Stat 344 – HW 11

Trey Tipton

April 07, 2022

Problem 7.14

a.)

```
gpa.lm <- lm(gpa ~ satm + satv + act, data = GPA)
msummary(gpa.lm)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.2295888  0.1968034   6.248 1.64e-09 ***
## satm        0.0006913  0.0004316   1.602  0.11042
## satv        0.0012229  0.0004717   2.593  0.01005 *
## act         0.0339602  0.0125115   2.714  0.00707 **
##
## Residual standard error: 0.4219 on 267 degrees of freedom
## Multiple R-squared:  0.3259, Adjusted R-squared:  0.3183
## F-statistic: 43.02 on 3 and 267 DF,  p-value: < 2.2e-16
```

We want to do model comparison that will lead to each of the p-values 1.64e-09, 0.11042, 0.01005, and 0.00707.

```
gpa.lm1 <- lm(gpa ~ -1 + satm + satv + act, data = GPA)
anova(gpa.lm1, gpa.lm)
```

```
## Analysis of Variance Table
##
## Model 1: gpa ~ -1 + satm + satv + act
## Model 2: gpa ~ satm + satv + act
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    268 54.464
## 2    267 47.517  1    6.9469 39.035 1.636e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gpa.lm2 <- lm(gpa ~ satv + act, data = GPA)
anova(gpa.lm2, gpa.lm)
```

```
## Analysis of Variance Table
##
## Model 1: gpa ~ satv + act
## Model 2: gpa ~ satm + satv + act
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    268 47.973
## 2    267 47.517  1   0.45652 2.5652 0.1104
```

```
gpa.lm3 <- lm(gpa ~ satm + act, data = GPA)
anova(gpa.lm3, gpa.lm)
```

```
## Analysis of Variance Table
##
## Model 1: gpa ~ satm + act
## Model 2: gpa ~ satm + satv + act
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    268 48.713
## 2    267 47.517  1    1.1964 6.7223 0.01005 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gpa.lm4 <- lm(gpa ~ satv + satm, data = GPA)
anova(gpa.lm4, gpa.lm)
```

```
## Analysis of Variance Table
##
## Model 1: gpa ~ satv + satm
## Model 2: gpa ~ satm + satv + act
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    268 48.828
## 2    267 47.517  1    1.3112 7.3675 0.007074 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.) Model Comparison Test:

Hypotheses:

H_0: $\beta_1 + \beta_2 = \beta_3$

H_a: $\beta_1 + \beta_2 \neq \beta_3$

where:

$\Omega : E(Y) = \beta_0 + \beta_1 x_{satm} + \beta_2 x_{satv}$

$\omega : E(Y) = \beta_0 + \beta_3 x_{SAT}$

Anova Test:

```
GPA <- GPA %>%
  mutate(SAT = satv + satm)

gpa.lmboth <- lm(gpa ~ satm + satv, data = GPA)

gpa.lmSAT <- lm(gpa ~ SAT, data = GPA)

anova(gpa.lmSAT, gpa.lmboth)
```

```
## Analysis of Variance Table
##
## Model 1: gpa ~ SAT
## Model 2: gpa ~ satm + satv
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    269 49.081
## 2    268 48.828  1   0.25255 1.3862 0.2401
```

p-value and conclusion: With a p-value of 0.2401, we fail to reject the null hypothesis that $\beta_1 + \beta_2 = \beta_3$. Since our p-value is high, we cannot say that there is a difference between the model with SAT as a single score and the model with the separate subscores.

c.)

