = \text{error rate} = \dfrac{\#\text{ of wrong pred}}{\text{Total}}$

Example 1 - MPG as a response

1. We want to determine which certain predictors are related to 'mpg'. Do we have in inference or prediction problem?

   Inference

2. The response variable is 'mpg', we will use 'wt' (Weight per 1000lbs), 'qsec' (1/4 mile time), and 'am' (Transmission 0 = automatic, 1 = manual), and 'vs' (Engine 0 = V-shaped, 1 = straight) as predictors. Do we have a regression or classification problem?

   Regression because response = mpg is quantitative

3. Write the model that we will use.

$$mpg = \overbrace{\beta_0 + \beta_1 \times wt + \beta_2 \times qsec + \beta_3 \times am + \beta_4 * vs}^{f(x)} + \varepsilon$$

$$\varepsilon \sim N(0,1)$$

4. Given the following output, write the model with the coefficients and interpret the coefficient for `wt`.

```
Call:
lm(formula = mpg ~ wt + qsec + am + vs, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4780 -1.5520 -0.7256  1.4095  4.6626 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.58206    8.21022   1.167   0.2534    
wt          -3.91964    0.81060  -4.835 4.74e-05 ***
qsec         1.22881    0.44867   2.739   0.0108 *  
am           2.93655    1.43919   2.040   0.0512 .  
vs          -0.01512    1.75451  -0.009   0.9932    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.504 on 27 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8274 
F-statistic: 38.15 on 4 and 27 DF,  p-value: 9.688e-11
```

$$\hat{mpg} = \begin{cases} 9.582 - 3.9196 * wt + 1.2288 * qsec & \text{if } am=0 \,\&\, vs=0 \\ 12.5186 - 3.9196 * wt + 1.2288 * qsec & \text{if } am=1 \,\&\, vs=0 \\ 9.5669 - 3.9196 * wt + 1.2288 * qsec & \text{if } am=0 \,\&\, vs=1 \\ 12.5035 - 3.9196 * wt + 1.2288 * qsec & \text{if } am=1 \,\&\, vs=1 \end{cases}$$

5. What is the R code that will give us the summary output?

cars.lm = lm(mpg ~ wt + qsec + am + vs, data = mtcars)

summary(cars.lm)

6. Test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ versus H_a : at least one $\beta_j \neq 0$. Describe in words what this means, give the statistic, p-value, and conclusion of this test.

Testing if any of the attributes contribute to predicting mpg.

F = 38.15, p-value ≈ 0 reject H_0.

Conclusion: At least one of the attributes have a significant effect on mpg.

7. Test $H_0 : \beta_2 = 0$, versus $H_a : \beta_2 \neq 0$. Describe in words what this means, give the statistic, p-value, and conclusion of this test.

Do we need qsec in the model, given the other variables are in the model.

$t = 2.739$, p-value $= 0.0108$, reject H_0.

Conclusion: There is evidence that qsec is significant in the model.

8. Based on the t-tests in the summary, name any predictors that may not be needed to predict `mpg`.

Yes mainly vs, since p-value $= 0.9932$.