```
new_data <- data.frame(satv = 550, satm = 650)
conf_int <- predict(gpa.lmboth, newdata = new_data,
        interval = 'confidence',
        level = 0.95)
conf_int
```

```
##        fit     lwr      upr
## 1 3.253518 3.17607 3.330967
```

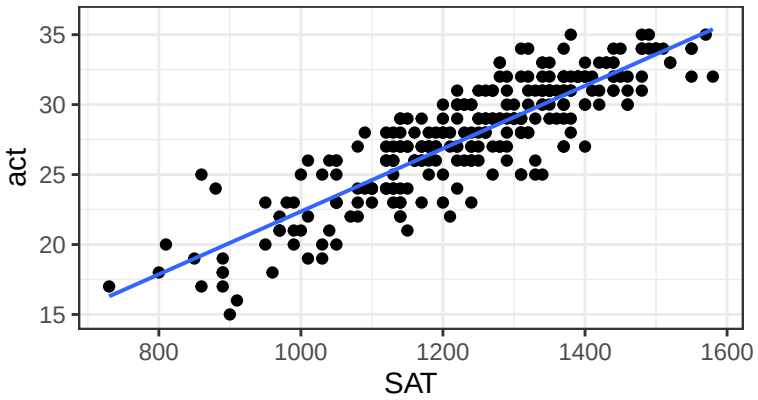
A 95% confidence interval for the mean GPA of students who have SATV and SATM scores of 550 and 650 respectively is (3.17607, 3.330967).

d.)

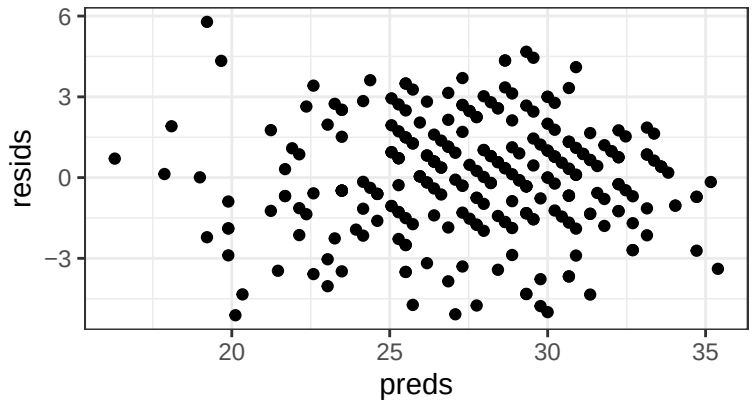
```
actSAT.lm <- lm(act ~ SAT, data = GPA)
msummary(actSAT.lm)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1040247  0.9652404  -0.108    0.914
## SAT          0.0224646  0.0007733  29.051   <2e-16 ***
##
## Residual standard error: 2.073 on 269 degrees of freedom
## Multiple R-squared:  0.7583, Adjusted R-squared:  0.7574
## F-statistic: 843.9 on 1 and 269 DF,  p-value: < 2.2e-16
```

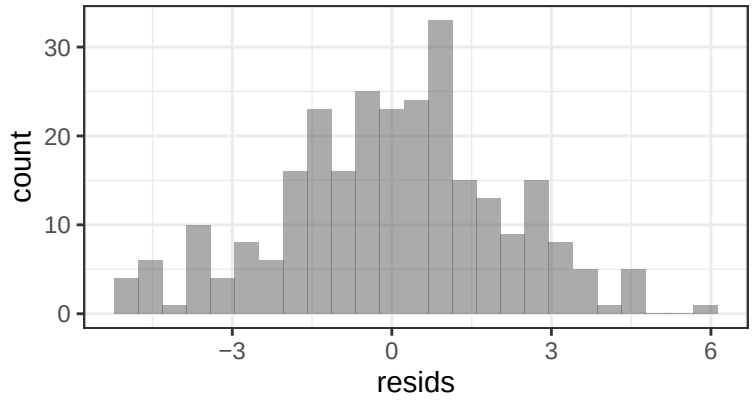
```
GPA <- GPA %>%
  mutate(preds = predict(actSAT.lm), resids = resid(actSAT.lm))