R^2 AIC and C_p

adjus R^2

9. Given the output below calculate the R^2, AIC and C_p.

```
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq  F value    Pr(>F)
wt         1 847.73  847.73 140.2143 2.038e-12 ***
qsec       1  82.86   82.86  13.7048 0.0009286 ***
am         1  26.18   26.18   4.3298 0.0467155 *
Residuals 28 169.29    6.05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$n = 32$

$31 = n - 1$

$RSS = 169.29 \qquad SSR = 847.73 + 82.86 + 26.18 = 956.77$

$TSS = RSS + RSS = 169.29 + 956.77 = 1126.06$

$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{169.29}{1126.06} \approx 0.8497$

$Ad.R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - \frac{169.29/28}{1126.06/31} = 0.8336$

$AIC = 2(p+1) + n\ln\left(\frac{RSS}{n}\right) = 2(3+1) + 32\ln\left(\frac{169.29}{32}\right) = 61.308$

$$C_p = \frac{RSS_p}{MSE_{All}} + 2(p+1) - n = \frac{149.2}{6.27} + 2(3+1) - 32 = 3$$

Asumptions

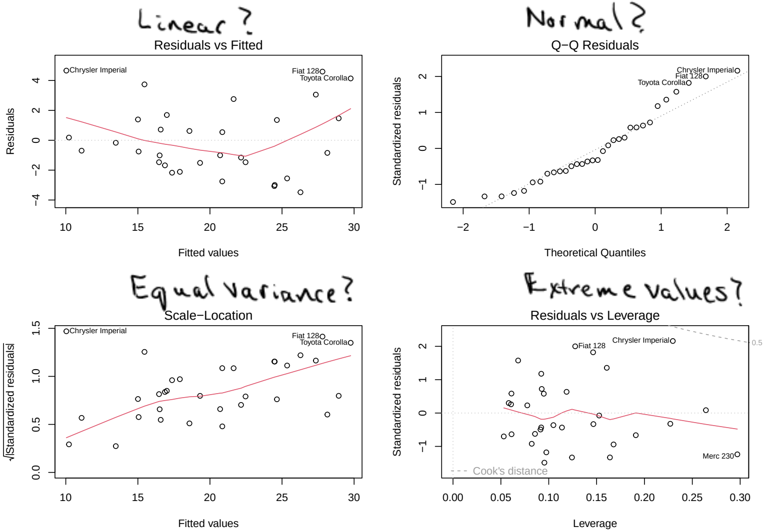
10. Give the assumptions of this model.

Linear
Independent opservation
Normal error term
Equal variance of errors for each value of X.
No extreme values

11. How do we determine if these assumptions are met?

diagnostic plots

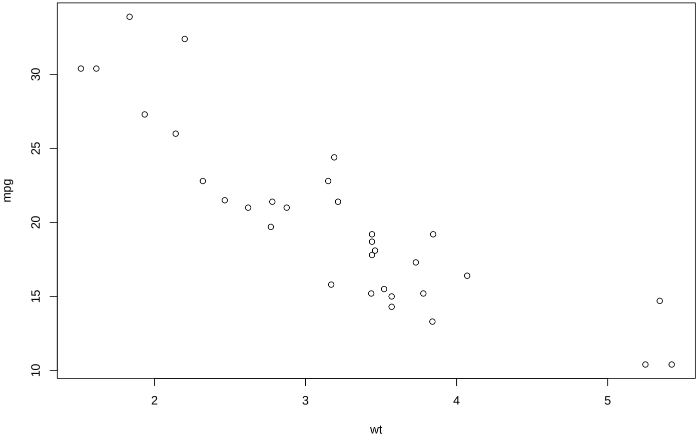
12. Are the assumptions met?



No, this is not linear

13. Using the plot below, what type of relationship do we have between `mpg` and `wt`?

Negative, maybe linear



Predictions and Intervals

14. Based on the output below, what is the predicted mpg if $wt = 3,200$ lbs?

```
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

$\hat{mpg} = 37.2851 - 5.3445 * 3.2$
$= 20.1827$

15. Give a 95% confidence interval for the slope coefficient. Interpret this interval.

$t_{0.025,30} = qt(0.975, 30) = 2.0423$

$-5.3445 \pm 2.0423 * 0.5591 = [-5.9036, -4.7854]$

As the wt increases per 1000 lbs the mpg will decrease between 4.8 and 5.9 mpg with 95% confidence.

16. The following is a 95% prediction interval for `mpg` when $wt = 3,200$ lbs. Give the `R` code to get this interval, what does this mean?

```
       fit      lwr      upr
1 20.18282 13.86582 26.49982
```

Given another automobile with weight of 3,200 lbs we predict the mpg of one car to be between 13.865 and 26.5.

17. The following is a 95% confidence interval for `mpg` when $wt = 3,200$ lbs. Give the `R` code to get this interval, what does this mean?

```
       fit    lwr      upr
1 20.18282 19.083 21.28264
```

Given automobiles with weight of 3,200 lbs we predict the average mpg to be between 19.083 and 21.283