gf_point(act ~ SAT, data = GPA) %>%
  gf_lm()
```



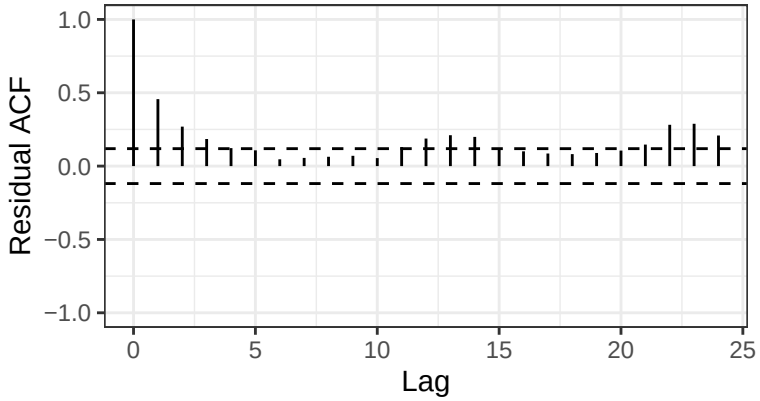

```
gf_point(resids ~ preds, data = GPA)
```




```
gf_histogram(~resids, data = GPA)
```




```
s245::gf_acf(~actSAT.lm) %>%
  gf_lims(y = c(-1, 1))
```



Our model seems to pass the conditions with some leniency. The scatterplot shows that the relationship does appear to be linear, so it makes sense to fit a linear model. The scatterplot of residuals vs. predictions is scattered randomly so the constant residual variance condition is passed. The histogram of residuals is normal as well. The independence condition, checked by the ACF plot, appears to pass with some leniency.

The linear model equation obtained from the lm():

$$y_{act} = -.1040247 + 0.0224646x_{SAT} + \epsilon$$

The best way to estimate uncertainty is to create a confidence or prediction interval.

New formula using CI or PI: In the following code, replace "SAT = 1350" with the SAT score of your choice.

If you want to predict the ACT score of a particular student based on their SAT score, change "confidence" to "prediction". If you would like to get the mean of ACT scores based on the SAT score you provided, leave it as is. You can also change the 95% confidence interval from 0.95 to a confidence level of your choice.

```
new_data <- data.frame(SAT = 1350)
conf_int <- predict(actSAT.lm, newdata = new_data,
        interval = 'confidence',
        level = 0.95)
conf_int
```