Polynomial Regression

1. The output below is a polynomial regression with degree 3. Write out this model using the coefficients.

```
Call:
lm(formula = mpg ~ poly(wt, 3), data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.506 -1.999 -0.768  1.490  6.188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.0906     0.4768  42.139  < 2e-16 ***
poly(wt, 3)1 -29.1157     2.6970 -10.796 1.73e-11 ***
poly(wt, 3)2   8.6358     2.6970   3.202  0.00339 **
poly(wt, 3)3   0.2749     2.6970   0.102  0.91954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.697 on 28 degrees of freedom
Multiple R-squared:  0.8191,    Adjusted R-squared:  0.7997
F-statistic: 42.27 on 3 and 28 DF,  p-value: 1.585e-10
```

$$\hat{mpg} = 20.091 - 29.1157\, wt + 8.6358\, wt^2 + 0.2749\, wt^3$$

2. Write out the R code to get this output.

```
poly.car = lm(mpg ~ poly(wt, 3), data=mtcars)
summary(poly.car)
```

3. Write out the best model based on this output.

$$mpg = \beta_0 + \beta_1 wt + \beta_2 wt^2 + \varepsilon \quad , \quad \varepsilon \sim N(0,1)$$

Example 2 - Predicting Type of Engine

1. We want to predict the type of engine based on `disp`, `hp` and `wt`. Do we mainly have in inference or prediction problem?

   prediction

2. Is this a classification or regression problem?

   classification since response = vs = v = V-engine
   s = straight

3. Write the model for this type of problem.

   Logistic

$$p(x) = \frac{\exp(\beta_0 + \beta_1 disp + \beta_2 hp + \beta_3 wt)}{1 + \exp(\beta_0 + \beta_1 disp + \beta_2 hp + \beta_3 wt)} + \varepsilon$$

4. Given the output below write out the model with the coefficients.

```
Call:
glm(formula = vs ~ disp + hp + wt, family = "binomial", data = mtcars)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.32436 =β0 3.86897   1.376   0.1688
disp        -0.01787 =β1 0.01774  -1.007   0.3140
hp          -0.06624 =β2 0.03679  -1.800   0.0718 .
wt           2.04416 =β3 1.65978   1.232   0.2181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.860  on 31  degrees of freedom
Residual deviance: 14.987  on 28  degrees of freedom
AIC: 22.987
```

$= 2(p+1) + Dev_R = 2(4) + 14.987$

```
Number of Fisher Scoring iterations: 7
```

$$R^2 = 1 - \frac{14.987}{43.86} = 0.6583$$

5. Write out the R code to get this summary.

```
glm.cars = glm(vs ~ disp + hp + wt, family = "binomial",
               data = mtcars)
Summary(glm.cars)
```

6. Interpret the coefficient with hp, $\hat{\beta}_2$.

With one unit increase in hp, the probability of a straight engine decreases, with fixed values of disp and wt.

7. If we were to determine we want inference is there justification to not use all of the predictors?

Yes, since some of the p-values to test $H_0: \beta_j = 0$ is greater than 0.05.

8. Determine and interpret R^2.

$1 - \frac{14.987}{43.86} = 0.4583$ This determines how well the model "fits" the data

9. Name a function in `R` that will allows to best determine which predictors to use.

step( )

10. Based on the summary below, give the predicted response when $hp = 146.7$. Interpret this response.

```
Call:
glm(formula = vs ~ hp, family = "binomial", data = mtcars)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.37802    3.21593   2.605  0.00918 **
hp          -0.06856    0.02740  -2.502  0.01234 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.860  on 31  degrees of freedom
Residual deviance: 16.838  on 30  degrees of freedom
AIC: 20.838

Number of Fisher Scoring iterations: 7
```

$$R^2 = 1 - \frac{16.838}{43.860} = 0.616$$

$$P(Y=1 \mid hp = 146.7) = \frac{\exp(8.378 - 0.06856 * 146.7)}{1 + \exp(8.378 - 0.06856 * 146.7)} = 0.1564$$

The probability that an automobile has a straight engine, given hp=147.7 is 15.6% chance.

Determine the R^2 for this model

12. Given the confusion matrix below, what is the error rate?

		True Response	
		0 = V-shaped	1 = straight
Predicted	0 = V-shaped	15	2
Response	1 = straight	3	12

18

error rate = $\frac{3+2}{32}$ = 0.15625

13. Write the R code to get this confusion matrix.

```
perc.cars = predict.glm( glm.cars, type="response")
pred.cars = ifelse(perc.cars<0.5, 0, 1)
table(pred.cars, mtcars$vs)
```

14. What is the specificity rate?

$\frac{15}{18}$ = 0.8333