```
##        fit      lwr      upr
## 1 30.22324 29.92199 30.52449
```

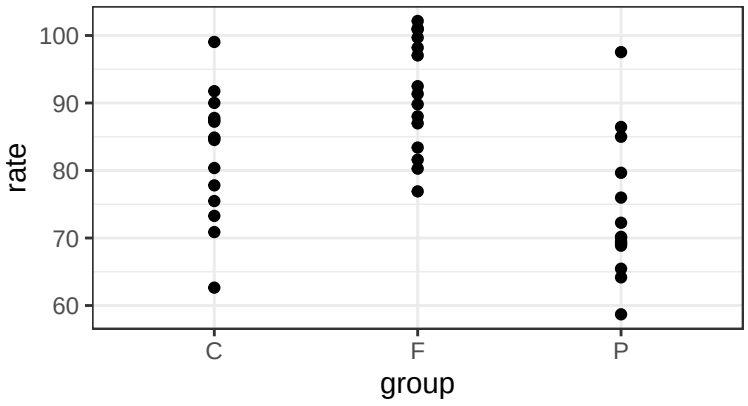
Problem 7.25

a.)

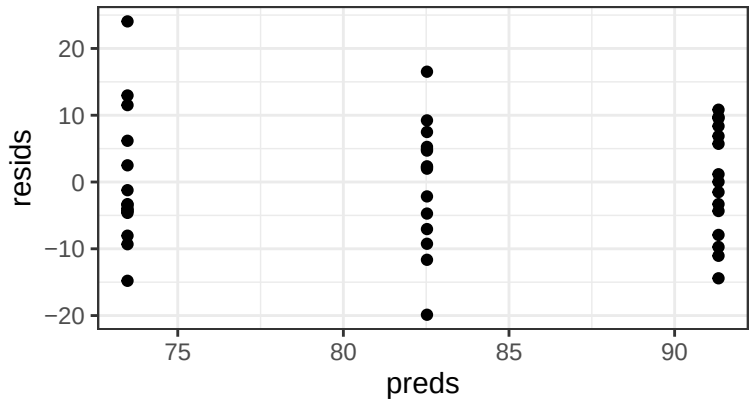
```
PetStress <- PetStress
m1 <- lm(rate ~ group, data = PetStress)
msummary(m1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.524      2.378  34.709   <2e-16 ***
## groupF         8.801      3.362   2.617   0.0123 *
## groupP        -9.041      3.362  -2.689   0.0102 *
##
## Residual standard error: 9.208 on 42 degrees of freedom
## Multiple R-squared:  0.4014, Adjusted R-squared:  0.3729
## F-statistic: 14.08 on 2 and 42 DF,  p-value: 2.092e-05
```

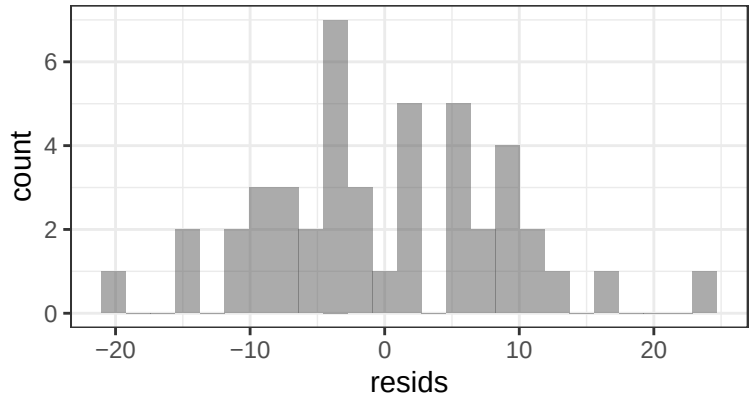
```
PetStress <- PetStress %>%
  mutate(preds = predict(m1), resids = resid(m1))

gf_point(rate ~ group, data = PetStress) %>%
  gf_lm()
```



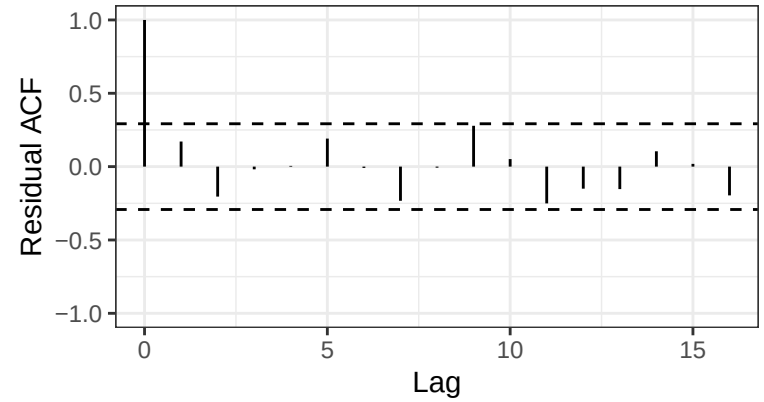
```
gf_point(resids ~ preds, data = PetStress)
```




```
gf_histogram(~resids, data = PetStress)
```




```
s245::gf_acf(~m1) %>%
  gf_lims(y = c(-1, 1))
```



The scatter plot of rate and group appears to be a linear trend so it makes sense to fit a linear model. The histogram of residuals is normal, so the normality of residuals condition is passed. The ACF plot shows that the model passes the independence condition. The scatter plot of residuals and predictions is randomly scattered so it passes the error variance condition. Since the model passes LINE conditions, it is appropriate to proceed with a model utility test.

b.) Hypotheses:

Null Hypothesis: $\beta_1 = 0$, The predictor group does not change anything, and the intercept only model is the

same.

Alternate Hypothesis: $\beta_1 \neq 0$, The predictor group has an affect on the rate, and the intercept only model is different.

where:

$\Omega : E(Y) = \beta_0 + \beta_1 x_{group}$

$\omega : E(Y) = \beta_0$

```
car::Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: rate
##           Sum Sq Df F value    Pr(>F)
## group     2387.7  2  14.079 2.092e-05 ***
## Residuals 3561.3 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a low p-value of 2.092e-05, we reject the null hypothesis that the group has no affect on rate. Therefore, we have evidence to say that the intercept only model is different and the full model with group as a predictor is a good model.

Problem 7.46

Backwards Step-wise Selection

```
require(faraway)
```

```
## Loading required package: faraway
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:mosaic':
##
##     ilogit, logit
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
uswages <- uswages

mod <- lm(wage ~ educ + exper + race + smsa + ne + mw + so + we + pt, data = uswages)

car::Anova(mod)
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
```

```
## Anova Table (Type II tests)
##
## Response: wage
##              Sum Sq   Df F value    Pr(>F)
## educ       38320052    1 225.649 < 2.2e-16 ***
## exper      26872523    1 158.240 < 2.2e-16 ***
```

```
## race        1946924    1  11.464 0.0007232 ***
## smsa        4808776    1  28.317 1.146e-07 ***
## ne                     0
## mw                     0
## so                     0
## we                     0
## pt         18819536    1 110.820 < 2.2e-16 ***
## Residuals 338114663 1991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the original model, ne is zero in the Anova test, so we should remove it. Remove ne

```
mod2 <- lm(wage ~ educ + exper + race + smsa + mw + so + we + pt, data = uswages)

car::Anova(mod2)
```

```
## Anova Table (Type II tests)
##
## Response: wage
##               Sum Sq   Df  F value    Pr(>F)
## educ       38320052    1 225.6490 < 2.2e-16 ***
## exper      26872523    1 158.2398 < 2.2e-16 ***
## race        1946924    1  11.4645 0.0007232 ***
## smsa        4808776    1  28.3166 1.146e-07 ***
## mw             9238    1   0.0544 0.8156060
## so             3112    1   0.0183 0.8923327
## we           631096    1   3.7162 0.0540279 .
## pt         18819536    1 110.8195 < 2.2e-16 ***
## Residuals 338114663 1991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next the largest large p-value is so: Remove so

```
mod3 <- lm(wage ~ educ + exper + race + smsa + mw + we + pt, data = uswages)
car::Anova(mod3)
```

```
## Anova Table (Type II tests)
##
## Response: wage
##               Sum Sq   Df  F value    Pr(>F)
## educ       38351594    1 225.9461 < 2.2e-16 ***
## exper      26874280    1 158.3282 < 2.2e-16 ***
## race        1964866    1  11.5759 0.0006813 ***
## smsa        4842872    1  28.5315 1.027e-07 ***
## mw            22978    1   0.1354 0.7129652
## we           804622    1   4.7404 0.0295796 *
## pt         18820785    1 110.8815 < 2.2e-16 ***
## Residuals 338117775 1992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, the largest large p-value from the model is mw: Remove mw

```
mod4 <- lm(wage ~ educ + exper + race + smsa + we + pt, data = uswages)
car::Anova(mod4)
```

```
## Anova Table (Type II tests)
##
## Response: wage
##              Sum Sq   Df  F value    Pr(>F)
## educ       38328908    1 225.9104 < 2.2e-16 ***
## exper      26933961    1 158.7486 < 2.2e-16 ***
## race        1943300    1  11.4538 0.0007273 ***
## smsa        4899804    1  28.8794 8.604e-08 ***
## we           980842    1   5.7811 0.0162903 *
## pt         18889261    1 111.3332 < 2.2e-16 ***
## Residuals 338140752 1993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since each predictor's p-value is small, each predictor now plays a significant role in the model. Our final model includes educ, exper, race, smsa, we, and pt as predictors for wage.